NNPDF Final Report

Stephanie M. Cologna, Ph.D.

Mentor: Forbes D. Porter

CONFIDENTIAL

Differential Protein Expression Profiling in Cerebrospinal Fluid from Niemann-Pick Disease, type C1 Patients

Summary of the Research: The enclosed research project aims to identify proteins that are elevated and decreased in cerebrospinal fluid (CSF) from NPC1 patients compared to controls. Using this discovery-based proteomics approach, proteins that show different levels may be used as markers of NPC1 and can be used to develop appropriate therapies and to monitor therapeutic options currently being evaluated. While NPC1 is known to have a heterogeneous clinical phenotype, the most severe and currently untreatable aspect is the progressive neurodegeneration. Building on our previous proteomics work carried out in the mouse model, we sought to extend or efforts, focused on the NPC1 patient CSF proteome. This project incorporates an existing mass spectrometry-based quantitative proteomics strategy termed isobaric tags for relative and absolute quantitation (iTRAQ). The commercially available iTRAQ chemical labels are provided in an 8-plex system. We further extend the multiplexing capabilities to a theoretical unlimited number of samples. Additionally, complete characterization of the iTRAQ strategy was carried out including new data analysis approaches for quantitative interpretation. Taking together, we have analyzed the CSF proteome from NPC1 patients and controls and have identified proteins that are different between the two groups which may serve in future studies and clinical trials.

Results and Interpretation:

1) **Evaluation of iTRAQ:** Our initial experiments were geared to understand the variability of the iTRAQ reporter ions that are observed in the mass spectrometer and used for quantification. Therefore, using a cell lysate (HepG2) at equal molar concentration, the peptides generated from tryptic digestion of proteins were labeled with the iTRAQ labels and combined and analyzed. Figure 1 illustrate an (A) example fragmentation mass spectrum with the iTRAQ reporter ions included from a peptide measured in the HepG2 cell lysate. (B) Enlarged view of the iTRAQ reporter ion region in spectrum. Taken together the sum of these data, the reporter ion ratios appear to be close to one with some variability.

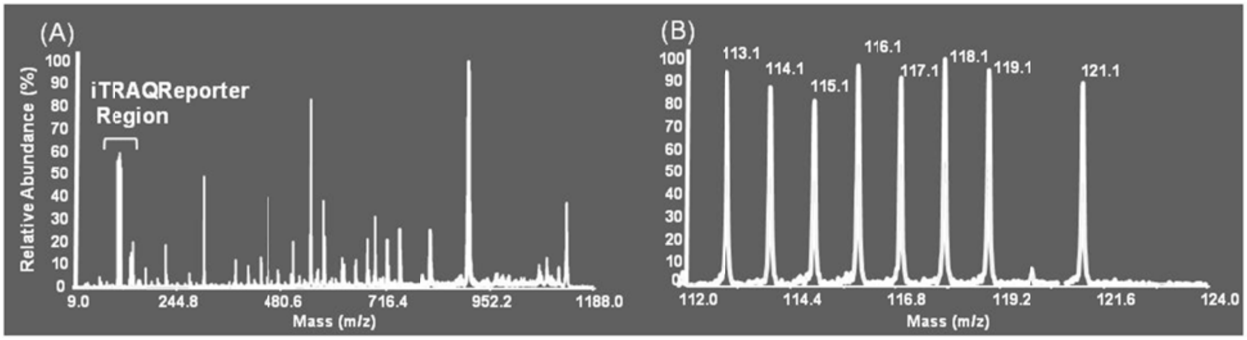

Figure 1. Tandem mass spectrum displaying peptide fragmentation (A) and iTRAQ reporter ions for quantification (B).

2) **Data Analysis Tools:** The HepG2 analyzed peptides provided an excellent resource to evaluate the variability of the technique and also to explore new approaches to interpret mass spectrometry generated data for protein quantification using iTRAQ. One area of interest to our research group is data normalization. The majority of commercially developed software platforms used for interpreting iTRAQ data assumes that the data is not normally distributed therefore; conversion into log-space is performed initially followed by non-parametric statistics. Regarding normalization, each iTRAQ reporter ion is normalized within the dataset such that a Gaussian distribution is expected and a correction factor is applied resulting in the median of the distribution representing no change in peptide and therefore protein levels. The dataset is then processed in this manner iteratively across the eight iTRAQ labels. Protein fold-change values then are represented from the mean values obtained for all of the unique peptide identified from a protein by comparison of one iTRAQ label to another (e.g., 113 vs. 115).

We initially evaluated options to normalize the datasets that did not require column-based normalization with the argument that normalization within a spectrum is more reliable given that the iTRAQ reporter ion spectra are collected under the same conditions therefore, not having to account for differences in fragmentation efficiency. As such, we considered the concept of row-normalization (or within spectrum normalization). This is shown in Equation 1 below, in which the peak area of any given iTRAQ reporter ion is then divided by the sum of all of the iTRAQ reporter ions in the sample. For the equal molar mixture, we would expect normalized area values then to be 1/8th of the initial peak area, or 0.125.

$$\text{Normalized Reporter Signal} = \frac{i}{\Sigma\ (i=1\text{-}8)} \qquad \text{(Eq. 1)}$$

Using the row-based normalization, several aspects of the dataset were then evaluated. Initially, we extracted the distribution of normalized peak areas for a single reporter ion, the 113 m/z ion. Figure 2 shows this data in which a Gaussian distribution is observed. Take note that this data was not logarithmically transformed. Additionally, some datasets suffer from truncation at the low ratio range, which is not the case for the currently explored dataset.

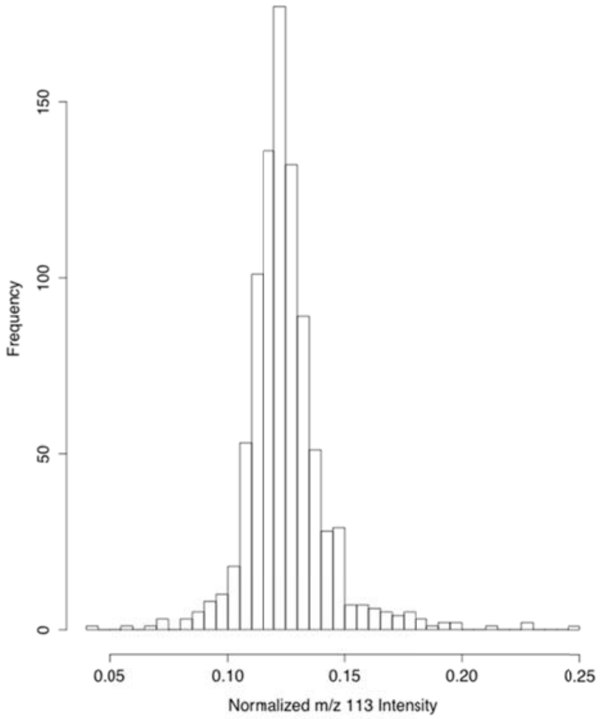

Figure 2. Distribution of iTRAQ 113 label peak area following row-based normalization.

Taking all eight reporter ions into account, the row-normalized distribution is shown in Figure 3. A normal distribution is observed and the mean is centered on 0.125, left panel. The data was also extracted such that each reporter ion was evaluated individually, and the box and whisker plot of the distribution of each reporter ion is provided in Figure 3, right panel.

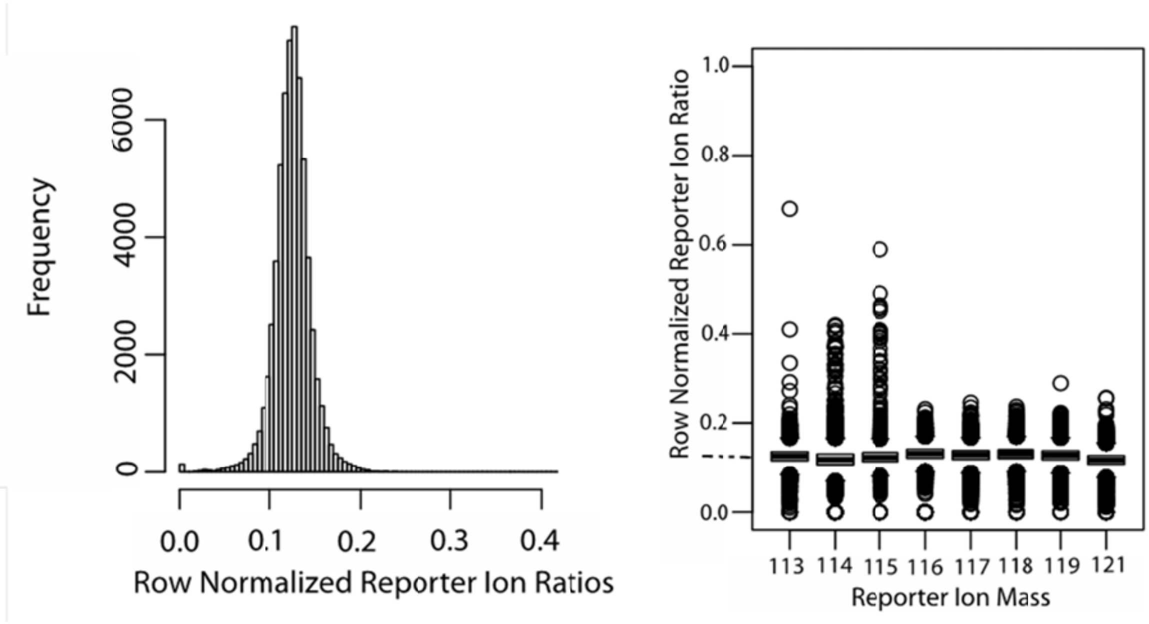

Figure 3. Row-normalized iTRAQ reporter ion collected from HepG2 cell lysate analysis. Data is shown for entire dataset (left) and broken out by each reporter ion (right).

The row-normalization was carried out using the R programming language and written in our laboratory. We are currently working to implement this step prior to database searching to obtain streamlined datasets that include protein identification and quantification information. The standard method for protein identification in our laboratory is the submission of the raw dataset to the Mascot Search Engine. The Mascot output file is then read into secondary filtering software called Scaffold which only includes high confidence peptide assignments and also performs quantitation as described above using the iTRAQ reporter ions. Based on the initial row-normalized strategy, we established a fold-change cut off value of peptides to be 1.7 in order to be outside of the 99% confidence interval, thereby displaying peak areas beyond the normal distribution for this chemistry. Ongoing work is focused on combining both row-normalization and column-normalization for analysis of these datasets. Additionally, we are working with our collaborators from Scaffold to add additional features into the software based on these studies including custom fold-change requirements to implement these findings.

3) **Protein Markers in NPC1:** The final aspect of this project was to incorporate a reference sample which allows extension of the eight iTRAQ labels and provides the opportunity to evaluate data across the NPC1 patient cohort. An illustration of this method is provided in Figure 4. CSF from adult controls and NPC1 patients is processed to remove the high abundant protein, albumin. The resulting proteome is then used for the mass spectrometry based proteomics experiments.

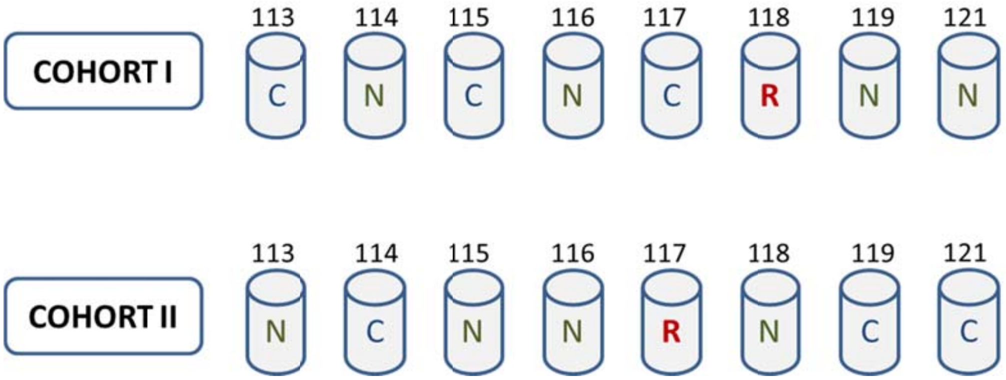

Figure 4. Experimental design for two NPC1 patient and control CSF cohorts analyzed via iTRAQ labeling and mass spectrometry detection. The bold R represents a pooled sample which is meant to represent the biological samples and is analyzed in each experiment. The N represents an NPC1 patient sample and C denotes a control CSF samples. The numbers 113-121 denote the iTRAQ label that is available in each commercial kit. Note that the labels are randomized in different cohort to prevent bias.

Our most up to date dataset has revealed that proteins altered in NPC1 are primarily secreted proteins. It should be noted that these data required a 1.5 fold-change cutoff value and a $p<0.05$ as determined by permutation testing (also referred to as a randomization test). Pathway analysis revealed that pathways that may be perturbed based on these findings include those involved in lipoprotein metabolism, arachidonic acid metabolism and the complement cascade among others. Examples of proteins that differ in NPC1 CSF versus controls are provided in Figure 5. Apolipoprotein E is a protein involved in cholesterol transport and metabolism and disturbances have been previously shown in NPC1. Superoxide dismutase is an oxidative stress marker and these results are consistent with previous data collected in our laboratory. Ceruloplasmin is a copper binding protein that has also recently been shown to be perturbed in NPC1. Finally, transthyretin is a protein involved in binding and transport of thyroid hormones and retinol. We previously found this protein to be increased in cerebellar tissue from *Npc1* mutant mice and observe a slight elevation in our current dataset. These proteins along with others are the current focus of our validation studies to confirm altered levels in NPC1 patients.

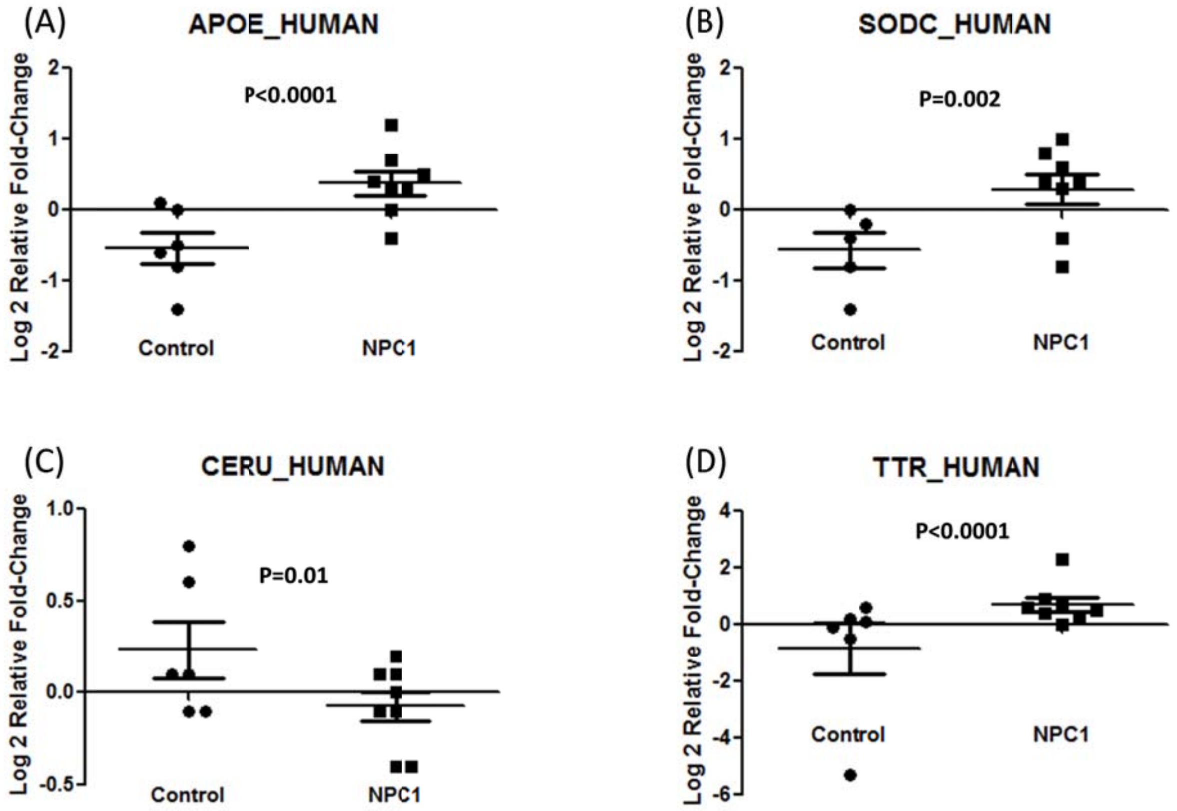

Figure 5 (A-D). Log2 relative fold change values for various proteins in CSF from controls and NPC1 patients relative to a pooled reference samples. The pooled reference sample was composed of CSF from 10 NPC1 patients therefore; the NPC1 values will typically be near zero. Each protein level was determined statistically significant via permutation testing and p-value is noted.

Conclusions: This project has contributed both to the quantitative proteomics field as well as the NPC1 field. Specifically, we have established methods for routine characterization of large mass spectrometry datasets using iTRAQ and have identified proteins that are altered in NPC1. Further efforts on this project are geared to validation of initial findings as well as evaluating the effects of various drug treatments on these proteins. Additionally, we are considered statistical approaches which account for other variables of the disease such as age of onset and severity.

Lay Summary: The goal of this project is to identify proteins in the cerebrospinal fluid (CSF) of NPC1 patients and compared the protein levels with unaffected controls. Our hypothesis is that identifying proteins that have different levels in NPC1 patients compared to controls will inform us about the changes in the brain that occur in NPC disease. The beginning of this project was to use an available technology to do protein identification and quantification. We have chosen to make modifications to the existing protocols which allow us to look at a large number of patient samples. To assist with this project, we have developed custom data analysis tools and methods to evaluate the resulting data. We have analyzed CSF from both groups and have identified several proteins that are elevated or decreased in the NPC1 patients compared to the controls. These potential protein marker candidates are currently being evaluated to further confirm their different levels. Going forward, the validated protein markers can be used to evaluate future therapeutic options in NPC1. The results of these studies will be published in peer-reviewed journals.