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

Xun Zhu[1,2], Breck Yunits[1], Thomas Wolfgruber[1], Olivier Poirion[1], Cédric Arisdakessian[1,2], Lana Garmire[1,2] *

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

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

* To whom correspondence should be addressed. Email address: lgarmire@cc.hawaii.edu

Abstract

GranatumX is a next-generation software environment for single-cell data analysis. It offers biologists access to the latest single-cell bioinformatics methods, and software developers the opportunity to rapidly promote and combine their tools with various others in customizable pipelines. The architecture of GranatumX allows for easy inclusion of plugin modules made by bioinformatics tool developers. These modules can be flexibly arranged into customized workflows by biologists entirely in a graphical environment. Using novel modularized design with"Gbox" GranatumX enables seamless integration of bioinformatics tools written in R, Python, or any other language together. Granatum can be run in the cloud, private servers or mobile devices and generate reproducible results thanks to Gbox containerization. In summary, GranatumX is expected to become a community-engaging, flexible, and evolving software ecosystem for scRNA-Seq analysis, connecting developers with bench scientists. GranatumX is 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 wide adoption by many experimental biologists. A few platforms, such as ASAP [7] and Granatum [8], provide an intuitive graphical user interface. However, these platforms lack the flexibility to incorporate a continuously r-growing list of new computational tools. Furthermore, these tools have limited scalability and cannot handle extremely large datasets. Here we present GranatumX, an scRNA-Seq analysis platform that aims to solve these issues systematically. Its architecture facilitates the rapid incorporation of cutting-edge tools and enables 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. In contrast, GranatumX focuses on new tool incorporation, using a 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 and appropriate Gboxes. As an example, it took GranatumX 14.5 minuteswe to finish the entire pipeline on a Google Cloud with a 4 virtual CPUs and 60G memory, using a downsampled

100K cells from the dataset of "1.3 Million Brain Cells from E18 Mice" provided by 10x Genomics website.

Gbox is a unique software concept of GrantumX, it represents a containerized version of a scientific package that handles its input and output in a format that is understood by the GrantaumX core (Figure 2B). It enables scRNA-Seq analysis in GrantumX out of the box. Various Gboxes for data entry, preprocessing and processing form a complete analysis pipeline (Figure 2C). The input files of GranatumX include expression matrices and (optionally) sample metadata tables, 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 or t-SNE plot for visualization. A set of built-in modules are implemented to perform tasks such as imputation, normalization 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 existing methods for users to choose. Additionally, GranatumX provides a comprehensive list of methods for dimension reduction and visualization (including PCA and t-SNE), clustering, differential expression and marker gene identification, Gene Set Enrichment Analysis, and pseudo-time construction.

GranatumX makes customizing and analyzing the results of workflows simple. Along with the above major features, GranatumX provides the flexibility of 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 revisited and downloaded for future references. As a complex software environment, GranatumX can create multiple concurrent projects for a single user, with each project having its data and steps organized independently. To ensure reproducibility, GranatumX can automatically generate a human-readable report detailing the inputs, running arguments, and results of all steps. An exampliary analysis report is provided as the Supplementary File 1, using a dataset with the Patient-derived Xenografts (PDX) [11] model from primary and metastasized lung cancer. All these features are desirable for research labs with multiple users, or genomics cores or which coordinate with customers' projects.

Code availability

The webtool of GranatumX can be found at http://garmiregroup.org/granatumx/app.

Acknowledgement

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

Figures

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 device 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).

Supplemental File 1

An exampliary analysis report using a dataset with the Patient-derived Xenografts (PDX) models from primary and metastasized lung cancer.

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).

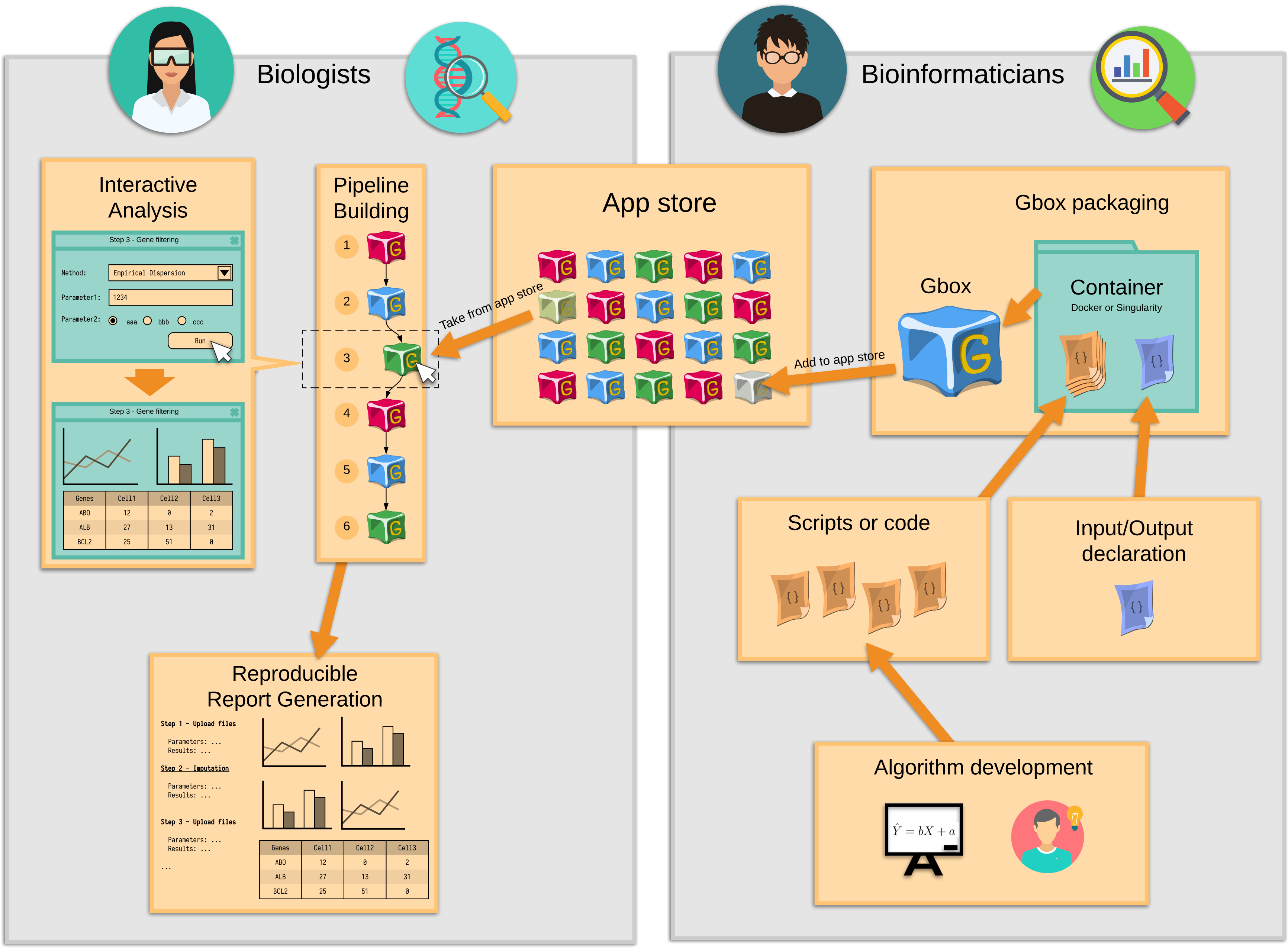
8. Zhu, X. *et al.* Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* **9,** 108 (2017).

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. Kim, K.-T. *et al.* Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* **17,** 80 (2016).

GranatumX



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

Step 1 - Upload files

Parameters: ...
Results: ...

Step 2 - Imputation

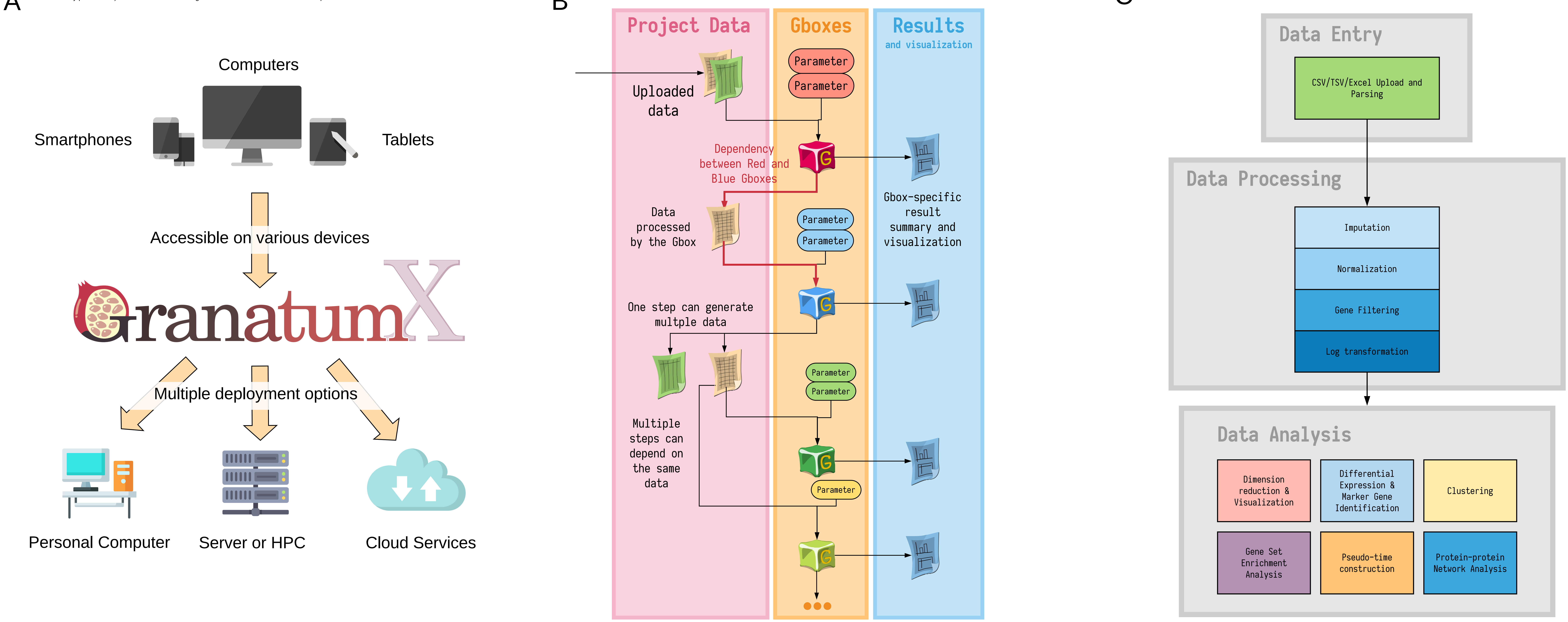
Parameters: ...
Results: ...

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
G
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