Step-by-step Guide to Preprocess Expression Data using FlexStat Pipeline - Upload experimental results

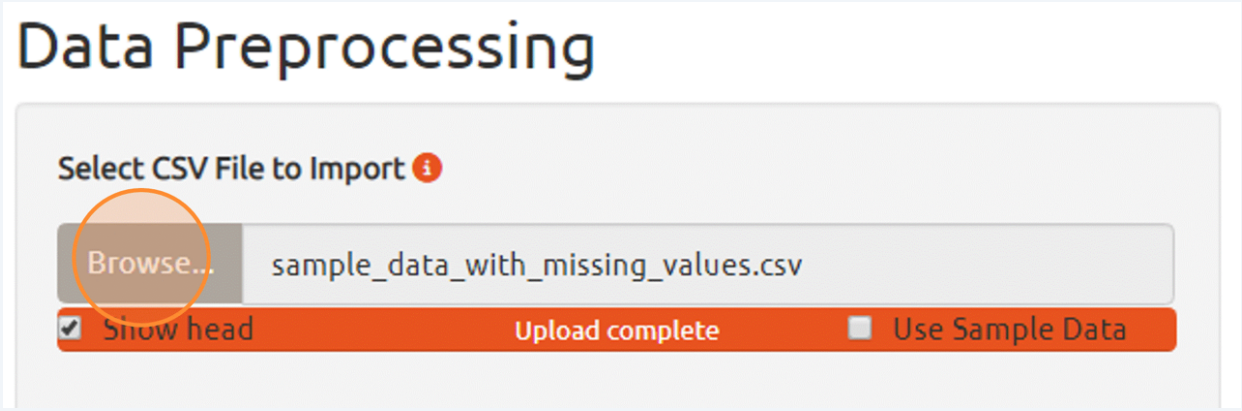
This feature facilitates preprocessing expression data with experimentally generated data. It involves missing value imputation and data normalization where users can specify the algorithm, and method to be used.

This tutorial is based on uploading an experimentally generated protein expression profile into the application.

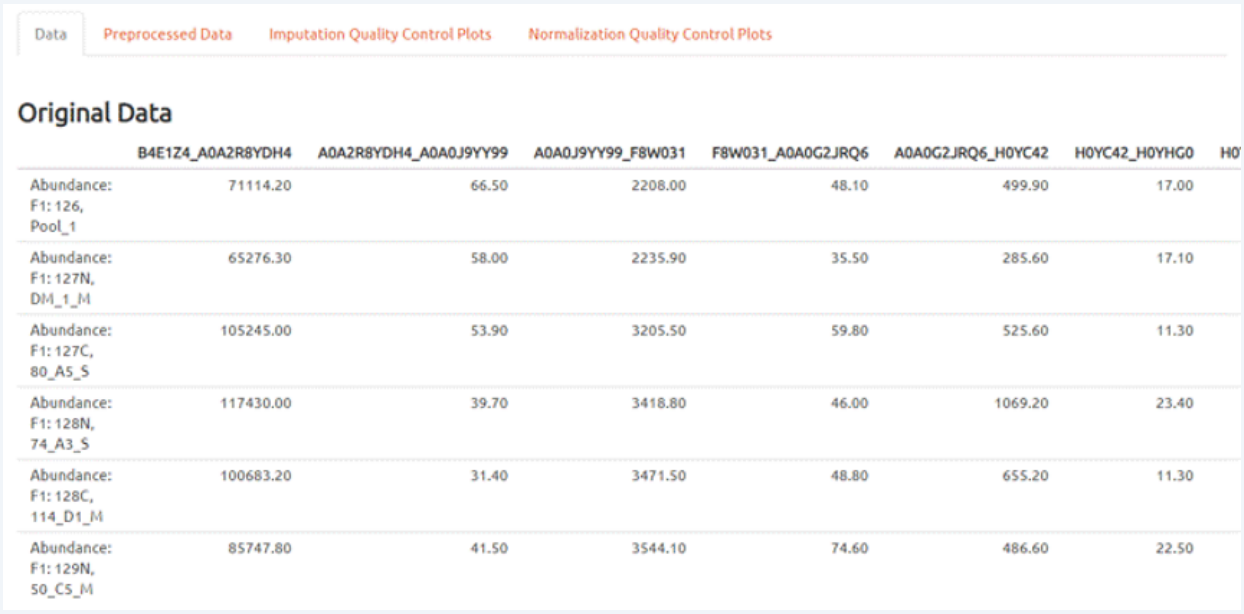
1 Navigate to https://jglab.shinyapps.io/flexstatv1-pipeline-only/

2 Go to "Data Preparation" tab.

3 Upload the expression file having rows as the samples and protein names in the columns and having a column for experimental condition/class as shown in the right-side panel.



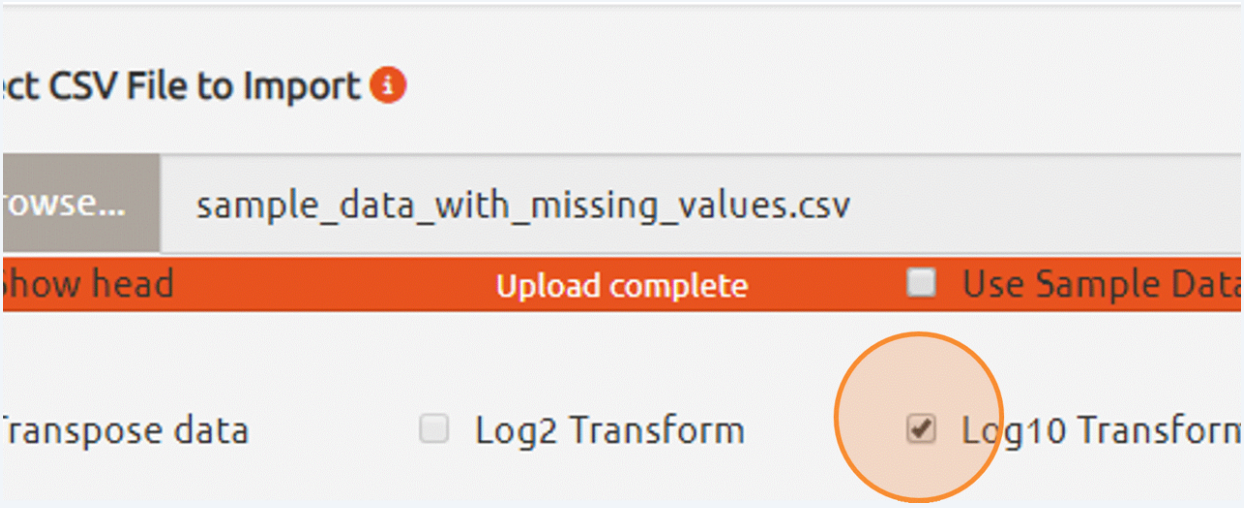
4 The uploaded data should be shown in the right-side panel.



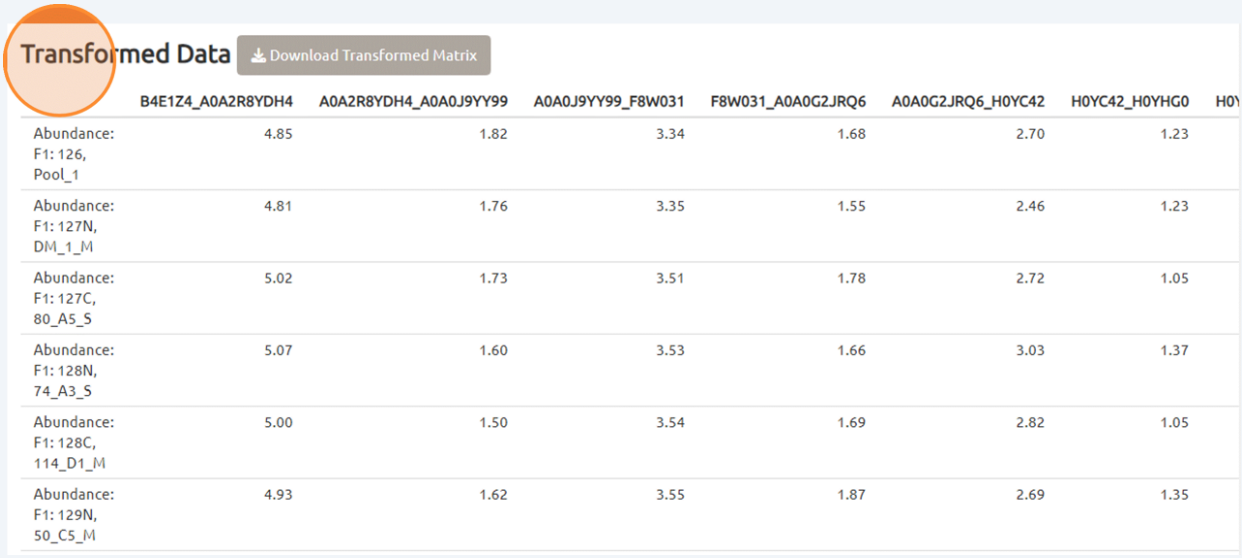
	B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG0	H0
Abundance: F1: 126, Pool_1	71114.20	66.50	2208.00	48.10	499.90	17.00	
Abundance: F1: 127N, DM_1_M	65276.30	58.00	2235.90	35.50	285.60	17.10	
Abundance: F1: 127C, 80_A5_S	105245.00	53.90	3205.50	59.80	525.60	11.30	
Abundance: F1: 128N, 74_A3_S	117430.00	39.70	3418.80	46.00	1069.20	23.40	
Abundance: F1: 128C, 114_D1_M	100683.20	31.40	3471.50	48.80	655.20	11.30	
Abundance: F1: 129N, 50_C5_M	85747.80	41.50	3544.10	74.60	486.60	22.50	

5 [Optional] Transform data into log scale if needed.

NOTE: This example data is raw abundance values, therefore here we are applying log10 scale to transform.

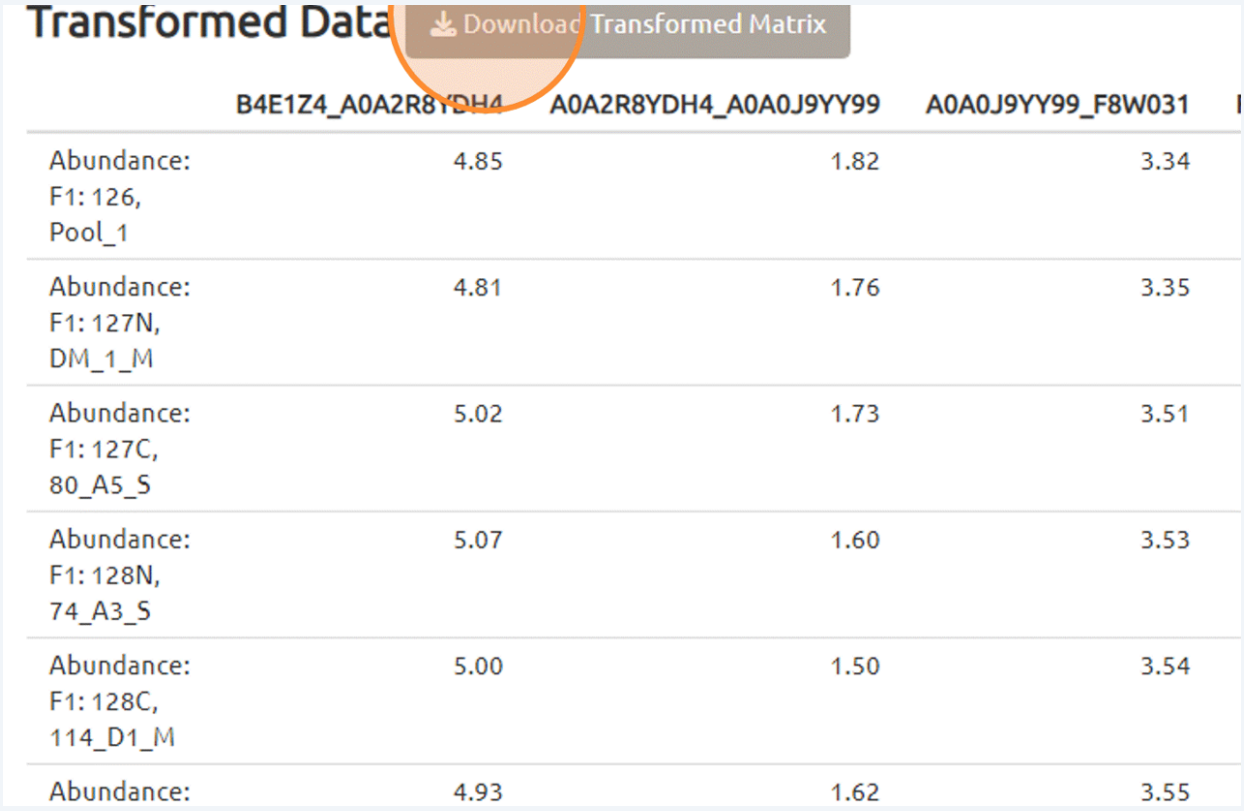


6 Transformed data is shown in the bottom right panel.



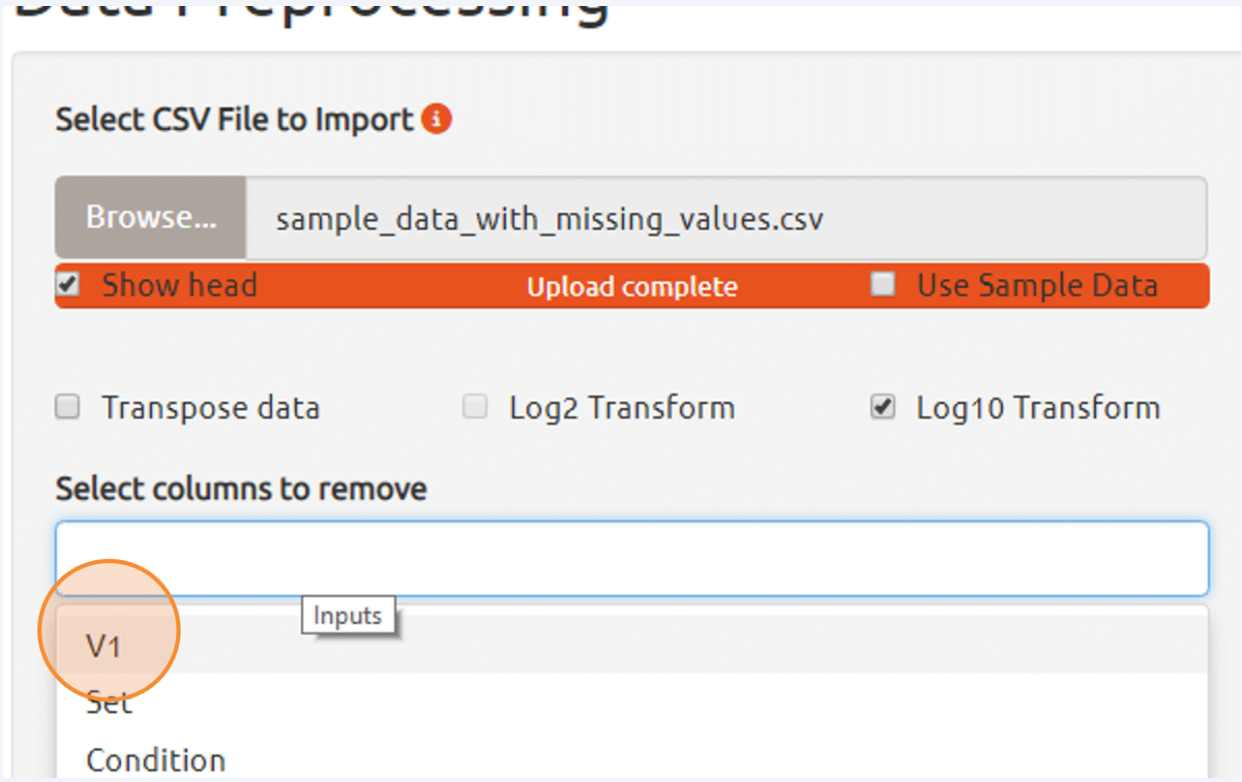
	B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG0	H0\
Abundance: F1: 126, Pool_1	4.85	1.82	3.34	1.68	2.70	1.23	
Abundance: F1: 127N, DM_1_M	4.81	1.76	3.35	1.55	2.46	1.23	
Abundance: F1: 127C, 80_A5_S	5.02	1.73	3.51	1.78	2.72	1.05	
Abundance: F1: 128N, 74_A3_S	5.07	1.60	3.53	1.66	3.03	1.37	
Abundance: F1: 128C, 114_D1_M	5.00	1.50	3.54	1.69	2.82	1.05	
Abundance: F1: 129N, 50_C5_M	4.93	1.62	3.55	1.87	2.69	1.35	

7 Transformed data can be downloaded for future reference.



	B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	I
Abundance: F1: 126, Pool_1	4.85	1.82	3.34	
Abundance: F1: 127N, DM_1_M	4.81	1.76	3.35	
Abundance: F1: 127C, 80_A5_S	5.02	1.73	3.51	
Abundance: F1: 128N, 74_A3_S	5.07	1.60	3.53	
Abundance: F1: 128C, 114_D1_M	5.00	1.50	3.54	
Abundance:	4.93	1.62	3.55	

8 [Optional] Select columns to be removed.



9 Select 'Class' variable to generate quality control plots and "Experimental batch variable" to be used in normalization.

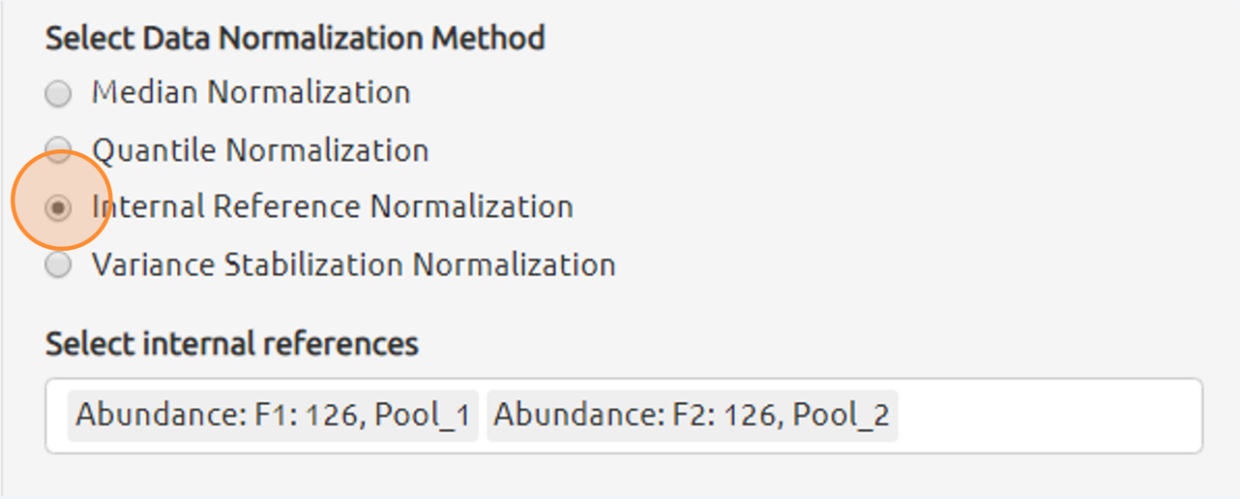
10 Select Data Normalization method

1. Median normalization

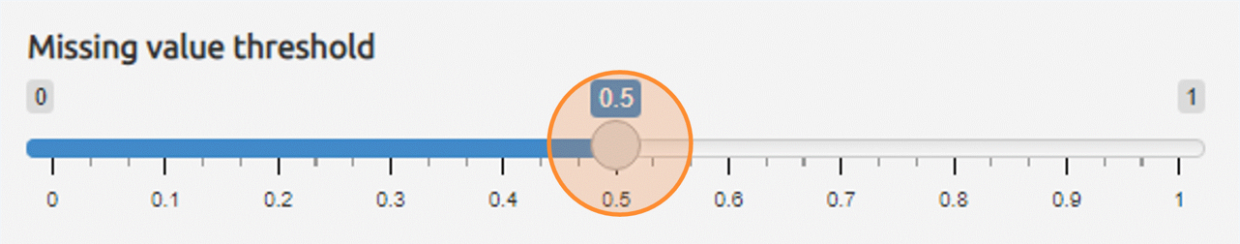
2. Quantile normalization

3. Internal Reference Normalization: if selected, select the corresponding internal reference to be used

4. Variance stabilization normalization

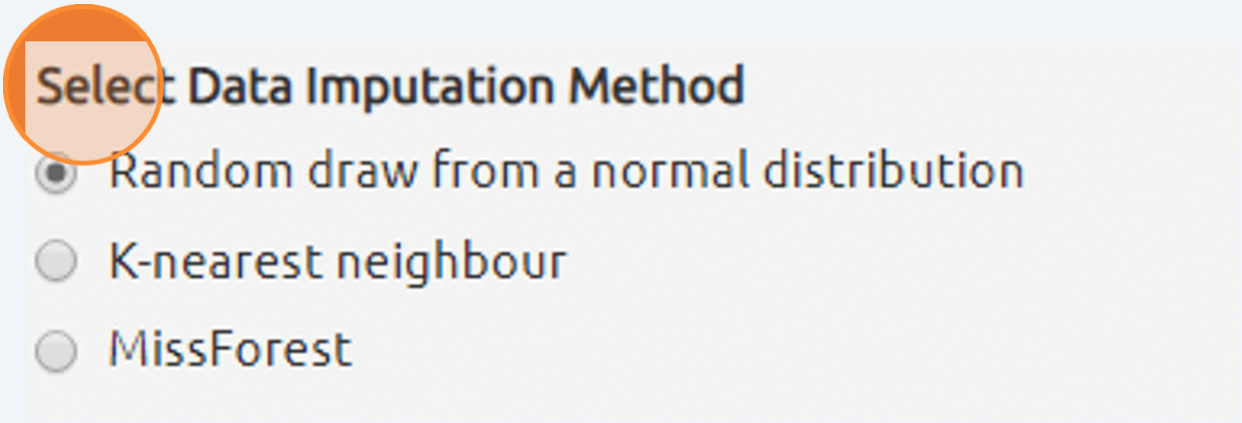


11 Select missing value threshold. Default is 0.5 which means proteins with 50% or less missing values will be considered for imputation.

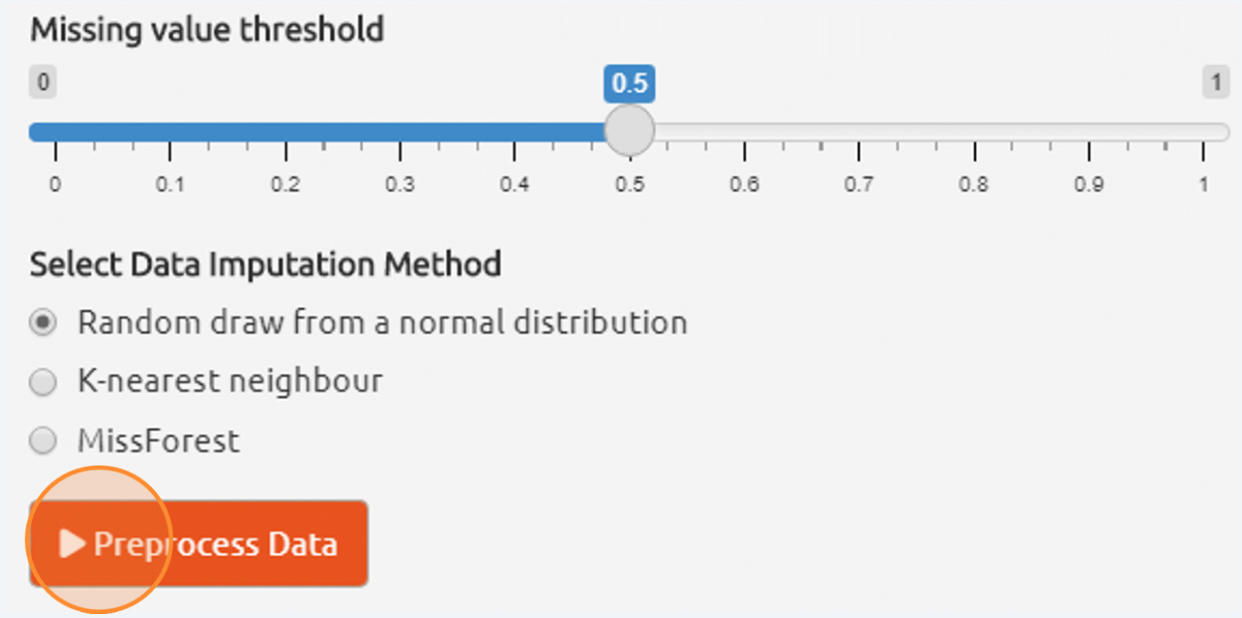


12 Select data imputation method. This include widely used imputation methods,

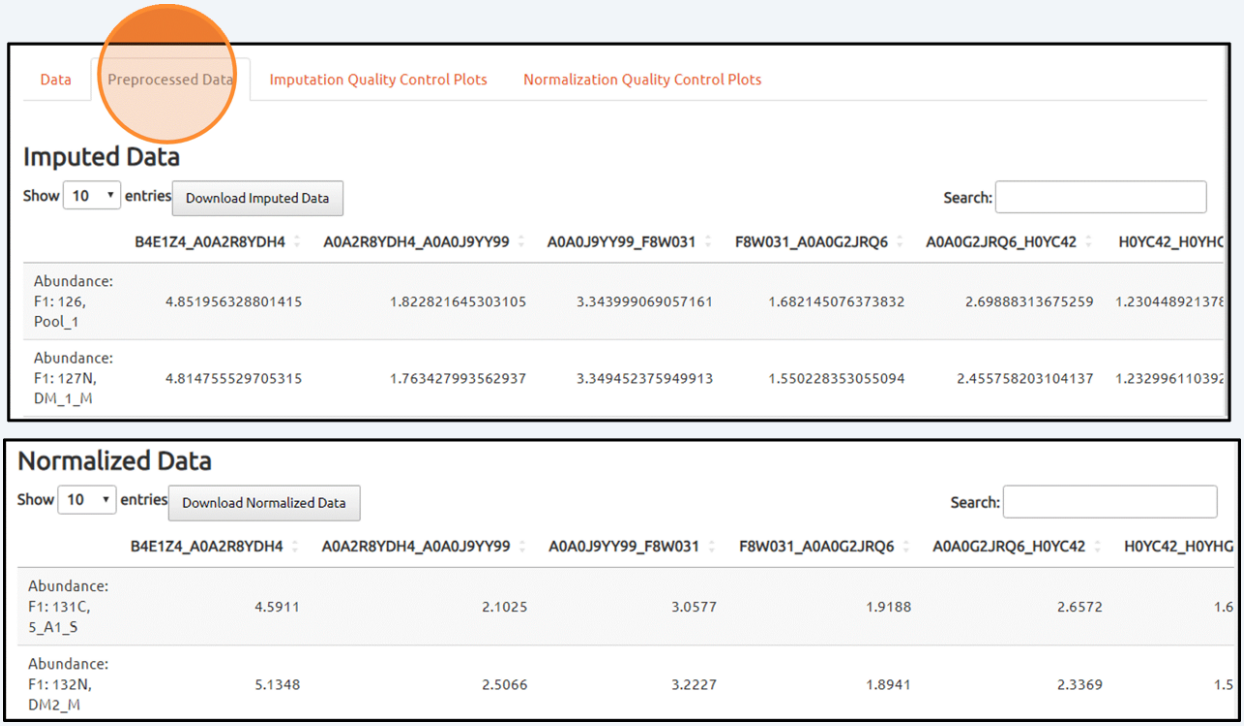
1. Random draw from a normal distribution
2. K-nearest neighbour imputation (**Hastie, Trevor, et al., 1999)**
3. MissForest imputation: (**Stekhoven and Peter 2012)**https://doi.org/10.1093/bioinformatics/btr597



13 Click "Preprocess Data" to start preprocessing.

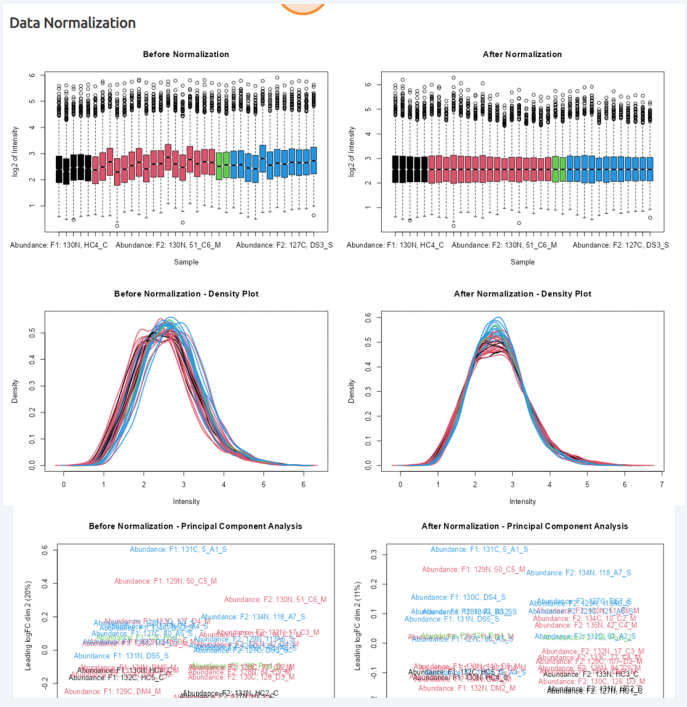


14 The imputed and normalized matrices can be checked and downloaded in "Preprocessed Data" tab.



15 "Normalization Quality Control Plots" tab displays, boxplots, density plots and principal component analysis plots before and after normalization.

The coloring is based on the "Class Variable".



16 "Imputation Quality Control Plots" tab displays, boxplots, density plots and principal component analysis plots before and after imputation.

The coloring is based on the "Class Variable".