Genomics Proteomics Bioinformatics 19 (2021) 452–460

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

WEB SERVER

GranatumX: A Community-engaging, Modularized, and Flexible Webtool for Single-cell Data Analysis

David G. Garmire[1,#], Xun Zhu[2,#], Aravind Mantravadi[3], Qianhui Huang[4], Breck Yunits[2], Yu Liu[4], Thomas Wolfgruber[2], Olivier Poirion[2], Tianying Zhao[5], Cédric Arisdakessian[2,5], Stefan Stanojevic[4], Lana X. Garmire[4,*]

[1] *Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109, USA*
[2] *Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA*
[3] *Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48105, USA*
[5] *Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA*

Received 24 June 2021; accepted 24 September 2021
Available online 30 December 2021

Handled by Wei Lin

Abstract We present GranatumX, a next-generation software environment for single-cell RNA sequencing (scRNA-seq) data analysis. GranatumX is inspired by the interactive webtool Granatum. GranatumX enables biologists to access the latest scRNA-seq bioinformatics methods in a web-based graphical environment. It also offers software developers the opportunity to rapidly promote their own tools with others in customizable pipelines. The architecture of GranatumX allows for easy inclusion of plugin modules, named Gboxes, which wrap around bioinformatics tools written in various programming languages and on various platforms. GranatumX can be run on the cloud or private servers and generate reproducible results. It is a community-engaging, flexible, and evolving software ecosystem for scRNA-seq analysis, connecting developers with bench scientists. GranatumX is freely accessible at http://garmiregroup.org/granatumx/app.

KEYWORDS Single-cell RNA sequencing; Analysis; Pipeline; Webtool; Module

Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have advanced our understanding of cell-level biology significantly [1]. Many exciting scientific discoveries are attributed to new experimental technologies and sophisticated computational methods [2,3]. Despite the progress in cultivating professionals with cross-discipline training, a gap continues to exist between the wet-lab biology and the bioinformatics community. Moreover, with the rapid development of many varieties of modules handling different parts of scRNA-seq analysis [4–6], it becomes increasingly challenging for bioinformaticians themselves to decide which method to choose. Although some analytical packages such as SINCERA [7], Seurat [8], and Scanpy [9] provide complete scRNA-seq pipelines, they require users to be familiar with their corresponding programming language (typically R or Python), installation platform, and command-line interface. This overhead hinders wide

*Corresponding author.
E-mail: lgarmire@med.umich.edu (Garmire LX).

[#] Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2021.07.005

adoption by experimental biologists, especially those newly adopting scRNA-seq technologies. A few platforms, such as ASAP [10] and our own previous tool Granatum [11], provide intuitive graphical user interfaces (GUIs) and may be useful for a first-hand exploratory check. However, Granatum does not allow for modularity, while ASAP lacks flexibility and restricts the user to a set number of computational tools. Here we present GranatumX, the new generation of scRNA-seq analysis platform that aims to solve these issues systematically. Its architecture facilitates the rapid incorporation of cutting-edge tools and enables the efficient handling of large datasets aided by virtualization [12].

Method

Architectural overview

GranatumX consists of three independent components: central data storage (CDS), user interface (UI), and task runner (TR). CDS stores all data and metadata in GranatumX, including the uploaded files, processed intermediate data, and final results. The other two components of GranatumX both have controlled access to CDS, which allows them to communicate with each other. CDS is implemented using a PostgreSQL database and a secure file system-based data warehouse. UI is the component with which wet-lab biologists interact. The layout is intuitive with Gbox settings while providing a flexible and customizable analysis pipeline. UI also allows for the asynchronous submission of tasks before they can be run by the backend. UI is implemented using JavaScript, with the ReactJS framework. The submitted jobs queue up in the database and can be retrieved in real time by TR. TR monitors the task queue in the CDS in real time, actively retrieves the high-priority tasks (based on submission time), initializes the corresponding Gboxes, and prepares the input data by retrieving relevant data from CDS.

Deployment

GranatumX uses Docker to ensure that all Gboxes can be installed reproducibly with all their dependencies. As a result, GranatumX can be deployed in various environments including personal computers (PCs), dedicated servers, high-performance computing (HPC) platforms, and cloud services. The installation instructions are detailed in the README file of the source code.

Responsive UI

The web-based UI offers different device-specific layouts to suit a wider range of screen sizes. On desktop computers, the UI takes advantage of the screen space and uses a panel-based layout, and maximizes the on-screen information. On small tablets and mobile devices with limited screen space, a collapsible sidebar-based layout is used to allow the most important information (the results of the current step) to show up on the screen.

Recipe system

Most studies can use similarly structured pipelines, which typically consist of data entry (upload and parsing), data processing (imputation, gene filtering, normalization, *etc*.), and finally data analysis functionalities (clustering, differential expression and marker gene identification, pseudotime construction, *etc*.). GranatumX allows users to save a given pipeline into a "recipe" for the future. GranatumX comes with a set of built-in recipes, which cover many of the most common experiment pipelines.

Software development kits

Software development kits (SDKs) in GranatumX are made for Python and R. These SDKs provide a set of application programming interfaces (APIs) and helper functions that connect Gbox developer's own code with the core of GranatumX. The detailed documentation can be found in the Github repository.

There are three steps to build a new Gbox from the existing code. 1) An entry point is written in the language of the developer's choice. The entry point uses the SDK to retrieve necessary input from the core of GranatumX and send back output to the core after the results are computed. 2) The entry point, the original package source code, and any dependencies are packed into a Docker image using a Docker file and the "docker build" command. 3) A UI specification is written for the Gbox. The specification is a simple YAML file that declares the data requirements of the Gbox.

Pipeline customization

GranatumX allows for full customization of the analysis pipeline. An analysis pipeline has a number of Gboxes organized in a series of steps. Note that two different steps can have the same underlying Gbox. For example, two principal component analysis (PCA) Gboxes can appear before and after imputation, to evaluate its effect. Because the data are usually processed in a streamlined fashion, later steps in the pipeline usually depend on data generated by the earlier steps. Steps can be added from the App store into the current project and can be removed from the pipeline at any time. A newly added step can be inserted at any point in the pipeline and can be reordered in any way, as long as such re-arrangement does not violate the dependency relationships.

Current GranatumX cloud server setup

The current GranatumX web version is hosted on OVH-cloud, with specs: Intel Haswell vCPU 128 GB RAM Xeon E5-1650 4GHz. Additionally the https protocol is verified with Let's Encrypt (https://letsencrypt.org) with an Apache 2 server (https://httpd.apache.org/) and a site registered with No-IP (https://www.noip.com/). This server uses a proxy implementation to pass a user to the Node.js web service. In this manner, Node does not have to manage the security or https connections which allows setup to occur efficiently in an enterprise system. Additionally, an optional fast compute system may be connected to the OVH cloud server through ssh tunneling which allows the local port to be mapped to the remote connection. In this manner, a high-speed rig can be connected—in this case, the AMD 3590x can be connected without having to procure a new cloud system.

Project management

The studies in GranatumX are organized as projects. Each user can manage multiple concurrent projects. The automatic customer's report can be generated per project using the parameters and results stored in the CDS.

Example datasets

Three datasets are used in this report. One dataset is downloaded from Gene Expression Omnibus (GEO: GSE117988), a study on a patient with metastatic Merkel cell carcinoma, treated using T cell immunotherapy as well as immune-checkpoint inhibitors (anti-PD1 and anti-CTLA4) but later developed resistance [13]. A second dataset is *Tabula Muris* dataset, which contains 54,865 cells from 20 organs and tissues of mouse [14]. Another dataset is the 1.3 Million Brain Cells from E18 Mice, downloaded from 10x Genomics website: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons (accessed on date 05/09/2020). This dataset contains 1,308,421 cells from embryonic mouse brains, done by Chromium™ Single Cell 3′ Solution (v2 Chemistry).

GranatumX plugin development

The detailed instruction document and the tutorial YouTube or Youku videos for writing Gbox plugin are on the project website: http://garmiregroup.org/granatumx/app. Additionally, we created a slack group named "GranatumX-Developer" to facilitate plugin development from the 3rd party.

Results

Overview of GranatumX

The objective of GranatumX is to provide scRNA-seq biologists better access to bioinformatics tools and the ability to conduct single-cell data analysis independently (**Figure 1**). Currently other scRNA-seq platforms usually only provide a fixed set of methods implemented by the authors themselves. It is difficult to add new methods developed by the community due to programming language lock-in as well as monolithic code architectures. If a pipeline is assembled between heterogeneous tools, it is manually crafted and inhibits a repeatable execution of data analysis tools by other wet-lab scientists. As a solution, GranatumX uses the plugin and virtualized framework that provides an easy and unified approach to add new methods in a data-analysis pipeline. The plugin system is agnostic to developer code and the choice of the original scripting language. It also eliminates inter-module incompatibilities, by isolating the dependencies of each module (**Figure 2**A). As a data portal, GranatumX provides a GUI that requires no programming experience.

Deployment of GranatumX

The web-based GUI can be accessed on various devices including desktops, tablets, and smartphones (Figure 2A). In addition to the web-based format, GranatumX is also deployable on a broad variety of computational environments, such as PCs, cloud services, servers, and HPC platforms with minimal effort by system administrators. The deployment process is unified on all platforms because all components of GranatumX are containerized in Docker [15] (also portable to Singularity [16]). GranatumX can handle larger-scale scRNA-seq datasets coming online, with an adequate cloud configuration setup and appropriate Gboxes. For example, after uploading data, it took GranatumX ~ 12 min to finish the recommended pipeline with xxx modules on an AMD 3950x with 16 cores and 128 GB of DRAM memory running Ubuntu 20.04, using 10,000 cells downsampled from the dataset of "1.3 Million Brain Cells from E18 Mice" on the 10x Genomics website. The most time-consuming step is imputation using neural-network model DeepImpute (~ 2/5 time), and the detailed breakdown of time consumption is shown in Table S1.

Unique Gbox modules

Gbox is a unique concept of GranatumX. It represents a containerized version of a scientific package that handles its input and output by a format understood by the GranatumX core (Figure 2B). GranatumX has a set of pre-installable Gboxes that enable complete scRNA-seq analysis out of the

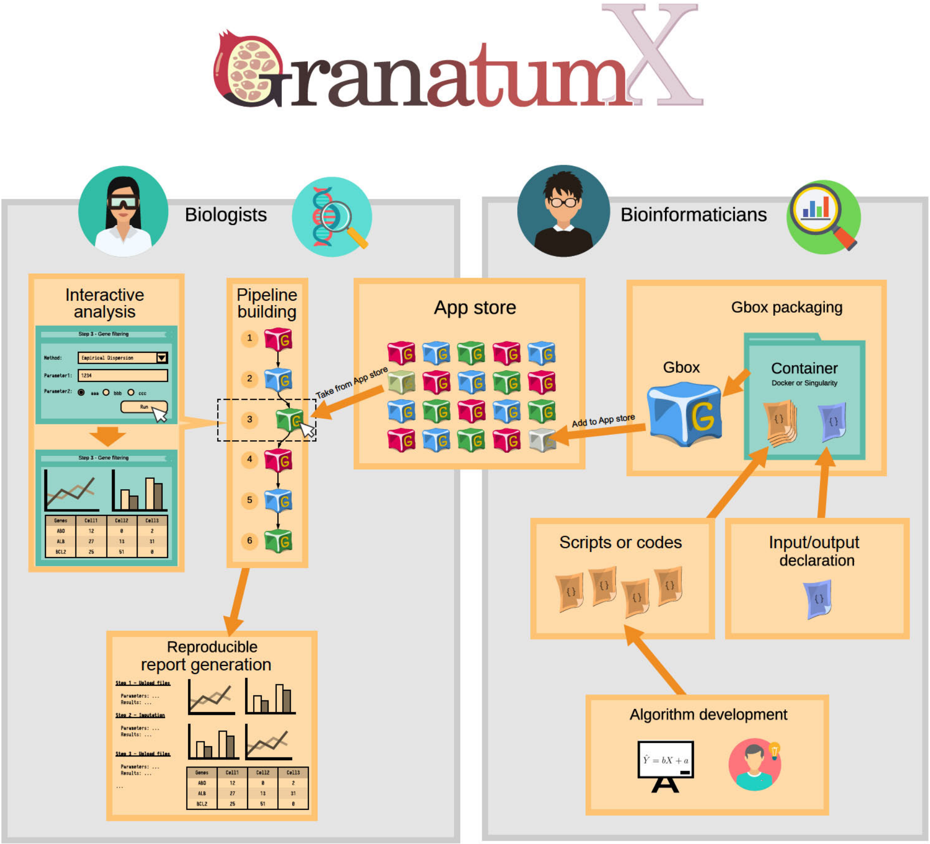

Figure 1 Overview of the Granatum X platform

Granatum X aims to bridge the gap between the computational method developers (the bioinformaticians) and the experiment designers (the biologists). It achieves this by building end-to-end infrastructure including the packaging and containerization of the codes (Gbox packaging), organization and indexing of the Gboxes (Apps), customization of the analysis steps (pipeline building), visualization and result downloading (interactive analysis), and finally the aggregation and summarization of the study (report generation).

box. Various Gboxes for data entry, processing, and analysis can be customized and organized together, to form a complete analysis pipeline (Figure 2C). One highlight feature of the Gbox is that it stands alone, and the user can assume any Gbox without the need to restart the full pipeline, in case one implemented by the user fails. Another highlight of the Gbox feature is that the entire GranatumX platform is fully interactive, with addition or removal of some Gboxes or parameter changes on the go, while some other Gboxes are being executed.

A comprehensive set of over 30 Gboxes are implemented in GranatumX to perform tasks all the way from data entry and processing to downstream functional analysis. The data processing tasks help to minimize the biases in the data and increase the signal-to-noise ratio. For each of these quality improvement categories, GranatumX provides multiple popular methods from which users can pick. To assist functional analysis, GranatumX provides a core list of methods for dimension reduction, visualization [including PCA, t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP)], clustering, differential expression, marker gene identification, gene set enrichment analysis (GSEA), network analysis, and pseudo-time construction. Versioning for each of these Gboxes has been implemented so that users can use a specific tested version of a Gbox. Developers on the other hand can work on newer versions separately before the official upgrade. Gboxes can be stored on Docker Hub for public use which maintains its own versioning system (https://hub.docker.com/u/granatumx). Detailed step-by-step tutorials for writing and building Gboxes are on GranatumX website http://garmiregroup.org/granatumx/app.

Input files

The input files of GranatumX include expression matrices and optional sample metadata tables, acceptable in a variety of formats such as CSV, TSV, or Excel format. GranatumX even accepts zip files and GNU zip (gz) files, which the user can choose for large expression matrices. Expression matrices are raw read counts for all genes (rows) in all cells (columns). The sample metadata tables annotate each cell with a pre-assigned cell type, state, or other quality information. The parsing step creates a sparse matrix using

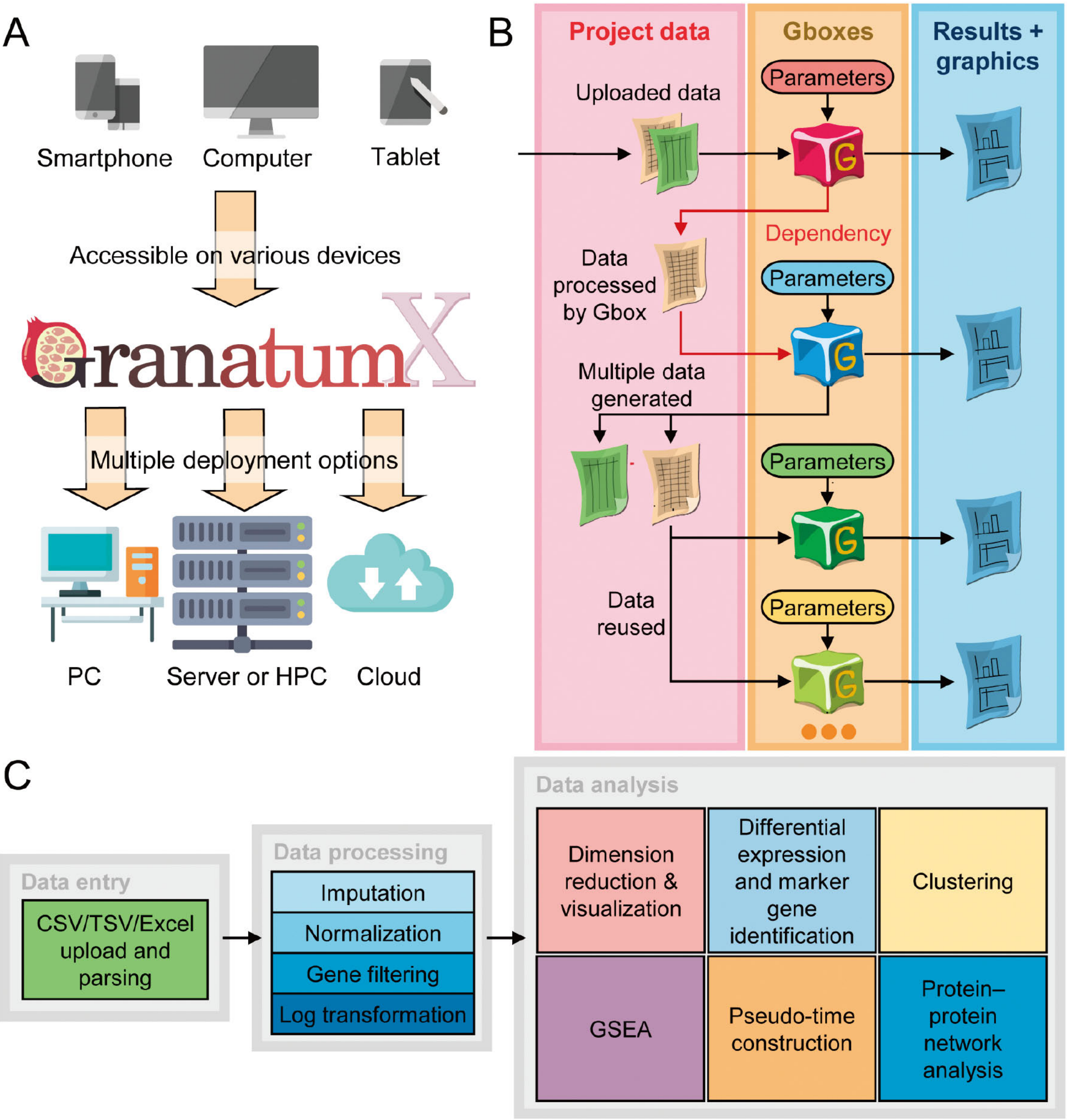

Figure 2 GranatumX deployment, data management, and analysis flow
A. Granatum X can be deployed on various computational environments, from PCs, servers, HPC systems, to cloud services. Granatum X's web UI is adaptable to devices with various screen sizes, which allows desktop and mobile access. **B.** Granatum X's data management. Each Gbox (labeled by a particular color to represent a certain functionality) with order dependency on the pipeline, may take some project data and some user-specified parameters as input and may generate results (interactive visualization, plots, tables, or even plain text) and new project data. All project data and results, as well as the specified parameters, are recorded and saved into the CDS and can be used for reproducibility control. **C.** A scRNA-seq computational study typically consists of three phases: the upload and parsing of the expression matrices and metadata (data entry), the quality improvement and signal extraction of the data (data processing), and finally the assorted analyses on the processed data which offer biological insights (data analysis). PC, personal computer; HPC, high-performance computing; UI, user interface; CDS, central data storage; GSEA, gene set enrichment analysis.

the coordinate list (COO) format, and this representation ensures swift upload onto the backend, even for large input datasets (> 10,000 cells). Such information will either be used to generate computational results (such as GSEA or be mapped onto the PCA, t-SNE, or UMAP plot for visualization (see Figure 2C for the workflow). Once the user uploads the gene expression matrix, the data are read into a dataframe using *Pandas*, and the step updates the user with a "preview", consisting of the first few rows and columns of the gene expression matrix, along with the number of genes and samples present.

User-centric design

As a user-friendly tool, GranatumX allows multiple users to be affiliated with the same project for data and result sharing, while restricting one user to run the pipeline at a time to avoid data conflicts. It allows dynamically adding, removing, and reordering pipeline steps on the go. It also allows users to reset the current step. All relevant data in the analysis pipeline and all results generated by each module are stored in a database, allowing users to access and download them. To ensure reproducibility, GranatumX can automatically generate a human-readable report detailing the inputs, running arguments, and the results of all steps (see examples in Files S1 and S2). All of these features are designed with the mindset of "consumer reports" to facilitate research in experimental labs or genomics cores.

Case studies using GranatumX

In the following section, we demonstrate two case studies of GranatumX. The first dataset was downloaded from Gene Expression Omnibus (GEO: GSE117988), including 7431 single cells generated by the 10x Genomics 3′ Chromium platform. It was obtained from a patient with metastatic Merkel cell carcinoma, treated using T cell immunotherapy as well as immune-checkpoint inhibitors (anti-PD1 and anti-CTLA4) but later developed resistance [13]. We used a customized pipeline to analyze the scRNA-seq data (**Figure 3**A). The pipeline comprises all common analysis steps, including 1) file upload, 2) imputation (based on DeepImpute [6]), 3) cell normalization, 4) gene filtering, 5) log transformation, 6) PCA, 7) t-SNE/UMAP visualization, 8) clustering, 9) sample coloring, 10) marker gene identification, 11) GSEA, and 12) pseudo-time construction. The analysis report of the entire pipeline is included as File S1. The clustering step identifies 7 clusters on the UMAP plot (Figure 3B). The exemplary GSEA results (**Table 1)** show the significance in many important immune-related pathways, including the MAPK signaling pathway and antigen processing and presentation pathway (cluster 5 *vs*. rest), cell cycle genes (cluster 3 *vs*. rest), and ubiquitin-mediated proteolysis (cluster 3 *vs*. rest).

We also used GranatumX to analyze *Tabula Muris* dataset, which contains 54,865 cells from 20 organs and tissues of mouse [14]. Again, we used the same pipeline as shown in Figure 3A. GranatumX offers multiple popular clustering algorithms, and for this dataset we used the Louvain algorithm. For illustration purposes, we focus on the viewing and clustering of this large graph-based clustering method implemented by Scanpy. A total of 44 clusters are assigned on the UMAP plot (Figure 3C). We also superimposed the metadata that contain tissue types for each cell on the same plot for visualization (Figure S1). The complete analysis report of this dataset is included as File S2.

Discussion

With the ever-increasing popularity of scRNA-seq, more and more experimental biologists will adopt this technology. At the same time, new bioinformatics tools are being developed rapidly. The development of GranatumX fills in a unique niche that enables both scientific and technical advancements. It is a "common ground" that connects scRNA-seq tool developers with the end-users, together for new discoveries. Domain experts can use GranatumX for the initial exploratory analysis. Additionally, with more Gboxes to be implemented on model performance metrics, GranatumX could be a vessel to enable benchmark studies to compare existing computational modules and pipelines, as well as assess the performance of a new method or pipeline relative to the existing ones. Moreover, it can also serve as the test engine to probe the source of variations in different modules, so as to optimize a pipeline for given datasets.

Table 1 GSEA results on clusters from UMAP plot in Figure 3B

Comparing pair	KEGG gene set name	Gene set size	NES	*P* value	FDR
Cluster 2 *vs*. rest	Glycolysis gluconeogenesis	13	4.23	0	0
	Pathogenic *Escherichia coli* infection	14	3.63	0	0
	Alzheimer's disease	20	3.97	0	0
	Tight junction	16	3.03	0.004	0.0063
Cluster 3 *vs*. rest	Oocyte meiosis	15	3.75	0	0
	Pathogenic *Escherichia coli* infection	14	3.48	0.001	0.0315
	Cell cycle	20	3.13	0.002	0.042
	Ubiquitin-mediated proteolysis	12	3.29	0.005	0.0787
Cluster 4 *vs*. rest	Spliceosome	13	3.84	0	0
	Viral myocarditis	25	3.23	0.002	0.042
Cluster 5 *vs*. rest	Alzheimer's disease	20	3.08	0	0
	Antigen processing and presentation	30	2.51	0.003	0.0472
	MAPK signaling pathway	45	2.91	0	0
	Glycolysis gluconeogenesis	13	2.49	0.043	0.198
	Spliceosome	13	2.99	0.004	0.0504

Note: GSEA, gene set enrichment analysis; UMAP, uniform manifold approximation and projection; KEGG, Kyoto Encyclopedia of Genes and Genomes; NES, normalized enrichment score; FDR, false detection rate.

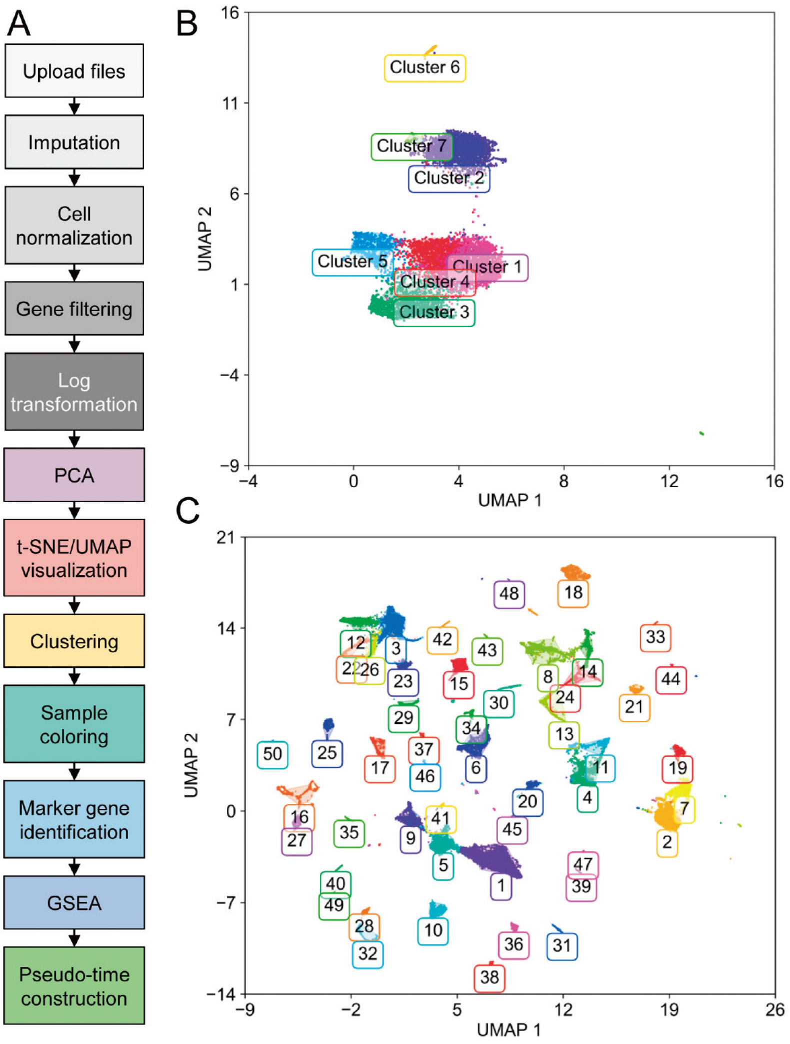

Figure 3 Case studies using an exemplary workflow of GranatumX

A. An exemplary workflow of a customized scRNA-seq pipeline set by the user. **B.** UMAP plot showing clusters on metastatic Merkel cell carcinoma data from the 10x Genomics platform [13]. **C.** UMAP plot showing clusters of *Tabula Muris* Consortium data [14]. PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

To demonstrate the uniqueness of GranatumX, we also compare it with other similar tools for comprehensive scRNA-seq analysis, such as SC1 [17], ASAP [10] and Single Cell Explorer [18] in **Table 2.** While all these tools aim for simple report and interaction with biologists, GranatumX is the only framework that supports bioinformatics developers to contribute their own plugins (Table 2). This significantly enhances the adaptability of GranatumX among the developer community. The web-tool that is closest to GranatumX is Single Cell Explorer, still with significant differences in the functionalities. Single Cell Explorer begins from raw data processing including reading mapping alignment. GranatumX is a much lighter-weight tool that starts from a cell read count table since the alignment/tag counting step is readily done by the popular Cell Ranger software of the 10x Genomics platform. Instead, GranatumX put more efforts on downstream analysis, such as gene enrichment analysis, protein-protein interaction and pseudo-time construction (Table 2). For ASAP, besides lacking modules to perform functions such as imputation, protein-protein interaction and pseudo-time construction, it also does not allow reconfiguration of the pipeline like GranatumX. SC1 lacks the flexibility and functionalities similar to ASAP and is restricted by Shiny, an R programming language-based web-interface, whereas GranatumX accepted containerized Gboxes packaged written in R, Python, or other languages.

As an inclusive and open software environment that employs other third-party tools, GranatumX has some challenges. One of them is handling the upgrade of underlying 3rd party libraries and resources. Accompanying with updated 3rd party tools which may not be tested extensively by the original developers, errors from these packages may propagate into GranatumX. To deal with this issue, we

Table 2 **Comparison on multiple user-friendly webtools**

Feature	Platform			
	SC1	ASAP	Single Cell Explorer	GranatumX
Simple report and interactivity for biologists	Yes	Yes	Yes	Yes
Configurable* pipeline	No	No	Yes	Yes
Supporting computational developers to plug in their own containers	No	No	No	Yes
Programming languages allowed in plug ins	NA	NA	NA	Multiple languages (*e.g.*, Python, R)
Default pipeline supporting imputation	No	No	No	Yes
Default pipeline supporting pseudo-time analysis	No	No	No	Yes
Supporting protein–protein interaction network	No	No	No	Yes

Note: * configurable refers to the ability to customize the analytical steps and orders. NA, not available.

implement versioning through the use of Docker which helps to maintain system-level dependencies as well as software dependencies in a complete package. The Gbox Docker containers for this release are listed in the Table S2 with a version number of 1.0.0. New versions can update the minor and major revision numbers so that users know exactly which code is being executed for a given pipeline. The source code for the Docker containers which represent the Gboxes are stored in the corresponding GitHub repositories. For example, https://hub.docker.com/r/granatumx/gbox-differentialexpression is stored in https://github.com/granatumx/gbox-differentialexpression. Such an endeavor provides safety and reliability in maintaining the stability of the software not just in the source of the software but in the configuration of the system required to run the computational elements. Due to its openness, GranatumX cannot prevent p-hacking or manipulating data analysis to improve the statistical significance of the desired result [19]. One way to discourage p-hacking is to suggest using standard pipeline and default parameters. If the user chooses values other than defaults, the reproducible design of GranatumX allows one to compare the outputs from the users (if they are recorded) with those from the default setting.

Conclusion

We present an open-source, shareable and evolvable single cell analysis tool called GranatumX. It not only enables the domain experts to independently conduct single cell analysis, but also promotes bioinformatics tool developers to contribute and develop their own single cell analysis methods through Gbox plugin setup. We hope that GranatumX will engage the single cell analysis community broadly and continuously for scientific discoveries.

Code availability

The webtool of GranatumX can be found at http://garmiregroup.org/granatumx/app. On this website, users can also find YouTube or Youku tutorial videos that demonstrate how to use GranatumX webtool. The source code for GranatumX is available at https://github.com/granatumx under MIT license. All builds are deployed via Docker Hub at https://hub.docker.com/u/granatumx.

CRediT author statement

David G. Garmire: Software, Methodology, Visualization, Writing - original draft, Writing - review & editing, Investigation, Validation, Resources, Data curation, Supervision. **Xun Zhu:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Validation, Writing - original draft, Data curation, Visualization. **Aravind Mantravadi:** Software, Investigation. **Qianhui Huang:** Investigation, Data curation, Validation. **Breck Yunits:** Software, Validation, Data curation. **Yu Liu:** Software, Validation, Data curation. **Thomas Wolfgruber:** Software, Validation, Data curation. **Olivier Poirion:** Software, Validation, Data curation. **Tianying Zhao:** Software, Validation, Data curation. **Cédric Arisdakessian:** Software, Validation, Data curation. **Stefan Stanojevic:** Investigation, Validation, Writing - review & editing. **Lana X. Garmire:** Conceptualization, Methodology, Formal analysis, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors have no potential competing interests.

Acknowledgments

We thank Dr. Larry Reiter for providing valuable feedback while using GranatumX to test his single-cell dataset. This research was supported by grants from the National Institute of Environmental Health Sciences (NIEHS) through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov; Grant No. K01ES025434), the National Institutes of Health / National Institute of General Medical Sciences (NIH/NIGMS; Grant No. P20 COBRE GM103457), the National Library of Medicine (NLM; Grant No. R01 LM012373), and the National Institute of Child Health and Human Development (NICHD; Grant No. R01 HD084633) awarded to LXG. Cartoon icons in Figures 1 and 2 are downloaded from https://www.flaticons.com/.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.07.005.

ORCID

0000-0002-6913-6550 (David G. Garmire)
0000-0001-9097-7824 (Xun Zhu)
0000-0001-9217-913X (Aravind Mantravadi)
0000-0002-3292-7519 (Qianhui Huang)
0000-0001-7033-3538 (Breck Yunits)
0000-0002-8631-4347 (Yu Liu)
0000-0001-8770-1800 (Thomas Wolfgruber)
0000-0002-0429-7003 (Olivier Poirion)
0000-0002-0103-1867 (Tianying Zhao)
0000-0001-5255-0942 (Cédric Arisdakessian)
0000-0002-9692-5301 (Stefan Stanojevic)
0000-0002-4654-2126 (Lana X. Garmire)

References

[1] Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 2014;42:8845–60.

[2] Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 2018;14:e1006245.

[3] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 2018;13:599–604.

[4] Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq data. Genomics Proteomics Bioinformatics 2021;19:267–81.

[5] Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics and computational challenges. Front Genet 2016;7:163.

[6] Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol 2019;20:211.

[7] Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol 2015;11:e1004575.

[8] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.

[9] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.

[10] David FPA, Litovchenko M, Deplancke B, Gardeux V. ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. Nucleic Acids Res 2020;48:W403–14.

[11] Zhu X, Wolfgruber TK, Tasato A, Arisdakessian C, Garmire DG, Garmire LX. Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. Genome Med 2017;9:108.

[12] Felter W, Ferreira A, Rajamony R, Rubio J. An updated performance comparison of virtual machines and Linux containers. 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) 2015. https://doi.org/10.1109/ispass.2015.7095802.

[13] Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, Perdicchio M, et al. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. Nat Commun 2018;9:3868.

[14] The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. Nature 2018;562:367–72.

[15] Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014. http://dl.acm.org/citation.cfm?id=2600239.2600241 (accessed June 22, 2021).

[16] Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One 2017;12:e0177459.

[17] Moussa M, Mandoiu II. SC1: a web-based single cell RNA-seq analysis pipeline. 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) 2018. https://doi.org/10.1109/iccabs.2018.8542088.

[18] Feng D, Whitehurst CE, Shan D, Hill JD, Yue YG. Single Cell Explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data. BMC Genomics 2019;20:676.

[19] P-HACKING. Oxford Online Dictionary n.d. https://www.lexico.com/definition/p-hacking (accessed June 22, 2021).