Digitizing the Proteomes From Big Tissue Biobanks

Analyzing 24 Proteomes Per Day by Microflow SWATH® Acquisition and Spectronaut Pulsar Analysis

Jan Muntel[1], Nick Morrice[2], Roland M. Bruderer[1], Lukas Reiter[1]
[1]Biognosys, Switzerland, [2]SCIEX, USA

Tissue biopsies have been preserved and stored in biobanks for more than a century in the hope that their future analysis will provide a better understanding of health and disease. One of the most common methods of preserving these tissue samples is by formalin-fixed paraffin-embedded (FFPE). These samples are often very well characterized by classical pathological methods and provide great potential for precision medicine and the discovery of new diagnostic/stratification markers and therapeutic targets.

A powerful way to take advantage of this repository is to quantify large numbers of proteins across all the samples so that correlations can be made with respect to various health and disease states. Such an endeavor would require highly reproducible sample preparation, a robust analytical platform for high throughput sample analysis, as well as robust data analysis. Current LC-MS/MS proteomics tools now allow for the reproducible quantitation of 1,000s of proteins in a single run. In particular, SWATH® Acquisition has been shown to provide the very good data completeness, reproducibility, and quantitative precision in comparative studies.[1] When coupled with microflow chromatography, sample throughput and assay robustness are improved while still maintaining similar overall workflow sensitivity to nanoflow separations.

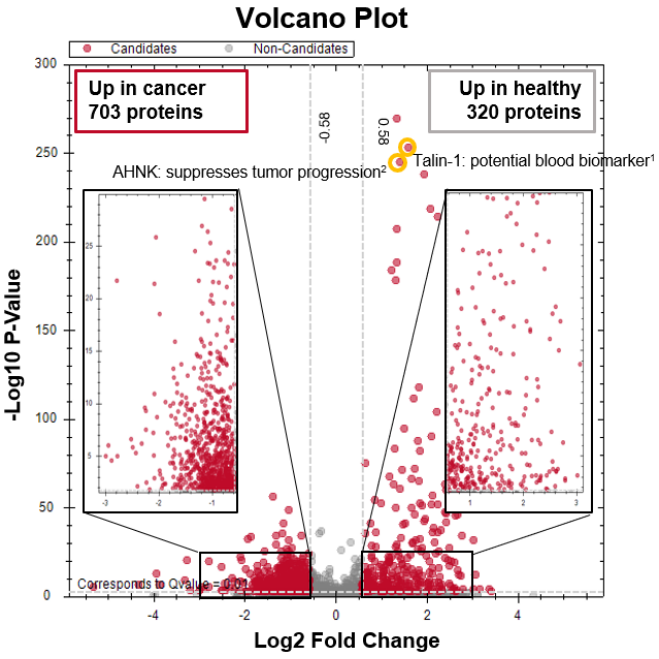

Figure 1. Differential Proteins Between Healthy and Cancer Samples. Using a t-test, 1,023 proteins were found to be significantly altered in abundance between healthy and cancer samples with the majority of those (703) found in increased amount in cancer tissue.

Here, microflow SWATH Acquisition[2] was used to generate quantitative proteomics data on a cohort of colon cancer samples from a biobank. This study demonstrates how high throughput proteomics can be used to interrogate these precious samples from biobanks and how this research can pave the way to a better understanding of health and disease.

Key Features of Biobank Proteomics using SWATH® Acquisition

- FFPE tissue samples can be reproducibly prepared and analyzed to identify and quantify all proteins within samples
- Results can be compared across large cohorts to identify potential diagnostic/stratification and therapeutic markers
- SWATH® Acquisition on TripleTOF® 6600 System provides high analytical depth and reproducibility for protein quantitation
- Microflow chromatography provides increased assay throughput and robustness
- Spectronaut Pulsar software provides fast data processing for protein quantification and advanced results processing including various statistical and gene ontology analyses

Methods

Sample Preparation: The formalin fixed paraffin embedded (FFPE) colon tissue samples were ordered from a public repository. These samples were classified as healthy or disease according to clinically accepted protocols. Protein extraction and tryptic digestion of a 10 µm slice were performed using an adapted protocol[3] and resulted on average in 140 µg of protein per slice. Prior to LC-MS analysis digests were spiked with iRT peptides (Biognosys) for retention time normalization. Six µg of protein digest was used per run.

Chromatography: Separation was performed using a Triart C18 150 x 0.3 mm column (YMC) coupled to a NanoLC™ 425 system (SCIEX). A non-linear 43 min gradient was used at a flow rate of 5 µL/min.

Mass Spectrometry: Data acquisition was performed using a TripleTOF® 6600 system (SCIEX) with Turbo V™ Source plumbed with microflow hybrid electrodes. SWATH Acquisition method consisted of 120 variable windows, 18 msec MS/MS accumulation time, and one 250 msec MS scan. The cycle time of the method was 2.4 sec resulting in 6 data points per LC peak. Total run time per sample was ~1 hour such that the whole project was completed in <5 days. A spectral library was generated by fractionating a pooled sample using high pH reverse phase fractionation (pooled digests from 10 healthy and 20 cancer samples). These samples were analyzed with the same LC setup using a standard data dependent acquisition (DDA) method.

Data Processing: DDA data were searched against the human UniProt database using the Pulsar search engine (Biognosys) and a library was generated using 3-6 fragment ions per precursor. The library comprised 49,176 precursors, 44,807 peptides and 5,499 protein groups. This library was then used for data analysis of the SWATH Acquisition data in Spectronaut Pulsar (Biognosys) using default parameters. The analysis of the whole dataset took ~30 hours. All data were filtered by 1% FDR on precursor and protein level. Protein grouping was performed based on the ID picker algorithm. Data were normalized by local regression normalization. Statistical testing for differential protein abundance was done using the Spectronaut pipeline (t-test, multiple testing correction after Storey).

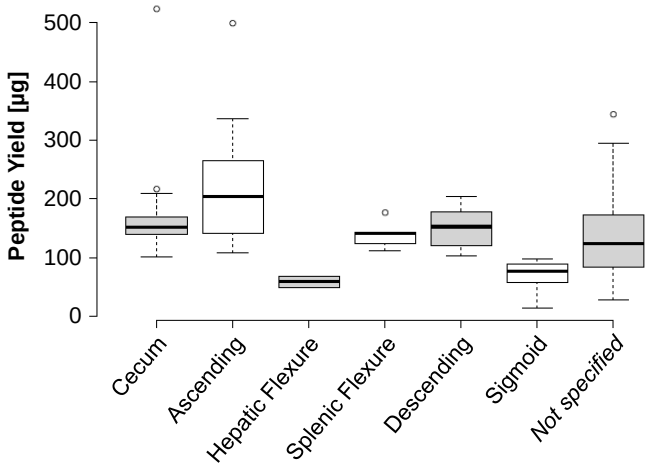

Figure 3. Assessing Sample Preparation Reproducibility. Boxplots of peptide yields from the healthy and cancerous FFPE sample preparations. Sample preparation was highly reproducible with CV = ~10%.

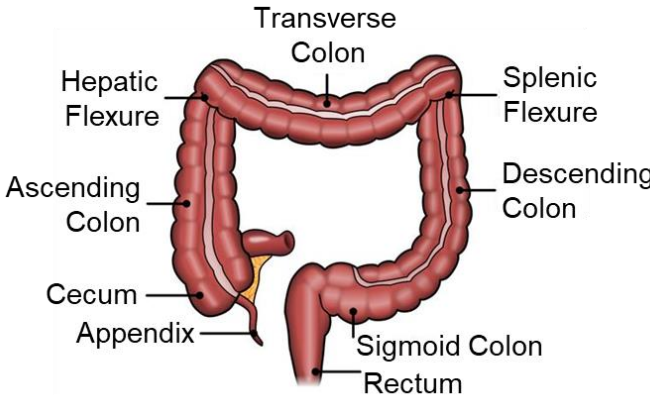

Figure 2. Overview of Colon and Resection Sites of the 105 Tissue Samples. Samples (cancer/healthy) were from: Cecum (16/1), Ascending (17/0), Right hepatic flexure (2/0), Left splenic flexure (5/0), Descending (12/0), Sigmoid (21/3), and Not specified (22/6).

Highly Reproducible Sample Preparation

Colon and rectal cancers rank third in the USA with respect to the number of new cases and number of deaths per year[4] with 1 out of 17 people developing colorectal cancer, highlighting the importance of performing biomarker research in this disease area. This current study was comprised of 105 FFPE tissue samples consisting of 95 cancer samples and 10 healthy samples from various resection sites of the colon (Figure 2). The sample preparation produced a high peptide yield without biases towards the colon region. On average, one slice yielded 140 µg of total protein with slightly lower yields for hepatic flexor and sigmoid region (Figure 3). Overall the sample preparation was highly reproducible with CV=~10%.

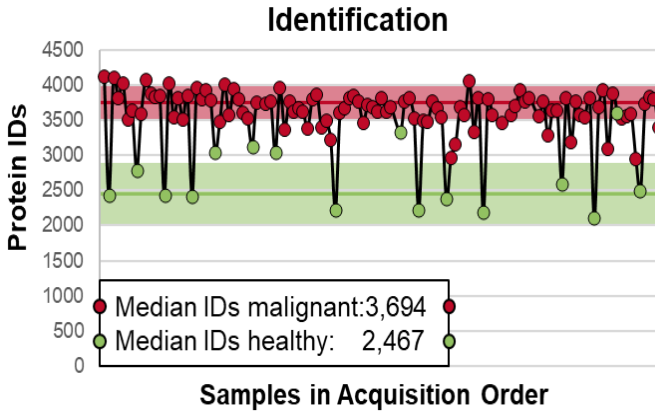

Figure 4. Protein Identification Results. Overview of protein group identifications across the sample set organized by acquisition order. Colored boxes indicate one standard deviation from the median IDs within the cohorts (green = healthy; red = cancer).

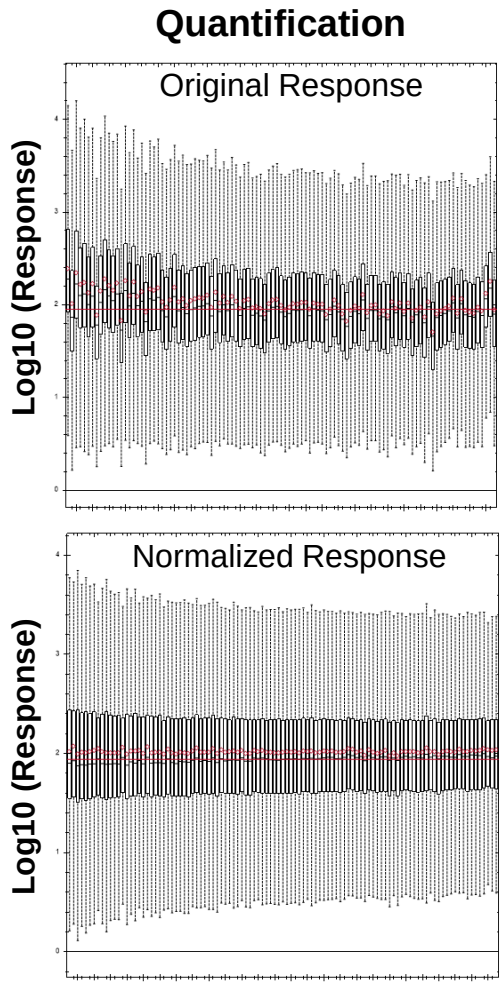

Figure 5. Normalization Provides Corrects for Small Signal Changes Across Study. Over time, a slight decrease in overall intensity was observed (top). This was corrected for by normalization in Spectronaut software (bottom).

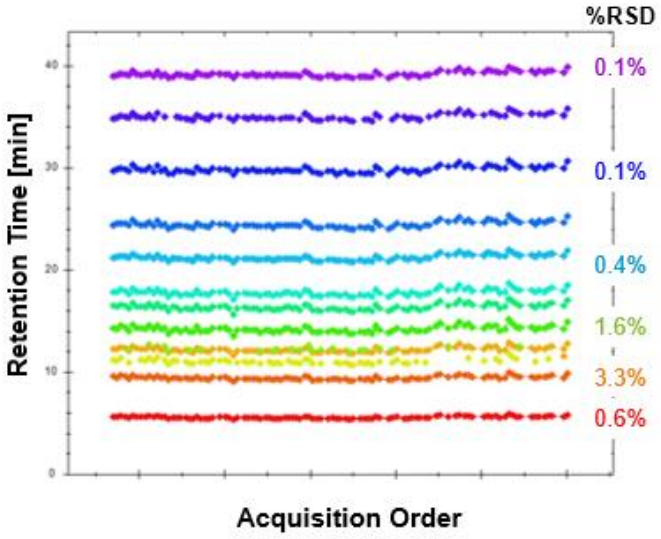

Figure 6. Chromatographic Reproducibility. Retention times were highly stable across the study, as highlighted by these select 12 peptides across the retention time range. The retention time RSD is shown for some peptides on the right. The median RSD for the (entire?) sample set was 0.4%.

A Highly Robust Workflow

SWATH Acquisition uses a library for data analysis. In this study, the spectral library that was generated comprised 5,499 protein groups, 44,807 peptides and 49,176 precursors. As highlighted in Figure 4, the median number of protein groups quantified from the cancer samples was 3,644 proteins, and a median of 2,882 proteins in the healthy samples and remained constant across the entire sample set within 1 standard deviation as indicated by the colored areas. In total, 4,565 proteins groups were quantified in this sample set across the two sample types.

Data normalization is often applied to correct for small differences in sample starting amounts, or variation in the sample preparation steps and LC-MS analysis. In this study, a small amount of variation was observed across the samples as well as a slight decrease in overall intensity over time (Figure 5, top). This variation was corrected for by normalization in Spectronaut software (Figure 5, bottom).

Finally, reproducibility of chromatography is important for targeted data extraction when using the spectral libraries. Using microflow LC, highly reproducible retention times were observed (Figure 6), with a median variation of 0.4% RSD.

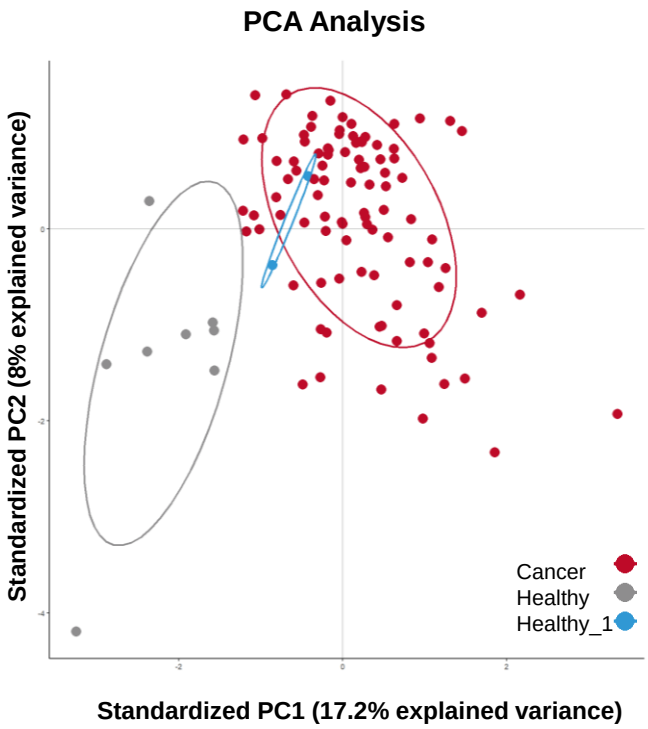

Figure 7. Principal Component Analysis (PCA). Cancer samples are labeled red. Healthy samples are labeled gray. Two samples of the healthy cohort clustered together with the cancer cohort. Therefore, they were analyzed separately (blue dots, labeled Healthy_1). PC1 clearly separates both cohorts.

Many Proteins Are Significantly Altered Between Cancer and Healthy Tissues

Figure 1 shows a volcano plot of the data after a t-test for this quantitative dataset. The results revealed 1,023 proteins in a significantly altered abundance between the healthy and the cancer samples (Q value < 0.01, absolute log2 fold-change > 0.58). The majority of these significantly altered proteins (703) were found in an increased amount in the samples from the cancer tissue.

Using principal component analysis (PCA), healthy and cancer samples were clearly separated (PC1 - Figure 7).

Gene Ontology Information Aids in Biological Interpretation

The PCA findings were supported by gene ontology analysis in Spectronaut in which proteins involved in translation initiation, translation and RNA metabolism were highly enriched in the cancer cohort (Figures 8 and 9). These findings demonstrated an increased protein synthesis capacity in the cancer cells compared to the healthy cells, which has been described as a key physiological task for cancer cells.[4]

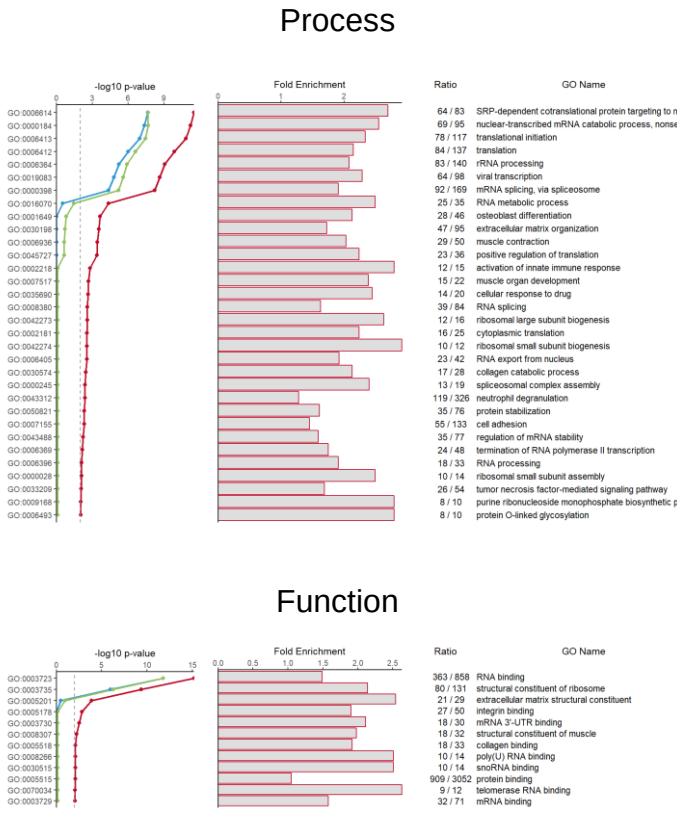

Figure 8. Biological Interpretation of the Data – Cancer vs Healthy. Gene ontology analysis for both Process (top) and Function (bottom) for proteins with an increased abundance in the cancer cohort. Red line indicates the -log10 p-value and the green and blue lines indicate the p-values after multiple testing correction (green: Bonferroni, blue: Benjamini-Hochberg).

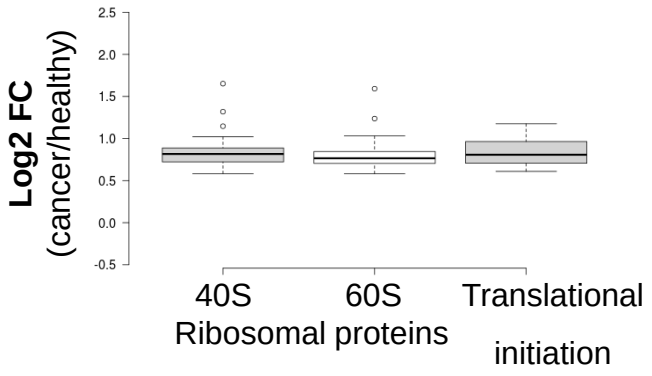

Figure 9. Increased Expression of Proteins Involved in Protein Synthesis. The log2 fold change shows significant up-regulation for 40s and 60s ribosomal proteins and proteins involved in translational initiation in cancer samples. Increased protein synthesis capacity has been described as a key physiological task of cancer cells and thus, proteins involved in translation become highly enriched.[5]

Proteomic Profiles Can Be Used to Identify Potential Cancer Subtypes

A cluster analysis was performed in order to find patterns in the data, and this revealed three cancer subtypes based on their proteome profiles (Figure 10). These potential sub-populations within the cancer cohort are labeled as proteomic subtype A, B and C. Healthy samples are clustered within the unboxed area to the left. Ribosomal proteins are boxed in red at the bottom of the plot and these show various levels of increasing abundance with respect to the healthy samples for all cancer sub-populations.

Another interesting protein cluster is labeled with a blue arrow. Expression levels were high in subtype A and healthy samples, but low in subtype B and C. This cluster primarily consisted of cell adhesion proteins which were previously shown to play a significant role in the metastatic potential of colon cancer.[6]

A previous colon cancer study showed that Hepatocyte Nuclear Factor 4-α (HNF4α) expression levels differ between subtypes.[7] In this study, HNF4α was significantly higher in abundance in subtype B (Figure 11) indicating an HNF4α amplification in this tumor subtype.

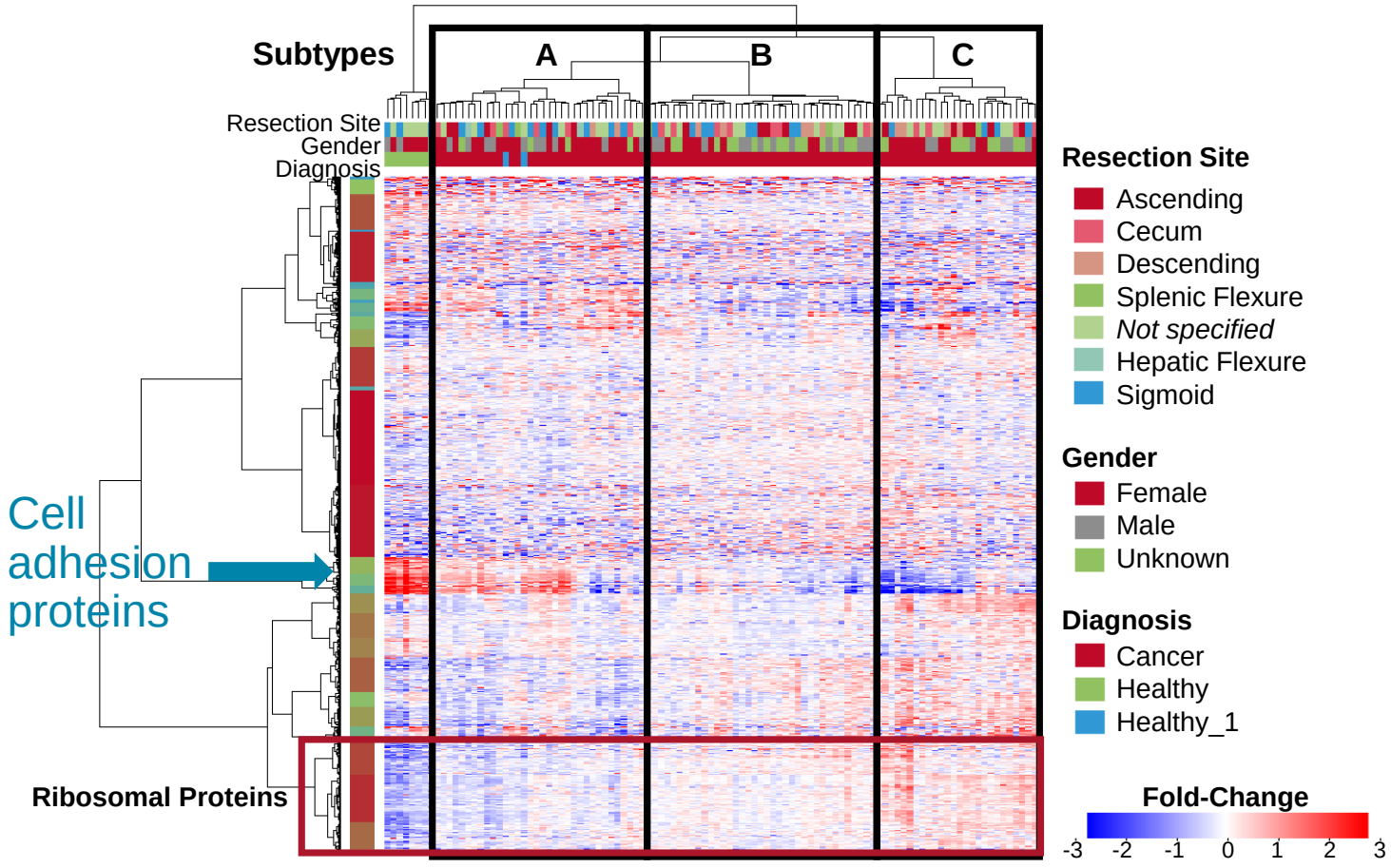

Figure 10. Proteomic Cancer Subtypes Revealed by Cluster Analysis. Unsupervised clustering of protein fold changes against median abundance across all samples was performed and this revealed 3 proteomic subtypes A, B, C shown by black boxes. Clustered at the bottom of the plot are the ribosomal proteins which show increased expression relative to the healthy samples. Another interesting protein cluster is shown with blue arrow and consisted primarily of cell adhesion proteins.

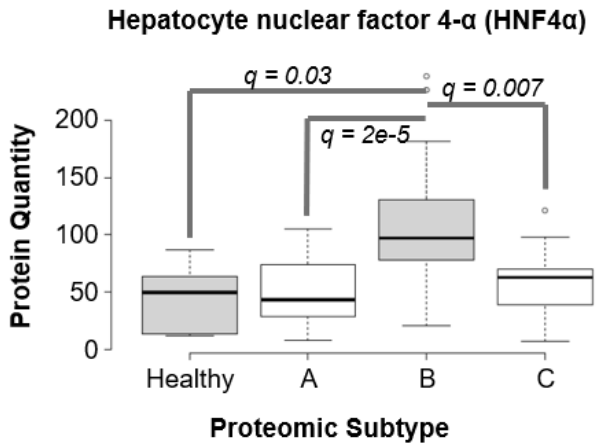

Figure 11. HNF4α Expression Levels. Expression levels of HNF4α for the three proteomic subtypes are shown. Previous work has shown that this protein can differentiate colon cancer subtypes[6].

Conclusions

This study demonstrates how high throughput proteomics can be used to analyze large sample sets from tissue biobanks. Microflow SWATH acquisition on TripleTOF 6600 system combined with Spectronaut data analysis generates data from these large resources in rapid fashion with high analytical depth. The proteomic analysis of large sample sets available in today's biobanks will enable a better understanding of the molecular pathways behind health and disease and pave the way in the future for a better personalized treatment of cancer.

- High throughput analysis with microflow SWATH acquisition, 105 samples analyzed in ~5 days
- High analytical depth - 4,500 proteins across the two sample types, healthy, and colon cancer
- Fast data analysis and results processing with Spectronaut software
- Three cancer subtypes were found in this study based on proteomic profiles

References

1. Collins, B. C.*, Hunter, C.*, Liu, Y.*, *et al*., "Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry", *Nature Communications* (2017), **8**.
2. Microflow SWATH Acquisition for Industrialized Quantitative Proteomics. SCIEX technical note RUO-MKT-02-3637-B.
3. Buczak *et al.,* Spatial Tissue Proteomics Quantifies Inter- and Intratumor Heterogeneity in Hepatocellular Carcinoma (HCC). *Mol Cell Prot,* (2018) **17(4)**: 810–825.
4. Cancer Facts & Figures 2018. American Cancer Society; 2018. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2018/cancer-facts-and-figures-2018.pdf
5. White-Gilbertson *et al.,* The role of protein synthesis in cell cycling and cancer. *Mol Oncol.,* (2009) **3(5-6)**: 402–408.
6. Paschos, *et al.,* The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cell Signal.,* (2009) **21(5),** 665-674
7. Zhang *et al.,* Proteogenomic characterization of human colon and rectal cancer. *Nature*, (2014) **513(7518),** 382-387.

SCIEX

Headquarters
500 Old Connecticut Path | Framingham, MA 01701 USA
Phone 508-383-7700
sciex.com

International Sales
For our office locations please call the division headquarters or refer to our website at
sciex.com/offices