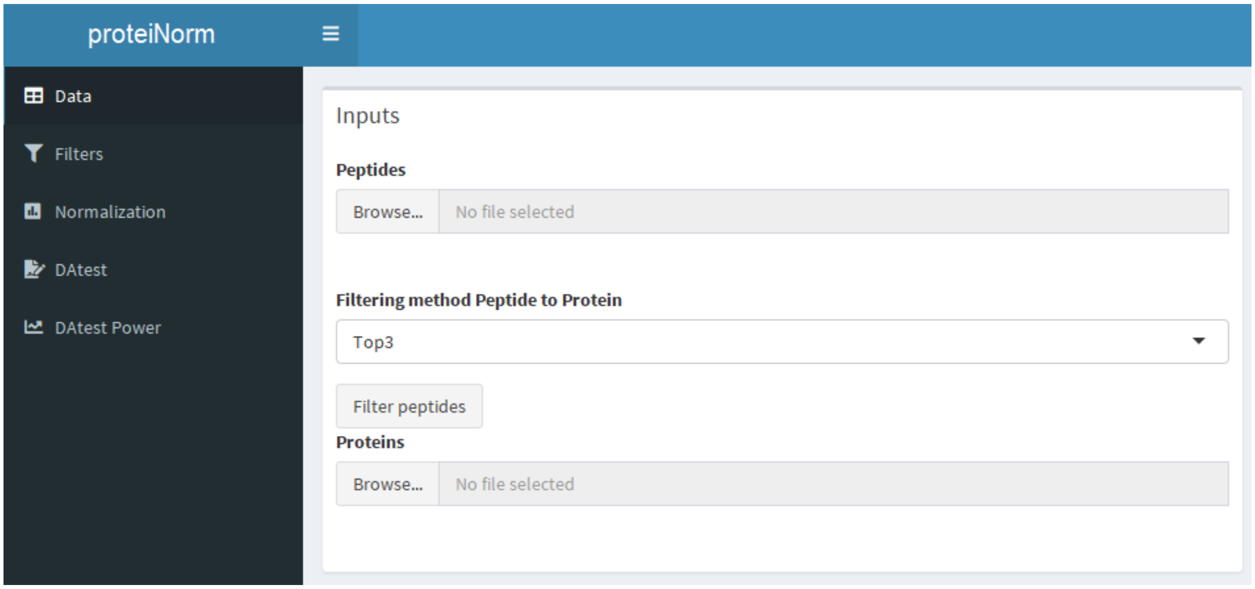
Data-tab (uploading data)

As input, proteiNorm expects tab-separated peptide (optional) and protein data (not on logarithmic scale) as produced by software such as MaxQuant, where each row represents a peptide or protein and the column names of the measured intensities (samples) beginning with “Reporter intensity corrected” followed by an integer and an optional label (e.g. “Reporter intensity corrected 5 TMT2”) for TMT experiments. The column names for samples in label free experiments should begin with “Intensity” followed by an integer (“Intensity 01”). In addition, for peptide file, proteiNorm expects the following columns: “id”, “Protein group IDs”, “Leading razor protein”, “Gene names”, “Reverse”, “Potential contaminant” (see Table 1 for an example). And for protein file, proteiNorm expects the following column: “id”, (optional: “Reverse”, “Potential contaminant”, “Only identified by site”) (see Table 2 for an example)



Inputs

Peptides

No file selected

Filtering method Peptide to Protein

Top3

Proteins

No file selected

Figure 1 Data-tab. Peptide and/or protein files (protein required) can be uploaded. Peptide file can be filtered using “Top3” method to create a filtered protein file.

Uploading a peptide-file will provide option to filter peptides using the “Top3” methods and export a filtered protein-file (using the “Filter peptide” button) (see Figure 1).

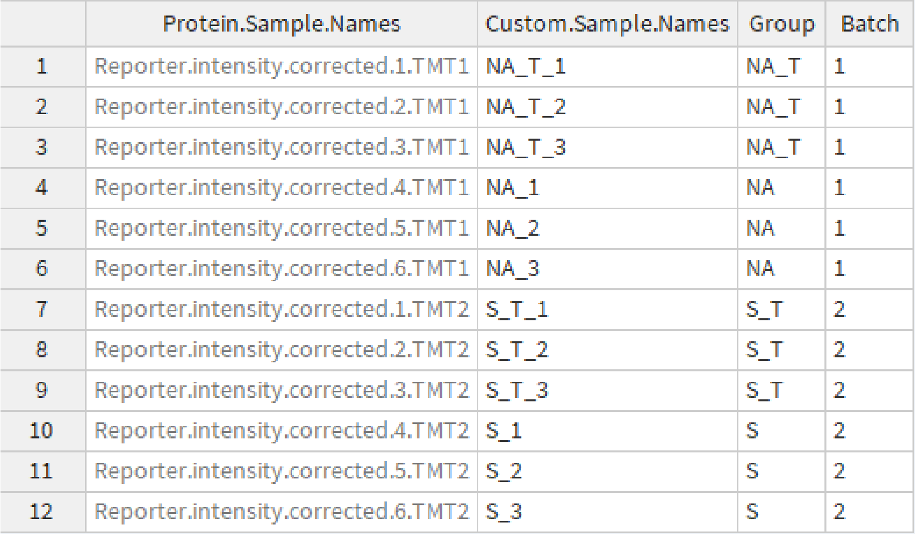
Table 1 Example of "peptide.txt" file

Leading razor protein	Gene names	Reporter intensity corrected 1 TMT1	Reporter intensity corrected 2 TMT1	Reporter intensity corrected 1 TMT2	Reporter intensity corrected 2 TMT2	Reverse	Potential contaminant	id	Protein group IDs
G3UY42	Pabpn1	10441	0	10354	16336			0	1302
A2ARS0	Ankrd63	221690	93505	0	0			1	665
P63085	Mapk1;Erk2	0	0	139070	69007			2	1872
D3Z3G6	Mapk3	0	0	0	0			3	1020
Q3TLR3	Trim32	0	0	15950	0		+	4	2128
Q8BKC5	Ipo5	0	0	0	0			5	2689
Q9CX34	Sugt1	0	0	0	7502.7			6	3017
D3Z4J5	Get4	725980	487850	0	0			7	1025
E9PU87	Sik3	0	0	0	0			8	1084
B1AX98	Lrrc47	5738.5	0	0	0	+		9	776

Table 2 Example of "proteinGroup.txt" file

id	Reporter intensity corrected 1 TMT1	Reporter intensity corrected 2 TMT1	Reporter intensity corrected 1 TMT2	Reporter intensity corrected 2 TMT2	Reporter intensity corrected 3 TMT2
G3UY42	141855.5	274230	70327	74433	69362
A2ARS0	1534300	688728.3	1018670	848550	934076.7
P63085	7850967	6062200	14272167	15162567	13325467
D3Z3G6	1325385	904430	3375350	4139050	3376250
Q3TLR3	17013	6679.7	235733.3	399395	391580
Q8BKC5	770910	729753.3	631380	488136.7	647203.3
Q9CX34	NA	NA	325425	187800.9	210552.7
D3Z4J5	725980	487850	NA	NA	NA
E9PU87	346169	217685.7	212900	212770	214720
B1AX98	583256.7	607100	289972	348293.3	301390

After uploading a protein-file, meta information for the recognized intensity columns can be specified. This include groups, batches and optional custom sample names (must be unique). See Figure 2 for an example.



	Protein.Sample.Names	Custom.Sample.Names	Group	Batch
1	Reporter.intensity.corrected.1.TMT1	NA_T_1	NA_T	1
2	Reporter.intensity.corrected.2.TMT1	NA_T_2	NA_T	1
3	Reporter.intensity.corrected.3.TMT1	NA_T_3	NA_T	1
4	Reporter.intensity.corrected.4.TMT1	NA_1	NA	1
5	Reporter.intensity.corrected.5.TMT1	NA_2	NA	1
6	Reporter.intensity.corrected.6.TMT1	NA_3	NA	1
7	Reporter.intensity.corrected.1.TMT2	S_T_1	S_T	2
8	Reporter.intensity.corrected.2.TMT2	S_T_2	S_T	2
9	Reporter.intensity.corrected.3.TMT2	S_T_3	S_T	2
10	Reporter.intensity.corrected.4.TMT2	S_1	S	2
11	Reporter.intensity.corrected.5.TMT2	S_2	S	2
12	Reporter.intensity.corrected.6.TMT2	S_3	S	2

Figure 2 Meta information example

Filter-tab

The filter tab is designed to identify outlier samples and exclude unwanted samples or samples with poor quality. Therefore, the user can evaluate the distribution of intensities of each sample from the peptide and protein data (Figure 3) and the PCA plots, which can be color-coded by group or batch (Figure 4). Samples can be excluded by unselecting them. In this example we want to focus the striatum (S) samples and exclude NA_T_1 - NA_T_3 and NA_1 - NA_3 by unselecting them. All figures will automatically update.

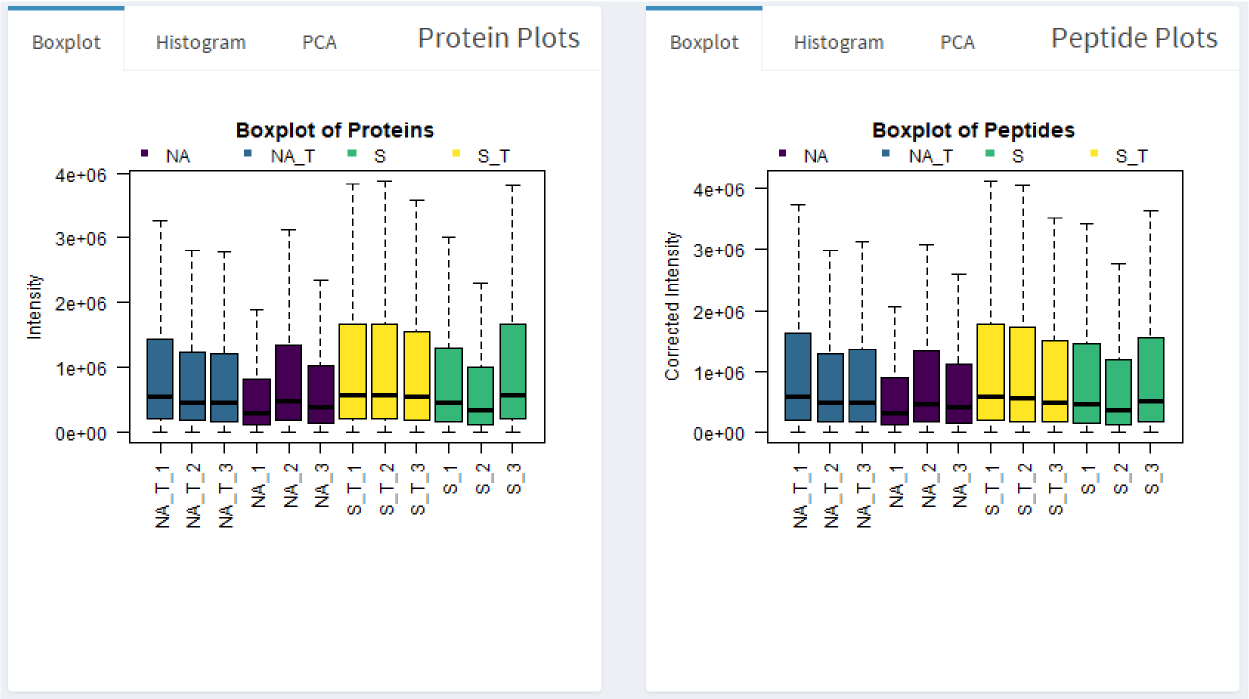
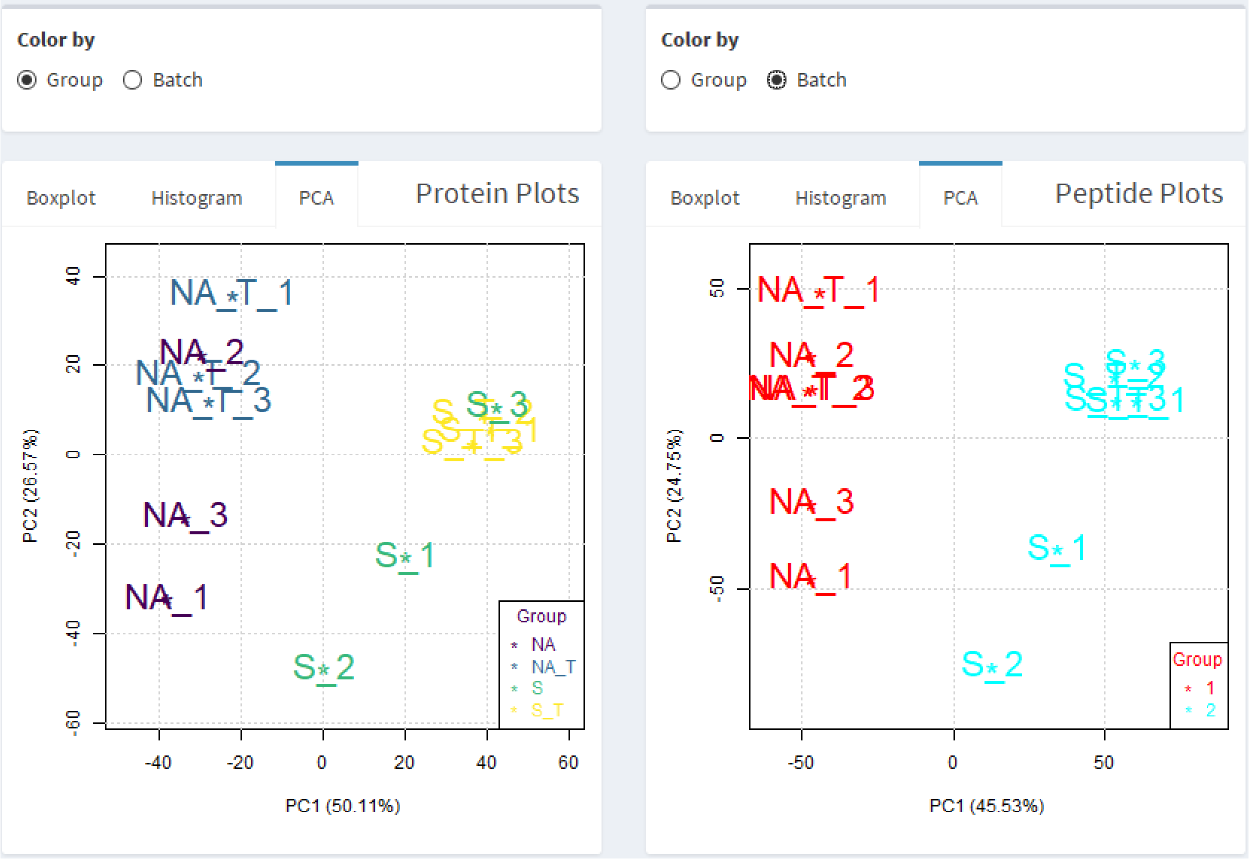


Figure 3 Intensity distribution by sample for peptide and protein data



Color by

● Group ○ Batch

Color by

○ Group ● Batch

Figure 4 PCA plots of peptides and proteins, color-coded by group or batch.

In addition, the user can the minimum number of samples with measurements in Y groups and the number of Y groups, meaning that only proteins are used that were measured in at least X samples in Y groups, where X is the number specified in “Minimum number of samples with measurements in Y groups” and Y is the number specified in “Number of Y groups”. For example, if “Minimum number of samples with measurements in Y groups” is set to 2 and “Number of Y groups” is set to 3, only proteins that were measured in at least 2 samples in 3 of the groups will be used downstream.

Normalization-tab

The purpose of the normalization tab is to identify a suitable normalization method by evaluating the following series of plots (all plots can be exported by the “Save Figures” button):

Total intensity

Figure 5 shows the total intensity plot that visualizes the sum of all intensities of a sample for different normalization methods. This plot can be useful to identify cases where a sample did not sequence well on the instrument and the total intensity for that sample will be lower than the rest (potentially indicating a bad sequencing run).

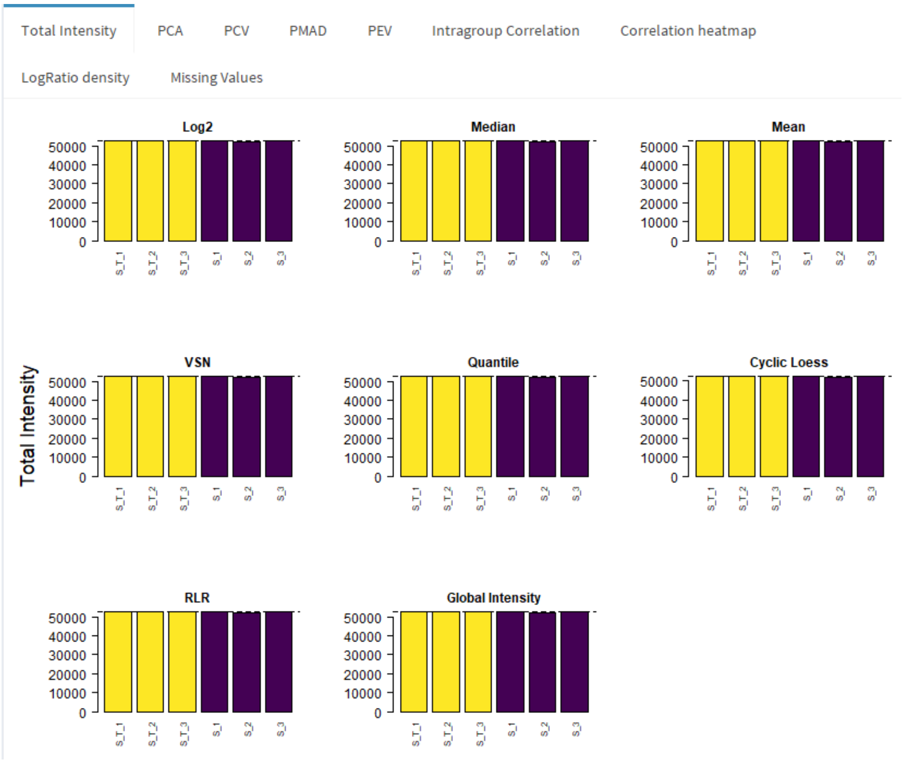
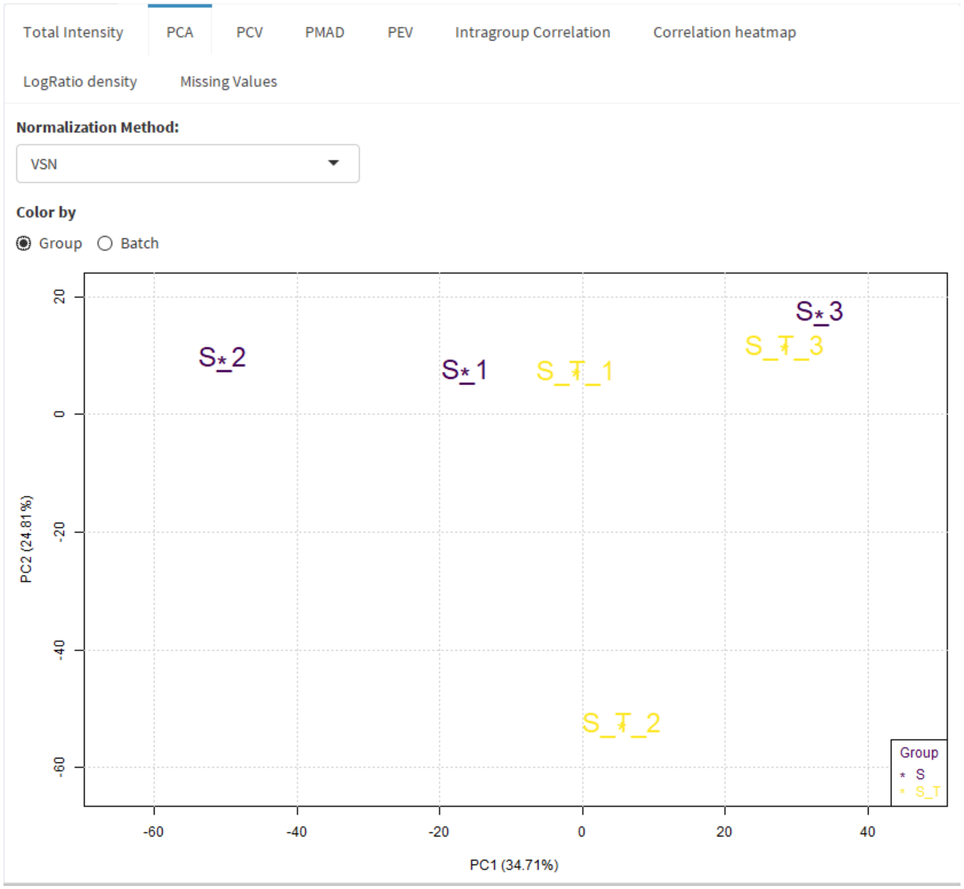


Figure 5 Total intensities of samples by normalization method

PCA

Figure 6 shows a PCA plot for a selected normalization methods. Samples can be color-coded by group or batch. Ideally, sample replicates of a group cluster together, while samples of different groups are separate from each other.



Normalization Method:

VSN

Color by

● Group ○ Batch

Figure 6 PCA plot

PCV, PMAD, PEV

Pooled coefficient of variation (PCV), pooled median absolute difference (PMAD), pooled estimate of variance (PEV) are measurements of intragroup variations (Figure 7). Small values are desirable here as they indicate limited variation of sample replicates. In this data set, the normalization method “VSN” seems to perform the best.

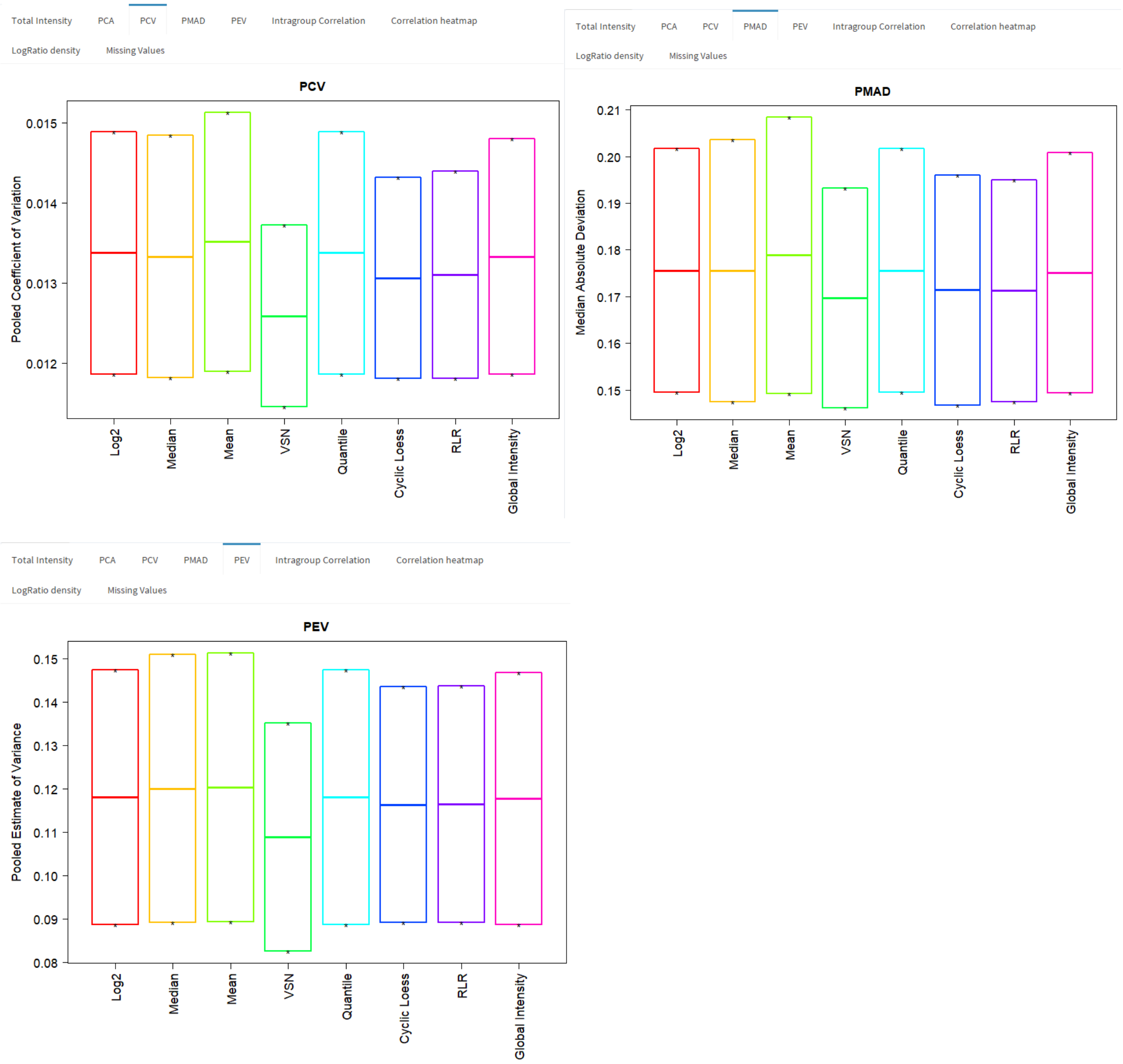


Figure 7 PCV, PMAD and PEV

Intergroup correlation

Figure 8 shows all pair-wise intra-group correlations for different normalization methods. High intragroup correlation values indicate high correlation of replicates within a treatment groups and are desirable. It can be observed, that “VSN” seems to produce the highest intragroup correlations.

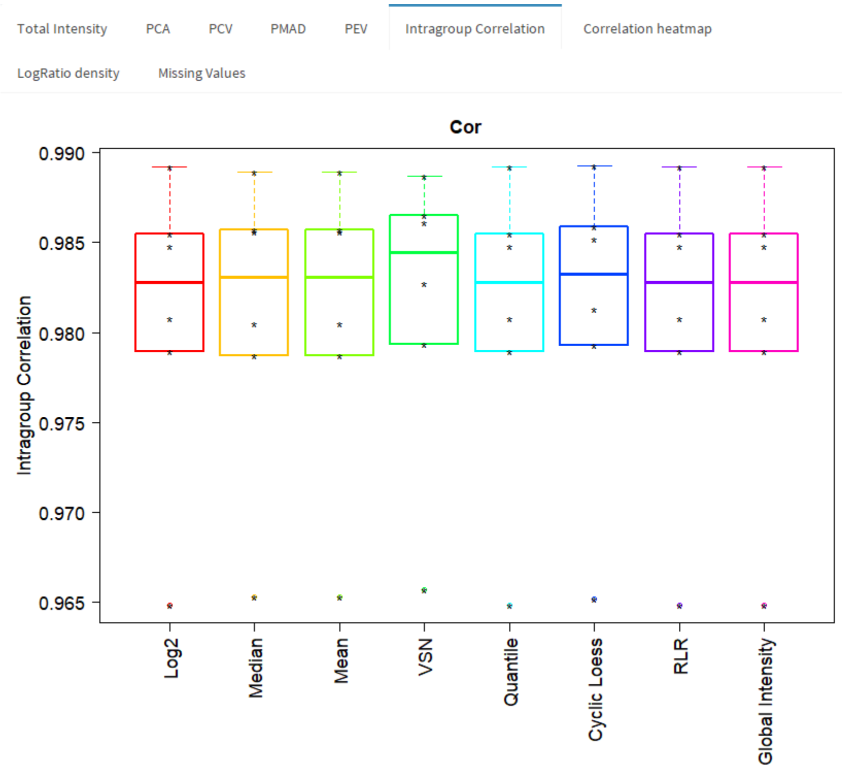
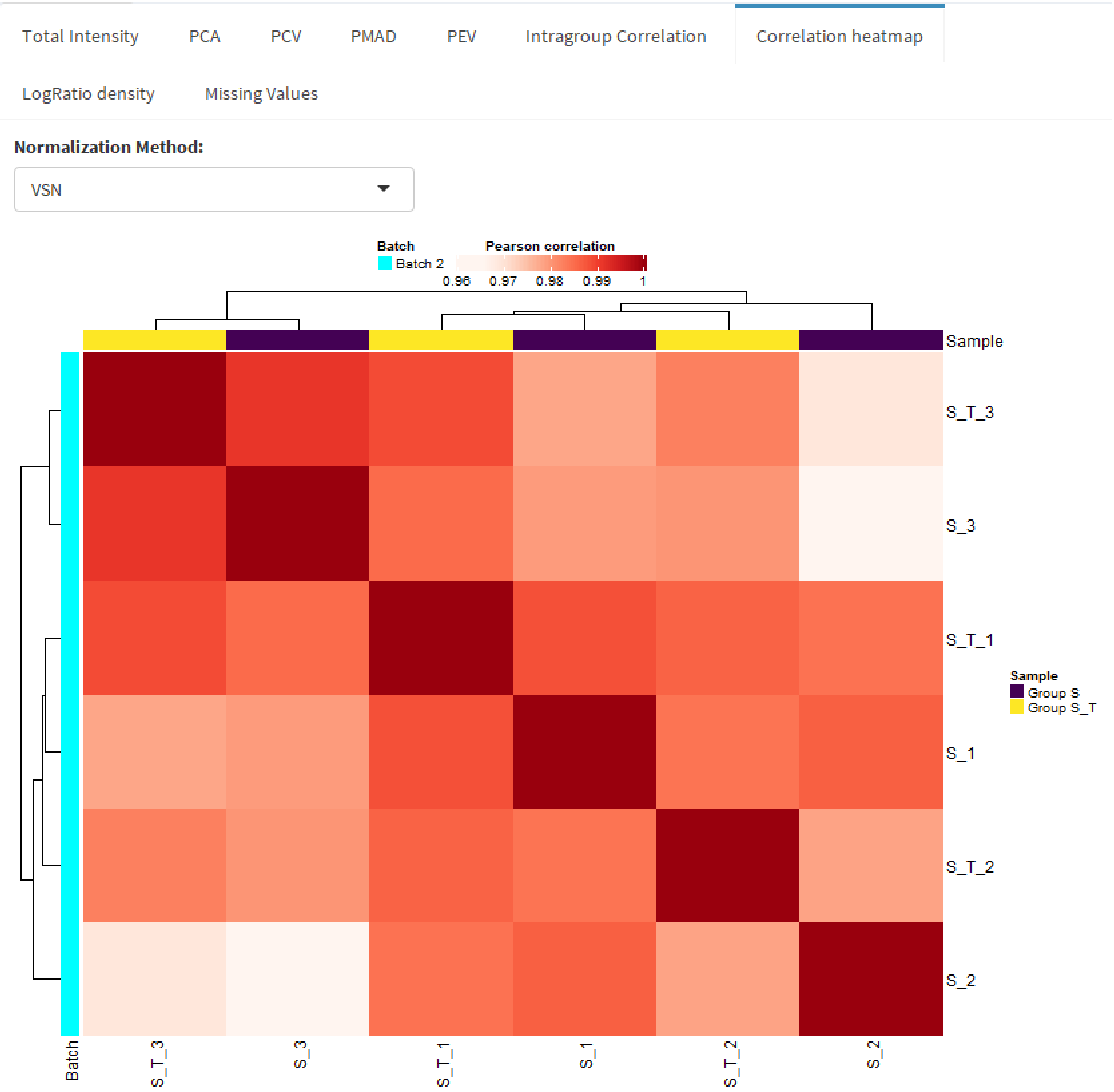


Figure 8 Intragroup correlations

Correlation heatmap

Figure 9 shows a clustered correlation heatmap for a selected normalization method. The x-axis color codes the sample group and the y-axis is color coded by batch.



Normalization Method: VSN

Figure 9 Correlation heatmap

Log2-ratio distribution

Figure 10 shows the distribution of log2-ratios for different normalization methods. In case of more than 2 groups, all pair-wise group comparisons are combined. The log2 ratio should be centered on zero and not introduce bias towards up or down regulation of proteins. In this data set it can be observed that the distribution of log2-ratios of most normalization methods is not centered around zero, with the exception of VSN and RLR.

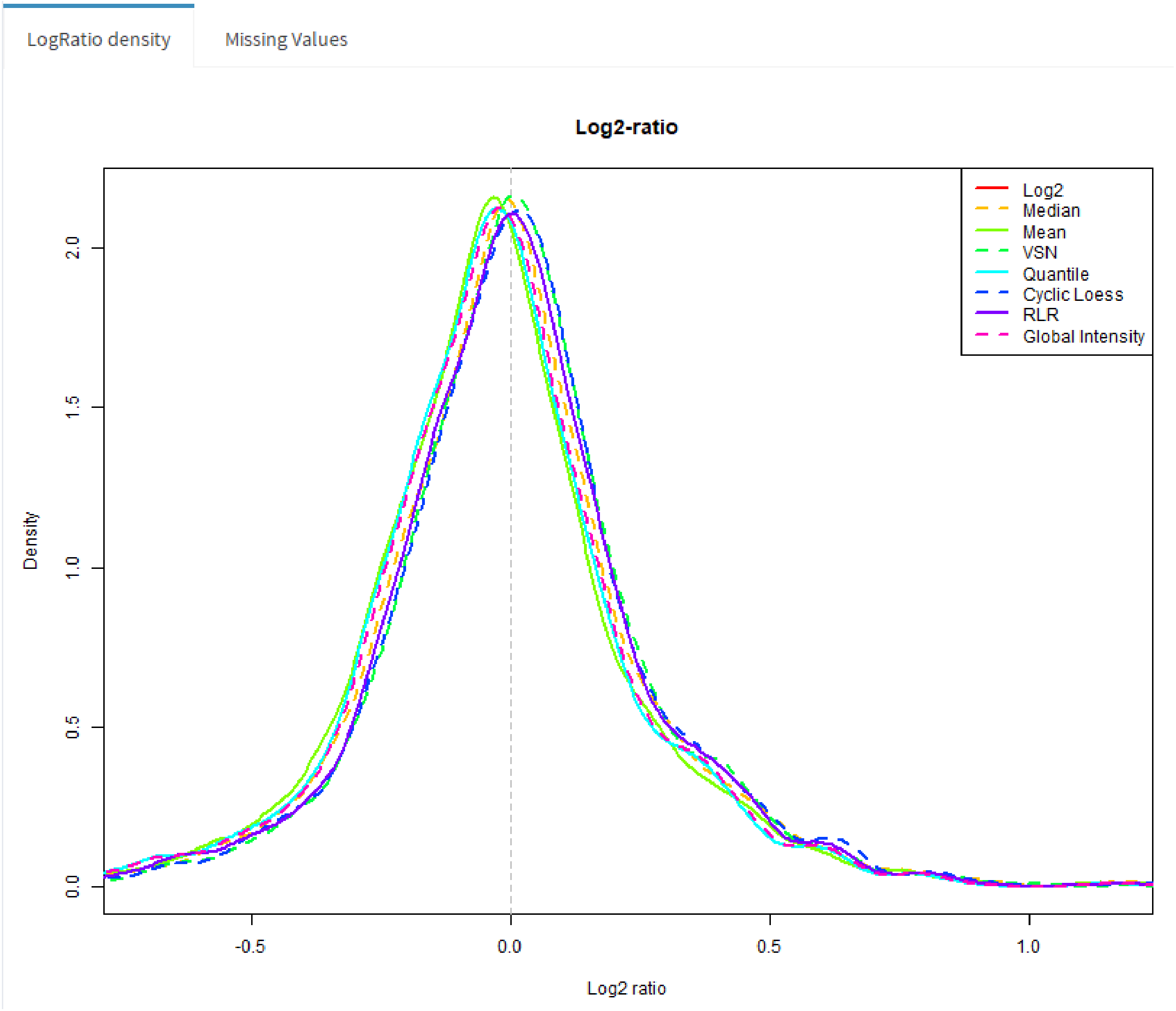
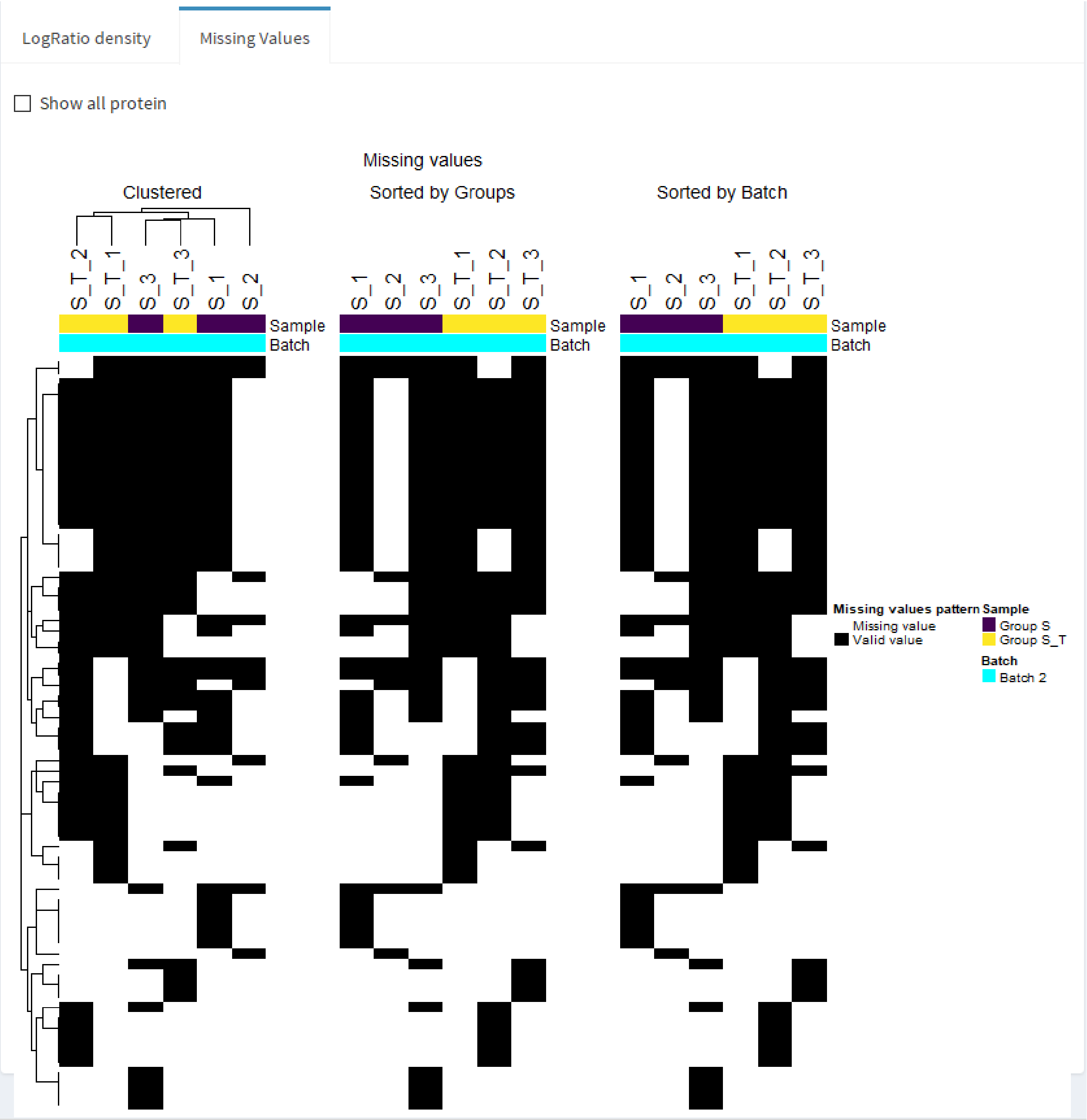


Figure 10 Log2-ratio distribution

Missing values

Figure 11 shows the three differently organized heatmaps of missing values (clustered, sorted by groups, and sorted by batch). Only proteins with at least one measured value and one missing value are shown (proteins with complete measurement or only missing values are not shown). The purpose of this heatmap is to assist researchers to identify patterns of missing values (i.e., MAR or MNAR) and guide the choice of imputation method, if desired.

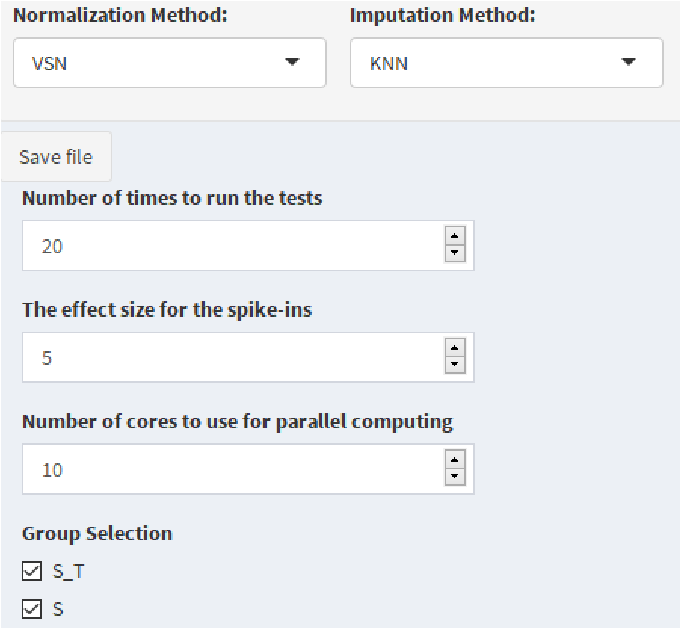


☐ Show all protein

Figure 11 Clustered heatmap of missing values

DAtest-tab

Once the user has identified a suitable normalization method, the protein data can be imputed (optional) and exported (“Save file” button) for further downstream analyses. If the user is uncertain about the choice of method for the differential abundance analysis, proteiNorm provides the option to compare different methods for a differential abundance analysis provided by the R package “DAtest”. As the DAtest package requires complete data (no missing values), the user can select an appropriate method for normalization and imputation for the differential abundance analysis. Using VSN and KNN to normalize and impute the protein data and the default setting for DAtest (Figure 12), different methods for a differential abundance analysis are compared for a two group comparison (S_T vs S). Comparisons of more than two groups is possible and appropriate statistical methods are automatically selected. Results of DAtest are summarized in a table and figure as shown in Figure 13. DAtest expects raw intensities, therefore intensities are provided as following: $DAinput = 2^{normalized\ intensities}$



Normalization Method: VSN

Imputation Method: KNN

Number of times to run the tests: 20

The effect size for the spike-ins: 5

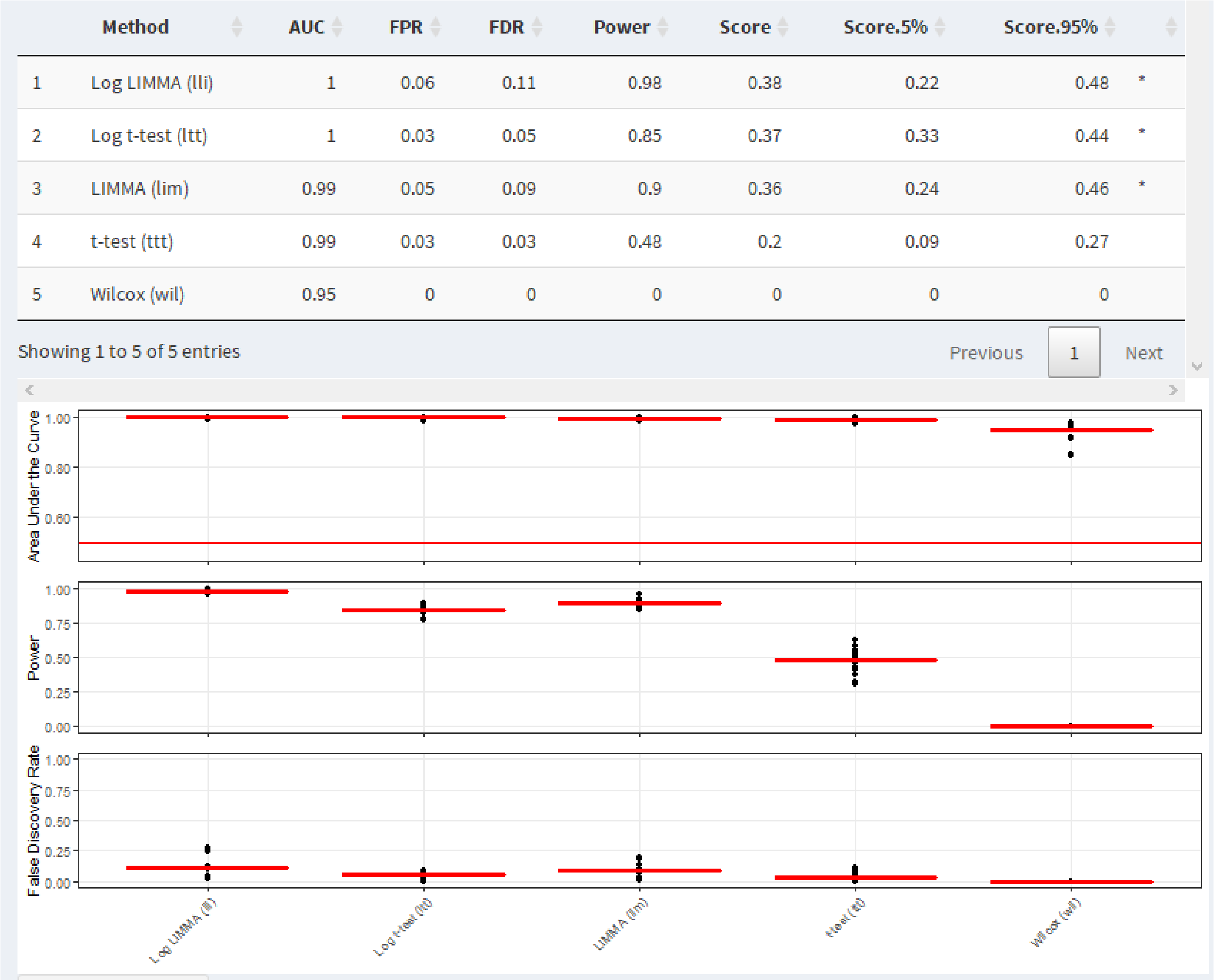
Number of cores to use for parallel computing: 10

Group Selection

☑ S_T

☑ S

Figure 12 DAtest settings



	Method	AUC	FPR	FDR	Power	Score	Score.5%	Score.95%	
1	Log LIMMA (lli)	1	0.06	0.11	0.98	0.38	0.22	0.48	*
2	Log t-test (ltt)	1	0.03	0.05	0.85	0.37	0.33	0.44	*
3	LIMMA (lim)	0.99	0.05	0.09	0.9	0.36	0.24	0.46	*
4	t-test (ttt)	0.99	0.03	0.03	0.48	0.2	0.09	0.27	
5	Wilcox (wil)	0.95	0	0	0	0	0	0	

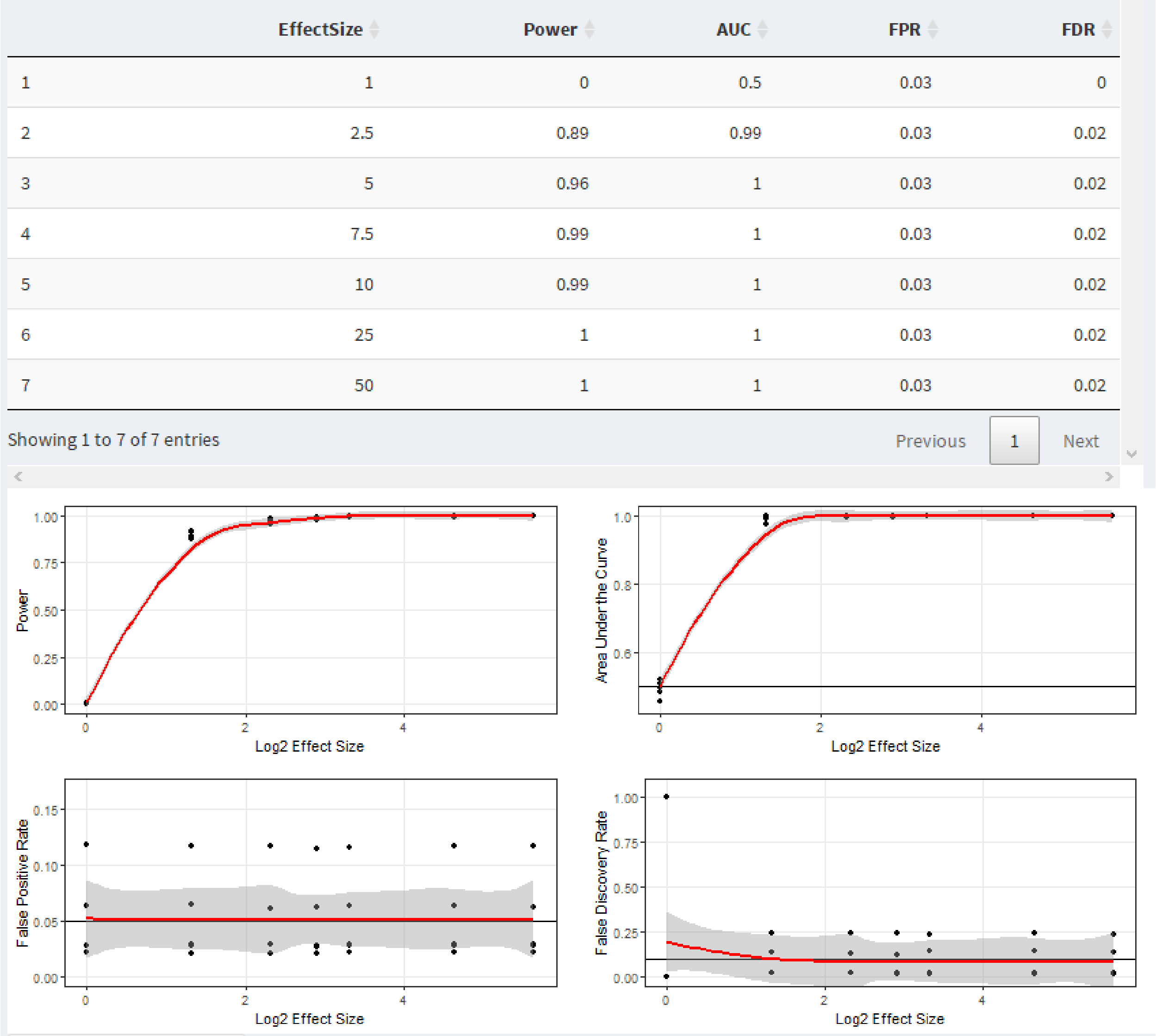
Showing 1 to 5 of 5 entries

Figure 13 DAtest results

As shown in Figure 13, Log LIMMA seems to perform the best. Because the normalized protein data is exported on the log-scale, LIMMA should be used to perform a differential analysis here.

DAtest power-tab

The last tab provides the option to evaluate the power of a selected differential abundance analysis method for a range of effect sizes. The results are summarized in a table and a figure as shown in Figure 14 for “Log LIMMA”.



	EffectSize	Power	AUC	FPR	FDR
1	1	0	0.5	0.03	0
2	2.5	0.89	0.99	0.03	0.02
3	5	0.96	1	0.03	0.02
4	7.5	0.99	1	0.03	0.02
5	10	0.99	1	0.03	0.02
6	25	1	1	0.03	0.02
7	50	1	1	0.03	0.02

Showing 1 to 7 of 7 entries

Figure 14 Power evaluation of a selected differential abundance analysis method. Here, “Log LIMMA” was selected.