MS & Proteomics Resource

Yale School of Medicine
Keck Biotechnology Resource Laboratory

Application Note 5: Understanding Label Free Quantitation (LFQ) Analysis Results

Label Free Quantitation (LFQ) is a technique utilized for locating protein expression changes across numerous samples. For each sample, an estimated 0.25ug is loaded on the LC column for LC-MS/MS analysis on a Thermo Scientific LTQ Orbitrap Mass spectrometer which is equipped with a Waters nanoACQUITY UPLC system. Samples are typically run in triplicate after block randomizing with 2 blanks after each run. The LC-MS/MS data is processed with Progenesis LCMS software (Nonlinear Dynamics, LLC. (www.nonlinear.com)) and protein identification is performed using the Mascot search algorithm(Matrix Science www.matrix-science.com). Results are returned using the Yale Protein Expression Database or YPED software tool. In addition, Excel tables can be supplied upon request by the customer.

Data analysis:
The Progenesis LCMS software performs feature/peptide extraction, chromatographic/spectral alignment, data filtering, and statistical analysis. First, the raw data files are imported into the program. A sample run is then chosen as a reference (usually at or near the middle of all runs in a set), and all other runs are automatically aligned to that run in order to minimize retention time (RT) variability between runs. (note that due to the very high reproducibility of the nanoACQUITY, the RT shifts are very minimal if at all). No adjustments are necessary in the m/z dimension due to the high mass accuracy of the mass spectrometer (typically <3ppm). All runs are selected for detection with an automatic detection limit. Features within RT ranges of 0-10 minutes are often filtered out, as are features with charge ≥ +7 and +1. Progenesis calculates a normalization factor for each run to account for differences in sample load between injections, and differences in ionization. The normalization factor is determined by calculating a quantitative abundance ratio between the reference run and the run being normalized. The basic assumption is that most proteins and therefore peptides are not changing in the experiment so the quantitative value should equal 1. A standard peptide mixture is also spiked into the samples and this can be used to normalize across the set, as can any identified protein post data analysis.

The experimental design is setup to group multiple injections from each run. The algorithm then calculates tabulated raw and normalized abundances, max fold change, and Anova values for each feature in the data set. The MS/MS spectra (a combined list of all LC-MS/MS runs) are exported to an .mgf (Mascot generic file) for database searching. The Mascot search results are exported to an .xml file using a significance cutoff of $p<0.05$ and are then imported into the Progenesis LCMS software, where search hits are assigned to corresponding features or peptides. The features are tagged in sets based on characteristics such as MS/MS>1, $p<0.05$. Features or peptides that show protein expression changes but were not identified can often be identified by re-running the sample using an exclude list of the identified peptides/proteins.

Using the Mascot database search algorithm, the Keck Facility considers a protein identified when Mascot lists it as significant and more than 2 unique peptides match the same protein. The Mascot significance score (similar to the “Confident scores” column in the Excel Progenesis LCMS protein features spreadsheet – see below)

match is based on a MOWSE score and relies on multiple matches to more than one peptide from the same protein. Only protein identifications at a false discovery rate (FDR) of 1% or less are included.

Excel Sheet Results Dissemination:
The Progenesis LCMS analyses are exported into an Excel spreadsheet containing the protein list which contains all the protein identifications along with their corresponding scores and quantitation values (normalized and raw), Anova p-value, and number of peptides matched to each protein. The maximum fold change column shows the fold change between the normalized values in the different groups and will thus, always be a positive number. A second Excel table with peptides (also called features) is also sent. This includes the retention time of the peptide or feature, m/z and peptide charge for each peptide.

YPED Results Dissemination:
In addition, the results are loaded into our online viewing system called The Yale Protein Expression Database (YPED). YPED allows you to view the data, with additional protein links to Entrez, from your lab. YPED will allow you to choose the denominator you want, and then calculates the average normalized abundance and Log 2 values. This is done for both the protein and peptide lists. In sample sets comparing only 2 groups, a Volcano plot is also available – simply click on the link and choose the sample. The Red dots indicate proteins identified with only 1 peptide identified; blue dots, have 2 or more peptides identified. And, by placing the cursor on the dot, the name of the identified protein will appear.