Development of Targeted Proteome Assays

Yale SCHOOL OF MEDICINE

Christopher Colangelo[1]; Thomas Abbott[1]; Mark Shifman[1]; Gordana Ivosev[2]; Lisa Chung[1]; Fumika Sakaue[1]; David Cox[2]; Stephen A Tate[2]; Erol Gulcicek[1],Ron Bonner[2]; Jesse Rinehart[3];Angus Nairn[1]; Kenneth Williams[1]
[1]Yale NIDA Proteomics Center, New Haven, CT; [2]AB Sciex, Concord, ON; [3]Department of Cellular & Molecular Physiology and Systems Biology Institute, Yale University School of Medicine, New Haven, CT

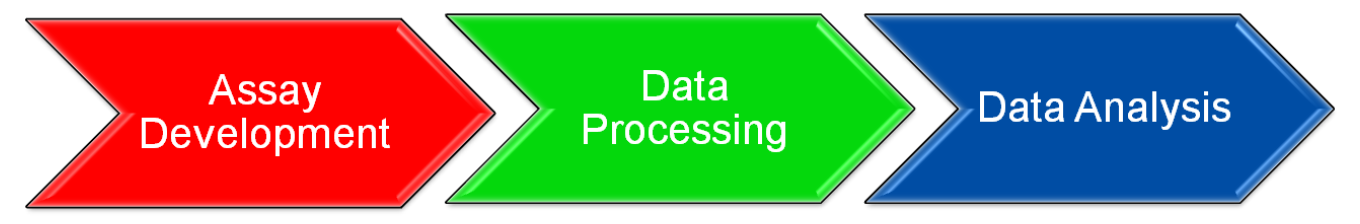
Overview

We have designed a comprehensive workflow for the development of routine LC-MRM proteome assays. Our initial efforts have resulted in two robust 90 minute LC-MRM assays for Human Red Blood Cell Membrane (RBC) and Mouse Post-Synaptic Density (PSD) fractions, which enable us to routinely quantify 57 and 119 proteins from 149 and 357 peptides respectively using 5 transitions/peptide (725-1785 data points). Since each sample is run in triplicate, a total of 2175 and 5355 quantitative data points were generated for each sample. In addition, we translated this same workflow to LC-SWATH targeted proteome assays that expand these quantitative assays four to ten fold. A direct comparison between LC-MRM and LC-SWATH is provided as measurement on the performance of multiplex quantitative assays, such as LC-SWATH.

Introduction

For ~15 years, large scale proteomic discovery has relied on massive LC-MS/MS to profile proteins in complex extracts. Problems with this approach are the limited dynamic range, poor run to run protein identification reproducibility, and the wide range in the number of peptides isolated from each identified protein. The latter results in MS/MS sequencing of many more peptides (>3) from some proteins than are needed to identify the parent protein. With complex mixtures this approach also must be coupled with off-line fractionation which results in numerous LC/MS/MS runs that require tens of hours of MS instrument time to detect and quantify 100-1,000 proteins in a complex mixture. As an example of the enormous duplication of effort with this approach, since 2007 the Keck MS/Proteomics Resource sequenced and stored in the Yale Protein Expression Database (YPED, Shifman et. al. JPT 2007) 18 million (p <0.01) with these peptides containing only 460,000 unique sequences or 2.5% of all YPED data. If we continue to use the same LC-MS/MS approach then 95% of our instrument time will be wasted by resequencing the same abundant peptides in each experiment. As a better approach, we are developing 90 min LC-MRM assays that relatively or absolutely quantify at least 100 targeted proteins by quantifying 5 MS/MS transitions from each of 3 peptides/protein that provide the best MS response. Among the first MRM Proteome assays that will be implemented are those for rat post-synaptic density (PSD) and Human Red Blood Cells (RBC).

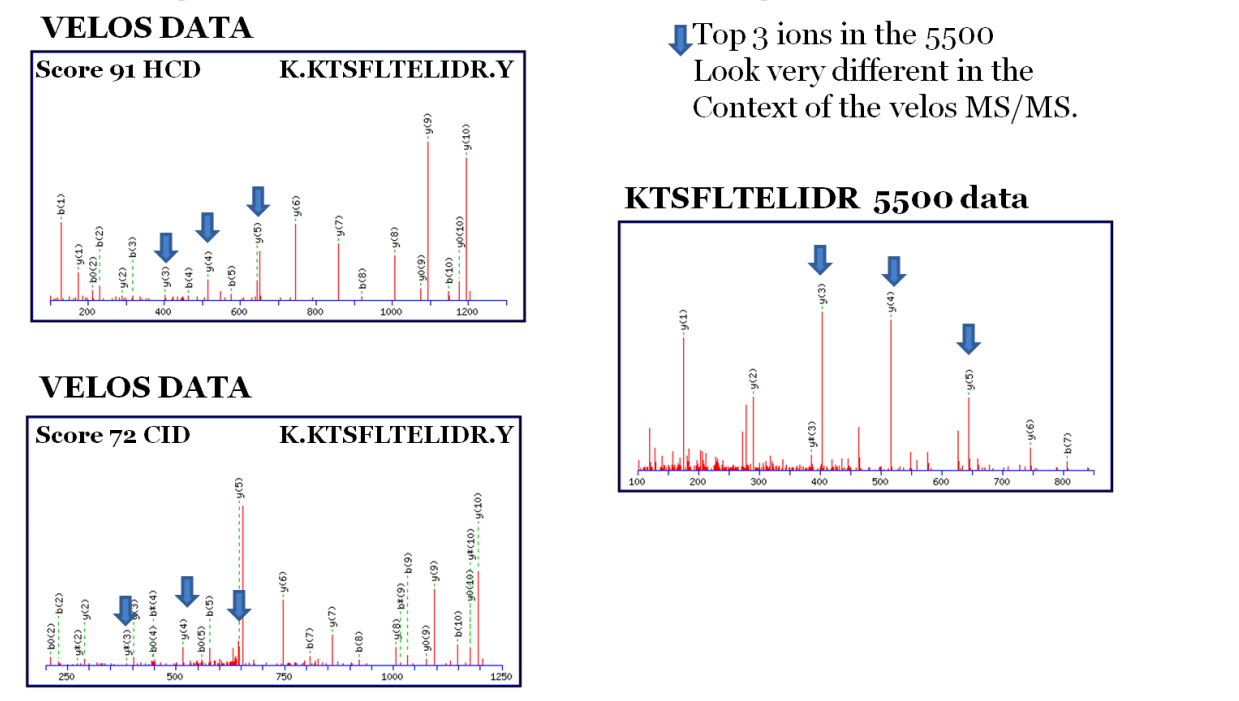
Methods/Workflow


Assay Development
Data Processing
Data Analysis

Assay Development	Data Processing	Data Analysis
Yale Protein Expression Database (YPED)	Multiquant – MRM Peakview – SWATH	R-scripts and Matlab
1 . Protein Identification 2. Spectral Library 3. MRM Transitions 4. Data Collection sMRM/SWATH	1 . Peak Integration 2. Data Metrics 3. Confidence Values	1 . Normalization 2. Cluster Dendrograms 3. PCA analysis 4. Correlation plots 5. Fold change analysis

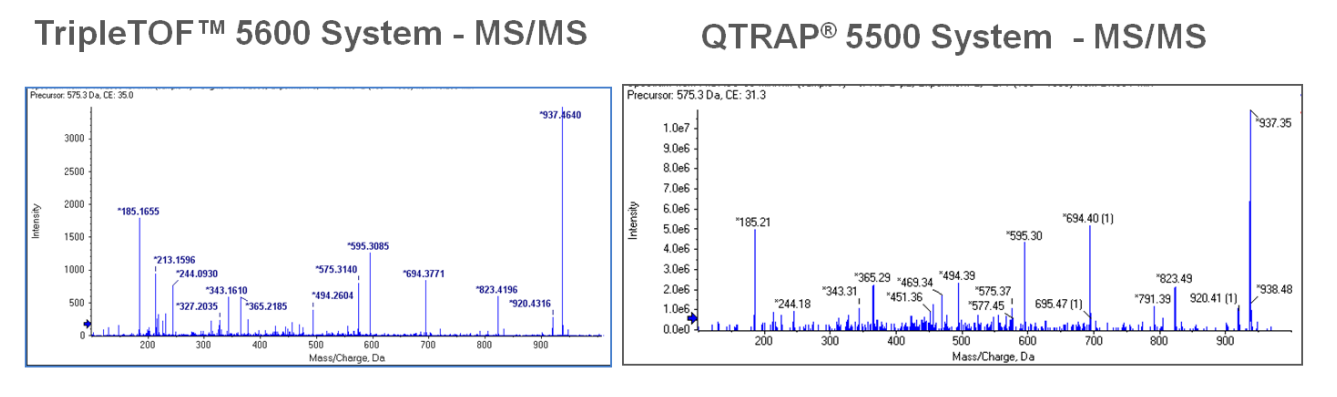
The TripleTOF 5600 MS was used to "sequence" peptides from a tryptic digest. Yale Protein Expression Database (YPED) was then used to translate the "learned" peptide sequences into a triggered LC-MRM (xMRM) assay that were run in triplicate on a QTRAP 5500 MS. The resulting LC-xMRM data was processed with MultiQuant software utilizing a newly developed SignalFinder Research algorithm and exported to Excel. Peak areas from corresponding LC-SWATH analysis were extracted in Peakview using transitions from corresponding IDA runs. A suite of bioinformatics tools then provided assay metrics, data normalization, and peptide and protein fold change calculations. The resulting data were imported into YPED where users can view, subset and download their data through a secure Web interface.

Challenges in translating IDA datasets



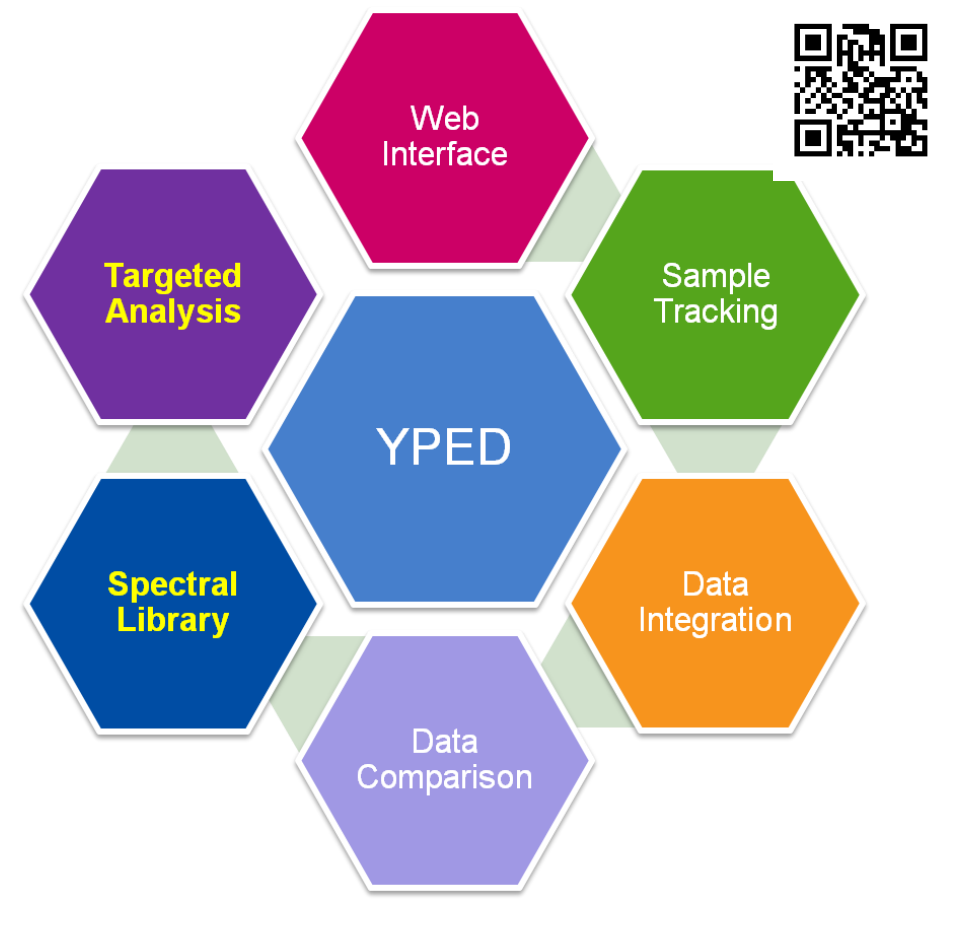
Top 3 ions in the 5500 Look very different in the Context of the velos MS/MS.

DATA acquired on the 5600 TripleTOF translates to peptide quantitation by MRM on 5500 QTRAP®Systems resulting in identical CID fragmentation profiles



Peptide LVNEVTEFAK; m/z 575.3 [2^+]

Yale Protein Expression Database (YPED)



We have developed an integrated, web-accessible software system called the Yale Protein Expression Database, or YPED, to address the storage, retrieval, and integrated analysis of high throughput proteomic and small molecule analyses. The interface supports sample submission, project management, sample tracking, data import, sample administration, and user billing. For data integration, YPED handles data from the following: LC-MS/MS protein identifications and protein post-translational modifications (phosphorylation, ubiquitination, acetylation, methylation, and others); identification/quantitation results from label-based proteomics experiments (DIGE, iTRAQ, ICAT, and SILAC); LC-MS based label-free quantitative proteomics; and targeted proteomics (MRM). YPED also serves as a peptide spectral library for all our protein database search identification results. As of March 1, 2012, YPED contained 12,113 datasets from 1,086 users, resulting in a current database of 441,950 unique LC-MS proteins and 2,304,375 distinct Mascot LC-MS peptides.

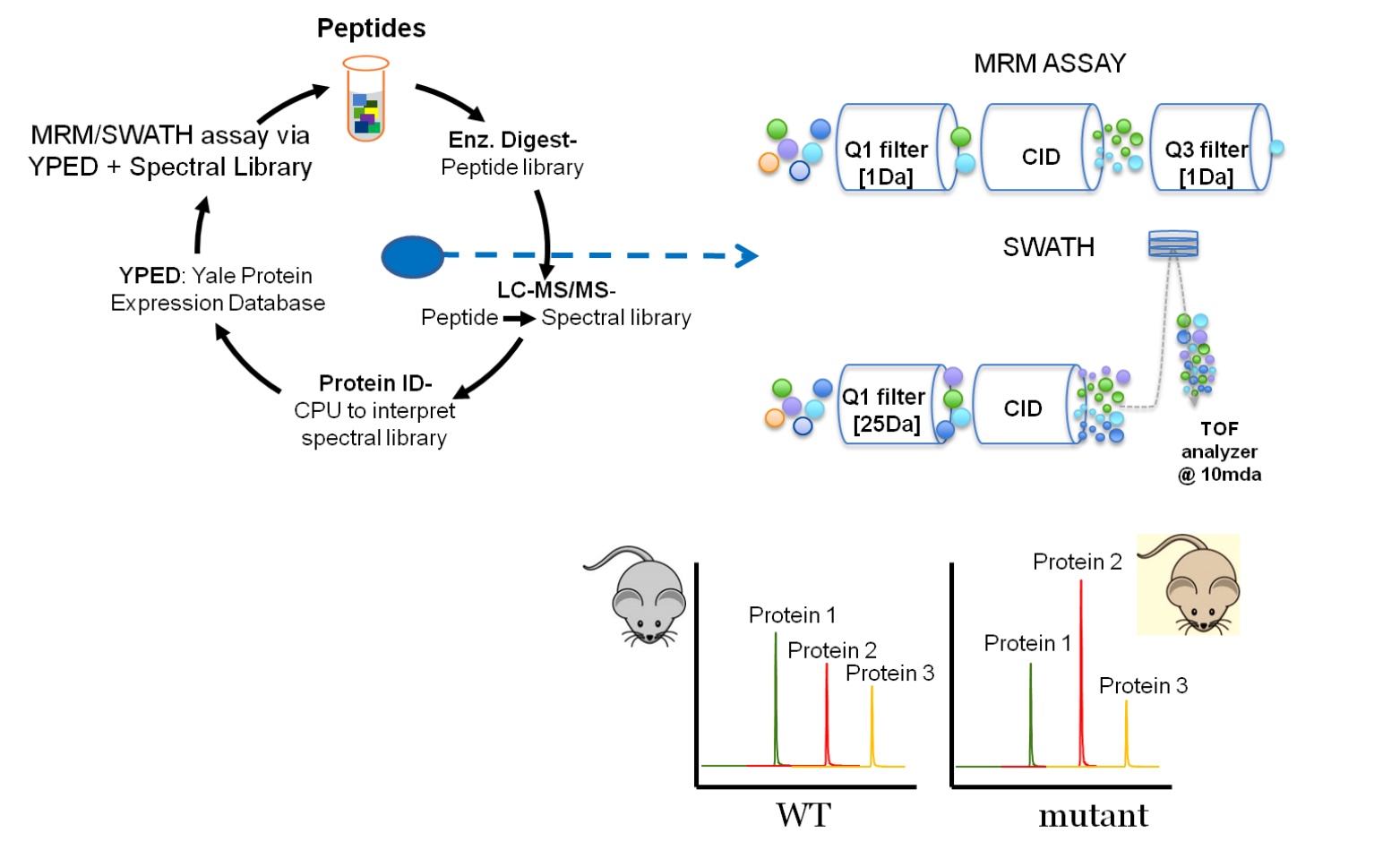
YPED Spectral Library

The table to the right summarizes the current size of our YPED peptide spectral library. The library is generated from a continual processing pipeline which periodically submits all the distinct protein identification peptides in YPED to BLAST for comparison against then current standard references such as UniProtKB/Swiss-Prot database.

Organism	Blast Protein Count	Blast Peptide Count
E.Coli	3,515	30,059
Yeast	5,344	40,368
Rat	10,544	75,080
Mouse	17,777	122,153
Human	18,366	129,289

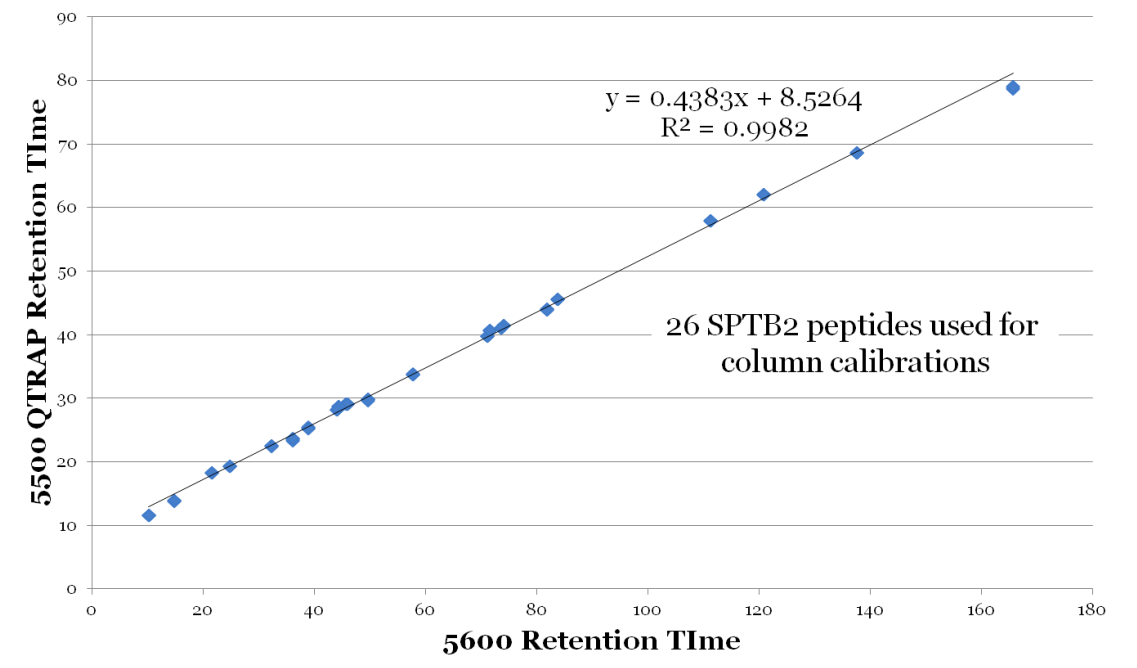
Assay Development

We developed a YPED tool to automatically transform discovery data into targeted MRM methods to construct a targeted MRM proteome assay. Data from discovery runs on a TripleTOF 5600 MS were database searched and peptide identifications were uploaded to YPED which outputs either a scheduled LC-MRM method for the QTRAP 5500 MS or a Peakview input file for SWATH acquisition on our 5600 TripleTOF.



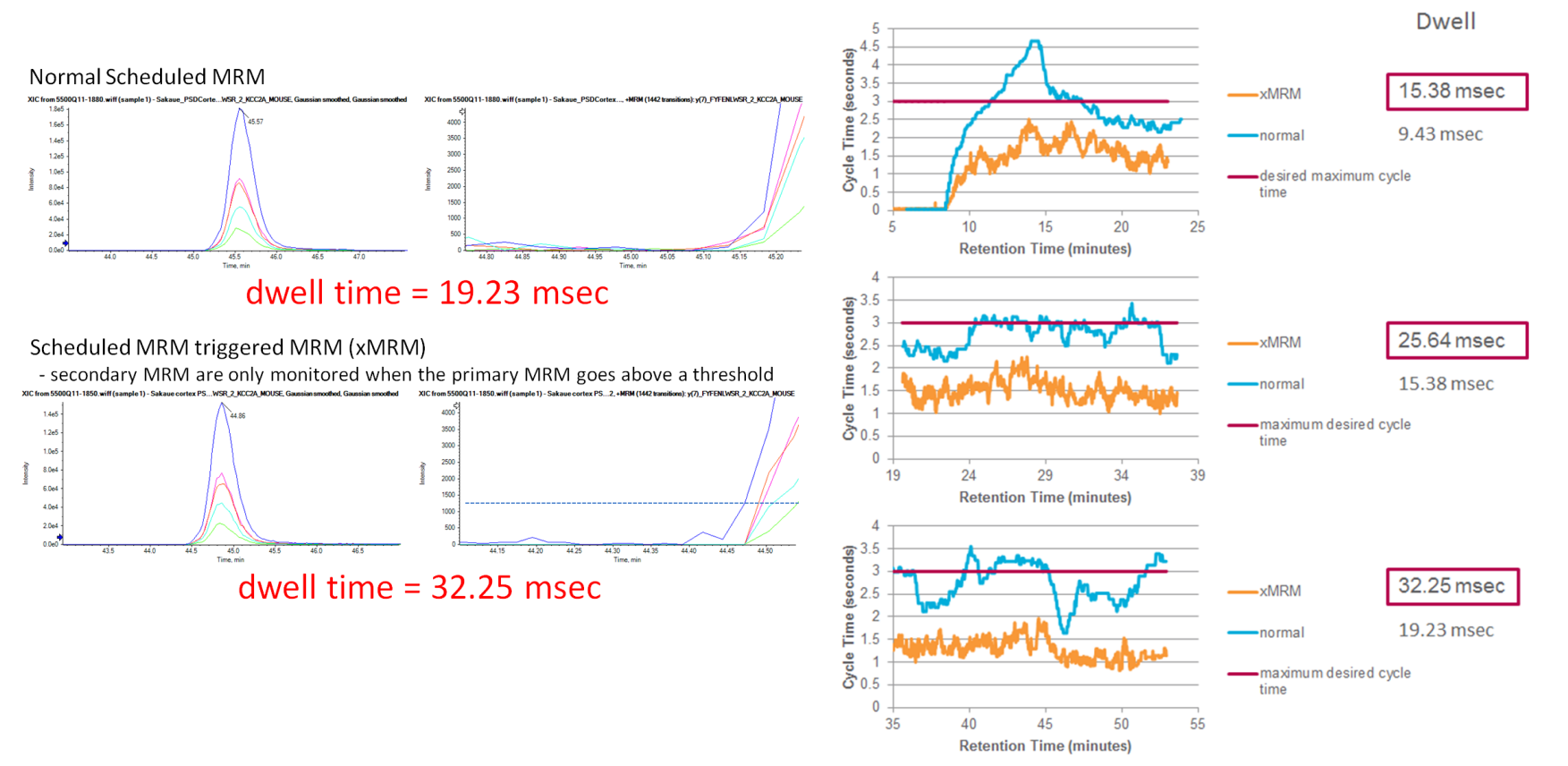
LC-MS Column Calibration

A key requirement for creating the scheduled MRM methods were consistent LC retention times across samples for each peptide. Since we use the same LC system and methods for LC/MS/MS discovery, LC-xMRM, and LC-SWATH validation runs, this allows us to assign precise retention times to each peptide identified from each protein, to incorporate this information into YPED, and to then generate a scheduled MRM method. The figure below shows the retention time correlation of 5600 TripleTOF and 5500 QTRAP on a Red Blood Cell sample when monitoring 26 Spectrin Beta Chain Peptides (SPTB2) monitored throughout our gradient.



Triggered xMRM

A novel triggered LC-MRM (xMRM) assay (research version of Analyst software) is critical to maximize cycle and dwell times since only the most intense group transition is designated as the primary transition and monitored throughout its entire scheduled window. Secondary MRMs for each peptide are only monitored when the primary MRM exceeds a predetermined threshold, enabling such large scale MRM assays. This reduced the number of MRM being monitored at any given time, improving dwell time while decreasing cycle time. The triggered xMRM transitions were used to obtain quantitative results.



The figures above show the cycle time for three transitions from the beginning, middle, and end of the gradient. The orange line is the cycle time using the xMRM assay and the blue line is from the normal scheduled MRM (sMRM) assay. The plots clearly indicate that by using xMRM the cycle time for each transition significantly decreases resulting in an increase in collection of more points across the peak. On the right hand side of the figure, the number indicating the dwell time for each of these transitions in either xMRM or sMRM mode, is indicated. Using xMRM also clearly showed improved dwell time for each transition, which means better signal to noise measurements using xMRM vs. sMRM.

Peak Integration

One main difficulty we had with Peak integration was that the MQ4 algorithm of Multiquant required visual confirmation of the integrated area of *each* transition. This manual intervention was needed at the time to correctly integrate many transitions in each analysis. Thus, we worked to develop Signal Finder research, which as shown below significantly reduces the % of peaks required for area correction.

Integration algorithm	% of peaks that required group RT correction	% of peaks that required area correction	Notes
MQ4	5	20	Extended peak end caused increased peak area on more than 5% of peaks
SignalFinder MQ 2.1	5	20	Baseline issue and bad peak model caused missed peaks/underestimated area. Long Processing time
SignalFinder research	5	9	Some low S/N peaks needed correction, but processing time was reduced 3-fold

Signal/Noise Peptide Metric Plot

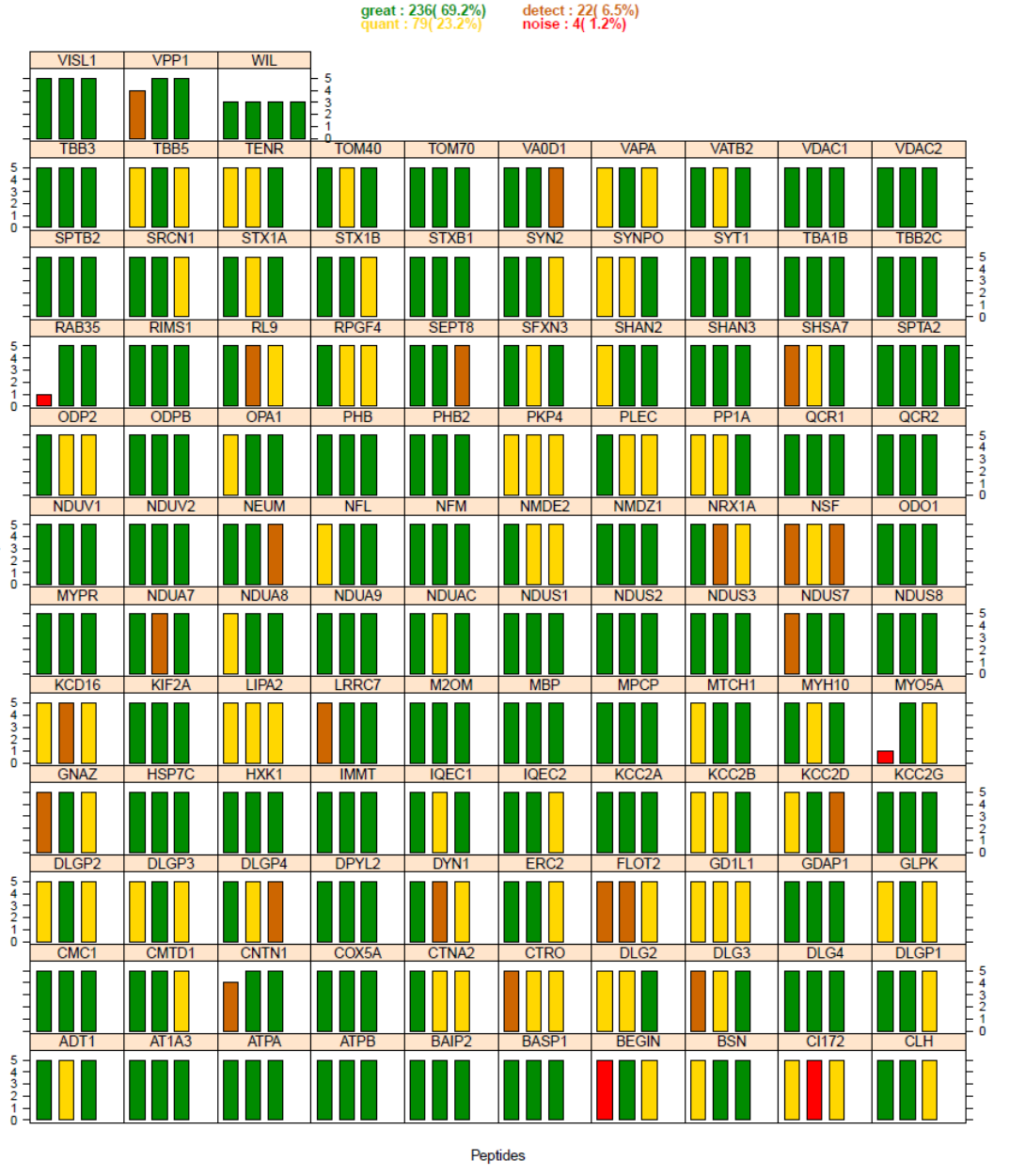


Table 2: Criteria for Categorizing Quantitative Protein MRM Data

S/N	S/N Category		
	Description	Bar Graph Color	Number Designation
Greater than 10	Great	Green	5
Between 5 and 10	Quant	Yellow	4
Between 3 and 5	Detect	Orange	3
Below 3	Noise	Red	2
Missing peak	N/A	N/A	1

Metric Bar Graph of PSD xMRM assay quality. The plot above is broken into 113 individual bar graphs for each protein and internal standard (WIL) in the PSD assay and each bar graph is further sub divided into individual bars which represents the data quality of single peptide with 5 transitions/peptide (except the WIL internal standard peptides which have 3 transitions/peptide). The height of each bar corresponds to the number of MS/MS transitions observed for the corresponding peptide with all proteins potentially having 5 transitions as compared to the maximum of 3 transitions that were monitored for each of the internal standard peptides.Finally, the color of each bar depicts the average signal/noise ratio for the underlying, usually, 5 transitions as described in Table 2.

PSD Cortex—xMRM and SWATH Fold Change Comparisons

We ran three rat brain cortex post-synaptic density factions on our 5600 TripleTOF mass spectrometer using 180 min LC-MS runs. After MASCOT database search we identified 1,574 unique rodent proteins. Using this list we generated one scheduled MRM assay of 1697 transitions for 111 proteins from 337 peptides. We then ran six PSD cortex biological replicates in triplicate across the 1697 transitions PSD MRM proteome assay described to the right for total of 30,546 transitions. We also ran the same six PSD cortex biological replicates in triplicate with SWATH aquisition. For comparison analysis between xMRM and SWATH extracted the same 1697 transitions as our xMRM assay and use our R and Matlab fold change analysis tools for the plots and calculations. **[shown below]**

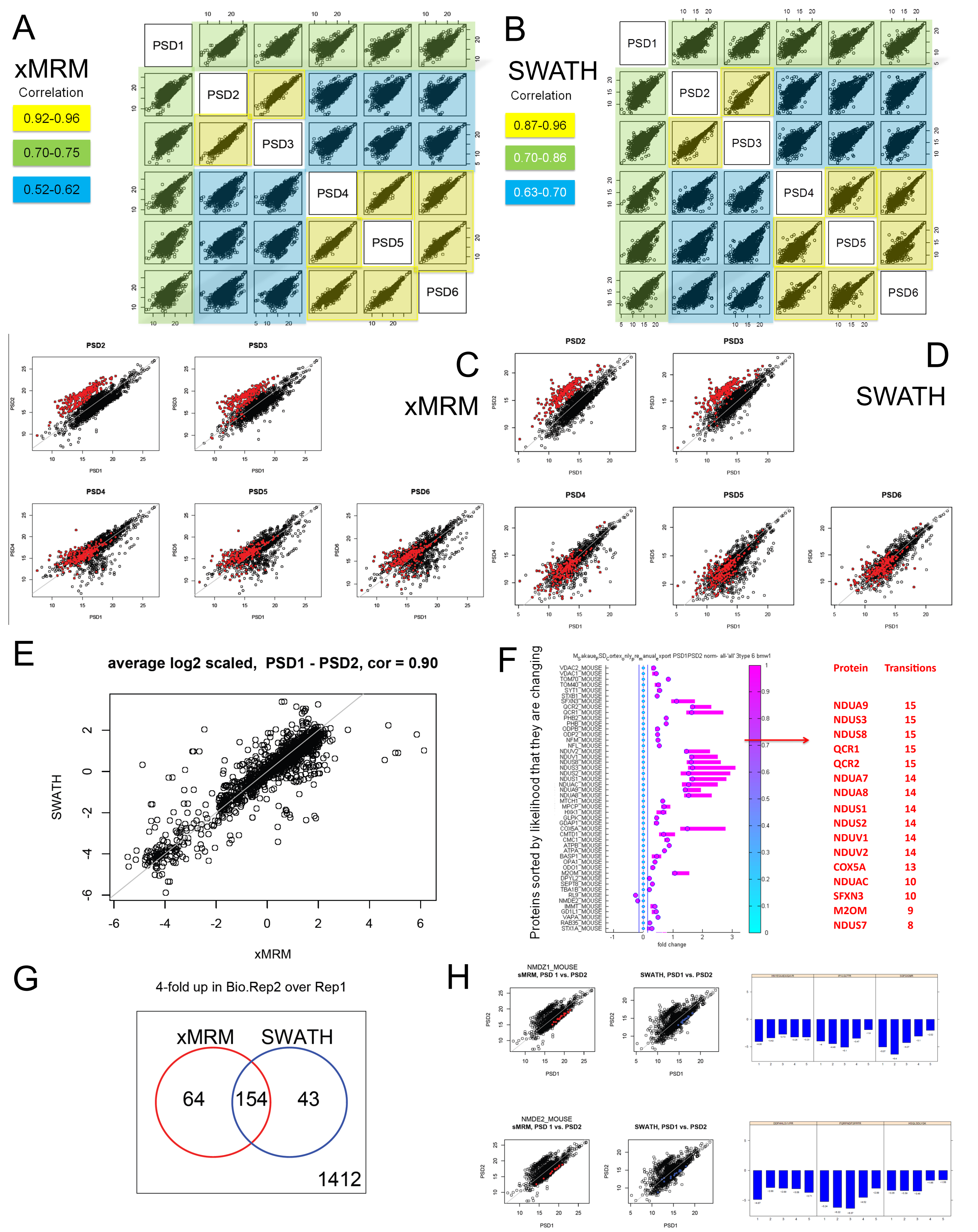
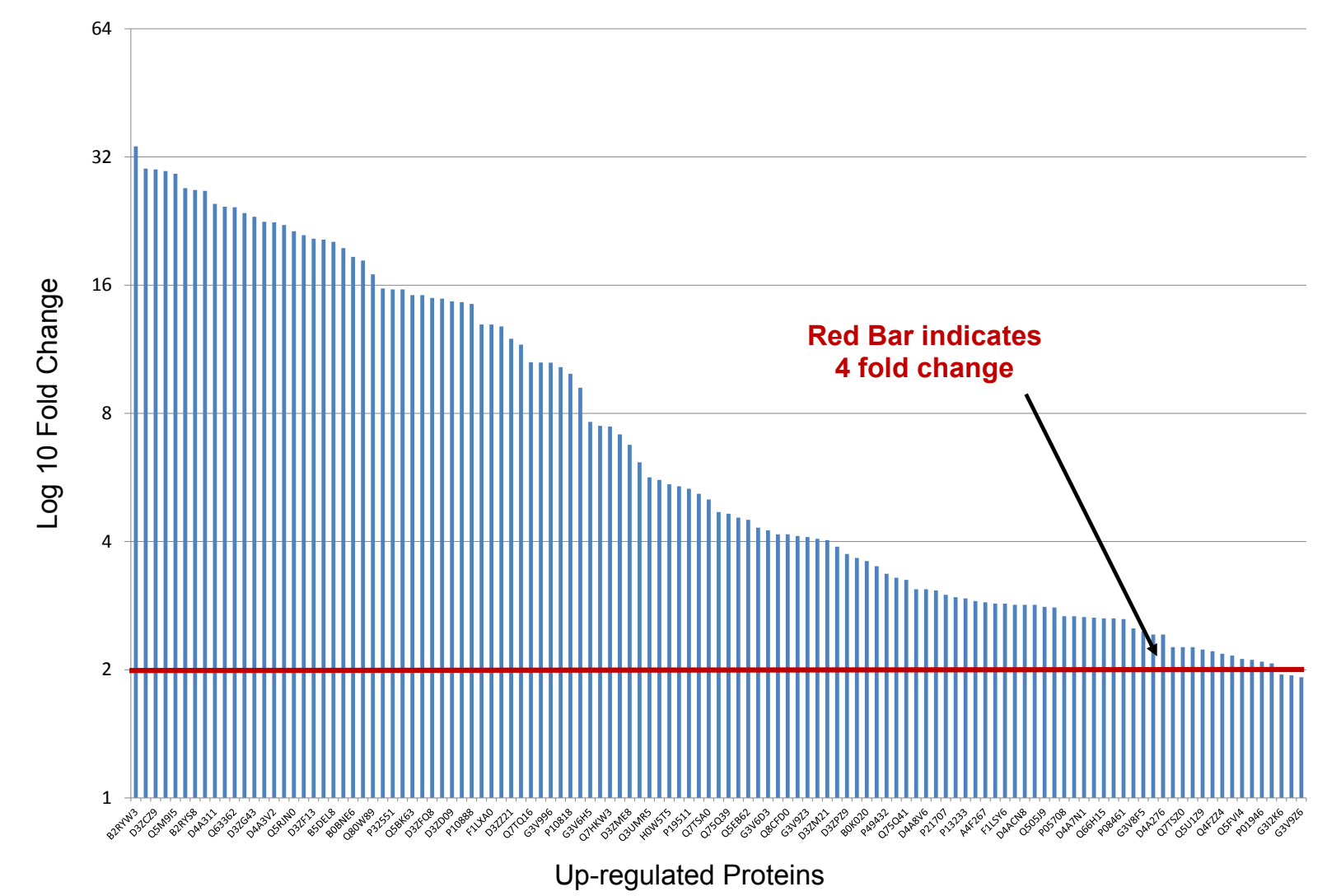


Figure A and B are xMRM and SWATH log2 scatter plots between six PSD samples averaged over three technical replicates/sample. The color shading indicates the Pearson correlation between paired samples. **Figure C and Figure D** are xMRM and SWATH log 2 scatter plots between PSD 1 and the other five PSD biological replicates. The red dots indicates transitions that are four fold up-regulated in PSD 2 vs. PSD 1. These same transitions are then mapped onto the other four scatter plots. **Figure E** is a log 2 fold change scatter plot between xMRM and SWATH for PSD2 vs. PSD 1. A correlation of 0.90 shows that the two methods produce consistent fold-changes between each other. **Figure F** is a Weighted Protein Fold Change plot for PSD 2 vs. PSD 1 for xMRM. The proteins listed to the right are all mitochondrial proteins and the number of transitions that were four fold up-regulated are also displayed. Restricting the analysis to only those peptide quantified in all samples should decrease the error bars and then allow fold change calculations. The pink dots represent highly confident weighted fold-change values and since the dot is a combined weighted fold-change value it represents a more accurate measure of the actual fold-change between the samples. **Figure G** is a Venn Diagram of 4-fold or greater transitions for PSD 2 vs. PSD 1. 75% of the transitions with > 4-fold up-regulation in the venn diagram overlap. The 1412 remaining transitions in the assay were below 4-fold in both xMRM and SWATH. **Figure H** are two examples of down-regulated proteins in PSD 2 vs. 1 in both xMRM and SWATH.

Expanding PSD Protein Quantitation with SWATH

From six LC-SWATH PSD Cortex samples (18 runs) we extracted 56,000 transitions (>1,200 proteins) in each run for a total of 1,000,800 data points. After minimum variance normalization and fold change analysis of PSD 2 vs. PSD 1 we are able to expand the number of >4 fold up-regulated proteins from PSD Cortex xMRM assay to 101 proteins for a 6 fold increase. As a result of this finding we carefully re-examined the protocols used to prepare these control samples and found experimental variables that may well explain the apparent "up-regulation" of proteins in PSD 2. The plot below shows Log 10 Fold Change for SWATH PSD1 vs. PSD 2 with proteins that had 5 or more peptides and a weighted confidence of 0.80 or greater (Bisson et. al. Nature Bio 2011). There are 101 proteins on the bar chart but due to space only a portion of the proteins are labeled.

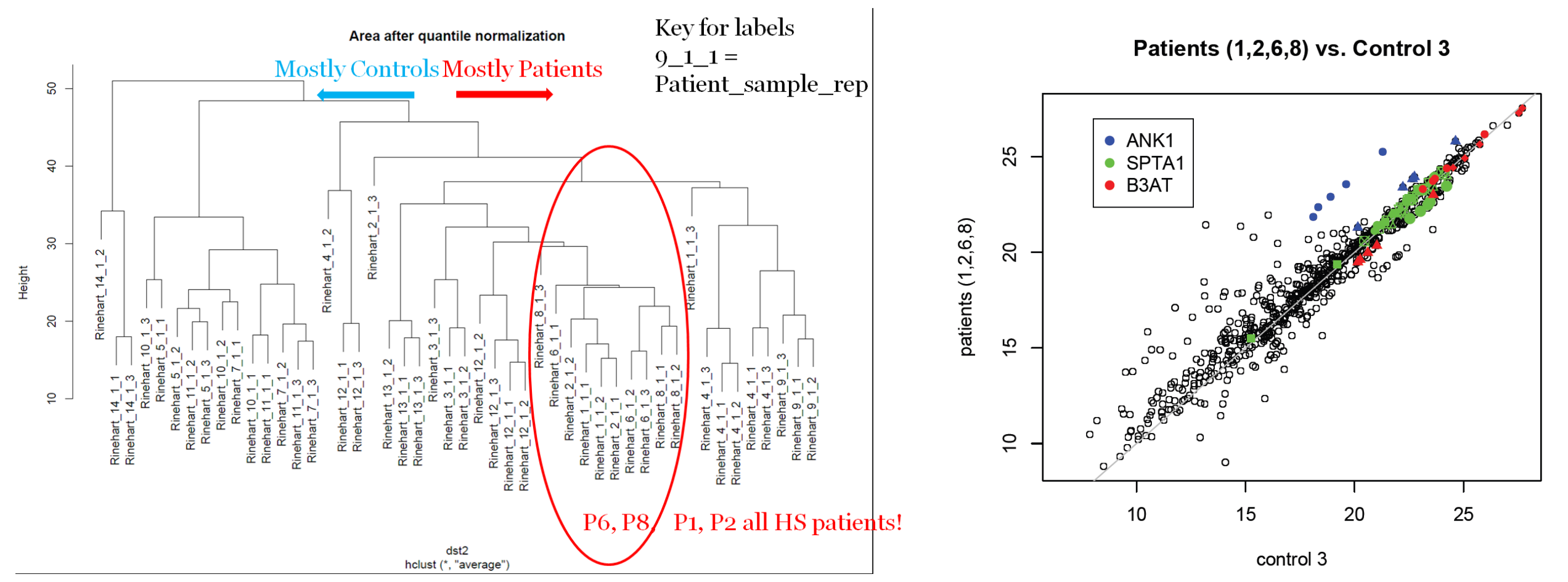
LC-SWATH Fold Change for PSD2 vs. PSD1



Up-regulated Proteins
(with confidence >0.8 and 5 or more peptides)

Red Blood Cell —xMRM

The RBC Proteome is estimated to contain approximately 1000 proteins. From our TripleTOF 5600 platform, we identified 750 proteins in a single 90 min IDA LC-MS/MS run. Rinehart and colleagues recently used the RBC discovery data to identify a splice variant of the low abundant RBC membrane protein Piezo1 (Zarvchanski et al., Blood 2012). This data was translated to our current LC-xMRM assay, which enabled the rapid quantitation 54 proteins (149 peptides and 725 transitions). We ran a control and patient sample on our 5600 TripleTOF mass spectrometer using 180 min LC-MS runs and using a similar approach as described for RAT PSD we ended up with one scheduled MRM assay of 807 transitions for 63 proteins from 171 peptides. We ran the targeted assay on 14 samples (9 patients and 6 control) in triplicate with randomized order. After data processing and quantile normalization, we plotted a peak area cluster dendrogram below. The results show classification of patients from controls with a cluster of four HS (Heredity Spherocytosis) patients. Interestingly, the proteome data cluster when genetic analysis shows a diverse set of primary mutations in the patients (genome data not shown). Also plotted are the log 2 scatter plots for four HS patients vs. RBC Control sample, showing evidence of Ankyrin-1 (ANK1) up-regulation as compared to no fold change for Spectrin Alpha Chain (SPTA1) and Band-3 (B3AT). Interestingly, Ankyrin gene defects and protein expression abnormalities are common in HS.



Conclusions

- Developed a Pipeline which includes Targeted Assay Development, Data Processing, and Data Analysis tools for both LC-MRM and LC-SWATH assays.
- Utilized Triggered xMRM (research version of Analyst) improving both dwell time while decreasing cycle time during LC-MRM runs.
- Improved Peak integration using Signal Finder Research that significantly reduces the % of peaks required for area correction.
- Developed new data metrics and automated R plots, such as Signal/Noise Peptide Metric Plot.
- In-house developed robust normalization algorithms and confidence weighted fold-change analysis for transition to peptide to protein level.
- Utilized entire pipeline to show virtually identical fold-change values for both LC-MRM and LC-SWATH on Rat PSD Cortex samples
- Demonstrated the ability of SWATH to expand our targeted proteomic assays 6-fold in PSD Cortex samples
- Analyzed 14 Human Red Blood Cell Patients in a randomized study and were able to effectually classify patients vs. control as well cluster common disease phenotypes based on common proteome variations.

Funding

Yale NIDA Proteomics Center
1 P30 DA018343 (NIDA/NIH)

NIDDK
NIH/NIDDK K01 DK089006

Yale Center for Clinical Investigation
ULRR024139 (CTSA/NCRR/NIH)