

SOFTWARE

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

1 Xun Zhu<sup>1,2</sup>, Thomas Wolfgruber<sup>1,2</sup>, Austin Tasato<sup>3</sup>, David G. Garmire<sup>3</sup>, Lana X Garmire<sup>1, 2\*</sup>

2 \_\_\_\_\_

3 \*Correspondence:

4 LGarmire@cc.hawaii.edu

5 1 Graduate Program in Molecular  
6 Biology and Bioengineering,  
7 University of Hawaii at Manoa,  
8 Honolulu, HI 96816

9 2 Epidemiology Program, University  
10 of Hawaii Cancer Center, Honolulu,  
11 HI 96813

12 3 Department of Electrical  
13 Engineering, University of Hawaii at  
14 Manoa, Honolulu, HI 96816

15

## 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 have limited accessibility to bench scientists as they require significant amounts of bioinformatics 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 filtering, gene-expression normalization, 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 the 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; differential expression; pathway; pseudo-time; software

## 16 **Background**

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

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

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

## 41 **Implementation**

### 42 **Overview**

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

### 51 **Interactive widgets**

52 Layout and interactivity for the protein-protein interaction (PPI) network modules is implemented  
53 using the visNetwork package [11]. Preview of user-submitted data and display of tabular data in  
54 various modules is implemented using DataTables [12]. The interactive outlier-identification step

55 uses Plotly [13]. Scatter-plots, box-plots, and pseudo-time construction in Monocle are done by  
56 the ggplot2 package [3,14].

## 57 **Back-end variable management**

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

## 64 **Batch-effect removal**

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

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

72 **Median alignment:** First, this method calculates the median expression of each sample, denoted as  
73  $med_i$  for sample  $i$ . Second, it calculates the mean of  $med_i$  for each batch, denoted as

74  $batchMean_b$  for batch  $b$ ,

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

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

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

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

## 79 **Outlier detection and gene filtering**

80 Z-score threshold is used to automatically detect outliers. The z-score of a cell is calculated by  
81 calculating the Euclidean norm of the cell's vector of expression levels, after scaling all genes to  
82 have unit standard deviation and zero mean [21]. Over-dispersion gene filtering is done as  
83 recommended by Brennecke et al. 2013 [4]. The output of the Monocle package [3] is modified to  
84 calculate dispersion and fit a negative binomial model to the result.

## 85 **Clustering methods**

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

88 **Non-negative matrix factorization (NMF):** The log-transformed expression matrix ( $n$ -by- $m$ ) is  
89 factorized into two non-negative matrices  $H$  ( $n$ -by- $k$ ) and  $W$  ( $k$ -by- $m$ ). The highest-valued  $k$  entry

90 in each column of  $W$  determines the membership of each cluster [22,23]. The NMF computation is  
91 implemented in the NMF R-package, as reported earlier [22,24].

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

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

## 97 **Dimension reduction methods**

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

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

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

103 **PCA:** The Principal Component Analysis algorithm, implemented as “prcomp” function in R,  
104 decomposes the original data into linearly uncorrelated variables (components) using orthogonal  
105 transformation. The components are then sorted by their variance. The two components with the  
106 largest variances (PC1 and PC2) are extracted for visualization [28].

## 107 **Elbow-point finding algorithm in clustering**

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

## 116 **Differential expression analysis**

117 We include four differential expression (DE) algorithms in Granatum: NODES[30], SCDE[31], EdgeR  
118 [32], and Limma [33]. Among them, NODES and SCDE are designed for single-cell RNA-Seq  
119 specifically. EdgeR and Limma are conventional bulk cell RNA-Seq DE tools that have also been  
120 used in single-cell RNA-Seq studies [34,35]. When more than two clusters are present, we perform  
121 pairwise DE analysis on all clusters. We use default parameters for all packages. Their versions are:  
122 NODES (0.0.0.9010), SCDE (1.99.2), EdgeR (3.18.1) and Limma (3.32.2)

## 123 **Gene-set enrichment analysis**

124 The *fgsea* R-package implements the Gene Set Enrichment Analysis (GSEA) algorithm with  
125 optimizations for speedup [36,37]. GSEA calculates an *enrichment score*, which quantifies the  
126 relevance of a gene set (for example, a KEGG pathway or a GO term) to a particular group of

127 selected genes (e.g., DE genes called by a method). The  $p$ -value is calculated for each gene set  
128 according to the empirical distribution, followed by Benjamini–Hochberg multiple hypothesis tests  
129 [38].

## 130 **Pseudo-time construction**

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

# 134 **Results**

## 135 **Overview of Granatum**

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

141 We have systematically compared Granatum with 12 other existing tools to demonstrate its  
142 versatile functions (Table 1). Popular packages such as SCDE / PAGODA and Flotilla are developed  
143 for programmers and require expertise in a particular programming language. In contrast,  
144 Granatum with its easy-to-navigate graphical interface requires no programming specialty. The



145 current version of Granatum neatly presents nine modules, arranged as steps and ordered by their  
146 dependency. It starts with one or more expression matrices and corresponding sample metadata  
147 sheet(s), followed by data merging, batch-effect removal, outlier removal, normalization, gene  
148 filtering, clustering, differential expression, protein-protein network, and pseudo-time  
149 construction.

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

## 162 **Testing of the software**

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

166 To estimate the total running time of Granatum (with default parameters) at different sizes of  
167 datasets, we first simulate expression matrices with 200, 400, 800, or 1600 cells using the Splatter  
168 package, based on the parameters estimated from the K-dataset [40]. Additionally, we also use  
169 down-sample approach (200, 400, 800, 1600, 3200 and 6000 cells) on a dataset (P-dataset)  
170 provided by 10x Genomics, which has 6,000 peripheral blood mononuclear cells (PBMCs)  
171 (<https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k>). The  
172 running time scales linearly with the number of cells, regardless of platform (Suppl. Figure 1). The  
173 most time-consuming step is Monocle based pseudo-time construction, which takes about 80% of  
174 all computing time.

175 In the following sections, we use K-dataset to elaborate the details of each step in Granatum in  
176 chronological order.

## 177 **Upload data**

178 Granatum accepts one or more expression matrices as input. Each expression matrix may be  
179 accompanied by a *metadata sheet*. A metadata sheet is a table describing the groups, batches, or  
180 other properties of the samples in the corresponding expression matrix. Users may upload  
181 multiple matrices sequentially. Currently, Granatum accepts either human or mouse species, for

182 downstream functional analysis. After uploading the input files, users can preview the matrix and  
183 metadata tables to validate that the dataset is uploaded correctly.

## 184 **Batch-effect removal**

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

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

## 198 **Outlier identification**

199 Computationally abnormal samples pose serious problems for many downstream analysis  
200 procedures. Thus, it is crucial to identify and remove them in the early stage. Granatum's outlier  
201 identification step features PCA and t-SNE [41] plots, two connected interactive scatter-plots that

202 have different computational characteristics. A PCA plot illustrates the Euclidean distance between  
203 the samples, and a correlation t-SNE plot shows the associative distances between the samples.  
204 Granatum generates these two plots using top genes (default 500). Using the Plotly library [13],  
205 these plots are highly interactive. It is an example of thoughtful tool design that empowers users  
206 to explore the data. Outliers can be identified automatically by using a z-score threshold or setting  
207 a fixed number of outliers. In addition, each sample can be selected or de-selected, by clicking,  
208 boxing or drawing a lasso on its corresponding points.

209 The original K-dataset has one sample with abnormally low expression level. This potential outlier  
210 sample can affect downstream analyses. Using Granatum, users can easily spot such outliers in the  
211 PCA plot or in the correlation t-SNE plot (Figure 3A and B). After removal of the outliers, the top-  
212 gene based PCA and correlation t-SNE plots are more balanced (Figure 3C and D).

## 213 **Normalization**

214 Normalization is essential to most scRNA-Seq data before the downstream functional analyses  
215 (except those with the UMI counts). Granatum includes four commonly used normalization  
216 algorithms: quantile normalization, geometric mean normalization, size-factor normalization  
217 [42,43], and Voom [44]. A post-normalization box-plot helps illustrate the normalization effect to  
218 the median, mean, and extreme values across samples.

219 The box-plots allow observation of various degrees of stabilization (Figure 4). The original dataset  
220 has high levels variations among samples (Figure 4A). Quantile normalization unifies the  
221 expression distribution of all samples, thus renders the box-plots identical (Figure 4B). Mean

222 alignment tries to unify all means of the samples by multiplying the expression levels in each  
223 sample by a factor, thus visually all means (the red dots) are the same (Figure 4C). Size-factor and  
224 Voom normalization use more sophisticated procedures to normalize the data, but the variation of  
225 distribution across samples is evidently reduced (Figure 4D and E). According to our experience  
226 and others [45,46], quantile normalization is recommended.

## 227 **Gene filtering**

228 Due to high noise levels in scRNA-Seq, Brennecke et al. [4] recommended removing lowly-  
229 expressed genes as well as lowly-dispersed genes. To this end, Granatum includes a step to  
230 remove these genes. Both the average expression-level threshold and the dispersion threshold can  
231 be adjusted interactively. Granatum displays the threshold selection sliders and the number-of-  
232 genes statistics message to enhance integration with the other components. On the mean-  
233 dispersion plot, a point represents a gene, whose x-coordinate is the log transformed mean of the  
234 expression levels of that gene, and the y-coordinate is the dispersion factor calculated from a  
235 negative binomial model. The plot highlights the preserved genes as black and the filtered genes as  
236 gray (Suppl. Figure 2).

## 237 **Clustering**

238 Clustering is a routine heuristic analysis for scRNA-Seq data. Granatum selects five commonly used  
239 algorithms: non-negative matrix factorization [22], k-means, k-means combined with correlation t-  
240 SNE, hierarchical clustering (hclust), and hclust combined with correlation t-SNE. The number of  
241 clusters can either be set manually, or automatically using an elbow-point finding algorithm. For

242 the latter automatic approach, the algorithm will cluster samples with the number of clusters ( $k$ )  
243 ranging from 2 to 10, and determine the best number as the elbow-point  $k$ . the starting point of  
244 the plateau for explained variance (EV). If hclust is selected, a pop-up window shows a heatmap  
245 with hierarchical grouping and dendrograms.

246 Next, the two unsupervised PCA and correlation t-SNE plots superimpose the resulting  $k$  cluster  
247 labels on the samples (Suppl. Figure 3). Users can also chose to use their pre-defined labels  
248 provided in the sample metadata. By comparing the two sets of labels, one can check the  
249 agreement between the prior metadata labels and the computed clusters. We perform the K-  
250 means clustering ( $k = 2$ ) on the correlation t-SNE plot, using K-dataset. The generated clusters  
251 perfectly correspond to the original cell type labels in this case.

## 252 **Differential expression**

253 After the clustering step, Granatum allows DE analysis on genes between any two clusters. It  
254 currently includes four commonly used differential expression methods, namely NODES [30], SCDE  
255 [31], Limma [33] and edgeR [32]. The DE analysis is performed in a pair-wise fashion when more  
256 than two clusters are present. To shorten the computation time, the number of cores for  
257 parallelization on multi-core machines can be selected. When the DE computation is complete, the  
258 results are shown in a table with DE genes sorted by their Z-scores, along with the coefficients. As  
259 another feature to empower the users, the gene symbols are linked to their corresponding  
260 GeneCards pages ([www.genecards.org](http://www.genecards.org)) [47]. The "Download CSV table" button allows saving the  
261 DE results as a CSV file.

262 Next, Gene Set Enrichment Analysis (GSEA) with either KEGG pathways or Gene Ontology (GO)  
263 terms [37,48–50] can be performed, to investigate the biological functions of these DE genes. The  
264 results are plot in an intuitive bubble-plot (Figure 5D). In this plot, the y-axis represents the  
265 enrichment score of the gene sets, x-axis shows gene set names, and the size of the bubble  
266 indicates the number of genes in that gene set.

## 267 **Comparison with other Graphical web tools of scRNA-Seq**

268 To evaluate the differences between Granatum and a similar graphical scRNA-Seq pipeline ASAP  
269 [39], we compare the DE genes (primary vs. metastasized patient) in K-dataset obtained by both  
270 pipelines (Figure 5). While Granatum uses quantile normalization, ASAP uses Voom normalization  
271 as default method. We use SCDE as it is the common DE method for both pipelines.

272 Both pipelines agree on most DE genes called (Figure 5A), but each identifies a small number of  
273 unique DE genes (Figure 5B). In Granatum, the number of up or down regulated DE genes detected  
274 by Granatum are closer. Whereas in ASAP, a lot more genes are higher regulated in the primary  
275 cells, compared to those in metastasized cells (Figure 5C). Further, KEGG pathway based GSEA  
276 analysis on the DE genes shows that Granatum identified more significantly (Enrichment Score >  
277 1.5) enriched pathways than ASAP (Figure 5C). The top pathway enriched in Granatum's DE genes  
278 is the NOD-like receptor-signaling pathway, corresponding to its known association with immunity  
279 and inflammation [51]. In ASAP "African trypanosomiasis" is the top pathway, which describes the  
280 molecular events when parasite *Trypanosoma brucei* pass through the blood-brain barrier and  
281 cause neurological damage by inducing cytokines. Despite the differences, some signaling

282 pathways are identified by both pipelines with known associations with tumorigenesis, such as  
283 PPAR signaling pathway [52] and Epithelial cell signaling pathway [53].

## 284 **Granatum-specific Steps: Protein network visualization and** 285 **Pseudo-time construction**

286 Unlike ASAP, SAKE and SCRAT, Granatum implements a Protein-protein interaction (PPI) network  
287 to visualize the connections between the DE genes (Figure 6A). By default, up to 200 genes are  
288 displayed in PPI network. We use visNetwork to enable the interactive display of the graph [11], so  
289 that users can freely rearrange the graph by dragging the nodes to the desired location. Users can  
290 also reconfigure the layout to achieve good visualization via an elastic-spring physics simulation.  
291 Nodes are colored according to their regulation direction and the amount of change (quantified  
292 using Z-score), where red indicates up-regulation and blue indicates down-regulation. As an  
293 example, Figure 6A shows the PPI network result from PDX primary to metastatic cells in the K-  
294 dataset. A large, closely connected module exists in PPI network, which contains many heat shock  
295 protein genes including down-regulated HSP90AB1, HSPA6, HSPA7, HSPA8, HSPA1A, HSPA1B and  
296 HSPA4L as well as up-regulated HSP90AA1 and HSPH1 in metastasized cells. Heat shock genes have  
297 been long recognized as a stress response genes [54], and inhibiting heat shock protein genes can  
298 control metastasis in various types of cancers [55,56].

299 Lastly, Granatum has included the Monocle algorithm[3], a widely-used method to reconstruct a  
300 pseudo-timeline for the samples (Figure 6B). Monocle uses the Reversed Graph Embedding  
301 algorithm to learn the structure of the data, as well as the Principal Graph algorithm to find the



302 timelines and branching points of the samples. The user may map any pre-defined labels provided  
303 in the metadata sheet onto the scatter-plot. In the K-dataset, the three (PDX primary, PDX  
304 metastasized, and patient metastasized) types of cancer cells are mostly distinct (Figure 6B).  
305 However, small portions of cells from each type appear to be on intermediate trajectory.

## 306 Discussion

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

318 Granatum is our attempt to fill this void. It is to our knowledge the most comprehensive solution  
319 that aims to cover the entire scRNA-Seq workflow with an intuitive graphical user interface.  
320 Throughout the development process, our priority has been to make sure that it is fully accessible

321 to researchers with no programming experiments. We have strived to achieve this, by making the  
322 plots and tables self-explanatory, interactive and visually pleasant. We have sought inputs from  
323 our single-cell bench-side collaborators to ensure that the terminologies are easy to understand by  
324 them. We also supplement Granatum with a manual and online video that guide the users through  
325 the entire workflow, using example datasets. We also seek feedback from community via Github  
326 pull-requests, emails discussions and user survey.

327 Currently, Granatum targets bench scientists who have their expression matrices and metadata  
328 sheets ready. However, we are developing the next version of Granatum, which will handle the  
329 entire scRNA-Seq data processing and analysis pipeline including FASTQ quality control, alignment,  
330 and expression quantification. Another caveat is the lacking of benchmark dataset in single-cell  
331 analysis field currently, where the different computational packages can be evaluated unbiasedly.  
332 We thus resort to empirical comparisons on packages between Granatum and ASAP. In the future,  
333 we will enrich Granatum with capacities to analyze and integrate other types of genomics data in  
334 single cells, such as exome-seq and methylation data. We will closely update Granatum to keep up  
335 with the newest development in the scRNA-Seq bioinformatics field. We welcome third-party  
336 developers to download the source-code and modify Granatum, and will continuous integrate and  
337 innovate this tool as the go-to place for single-cell bench scientists.

## 338 **Conclusions**

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

## 344 **Figure legends**

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

350 **Figure 2: The batch-effect removal.** The PCA plots show the before/after median alignment  
351 comparison. The colors indicate the two batches 1 and 2, and the shapes indicate the three cell  
352 types reported from the original data. (A) Before and (B) After batch-effect removal.

353 **Figure 3: The outlier removal using PCA plot.** (A) Before outlier removal. (B) After outlier removal.

354 **Figure 4: Box-plot comparison of normalization methods.** The cells size is down-sampled to  
355 representatively show the general effect of each method. The colors indicate the three cell types

356 reported from the original data. (A) The original (no normalization) (B) Quantile normalization (C)  
357 Geometrical mean normalization (D) Size-factor normalization (E) Voom normalization.

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

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

## 371 **Supplementary Figures**

372 **Suppl. Figure 1: Granatum total running time with various numbers of cells.** Datasets with various  
373 sizes from two single-cell platforms (Fluidigm C1 and 10x Genomics) are used. To generate  
374 expression data up to 6000 cells, the Fluidigm C1 datasets are simulated using Splatter, with

375 parameters estimated from the K-dataset (118 cells). The 10x Genomics datasets are down-  
376 sampled from the original 6000-cell PBMC dataset. The x-axis represents the size of the dataset,  
377 and the y-axis represents the total running time (in minutes) of Granatum. Monocle based pseudo-  
378 time construction step takes about 80% of total running time.

379 **Suppl. Figure 2: The Gene filtering step.** The y-axis of the scatter-plot is the empirical dispersion,  
380 estimated by a negative binomial model. The x-axis is the log mean expression of each gene. The  
381 red line is the fit of a negative binomial model onto the data. Black points represent gene to be  
382 kept and gray points are filtered genes.

383 **Suppl. Figure 3: The Clustering step.** (A) PCA and (B) Correlation t-SNE plots of single cells (dots)  
384 are shown, with colors indicating the cell types reported in the original dataset and cluster number  
385 (1, 2) super-imposed on the cells.

## 386 **Tables**

387 **Table 1: Comparison of existing single-cell analysis pipelines.**

## 388 **Declarations**

## 389 **Ethics approval and consent to participate**

390 Not Applicable.

391 **Consent for publication**

392 Not Applicable.

393 **Availability of data and material**

394 All datasets used in the comparisons are reported by previous studies. The K-dataset has the NCBI  
395 Gene Expression Omnibus (GEO) accession number GSE73122. The 6,000 cells PBMCs dataset is  
396 retried from 10x Genomics website [https://support.10xgenomics.com/single-cell-gene-](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k)  
397 [expression/datasets/1.1.0/pbmc6k](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k).

398 Granatum can be visited at: <http://garmiregroup.org/granatum/app>

399 Granatum source-code can be found at: <http://garmiregroup.org/granatum/code>

400 A demonstration video can be found at: <http://garmiregroup.org/granatum/video>

401 **Competing interests**

402 The authors declared no conflict of interest.

403 **Funding**

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

408 **Authors' contributions**

409 LXG envisioned the project. XZ developed the majority of the pipeline. TW and AT assisted in

410 developing the pipeline. TW documented the user manual and performed packaging. XZ, TW and  
411 LXG wrote the manuscript. All authors have read, revised, and approved the final manuscript.

## 412 **Acknowledgements**

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

## 416 **List of abbreviations**

417 **scRNA-Seq:** Single-cell high-throughput RNA sequencing

418 **DE:** differential expression

419 **GSEA:** Gene-set enrichment analysis

420 **KEGG:** Kyoto Encyclopedia of Genes and Genomes

421 **GO:** Gene ontology

422 **PCA:** Principal component analysis

423 **t-SNE:** t-Distributed Stochastic Neighbor Embedding

424 **NMF:** Non-negative matrix factorization

425 **Hclust:** Hierarchical clustering

426 **PPI:** Protein-protein interaction

## 427 **References**

- 428 1. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq  
429 highlights intratumoral heterogeneity in primary glioblastoma. *Science* (80-. ). 2014;344:1396–401.
- 430 2. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates  
431 that thousands of human genes are microRNA targets. *Cell*. Elsevier; 2005;120:15–20.
- 432 3. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and  
433 regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat.*  
434 *Biotechnol.* Nature Research; 2014;32:381–6.
- 435 4. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for  
436 technical noise in single-cell RNA-seq experiments. *Nat. Methods*. Nature Publishing Group; 2013;
- 437 5. Poirion OB, Zhu X, Ching T, Garmire L. Single-Cell Transcriptomics Bioinformatics and  
438 Computational Challenges. *Front. Genet.* 2016. p. 163.
- 439 6. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical  
440 Computing, Vienna, Austria. 2015, URL [http. www. R-project. org](http://www.R-project.org). 2016;
- 441 7. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. scater: pre-processing, quality control,  
442 normalisation and visualisation of single-cell RNA-seq data in R. *bioRxiv [Internet]*. Cold Spring



- 443 Harbor Labs Journals; 2016; Available from: <http://biorxiv.org/content/early/2016/08/15/069633>
- 444 8. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*
- 445 Taylor & Francis; 1996;5:299–314.
- 446 9. RStudio, Inc. Easy web applications in R. 2013.
- 447 10. Attali D. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds [Internet].
- 448 2016. Available from: <https://cran.r-project.org/package=shinyjs>
- 449 11. Almende B.V., Thieurmel B. visNetwork: Network Visualization using “vis.js” Library [Internet].
- 450 2016. Available from: <https://cran.r-project.org/package=visNetwork>
- 451 12. Xie Y. DT: A Wrapper of the JavaScript Library “DataTables” [Internet]. 2016. Available from:
- 452 <https://cran.r-project.org/package=DT>
- 453 13. Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, et al. plotly: Create
- 454 Interactive Web Graphics via “plotly.js” [Internet]. 2016. Available from: [https://cran.r-](https://cran.r-project.org/package=plotly)
- 455 [project.org/package=plotly](https://cran.r-project.org/package=plotly)
- 456 14. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York;
- 457 2009. Available from: <http://ggplot2.org>
- 458 15. Hicks SC, Teng M, Irizarry RA. On the widespread and critical impact of systematic bias and
- 459 batch effects in single-cell RNA-Seq data. *bioRxiv. Cold Spring Harbor Labs Journals*; 2015;25528.
- 460 16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using

- 461 empirical Bayes methods. *Biostatistics*. Biometrika Trust; 2007;8:118–27.
- 462 17. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing  
463 identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells.  
464 *Genome Biol*. 2015;16:127.
- 465 18. Kim K-T, Lee HW, Lee H-O, Song HJ, Shin S, Kim H, et al. Application of single-cell RNA  
466 sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma.  
467 *Genome Biol*. BioMed Central; 2016;17:80.
- 468 19. Petropoulos S, Edsgård D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq  
469 Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*. Elsevier;  
470 2016;
- 471 20. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable  
472 analysis. *PLoS Genet*. Public Library of Science; 2007;3:e161.
- 473 21. Iglewicz B, Hoaglin DC. How to detect and handle outliers. Asq Press; 1993.
- 474 22. Zhu X, Ching T, Pan X, Weissman S, Garmire L. Detecting heterogeneity in single-cell RNA-Seq  
475 data by non-negative matrix factorization. *PeerJ Prepr*. PeerJ Inc. San Francisco, USA;  
476 2016;4:e1839v1.
- 477 23. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using  
478 matrix factorization. *Proc. Natl. Acad. Sci*. 2004;101:4164–9.

- 479 24. Gaujoux R, Seoighe C. Algorithms and framework for nonnegative matrix factorization (NMF).  
480 2010.
- 481 25. Lloyd S. Least squares quantization in PCM. *IEEE Trans. Inf. theory.* IEEE; 1982;28:129–37.
- 482 26. Murtagh F, Contreras P. Methods of hierarchical clustering. *arXiv Prepr. arXiv1105.0121.* 2011;
- 483 27. Krijthe J. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut  
484 Implementation. R Packag. version 0.10, URL <http://CRAN.R-project.org/package=Rtsne>. 2015;
- 485 28. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. London,  
486 Edinburgh, Dublin *Philos. Mag. J. Sci.* Taylor & Francis; 1901;2:559–72.
- 487 29. Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT. *Bioinformatics.* Oxford  
488 University Press; 2017;btx315.
- 489 30. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential  
490 expression analysis for single cells. *bioRxiv. Cold Spring Harbor Labs Journals;* 2016;49734.
- 491 31. Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential  
492 expression analysis. *Nat. Methods.* Nature Publishing Group; 2014;11:740–2.
- 493 32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential  
494 expression analysis of digital gene expression data. *Bioinformatics.* Oxford Univ Press;  
495 2010;26:139–40.
- 496 33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression

- 497 analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res. Oxford University Press*;  
498 2015;43:e47–e47.
- 499 34. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of  
500 linear and circular RNAs in mouse preimplantation embryos. *Genome Biol. BioMed Central*;  
501 2015;16:148.
- 502 35. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy  
503 by single cell transcriptomics. *Nat. Neurosci. NIH Public Access*; 2016;19:335.
- 504 36. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative  
505 statistic calculation. *bioRxiv [Internet]. Cold Spring Harbor Labs Journals*; 2016; Available from:  
506 <http://biorxiv.org/content/early/2016/06/20/060012>
- 507 37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set  
508 enrichment analysis: a knowledge-based approach for interpreting genome-wide expression  
509 profiles. *Proc. Natl. Acad. Sci. National Acad Sciences*; 2005;102:15545–50.
- 510 38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful  
511 approach to multiple testing. *J. R. Stat. Soc. Ser. B. JSTOR*; 1995;289–300.
- 512 39. Gardeux V, David F, Shajkofci A, Schwalie PC, Deplancke B. ASAP: a Web-based platform for the  
513 analysis and inter-active visualization of single-cell RNA-seq data. *bioRxiv. Cold Spring Harbor Labs  
514 Journals*; 2016;96222.
- 515 40. Zappia L, Phipson B, Oshlack A. Splatter: Simulation Of Single-Cell RNA Sequencing Data.

- 516 bioRxiv. Cold Spring Harbor Labs Journals; 2017;133173.
- 517 41. Maaten L van der, Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 2008;9:2579–  
518 605.
- 519 42. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high  
520 density oligonucleotide array data based on variance and bias. *Bioinformatics.* Oxford Univ Press;  
521 2003;19:185–93.
- 522 43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq  
523 data with DESeq2. bioRxiv. Cold Spring Harbor Labs Journals; 2014;
- 524 44. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools  
525 for RNA-seq read counts. *Genome Biol. BioMed Central*; 2014;15:R29.
- 526 45. Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, et al. Genetic programs in human and mouse early  
527 embryos revealed by single-cell RNA sequencing. *Nature. NIH Public Access*; 2013;500:593.
- 528 46. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional  
529 quantile normalization. *Biostatistics.* Oxford University Press; 2012;13:204–16.
- 530 47. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about  
531 genes, proteins and diseases. *Trends Genet. Elsevier Current Trends*; 1997;13:163.
- 532 48. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on  
533 genomes, pathways, diseases and drugs. *Nucleic Acids Res. Oxford Univ Press*; 2017;45:D353--

534 D361.

535 49. Consortium GO, others. Gene ontology consortium: going forward. *Nucleic Acids Res. Oxford*  
536 *Univ Press*; 2015;43:D1049--D1056.

537 50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for  
538 the unification of biology. *Nat. Genet. Nature Publishing Group*; 2000;25:25–9.

539 51. Fritz JH, Ferrero RL, Philpott DJ, Girardin SE. Nod-like proteins in immunity, inflammation and  
540 disease. *Nat. Immunol. Nature Publishing Group*; 2006;7:1250–7.

541 52. Belfiore A, Genua M, Malaguarnera R. PPAR-agonists and their effects on IGF-I receptor  
542 signaling: implications for cancer. *PPAR Res. Hindawi Publishing Corporation*; 2009;2009.

543 53. Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB. Hedgehog signalling  
544 within airway epithelial progenitors and in small-cell lung cancer. *Nature. Nature Publishing Group*;  
545 2003;422:313–7.

546 54. Santoro MG. Heat shock factors and the control of the stress response. *Biochem. Pharmacol.*  
547 *Elsevier*; 2000;59:55–63.

548 55. Tamura Y, Peng P, Liu K, Daou M, Srivastava PK. Immunotherapy of tumors with autologous  
549 tumor-derived heat shock protein preparations. *Science (80-. ). American Association for the*  
550 *Advancement of Science*; 1997;278:117–20.

551 56. Eccles SA, Massey A, Raynaud FI, Sharp SY, Box G, Valenti M, et al. NVP-AUY922: a novel heat

Zhu et al.

Page 31 of 31

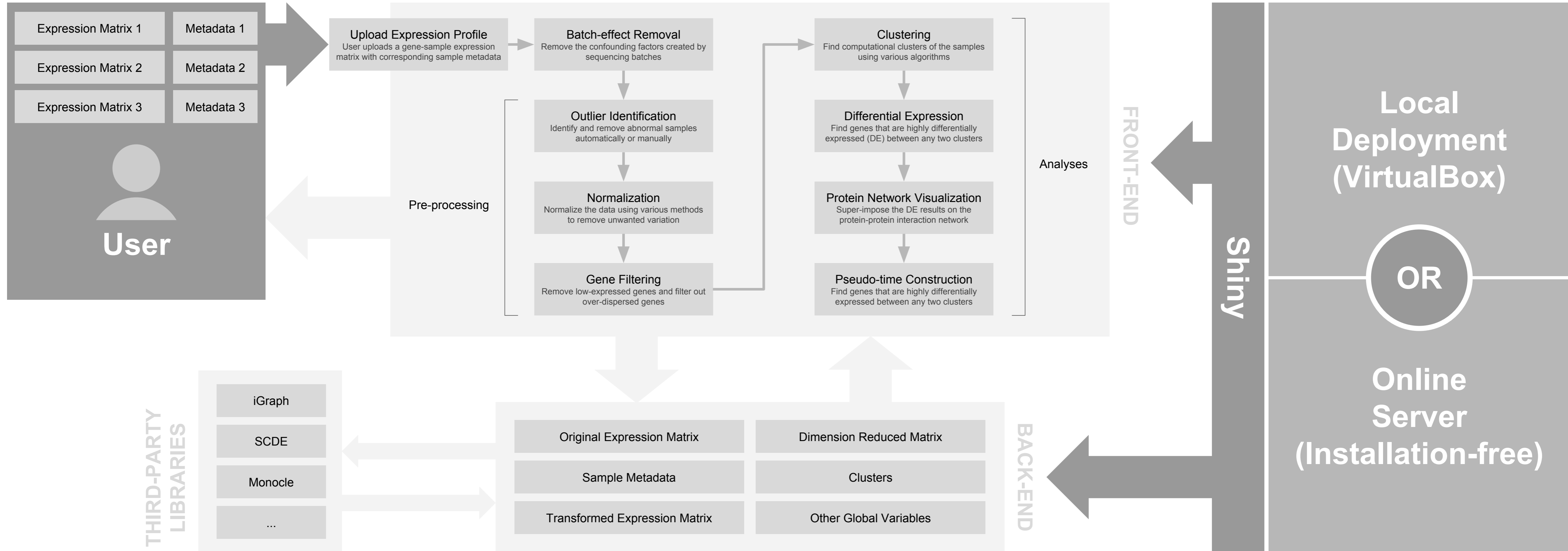
552 shock protein 90 inhibitor active against xenograft tumor growth, angiogenesis, and metastasis.

553 Cancer Res. AACR; 2008;68:2850–60.

554

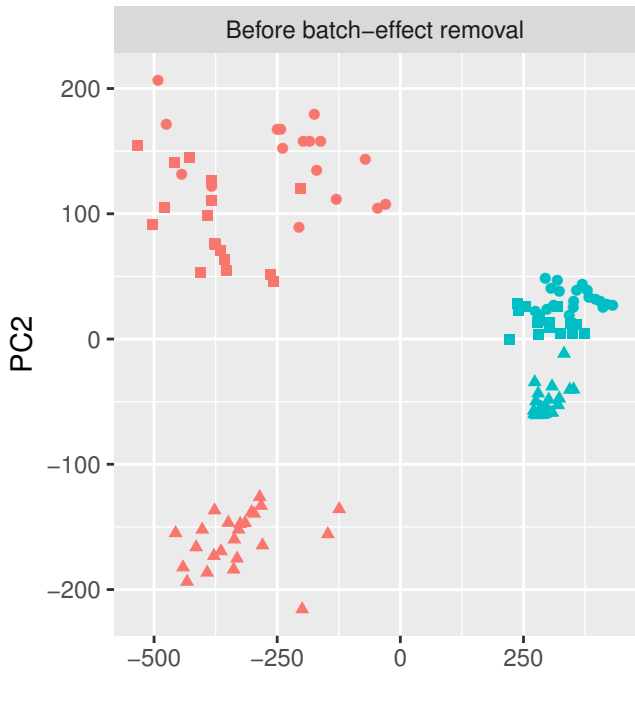
555

# Granatum

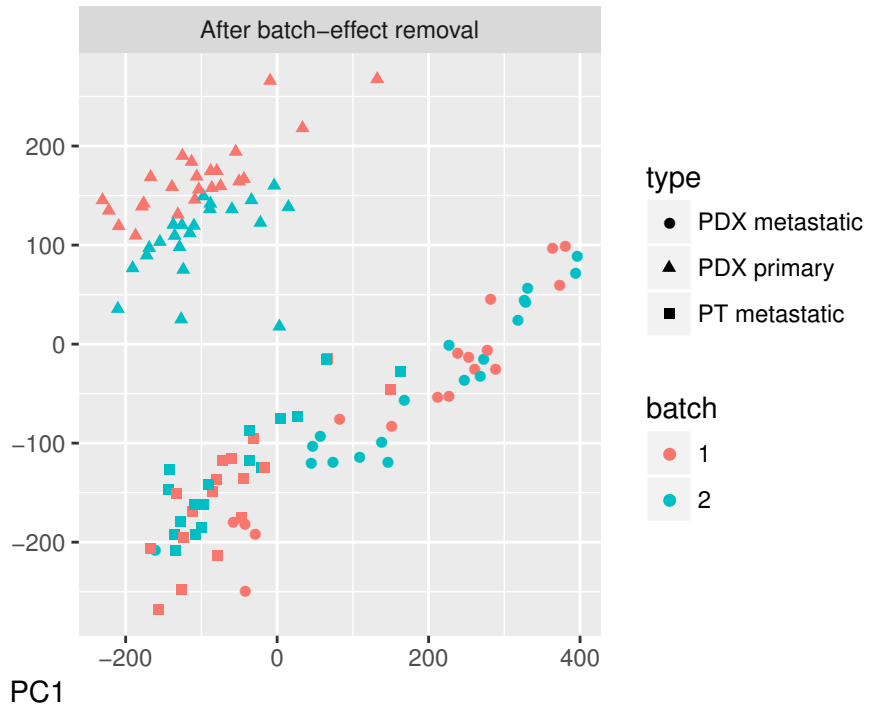




A



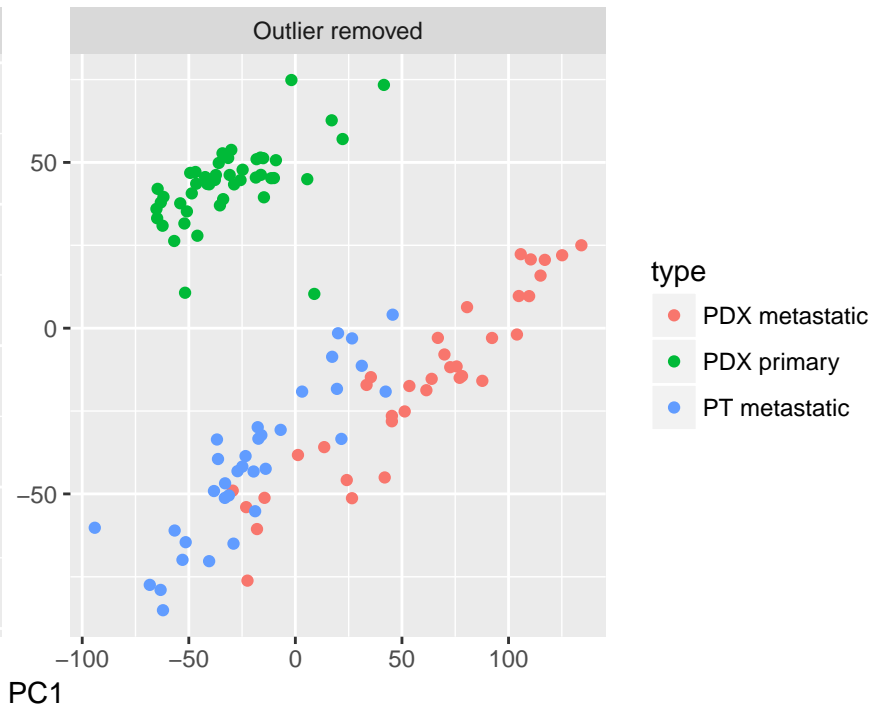
B

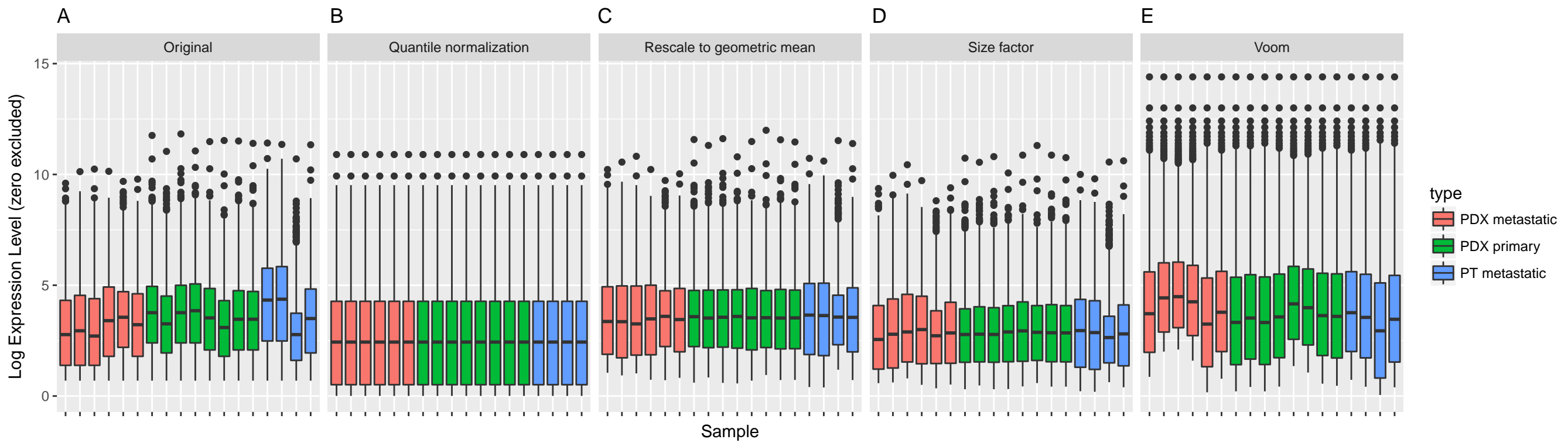


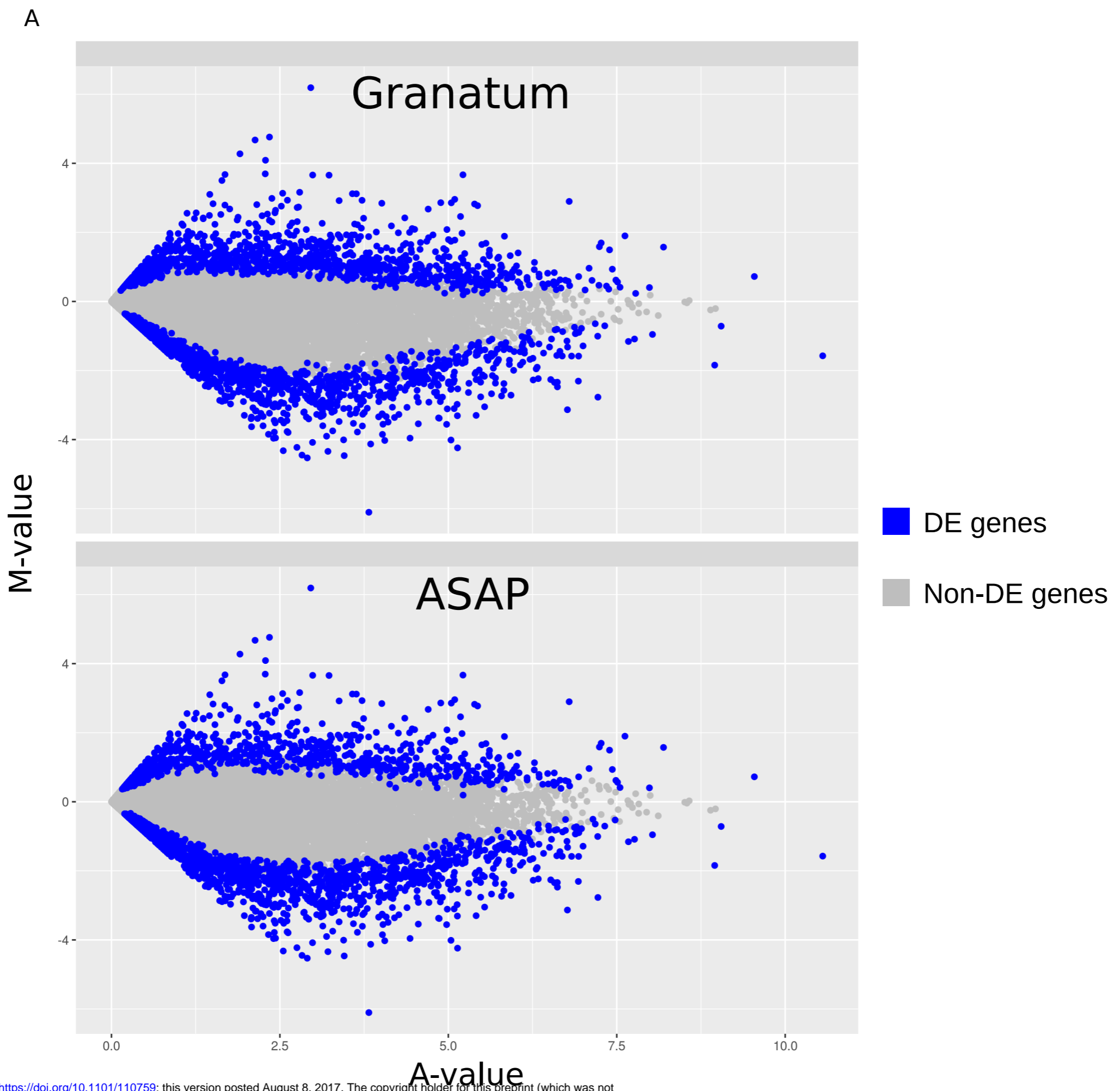
A



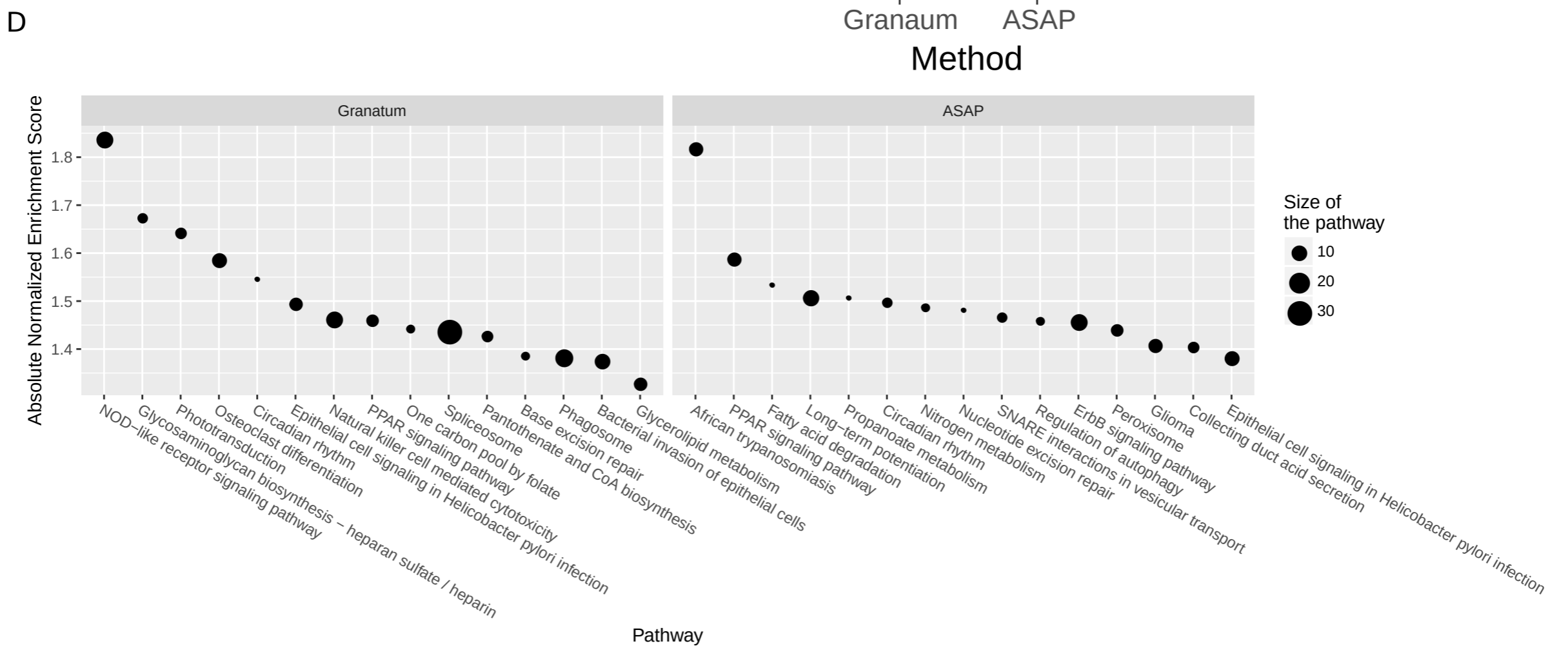
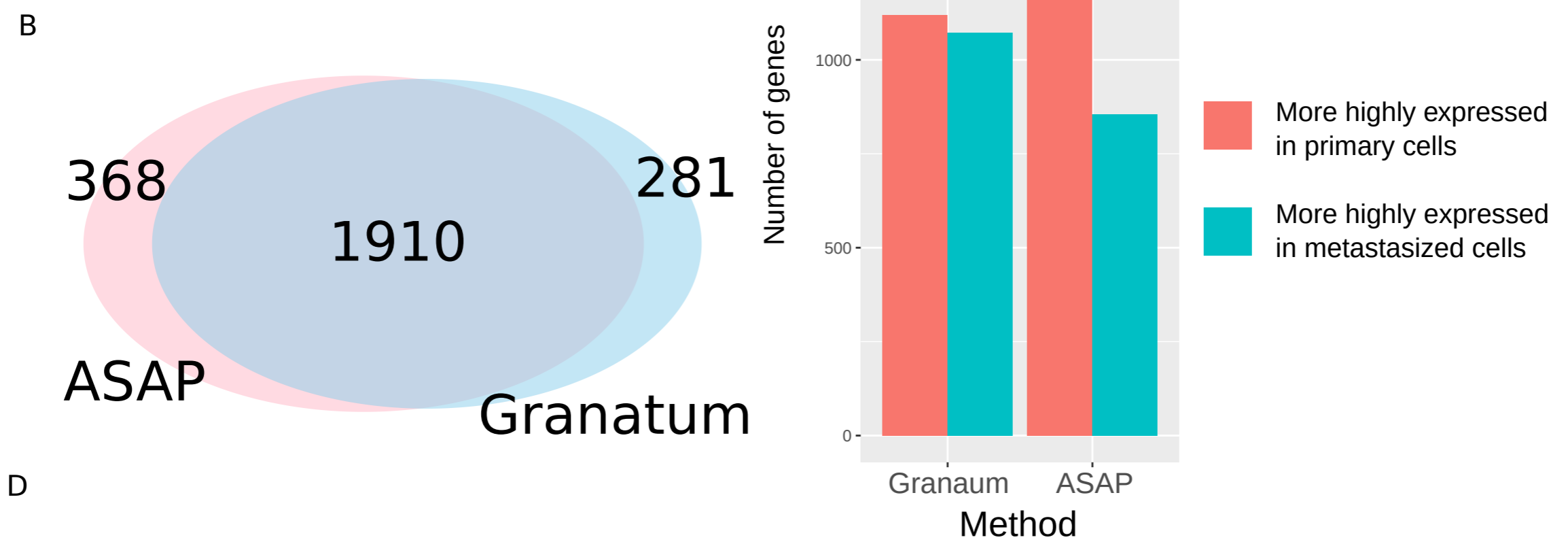
B



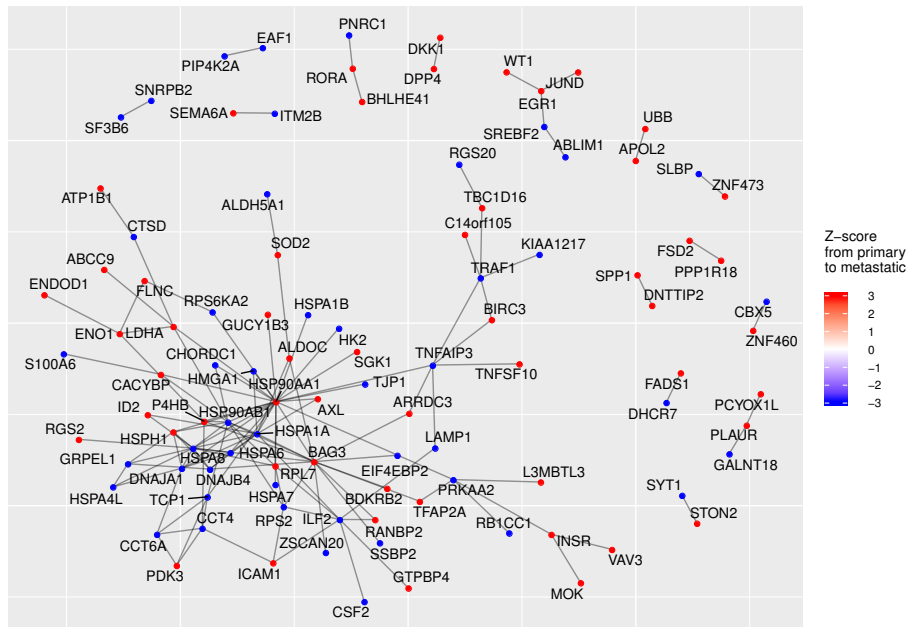




bioRxiv preprint doi: <https://doi.org/10.1101/110759>; this version posted August 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



A



B

