A novel translational analysis pipeline applied to the discovery and verification of potential biomarkers for type 2 diabetes.

Amol Prakash[1], Scott Peterman[2], Maryann Vogelsang[2], Dave Sarracino[2], Bryan Krastins[2], Patrick Muraca[3], Allison. B Goldfine[4], Mary Elizabeth Patti[4], Mary Lopez[2]

[1]Optys Tech Corporation, Philadelphia PA 19146

[2]Thermo Scientific BRIMS, Cambridge MA 02139

[3] Nuclea Biotechnologies, Cambridge, MA

[4] Joslin Diabetes Center and Harvard Medical School, Boston, MA

The translational proteomics pipeline requires a workflow that can provide unbiased discovery through targeted verification of biomarkers. This process can be long and painful since it typically requires integration of multiple pieces of data and, to date, has required a patchwork of different analysis tools that must be manually integrated since a complete solution has not been available. We have created a novel analysis package to address these challenges and applied it to a large set of plasma samples (obtained with full donor and IRB approval) with increased insulin resistance and type 2 diabetes. The complete dataset was acquired on a Thermo Scientific Q-Exactive mass spectrometer. In addition, we incorporated ultra-high performance liquid chromatography (UHPLC) to provide increased loading capacity, flow rates, peak capacity and superior (to nanoflow) robustness. In our analysis, we used a plasma-specific spectral library that was created using the current and also previous data sets acquired on a Thermo Scientific Orbitrap Fusion mass spectrometer. Over 500GB of raw data were obtained from the samples and searched using Proteome Discoverer v1.4 with SEQUEST search engine as well as a new spectral library-based searching algorithm. Using our novel tools, all the protein and peptide data were analyzed for quantification and differential abundance followed by unsupervised and supervised statistical analysis. Protein and peptide candidates with high statistical significance were chosen for the next stage of targeted verification and the targeted method export was automated and instrument specific.

The accurate identification of good biomarkers in a translational experiment requires very high data quality. To this end, we employed multiple steps to ensure robust, reproducible and high-confidence data. Before starting the project, we optimized multiple LC-MS platform parameters

to ensure that the system was suitable for long-term (weeks to months) high-quality and very stable data acquisition without interruption. This process was facilitated by the application of state-of-the-art algorithms that objectively evaluated the platform performance. Once platform stability was optimized and ensured, donor sample data were acquired and during the entire process multiple algorithms were applied to monitor system suitability by using internal standards.

The analysis tools and algorithms described above create a new translational proteomics pipeline that will be presented at the conference.