GranatumX: A community engaging and flexible software environment for single-cell analysis

Xun Zhu[1], Breck Yunits[1], Thomas Wolfgruber[1], Yu Liu[2] , Qianhui Huang[3], Olivier Poirion[1], Cédric Arisdakessian[1,4], Tianying Zhao[4], David Garmire[5] , Lana Garmire[2 *]

[1]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA.

[2]Deparment of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, 48105, USA.

[3]Deparment of Biostatistics, University of Michigan, Ann Arbor, 48109, USA.

[4]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA.

[5]Deparment of Electrical & Computer Engineering, University of Michigan, Ann Arbor, 48105, USA.

[*] To whom correspondence should be addressed. Email address: lgarmire@med.umich.edu

Abstract

We present GranatumX, the next-generation software environment for single-cell data analysis. It enables biologists access to the latest single-cell bioinformatics methods in a 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", that wrap around bioinformatics tools written in various programming languages. GranatumX can be run in the cloud or private servers, and generate reproducible results. It is expected to become 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-Seq, analysis, software, pipeline, plugin

Main

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 on both sides, it has become obvious that an increasingly larger gap exists between the wet-lab biology and the bioinformatics

community. Although some analytical packages such as SINCERA [4], Seurat [5], and Scanpy [6] provide complete scRNA-Seq pipelines, they require users to be familiar with their corresponding programming language (typically R or Python) and/or command line interface, hindering a wide adoption experimental biologists. A few platforms, such as ASAP [7] and our own tool Granatum [8], provide an intuitive graphical user interface. However, these platforms are not modularized and lack the flexibility to incorporate a continuously growing list of new computational tools. Furthermore, these tools have limited scalability and cannot handle extremely large datasets. 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 handling of large datasets very efficiently.

The objective of GranatumX is to provide scRNA-Seq biologists better access to bioinformatics tools and ability to conduct single cell data analysis independently (Figure 1). Currently other single-cell RNA-Seq platforms usually only provide a fixed set of methods implemented by the authors themselves. Adding new methods developed by the community is difficult, due to programming language lock-in as well as monolithic code architectures. As a solution, GranatumX uses the plugin framework that provides an easy and unified approach to add new methods. The plugin system is developer code/scripting language agnostic. It also eliminates inter-module incompatibilities, by isolating the dependencies of each module (Figure 2A). As a data portal, GranatumX provides a graphical user interface (GUI) that requires no programming experience. Its web-based GUI can be accessed on various devices including desktop, 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 private PCs, cloud services, and High Performance Computing (HPC) platforms. The deployment process is unified on all platforms because all components of GranatumX are containerized in Docker [9] (also portable to Singularity [10]). GranatumX can handle larger-scale scRNA-seq datasets coming online, with an adequate cloud configuration setup and appropriate Gboxes. For example, it took GranatumX 14.5 minutes to finish the entire pipeline on a

Google Cloud with a 4 virtual CPUs and 60G memory, using 100K cells downsampled from the dataset of "1.3 Million Brain Cells from E18 Mice" on the 10x Genomics website.

Gbox is a unique concept of GrantumX, it represents a containerized version of a scientific package that handles its input and output by a format understood by the GrantaumX core (Figure 2B). GranatumX has a set of pre-installed Gboxes that enable complete scRNA-Seq analysis out of the box. Various Gboxes for data entry, preprocessing and processing together form a complete analysis pipeline (Figure 2C). The currently implemented Gboxes are listed in **Supplementary Table 1**. We also provide templates (Supplementary File 3) and tutorials for writing gboxes (Supplementary File 4). The input files of GranatumX include expression matrices and (optionally) sample metadata tables, acceptable in a variety of formats such as CSV, TSV, or Excel format. Expression matrices are raw read counts for all genes (rows) in all cells (columns). The sample metadata tables annotate each cell with pre-assigned cell type/state or other quality information. Such information will either be used to generate computational results (such as Gene Set Analysis), or be mapped onto PCA plot, t-SNE, or UMAP plot for visualization. A set of built-in modules are implemented to perform pre-processing tasks such as imputation and gene filtering. These 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 for users to choose. To assist functional analysis, GranatumX provides a comprehensive list of methods for dimension reduction, visualization (including PCA, t-SNE, and UMAP), clustering, differential expression and marker gene identification, Gene Set Enrichment Analysis, and pseudo-time construction.

As a user-friendly tool, GranatumX allows multiple users to create different projects, and it makes customizing and analyzing the results of workflows very simple. It allows dynamically adding/removing/reordering steps in a pipeline. All relevant data in the analysis pipeline and all results generated by each module, are stored in a database when deployed locally. These data can be accessed

and downloaded by users. To ensure reproducibility, GranatumX can automatically generate a human-readable report detailing the inputs, running arguments, and the results of all steps. All these features are designed with the mindset of "consumer reports" to facilitate research labs with multiple users or genomics cores. In the following section, we will demonstrate two case studies.

The first data set was downloaded from GSE117988, including 7431 single cells generated by 10x Genomics 3' Chromium platform. It was 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 [11]. We used the "Comprehensive pipeline" to analyze the scRNA-seq data (**Figure 3A**). The pipeline comprises all common analysis steps, including 1) File upload, 2) imputation (based on DeepImpute [12]), 3) normalization, 4) gene filtering, 5) log transformation, 6) principal component analysis (PCA), 7) t-SNE/UMAP plot, 8) sample coloring, 9) clustering, 10) marker gene identification, 11) GSEA analysis, and 12) pseudotime construction. The analysis report of the entire pipeline is included as **Supplementary File 1.** In the exemplary GSEA analysis results (**Figure 3B)**, many important immune-related pathways show significance, including the MAPK signaling pathway and antigen processing and presentation pathway (cluster_0 vs. rest), cell cycle genes (cluster_2 vs. rest), and ubiquitin mediated proteolysis (cluster_7 vs. rest).

To test the size of the data that GranatumX web version can handle with the default setup (Google Cloud Intel Haswell vCPU 64 GB RAM Xeon E5 2.4GHz), we next used it to analyze Tabula Muris data, which contain 54,865 cells from 20 organs and tissues [13]. Again we used the "Comprehensive pipeline" (**Figure 3A**). For illustration purpose, we focus on viewing and clustering of this large scRNA-Seq dataset. GranatumX offers multiple popular clustering algorithms, here we used Louvain graph-based clustering methods implemented by Scanpy. The 44 clusters assigned in this step are visualized and co-localized on

the UMAP plot (**Figure 3C**). We also imposed the metadata that contain tissue types for each cell on the same plot. The complete analysis report of the pipeline is included as **Supplementary File 2.**

With its 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 offers a unified software environment that enables many scientific and technical advancements. It is an ideal "common ground" that connects scRNA-seq tool developers with the end-users, to enable new discoveries. Additionally, with more Gboxes to be implemented on model performance metrics, GranatumX could also allow 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.

Methods

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 the central data storage, 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 asynchronous submission of tasks before they can be run by the back-end. 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 reproducibly installed with all their dependencies. As a result, GranatumX can be deployed in various environments including personal computers, 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 pre-processing (imputation, filtering, normalization, etc.), and finally data analysis functionalities (clustering, differential expression, pseudo-time construction, network analysis, 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 (SDKs)

GranaumX SDKs 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) Write an entry point 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) Package the entry point, the original package source code and any dependencies into a docker image using a Dockerfile and the "docker build" command. 3) Write a UI specification 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 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.

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.

Code availability

The webtool of GranatumX can be found at http://garmiregroup.org/granatumx/app. The source code for GranatumX is available at https://github.com/lanagarmire/granatumx under MIT license. The template for the gbox wrapper is provided as Supplementary File 3, and the tutorial for writing gboxes is documented in detail in Supplementary File 4.

Copyright

Some of the cartoon icons in Figure 1 and Figure 2 are downloaded from https://www.flaticon.com/.

Acknowledgement

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, R01 HD084633 awarded by NICHD to L.X. Garmire.

Figure legends

Figure 1: Overview of the Granatum X platform. Granatum X aims to bridge the gap between the computational method developers (the bioinformaticians) with the experiment designers (the biologists). It achieves this by building end-to-end infrastructure including the packaging and containerization of the code (**Gbox Packaging**), organization and indexing of the Gboxes (**App Store**), customization of the analysis steps (**Pipeline building**), visualization and results downloading (**Interactive Analysis**), and finally the aggregation and summarization of the study (**Report Generation**).

Figure 2: A) Due to its heavy usage of dependency locking and containerization, Granatum X can be deployed on various computational environments, from personal computers, private servers, High Performance Computation 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 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 central data storage, and can be used for reproducibility control. C) An scRNA-Seq computational study typically consists of three phases: the uploading 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).

Figure 3: A) The workflow of a customized scRNA-seq pipeline, called the comprehensive pipeline. B) T-SNE clustering plot and Gene Set Enrichment Analysis results on Merkel cell carcinoma data from 10x genomics platform. C) The clustering results of Tabula Muris Consortium data.

Supplementary Materials

Supplementary Table 1. The list of currently implemented Gboxes.

Supplementary File 1: The analysis report using a dataset with metastatic Merkel cell carcinoma from 10x genomics platform.

Supplementary File 2: The analysis report using a dataset with Tabula Muris Consortium data.

Supplementary File 3: The template for creating a new Gbox for GranatumX.

Supplementary File 4: The tutorial for writing a Gbox for GranatumX.

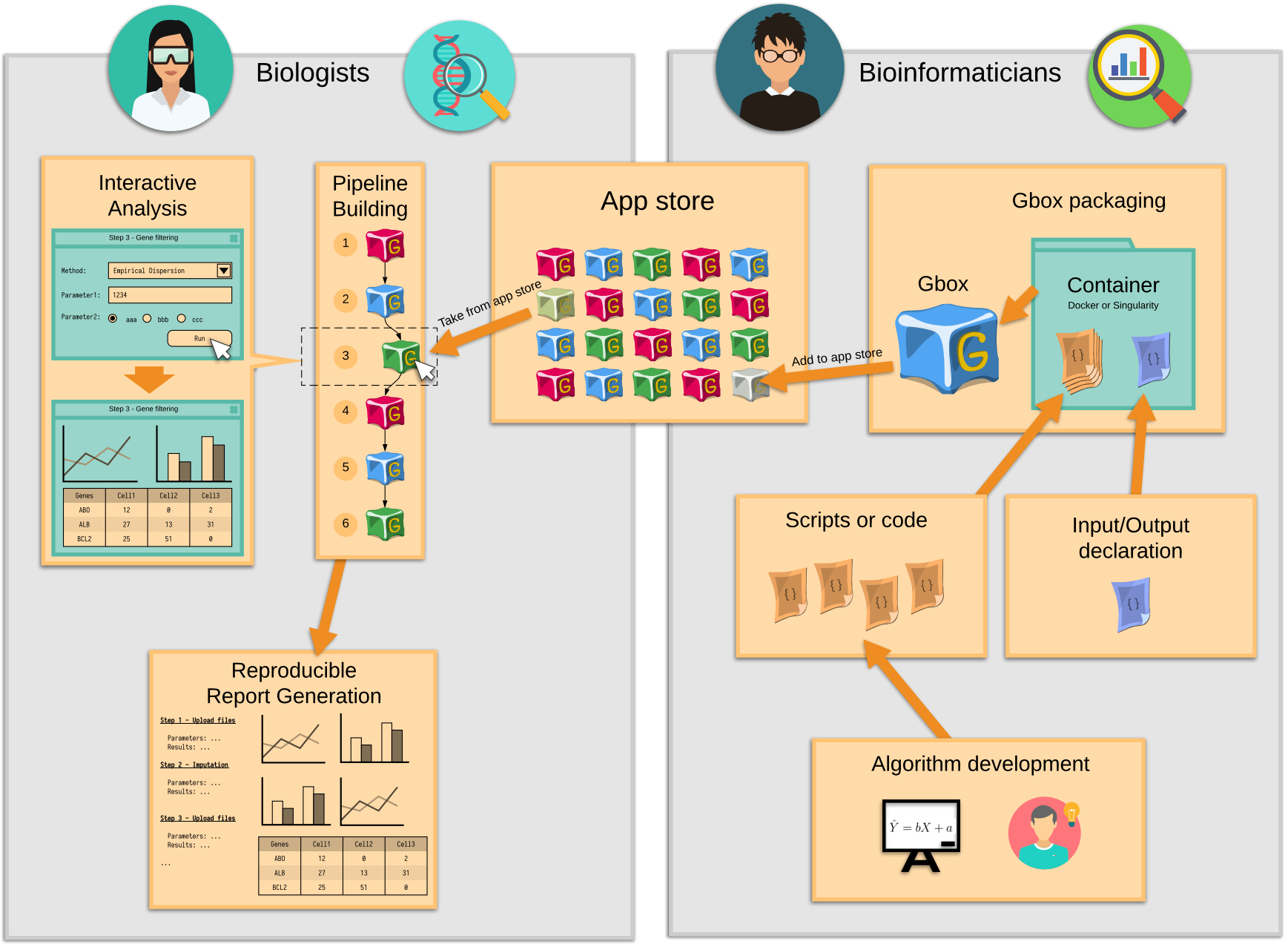
References

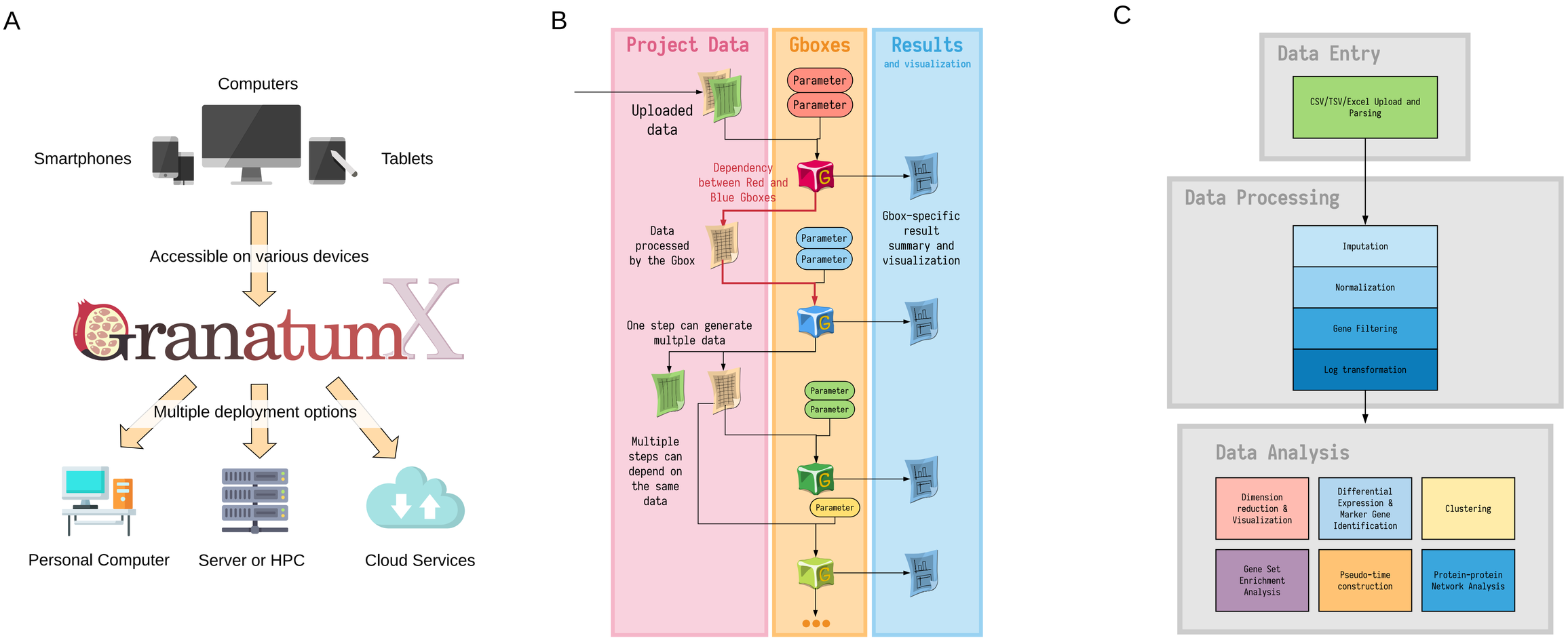
1. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).
2. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
3. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
4. Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 (2015).
5. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
6. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data

analysis. *Genome Biol.* **19**, 15 (2018).

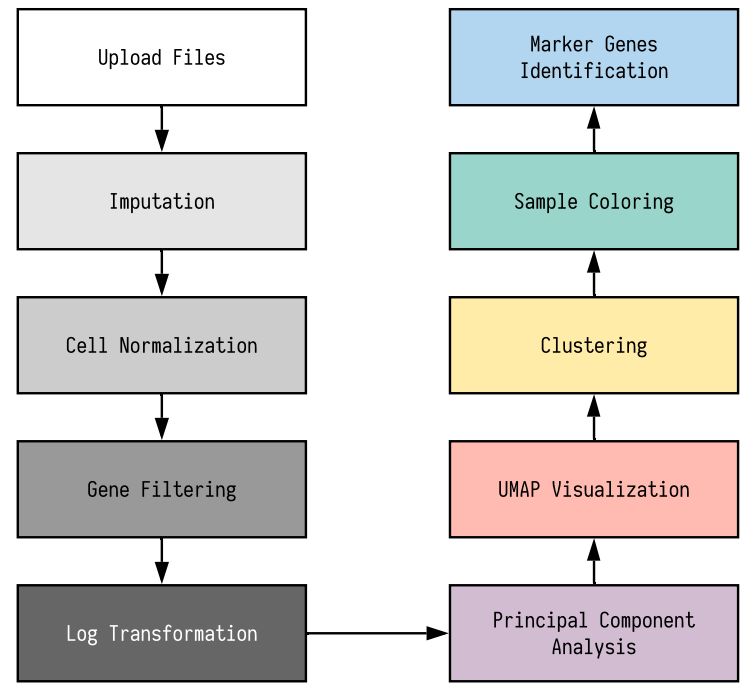
7. Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C. & Deplancke, B. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* **33**, 3123–3125 (2017).
8. Zhu, X. *et al.* Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* **9**, 108 (2017).
9. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.* **2014**, (2014).
10. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**, e0177459 (2017).
11. Paulson, K. G. *et al.* Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* **9**, 3868 (2018).
12. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. & Garmire, L. X. DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-Seq data. *bioRxiv* 353607 (2018). doi:10.1101/353607
13. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

GranatumX

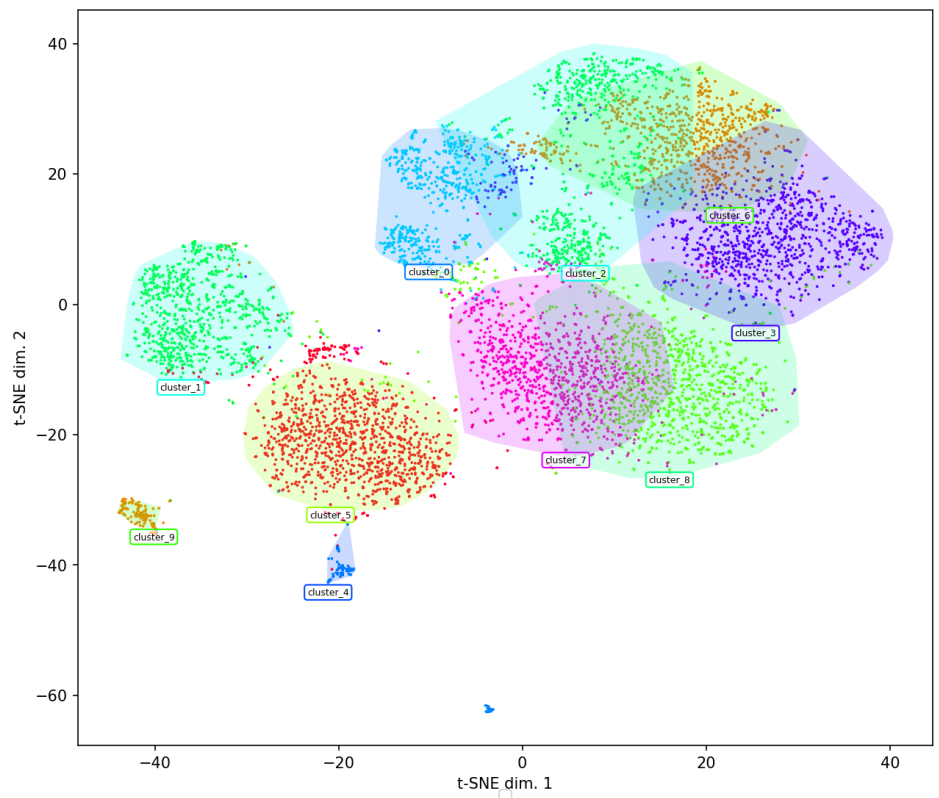

Biologists
Bioinformaticians
Interactive Analysis
Step 3 - Gene filtering
Method:
Empirical Dispersion
Parameter1:
1234
Parameter2:
aaa
bbb
ccc
Run
Step 3 - Gene filtering
Genes
Cell1
Cell2
Cell3
ABO
12
0
2
ALB
27
13
31
BCL2
25
51
0
Pipeline Building
1
2
3
4
5
6
Take from app store
App store
Add to app store
Gbox packaging
Gbox
Container
Docker or Singularity
Scripts or code
Input/Output declaration
Algorithm development
$\hat{Y} = bX + a$
Reproducible Report Generation
Step 1 - Upload files
Parameters: ...
Results: ...
Step 2 - Imputation
Parameters: ...
Results: ...
Step 3 - Upload files
Parameters: ...
Results: ...
...
Genes
Cell1
Cell2
Cell3
ABO
12
0
2
ALB
27
13
31
BCL2
25
51
0

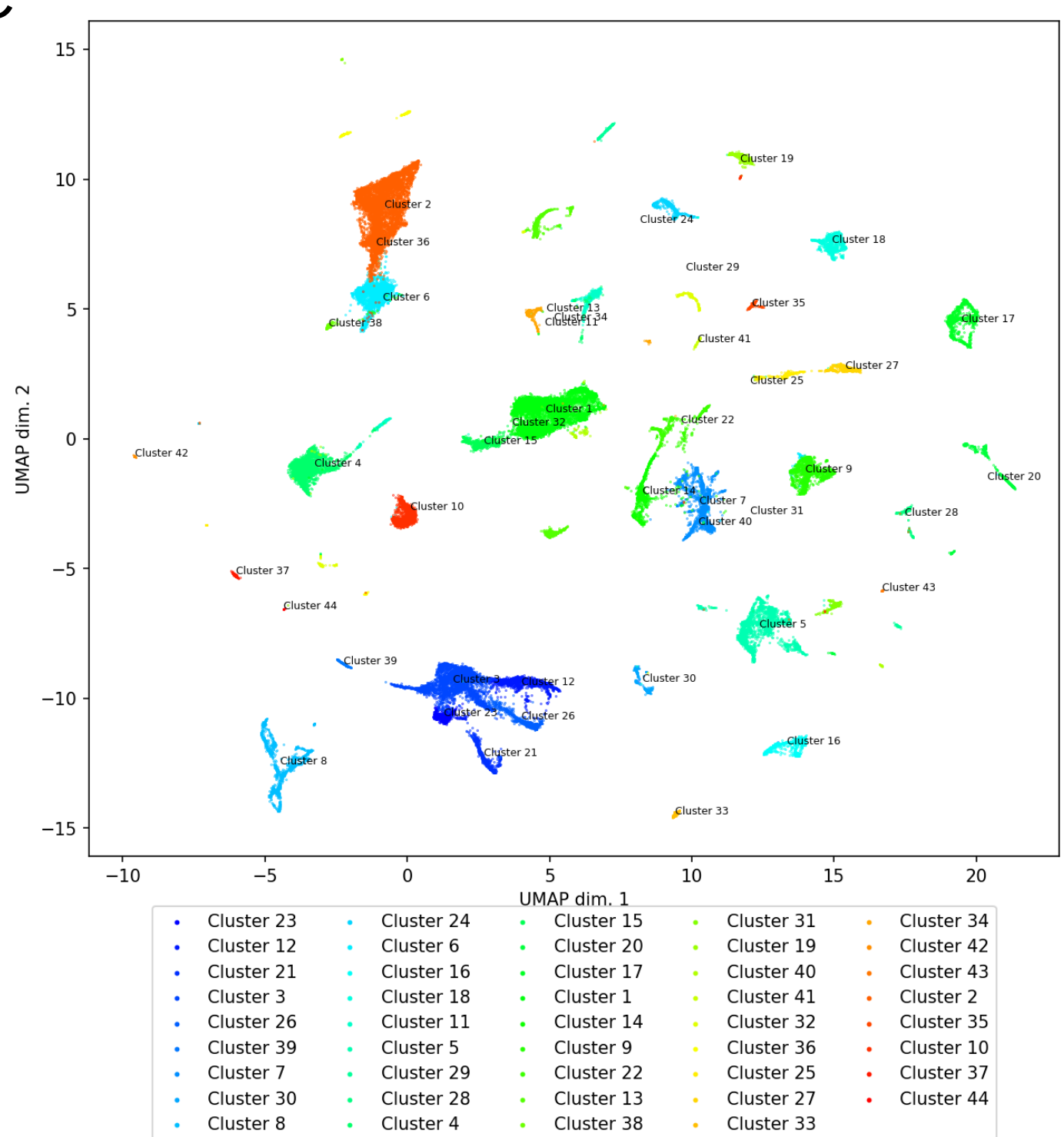

A
Computers
Smartphones
Tablets
Accessible on various devices
GranatumX
Multiple deployment options
Personal Computer
Server or HPC
Cloud Services
B
Project Data
Gboxes
Results
and visualization
Uploaded data
Parameter
Parameter
Dependency between Red and Blue Gboxes
Gbox-specific result summary and visualization
Data processed by the Gbox
Parameter
Parameter
One step can generate multple data
Parameter
Parameter
Multiple steps can depend on the same data
Parameter
C
Data Entry
CSV/TSV/Excel Upload and Parsing
Data Processing
Imputation
Normalization
Gene Filtering
Log transformation
Data Analysis
Dimension reduction & Visualization
Differential Expression & Marker Gene Identification
Clustering
Gene Set Enrichment Analysis
Pseudo-time construction
Protein-protein Network Analysis

A



B



Cluster_0 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_ALZHEIMERS_DISEASE	18	3.064463	0	0
KEGG_MAPK_SIGNALING_PATHWAY	46	2.65772	0.001	0.0315
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	28	2.720225	0.002	0.0315
KEGG_SPLICEOSOME	14	2.931969	0.002	0.0315
Cluster_2 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_OOCYTE_MEIOSIS	14	3.126715	0	0
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTIO N	14	2.938216	0	0
KEGG_CELL_CYCLE	21	2.901928	0.001	0.021
KEGG_LYSOSOME	23	2.705149	0.002	0.0315
Cluster_3 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	12	5.241768	0	0
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	24	4.054828	0	0
KEGG_TIGHT_JUNCTION	17	4.498274	0	0
Cluster_5 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTIO N	14	2.869346	0	0
Cluster_6 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_OOCYTE_MEIOSIS	14	3.819167	0	0
Cluster_7 vs. rest				
gset_name	gset_size	nes	p_val	fdr
KEGG_SPLICEOSOME	14	3.021011	0	0
KEGG_TYPE_I_DIABETES_MELLITUS	16	2.844379	0.001	0.0315
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	13	2.731589	0.003	0.042
KEGG_PRION_DISEASES	10	2.773909	0.003	0.042
KEGG_HUNTINGTONS_DISEASE	14	2.723391	0.004	0.042
KEGG_GRAFT_VERSUS_HOST_DISEASE	17	2.68223	0.004	0.042
KEGG_ALLOGRAFT_REJECTION	16	2.614118	0.005	0.045

C