Data Analysis Report

Global Proteomics Analysis of Serum from Healthy Individuals and Colon Cancer Patients

Client:	Panome Bio
Contact:	info@panomebio.com

Contents

Project Summary

Sample Description

Twenty human serum samples were received. Ten of these samples were from colorectal cancer patients (CRC, Stage IV). The remaining ten samples were from non-affected individuals (referred to as healthy). Individuals in both the CRC and healthy groups were between 40 and 80 years of age at the time of sample collection.

Goal

The goal of the present study is to identify circulating protein CRC biomarkers that will be subsequently validated in an expanded sample cohort. In addition, the biological significance of detected protein abundance changes should be interpreted, and peptide-level data should be assessed to elucidate potential proteoform changes in CRC.

Assay Summary

Untargeted bottom-up, label-free proteomics analysis was performed on all samples. Protein isolation and digestion was performed with the Seer Proteograph XT assay. Peptide digests were analyzed with DIA LC/MS/MS utilizing the Orbitrap Astral mass spectrometer.

Analysis Summary

Data were processed with a library-free search and the relative protein and peptide abundance was quantified. Unsupervised and supervised statistical analyses were performed to assess group separation and protein and peptide-level differences between healthy individuals and CRC patients. Results were aggregated through pathway analysis.

Results Summary

Global proteomics analysis identified over >5k proteins and 50k peptides that provided separation of CRC and healthy groups. Statistical analysis revealed 25 proteins and 47 peptides that were differentially expressed in CRC. Global interpretation of the proteomics data revealed three primary biochemical pathways that are dysregulated in CRC: 1) upregulation of glycine, serine, and folate metabolism, 2) Downregulation of beta oxidation and the electron transport chain, and 3) upregulation of tyrosine and catecholamine catabolism. To better contextualize these results, we recommend the following: 1) untargeted metabolomics profiling of serum samples that includes early-stage and inflammatory bowel disease (IBD) patients, 2) expanding the sample cohort for proteomics to validate potential biomarkers and control for IBD effects, and 3) profiling of primary tumor tissue and healthy tissue through integration of publicly available transcriptomics data.

Results

Analyte Summary

In the comprehensive proteomic profiling of the samples, two protein fractions for each sample were created through the Proteograph XT assay (see details in the Experimental Methods). These fractions were profiled separately with LC/MS/MS for each sample. In total, **68,238 peptides** corresponding to **7,027 protein groups** were profiled across the samples. In data pre-processing, the abundance of the detected peptides were mapped to their corresponding protein groups. A protein group consists of 1+ protein(s), which are generally isoforms whose peptides cannot be distinguished with LC/MS/MS alone. The abundance of the measured peptides for a protein are normalized and aggregated into a single abundance measurement for the protein group, see **Experimental Methods**. The breakdown of the peptide and protein group detections are summarized in Figure 1 and Figure 2, respectively. Proteins and peptides that had a missing value rate greater than 75% were excluded from downstream analysis. For this study, this resulted in **50,079 peptides and 5,724 proteins that were available for statistical analysis**.

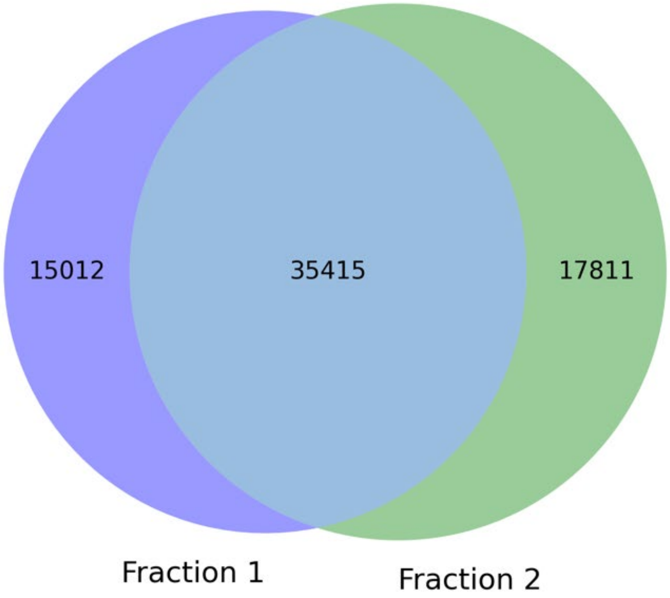

Figure 1: Peptide detection breakdown. The peptides detected in each fraction (Fraction 1 and Fraction 2) are shown in the Venn diagram. These values are for peptides detected in 1+ of the research samples.

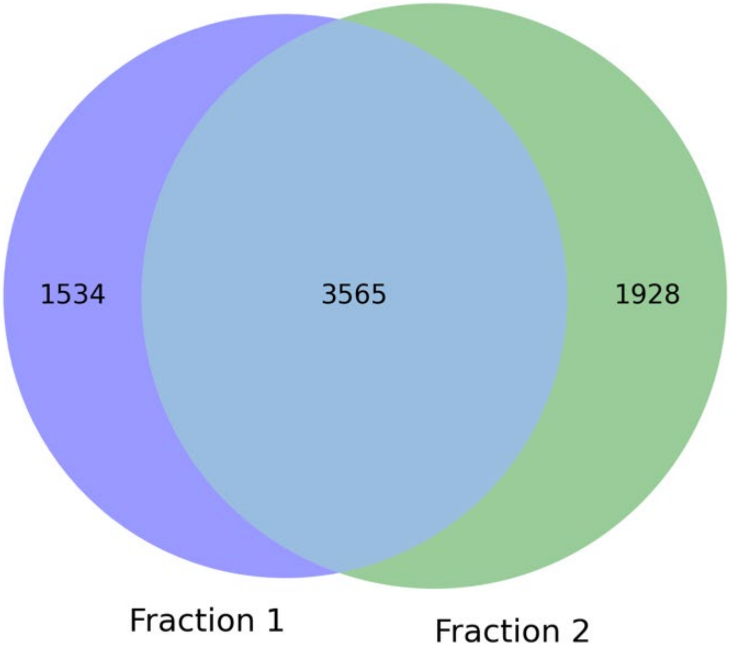

Figure 2: Protein detection breakdown. The proteins detected in each fraction (Fraction 1 and Fraction 2) are shown in the Venn diagram. These values are for proteins detected in 1+ of the research samples.

Protein Associations

In the following sections, the trends in individual peptide measurements and aggregated protein-group abundance values are assessed with both unsupervised (agnostic of sample groups) and supervised (discriminating between sample groups).

Unsupervised Analyses

The global trends in the proteomics data are assessed through unsupervised principal component analysis (PCA) and hierarchical clustering analysis (HCA, visualized through a heatmap), see **Experimental Methods**. These results are presented at the peptide level in Figures 3 and 4 and the protein level in Figures 5 and 6. When considering both peptide-level and protein-level abundance values, we see clustering of samples that correspond to the experimental groups (healthy and CRC). These results demonstrate the dysregulation of the human proteome that occurs in CRC. In the subsequent sections, we will perform supervised statistical analyses to identify proteins biomarkers of CRC and interpret the biological significance of these proteins through pathway analysis.

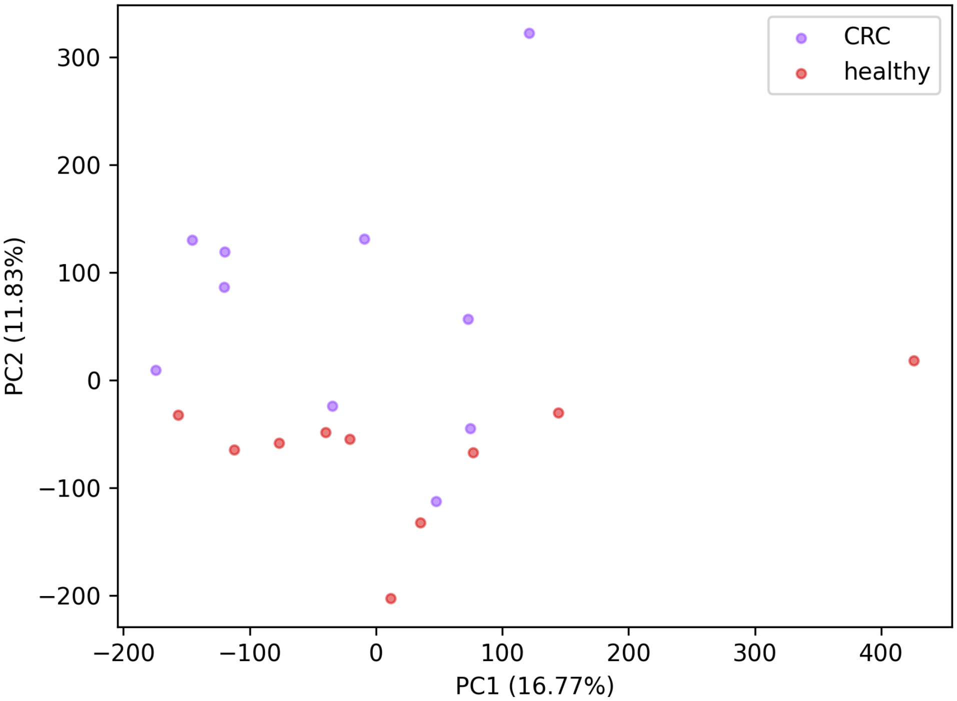

Figure 3: Principal components analysis (PCA) of peptide-level data. The peptide profiles for each sample are visualized in the scatter plot above. Samples are color-coded based on their respective groups. The specific colors for each group are provided in the legend.

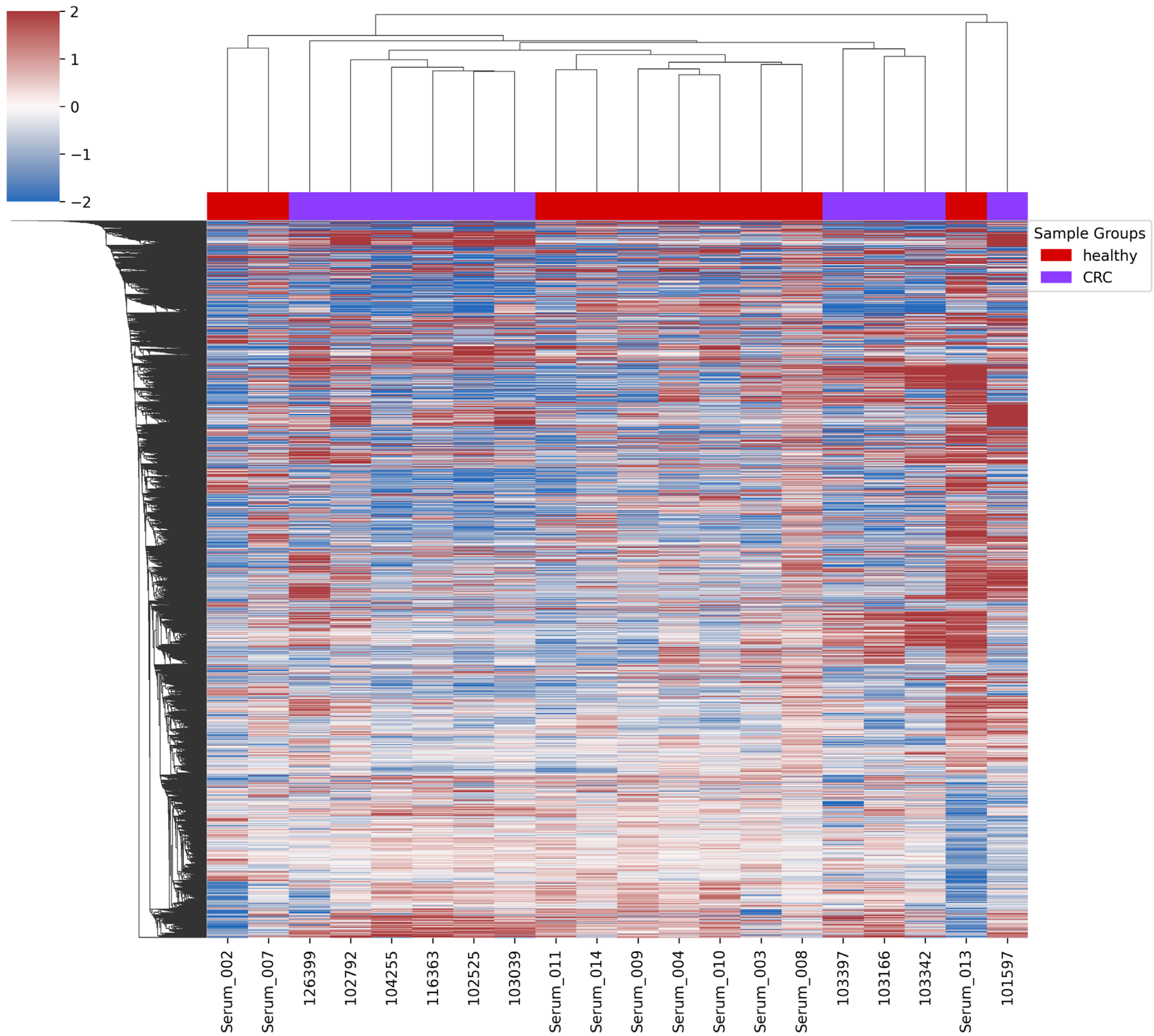

Figure 4: Heatmap and hierarchical clustering analysis of peptide-level data. The peptide profiles for each sample are visualized in the heatmap above and clustered. Each column represents a sample, and each row represents a peptide. Columns are colored according to experimental type. The color of each cell indicates the log2(fc) relative to the mean level of each peptide in the healthy group.

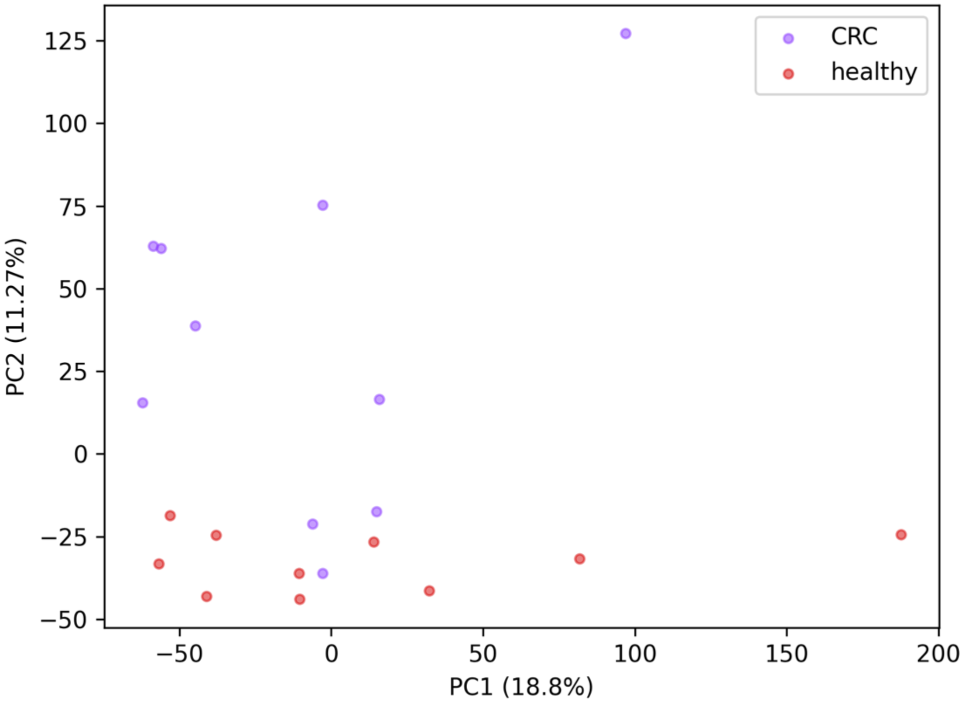

Figure 5: Principal components analysis (PCA) of protein-level data. The protein profiles for each sample are visualized in the scatter plot above. Samples are color-coded based on their respective groups. The specific colors for each group are provided in the legend.

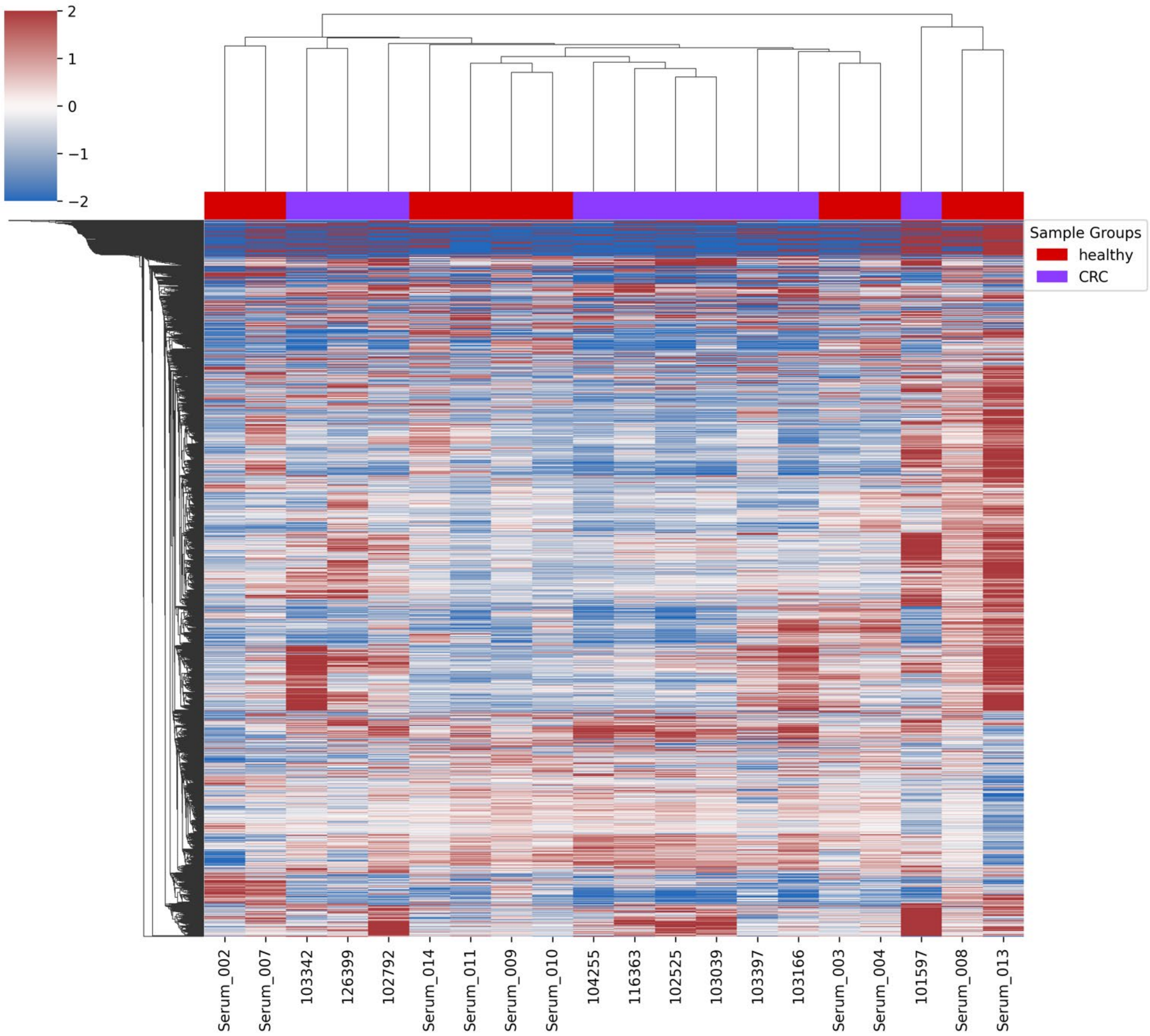

Figure 6: Heatmap and hierarchical clustering analysis of protein-level data. The protein profiles for each sample are visualized in the heatmap above and clustered. Each column represents a sample, and each row represents a protein. Columns are colored according to experimental type. The color of each cell indicates the log2(fc) relative to the mean level of each protein in the healthy group.

Differential Expression Analysis

To determine specific differences in protein and peptide levels between the healthy and CRC samples, a one-way ANOVA was conducted on all profiled analytes. This analysis identified **47 peptides** and **25 proteins** that were differentially expressed ($q < 0.05$). The associated volcano plots for the peptide and protein measurements are shown below in Figures 7 and 8, respectively.

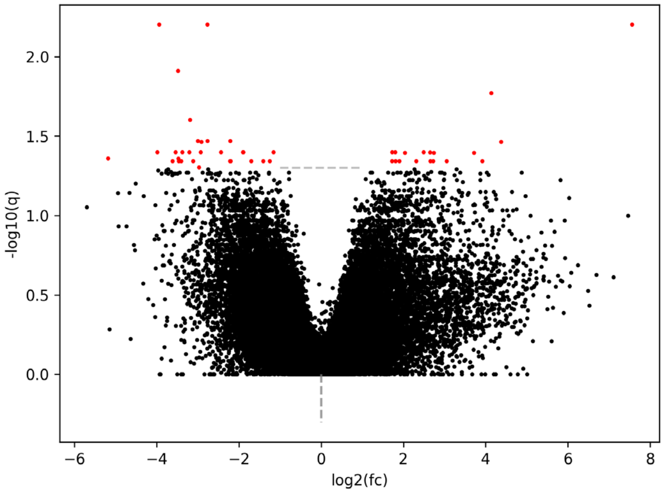

Figure 7: Volcano plot of peptides. The q-values (q) and $\log_2$ fold-changes of the peptide levels between healthy and CRC samples are plotted against each other in the plot above. Red dots indicate peptides that pass the $q < 0.05$ cutoff.

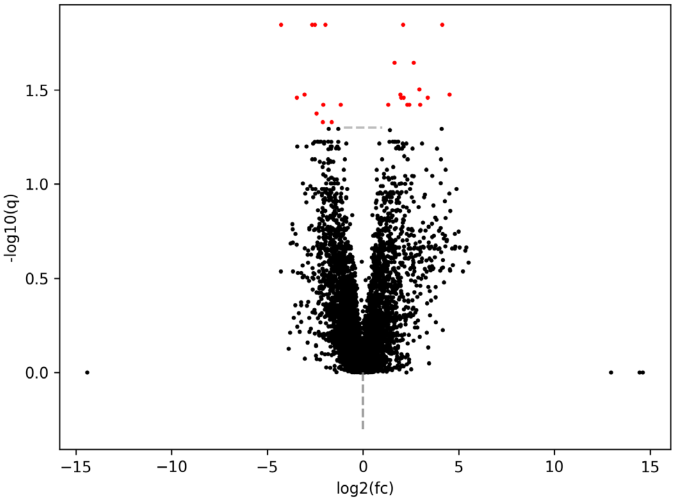

Figure 8: Volcano plot of proteins. The q-values (q) and $\log_2$ fold-changes of the protein levels between healthy and CRC samples are plotted against each other in the plot above. Red dots indicate proteins that pass the $q < 0.05$ cutoff.

The statistically significant peptides (Figure 9) and proteins (Figure 10) are shown in the heatmaps below. Notably, multiple peptides were able to reach statistical significance, whose aggregating protein abundance did not. Peptide measurements are often more sensitive and precise (e.g., measuring specific proteoform) when compared to the summarized protein abundance, which leads peptide-level measurements to be more accurate biomarkers in some cases.

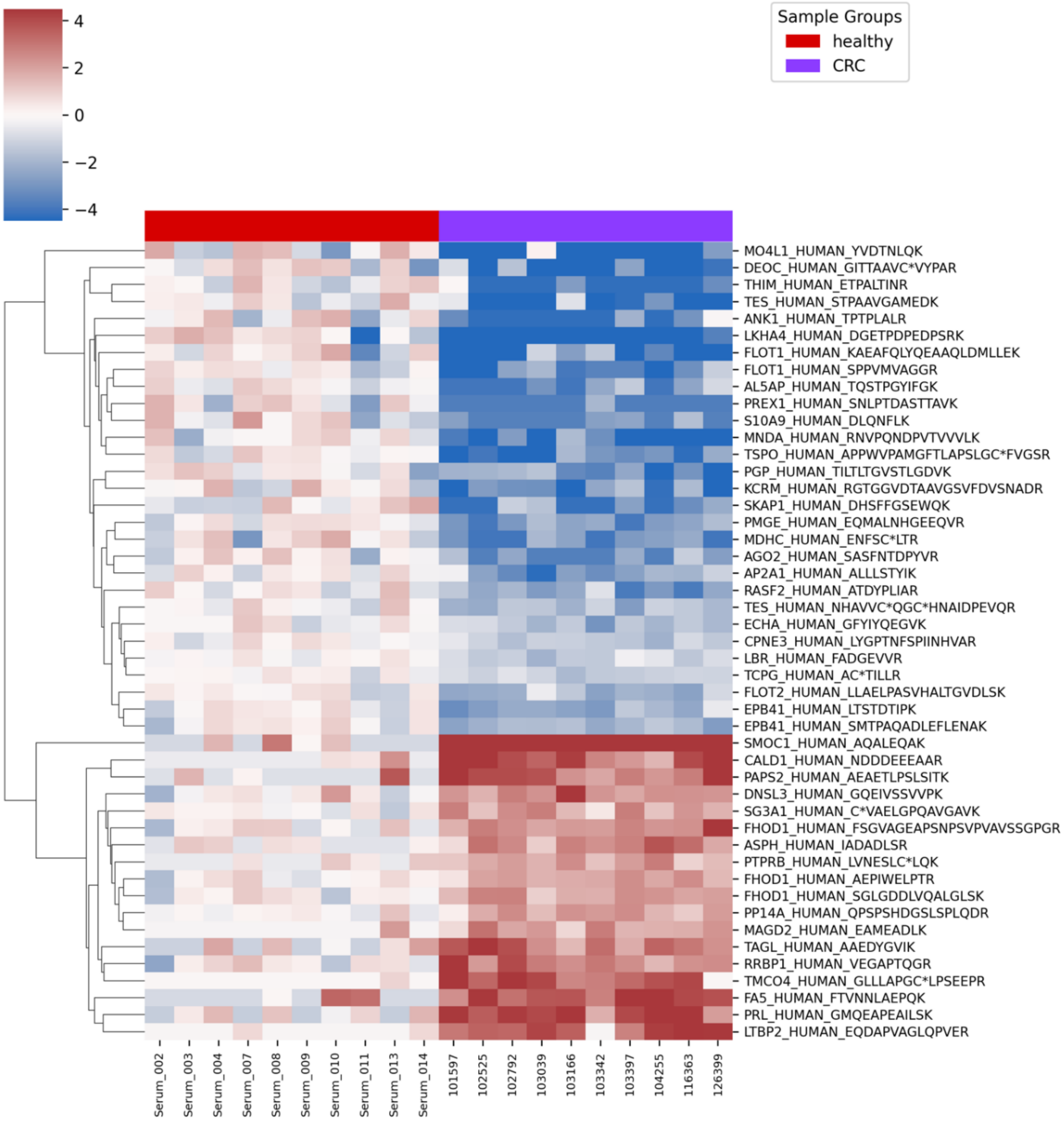

Figure 9: Heatmap of significantly altered peptides. The normalized abundance of significantly altered peptides are plotted for each experimental group. The color of cells indicates the log2 fold change relative to the healthy group. The peptides are labeled by their protein group concatenated with their peptide modified sequence. * indicates carbamidomethylation modification on cysteine residue.

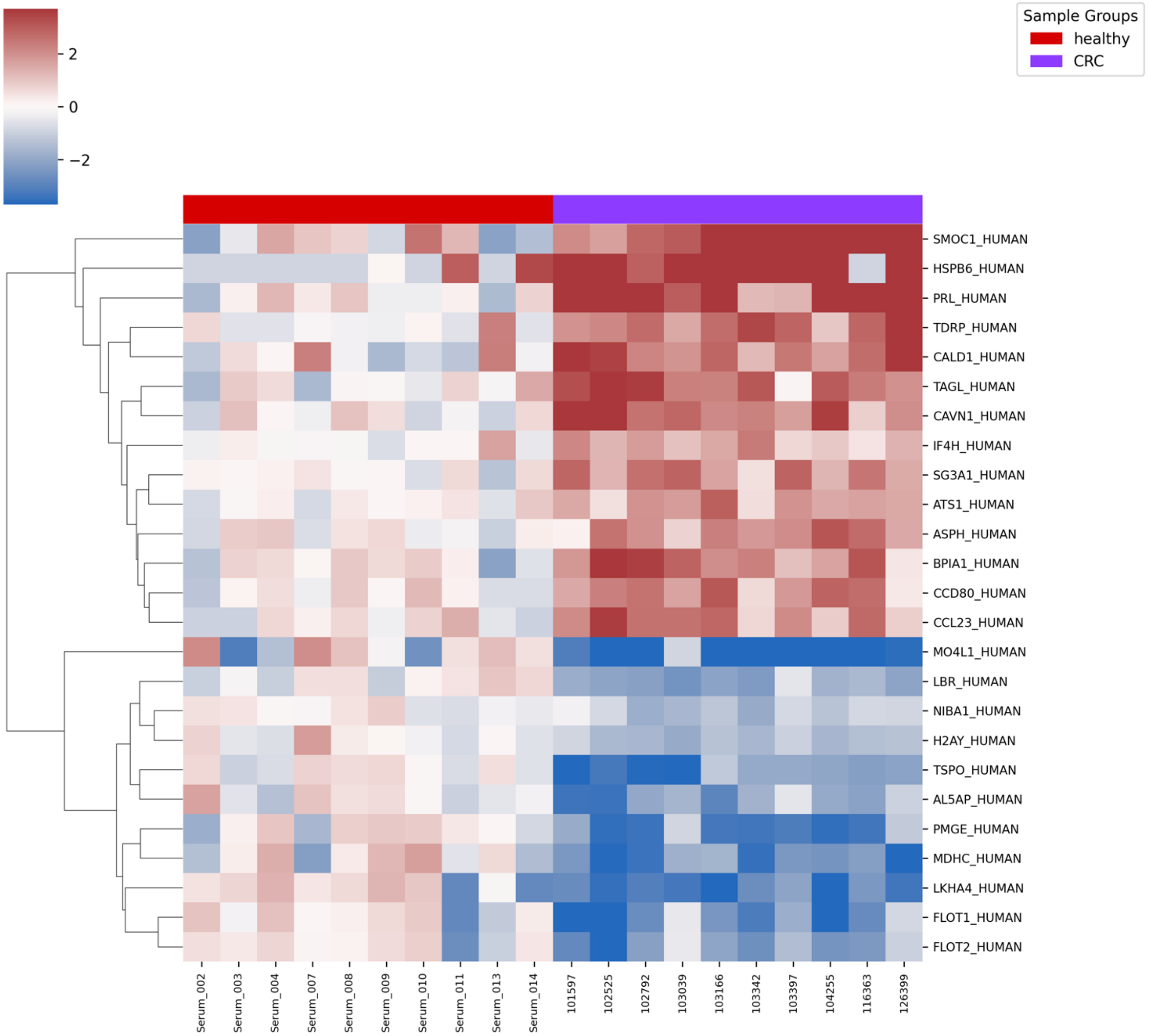

Figure 10: Heatmap of significantly altered proteins. The normalized abundance of significantly altered proteins are plotted for each experimental group. The color of cells indicates the log2 fold change relative to the healthy group.

Interpretation

Pathway Analysis

To gain biological insights from the dysregulated proteins, an over-representation analysis was conducted. This identified key biochemical pathways (signaling, metabolic, protein, physiological, and disease pathways) that were altered between the healthy and CRC samples (Figure 11). This analysis identified **28 pathways that were enriched** for proteins showing statistically significant changes in CRC. The diversity of these pathways underscores the magnitude of molecular dysregulation in CRC as well as breadth of proteome coverage achieved.

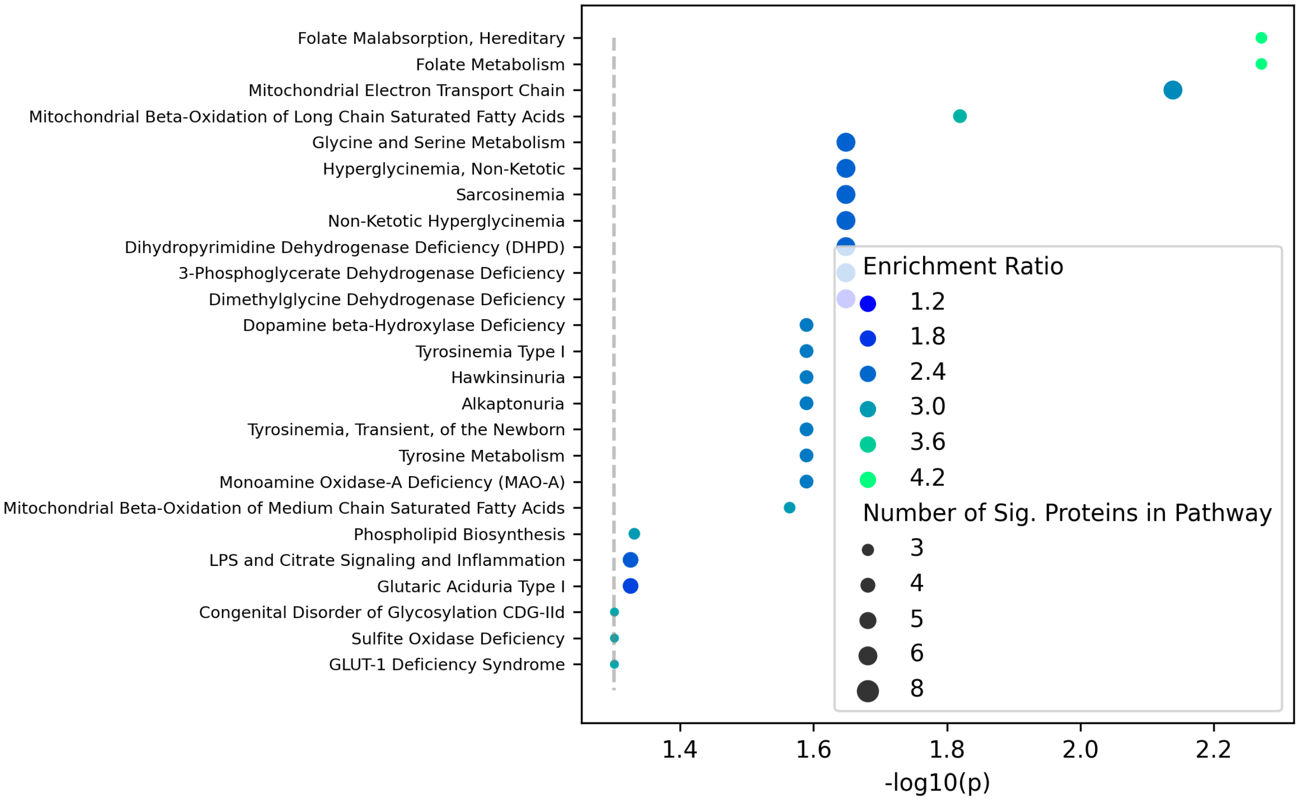

Figure 11: Pathway analysis of significant proteins. The dot plot displays the enriched pathways identified in the analysis, with their corresponding significance levels represented on the x-axis. The significance threshold (p = 0.05) is denoted by the grey dashed line. The size of the dots reflects the number of proteins that showed statistical differences between the sample groups. Additionally, the color of the dots represents the enrichment ratio, which reflect the extent to which the observed number of statistically significant proteins in a pathway deviate from what would be expected by random chance.

The two most enriched pathways relate to folate metabolism and are driven by the upregulation of the same four proteins (FTCD, MTHFS, AL1L1, C1TC). A heatmap of these proteins are shown in Figure 12. These upregulated proteins are all involved with the conversion of tetrahydrofolic acid species. These interconversions provide one-carbon units for purine and pyrimidine synthesis as well as change the redox balance through the production/consumption of NADPH. The production of NAPDH is critical for oxidative stress and upregulation of one-carbon metabolism has been shown to provide protection against chemotherapeutics (PMID: 36973440).

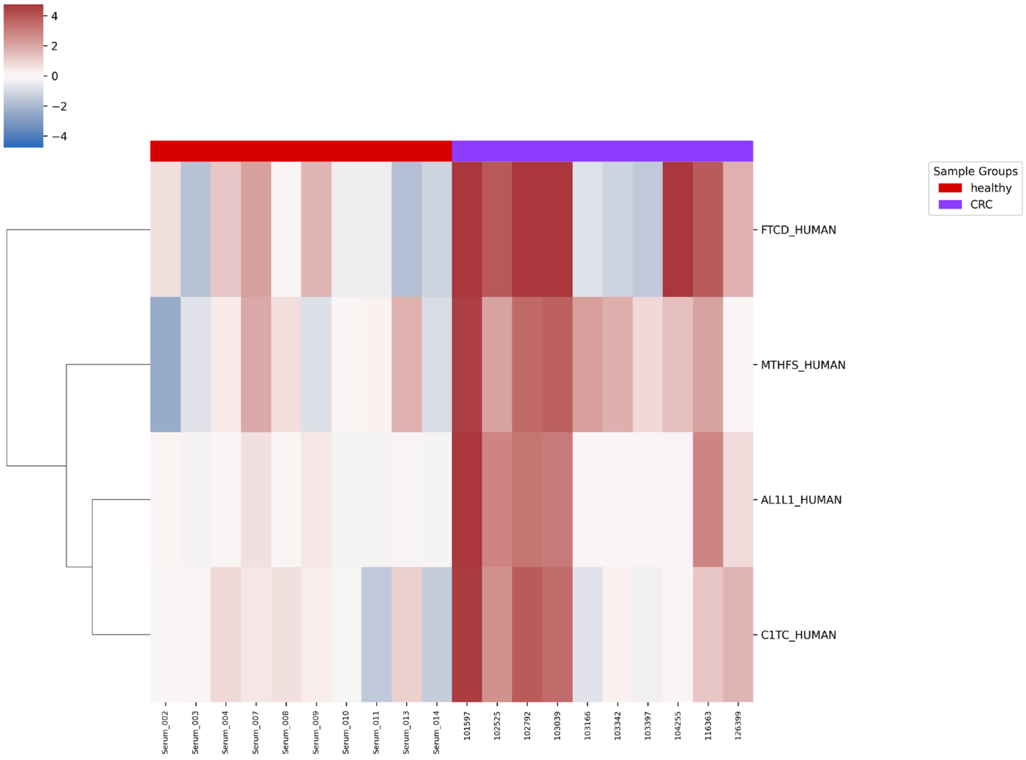

Figure 12: Heatmap of folate metabolism proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

In addition, seven of the enriched pathways relate to serine and glycine metabolism and are all driven by the enrichment of the same 8 proteins (GNMT, DLDH, SERA, GLCTK, SDHL, SARDH, GAMT, and GLYC) that are upregulated in CRC. These proteins are shown in Figure 13. Upregulation of these proteins is consistent with several diseases, including hyperglycinemia, sarcosinemia, and 3-phophoglycerate dehydrogenase deficiency. Upregulation of glycerate kinase (GLCTK) and D-3-phosphoglycerate dehydrogenase (SERA) suggests increased synthesis of serine from glycerate. Serine can then be catabolized to produce the essential amino acid cysteine and glycine through the homocysteine cycle or used for nucleotide synthesis through one-carbon metabolism. The high demand for amino acids and nucleotides to support cancer cell proliferation is consistent with the upregulation of this pathway and is further supported by the changes in folate metabolism discussed above. One-carbon metabolism is also connected to epigenetic modifications such as DNA methylation.

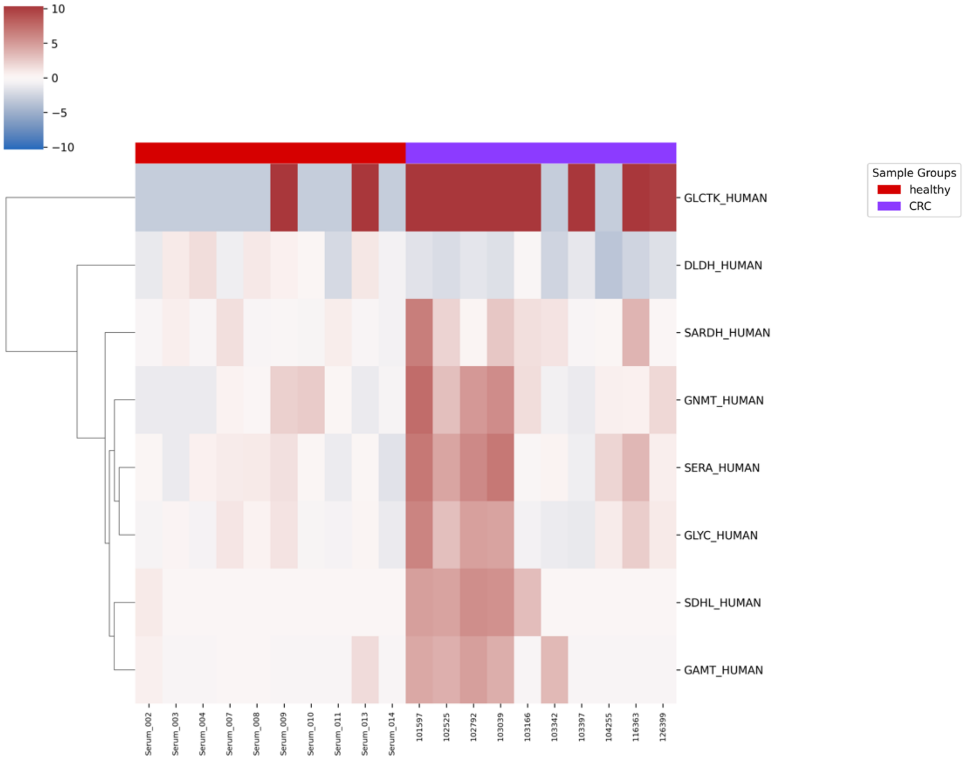

Figure 13: Heatmap of serine and glycine-associated proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

Two of the most enriched pathways both relate to mitochondrial metabolism: the mitochondrial electron transport chain (Figure 14) and mitochondrial beta-oxidation of long-chain saturated fatty acids (Figure 15). In both cases, we see a downregulation of proteins in this pathway. These pathways are naturally connected as beta oxidation produces acetyl-CoA units that can be metabolized to produce reducing equivalents used to generate ATP in the electron transport chain. Downregulation of both pathways suggests that, systemically, beta-oxidation is reduced, causing a corresponding downregulation of the electron transport chain. This is further supported by enrichment in beta-oxidation of medium-chain fatty acids as well (Figure 16).

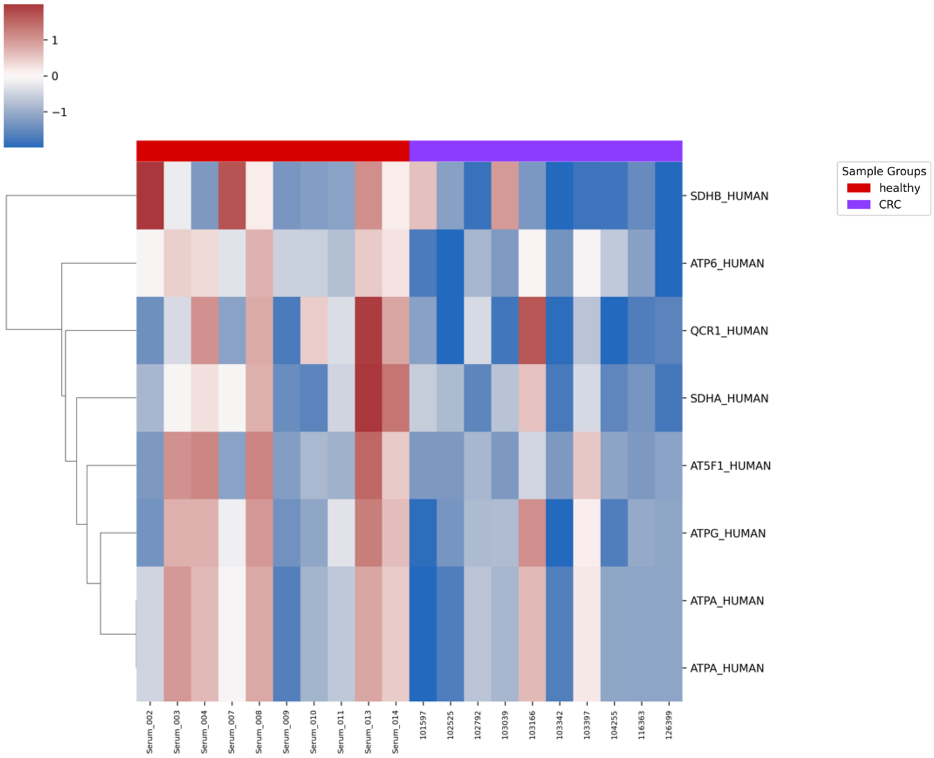

Figure 14: Heatmap of mitochondrial electron transport chain proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

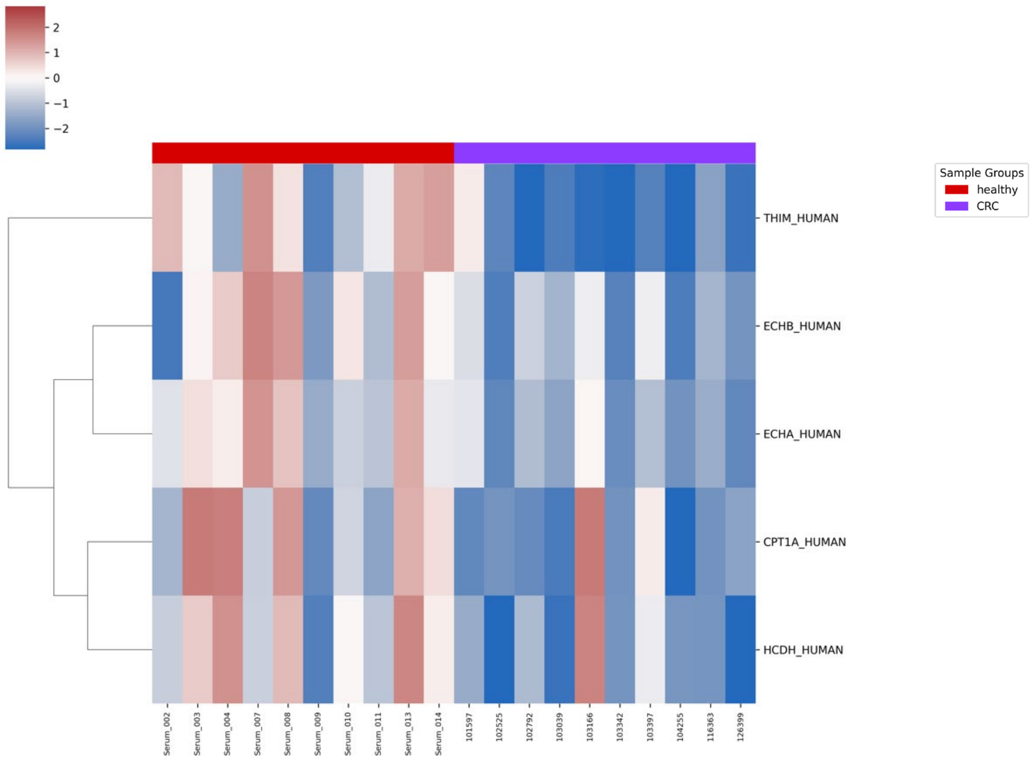

Figure 15: Heatmap of mitochondrial beta-oxidation of long-chain saturated fatty acid associated proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

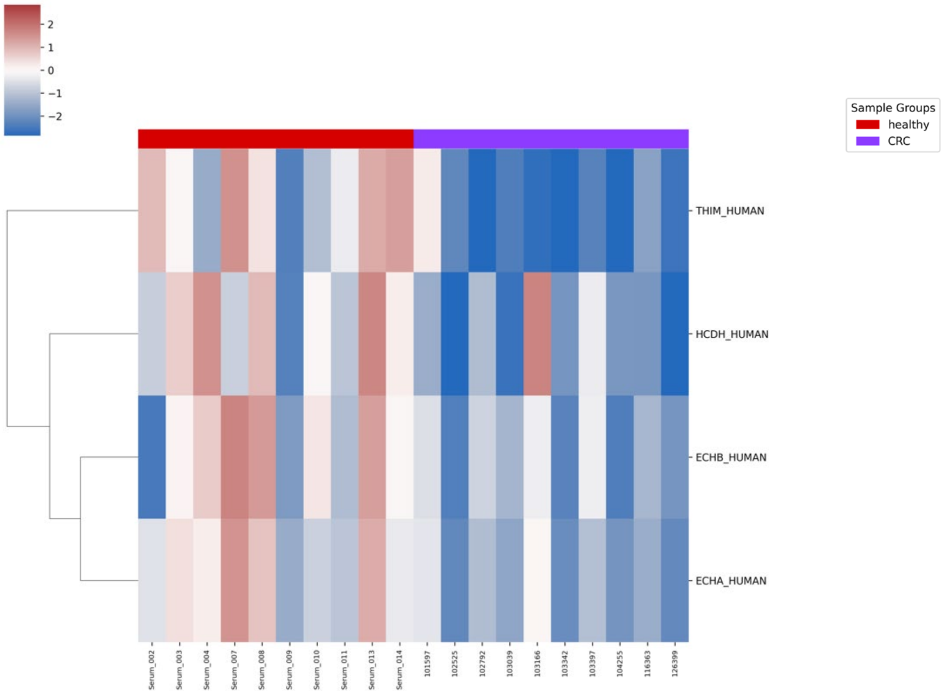

Figure 16: Heatmap of mitochondrial beta-oxidation of medium-chain saturated fatty acid associated proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

Seven pathways that were enriched largely relate to tyrosine and monoamine metabolism and were all driven by changes in five proteins (MIF, COMT, MAAI, ATTY, HPPD) that are upregulated in CRC. These proteins are shown in the heatmap provided in Figure 17. These changes are associated with multiple diseases, including dopamine beta hydroxylase deficiency, monoamine oxidase-A deficiency, and tyrosinemia. The most upregulated proteins are tyrosine aminotransferase (ATTY) and 4-hydroxyphenylpyruvate dioxygenase (HPPD), which have a >1000-fold abundance in CRC relative to healthy. HPPD is a key enzyme in regulating tyrosine catabolism and has been implicated in many diseases, including colon and breast cancer. For colon cancer in particular, it has been suggested as a prognostic biomarker, and small molecular inhibitors of HPPD are currently in development (PMID: 37794595). ATTY, HPPD, and MAAI (maleylacetoacetate isomerase) are part of linear pathway that converts tyrosine to fumarate and acetoacetate. Catechol O-methyltransferase (COMT) was also upregulated in CRC and is one of the enzymes that degrades catecholamines (dopamine, norepinephrine, etc.). COMT has been shown to be upregulated in primary colon cancer tumor tissue (PMID: 20646666). Together, these results suggest increased catabolism of tyrosine and catecholamines potentially to fuel tumor metabolism.

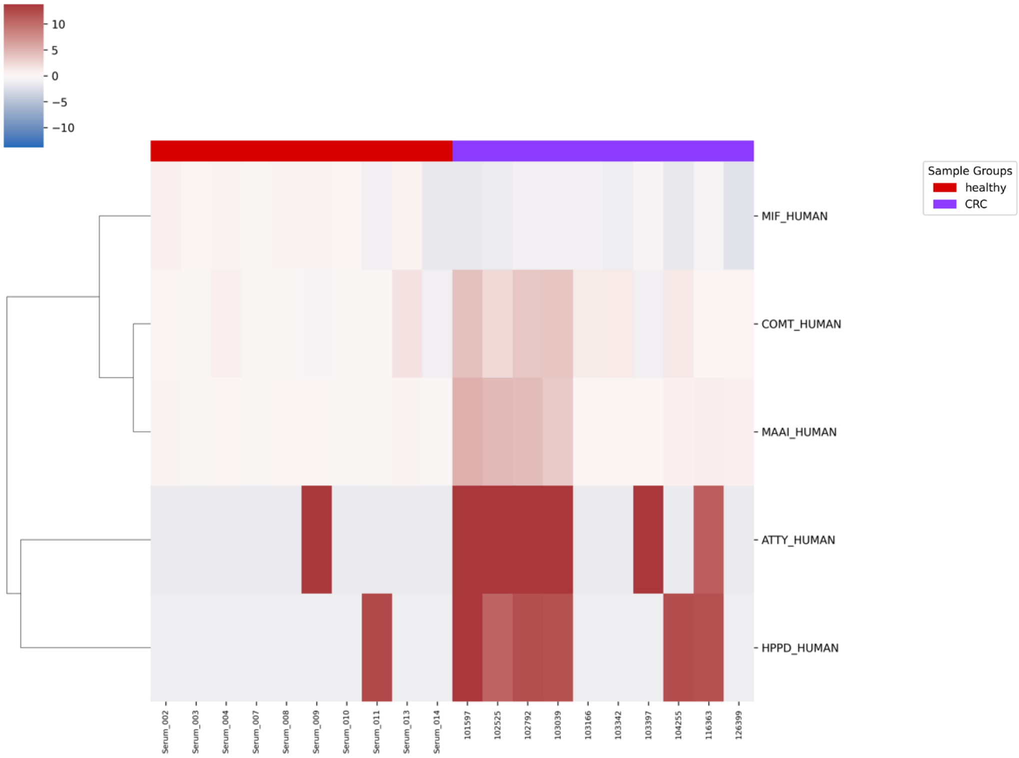

Figure 17: Heatmap of tyrosine and monoamine metabolism proteins across healthy and diseased samples. Each row represents a protein within the pathway, while columns correspond to the experimental conditions.

Heatmaps for all 314 biochemical pathways assessed as well as the detailed results of the statistical analysis is provided in the Supplementary Information.

Candidate Biomarkers

Twenty-five proteins reached statistical significance between CRC and healthy groups. To compare the ability of these proteins to stratify individuals by disease status, we computed receiver-operating characteristic (ROC) curves for the top 10 most statistically significant proteins detected (Figure 18). These curves plot the false positive rate (FPR) vs the true positive rate (TPR) of predicting CRC status when applying different abundance cutoffs for a protein. The FPR is fraction of healthy individuals that were incorrectly classified as CRC. The TPR is the fraction of CRC individuals that were correctly classified as CRC. The ability of a protein to separate sample groups can be quantified by computing the area under the ROC curve (AUROC), which ranges from 0.0-1.0. Perfect separation of sample groups is achieved with an AUROC of 1.0, where, at FPR of 0.0, a TPR of 1.0 is achieved. Random performance has a AUROC of 0.5.

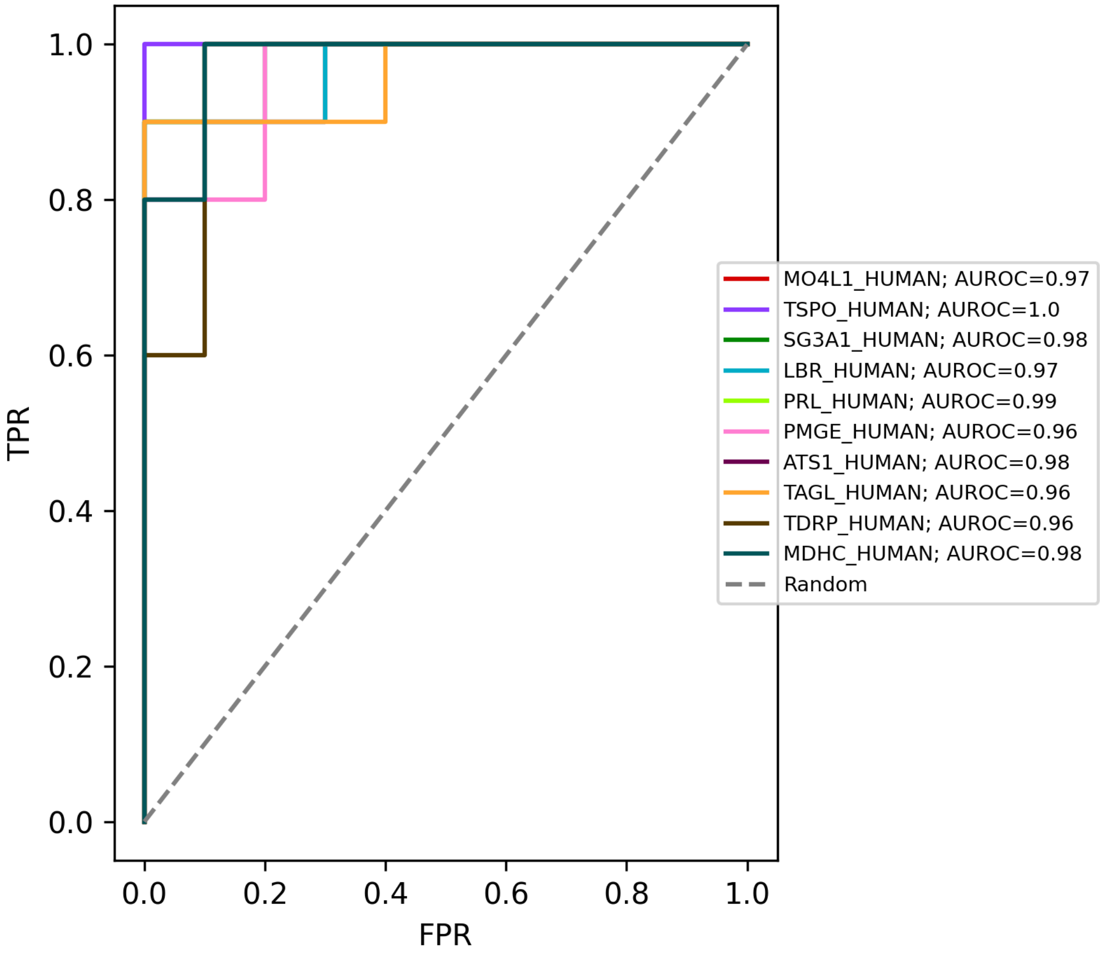

Figure 18: ROC curves for top 10 most statistically significant proteins. ROC curves are shown in the plot above for the top 10 most statistically significant proteins. Random performance is shown in the grey dashed line. Proteins are denoted with different colors. The AUROC is provided for each protein in the legend. The greater the AUROC, the better separation is achieved between sample groups.

When examining the ROC curves for the top 10 most statistically significant proteins, all are able to separate sample groups with an AUROC > 0.95, demonstrating the ability of untargeted proteomics to discover circulating biomarkers. The most stratifying protein was translocator protein (TSPO), which achieved perfect separation of sample groups. The levels of TSPO across the sample groups are shown in Figure 19 and clearly show the strong downregulated of TSPO in CRC. TSPO is the mitochondrial translocator protein that has high binding affinity for many small-molecule drugs and cholesterol. TSPO is expressed in the colon and has been showed to be over-expressed in inflammatory bowel disease (IBD) as well as colon cancer cells (PMID: 20222126). Given TSPO biological function is to regulate molecular transport to the mitochondria and mitochondrial metabolism is likely perturbed in CRC from the pathway analysis results, TSPO may be a potential therapeutic target for CRC.

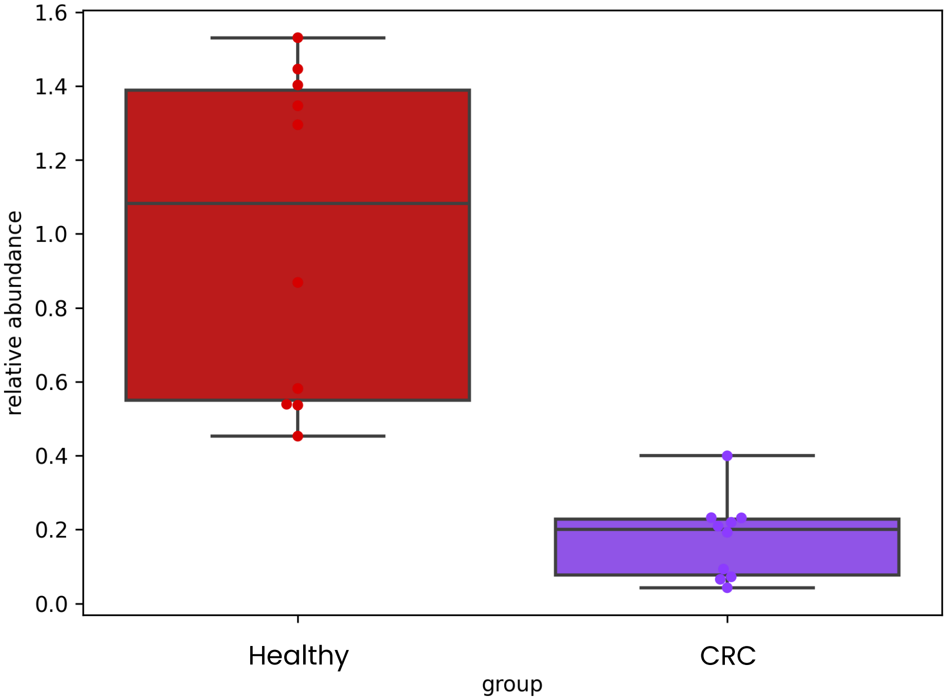

Figure 19: TSPO protein levels. The boxplot above shows the abundance of TSPO, the protein achieving the highest AUROC, across the healthy and diseased sample groups.

.

Conclusions

Global proteomics analysis of human serum from healthy individuals and CRC patients identified over >5k proteins and 50k peptides that were present in >25% of samples. Unsupervised analysis of the protein and peptide-level data revealed dramatic proteome dysregulation and separation of CRC and healthy groups in PCA and HCA analysis. Statistical analysis revealed 25 proteins and 47 peptides that showed a statistically significant change in abundance in CRC relative to healthy.

One of the statistically significant proteins, TSPO, perfectly stratified the healthy and CRC populations. TSPO is known to be upregulated in colon cancer and IBD. Given its role as mitochondrial transporter, it may have a functional role in the changes to lipid metabolism occurring inside mitochondria that were identified.

Global interpretation of the dysregulated proteins through pathway analysis offered additional biological insights into biochemical dysregulation in CRC. These findings are summarized below:

- Upregulation of glycine, serine, and folate metabolism in the cytosol. These alterations were consistent with increased one-carbon metabolism used to fuel nucleotide synthesis and substrates for epigenic modifications. Upregulation of these pathways is also consistent with increased NADPH production to buffer oxidative stress, such as that induced with chemotherapeutics.
- Downregulation of the electron transport chain (ETC) and beta oxidation. Proteins involved with ATP production via the ETC are downregulated in CRC compared to healthy individuals. Additionally, enzymes involved with the catabolism of medium and long-chain fatty acids through beta oxidation were downregulated. Beta oxidation produces acetyl units that can be used to generate reducing equivalents that fuel the ETC. These results suggest that in CRC, energy production from fats is reduced systemically.
- Upregulation of tyrosine and mono-amino metabolism. Multiple up-regulated proteins relate to tyrosine and catecholamine catabolism pathways. These results suggest increased catabolism to support tumor metabolism. One of the most upregulated proteins, HPPD, has been implicated in CRC. ATTY, HPPD, and MAAI are part of linear pathway that converts tyrosine to fumarate and acetoacetate, which provide substrates to produce ATP via the TCA cycle.

Next Steps

To better contextualize these results, we recommend the following experiments and analyses:

- Perform untargeted metabolomics profiling of serum samples to provide a greater mechanistic link between the dysregulated metabolic catabolism and synthesis pathways with altered protein abundance. This would provide considerable clarity into the biological consequences of one-carbon and tyrosine metabolism upregulation as well as the nutrient landscape that may influence the fatty acid oxidation downregulation.

- Profiling (both metabolomic and proteomic) of serum from patients with IBD, but not CRC, would help stratify the protein changes from IBD with CRC.

- Profiling of primary tumor tissue and healthy tissue to determine which proteome alterations are specific to the tumor as opposed to systemic changes. Public transcriptomics data would provide a solid starting point to determine the relevance of these findings to tumor biochemistry.

- Expand the sample cohort to validate candidate biomarkers and build multivariate models of CRC. By using this study to perform a power analysis, we recommend a n=100 per sample group to detect a 50% change in protein abundance with a power of 85%. In such a follow-up study, including early-stage CRC patients would provide insight into disease progression and enable identification of early disease biomarkers to improve patient outcomes.

- Proteoform analysis of discordant peptides to identify peptide variants and proteoform (phosphorylation, ubiquitination, etc.) changes between CRC and healthy serum.

Experimental Methods

Sample Preparation

Sample handling and storage

Samples were frozen at -80°C after receipt.

Bottom-up proteomics sample preparation

Serum samples are thawed and prepared for bottom-up LC/MS/MS-based proteomics analysis with the SP100 automation instrument. 240 µL of serum samples are loaded on the instrument each of which is then aliquoted into two wells (100 µL each), and each tube is incubated with a unique nanoparticle (NP) suspension. After incubation with the NPs and subsequent washing steps, to remove non-specific and weakly bound proteins, bound proteins were then reduced, alkylated, and digested with trypsin protease to generate tryptic peptides for downstream LC/MS analysis. The peptides are then desalted**,** and all detergents are removed using a mixed media filter plate and a positive pressure (MPE) system. The peptides are eluted, and peptide quantitation is performed on the SP100 Automation instrument using the Piece Fluorescent Assay Kit. Peptides are then dried and stored at -80 °C until LC/MS analysis.

Data acquisition and pre-processing

LC/MS/MS analysis of digested peptides

LC/MS/MS mobile phases A and B were prepared as follows:

A) 0.1% FA, 100% water

B) 0.1% FA, 100 % ACN

7 µL of the reconstituted sample was injected onto a Vanquish Neo HPLC System with a 50 cm mPAC HPLC column and analyzed by Reversed Phase (RP) LC-MS/MS by using the following gradient at a flow rate of 1 µL/min: 0-1.5 min: 5% B, 1.5-23.5 min: 25% B, 23.5-29.0 min: 40-90% B, 29-29.75 min: 1% B. MS data acquisition was performed with a Thermo Fisher Orbitrap Astral mass spectrometer operating in data-independent acquisition (DIA) mode with 3 m/z isolation windows ranging from 380-980 m/z.

Data preprocessing

Spectral Search and Peptide Identification

LC/MS/MS data acquired on both protein fractions for each sample were first processed with the DIANN (v1.81.) software with a library-free spectral search. The software performs an in-silico protein digestion based on the input proteome (human proteins and common contaminant sequences downloaded from UniProt). MS/MS spectra for each peptide predicted from the in-silico digestion are then computed and used to identify peptides measured in the acquired LC/MS/MS data and assign them to defined protein groups. Peptide identifications are then scored and filtered with a neural network-based approach. A protein-group, peptide q-value, and library q-value of 0.01 was applied. The raw peptide abundance is calculated using modeled chromatographic peaks. Further details are provided in the DIANN publication. DIANN parameters are included in the Supplementary Data.

Peptide and Protein Quantification

After peptide identification, peptide abundance values are normalized and summed across fractions for each sample through the application of delayed normalization (as implemented within the maxLFQ software). In brief, a scaling factor is determined for each LC/MS/MS run that minimizes the variance across all detected peptides for all samples. This is performed to avoid systematic errors that are derived from pipetting, sample collection, or other pre-analytical variation. After performing delayed normalization, the normalized peptide values are summed across fractions to yield a single peptide abundance value for each peptide for each sample. To remove batch effects, ComBat normalization is applied to address batch effects derived from the Proteograph assay and LC/MS/MS runs. Lastly, missing peptide abundance values are imputed for peptides measured in >25% of samples by estimating missing peptide values as ½ of the minimum detected intensity for that peptide across all research samples.

After peptide abundance calculation, the protein-group abundance is inferred through the maxLFQ algorithm. In brief, a peptide ratio matrix is computed for each protein by computing the median ratio of peptide abundance for each pair of samples when considering the peptides measured in both samples. This matrix then represents an over-determined systems of linear equations which are then solved to yield a single protein abundance value for each sample. These values are then scaled such that the mean protein abundance is equal to the mean summed peptide abundance for each sample.

Prior to downstream analysis, analytes missing in >25% of samples are removed. Data is then log2 transformed.

Statistical Analysis

Hypothesis testing was executed utilizing a one-way ANOVA, which does not assume equal variances among the groups. Log2 transformed analyte (protein/peptide) intensities were used for null hypothesis testing. Fold-changes were computed from non-log2 transformed intensities.

Hierarchical clustering analyses was performed based on log2(fc) transformed values and Euclidian distance. Clusters were calculated with the UPGMA algorithm.

Over-representation analysis was performed with a Fisher’s Exact test that compares the expected number of analytes found to be statistically significant in each pathway with the number of significant analytes found in each pathway. This analysis was performed with the PathBank database and includes metabolic, protein, signaling, disease, and physiological pathways.

Appendix

Supplementary Files

1. Raw_data.zip: raw LC/MS/MS data (.raw) format
2. SupplementaryTables.xlxs: excel file containing sample metadata, technical variation metrics for each peptide and protein, all detected peptide and protein features and their raw intensities and normalized intensities, the detailed results of the statistical analyses.
3. Plots.zip: high-resolution images for all figures in this report as well as boxplots showing the intensity of each protein and peptide across all experimental conditions.

Glossary

1. LC/MS/MS: Liquid chromatography coupled to tandem mass spectrometry. The analytical technique used for the proteomics assays.
2. Peptide: A tryptic peptide that was measured in the LC/MS experiment. These are portions of the proteins present in the sample.
3. Protein Group: A group of proteins measured in the proteomics analysis. Proteins within a protein group consist of proteins whose peptides cannot be distinguished with LC/MS/MS.
4. Post-translational modification (PTM): a chemical modification made to a protein after translation.
5. Proteoform: A single protein species complete with PTM and full sequence specification.
6. Intensity: The relative abundance of the protein/peptide in the sample.

Contact us to initiate a project

www.panomebio.com

info@panomebio.com

4340 Duncan Avenue, St. Louis, MO 63110