

PIVOT User Manual

Qin Zhu, Junhyong Kim Lab, University of Pennsylvania

July 10th, 2015

Contents

1	Introduction	1
2	Launch PIVOT	2
2.1	Launch with PIVOT Launcher	2
2.2	Launch using R command	2
3	Data Management	2
3.1	File Input	2
3.2	Design Information	3
3.3	Gene Filtering	4
3.4	Data Subsetting	6
4	Analysis Modules	6
4.1	Data Scale	6
4.2	Data Table	6
4.3	Basic Data Statistics	6
4.4	ERCC	7
4.5	Differential Expression Analysis	7
4.6	Monocle Cell State Ordering	9
4.7	Correlation Analysis	9
4.8	Feature Heatmap	9
4.9	Dimension Reduction	10
4.10	Classification	12
5	Report Module	13
6	System Control	13
6.1	Program State Saving and Loading	13
6.2	Launch New Session	15
7	Other Useful Information	16
7.1	Gene Expression Plot	16
7.2	Size of the Plot	16
7.3	Citation Infomation	16
7.4	Built-in Help Infomation	16

1 Introduction

This program is developed based on the Shiny framework, a set of R packages and a collection of scripts written by members of Junhyong Kim Lab at University of Pennsylvania. Its goal is to facilitate fast and interactive RNA-Seq data analysis and visualization. Current version of PIVOT supports routine RNA-Seq data analysis including normalization, differential expression analysis, dimension reduction, correlation analysis, clustering and classification. Users can complete workflows of DESeq2, monocle and scde package with just a few button clicks. All analysis reports can be exported, and the program state can be saved, loaded and shared.

See <http://kim.bio.upenn.edu/software/pivot.shtml> for more details.

2 Launch PIVOT

In R, use command `library(PIVOT)` to attach the package. Once loaded, you can use one of the following methods to launch PIVOT.

2.1 Launch with PIVOT Launcher

- By calling `pivot()`, you will see the window like the left image. This launcher serves as a gateway to multiple pivot modules you can use. Just check the boxes of modules that you want to load, and then press launch module.
- It is recommended that a user only launch needed modules, because using too many modules will lead to too many loaded DLLs (current R only support up to 100 DLLs per session) and potential namespace conflicts. The launcher also serves as a DLL monitor and cleaner. Simply press the clean DLLs button when there are too many DLLs loaded. Note that the best way to clean loaded DLLs is using Rstudio -> Session -> Restart R.

2.2 Launch using R command

- `pivot_module()` command will list out current available modules in PIVOT. Use `pivot(#ID_vector)` such as `pivot(c(1,3,5))` to launch selected modules.

3 Data Management

3.1 File Input

The input file format can be a set of count files in a folder (choose “directory”), or a single file containing all counts (choose “counts table”).

3.1.1 Input gene expression matrix

PIVOT support expression matrix in csv, txt, xls or xlsx formats. Choose proper settings on the left file input panel until the right “Loaded File Preview” correctly shows the data frame.

you need to make sure that the data table:

+ *Contains no NA or non-numeric values.*

+ *Does not have duplicated feature or sample names (PIVOT will alert the user if it detects any).*

+ *Sample names are column names (first row) and gene names are row names (first column). (Check transp

3.1.2 Input gene counts files as a directory

- If you specifies **directory**, then the next step is to choose which files you want to upload.
 - You can type key words in the filter to quickly select those files you are interested in.

File filter

CM20

- Sample_CM206-1b.htseq.exons.cnts.txt
- Sample_CM206-1c.htseq.exons.cnts.txt
- Sample_CM206-2b.htseq.exons.cnts.txt
- Sample_CM206-2c.htseq.exons.cnts.txt

Then, set proper normalization method and feature filtering criteria before pressing the submit button.

- If your data has already been processed by DESeq or other methods, please specify “none” in the normalization method.
- If DESeq failed on your data, one possibility is that you have low counts samples, which leads to all the genes contain at least one 0 in the counts matrix. You can either find out and remove these samples, or choose the “modified DESeq” normalization method.
- Before submitting data, please check the threshold. The program will first filter input counts based on row means or row sums, and then apply normalization. By default features with zero count will be filtered out.

Choose pre-filtering type:

☒ Row Mean ☐ Row Sum

Row Mean Threshold

0

Figure 1: alt text

3.2 Design Information

- The design information are used for sample point coloring and differential expression analysis. The group information can be conditions (treated, control), cell types or other metafeatures that the user is interested in testing. The batch information are used for control of batch effects in differential expression analysis.
- You can add group and/or batch info manually or using a **design-info** file. A design-info file should contain one ‘Sample’ column, and one ‘Group’ column and/or one ‘Batch’ column (column names must be ‘Sample’, ‘Group’, ‘Batch’, case-sensitive). You can make a design-info file in the manual mode, and download it for later use.
- Most modules allow you to color the samples with the design info.

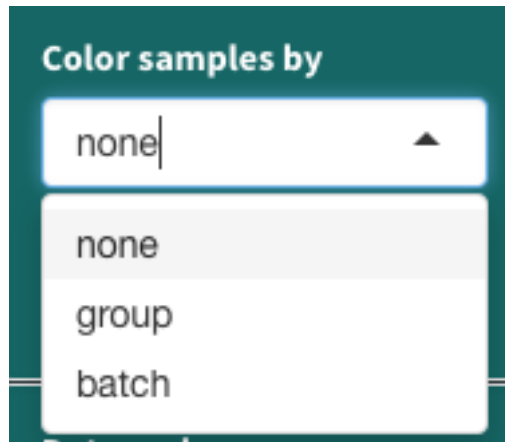
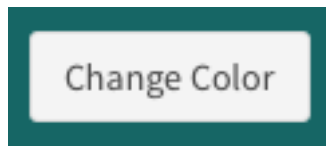


Figure 2: alt text



- You can press  to try different color sets!

3.3 Gene Filtering

- Filtration allows you to remove genes with too low or too high counts, or only perform analysis on a set of marker genes. The former requires the use of the expression filter, and the latter can be done using the marker feature filter.

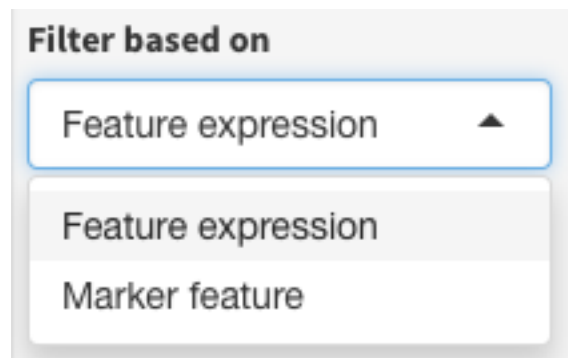


Figure 3: alt text

- Filter can be either positive (selecting the genes that satisfy the condition) or negative (delete the genes).
- Keep filtering the dataset means that the effect of filtration is additive. This mode is useful if you want to apply multiple criteria, e.g., first filter with marker features, then remove low count features, and finally remove features that's not expressed in a certain proportion of your samples. If this option is unchecked, you are always filtering the input dataset.
- Filter with renormalization means that after filtration, DESeq normalization will be re-performed on the filtered raw count dataset.

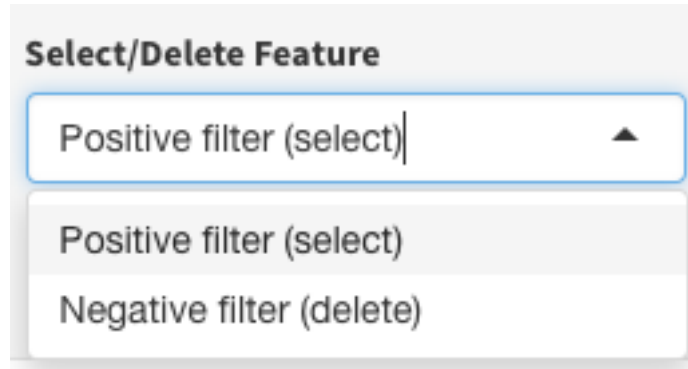


Figure 4: alt text

- The count range filter can be based on either raw counts or normalized counts. You can either input the range manually or use the slider.

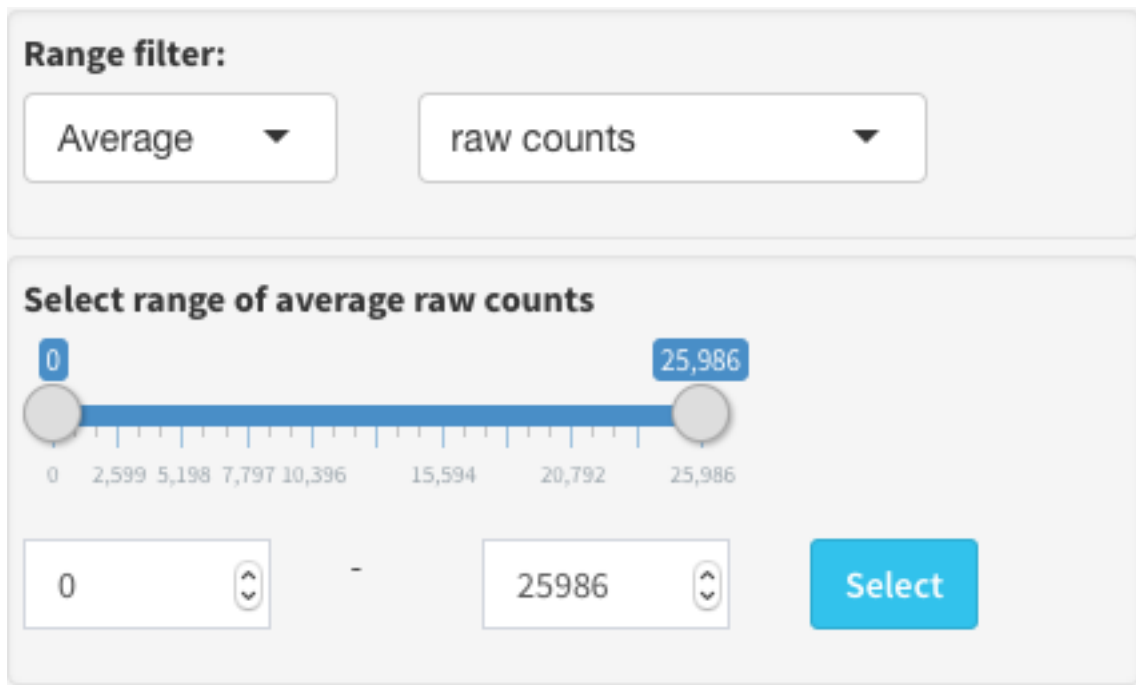



Figure 5: alt text

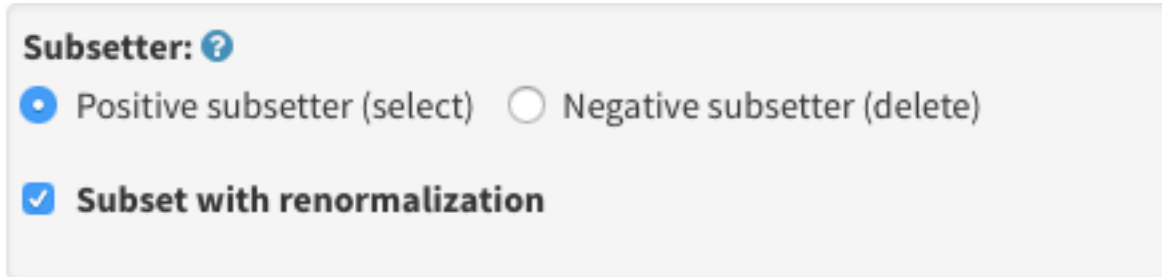
- You can provide a marker feature list to get a marker feature expression dataset.
 - The marker features should appear as the first column in your file.
 - Please note that some of the genes in your marker list may not be found in the dataset, because they may have already been removed due to 0 expression in all your samples.



- Applying  will return the data to the unfiltered state. Note that if you have performed subsetting prior to filtering, you will get a subsetting dataset with a non-zero expression feature set.

3.4 Data Subsetting

- Subsetter allows you to choose a subset of samples for analysis.
- You can choose whether or not the subset should be renormalized with DESeq.



The image shows a user interface for the 'Subset' function. It has a title 'Subset: ?' with a help icon. Below the title are two radio buttons: 'Positive subsetter (select)' which is selected, and 'Negative subsetter (delete)'. Below these is a checked checkbox labeled 'Subset with renormalization'.

Figure 6: alt text

- An implicit filtration will occur to get nonzero count genes for the subset. This procedure prevents some downstream analysis from breaking on 0s.

4 Analysis Modules

4.1 Data Scale

- For most analysis modules, you can choose one of the four data scales:
 - **counts (normalized)** : DESeq normalized counts;
 - **Log10** : $\log_{10}(\text{normalized_counts} + 1)$. Plus one to include zeros;
 - **Standardized** : Standardization (calculate Z-scores) is performed on the DESeq normalized counts;
 - **Log10 & Standardized** : Standardization (calculate Z-scores) is performed on $\log_{10}(\text{normalized_counts} + 1)$, assuming log-normal distribution.
- For each individual analysis, please choose the most proper data scale. Some modules have fixed data scale choice (e.g., raw counts input for DESeq differential analysis) so this option is not available.

4.2 Data Table

- You can download data table with different data scales and ordering. If your original data is multiple counts files in a folder, you can also download the combined raw count table or normalized table for single file input.
- The **relative frequency** of a gene is defined as its raw count divided by the total counts of the sample.

4.3 Basic Data Statistics

- This module contains simple statistics of your data, including some useful plots and table for quality control.

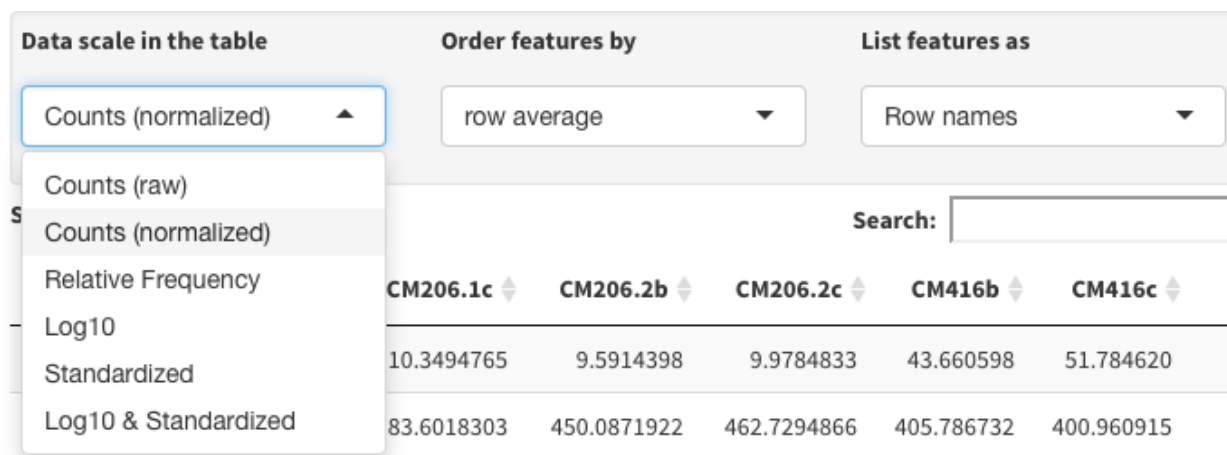


Figure 7: alt text

4.4 ERCC

- If your data contains ERCC spike-in, PIVOT will plot the ERCC read distribution and the estimated molecules based the standard table (https://tools.thermofisher.com/content/sfs/manuals/cms_095046.txt).
- The default parameter is 0.9 μ L ERCC with 1:4000000 dilution added per sample. The user will need to adjust the parameters according to the protocol used.

4.5 Differential Expression Analysis

4.5.1 DESeq Differential Expression Analysis

- This module is a graphical interface for the DESeq2 package (<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>). Because DESeq requires raw counts input, if the input file is a normalized counts table, this analysis will not be available.

4.5.2 SCDE

- This module is a graphical interface for the SCDE package (<http://hms-dbmi.github.io/scde/diffexp.html>).
- The SCDE error modeling must be performed first before you can use other SCDE analysis. For large dataset the modeling process can be very slow. You can monitor the progress in the background R session.
- You can use SCDE distance for hierarchical clustering and minimum-spanning-tree generation. There are three types of adjustment method you can choose: direct drop-out, reciprocal weighting and mode relative weighting. For details of these methods please check the SCDE website. Once a distance has been computed, it is loaded into PIVOT to be used in other modules.

4.5.3 Mann-Whitney Test

- Also known as wilcoxon rank sum test. The null hypothesis is that the distributions of the gene expression in the two groups has no difference and the alternative is that they differ by some non-zero location shift.
- You can choose the P adjustment method. The default method is Bonferroni correction.

Data scale in this module is
deseq normalized counts.

**Amount of ERCC added
(μ L)**

0.9

ERCC dilution, 1 :

4000000

Recompute

Figure 8: alt text

Compute Distance




-  Direct drop-out adjusted distance has been successfully loaded.
-  Reciprocal weighting adjusted distance has been successfully loaded.
-  Mode-relative weighting adjusted distance has been successfully loaded.

Figure 9: alt text

4.5.4 Monocle Differential Expression Analysis

- This module is a graphical interface for the Monocle package (<http://cole-trapnell-lab.github.io/monocle-release/>).
- This analysis is most proper for groups that represent the progress of a biological process, such as time of cell collection, cell state or media change. For details of this analysis please check the monocle paper and website.

4.6 Monocle Cell State Ordering

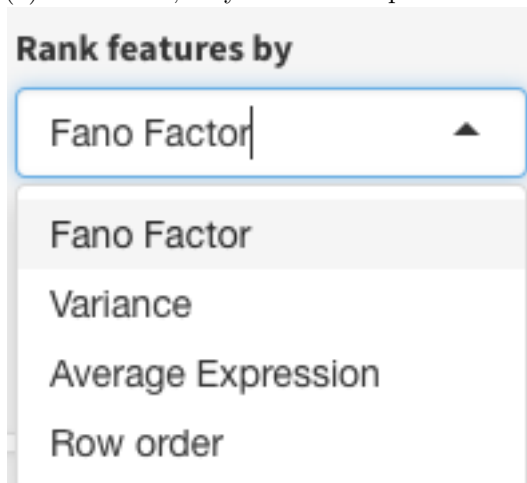
- Currently you can only use less than 1000 features for cell state ordering. It is recommended that you use a marker feature list, the top genes ranked by overdispersion or the significant differentially expressed genes for ordering.
- If you have provided group information, the comparison between monocle assigned state and your group will be available.
- You can click the gene list to view the expression level of the selected gene plotted on the graph or as a function of pseudotime.

4.7 Correlation Analysis

- You can generate pairwise scatterplot if you have less than 50 samples. This module will fail if you have too many samples.
- The sample correlation heatmap provides a more intuitive way of visualizing the correlation between your samples. If you specifies color by group, a color bar will be added to the heatmap to show the group info. You can only use the correlation distance for hierarchical clustering in a correlation heatmap.

4.8 Feature Heatmap

- You can rank features by fano factor, variance or mean expression. You are not allowed to use standardized data for ranking because all feature will have the same mean (0) and variance (1). However, if you want to plot in standardized scale, you can choose rank by “Row order”.



- The ranking of the heatmap is used for choosing the top ranked features to be plotted in the heatmap. You can use the slider to choose the features to be plotted. By default if the number of features is

greater than 500, only the top 500 features will be plotted. Using the manual input, you can choose any rank range other than the top 500. The only limit is that you can only plot 1000 features at a time.



Figure 10: alt text

- By default, the order of the samples (columns) and features (rows) shown in the heatmap will be determined by hierarchical clustering. Alternatively, if “Do not cluster sample/feature” is specified, the samples/features will be ordered by the rank. If the user also specifies “rank by row order”, then the heatmap order will be exactly the same as the input row/column order.

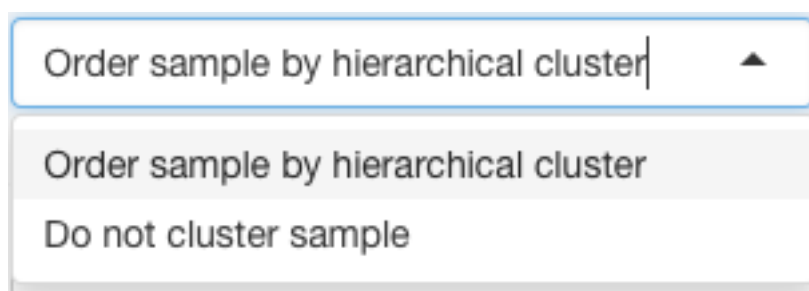
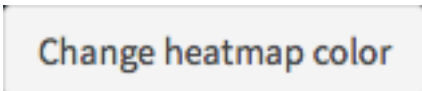


Figure 11: alt text

- d3heatmap package allows zoom in (click and drag) and zoom out (double click).

- Press  to try different colors!

4.9 Dimension Reduction

4.9.1 PCA

- You can look at any principal component. Percentage shown on each axis is the percent variance explained by that principal component.

4.9.2 T-SNE

- 1D, 2D and 3D T-SNE are results of 3 different t-SNE processes (dims = 1, 2 or 3).
- According to <http://lvdmaaten.github.io/tsne/>,

“Perplexity is a measure for information that is defined as 2 to the power of the Shannon entropy. The perplexity of a fair die with k sides is equal to k . In t-SNE, the perplexity may be viewed as a knob that sets the number of effective nearest neighbors. It is comparable with the number of nearest neighbors k that is employed in many manifold learners.”

PC1 (14.7%)

PC on X axis

PC1

PC1

PC2

PC3

PC4

PC5

PC6

PC7

PC on Y axis

PC2

Figure 12: alt text

Perplexity

1

Perform initial PCA step?

☒ Yes ☐ No

Figure 13: alt text

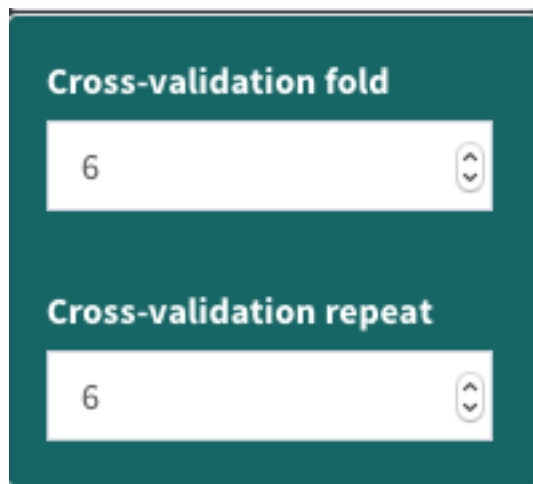
4.10 Classification

- This module is a graphical interface for the caret package (<http://topepo.github.io/caret/index.html>). Currently it allows the user to choose almost all classification models listed in (<http://topepo.github.io/caret/modelList.htm>).
- Some methods may require new packages to be installed. In such cases, the background R session will ask you to install it. Choose yes if you want to proceed.

```
1 package is needed for this model and is not installed. (kknn). Would you like to try to install it now?  
1: yes  
2: no  
Selection: 1|
```

Figure 14: alt text

- Users are expected to have the knowledge of which model is most suitable for their data. The details of these models can be found in the caret website.
- Many methods contains built-in feature selection. For those having explicit coefficients for the features, the feature coefficient table will be available for download.
- You can specify the cross-validation parameters to be used in the training.



The image shows a dark teal background with two white input fields. The first field is labeled "Cross-validation fold" and contains the number "6". The second field is labeled "Cross-validation repeat" and also contains the number "6". Both fields have a small circular icon with up and down arrows on the right side, indicating they are dropdown menus.

Figure 15: alt text

- For methods containing internal feature selection and explicit feature coefficients, the information will be available in the “Feature Coefficient” table.
- The trained model can be used as a classifier for new data, or be used for the testing of the model. For example, if you train the classifier using data of group A and group B, and then perform testing on group A-t and group B-t. Then samples in group A-t and B-t will be classified as group A or group B, so that you can tell how well the model is performing. Alternatively, you can use the model to classify the unknow samples in e.g. group C.

5 Report Module

- The code has been pre-written for you – you only need to add it. Please always run the module before you add it to report, because PIVOT requires the module has a last state to capture all the parameters.

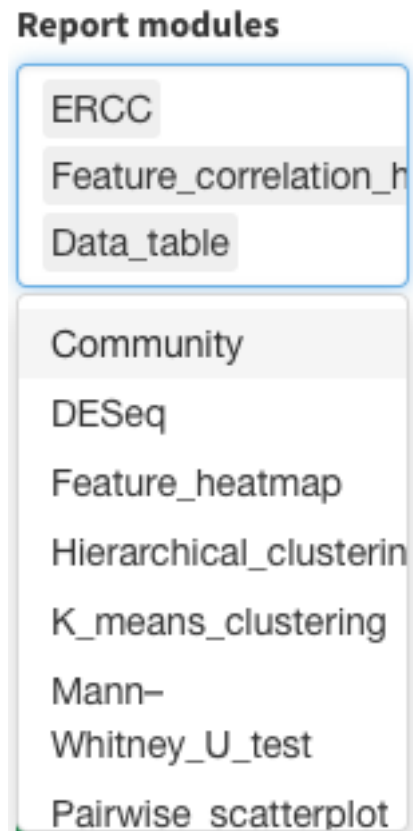


Figure 16: alt text

- The report code is based on R-markdown v2 (<http://rmarkdown.rstudio.com/>). You can add comments, change titles or modify the code (not recommended) according to the syntax.

6 System Control

- The system control menu is located at the top right corner. Users can use this panel to save the program state, launch new session or go to data management panels.

6.1 Program State Saving and Loading

- You can save the program state as an R data object. In this way you can make sure that you won't lose your analysis progress, and you can share the state with others.
- To load the saved state, go to File panel and choose PIVOT state in input file type. The session will auto refresh and immediately switch to the loaded state when the state uploading is complete.
- Please note that although every analysis result will be kept in the state, PIVOT is not able to return the parameter choices to the ones you chose when the analysis was performed. In other words, the

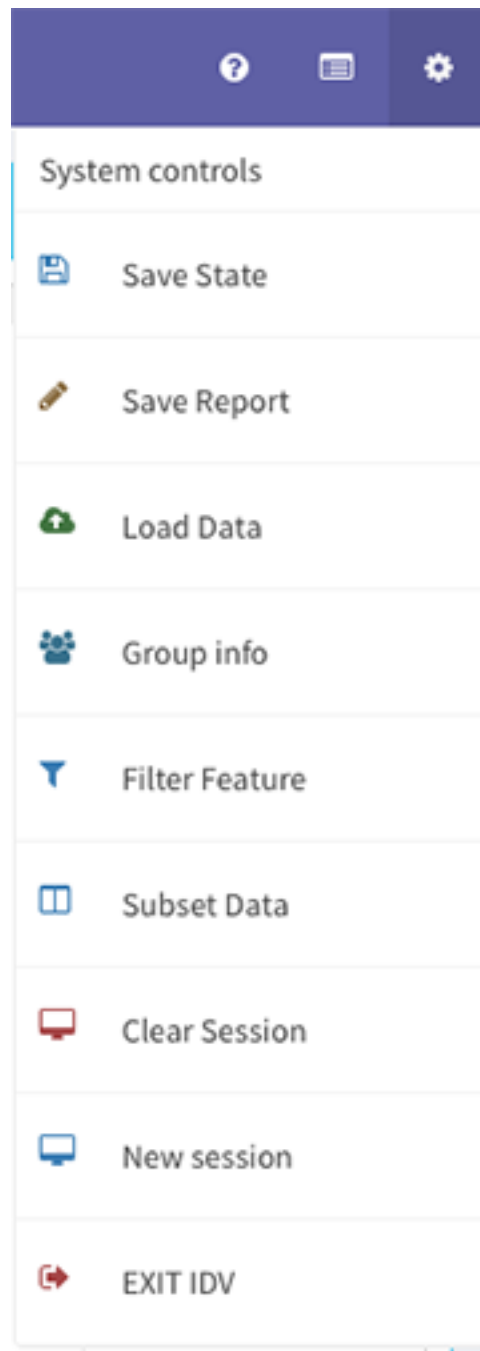


Figure 17: alt text

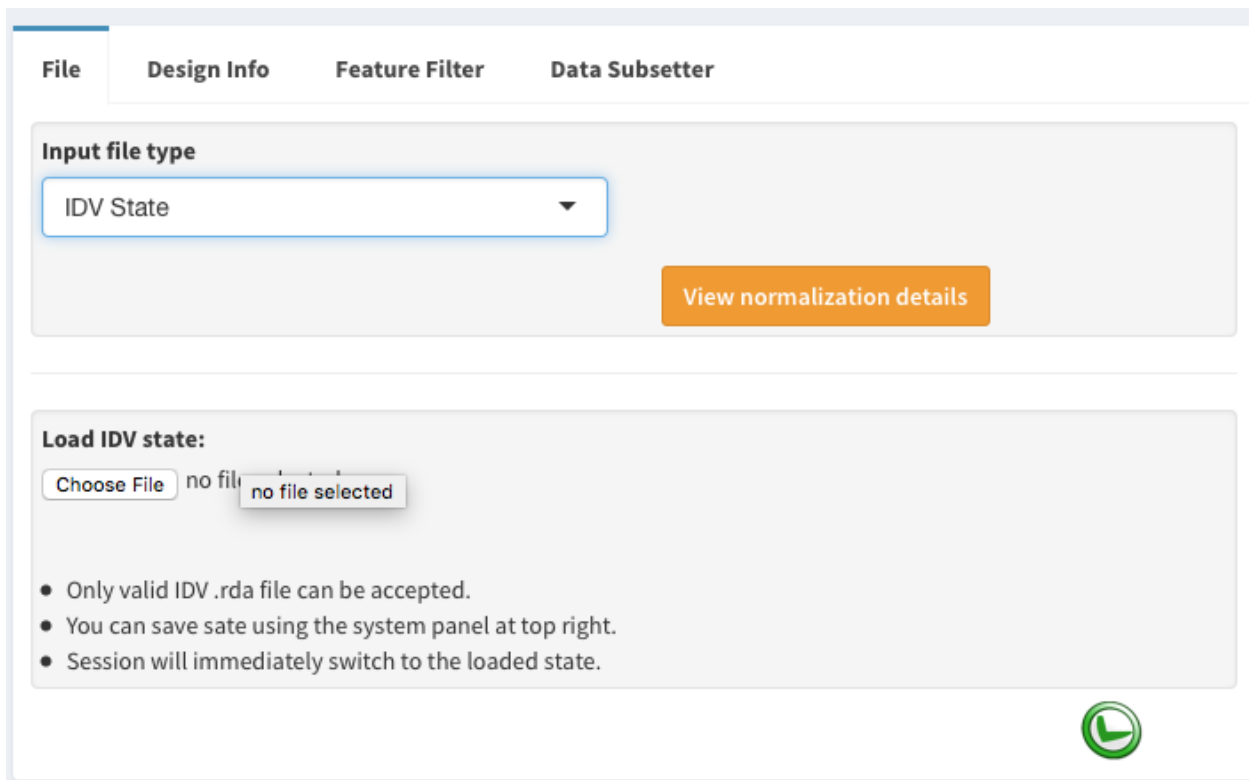


Figure 18: alt text

results and plots shown in the loaded state will stay the same, but the new parameter UI will return to the defaults.

- Each time you exit, PIVOT will automatically save the state into the background R session. If you don't close the R session, or save the workspace image before exiting R, the next time you launch PIVOT it will automatically load the state for you.



- To clear the state, click

6.2 Launch New Session



- Clicking the button will launch new PIVOT session for you. In this way you can do different things in different sessions. There is a limitation for this: because R is single-thread, you can only perform one analysis at a time. While R is busy computing in one session, the other sessions will have to wait in a queue.
- If you really want to simultaneously perform multiple analysis in multiple sessions, currently the only way is to open multiple copies of R, and launch PIVOT using the command "pivot()" in each R session.

7 Other Useful Information

7.1 Gene Expression Plot

- In all differential expression analysis modules, you can click the result table to view the expression plot of individual genes.

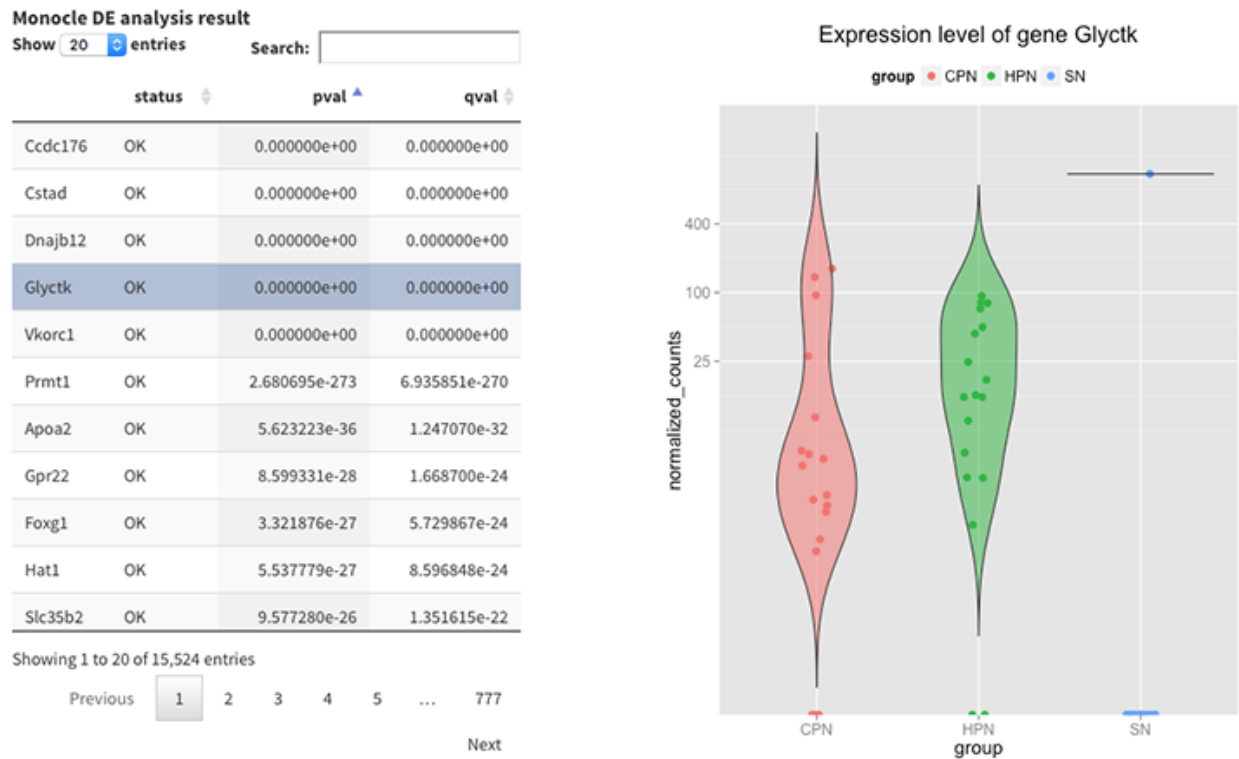


Figure 19: alt text


7.2 Size of the Plot

- You can change the width of most plots by resizing the window.
- You can resize ggvis plots by using the bottom right triangle, and download the plot by clicking the cog icon on top right.

7.3 Citation Infomation

- The citation and licensing information can be found at the bottom of each module.

7.4 Built-in Help Infomation

- You can find  in many places in PIVOT. You can click it to view the relavant information of the module.

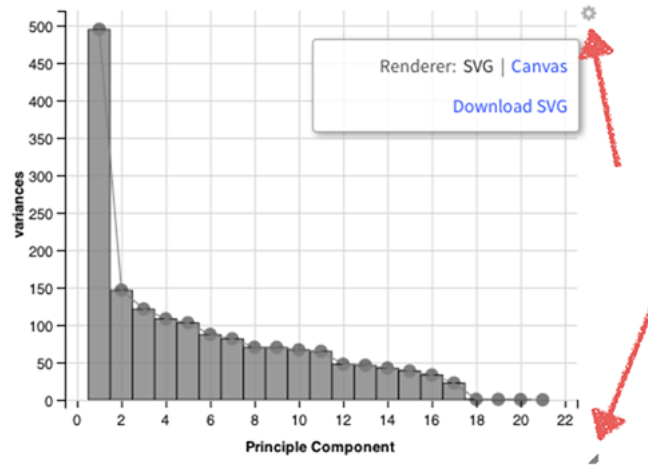


Figure 20: alt text

Citation

1. Cole Trapnell and Davide Cacchiarelli et al (2014): The dynamics and regulators of cell fate decisions are revealed by pseudo-temporal ordering of single cells. *Nature Biotechnology*
2. Monocle Website: <http://cole-trapnell-lab.github.io/monocle-release/>
3. Monocle was written by Cole Trapnell with input from Davide Cacchiarelli and is provided under the OSI-approved Artistic License (version 2.0).

Figure 21: alt text

- Many parameter inputs has tooltips containing the relavant info.

Dispersion bottom cutoff

2

E.g, Setting the cutoff to 2 identifies genes that are more than two standard deviations away from the average dispersion within a bin.

Figure 22: alt text