Data analysis workflow for each bio replicate database search

[\\mchampion-nas.esc.nd.edu\DATA\Simon\02\_N-terminal\_Acetylation\Data\_Analysis\2022\_05\_23](file:///\\mchampion-nas.esc.nd.edu\DATA\Simon\02_N-terminal_Acetylation\Data_Analysis\2022_05_23)

Simon Weaver

1. Normalized each peptide area by dividing by the LFQ are of RpoA from that corresponding sample type/replicate.
2. Removed peptides that had 3 or less area values for the 6 technical replicates of WT + Comp (required at least 4 total values).
3. Calculated average, standard deviation, and %RSD for each peptide/each condition based on three replicates, ignoring NA values.
4. Stdev set to NA temporarily if there is 1 or 0 replicates of a certain condition.
5. Checked for “Acetylation (Protein N-term)” modification
6. Calculated the start position of each peptide by importing protein database, and returning the start index of the peptide when mapped to the correct protein accessioin.
7. Removed peptides that had a star position greater than 2.
8. Removed peptides that had an invalid accession ID (NA)
9. Calculated an average %RSD for each condition based on the remaining peptides where the %RSD was not NA.
10. For peptides that had a %RSD of NA, set their %RSD to the average %RSD for that condition.
11. Back-calculated a stdev for peptides with stdev of NA based on the new %RSD and the existing area value.
12. Separated peptides that contained N-term acetylation from those that did not.
13. For the acetylated peptide, calculated an area ratio for Del/WT and Del/Comp
14. Removed peptides where either of those ratios were greater than or equal to 1.
15. Combined acetylated and non acetylated peptide groups by matching the peptide variant sequence (for a match, the only difference could be the acetylation modification at the N-term, all other variants were considered different peptides).
16. Set all remaining NA area values to 0.
17. Calculated total area (avg acetylated + avg non acetylated) for each peptide, and propagated the stdev error.
18. Collapsed peptide rows originating from the same protein accession ID (different pep variants):
    1. Summed area for acetylatated + error prop for each condition
    2. Summed area for non acetylated + error prop for each condition
    3. Counted variants for acetylated and non acetylated + total variants
    4. Summed total area + error prop for each condition
19. Calculated %RSD for all values with standard deviation
20. Calculated % acetylation for each protein in each condition: acetylation area / total area + error propagation (this is a weighted average of all peptide variants based on normalized area)
21. Calculated %RSD for % acetylation for each condition.

Clustering (only performed for the acetylated November replicate):

1. Remove peptides that don’t have values in any 3 conditions for acetylation.
2. Create three new ratio columns, based on the average normalized areas
   1. Del / Comp
   2. Del / WT
   3. Comp / WT
3. Scale these columns without centering using the default scale function in R (Reduces the effect of outliers on the clustering)
4. Use the kmeans function in R to cluster based on the three ratio columns above, nstart = 25, number of clusters = 3
5. Use fviz\_cluster to visualize the clusters with a PCA
6. Separate each cluster into its own dataframe, and plot each based on original normalized areas for WT, Del, and Comp. (For plotting on a log scale, the 0 values were replaced with 0.00001 to eliminate NaN values. This was done only for visualization, and does not affect the clustering).