Filter features

A common task in bulk and single-cell RNA-Seq analysis is to filter the data to include only informative genes. Because there is no gold standard for what makes a gene informative or not, and ideal gene filtering criteria depend on your experimental design and research question, Partek Flow has a wide variety of flexible filtering options.

Filter features task can be invoked from any counts or single cell data node. Noise Reduction and Statistics Based filters take each feature and perform the specified calculation across all of the cells. The filter is applied to the values in the selected data node and the output is a filtered version of the input data node.

In the task dialog, select the filter option to activate the filter type and configure the filter, then click **Finish** to run.

Noise reduction filter

The **Noise reduction filter** lets you exclude features that meet basic criteria (Figure 1).
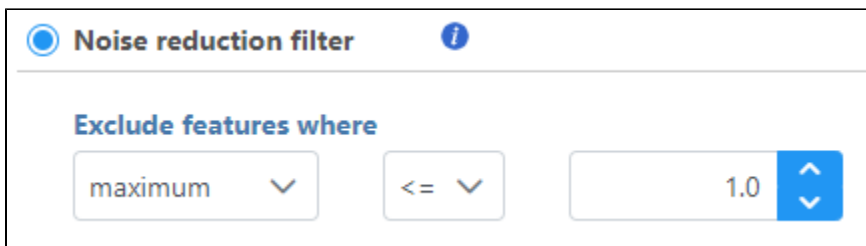

Figure 8. Noise reduction filter

Descriptive statistics you can choose are:

- Coefficient of variation: std. dev divided by mean of the feature
- Geometric mean: nth root of the product of the n numbers, n is the number of features
- Maximum: the highest value of a feature
- Mean: the average value of a feature
- Median: value of the mid point of a feature
- Minimum: lowest value of a feature
- Range: the difference between the highest and lowest values of a feature
- Std dev.: the square root of the variance
- Sum: total value of the feature
- Variance: the average of the squared differences from the mean
- Dispersion: variance divided by mean of the feature

For each of these you can choose to exclude features that are:

- <: less than
- <=: less than or equal to
- == equal to
- >: greater than
- >=: greater than or equal to

The threshold is set using the text box. The input must be a number; it can be an integer or decimal, positive or negative.

If you select value, you can also choose a percentage of samples or cells that must meet the criteria for the feature to be excluded (Figure 2).
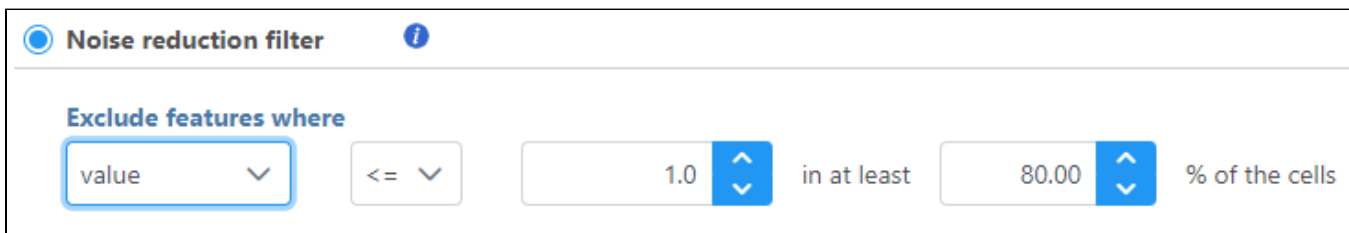

Figure 9. Selecting value exposes additional options

Statistics based filter

The **Statistics based filter** lets you include a number or percentile of genes based on descriptive statistics (Figure 3).
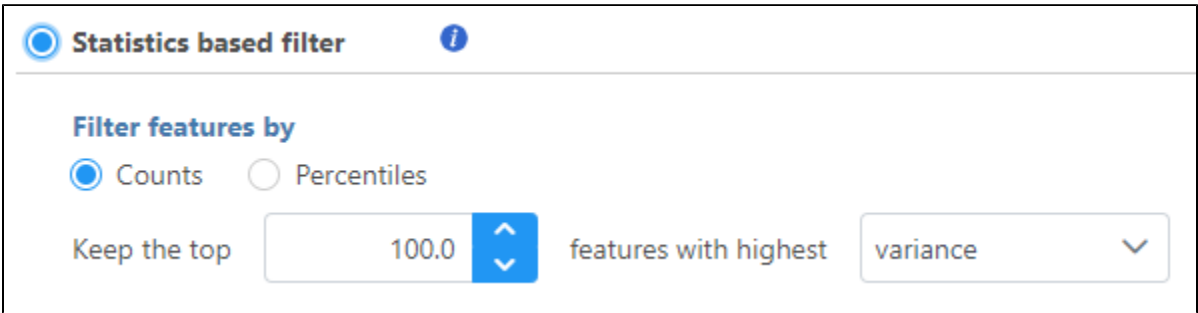

Figure 10. Statistics based filter

Select Counts to specify a number of top features to include or select Percentiles to specify the top percentile of features to include.

Descriptive statistics you can choose are:

- Coefficient of variance
- Geometric mean
- Maximum
- Mean
- Median
- Minimum
- Range
- Standard deviation (std dev)
- Sum
- Variance
- Dispersion

Feature metadata filter

If the data linked to feature (gene) annotation, different fields in the annotation can be used to filter, e.g. genomic location information, gene biotype information etc (Figure 4)
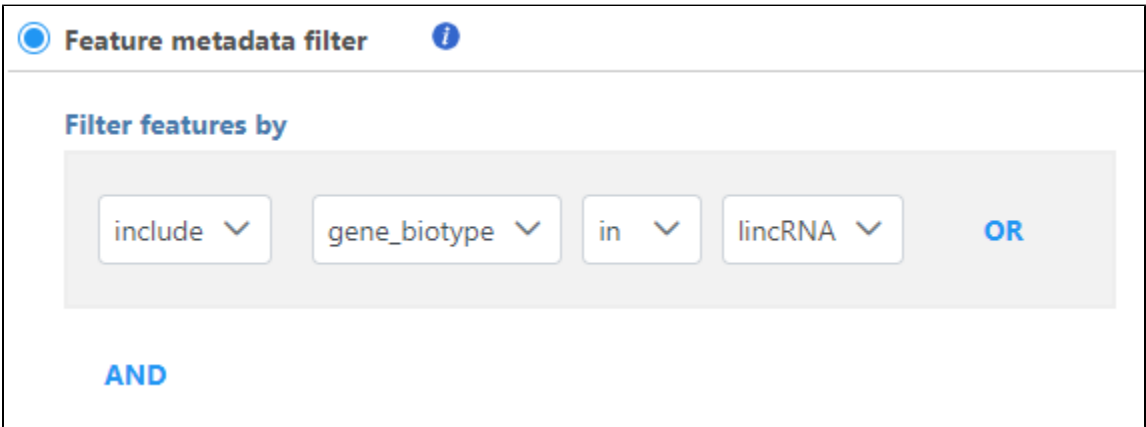

Figure 11. Filter features based on feature annotation fields

You can specify logical operation using different annotation field information.

Feature list filter

You can filter features based on a feature lists (Figure 5).

Figure 12. Feature list filter

If you have added feature lists in Partek Flow using the List management feature, the filter using **Saved list** option will be available. Otherwise, you can specify a **Manual list** by typing in the *Filter criteria* box.

If you choose *Saved list*, the drop-down list will display all the feature lists added in List management; If you choose *Manual list*, you can manually type in the feature IDs/names in the box, one feature per row.

You can choose to include or exclude features in any list that you have added.

Use the Feature identifier option to choose which identifier from your annotation matches the values in the feature list.

Filter features task report

The filter features task report lists the filter criteria, reports distribution statistics for the remaining features, and indicates the number and percentage of features that passed the filter (Figure 6).

Filter criteria   Exclude features where maximum <= 10.0

Feature distribution (9 samples; 427 genes)

Optional columns

Sample name	Min	2nd min	Max	Mean	Median	Q1	Q3	Missing
SRR592573	0	2.1E-6	3,148.0	149.4	53.0	11.0	161.0	0
SRR592574	0	1.1E-8	4,916.0	223.0	78.0	18.1	229.0	0
SRR592575	0	0.1	5,489.0	284.5	105.9	23.5	310.5	0
SRR592576	1.0	3.0	4,479.0	280.9	126.5	31.8	321.5	0
SRR592577	2.1	3.0	4,052.0	270.3	109.0	33.0	325.5	0
SRR592578	0.7	2.2	4,845.0	312.4	133.0	36.0	370.0	0
SRR592579	0	3.5E-22	2,392.0	124.4	53.0	13.5	134.5	0
SRR592580	0	5E-9	4,475.0	261.3	106.0	25.0	294.5	0
SRR592581	0	0.7	3,087.0	177.9	84.0	19.0	209.0	0
All samples	0	3.5E-22	5,489.0	231.6	88.5	21.2	253.0	0

Rows per page 25 ▼   ‹‹ ‹   (1 of 1)   › ››

⬇ Download

Of the 506 genes, 84.39% passed the filter (427 passed, 79 filtered out)

Figure 13. Feature filter report table

If the input was a count matrix data node, sample box plot and sample histograms are provided to show the distribution of features after filtering. These plots are not available if the input was a single cell counts data node.
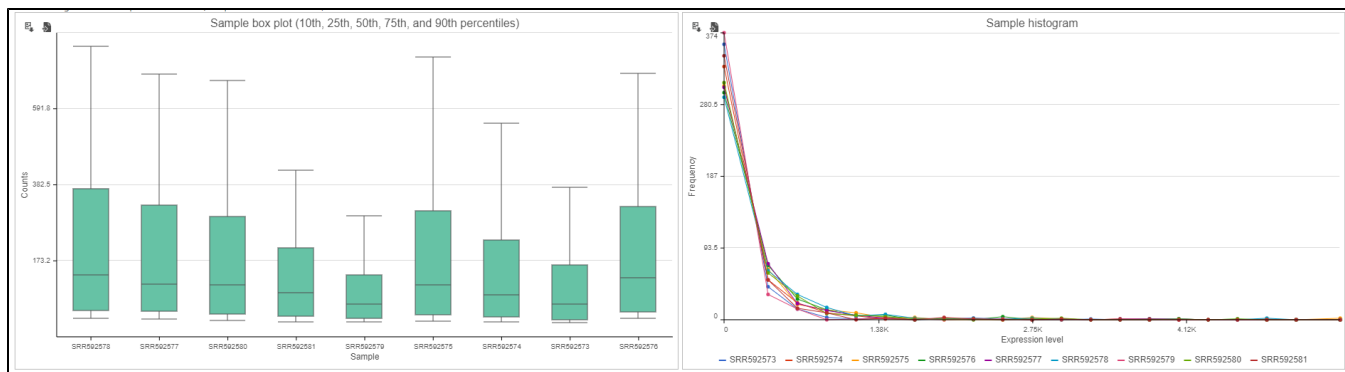
Figure 14. Filter features report figures

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.

Your Rating: ☆☆☆☆☆   Results: ★★★★★ 30 rates