

# KimLabIDV User Manual

*Junhyong Kim Lab, University of Pennsylvania*

*Dec 9th, 2015*

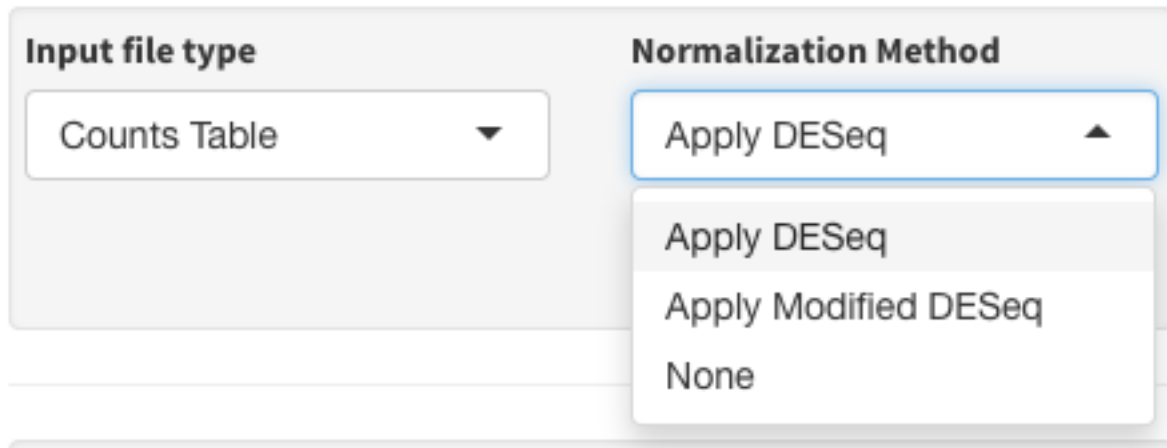
## Contents

<b>1</b>	<b>Data Management</b>	<b>2</b>
1.1	File Input . . . . .	2
1.2	Design Information . . . . .	3
1.3	Gene Filtering . . . . .	4
1.4	Data Subsetting . . . . .	5
<b>2</b>	<b>Analysis Modules</b>	<b>6</b>
2.1	Data Scale . . . . .	6
2.2	Data Table . . . . .	6
2.3	Basic Data Statistics . . . . .	7
2.4	ERCC . . . . .	7
2.5	Differential Expression Analysis . . . . .	7
2.6	Monocle Cell State Ordering . . . . .	8
2.7	Correlation Analysis . . . . .	9
2.8	Feature Heatmap . . . . .	9
2.9	Dimension Reduction . . . . .	10
2.10	Classification . . . . .	11
<b>3</b>	<b>Report Module</b>	<b>11</b>
<b>4</b>	<b>System Control</b>	<b>12</b>
4.1	Program State Saving and Loading . . . . .	13
4.2	Launch New Session . . . . .	14
<b>5</b>	<b>Other Useful Information</b>	<b>15</b>
5.1	Gene Expression Plot . . . . .	15
5.2	Size of the Plot . . . . .	15
5.3	Built-in Help Infomation . . . . .	16
5.4	Citation Infomation . . . . .	16

# 1 Data Management

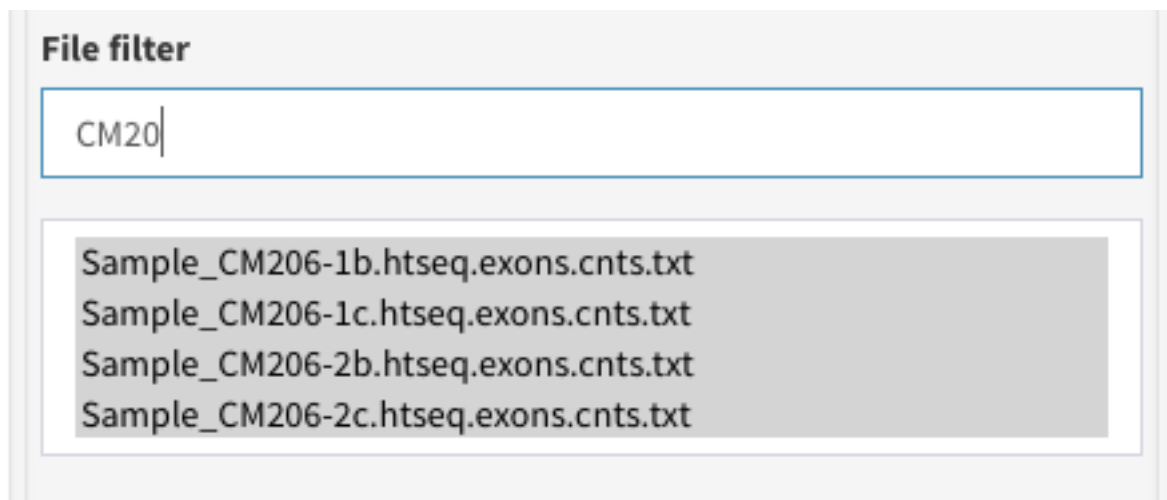
## 1.1 File Input

- 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”).
- 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.



The image shows two dropdown menus. The first, labeled 'Input file type', has 'Counts Table' selected. The second, labeled 'Normalization Method', has 'Apply DESeq' selected, and its dropdown menu is open, showing three options: 'Apply DESeq', 'Apply Modified DESeq', and 'None'.

- 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.



The image shows a 'File filter' section. It has a text input field containing 'CM20'. Below the input field is a list of files that match the filter:

- 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

- If you choose to upload a **single file** containing the raw counts table, you need to make sure that the data table are organized as follows:
  - *Sample names are column names (first row) and gene names are row names (first column).*
  - *No NA value is allowed.*
  - *Illegal R strings (e.g., sample-A(2)) are not allowed to be used as names.*
  - *The file must be a csv/txt file.*

Loaded Data Preview					
Show <input type="text" value="20"/> entries		Search: <input type="text"/>			
	CM206.1b	CM206.1c	CM206.2b	CM206.2c	C
A1BG	0	0	0	0	
A1BG-AS1	1	0	0	0	
A1CF	0	0	0	0	
A2M	1028	900	441	405	
A2M-AS1	4	4	2	1	
A2ML1	0	0	0	0	

- 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.

<b>Choose pre-filtering type:</b> <input checked="" type="radio"/> Row Mean <input type="radio"/> Row Sum	<b>Row Mean Threshold</b> <input type="text" value="0"/>
--	---

## 1.2 Design Information

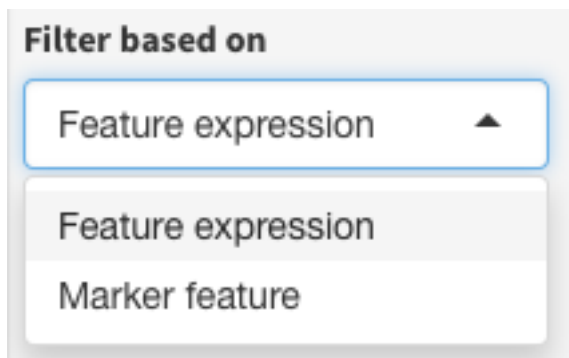
- The design information are used for sample point coloring and differential expression analysis. The group information can be conditions (treated, control), cell types or other metafeatures that the user is interested in testing. The batch information are used for control of batch effects in differential 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.



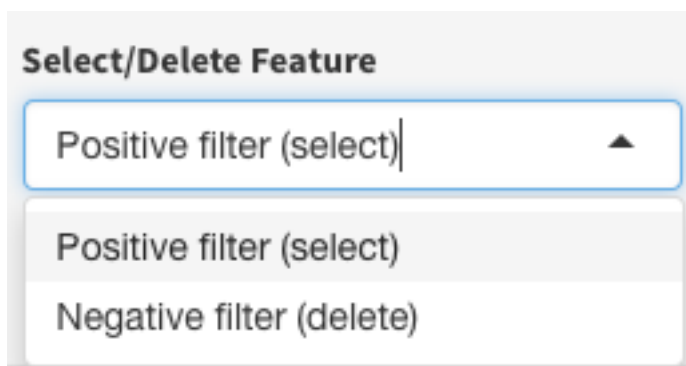
- You can press  to try different color sets!

### 1.3 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.



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

**Range filter:**

Average ▼ raw counts ▼

**Select range of average raw counts**


0 25,986

0 2,599 5,198 7,797 10,396 15,594 20,792 25,986

0 - 25986 Select

- 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.

## 1.4 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.

**Subsetter: ?**

☒ Positive subsetter (select) ☐ Negative subsetter (delete)

☒ **Subset with renormalization**

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

## 2 Analysis Modules

### 2.1 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.

### 2.2 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.

Data scale in the table	Order features by			List features as	
Counts (normalized) ▲	row average ▼			Row names ▼	
Counts (raw)					
Counts (normalized)					
Relative Frequency					
Log10					
Standardized					
Log10 & Standardized					

Search: <input type="text"/>					
CM206.1c	CM206.2b	CM206.2c	CM416b	CM416c	
10.3494765	9.5914398	9.9784833	43.660598	51.784620	:
83.6018303	450.0871922	462.7294866	405.786732	400.960915	:

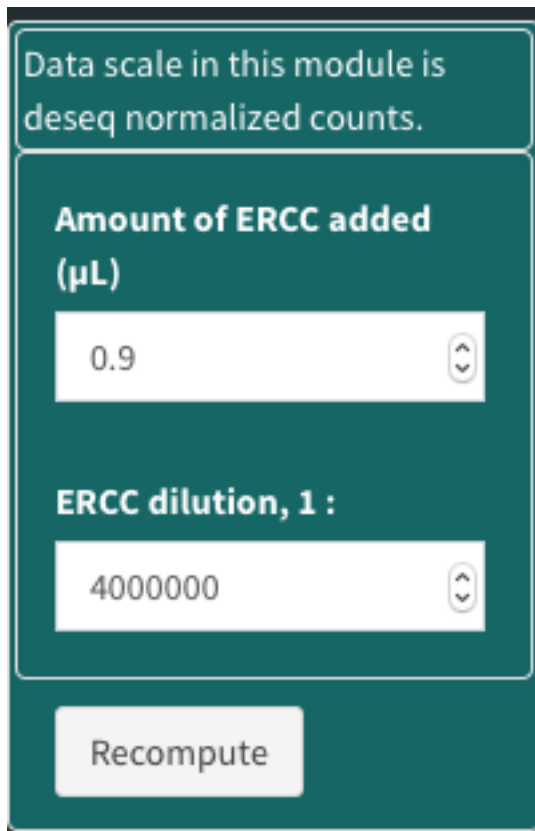
- The **relative frequency** of a gene is defined as its raw count divided by the total counts of the sample.

## 2.3 Basic Data Statistics

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

## 2.4 ERCC

- If your data contains ERCC spike-in, IDV will plot the ERCC read distribution and the estimated molecules based the standard table ([https://tools.thermofisher.com/content/sfs/manuals/cms\\_095046.txt](https://tools.thermofisher.com/content/sfs/manuals/cms_095046.txt)).
- The default parameter is 0.9  $\mu\text{L}$  ERCC with 1:4000000 dilution added per sample. The user will need to adjust the parameters according to the protocol used.



The screenshot shows a web interface for the ERCC module. It has a dark teal background. At the top, a light teal box contains the text "Data scale in this module is deseq normalized counts." Below this, there are two input fields. The first is labeled "Amount of ERCC added ( $\mu\text{L}$ )" and contains the value "0.9". The second is labeled "ERCC dilution, 1 :" and contains the value "4000000". Both fields have up and down arrow icons on their right sides. At the bottom of the interface is a light gray button labeled "Recompute".

## 2.5 Differential Expression Analysis

### 2.5.1 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.

### 2.5.2 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 IDV to be used in other modules.

---

**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.**

### 2.5.3 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.

### 2.5.4 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.

## 2.6 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.

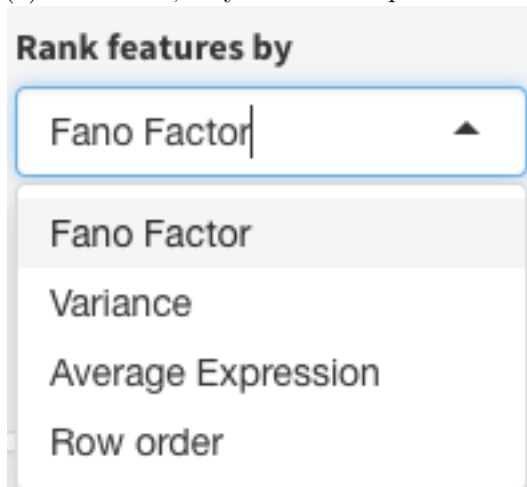


## 2.7 Correlation Analysis

- 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.

## 2.8 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”.



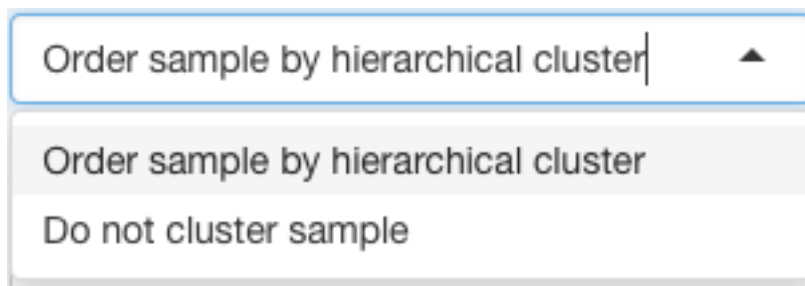
A dropdown menu titled "Rank features by". The selected option is "Fano Factor". Other visible options are "Fano Factor", "Variance", "Average Expression", and "Row order".

- 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.



A slider control for selecting the number of features to plot. The slider is labeled "Rank" and has a range from 1 to 500. The current selection is 1. To the right of the slider are input fields for "Min:" (1) and "Max:" (500), and an "Update Range" button.

- 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.



A dropdown menu titled "Order sample by". The selected option is "Order sample by hierarchical cluster". Other visible options are "Order sample by hierarchical cluster" and "Do not cluster sample".

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

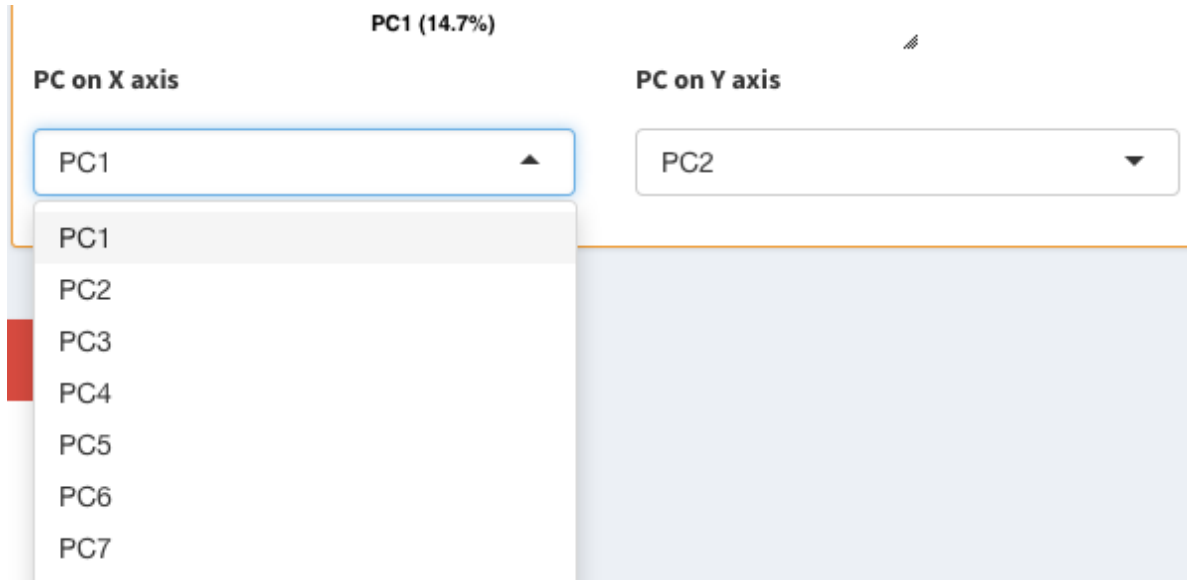
Change heatmap color

- Press to try different colors!

## 2.9 Dimension Reduction

### 2.9.1 PCA

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



### 2.9.2 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.”*

**Perplexity**

1

**Perform initial PCA step?**

☒ Yes ☐ No

## 2.10 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|
```

- 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.



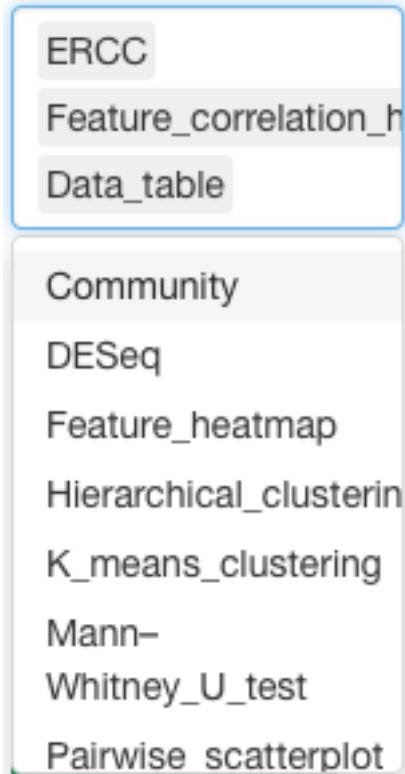
The image shows a dark green rectangular panel with two sections. The top section is titled "Cross-validation fold" in white text, and below it is a white input field containing the number "6" with a circular spinner icon to its right. The bottom section is titled "Cross-validation repeat" in white text, and below it is another white input field containing the number "6" with a circular spinner icon to its right.

- 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 unknown samples in e.g. group C.

## 3 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 IDV requires the module has a last state to capture all the parameters.

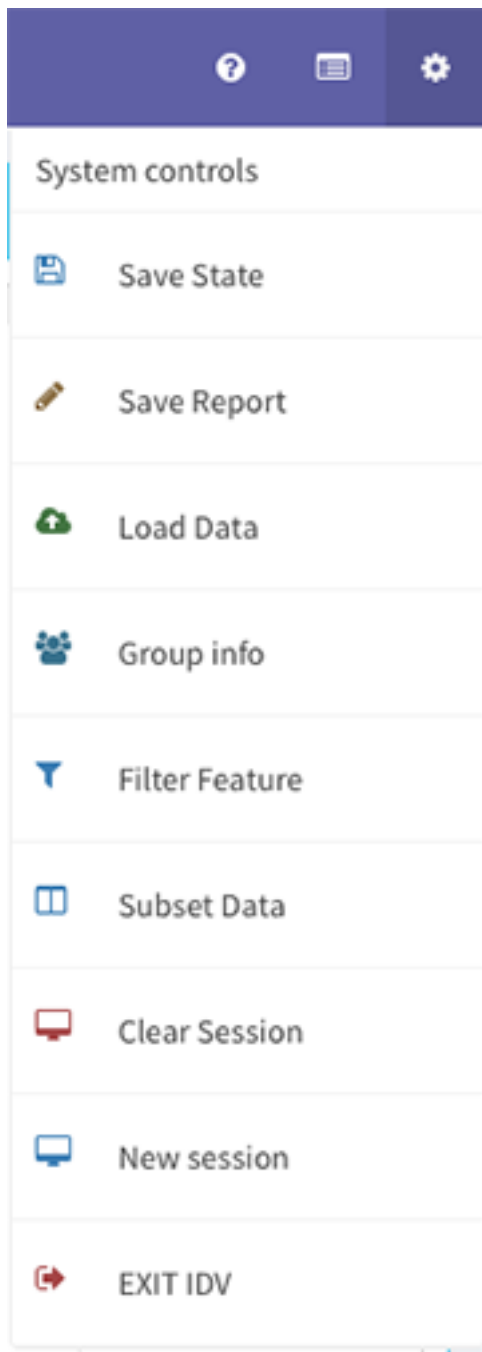
## Report modules



- 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.

## 4 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.



#### 4.1 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 IDV state in input file type. The session will auto refresh and immediately switch to the loaded state when the state uploading is complete.

The screenshot shows the IDV software interface with the 'File' tab selected. The 'Input file type' dropdown menu is set to 'IDV State'. To the right of this dropdown is an orange button labeled 'View normalization details'. Below this, the 'Load IDV state:' section contains a 'Choose File' button and a status indicator that says 'no file selected'. A list of instructions is provided: 'Only valid IDV .rda file can be accepted.', 'You can save state using the system panel at top right.', and 'Session will immediately switch to the loaded state.' A green circular icon is located at the bottom right of the interface.


- Please note that although every analysis result will be kept in the state, IDV is not able to return the parameter choices to the ones you chose when the analysis was performed. In other words, the 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, IDV 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 IDV it will automatically load the state for you.



- To clear the state, click  .

## 4.2 Launch New Session



- Clicking the  button will launch new IDV 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 IDV using the command “runIDV()” in each R session.

## 5 Other Useful Information

### 5.1 Gene Expression Plot

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

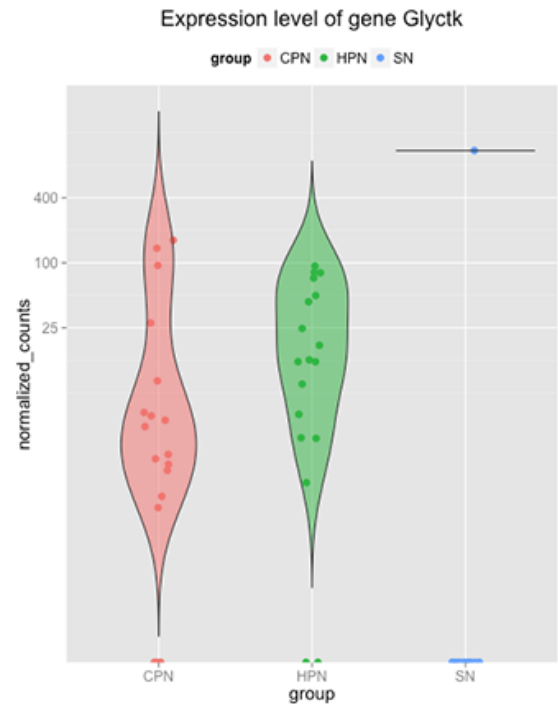
Monocle DE analysis result

Show 20 entries Search:

	status	pval	qval
Ccdc176	OK	0.000000e+00	0.000000e+00
Cstad	OK	0.000000e+00	0.000000e+00
Dnajb12	OK	0.000000e+00	0.000000e+00
Glyctk	OK	0.000000e+00	0.000000e+00
Vkorc1	OK	0.000000e+00	0.000000e+00
Prmt1	OK	2.680695e-273	6.935851e-270
Apoa2	OK	5.623223e-36	1.247070e-32
Gpr22	OK	8.599331e-28	1.668700e-24
Foxg1	OK	3.321876e-27	5.729867e-24
Hat1	OK	5.537779e-27	8.596848e-24
Slc35b2	OK	9.577280e-26	1.351615e-22

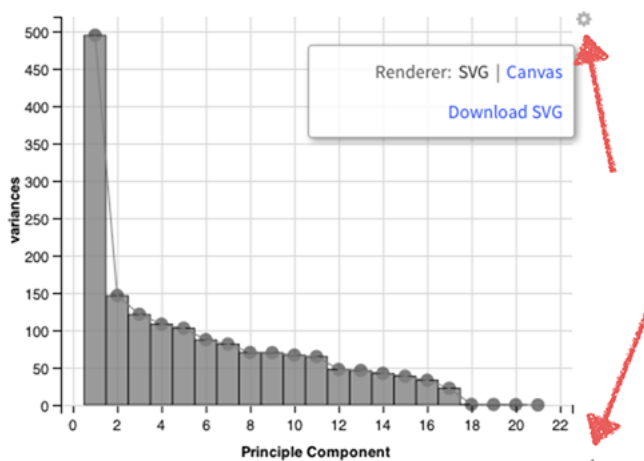
Showing 1 to 20 of 15,524 entries

Previous 1 2 3 4 5 ... 777 Next




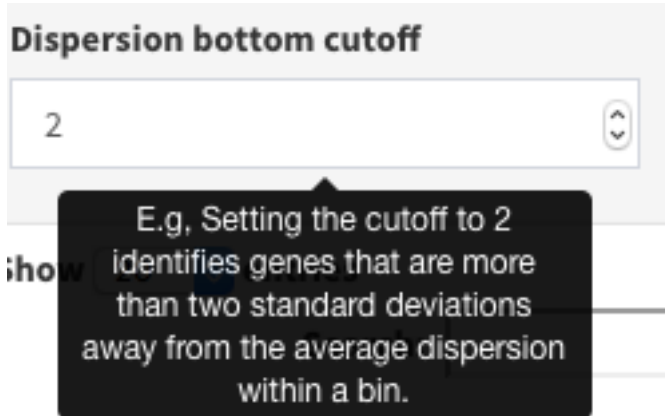
### 5.2 Size of the Plot

- 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.



### 5.3 Built-in Help Information

- You can find  in many places in IDV. You can click it to view the relevant information of the module.
- Many parameter inputs have tooltips containing the relevant info.



### 5.4 Citation Information

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

