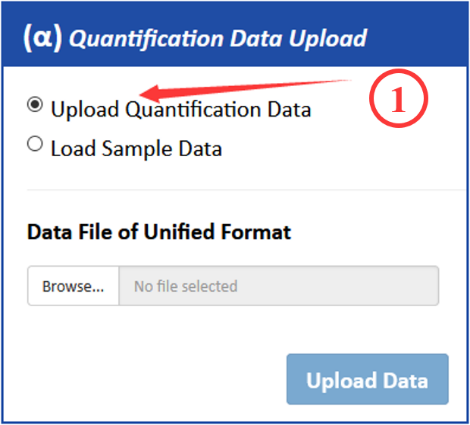
**One Example Illustrating The Whole Workflow Step By Step**

The example data PXD000672 was employed for illustrating the workflow step by step.

**Step 1. Quantification Data Upload**

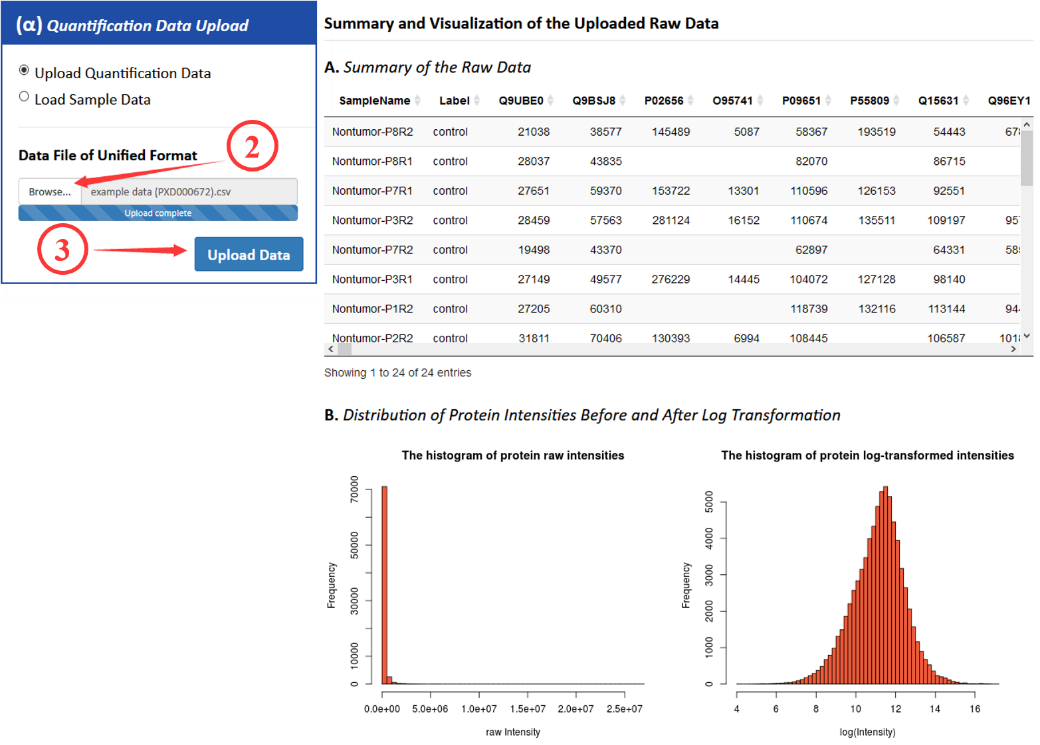
① When using MetaFS, users need to jump to the “Analysis” interface and select the “Upload Quantification Data” pattern.



② Then, by clicking the “Browse…” button, the user can upload the input file in the required format.

③ After a few seconds, the page will display “Upload complete” and the corresponding table and distribution on data will be presented. Users can click the “Upload Data” button and jump to step 2.

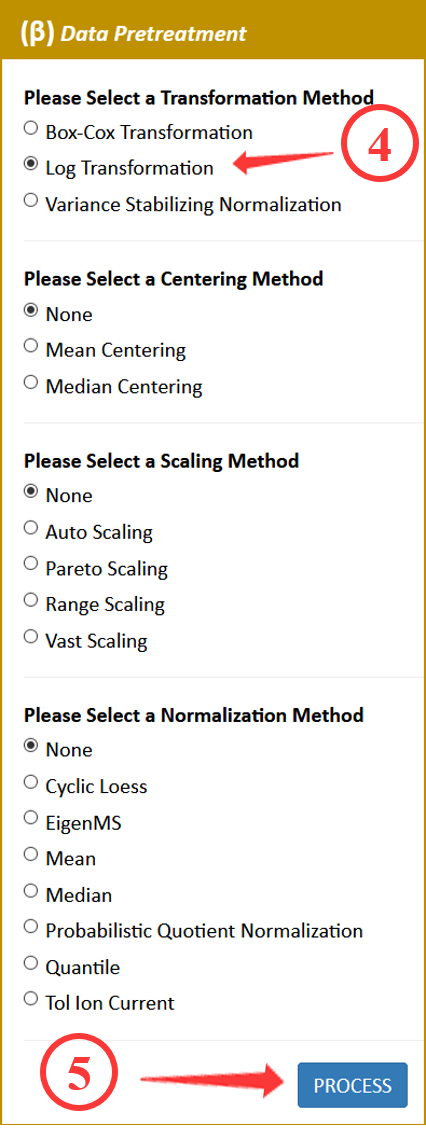
The exemplar required input file of PXD006224 was uploaded via the “Analysis” page of the website (https://idrblab.org/metafs) and the corresponding table and data distribution were present (shown below).



**Step 2. Data Pre-treatment**

④ In this step, users first need to select the transformation method to use, including box-cox transformation, log transformation and variance stabilizing normalization (VSN). If the first two transformation methods were chosen, further selection including centering method, scaling method and normalization method is needed. Otherwise, if you choose the VSN method, no more choices are required.

⑤ After selecting the pretreatment method, click the “PROCESS” button and the corresponding table, boxplot and qq plot after pre-treatment data will be presented.



⑥ Click the “NEXT” button to jump to step 3.

The pre-treatment of exemplar data was conducted by the log transformation, centering (none), scaling (none) as well as normalization (none). After clicking the “PROCESS” button, the corresponding table, boxplot and qq plot after pre-treatment were presented (shown below).

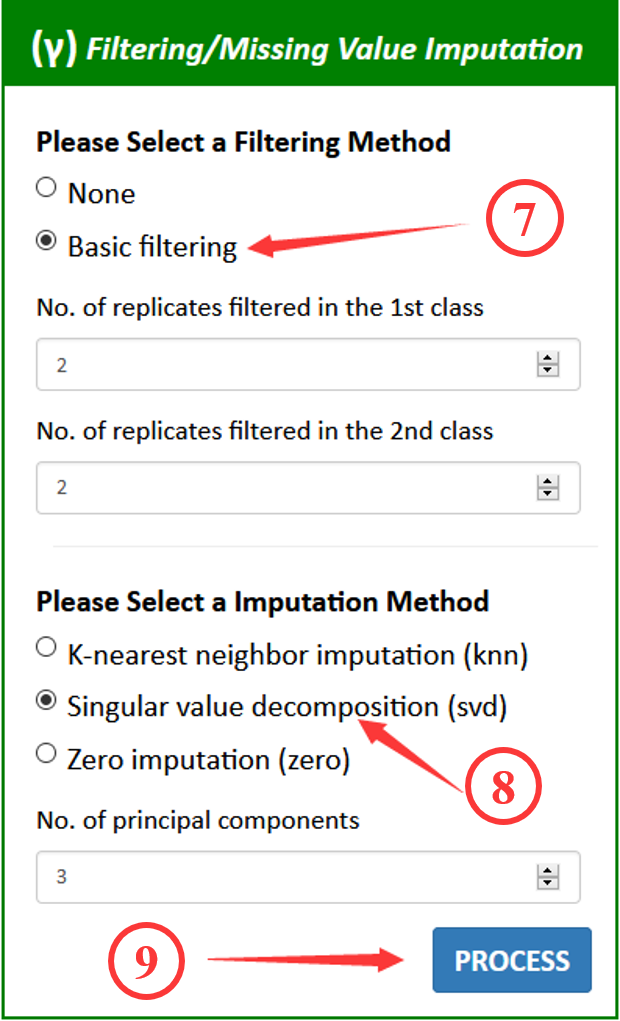


**Step 3. Filtering/Missing Value Imputation**

⑦ In this step, users should first decide whether to use the filtering method. If you choose the basic filtering method, you should also determine the corresponding parameters.

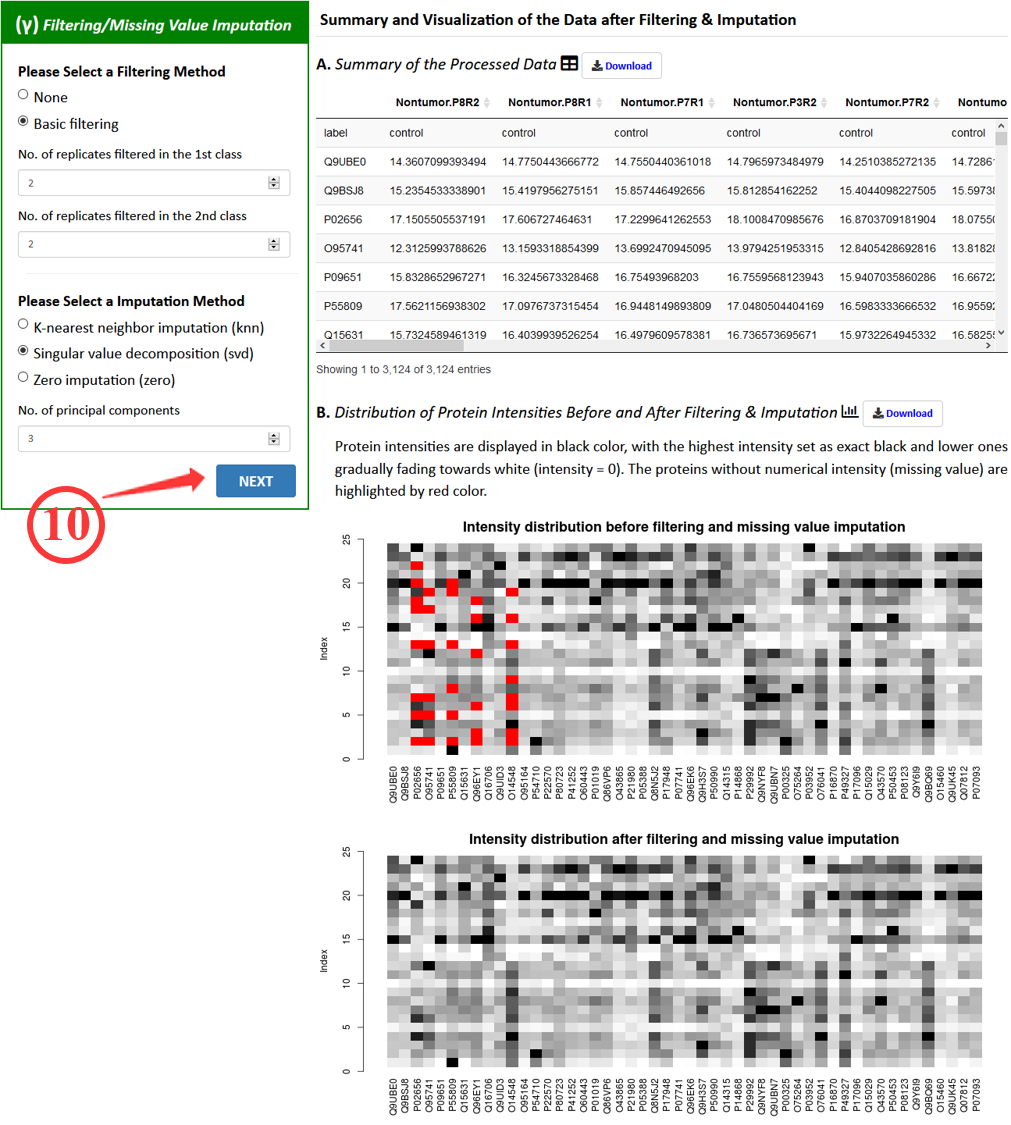
⑧ Then, users should select the imputation method and the corresponding parameters. The imputation methods include k-nearest neighbor imputation, singular value decomposition and zero imputation.

⑨ By clicking the “PROCESS” button, a summary of the processed data and a plot of the intensity distribution before and after data pre-treatment will be automatically generated.



⑩ Then click the “NEXT” button to jump to step 4.

The filtering/missing value imputation of the exemplar data was conducted using the basic filtering and singular value decomposition method, and the parameters were default. By clicking the “PROCESS” button, the corresponding table and matrix plot after imputed were presented immediately (shown below).

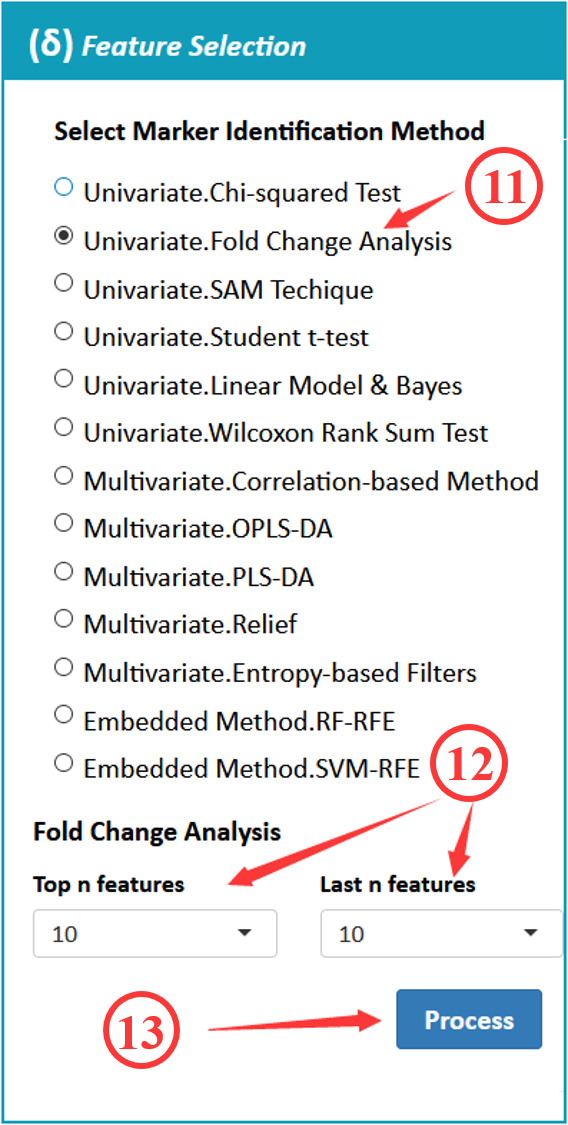


**Step 4. Feature Selection**

⑪ In sum, 13 feature selection methods were integrated and provided in this tool. Users should select a feature selection method in this step.

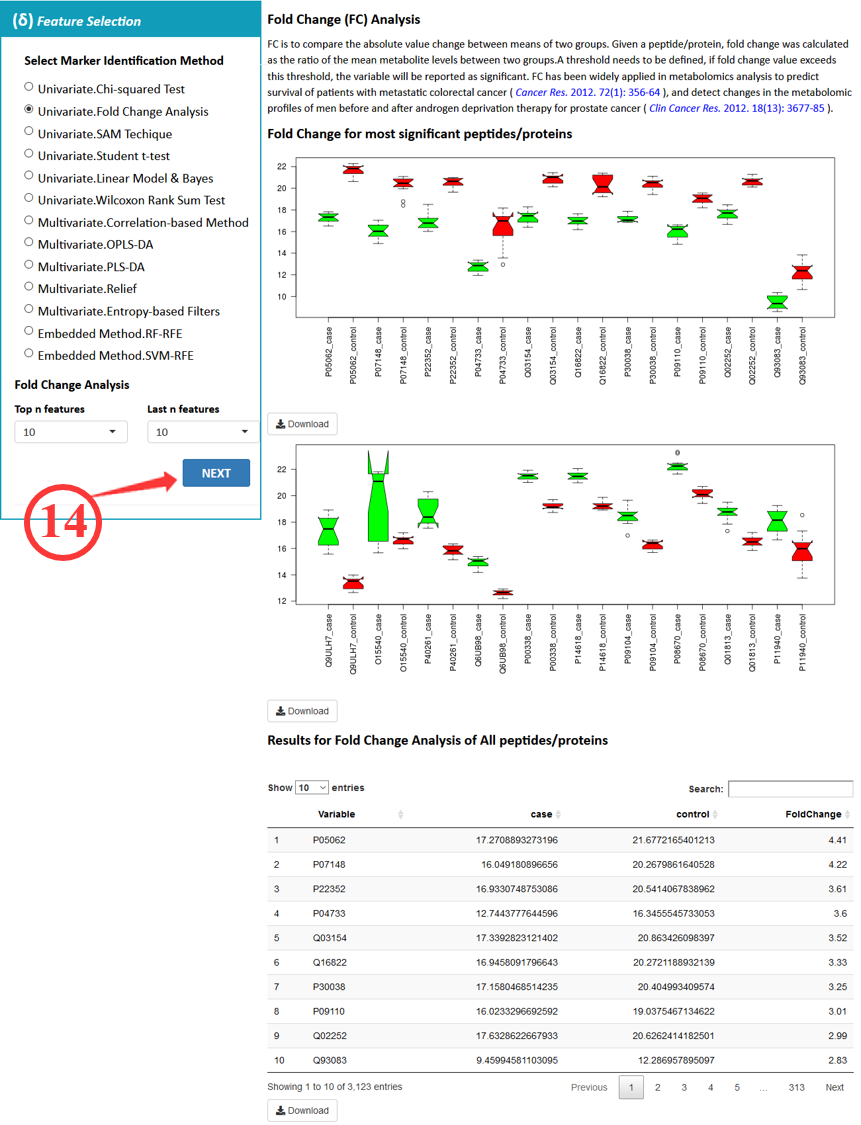
⑫ After selecting the feature selection method, users can set up corresponding parameters according to their needs.

⑬ By clicking the “PROCESS” button, the corresponding table and plot of differential proteins were presented.



⑭ Then click the “NEXT” button to jump to step 5.

We used the fold change method to identify the differential proteins in the example data. And the number of top features as well as the number of last features was set to 10. The corresponding table and plot of differential proteins were presented (shown below).

****

**Step 5. Performance Assessment**

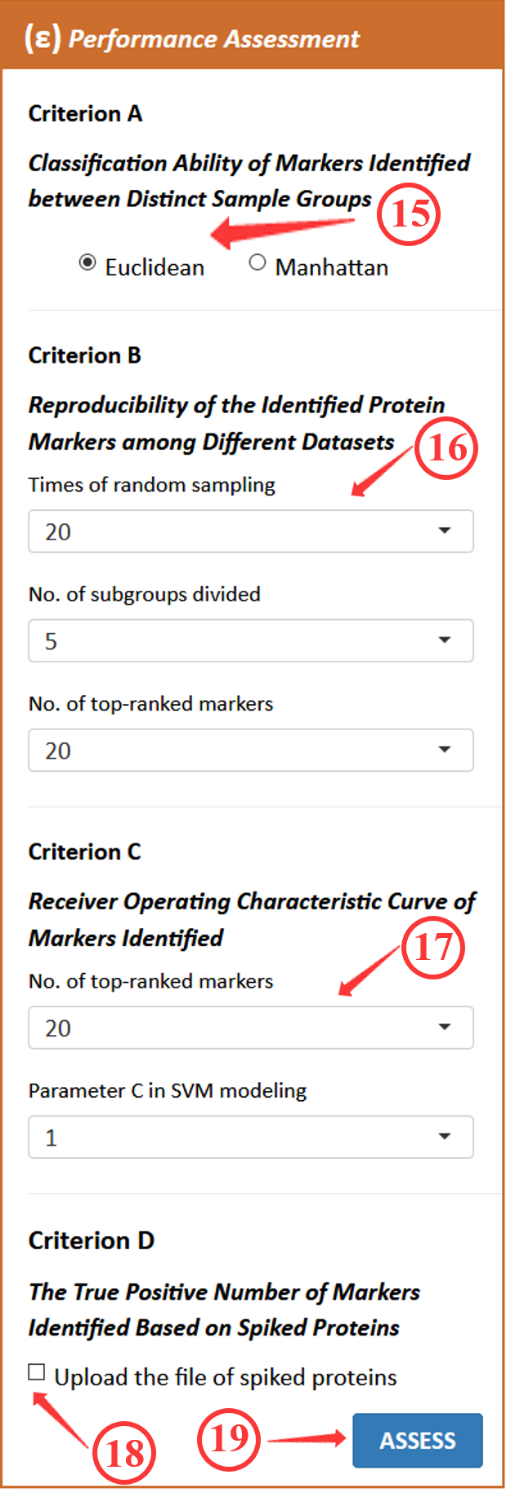
⑮ Four well-established criteria for comprehensively evaluating the performance of FS method are provided in MetaFS. In criterion A, users can choose Euclidean of Manhattan distance for analysis.

⑯ In criterion B, users need to set up three parameters, namely times of random sampling, number of subgroups divided and number of top-ranked markers.

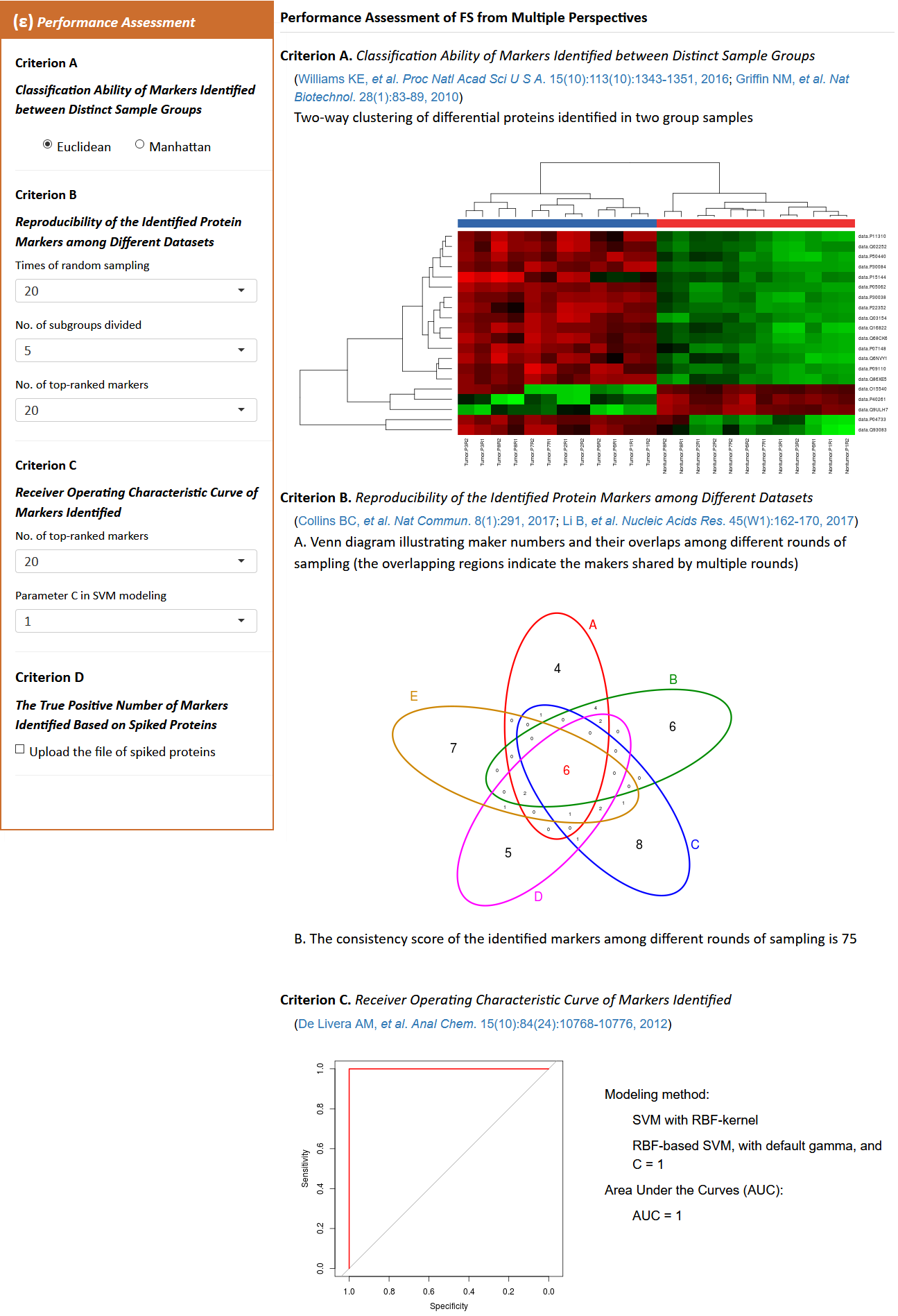
⑰ In criterion C, users need to determine the number of top-ranked markers and set up the parameter *C* in SVM modeling.

⑱ Before applying the criterion D, users need to up load the file of spiked proteins. Due to the lack of spiked proteins, the criterion D was not analyzed in this example data.

⑲ By clicking the “ASSESS” button, the corresponding plots of four well-established criteria were presented.



The example data in this step was assessed by criterion A, B and C, and the parameters were all default. The corresponding results of three well-established criteria were presented (shown below).

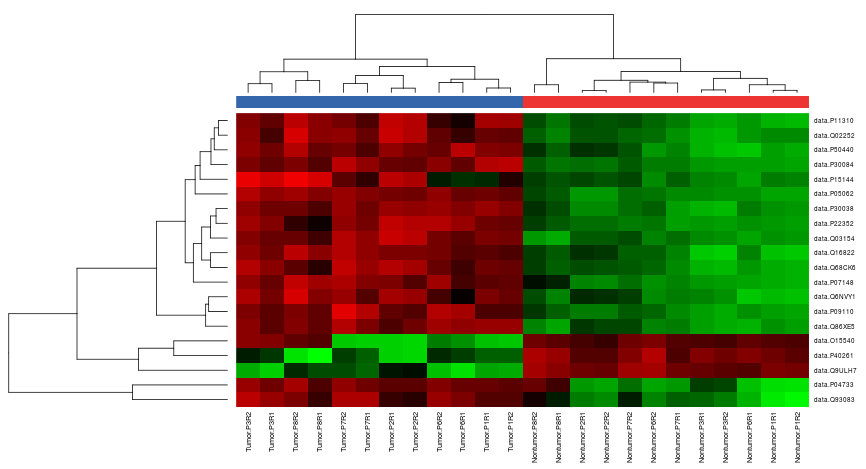


⑳ In order to select the appropriate feature selection method for this example data, the quantitative data (PXD000672) needed to be uploaded multiple times for generating results based on the different feature selection methods. To evaluate the performance of all feature selection methods, the same analysis procedure should be conducted repeatedly, except that feature selection method should be selected differently. Subsequently, the performance of the various FS methods was further analyzed. And the whole analysis process of metaproteomics data with MetaFS ends here.

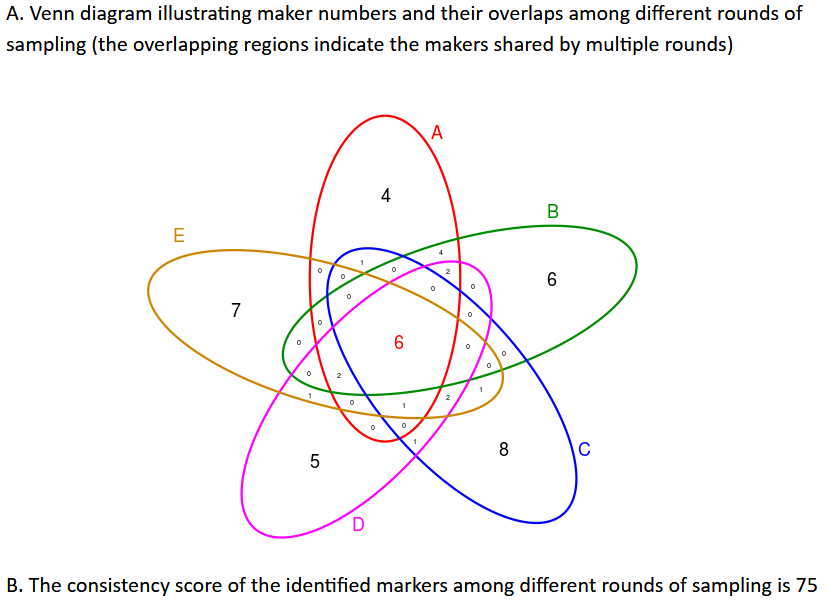
**The Evaluation Of The Output Based On Example Data**

**Figure S1 A**, **B** and **C** are the results of criteria A, B, and C based on the above example data (PXD000672) analysis process. Specifically, **Figure S1 A** illustrated the clustering analysis based on the differential proteins identified by feature selection method. These differential proteins successfully clustered samples according to the studied conditions so that distinct samples (control versus case groups) could be visually separated. The left half part represents the tumor samples, and the right part refers to non-tumor samples. The top proteins can clearly distinguish between the two groups, such as protein P11310, Q02252 and P50440. The last few proteins, such as Q93083, P04733 and Q15540, do not cluster samples efficiently, where the red and green cells intersect. **Figure S1 B** demonstrated the overlap among differential proteins sets from multiple samplings and the estimated consistency score. In this criterion, 6 identified makers were found to overlap in five rounds of sampling, and the consistency score equals to 75. **Figure S1 C** illustrated the curve of receiver operating characteristic (ROC) and the value of area under the curve (AUC). The specificity in x-axis represents the true negative rate and the sensitivity in x-axis refers to the true positive rate. The higher these two values, the better the feature selection method. The AUC value of the example data equals to 1, which indicates a perfect performance of fold change according to criterion C. Due to the lack of spiked proteins in this PXD000672 benchmark data, the performance of method’s capability of identifying the true positive markers was not analyzed in this example data.

**Figure S1 A**

****

**Figure S1 B**



**Figure S1 C**

