

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.

Data Preprocessing

Select CSV File to Import

Browse...

sample_data_with_missing_values.csv

Show head

Upload complete

Use Sample Data

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

Data	Preprocessed Data	Imputation Quality Control Plots	Normalization Quality Control Plots				
Original Data							
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.

Select CSV File to Import

sample_data_with_missing_values.csv

Upload complete

Use Sample Data

Transpose data

Log2 Transform

Log10 Transform

- 6 Transformed data is shown in the bottom right panel.

Transformed Data		Download Transformed Matrix						
		B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG0	H0YHG0_H0YHGO
Abundance:		4.85	1.82	3.34	1.68	2.70	1.23	
F1: 126, Pool_1								
Abundance:		4.81	1.76	3.35	1.55	2.46	1.23	
F1: 127N, DM_1_M								
Abundance:		5.02	1.73	3.51	1.78	2.72	1.05	
F1: 127C, 80_A5_S								
Abundance:		5.07	1.60	3.53	1.66	3.03	1.37	
F1: 128N, 74_A3_S								
Abundance:		5.00	1.50	3.54	1.69	2.82	1.05	
F1: 128C, 114_D1_M								
Abundance:		4.93	1.62	3.55	1.87	2.69	1.35	
F1: 129N, 50_C5_M								

- 7 Transformed data can be downloaded for future reference.

Transformed Data		Download Transformed Matrix						
		B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG0	H0YHG0_H0YHGO
Abundance:		4.85			1.82			3.34
F1: 126, Pool_1								
Abundance:		4.81			1.76			3.35
F1: 127N, DM_1_M								
Abundance:		5.02			1.73			3.51
F1: 127C, 80_A5_S								
Abundance:		5.07			1.60			3.53
F1: 128N, 74_A3_S								
Abundance:		5.00			1.50			3.54
F1: 128C, 114_D1_M								
Abundance:		4.93			1.62			3.55
F1: 129N, 50_C5_M								

- 8 [Optional] Select columns to be removed.

The screenshot shows the 'Select CSV File to Import' step in the QIIME 2 pipeline. A file named 'sample_data_with_missing_values.csv' has been uploaded. The 'Show head' checkbox is checked, and the status bar indicates 'Upload complete'. Transformation options include 'Transpose data', 'Log2 Transform', and 'Log10 Transform', with 'Log10 Transform' selected. A section titled 'Select columns to remove' contains a list of columns: 'V1', 'Set', and 'Condition'. The 'V1' column is highlighted with an orange circle.

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

The screenshot shows the 'Class Variable' and 'Experimental Batch Variable' selection steps. The 'Condition' variable is selected for the Class Variable, and the 'Set' variable is selected for the Experimental Batch Variable. Both selections are highlighted with orange circles.

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

Select Data Normalization Method

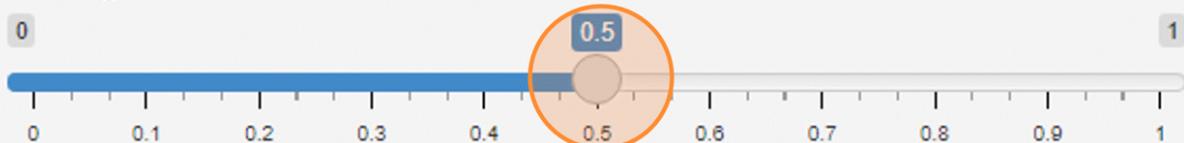
- Median Normalization
- Quantile Normalization
- Internal Reference Normalization
- Variance Stabilization Normalization

Select internal references

Abundance: F1: 126, Pool_1 Abundance: F2: 126, Pool_2

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

Missing value threshold



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

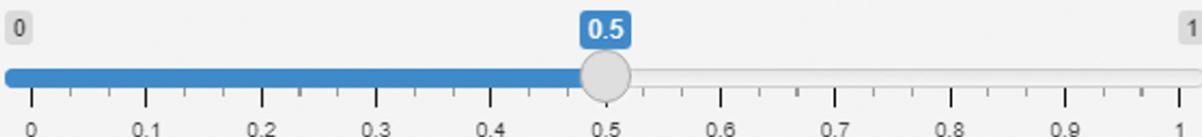
1. Random draw from a normal distribution
2. K-nearest neighbour imputation (**Hastie, Trevor, et al., 1999**)
3. MissForest imputation: (**Stekhoven and Peter 2012**)<https://doi.org/10.1093/bioinformatics/btr597>

Select Data Imputation Method

- Random draw from a normal distribution
- K-nearest neighbour
- MissForest

13 Click "Preprocess Data" to start preprocessing.

Missing value threshold



Select Data Imputation Method

- Random draw from a normal distribution
- K-nearest neighbour
- MissForest

▶ Preprocess Data

14

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

This screenshot shows the 'Preprocessed Data' tab. At the top, there are three tabs: 'Data' (selected), 'Preprocessed Data' (highlighted with an orange circle), and 'Imputation Quality Control Plots' and 'Normalization Quality Control Plots'. Below the tabs, the title 'Imputed Data' is displayed. There are buttons to 'Show 10 entries' and 'Download Imputed Data'. A search bar is also present. The data table lists abundance values for various peaks across different samples. Two rows are shown:

	B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG
Abundance: F1: 126, Pool_1	4.851956328801415	1.822821645303105	3.343999069057161	1.682145076373832	2.69888313675259	1.230448921378
Abundance: F1: 127N, DM1_M	4.814755529705315	1.763427993562937	3.349452375949913	1.550228353055094	2.455758203104137	1.232996110392

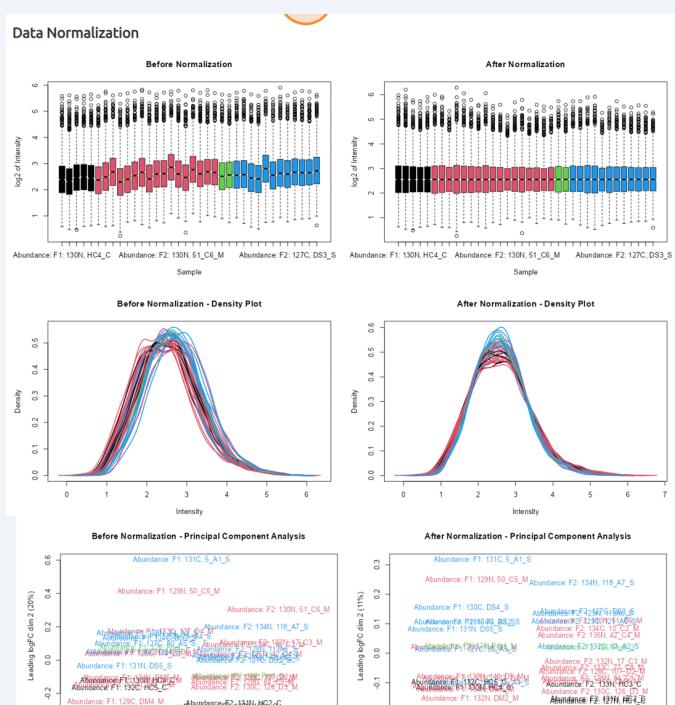
This screenshot shows the 'Normalized Data' tab. It has a similar layout to the 'Preprocessed Data' tab, with tabs for 'Data', 'Normalized Data' (selected), and 'Imputation Quality Control Plots' and 'Normalization Quality Control Plots'. The title 'Normalized Data' is displayed, along with buttons to 'Show 10 entries' and 'Download Normalized Data'. A search bar is also present. The data table lists normalized abundance values for the same peaks as the imputed data. Two rows are shown:

	B4E1Z4_A0A2R8YDH4	A0A2R8YDH4_A0A0J9YY99	A0A0J9YY99_F8W031	F8W031_A0A0G2JRQ6	A0A0G2JRQ6_H0YC42	H0YC42_H0YHG
Abundance: F1: 131C, S_A1_S	4.5911	2.1025	3.0577	1.9188	2.6572	1.6
Abundance: F1: 132N, DM2_M	5.1348	2.5066	3.2227	1.8941	2.3369	1.5

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".

