Zhu et al. Genome Medicine (2017) 9:108
DOI 10.1186/s13073-017-0492-3

Genome Medicine

SOFTWARE Open Access

Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists

Xun Zhu[1,2], Thomas K. Wolfgruber[1,2], Austin Tasato[3], Cédric Arisdakessian[1,2], David G. Garmire[3] and Lana X. Garmire[1,2*]

Abstract

Background: Single-cell RNA sequencing (scRNA-Seq) is an increasingly popular platform to study heterogeneity at the single-cell level. Computational methods to process scRNA-Seq data are not very accessible to bench scientists as they require a significant amount of bioinformatic skills.

Results: We have developed Granatum, a web-based scRNA-Seq analysis pipeline to make analysis more broadly accessible to researchers. Without a single line of programming code, users can click through the pipeline, setting parameters and visualizing results via the interactive graphical interface. Granatum conveniently walks users through various steps of scRNA-Seq analysis. It has a comprehensive list of modules, including plate merging and batch-effect removal, outlier-sample removal, gene-expression normalization, imputation, gene filtering, cell clustering, differential gene expression analysis, pathway/ontology enrichment analysis, protein network interaction visualization, and pseudo-time cell series construction.

Conclusions: Granatum enables broad adoption of scRNA-Seq technology by empowering bench scientists with an easy-to-use graphical interface for scRNA-Seq data analysis. The package is freely available for research use at http://garmiregroup.org/granatum/app

Keywords: Single-cell, Gene expression, Graphical, Normalization, Clustering, Imputation, Differential expression, Pathway, Pseudo-time, Software

Background

Single-cell high-throughput RNA sequencing (scRNA-Seq) is providing new opportunities for researchers to identify the expression characteristics of individual cells among complex tissues. From bulk cell RNA-Seq, scRNA-Seq is a significant leap forward. In cancer, for example, scRNA-Seq allows tumor cells to be separated from healthy cells [1], and primary cells to be differentiated from metastatic cells [2]. Single-cell expression data can also be used to describe trajectories of cell differentiation and development [3]. However, analyzing data from scRNA-Seq brings new computational challenges, e.g., accounting for inherently high drop-out or artificial loss of RNA expression information [4, 5].

Software addressing these computational challenges typically requires the ability to use a programming language like R [5, 6], limiting accessibility for biologists who only have general computer skills. Existing workflows that can be used to analyze scRNA-Seq data, such as Singular (Fluidigm, Inc., South San Francisco, CA, USA), Cell Ranger (10x Genomics Inc., Pleasanton, CA, USA), and Scater [7], all require some non-graphical interactions. They also may not provide a comprehensive set of scRNA-Seq analysis methods. To fill this gap, we have developed Granatum, a fully interactive graphical scRNA-Seq analysis tool. Granatum takes its name from the Latin word for pomegranate, whose copious seeds resemble individual cells. This tool employs an easy-to-use web browser interface for a wide range of methods suitable for scRNA-Seq analysis: removal of batch effects, removal of outlier cells, normalization of expression levels, imputation for dropout events, filtering of under-informative genes, clustering of cells, identification of differentially expressed genes, identification of

* Correspondence: LGarmire@cc.hawaii.edu
[1]Graduate Program in Molecular Biology and Bioengineering, University of Hawaii at Manoa, Honolulu, HI 96816, USA
[2]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA
Full list of author information is available at the end of the article

enriched pathways/ontologies, visualization of protein networks, and reconstruction of pseudo-time paths for cells. Our software empowers a much broader audience in research communities to study single-cell complexity by allowing the graphical exploration of single-cell expression data, both as an online web tool (from either computers or mobile devices) and as locally deployed software.

Implementation

Overview

The front-end and the back-end of Granatum are written in R [8] and built with the Shiny framework [9]. A load-balancer written in NodeJS handles multiple concurrent users. Users work within their own data space. To protect the privacy of users, the data submitted by one user is not visible to any other user. The front-end operates within dynamically loaded web pages arranged in a step-wise fashion. ShinyJS [10] is used to power some of the interactive components. It permits viewing on mobile devices through the reactivity of the Bootstrap framework. To allow users to redo a task, each processing step is equipped with a reset button. Bookmarking allows the saving and sharing of states.

Interactive widgets

Layout and interactivity for the protein–protein interaction (PPI) network modules is implemented using the visNetwork package [11]. Preview of user-submitted data and display of tabular data in various modules is implemented using DataTables [12]. The interactive outlier-identification step uses Plotly [13]. Scatter plots, box plots, and pseudo-time construction in Monocle are done by the ggplot2 package [3, 14].

Back-end variable management

The expression matrix and the metadata sheet are stored separately for each user. The metadata sheet refers to groups, batches, or other properties of the samples in the corresponding expression matrix. All modules share these two types of tables. Other variables shared across all modules include the log-transformed expression matrix, the filtered and normalized expression matrix, the dimensionally reduced matrix, species (human or mouse), and the primary metadata column.

Batch-effect removal

Batch effect is defined as the unwanted variation introduced in processing or sequencing in potentially different conditions [15]. To remove batch effects, we implement two methods in Granatum: ComBat and Median alignment.

ComBat

This method adjusts the batch effect using empirical Bayes frameworks, and is robust in the presence of outliers or for small sample sizes [16]. It is originally designed for batch-effect removal of microarray gene expression datasets but is commonly used in scRNA-Seq studies [17–19]. It is implemented by the "ComBat" function in the R package "sva" [20].

Median alignment

First, this method calculates the median expression of each sample, denoted as med_i for sample i. Second, it calculates the mean of med_i for each batch, denoted as $batchMean_b$ for batch b:

$$batchMean_b = geometricMean_{i \in batch_b}(med_i).$$

Finally, it multiplies each batch by a factor that pulls the expression levels towards the global geometric mean of the sample medians. When $i \in batch_b$ and m is the number of samples:

$$sample\_after_i = sample\_before_i \cdot \frac{geometricMean_{i \in 1,\ldots,m}(med_i)}{batchMean_b},$$

where $sample\_before_i$ and $sample\_after_i$ denote the expression levels for all genes within sample i before and after batch-effect removal.

Outlier detection and gene filtering

Z-score threshold is used to automatically detect outliers. The z-score of a cell is calculated by calculating the Euclidean norm of the cell's vector of expression levels, after scaling all genes to have unit standard deviation and zero mean [21]. Over-dispersion gene filtering is done as recommended by Brennecke et al. [4]. The output of the Monocle package [3] is modified to calculate dispersion and fit a negative binomial model to the result.

Clustering methods

The following description of clustering algorithms assumes that n is the number of genes, m is the number of samples, and k is the number of clusters.

Non-negative matrix factorization

The log-transformed expression matrix (n-by-m) is factorized into two non-negative matrices H (n-by-k) and W (k-by-m). The highest-valued k entry in each column of W determines the membership of each cluster [22, 23]. The non-negative matrix factorization (NMF) computation is implemented in the NMF R-package, as reported earlier [22, 24].

K-means

K-means is done on either the log-transformed expression matrix or the 2-by-*m* correlation t-SNE matrix. The algorithm is implemented by the *kmeans* function in R [25].

Hierarchical clustering

Hierarchical clustering (Hclust) is done on either the log-transformed expression matrix or the 2-by-*m* correlation t-SNE matrix. The algorithm is implemented by the *hclust* function in R [26]. The heatmap with dendrograms is plotted using the *heatmap* function in R.

Dimension reduction methods

Correlation t-SNE

The method assesses heterogeneity of the data using a two-step process. First, it calculates a distance matrix using the correlation distance. The correlation distance $D_{i,j}$ between sample i and sample j is defined as:

$$D_{i,j} = 1 - \text{Correlation}\left(S_i, S_j\right),$$

where S_i and S_j are the i-th and j-th column (sample) of the expression matrix. Next, Rtsne R package [27] uses this distance matrix to reduce the expression matrix to two dimensions.

PCA

The principal component analysis algorithm, implemented as "prcomp" function in R, decomposes the original data into linearly uncorrelated variables (components) using orthogonal transformation. The components are then sorted by their variance. The two components with the largest variances (PC1 and PC2) are extracted for visualization [28].

Elbow-point-finding algorithm in clustering

This method is inspired by a similar approach implemented in SCRAT [29]. In the clustering module with automatic determination of the number of clusters, the identification of the optimum number of clusters is done prior to presenting the clustering results. For each number of clusters $k = 2$ to $k = 10$, the percentage of the explained variance (EV) is calculated. To find the elbow-point $k = m$ where the EV plateaus, a linear elbow function is fit to the k-EV data points. This piecewise function consists of a linearly increasing piece from 0 to m, and a constant piece from m to 10. The algorithm iterates from $m = 1$ to 10 and identifies m which gives the best coefficient of determination (R^2) of linear regression as the "elbow point".

Differential expression analysis

We include four differential expression (DE) algorithms in Granatum: NODES [30], SCDE [31], EdgeR [32], and Limma [33]. Among them, NODES and SCDE are designed for scRNA-Seq specifically. EdgeR and Limma are conventional bulk cell RNA-Seq DE tools that have also been used in scRNA-Seq studies [34, 35]. When more than two clusters are present, we perform pairwise DE analysis on all clusters. We use default parameters for all packages. Their versions are: NODES (0.0.0.9010), SCDE (1.99.2), EdgeR (3.18.1) and Limma (3.32.2).

Gene set enrichment analysis

The *fgsea* R-package implements the gene set enrichment analysis (GSEA) algorithm with optimizations for speedup [36, 37]. GSEA calculates an *enrichment score*, which quantifies the relevance of a gene set (for example, a KEGG pathway or a Gene Ontology (GO) term) to a particular group of selected genes (e.g., DE genes called by a method). The p value is calculated for each gene set according to the empirical distribution, followed by Benjamini–Hochberg multiple hypothesis tests [38].

Pseudo-time construction

We use Monocle (version 2.2.0) in our pseudo-time construction step. When building the *CellDataSet* required for monocle's input, we set the *expressionFamily* to *negbinomial.size()*. We use *reduceDimension* function to reduce the dimensionality by setting *max_components* to 2.

Results

Overview of Granatum

Granatum is by far the most comprehensive graphic-user-interface (GUI)-based scRNA-Seq analysis pipeline with no requirement of programming knowledge (Table 1). It allows both direct web-based analysis (accessible through either desktop computers or mobile devices), as well as local deployment (as detailed in the front-page of http://garmiregroup.org/granatum/app). The project is fully open source, and its source code can be found at http://garmiregroup.org/granatum/code.

We have systematically compared Granatum with 12 other existing tools to demonstrate its versatile functions (Table 1). Popular packages such as SCDE/PAGODA and Flotilla are developed for programmers and require expertise in a particular programming language. In contrast, Granatum with its easy-to-navigate graphical interface requires no programming specialty. The current version of Granatum neatly presents nine modules, arranged as steps and ordered by their dependency. It starts with one or more expression matrices and corresponding sample metadata sheet(s), followed by data merging, batch-effect removal, outlier removal, normalization, imputation, gene filtering, clustering, differential expression, protein–protein network visualization, and pseudo-time construction.

Table 1 Comparison of existing single-cell analysis pipelines

Software	GUI driven workflow	Live web site	Video tutorial	Interactive plots	Batch-effect removal	Outlier removal	Normalization	Over-dispersed genes identification	Clustering analysis	Differential expression analysis	Gene-set enrichment analysis	Network analysis	Pseudo-time construction	Correcting for drop-outs	Spatial inference	Citation
Granatum	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	
SCRAT / TSCAN / GSCA	✓	(*)	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	Ji et al. 2017
ASAP	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗	✗	Gardeux et al. 2016
Sake	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	NA
Singular	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	Fluidigm Corp. 2015
Cell Ranger / Loupe	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	Zheng et al. 2017
Seurat	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	Satija et al. 2016
Scater	✗	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	McCarthy et al. 2016
Monocle	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	Trapnell et al. 2014
SCDE / PAGODA	✗	(**)	(***)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	Kharchenko et al. 2014
Flotilla	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	NA
Sincell	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗	✗	Juliá et al. 2015
Sincera	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	Guo et al. 2015

* The three components (SCRAT, TSCAN and GSCA) are not integrated.
** Results can be shown interactively using a web interface. However, the results themselves have to be pre-computed in R.
*** For the interactive interface only
Zheng et al. 2017 [60]; Satija et al. 2016 [61]; Juliá et al. 2015 [62]; Guo et al. 2015 [63]

Besides the features above, a number of enhanced functionalities make Granatum more flexible than other freely available tools (Table 1). (1) Unlike tools such as SCRAT (https://zhiji.shinyapps.io/scrat/), ASAP [39], and Sake (http://sake.mhammell.tools/), it is the only GUI pipeline that supports multiple dataset submission as well as batch effect removal. (2) Each step can be reset for re-analysis. (3) Certain steps (e.g., batch-effect removal, outlier removal, and gene filtering) can be bypassed without affecting the completion of the workflow. (4) Subsets of the data can be selected for customized analysis. (5) Outlier samples can be identified either automatically (by setting a pre-set threshold) or manually (by clicking/lassoing the samples from the PCA plot or the correlation t-SNE plot). (6) Multiple cores can be utilized in the differential expression module for speed-up. (7) Both GSEA and network analysis can be performed for the differentially expressed genes in all pairs of subgroups, following clustering analysis. (8) Pseudo-time construction is included, giving insights into relationships between the cells.

Testing of the software

In this report, we mainly use a previously published data set as an example [18]. This renal carcinoma dataset contains a total of 118 cells from three groups: patient-derived xenografts derived from the primary tumor (PDX primary), PDX metastatic cells, and patient metastatic cells [18]. We abbreviate this dataset as the K-dataset.

To estimate the total running time of Granatum (with default parameters) with different sizes of datasets, we first simulated expression matrices with 200, 400, 800, or 1600 cells using the Splatter package, based on the parameters estimated from the K-dataset [40]. Additionally, we also used a down-sample approach (200, 400, 800, 1600, 3200, and 6000 cells) on a dataset (*P*-dataset) provided by 10x Genomics, which comprises 6000 peripheral blood mononuclear cells (PBMCs; https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k). When the imputation step is not included, the running time scales linearly with the number of cells, regardless of platform (Additional file 1: Figure S1), among which Monocle based pseudo-

time construction is most time consuming (taking up 80% of all computing time).

There are ten main steps in Granatum (Fig. 1). In the following sections, we use the K-dataset to elaborate the details of each step in chronological order, unless notified otherwise.

Upload data

Granatum accepts one or more expression matrices as input. Each expression matrix may be accompanied by a *metadata sheet*. A metadata sheet is a table describing the groups, batches, or other properties of the samples in the corresponding expression matrix. Users may upload multiple matrices sequentially. Currently, Granatum accepts either human or mouse species for downstream functional analysis. After uploading the input files, users can preview the matrix and metadata tables to validate that the dataset is uploaded correctly.

Batch-effect removal

Samples obtained in batches can create unwanted technical variation, which confounds the biological variation [15]. It is therefore important to remove the expression level difference due to batches. Granatum provides a batch-effect removal step where two methods are included, namely ComBat [16] and median alignment. If multiple datasets are uploaded, by default, each dataset is assumed to be one batch. Alternatively, if the batch numbers are indicated in the sample metadata sheet, the user may select the column in which the batch numbers are stored. For datasets with a large number of cells, the box plot shows a random selection of 96 sub-samples for the visualization purpose and can be re-sampled freely.

To show that median alignment can effectively remove the batches, we randomly select half of the cells in K-dataset and multiply the expression levels by 3, thus creating two artificial batches 1 and 2. The PCA plot shows that, due to the batch effect, cells of the same type are separated by batch (the two colors; Fig. 2a). After performing median alignment, the batch effect is minimized, and cells from the same type but in two colors (batches) are now intermingled (Fig. 2b).

Outlier identification

Computationally abnormal samples pose serious problems for many downstream analysis procedures. Thus, it is crucial to identify and remove them in the early stage. Granatum's outlier identification step features PCA and t-SNE [41] plots, two connected interactive scatter plots that have different computational characteristics. A PCA plot illustrates the Euclidean distance between the samples, and a correlation t-SNE plot shows the associative distances between the samples. Granatum generates these two plots using top genes (default 500). Using the Plotly library [13], these plots are highly interactive. It is an example of thoughtful tool design that empowers users to explore the data. Outliers can be identified automatically by using a z-score threshold or setting a fixed number of outliers. In addition, each sample can be selected or de-selected by clicking, boxing, or drawing a lasso on its corresponding points.

The original K-dataset has one sample with an abnormally low expression level. This potential outlier sample

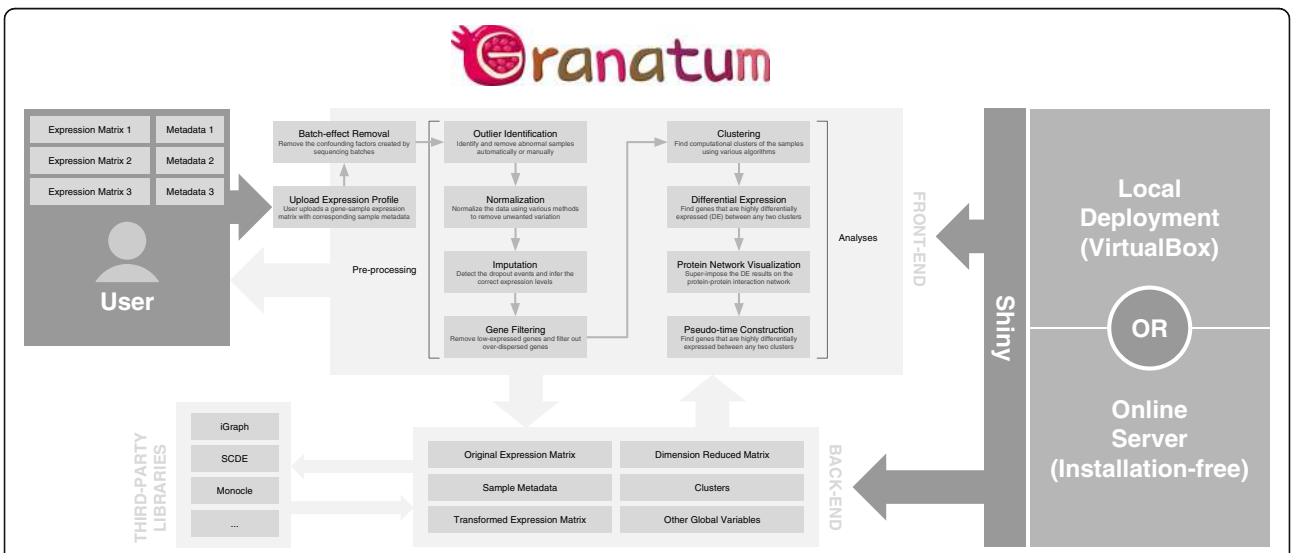

Fig. 1 Granatum workflow. Granatum is built with the Shiny framework, which integrates the front-end with the back-end. A public server has been provided for easy access, and local deployment is also possible. The user uploads one or more expression matrices with corresponding metadata for samples. The back-end stores data separately for each individual user, and invokes third-party libraries on demand

Fig. 2 Batch-effect removal. The PCA plots show the before/after median alignment comparison. The colors indicate the two batches 1 and 2, and the shapes indicate the three cell types reported from the original data. **a** Before batch-effect removal; **b** after batch-effect removal

can affect downstream analyses. Using Granatum, users can easily spot such outliers in the PCA plot or in the correlation t-SNE plot (Fig. 3a, b). After removal of the outliers, the top-gene-based PCA and correlation t-SNE plots are more balanced (Fig. 3c, d).

Normalization

Normalization is essential to most scRNA-Seq data before the downstream functional analyses (except those with the UMI counts). Granatum includes four commonly used normalization algorithms: quantile normalization, geometric mean normalization, size-factor normalization [42, 43], and Voom [44]. A post-normalization box plot helps illustrate the normalization effect to the median, mean, and extreme values across samples.

The box plots allow observation of various degrees of stabilization (Fig. 4). The original dataset has high levels of variation among samples (Fig. 4a). Quantile normalization unifies the expression distribution of all samples, thus rendering the box plots identical (Fig. 4b). Mean alignment tries to unify all means of the samples by multiplying the expression levels in each sample by a factor; thus, all means (the red dots) are visually the same (Fig. 4c). Size-factor and Voom normalization use more sophisticated procedures to normalize the data, but the variation of distribution across samples is evidently reduced (Fig. 4d, e). According to our and others' experience [45, 46], quantile normalization is recommended.

Imputation

A unique challenge in analyzing scRNA-Seq data is the dropout events, which introduce large number of false zeros in the expression matrix [4]. These erroneous zeros might affect many downstream analyses such as

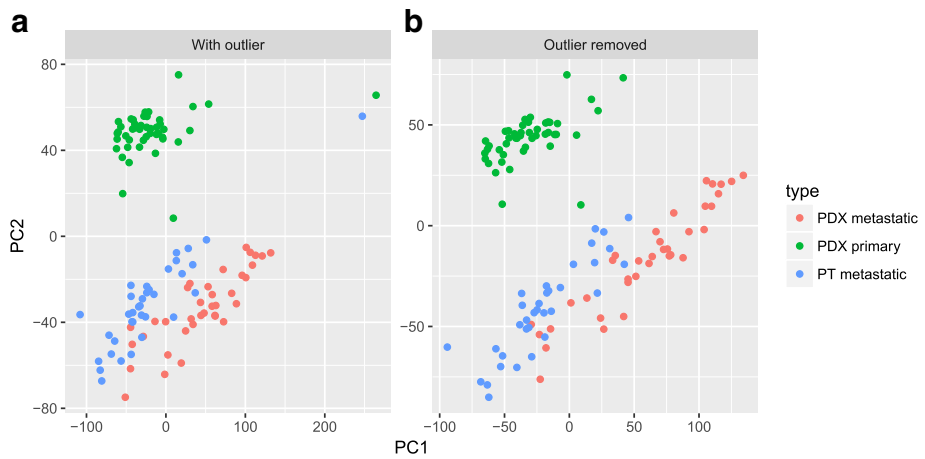

Fig. 3 Outlier removal using PCA plot. **a** Before outlier removal. **b** After outlier removal

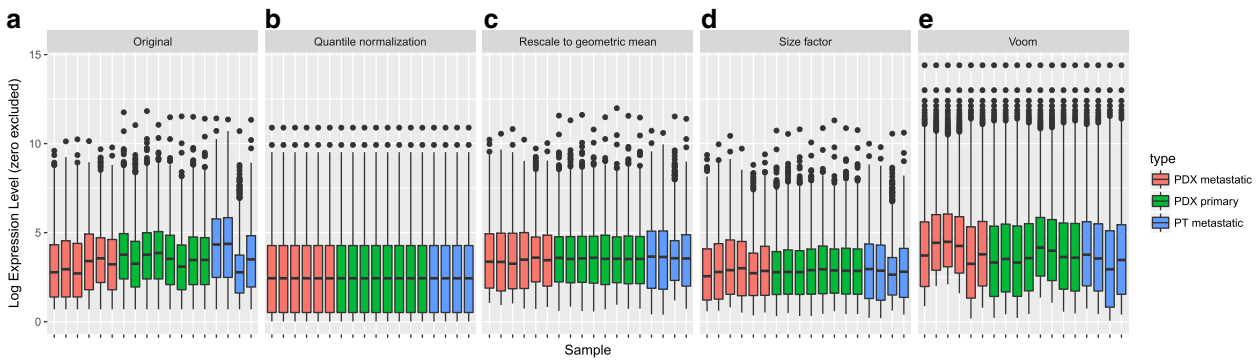

Fig. 4 Box-plot comparison of normalization methods. The cell size is down-sampled to representatively show the general effect of each method. The colors indicate the three cell types reported from the original data. **a** Original data (no normalization). **b** Quantile normalization. **c** Geometrical mean normalization. **d** Size-factor normalization. **e** Voom normalization

dimension reduction, clustering, and differential expression [47]. To resolve this issue, we include an "imputation" step to infer the true expression level of zero values in the input matrix. We choose the scImpute package [48] since it is the fastest among the imputation methods we have tested [48, 49]. It takes about 1 minute on K-dataset using four cores of an Intel Xeon CPU E5-2695 v3 (2.3 GHz). However, the running time grows exponentially and it took more than 15 h to impute the 6000-cell 10x Genomics dataset (Additional file 1: Figure S1).

Gene filtering

Due to high noise levels in scRNA-Seq data, Brennecke et al. [4] recommended removing lowly expressed genes as well as lowly dispersed genes. To this end, Granatum includes a step to remove these genes. Both the average expression level threshold and the dispersion threshold can be adjusted interactively. Granatum displays the threshold selection sliders and the number-of-genes statistics message to enhance integration with the other components. On the mean dispersion plot, a point represents a gene, where the x-coordinate is the log transformed mean of the expression levels of that gene and the y-coordinate is the dispersion factor calculated from a negative binomial model. The plot highlights the preserved genes as black and the filtered genes as gray (Additional file 1: Figure S2).

Clustering

Clustering is a routine heuristic analysis for scRNA-Seq data. Granatum selects five commonly used algorithms: non-negative matrix factorization [22], k-means, k-means combined with correlation t-SNE, hierarchical clustering (Hclust), and Hclust combined with correlation t-SNE. The number of clusters can be set either manually or automatically using an elbow-point-finding algorithm. For the latter automatic approach, the algorithm will cluster samples with the number of clusters (k) ranging from 2 to 10, and determine the best number as the elbow-point k, the starting point of the plateau for explained variance (EV). If Hclust is selected, a pop-up window shows a heatmap with hierarchical grouping and dendrograms.

Next, the two unsupervised PCA and correlation t-SNE plots superimpose the resulting k cluster labels on the samples (Additional file 1: Figure S3). Users can also chose to use their pre-defined labels provided in the sample metadata. By comparing the two sets of labels, one can check the agreement between the prior metadata labels and the computed clusters. We perform the K-means clustering ($k = 2$) on the correlation t-SNE plot, using K-dataset. The generated clusters perfectly correspond to the original cell type labels in this case.

Differential expression

After the clustering step, Granatum allows DE analysis on genes between any two clusters. It currently includes four commonly used DE methods, namely NODES [30], SCDE [31], Limma [33], and edgeR [32]. The DE analysis is performed in a pair-wise fashion when more than two clusters are present. To shorten the computation time, the number of cores for parallelization on multi-core machines can be selected. When the DE computation is complete, the results are shown in a table with DE genes sorted by their Z-scores, along with the coefficients. As another feature to empower the users, the gene symbols are linked to their corresponding GeneCards pages (http://www.genecards.org/) [50]. The "Download CSV table" button allows saving the DE results as a CSV file.

Next, gene set enrichment analysis (GSEA) with either KEGG pathways or Gene Ontology (GO) terms [37, 51–53] can be performed to investigate the biological functions of these DE genes. The results are plotted in an intuitive bubble plot (Fig. 5d). In this plot, the y-axis represents the enrichment score of the gene sets, the x-axis shows gene

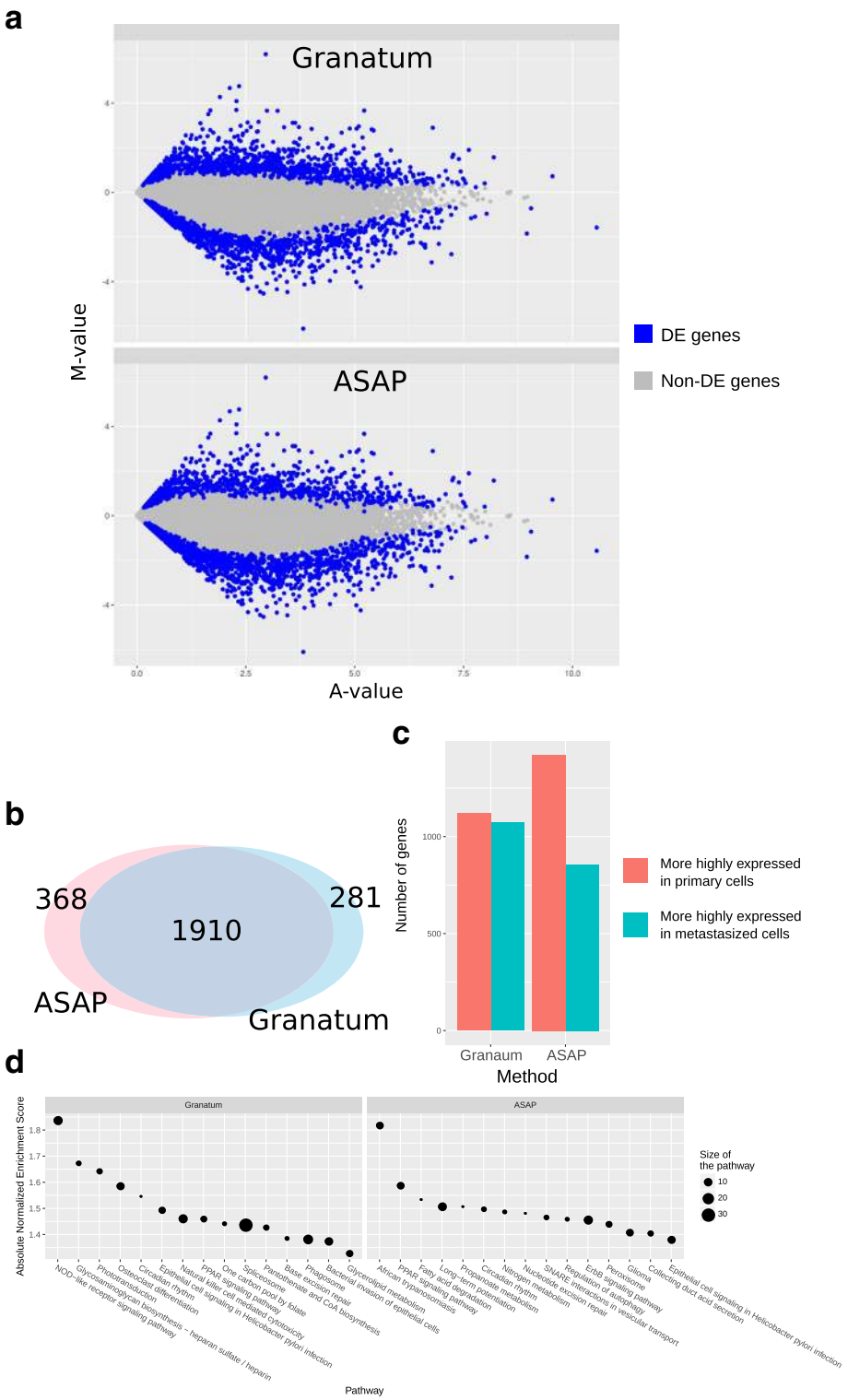

Fig. 5 Comparison of DE genes identified by Granatum or ASAP pipeline. **a** MA plot. *Blue color* labels DE genes, and *gray dots* are non-DE genes. **b** Venn diagram showing the number of DE genes identified by both methods, as well as those uniquely identified by either pipeline. **c** Bar chart comparing the number of genes up regulated in primary cells (*red*) or metastasized cells (*green*). **d** Bubble plots of KEGG pathway GSEA results for the DE genes identified by either pipeline. The y-axis represents the enrichment score of the gene sets, the x-axis shows gene set names, and the size of the bubble indicates the number of genes in that gene set

set names, and the size of the bubble indicates the number of genes in that gene set.

Comparison with other graphical web tools for scRNA-Seq data

To evaluate the differences between Granatum and a similar graphical scRNA-Seq pipeline, ASAP [39], we compare the DE genes (primary vs. metastasized patient) in K-dataset obtained by both pipelines (Fig. 5). While Granatum uses quantile normalization, ASAP uses Voom normalization as the default method. We used SCDE as it is the common DE method for both pipelines.

Both pipelines agree on most DE genes called (Fig. 5a) but each identifies a small number of unique DE genes (Fig. 5b). The numbers of up- or down-regulated DE genes detected by Granatum are closer, whereas in ASAP a lot more genes are more highly regulated in the primary cells compared to those in metastasized cells (Fig. 5c). Further, KEGG pathway-based GSEA analysis on the DE genes shows that Granatum identified more significantly (enrichment score > 1.5) enriched pathways than ASAP (Fig. 5c). The top pathway enriched in Granatum's DE genes is the NOD-like receptor signaling pathway, corresponding to its known association with immunity and inflammation [54]. In ASAP "African trypanosomiasis" is the top pathway, which describes the molecular events when the parasite *Trypanosoma brucei* passes through the blood–brain barrier and causes neurological damage by inducing cytokines. Despite the differences, some signaling pathways are identified by both pipelines with known associations with tumorigenesis, such as the PPAR signaling pathway [55] and the epithelial cell signaling pathway [56].

Granatum-specific steps: protein network visualization and pseudo-time construction

Unlike ASAP, SAKE, and SCRAT, Granatum implements a protein–protein interaction (PPI) network to visualize the connections between the DE genes (Fig. 6a). By default, up to 200 genes are displayed in the PPI network. We use visNetwork to enable the interactive display of the graph [11], so that users can freely rearrange the graph by dragging nodes to the desired locations. Uses can also reconfigure the layout to achieve good visualization via an elastic-spring physics simulation. Nodes are colored according to their regulation direction and the amount of change (quantified using Z-score), where red indicates up-regulation and blue indicates down-regulation. As an example, Fig. 6a shows the PPI network result from PDX primary to metastatic cells in the K-dataset. A large, closely connected module exists in the PPI network, which contains many heat shock protein genes, including down-regulated HSP90AB1, HSPA6, HSPA7, HSPA8, HSPA1A, HSPA1B, and HSPA4L, as well as up-regulated HSP90AA1 and HSPH1 in metastasized cells. Heat shock genes have been long recognized as stress response genes [57], and inhibiting heat shock protein genes can control metastasis in various types of cancers [58, 59].

Lastly, Granatum has included the Monocle algorithm [3], a widely used method to reconstruct a pseudo-timeline for the samples (Fig. 6b). Monocle uses the

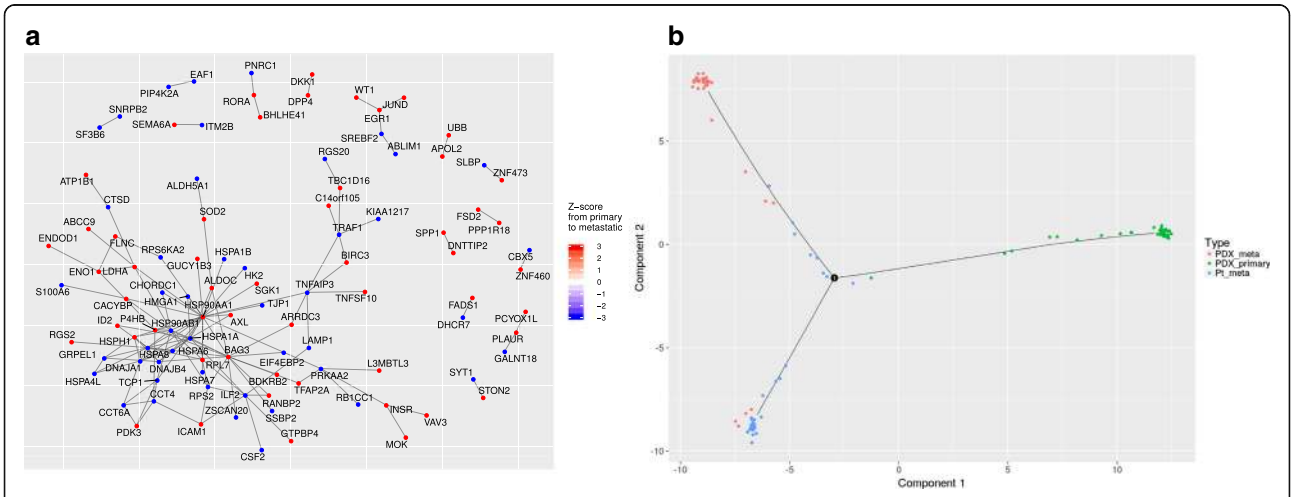

Fig. 6 Protein–protein interaction network and pseudo-time construction steps. **a** The PPI network derived from the DE results between PDX primary and metastasized cells in the K-dataset. The color on each node (gene) indicates its Z-score in the differential expression test. *Red* and *blue* indicate up- and down-regulation in metastasized cells, respectively. **b** The pseudo-time construction step. The Monocle algorithm is customized to visualize the paths among individual cells. Sample labels from the metadata are shown as different colors in the plot

reversed graph embedding algorithm to learn the structure of the data, as well as the principal graph algorithm to find the timelines and branching points of the samples. The user may map any pre-defined labels provided in the metadata sheet onto the scatter plot. In the K-dataset, the three (PDX primary, PDX metastasized, and patient metastasized) types of cancer cells are mostly distinct (Fig. 6b). However, small portions of cells from each type appear to be on intermediate trajectories.

Discussion

The field of scRNA-Seq is evolving rapidly in terms of both the development of instrumentation and the innovation of computational methods. However, it becomes exceedingly hard for a wet-lab researcher without formal bioinformatics training to catch up with the latest iterations of algorithms [5]. This barrier forces many researchers to resort to sending their generated data to third-party bioinformaticians before they are able to visualize the data themselves. This segregation often prolongs the research cycle time, as it often takes significant effort to maintain effective communication between wet-lab researchers and bioinformaticians. In addition, issues with the experimentations do not get the chance to be spotted early enough to avoid significant loss of time and cost in the projects. It is thus attractive to have a non-programming graphical application that includes state-of-the-art algorithms as routine procedures, in the hands of the bench scientists who generate the scRNA-Seq data.

Granatum is our attempt to fill this void. It is, to our knowledge, the most comprehensive solution that aims to cover the entire scRNA-Seq workflow with an intuitive graphical user interface. Throughout the development process, our priority has been to make sure that it is fully accessible to researchers with no programming experience. We have strived to achieve this by making the plots and tables self-explanatory, interactive, and visually pleasant. We have sought inputs from our single-cell bench-side collaborators to ensure that the terminologies are easy to understand by them. We also supplement Granatum with a manual and online video that guide users through the entire workflow, using example datasets. We also seek feedback from community via Github pull-requests, emails discussions, and user surveys.

Currently, Granatum targets bench scientists who have their expression matrices and metadata sheets ready. However, we are developing the next version of Granatum, which will handle the entire scRNA-Seq data processing and analysis pipeline, including FASTQ quality control, alignment, and expression quantification. Another caveat is the lack of benchmark datasets in the single-cell analysis field currently whereby the different computational packages can be evaluated in an unbiased fashion. We thus resort to empirical comparisons between Granatum and packages such as ASAP. In the future, we will enrich Granatum with capacities to analyze and integrate other types of genomics data in single cells, such as exome-seq and methylation data. We will closely update Granatum to keep up with the newest development in the scRNA-Seq bioinformatics field. We welcome third-party developers to download the source code and modify Granatum, and will continuously integrate and improve this tool as the go-to place for single-cell bench scientists.

Conclusions

We have developed a graphical web application called Granatum which enables bench researchers with no programming expertise to analyze state-of-the-art scRNA-Seq data. This tool offers many interactive features to allow routine computational procedures with a great amount of flexibility. We expect that this platform will empower bench-side researchers with more independence in the fast-evolving single cell genomics field.

Additional file

Additional file 1: Supplementary Figures S1, S2, and S3. (PDF 307 kb)

Abbreviations

DE: Differential expression; GO: Gene Ontology; GSEA: Gene-set enrichment analysis; Hclust: Hierarchical clustering; KEGG: Kyoto Encyclopedia of Genes and Genomes; NMF: Non-negative matrix factorization; PCA: Principal component analysis; PPI: Protein–protein interaction; scRNA-Seq: Single-cell high-throughput RNA sequencing; t-SNE: t-Distributed stochastic neighbor embedding

Acknowledgements

We thank Drs. Michael Ortega and Paula Benny for providing valuable feedback during testing of the tool. We also thank other group members in the Garmire group for suggestions for tool development.

Funding

This research is supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (http://datascience.nih.gov/bd2k), P20 COBRE GM103457 awarded by NIH/NIGMS, NICHD R01HD084633 and NLM R01LM012373 to LX Garmire.

Availability of data and materials

All datasets used in the comparisons are reported by previous studies. The K-dataset has the NCBI Gene Expression Omnibus (GEO) accession number GSE73122. The 6000-cell PBMC dataset was retrieved from the 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k).
Granatum can be visited at: http://garmiregroup.org/granatum/app.
Granatum source-code can be found at: http://garmiregroup.org/granatum/code.
A demonstration video can be found at: http://garmiregroup.org/granatum/video.

Authors' contributions

LXG envisioned the project. XZ developed the majority of the pipeline. TW, AT, DG, and AC assisted in developing the pipeline. TW documented the user manual and performed packaging. XZ, TW, and LXG wrote the manuscript. All authors have read, revised, and approved the final manuscript.

Ethics approval and consent to participate

Not sapplicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Graduate Program in Molecular Biology and Bioengineering, University of Hawaii at Manoa, Honolulu, HI 96816, USA. [2]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA. [3]Department of Electrical Engineering, University of Hawaii at Manoa, Honolulu, HI 96816, USA.

Received: 7 August 2017 Accepted: 7 November 2017
Published online: 05 December 2017

References

1. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.
2. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.
3. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
4. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods. 2013;10:1093–5.
5. Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics and computational challenges. Front. Genet. 2016;7:163.
6. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2015. http://www.R-project.org. Accessed 15 Oct 2017.
7. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. bioRxiv. 2016. http://biorxiv.org/content/early/2016/08/15/069633. Accessed 15 Oct 2017.
8. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.
9. RStudio, Inc. Easy web applications in R. 2013.
10. Attali D. shinyjs: easily improve the user experience of your shiny apps in seconds. 2016. https://cran.r-project.org/package=shinyjs.
11. Almende BV, Thieurmel B. visNetwork: network visualization using "vis.js" library. 2016. https://cran.r-project.org/package=visNetwork.
12. Xie Y. DT: a wrapper of the JavaScript library "DataTables". 2016. https://cran.r-project.org/package=DT.
13. Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, et al. plotly: create interactive web graphics via "plotly.js". 2016. https://cran.r-project.org/package=plotly.
14. Wickham H. ggplot2: elegant graphics for data analysis. 2009. http://ggplot2.org.
15. Hicks SC, Teng M, Irizarry RA. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv. 2015; 25528.
16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
17. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127.
18. Kim K-T, Lee HW, Lee H-O, Song HJ, Shin S, Kim H, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. 2016;17:80.
19. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell. 2016:165:1012–26.
20. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:e161.
21. Iglewicz B, Hoaglin DC. How to detect and handle outliers. Milwaukee: Asq Press; 1993.
22. Zhu X, Ching T, Pan X, Weissman S, Garmire L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ Prepr. 2016;4:e1839v1.
23. Chang, Winston, et al. Shiny: Web Application Framework for R, 2015. R package version 0.11 (2015). https://cran.r-project.org/package=shiny.
24. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization BMC Bioinformatics. 2010;11:367.
25. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory IEEE. 1982; 28:129–37.
26. Murtagh F, Contreras P. Methods of hierarchical clustering. arXiv prepr. arXiv1105.0121. 2011. https://arxiv.org/abs/1105.0121.
27. Krijthe J. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation. R Package version 0.10. 2015. http://CRAN.R-project.org/package=Rtsne. Accessed 15 Oct 2017.
28. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Lond Edinburgh Dublin Philos Mag J Sci. 1901;2:559–72.
29. Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT. Bioinformatics. 2017;33:2930–32.
30. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells. bioRxiv. 2016; 49734.
31. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.
32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
34. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16:148.
35. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy by single cell transcriptomics. Nat Neurosci. 2016;19:335.
36. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv. 2016. http://biorxiv.org/content/early/2016/06/20/060012. Accessed 15 Oct 2017.
37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300.
39. Gardeux V, David F, Shajkofci A, Schwalie PC, Deplancke B. ASAP: a web-based platform for the analysis and inter-active visualization of single-cell RNA-seq data. bioRxiv. 2016;96222.
40. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. bioRxiv. 2017;133173.
41. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
42. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 2014;15:550.
44. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15:R29.
45. Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500:593.
46. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13:204–16.
47. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:1–10.
48. Li WV, Li JJ. scImpute: accurate and robust imputation for single cell RNA-seq data. bioRxiv. 2017;141598.

49. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. Gene expression recovery for single cell RNA sequencing. bioRxiv. 2017;138677.
50. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13:163.
51. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61.
52. Consortium GO. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43:D1049–56.
53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
54. Fritz JH, Ferrero RL, Philpott DJ, Girardin SE. Nod-like proteins in immunity, inflammation and disease. Nat Immunol. 2006;7:1250–7.
55. Belfiore A, Genua M, Malaguarnera R. PPAR-γ agonists and their effects on IGF-I receptor signaling: implications for cancer. PPAR Res. 2009;2009: 830501.
56. Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB. Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. Nature. 2003;422:313–7.
57. Santoro MG. Heat shock factors and the control of the stress response. Biochem Pharmacol. 2000;59:55–63.
58. Tamura Y, Peng P, Liu K, Daou M, Srivastava PK. Immunotherapy of tumors with autologous tumor-derived heat shock protein preparations. Science. 1997;278:117–20.
59. Eccles SA, Massey A, Raynaud FI, Sharp SY, Box G, Valenti M, et al. NVP-AUY922: a novel heat shock protein 90 inhibitor active against xenograft tumor growth, angiogenesis, and metastasis. Cancer Res. 2008;68:2850–60.
60. Zheng, Grace XY, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications 8. 2017:14049.
61. Satija R, Butler, Andrew. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv. 2017:164889.
62. Juliá, Miguel, Telenti A, Rausell A. Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics 31.20. 2015:3380–3382.
63. Guo M, et al. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS computational biology 11.11. 2015:e1004575.