

**MASS SPECTROMETRY-CENTERED MULTI-
OMIC APPLICATIONS IN THE ANALYSIS OF
INFLAMMATION AND EXPOSURE**

A DISSERTATION

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Andrew Traver Rajczewski

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISORS: Drs. Timothy J. Griffin and Natalia Y. Tretyakova

October 2022

Acknowledgments

Conventional wisdom has it that the Acknowledgments section of a doctoral dissertation should not be the greatest part of the work, despite my feelings on the matter. For this reason, I will endeavor to keep my thanks brief. I have been extraordinarily fortunate in having a pair of the most knowledgeable, innovative, patient, and supportive advisors in Professors Natalia Y. Tretyakova and Timothy J. Griffin; I will always be grateful and indebted to them for their insight, their guidance, and their encouragement of an old dog struggling to learn new tricks.

I would also like to thank the members of my committee- Dr. Yue Chen, Dr. Chad Myers, and Dr. Christine Wendt- for their suggestions and advise during my PhD career. In this vein I would also like to thank Dr. Sharon Murphy for her extremely helpful criticism during my preliminary examination. In addition, I would like to offer my gratitude to the administrative staff in the BMBB program for their indefatigable patience and continuous help. I would like to thank Dr. Margareta Törnqvist, Dr. Isabella Karlsson, Lorena Ndreu, Efstathios Vryonidis, and Giulia Martella (Stockholm University) for their scientific advice and guidance during our collaborations as well as their gracious roles as hosts during my brief stay in Sweden, whom I consider to be friends as much as they are colleagues.

I am grateful to all members of the Griffin and Tretyakova laboratories past and present for their scientific acumen and comradery, especially Dr. Pratik Jagtap and Subina Mehta for their support and mentorship. I would like to thank Dr. Peter Villalta and Yingchun Zhao from the Masonic Cancer Center Analytical Biochemistry Core for their training and advice in conducting experiments. I would also like to thank I would also like to thank Robert Carlson at the Masonic Cancer Center for his endless patience and help during my Ph.D. career.

Finally, this dissertation represents a milestone in a literal dream come true and could not have been done on my own. I'd like to thank my family, especially my parents Stephen

and Sandy, my sister Kaelyn and brother-in-law Zach, and my in-laws Bill, Celeste, and Katharine for their love and support that allowed me to take the step and try again in a new home 1500 miles away. I'd also like to thank my other family, the network of friends that have graced me with their friendship and were integral to this process, especially Dani, Peter, Brittany, Jason, George, Andrew Eckel, Jenna, Nick, Makayla, Nicole, Julia, Rita Liz, and Gabi; you all walk with me daily, even from the other side of the world. Finally, and most importantly, I want to thank my wife Diane, who has laughed and cried with me throughout this entire process, believed in me when I couldn't myself, supported me when I was unsupportable, and whom I love with all my heart.

Thank you also to my dog Winnie, my favorite alarm clock, for all the walks and kisses.

Dedication

This work is dedicated to the memory of my late friends, Mikhail Morrison and Jesse McMichael, both of whom left far too early, before they could show the world what they could do. I am richer for having known you both and hope you both knew how special you were to everyone.

Abstract

Bottom-up proteomics represents an exciting technology which has found great utility across multiple fields of biological research. Using high-resolution mass spectrometry coupled with sophisticated bioinformatic software applications, bottom-up proteomics affords qualitative and quantitative information that reflects the actual molecular phenotype of a system in ways that next-generation sequencing technologies do not. Despite this, there are also blind spots in conventional bottom-up proteomics experiments; many of these limitations can be abrogated via the integration of bottom-up proteomics with other forms of ‘omics technologies and data. Through supplemental bioinformatic workflows, putative identifications of non-canonical peptides or non-host peptides (e.g microbial, viral) can be validated. The use of RNA-Seq data can be used to generate protein sequence databases files for proteogenomics, where non-canonical peptide sequences arising from genomic mutations, translocations, aberrant splicing events, etc. which are invisible to conventional proteomics experiments can be readily detected. In addition, by integrating and directly comparing proteomics data with transcriptomic data, levels of epigenetic and/or post-transcriptional control can be examined in a system in response to stimuli of interest that are invisible to both technologies. These supplemental approaches expand the power of bottom-up proteomics to where it becomes a highly useful tool for studying systems in which multiple levels of gene product expression response are regulated, including viral infections, tissues undergoing long-term inflammation, and exposure to endogenous and exogenous electrophiles.

The first chapter of this thesis provides an overview of the current state of proteomics-centered multi-omic technologies and their potential utility in biological research. The review begins with outlining the improvements to bottom-up proteomics technologies which have enabled a greater depth of information, such as isobaric peptide tagging, data-independent acquisition, ion mobility applications, and other instrumental advances. From there, bioinformatic tools are discussed that are of use in the analysis of proteomics data, with a focus on integrating mass spectrometry-based data with other forms of ‘omics data. Specific applications of using RNA-Seq data to inform the data analysis of proteomics, proteogenomics, are also addressed. The chapter concludes with notable instances of proteomics-centered multi-omics analysis as well as potential future applications of these technologies.

The second chapter of this thesis addresses the analysis of open-source proteomics datasets with customized multi-omic bioinformatic tools to determine the optimal targets for the detection of SARS-CoV-2 infections in patients. Through the use of *in vitro* and patient datasets, a panel of potential viral peptides were established, was used to search patient datasets. Ultimately, we found four peptides in the viral nucleocapsid which were reliably detected in patients and were unique to the SARS-CoV-2 virus.

The third chapter of this thesis utilizes proteogenomics workflows to examine the consequences of long-term inflammation in the proximal colon tissue of a murine model of inflammatory bowel disease. In this model, Rag2^{-/-}Il-10^{-/-} mice were subjected to five months of *Helicobacter hepaticus* infection in their colon to trigger chronic infection. For these analyses, RNA-Seq data acquired from these test subjects in an earlier study were converted into a FASTA protein sequence database containing variant sequences stemming

from this treatment. Through quantitative proteogenomic analysis we noted significant changes in abundances of proteins consistent with an inflammatory response; through bioinformatic analysis of our data, we also validated and confirmed the presence of 39 non-canonical peptides across our infected and control samples, demonstrating the importance of validation of targets of interest in proteogenomic studies.

The fourth chapter of this thesis integrates multiple levels of 'omic analysis to examine the effects of inflammation on murine Type II pneumocytes, a constituent cell within alveoli which serve as the source of lung adenocarcinomas. Mice were exposed to intranasal dosages of LPS or to whole-body cigarette smoke exposure for variable amounts of time before being sacrificed and the Type II cells isolated for analysis. Bottom-up proteomics of cells subjected to LPS for 3 weeks revealed a phenotype consistent with inflammation; this was reinforced when compared to transcriptomic data from the same cells, as these showed. Global proteomics analyses of Type II pneumocytes of mice subjected to exposure to cigarette smoke revealed significant changes in protein abundances occurring after after 10 weeks of exposure with a 4-week recovery period post exposure, encompassing biological processes such as nucleotide and amide metabolism as well as synthesis and acetyl CoA synthesis, which demonstrated a greater degree of disjuncture with the associated RNA-Seq data as compared to our LPS study.

The fifth chapter of this thesis examines the utility of bottom-up proteomics in examining the formation of amino acid adducts in hemoglobin, which serves as a valuable reservoir for exposome studies due to its longevity and high concentration within the blood. We were able to validate the presence of 4-hydroxybenzyl adducts at the N-terminal valine of hemoglobin and demonstrate their formation at nucleophilic side chains within the

protein. In addition, we compared bottom-up proteomics to the FIRE method, an experimental procedure which serves to isolate N-terminal adducts in hemoglobin for LC-MS detection, with a panel of electrophilic compounds incubated with blood at various concentrations and incubation times. Ultimately, we found that a proteomics-based approach to untargeted adductomics allowed for the detection of novel adducts at a number of sites within hemoglobin.

In this thesis we have applied mass spectrometry-based ‘omics technologies to complicated biological systems. We have utilized publicly available proteomics datasets to determine the optimal targets for MS-based detection of SARS-CoV-2 in patient samples. Using RNA-Seq data, we performed quantitative proteogenomic analysis of a murine model of IBD and validated the presence of several non-canonical peptide sequences. We also used multi-omic analyses to compare LPS-driven and cigarette smoke-driven inflammation of murine Type II pneumocytes. Finally, we demonstrated the utility of bottom-up proteomics in detecting and characterizing adducts in human hemoglobin as a record of the exposome. Overall, this work expands the utility of proteomics-centered analyses in characterizing systems subjected to viral infection, inflammatory stimuli, and exposure to environmental contaminants.

Table of Contents

Table of Contents

ACKNOWLEDGMENTS	I
DEDICATION	III
ABSTRACT	IV
TABLE OF CONTENTS	VIII
LIST OF TABLES	XIII
LIST OF FIGURES	XVI
LIST OF ABBREVIATIONS	XXVIII
I. LITERATURE REVIEW	33
1.1 The utility of proteomics, both alone and as a centerpiece of multi-omics analyses	34
1.2 Recent advances in MS-based proteomics enabling proteome-centered multi-omics	37
1.2.1 Sample preparation	39
1.2.2 Experimental methods for quantitative proteomics: advances in experimental methods, design, and instrumental analysis	41
1.2.3 Improvements in MS instrumentation	45

1.3. Bioinformatic integration of proteomics and other ‘omics	47
1.3.1 Common problems and strategies for data integration	47
1.3.2 Current software applications for integrative analysis of multi-omics results	50
1.3.3 User-friendly multi-omics platforms for increased access and flexibility	56
1.3.4 Prominent examples of proteomics-based multi-omics	59
1.4 Proteogenomics: genome- and transcriptome-driven proteomics	62
1.5. Projected future applications of proteomics-centered multi-omics	68
1.6 Concluding remarks and thesis goals	70
II. A RIGOROUS EVALUATION OF OPTIMAL PEPTIDE TARGETS FOR MS-BASED CLINICAL DIAGNOSTICS OF CORONAVIRUS DISEASE 2019 (COVID-19)	73
2.1. Introduction	74
2.2. Materials and Methods	76
2.3 Results	88
2.3.1. Sequence Database Searching Results.....	88
2.3.2. Peptide Validation Results.....	92
2.3.3. Identifying Detected Peptides with the Highest Quality Spectra	97
2.3.4 Viral Specificity of High-Quality Peptides Detected in SARS-CoV-2.....	108
2.4 Discussion	111

III. QUANTITATIVE PROTEOGENOMIC CHARACTERIZATION OF INFLAMED MURINE COLON TISSUE USING AN INTEGRATED DISCOVERY, VERIFICATION, AND VALIDATION PROTEOGENOMIC WORKFLOW118

3.1 Introduction 119

3.2 Materials and methods..... 121

3.3 Results 141

 3.3.1 Creation and sectioning of a custom RNA-Seq based FASTA database141

 3.3.2 Global proteogenomic analysis reveals inflammation-driven changes in protein abundance142

 3.3.3 Galaxy-P provides peptide-centric discovery of non-canonical sequences152

 3.3.4 PepQuery verifies the highest confidence non-canonical peptide candidates.....157

 3.3.5 Targeted proteomics experiments validate the presence of non-canonical peptides.....168

3.4 Discussion 173

IV. MULTI-OMIC ANALYSES OF LPS- AND CIGARETTE SMOKE-DRIVEN INFLAMMATION IN TYPE II PNEUMOCYTES183

4.1 Introduction 184

4.2. Materials and methods..... 189

4.3 Results 201

 4.3.1 Method validation of on-tip TMT labeling and STAGE-tip-based high pH fractionation201

 4.3.2 Bottom-up proteomics analyses of Type II cells of A/J mice intranasally treated with LPS demonstrates increased proteome changes consistent with inflammation206

 4.3.3 Multi-omic analysis to examine the correlation between the transcriptome and the proteome responses to LPS.....210

4.3.4 Multi-omic comparison of proteomics, transcriptomics, and epigenomics after 3 weeks of LPS exposure.....	214
4.3.5 Bottom-up proteomics of Type II pneumocytes of mice exposed to cigarette smoke	216
4.3.6 Multi-omic comparison of cigarette smoke-driven proteomic and transcriptomic changes in Type II cells of A/J mice treated with cigarette smoke	223
4.3.7 Direct comparison of LPS- and cigarette smoke-induced proteome changes shows distinct differences	229
4.3.9 A comparison of CPTAC lung cancer data shows potential therapeutic candidates of interest	232
4.4 Discussion.....	234

V. FURTHERING THE USE OF BOTTOM-UP PROTEOMICS FOR UNTARGETED ADDUCTOMICS OF HEMOGLOBIN238

5.1 Introduction	239
5.2. Materials and methods.....	247
5.3 Results	257
5.3.1 Global proteomic analysis of hemoglobin exposed to para-QM shows adduct formation at nucleophilic side chains.....	257
5.3.2 Targeted proteomic analysis verifies the presence of para-QM adducts at cysteine, histidine, lysine, serine, threonine, and tyrosine side chains	266
5.3.3 Differential reactivity of the hemoglobin side chains towards para-QM.....	273
5.3.4 4-Hydroxybenzyl adducts form at side chains characterized by low pKa values and/or high degrees of solvent accessibility.....	275
5.3.5 4-Hydroxybenzyl Adduct Formation in Hemoglobin Treated with 4-Hydroxybenzaldehyde.....	279
5.3.6 4-Hydroxybenzyl adduct formation in hemoglobin exposed to UV light.....	281
5.3.7 FIRE analysis of samples reveals linear dose-responses to most electrophiles	287

5.3.8 Bottom-up proteomics allows for the detection of a greater variety of Hb adducts as compared to FIRE	289
5.4 Discussion.....	296
VI. SUMMARY AND CONCLUSIONS.....	303
VII. FUTURE DIRECTIONS	307
7.1 Targeted detection and absolute quantitation of SARS-CoV-2 peptides in patient samples	307
7.2 Targeted detection and absolute quantitation of non-canonical peptides in proximal colon tissue at different stages of inflammation and oncogenesis	308
7.3 Evaluation of the roles of Pdhx, Psma6, Ruvbl1, and Ywhaq in cell proliferation and smoking-induced lung cancer	308
7.4 Automation of untargeted adductomics in Galaxy	309
BIBLIOGRAPHY.....	311

List of Tables

Table 1.1. A selection of bioinformatic tools for proteomic analyses.	51
Table 1.2. Multi-omics platforms with graphical user interfaces for ease of use.	57
Table 1.3. Prominent examples of proteomics-centered multi-omics.....	61
Table 2.2. Peptides generated from MS datasets in the construction of the library and validation in the patient datasets	89
Table 2.3. The presence of the four optimal SARS-CoV-2 target peptides in clinical datasets. The presence of the peptides MAGNGGDAALALLLDR, DGIIWVATEGALNTPK, RGPEQTQGNFGDQELIR, and IGMEVTPSGTWLTYTGAIK in COVID-19 positive patients.	107
Rag2 ^{-/-} -Il10 ^{-/-} mice were subjected to three oral gavages over the course of one week with either saline (control) or <i>Helicobacter hepaticus</i> culture, after which the infected mice developed chronic colorectal inflammation (Figure 3.1) ²⁶⁶	122
Table 3.1. Peptides generated from MS datasets in the construction of the library and validation in the patient datasets	126
Table 3.2. Acetonitrile concentrations of eluted TMT-labeled peptides.....	128
Table 3.3. Inclusion list for targeted detection of non-canonical peptides in proximal colon samples. Based on the initial global proteomics data, m/z values and charge states were determined for putative non-canonical peptides and used to create this inclusion list for targeted PRM analyses. Lower-case amino acid codes represent covalent modification by acetylation, oxidation, or phosphorylation.....	136
Table 3.4. Proteins identified as being increased in abundance in inflamed proximal colon tissue vs controls. Proteins shaded in red show increased abundance in inflamed proximal	

colon tissues, proteins shaded in blue show increased abundance in the control tissues.
..... 147

Table 3.5. Non-canonical sequence peptides identified, validated, and quantified in inflamed proximal colon tissues. Peptide precursors with at least three product ions (b- and/or y-ions) were detected in Skyline. A weighted contrast angle of the MS/MS spectra peaks against those of the reference library is reported in Skyline as the dot product, with a score of 1.0 representing a perfect match and 0.0 representing no match 158

Table 3.6. Human analogues of mouse non-canonical peptides. Human versions of murine non-canonical peptides found in TCGA datasets with PepQuery. Green-highlighted peptides show a decreased abundance in inflamed proximal colon samples while red-highlighted samples show an increased abundance in inflamed proximal colon samples.
..... 179

Table 4.1. Multi-omic comparison of genes with significantly altered methylation, gene expression, and protein abundance. 215

Table 4.2. Genes showing significant changes at the transcriptome and proteome levels after 10 weeks of cigarette smoke exposure. 228

Table 4.3. Genes showing sustained increases in protein abundance after 10 weeks of cigarette smoke exposure also seen upregulated in CPTAC data. Highlighted cells represent significant values..... 233

Table 5.1. Inclusion list for FTH-analytes in FIRE samples. The D₇-acrylamide adduct corresponds to the added internal standard. 255

Table 5.2. Peptides detected in hemoglobin treated with excess para-QM..... 262

Table 5.3. Peptides chosen for targeted proteomic analysis. 264

Table 5.4. Hemoglobin amino acid residues showing 4-hydroxybenzyl adduct formation before and after treatment with para-QM, 4-hydroxybenzaldehyde, and UV irradiation. a) Levels of adduct formation relative to control samples following incubation with potential 4-hydroxybenzyl adduct sources. Red highlighted values indicate increased adduct formation following treatment, green highlighted values indicate a loss of adduct following treatment. Residues annotated with an asterisk showed significant statistical differences to controls. The N-terminal peptides were not included in targeted experiments. b) Levels of endogenous adducts in the control hemoglobin samples..... 271

Table 5.5. N-terminal peptides detected by bottom-up proteomics following hemoglobin exposure to various electrophiles. Boxes highlighted in green represent manually validated peptides, red represent peptides that were detected but not validated, and grey represents peptides that were not detected. 291

List of Figures

Figure 1.1. An overview of methodologies that can be used to improve proteome coverage. Sections are highlighted in accordance with their provenance as sample preparation strategies (red), improvements to methods for protein quantitation (blue), or innovations to instrument design and/or operation (yellow). Generated using biorender.com..... 38

Figure 1.2. Proteogenomics workflows. a) For generating a FASTA library from genomic sequencing data, the FASTQ files are either aligned against a reference genome or assembled into contigs and then a working genome. In either case, the resulting assembled sequencing data is either translated into proteins in six open reading frames or submitted to analysis using gene identifying software, the results of which are translated into proteins. b) For generating a FASTA library from RNA sequencing data or exome sequencing data, FASTQ files are aligned to a reference genome. The assembled data is then searched against a variant sequence detector, the results of which are then converted into a FASTA library. Unaligned sequences from UTR transcription or novel RNA processing events are subjected to three-frame translation. c) Raw proteomic data is searched against bespoke FASTA libraries to detect non-canonical peptide sequences. Flowcharts made using lucidchart..... 65

Figure 2.1. MS/MS datasets used in the determination of optimal SARS-CoV-2 peptides for COVID-19 diagnosis. a) Cell culture, clinical, and bioinformatic datasets used to generate the SARS-CoV-2 peptide panel. b) Clinical datasets queried using the initially characterized peptide panel from a) to determine the feasibility of COVID-19 diagnosis via targeted proteomics as well as determine the optimal peptide targets for those assays. Figures were made using BioRender. 77

Figure 2.2. Workflows used in the interrogation of MS-data to identify and validate SARS-CoV-2 peptides a) Galaxy-based sequence database search workflow to detect and confirm SARS-CoV-2 peptides. MS/MS spectra from cell culture or clinical datasets were searched against appropriate protein sequence databases (protein sequences from COVID-19, contaminants, and Human Protein sequences) using SearchGUI/ Peptide Shaker. The peptide output was filtered to extract COVID-19 peptides, and the output was confirmed using PepQuery to extract confident peptides. mzidentML generated through this workflow was subsequently used for analysis in Lorikeet b) Workflow to verify detected SARS-CoV-2 peptides. A list of 639 Peptides (theoretical and validated peptides obtained from the cell-culture and clinical datasets) was subjected to PepQuery analysis of COVID-19 datasets to identify the presence of SARS-CoV-2 peptides. The quality of the peptide spectral matches (PSMs) was reviewed using Lorikeet visualization within the Multi-omics Visualization Platform for further validation. Peptides were also searched against NCBI-non redundant database and Unipept 4.3 for taxonomic annotation..... 81

Figure 2.3. Protein assignment of detected and validated SARS-CoV-2 peptides: Circos plot of peptides against SARS-CoV-2 proteins (outermost ring). Of the 639-peptide panel (2nd outermost ring), many peptides could be identified using our validation workflow in clinical and cell culture datasets (3rd outermost ring). Peptides derived from ORF9b, papain-like protease, Nsp4, Nsp10, uridylate endoribonuclease (Nsp15) and certain spike protein peptides were only found in cell culture datasets (2nd innermost ring). Peptides chosen for targeted analysis are annotated in the innermost ring. Circos plot was generated in Galaxy²³⁸..... 90

Figure 2.4. Alignment of the 639-peptide panel to viral proteins from SARS-CoV-2. Peptides detected from patient and cell culture datasets were aligned to the SARS-CoV-2 proteome. Proteins were colored in terms of their classification as structural proteins (green), non-structural proteins (maroon), or open reading frames (blue). 91

Figure 2.5. List of validated peptide spectral matches in the oro-pharyngeal and nasopharyngeal mass spectrometry dataset (PXD020394). The bar diagram above shows the peptide-spectral matches after running the validation workflow for 639 SARS-CoV-2 peptide-panel against the five COVID-19 positive patient samples (with replicates) and five COVID-19 negative patient samples (with replicates). Several SARS-CoV-2 peptides were detected in COVID-19 positive samples (See samples labeled ‘POS”) and only two peptides were detected in two of the negative samples (NEG5 Rep1 and NEG2 Rep2). The SARS-CoV-2 peptides detected in COVID-19 negative samples did not meet the threshold of acceptable spectral quality in subsequent spectral validation. 93

Figure 2.6. Guidelines for the manual validation of MS/MS spectra using the Multi-omic Visualization Platform (MVP). The MS/MS spectra of peptides that passed validation in PepQuery were manually annotated using MVP based on a test cohort of four peptides that passed validation in COVID-positive patient datasets and two peptides that passed validation in COVID-negative patient data. The signal-to-noise ratio of the product ions within MS/MS spectra was examined, and spectra containing product ions with at least a three-fold higher intensity than noise level were retained. Next, the degree of completeness of the b- and y-ion series was considered, with passing spectra determined to have at least three consecutive b- or y-ions in their series. Peptides with spectra that passed these criteria were considered valid peptide targets for the detection of SARS-CoV-2. 99

Figure 2.7. Peptide spectral matches (PSMs) of SARS-CoV-2 peptides in the upper respiratory clinical datasets are of higher confidence than deep lung datasets. PSMs validated in oro/nasopharyngeal datasets, saline gargling samples, lung biopsy samples, and bronchoalveolar lavage fluids (BALF) using PepQuery as grouped into the proteins they aligned to; columns correspond to those peptides that passed PepQuery validation with minimal required confidence (left) as well as those associated with higher confidence (right). 102

Figure 2.8. MS/MS spectra of SARS-CoV-2 peptides most confidently identified in PepQuery (p-value < 0.001) and across the most clinical samples. Spectral quality was interrogated using the Lorikeet viewer implemented within the Multi-Omics Visualization Platform (MVP); images for annotated PSMs for these peptides were created using the PDV platform from the Zhang lab. 104

Figure 2.9. Specificity of target peptides as for coronaviruses and for SARS-CoV-2 a) MetaTryp taxonomic analysis of the 4 most consistently found peptides. Coronaviruses with matches to peptides are highlighted in red and font size is correlated with the number of peptides that show a match in that coronavirus. Created with BioRender.com b) Sequence identity of peptides that show BLAST-P alignment with viral nucleocapsid protein. 109

Figure 3.1. Outline of the experimental procedure. Mice were infected with *Helicobacter hepaticus* to induce inflammation, and proximal colon proteins and mRNA were collected for proteogenomic analysis. Peptides were digested and labeled for differential proteomic analysis and variant discovery, with some unlabeled peptides reserved for quantitation of variants. Created with BioRender.com. 123

Figure 3.2. Galaxy-P-based bioinformatics workflows utilized in the study of inflamed colon proteogenomics a) Generation of RNA-Seq based custom protein FASTA database b) Sectioning workflow to reduce RNA-Seq FASTA database size c) Identification and verified of non-canonical variant peptides. All workflows created with BioRender.com 130

Figure 3.3. Differential proteogenomic analysis of inflamed proximal colon samples. a) Enrichment of proteins in proximal colon tissue in response to chronic inflammation, as demonstrated via volcano plot of log₂ fold-change of protein abundance against -log₁₀ of corrected p-value. Proteins showing significant increases in abundance in inflamed tissues are highlighted in red, proteins showing decreased abundance in inflamed tissues are highlighted in blue. b) Gene Ontology analysis of increased abundance proteins in inflamed proximal colon samples shows enriched molecular functions (blue), biological pathways (red), reactomes (orange), and CORUM complexes (green). c) Gene Ontology analysis of decreased abundance proteins in inflamed proximal colon samples shows enriched molecular functions (blue), WikiPathways (brown), and CORUM complexes (green)..... 144

Figure 3.4. Non-canonical protein with differential abundance in the mouse model of colonic inflammation. The protein chr8: 73261429-73261687+ in the sectioned proteogenomic FASTA database was shown to be enriched in inflamed proximal colon samples. a) Genomic coordinates associated with chr8: 73261429-73261687+, visualized with the UCSC Genome Browser. b) Peptides associated with chr8: 73261429-73261687+ detected in Proteome Discoverer 151

Figure 3.5. Validation of the non-canonical peptides results in the ultimate retention of 58 non-canonical peptides. a) The process of narrowing down the initial 14,491 non-canonical peptides using BLAST-P results in 235 peptides without matches to the conventional mouse proteome. Subsequent analysis by PepQuery results in 58 non-canonical proteins retained, with 130 peptides rejected by PepQuery. b) 130 non-canonical peptides rejected by PepQuery broken down along their reasons for failing PepQuery, specifically through finding a better match to a reference peptide, failing to pass the statistical barriers of the search engine, and/or matching to reference peptides with hypothetical post-translational modifications. c) Rejected non-canonical peptide spectral match (above) compared with a better scoring match to a reference proteome peptide (below). d) The use of the unrestrictive modification option demonstrates a superior match to a peptide with a modified sequence showing C-terminal a-type ionization, the loss of the alpha carbon and carboxyl group of the C-terminal lysine. e) PSM of a short rejected non-canonical peptide with repeated residues which can readily be matched to scrambled decoy peptides. Spectra generated using PDV²³⁷. 153

Figure 3.6. Examples of MS/MS spectra for non-canonical peptides. Spectra of selected highly confident (p-value < 0.001) non-canonical peptides which passed PepQuery are presented here. Spectra were visualized using the Proteomics Data Viewer (PDV)..... 161

Figure 3.7. Characteristics of non-canonical sequences validated by PepQuery a) Peptides with non-canonical sequences can be classified in several categories based on their altered sequence or location within a gene. b) Chromosomal locations of non-canonical sequence peptides correspond to mouse chromosomes throughout the genome..... 166

Figure 3.8. Differential abundance analysis of non-canonical peptides detected in inflamed proximal colon samples. a) Fold-changes of variant peptides in the inflamed and control proximal colon samples, as measured via targeted mass spectrometry b) Comparison of RNA-Seq, proteomics-derived change in peptide abundances c) Categories of non-canonical peptides in peptides that show increased and decreased abundance in the inflamed proximal colon samples. 170

Figure 3.9. The genomic coordinates of the peptide AASSANIPK, found in the 3' untranslated region of the Sor11 gene. This peptide was found to have a slightly increased abundance in inflamed proximal colon tissue. Genomic coordinates determined via UCSC Genome Browser. 172

Figure 4.1. Experimental designs for multi-omic analyses of Type II pneumocytes from exposed and control mice. a) LPS exposure of A/J mice to induce pulmonary inflammation. b) Cigarette smoke exposure of A/J mice (3 weeks, 10 weeks, 10 weeks with 4-week recovery) c) Scheme for cell isolation and processing for proteomics, transcriptomics, and epigenomics. 186

Figure 4.2. Combined protocol for quantitative proteomics for samples with low amounts of total protein. Proteins are extracted from cells and immobilized on single-pot, solid-phase-enhanced sample preparation (sp3) beads for digestion, followed by TMT-labeling and high pH fractionation prior to LC-MS analysis. 195

Figure 4.3. Grouping patterns for TMT-11plex labeling the LPS- and cigarette smoke-exposed samples. Each color represents an LC-MS experiment that was run separately, with the ¹³C channel representing pooled samples for normalization. “Con” = control, “Exp’ = experimental (LPS or cigarette smoke). 199

Figure 4.4. Labeling efficiencies of *in situ* and on-tip strategies for TMT labeling. Average labeling efficiencies for TMT labels were determined for both labeling strategies using triplicate samples of HeLa digest. Statistical comparison of the labeling efficiency values yields a p-value greater than 0.05, indicating a lack of statistical difference between these methodologies..... 203

Figure 4.5. Numbers of proteins identified in commercial HeLa peptides with and without high pH fractionation of TMT-labeled peptides. Two 6 µg aliquots of TMT-labeled HeLa digest were fractionated into seventeen fractions using increased acetonitrile, which were then concatenated into nine fractions and analyzed via LC-MS on a Fusion Tribrid Orbitrap Mass Spectrometer. Protein identifications from two injections of 1µg of labeled peptide were compared with the fractionated data. 205

Figure 4.6. Global proteomics analysis of LPS-treated Type II pneumocytes reveals characteristic changes to the protein content. a) Volcano plot of differentially abundant proteins in Type II pneumocytes following three weeks of LPS exposure. b) Protein-protein interactions of proteins increased in abundance. c) GO terms associated with proteins that were increased in abundance following LPS exposure..... 207

Figure 4.7. Multi-omic comparisons of proteomic and transcriptomic data of LPS-exposed Type II cells. a) Correlation plot between transcriptome and proteome generated in QuanTP b) GO Analysis of genes that show an increase in protein abundance without concurrent change in transcription (blue cluster in Figure 4.7a). c) GO Analysis of genes significantly increased at the transcriptomic and proteomic level..... 212

Figure 4.8. Global proteomics analysis of Type II pneumocytes isolated from cigarette smoke-exposed A/J mice reveals exposure-dependent phenotypes. a) Volcano plot of

differential proteome abundances after 3 weeks of cigarette smoke exposure b) Volcano plot of differential proteome abundances after 10 weeks of cigarette smoke exposure c) Selected Reactome GO terms enriched for proteins significantly increased after 10 weeks of cigarette smoke exposure d) Volcano plot of differential proteome abundances after 10 weeks of cigarette smoke exposure and 4 subsequent weeks of recovery in clean air e) GO terms enriched for proteins significantly increased after 10 weeks of cigarette smoke exposure with subsequent recovery. 218

Figure 4.9. Multi-omic comparisons of proteomic and transcriptomic data for cigarette smoke-exposed Type II pneumocytes using QuanTP a) 3 weeks of cigarette smoke b) 10 weeks of cigarette smoke c) 10 weeks of cigarette smoke with 4 weeks of post-exposure recovery..... 224

Figure 4.10. A comparison of proteins with significantly changed abundances following exposure. a) A Venn diagram comparing proteins increased in abundance following LPS exposure, 3 weeks of cigarette smoke exposure, 10 weeks of cigarette exposure, and 10 weeks of cigarette smoke exposure followed by a 4-week recovery period. b) GO analysis of proteins increased in abundance as a result of LPS exposure and 10 weeks of cigarette smoke exposure followed by 4 weeks of recovery. 230

Figure 5.1. Approaches for detection of hemoglobin adducts by mass spectrometry. a) Schematic representation of the use of fluorescein isothiocyanate, or FITC (FI) for the measurement of N-terminal protein adducts (R) via modified Edman (E) procedure (FIRE) b) Bottom-up proteomics approach for detection of hemoglobin adducts. 242

Figure 5.2. Proposed reaction mechanisms for the formation of 4-hydroxybenzyl adducts using histidine side chains as an example. a) Formation of para-QM from para-QM

precursor with potassium fluoride and subsequent reaction. b) Reductive amination of 4-hydroxybenzaldehyde with sodium cyanoborohydride. c) Generation of para-QM upon reaction with ultraviolet light and subsequent adduct formation..... 244

Figure 5.3. Study design to compare FIRE and bottom-up proteomics approaches for the detection of hemoglobin adducts. Donor blood is incubated with either acrylamide, acrylic acid, glycidic acid, 2-methyleneglutaronitrile (2-MGN), 2,3-epoxypropyl phenyl ether (PGE), or 1-chloro-2,4-dinitrobenzene (DNCB) at varying concentrations prior to analysis using both methodologies and data analysis. 246

Figure 5.4. Tandem mass spectra of a) Hemoglobin alpha subunit N-terminal peptide with N-terminal 4-hydroxybenzyl adduct, b) Hemoglobin beta subunit N-terminal peptide with N-terminal 4-hydroxybenzyl adduct, c) 4-hydroxybenzaldehyde adduct of histidine 45 in alpha subunit, d) 4-hydroxybenzaldehyde adduct of cysteine 93 in beta subunit. Spectra in a) and b) were sourced from Proteome Discoverer 2.2, c) and d) were taken from Skyline v20.2..... 259

Figure 5.5. Putative structures of 4-hydroxybenzyl amino acid adducts. 268

Figure 5.6. 4-Hydroxybenzyl adducted side chains identified in global proteomics analysis and confirmed via targeted mass spectrometry, highlighted in red. N-terminal valines are highlighted in green. 269

Figure 5.7. Side chain adduct occupancy of residues in hemoglobin with titration of increasing amounts of para-QM. Adduct formation on side chain residues follows either a) linear or b) saturation relationship with increasing addition of para-QM. 274

Figure 5.8. a) NMR structure of human hemoglobin³⁷³ (PDB ID: 2h35) with 4-hydroxybenzyl adducted side chains shown in red. Alpha and beta subunits are gray, with

the N-terminal valine residues shown in green and heme molecules in blue. Clusters of adduct sites in alpha subunits are shown in subfigures b) and c), clusters of adduct sites in beta subunits are shown in subfigures d) and e). 276

Figure 5.9. Biophysical properties of N-terminal valine and amino acid side chains within the a) alpha subunit and b) beta subunit. Labeled residues correspond to nucleophilic side chains. Relative solvent accessibilities are scaled relative to the theoretical values⁵². pKa values were not calculated for non-ionizable functional groups. Physiological pH is designated by the horizontal red line. 278

Figure 5.10. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA..... 280

Figure 5.11. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA..... 284

Figure 5.12. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA..... 285

Figure 5.13. Distances between nearest tyrosine and a) α Ser44 and b) β Thr84 in following exposure to UV radiation.....	286
Figure 5.14. Dose-response curves of blood samples incubated with electrophiles and analyzed via the FIRE method.....	288
Figure 5.15. MS/MS spectrum of N-terminal DNCB adduct.....	290
Figure 5.17. Dose-response curves derived from bottom-up proteomics. Adducted peptides were normalized to non-modified peptides at the a) alpha chain N-terminus and b) beta chain N-terminus.....	295
Figure 5.18. Consensus sequence of side chains showing adduct formation (empty position 8).....	298

List of abbreviations

2D-LC	two-dimensional liquid chromatography
2-MGN	2-methyleneglutaronitrile
4-HBA	5-hydroxybenzaldehyde
AA	acrylamide
AC	acrylic acid
AG	glycidic acid
AGC	automatic gain control
AIDS	acquired immunodeficiency syndrome
AUC	area under the curve
BALF	bronchiolar lavage fluid
BCA	bicinchoninic acid assay
BLAST	basic local alignment search tool
CBSF	cerebrospinal fluid
CCS	collision cross section
COPD	chronic obstructive pulmonary disorder
COVID19	coronavirus disease 2019
CPTAC	Clinical Proteomics Tumor Analysis Consortium
CV	counter voltage
Da	dalton
DDA	data dependent acquisition
DIA	data independent acquisition

DNCB	1-chloro-2,4-dinitrobenzene
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
ESI	electrospray ionization
FACS	fluorescence-assisted cell sorting
FAIMS	high-field asymmetric waveform ion mobility spectrometry
FASP	filter-assisted sample preparation
FASTA	FAST-all
FDA	Food and Drug Administration
FDR	false discovery rate
FIRE	FITC for the measurement of adducts (R) via modified Edman procedure
FITC	fluorescein isothiocyanate
FTH	fluorescein-2-thioxo-imidazolidin-5-one
GO	gene ontology
GPF	gas-phase fractionation
GUI	graphical user interface
H hepaticus	Helicobacter hepaticus
Hb	hemoglobin
HCC	hepatocellular carcinoma
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HIV	human immunodeficiency virus
IAA	iodoacetamide
IPA	ingenuity pathway analysis
IT	injection time
iTRAQ	isobaric tags for relative and absolute quantitation
KEGG	Kyoto encyclopedia of genes and genomes
LC	liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
LFQ	label-free quantitation
LPS	lipopolysaccharide
LUAD	lung adenocarcinoma
MERS-CoV	middle eastern respiratory syndrome coronavirus
MS	mass spectrometry
MSF	magellan storage file
MVP	multi-omics visualization platform
NADH	nicotinamide adenine dinucleotide
NCE	normalized collision energy
NGS	next generation sequencing
NMR	nuclear magnetic resonance spectroscopy
NNK	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)
ORF	open reading frame
para-QM	para-quinone methide

PASEF	parallel accumulation-serial fragmentation
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PDV	proteomics data viewer
PFOS	perfluorooctanesulfonic acid
PFP	pentafluorophenyl
PGE	2,3-epoxypropyl phenyl ether
PMSF	phenylmethanesulfonyl fluoride
ppm	parts per million
PRM	parallel reaction monitoring
PSM	peptide spectral match
PTM	post-translational modification
RF	radio frequency
RNA-Seq	RNA sequencing
RNS	reactive nitrogen species
ROS	reactive oxygen species
RT-LAMP	reverse transcription loop-mediated isothermal amplification
RT-qPCR	quantitative reverse transcription polymerase chain reaction
SARS-CoV	severe acute respiratory syndrome coronavirus
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2

SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SHERLOCK	specific high sensitivity enzymatic reporter unlocking
SILAC	stable isotope labeling by amino acids in cell culture
SP3	single-pot, solid-phase-enhanced sample preparation
SRM	selected reaction monitoring
TCGA	the Cancer Genome Atlas
TEAB	Triethylammonium bicarbonate
TIMS	trapped ion mobility spectrometry
TMM	trimmed mean of the m-values
TMT	Tandem mass tag
TOF	time of flight
UCSC	University of California Santa Cruz
UHPLC	ultra-high-performance liquid chromatography
UTR	untranslated region
UV	ultraviolet
WHO	World Health Organization

I. LITERATURE REVIEW

Adapted from:

Rajczewski AT, Jagtap PD, Griffin TJ. An overview of technologies for MS-based proteomics-centric multi-omics [published online ahead of print, 2022 May 2]. *Expert*

Rev Proteomics. 2022;1-17. doi:10.1080/14789450.2022.2070476

Pratik D. Jagtap and Timothy J. Griffin edited the text.

1.1 The utility of proteomics, both alone and as a centerpiece of multi-omics analyses

The twenty-first century has seen the rise of advanced nucleic acid sequencing technologies, driving the new discipline of systems biology, in which large-scale molecular data from a model system are examined in an integrated fashion to more holistically understand molecular networks underlying normal biological processes and dynamic response to stimuli^{1,2}. Beginning with the development of DNA sequencing in the 1970s³, improvements to sequencing technology allowed for the ability to sequence whole genomes of model organisms, eventually leading to the development of RNA sequencing (RNA-Seq) technology. These “next-generation” sequencing (NGS) approaches provide rapid genome sequencing and qualitative and quantitative information on transcribed messenger RNA⁴. This transcriptome sequencing information offers a picture of the genes expressed under a given set of conditions, providing insights into gene regulation mechanisms and potential biochemical functional response within the system.

While these powerful sequencing techniques have individually shown considerable utility in fields ranging from cancer research^{5,6} to microbiology⁷, these technologies lack a direct measurement of functional molecules responsible for the biochemistry driving phenotypic changes that occur in a cell, tissue, or organism. This is due in part to higher-level epigenetic regulatory mechanisms such as DNA methylation⁸, histone acetylation⁹, and siRNA and miRNA suppression of mRNA translation^{10,11}. To complete the molecular picture, it is essential to examine the expression of proteins present in a system (i.e., the proteome). This can be done using liquid chromatography (LC) coupled to mass spectrometry (MS). In bottom-up MS-based proteomics¹², proteins are isolated from a

system and enzymatically digested into their constituent peptides. Complex peptide mixtures are analyzed by LC-MS, collecting tandem mass spectra (MS/MS) which can be used as fragmentation signatures of each detected peptide. Each MS/MS spectrum is matched to sequences contained within a database of known or predicted proteins expressed by the organism(s) being studied, using customized bioinformatic software to identify the proteins present. Having advanced considerably with the introduction of high-resolution and high scan-rate instrumentation^{13,14,15}, MS-based proteomics is now a mature field with many research applications in the biomedical^{16,17}, biotechnological^{18,19}, and ecological research spaces²⁰.

Despite its advantages, proteomics is not without limitations. Historically, proteomics methodologies have only been able to identify a portion of the proteome within complex biological systems, largely due to the variable abundances of different proteins in a cell^{21,22} as well as chemical heterogeneity resulting in a complex array of proteoforms²³ expressed by any given coding gene. In addition, bottom-up proteomics is reliant on the use of genomic data of the organism under study to refine and predict proteins expressed. Although convenient, using a reference library of predicted canonical proteins does not allow for detection of potentially biologically relevant proteoforms expressed from sample-specific coding sequence variants and processing events that are not present within the reference proteome²⁴. Finally, conventional bottom-up proteomics data does not always measure activity of the many enzymes comprising signaling and metabolic pathways critical to living systems^{25,26}. The measurement of metabolites using LC-MS and/or other methods (e.g., NMR), known as metabolomics, can be used to investigate these changes in protein activity. However, metabolomics methodology has its own analytical challenges

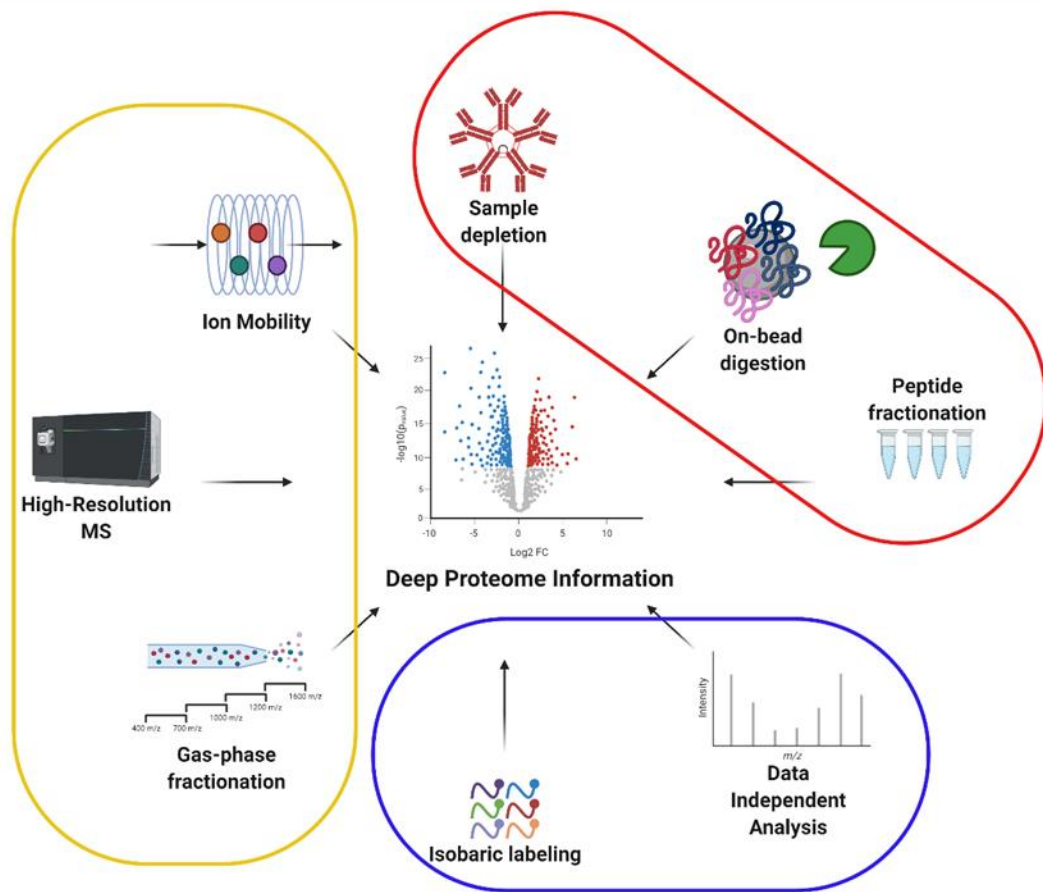
such as sensitivity limitations due to ion suppression²⁷, suitable database selection²⁸, sample variability²⁹, etc. It should be noted, however, that current advances in methodologies have helped researchers overcome some of the challenges in both MS-based proteomics and metabolomics and are now able to provide deep information at the protein and metabolite levels.

Given these now maturing, sensitive and accessible technologies across the ‘omic domains, the concept of multi-omic analysis has become a viable option for many researchers. Multi-omics seeks to integrate system-wide information generated by different ‘omic technologies to gain a more comprehensive molecular picture within biological systems. Given the array of ‘omic technologies now available, multi-omics can take on many flavors, depending on the types of information being generated and integrated^{30,31,32}. Here, we review multi-omic approaches which rely on MS-based proteomics as the centerpiece. We describe some of the recent advances in experimental methods and sample preparation, and MS instrumentation that have helped overcome some of the past limitations of MS-based proteomics, facilitating the generation of deep proteomic information necessary for multi-omic analysis. We also provide an overview of bioinformatic tools and approaches available for the integration of proteomic data with other ‘omics information. Collectively, this review should help to guide researchers seeking to integrate MS-based proteomics data with other ‘omic information to drive new discoveries across a wide variety of research fields.

1.2 Recent advances in MS-based proteomics enabling proteome-centered multi-omics

Despite its promise, traditional MS-based proteomics lacked the depth of information afforded by NGS technologies focused on sequencing DNA and/or expressed RNA transcripts³³. Lacking the ability to amplify low-abundance proteins, as well as the sheer chemical complexity of the potentially millions of expressed proteoforms^{23,34}, even the most cutting-edge MS-based proteomic methods still only reliably detected a portion of the proteome within complex samples. Fortunately, a combination of advances in the past several years have significantly improved this situation (**Figure 1.1**), dramatically increasing the depth of information now attainable by MS-based proteomics. Here we review some of these advances that are now available to the wider research community.

Figure 1.1. An overview of methodologies that can be used to improve proteome coverage. Sections are highlighted in accordance with their provenance as sample preparation strategies (red), improvements to methods for protein quantitation (blue), or innovations to instrument design and/or operation (yellow). Generated using biorender.com



1.2.1 Sample preparation

Increasing the number of detected and quantified proteins begins with optimizing the processing of samples prior to LC-MS analysis - especially important when analyzing precious, material-limited samples. A conventional protein sample preparation strategy involves the reduction of thiol groups of cysteine amino acid chains, followed by an alkylation step to prevent the reformation of disulfide bonds, after which the protein samples are digested in situ using trypsin³⁵. Following their digestion, samples can then be desalted to remove contaminants and injected into an LC-MS platform for analysis. Although standardized, this workflow of sample preparation can be altered to increase the sensitivity for detection of lower abundance proteins.

Peptides from highly abundant proteins suppress the detection of those from lower abundance proteins. Biofluids such as serum and plasma, as well as others (e.g., urine, lung lavage, cerebral spinal fluid, etc) are known to have high abundance proteins such as albumin which are many orders of magnitude more abundant than other proteins of interest³⁶; as such, many products and protocols based on immunoprecipitation strategies have been developed to remove albumin and other carrier proteins from blood³⁷. Similar strategies have been employed for cerebrospinal fluid (CSF)³⁸ and urine³⁹. Many epithelial tissues such as intestinal villi, lung tissue, etc. are also infused with blood, and can also be immunodepleted following homogenization⁴⁰ to improve the depth of detection of lower abundance proteins. Indeed, depletion methods coupled with the most sensitive MS-instrumentation can now detect proteins across ten orders of magnitude in abundance from serum or plasma samples^{41,42}.

Despite the utility of in situ protein digestion with trypsin, the standard protocols incorporate multiple sample-handling steps, making it less than ideal for material-limited samples. Handling steps introduce sample loss and make processing of large sample cohorts cumbersome. As a solution, Filter-Aided Sample Preparation (FASP)⁴³, was introduced, wherein reduction, alkylation, buffer exchange, and digestion can all occur in a single “pot”, using molecular weight cutoff filters within a single microcentrifuge tube as a reaction vessel. Other strategies have followed and extended the FASP methods, including sample processing and clean-up using small-scale solid-phase extraction stage tips^{44,45}, digestion of sequestered proteins in the three-dimensional S-Trap⁴⁶, immobilization of proteins onto solid sphere supports in enzymatic reactors⁴⁷, and precipitation of proteins onto magnetic⁴⁸ or glass beads⁴⁹ with subsequent digestion which allows for processing materials while minimizing sample handling steps.

Fractionation of complex peptide samples generated via protein digestion prior to LC-MS analysis also provides a means to increase sensitivity. Fractionation using orthogonal LC methods has long been known to increase sensitivity via simplification of mixtures introduced into the MS, thereby relieving ion suppression⁵⁰. SDS-PAGE gels, followed by in-gel protease digestion, allow for the pre-fractionation of proteins prior to digestion⁴², though this approach may be unfeasible for large cohorts of protein samples. A common alternative strategy is to perform high pH, reverse-phase pre-fractionation of peptide samples using either high-performance liquid chromatography⁵¹ or commercial centrifuge-based kits, as LC-MS analysis for bottom-up proteomics is generally performed using low pH on reverse-phase columns giving results akin to two-dimensional liquid chromatography (2D-LC) coupling orthogonal separation methods. Other stationary

phases useful for peptide pre-fractionation include ion exchange resins⁵², mixtures of ion exchange and reverse-phase modalities⁵³, and pentafluorophenyl (PFP) resin⁵⁴. For experiments involving extremely limited amounts of material, stage tip-based fractionation can be performed to increase the number of proteins identified^{55,56,57}.

1.2.2 Experimental methods for quantitative proteomics: advances in experimental methods, design, and instrumental analysis

The experimental design of proteomics experiments, which is guided by the experimental methods employed, has direct bearing on their inherent utility in the context of multi-omics analyses. To determine abundance changes in the proteome in response to stimuli and integrate these changes with other ‘omic information, it is important to get accurate and deep quantitative data on the proteome. The main strategies for quantitative proteomics break down along two lines, namely isotope labeling methods and unlabeled methods. With unlabeled quantitation, also called label-free quantitation (LFQ), the digested and desalted peptides are analyzed via LC-MS with no chemical modification to the peptides themselves. Quantitative information on peptide and protein levels in LFQ comes from the spectral counts (counting the number of peptide spectral matches, or PSMs, that map to a given protein), or through the area under the curve (AUC) in the MS1 chromatogram for peptides identified by PSMs⁵⁸. Label-free quantitation is now a mature methodology with numerous software options available for these analyses^{59,60}, though the method is not without its limitations. Since the peptide samples in LFQ proteomics experiments are analyzed in the mass spectrometer individually, stochastic variances in the intensity of the same species across multiple replicates can introduce uncertainty in

measurements⁶¹ or even the loss of signal between runs⁶², and must be accounted for in normalization strategies^{63,64}. Aside from potential methodological problems, LFQ can be impractical when processing many samples due to the large amount of instrument time needed to analyze each sample separately. In addition, the fractionation methodologies needed to perform deep sequencing and quantitation on LFQ samples may be difficult to do with small amounts of input sample.

In contrast to LFQ, isotope labeling mass spectrometry methods utilize labels containing stable, heavy isotopes to differentiate peptides labeled from different samples by mass signatures. One strategy is stable isotope labeling by amino acids in cell culture (SILAC), in which cell lines may be grown in media supplemented with isotopically labeled amino acids such as lysine and arginine, resulting in cells that constitutively express either normal (“light”) or stable isotope labeled (“heavy”) proteins⁶⁵. Cells labeled with “heavy” amino acids are treated or perturbed in some fashion, alongside “light” control cells after which the proteins from each cell population are isolated, digested, concatenated together, and analyzed in the same LC-MS experiment. Detected heavy or light labeled peptides have distinct masses, but similar chromatographic and ionization behavior. The heavy and light peptides are identified from their MS/MS spectra and are quantified using the AUC values from the MS1 chromatograms, minimizing the need for between-run normalizations. Comparison of the AUC values provides relative abundance measures for the peptides and inferred proteins. This technology is not limited to cell lines, as researchers can raise SILAC-labeled animals using isotopically labeled feed in animal models⁶⁶. The downsides of the SILAC methodology are the limited number of conditions

that can be tested at present⁶⁷ as well as the potential difficulty in producing cell lines or testing subjects with identical degrees of protein labeling.

Finally, another strategy uses stable-isotope labeling reagents, called isobaric tags. These tags usually are synthesized to react with the primary amines of N-termini and nucleophilic side chains of peptides, primarily lysines⁶⁸, thereby comprehensively and covalently tagging every peptide within a complex mixture generated by trypsin digestion, or potentially other proteases. The tags are isobaric, such that the overall mass added to peptides by the different labels is the same across the different samples being compared. Differentially labeled peptides are detected as a single MS1 peak by LC-MS. Relative quantities of peptides within each sample condition are determined by reporter ions that are generated from peptides selected for MS/MS analysis. Stable isotopes incorporated at different locations in the chemical tag give rise to mass differences in these reporter ions that distinguish the different samples being labeled. Comparison of their mass spectral intensity provides a relative abundance measure for each peptide subjected to MS/MS analysis and identified by sequence database searching. Identified peptides are then used to infer protein identities and associated relative abundance compared across experimental conditions. The use of isobaric tags such as iTRAQ⁶⁹ and TMT⁷⁰ allows for multiplexing samples processed across many conditions. This reduces instrumentation time needed for the analysis and increases the overall amount of digested peptides being handled due to pooling of labeled samples, allowing for the prefractionation of samples even with low amounts of material derived from each individually labeled sample. While initially limited to comparing only a few different sample conditions, current commercial isobaric labeling

strategies can multiplex as many as eighteen individual samples together⁷¹, with modifications being demonstrated for even higher levels of multiplexing⁷².

Historically, bottom-up proteomics has utilized data-dependent acquisition (DDA) mass spectrometry experiments, in which the most abundant peptides in every chromatogram peak detected in the MS1 scan are selected for fragmentation and detection in the MS/MS scan⁷³. The use of DDA experiments is still widespread and has had great utility in proteomics and multi-omics analyses⁷⁴. While this has been a largely successful approach, DDA can miss signals from very low abundance peptides in a complex sample, limiting the number of identified peptides and inferred proteins. Indeed, many studies have noted that the semi-stochastic sampling of DDA experiments results in irreproducible measurements of peptides across multiple samples^{61,62}. This can be mitigated using the isobaric labelling and fractionation strategies described previously, though not entirely. An alternative strategy to improve the depth and reproducibility of quantitative proteomics is data-independent acquisition (DIA). Here ions are continuously collected and fragmented by collecting MS/MS in overlapping m/z windows⁷⁵ across the entire range of expected peptide m/z values. Results are deconvoluted by extraction of co-eluting fragment peaks that belong to a single starting peptide detected within any m/z window. Spectral libraries of fragments derived from all the detectable peptides within a proteome are used to confirm the identity of co-eluting fragment ions. Quantification is achieved by AUC measurements of the peaks corresponding to peptide-specific fragment ions. While initially limited to DDA-generated spectral libraries⁷⁶, DIA-based bottom-up proteomics can now be performed using libraries generated using the DIA data itself⁷⁷ or wholly generated using deep learning prediction strategies^{78,79}. As an alternative to DDA, the DIA approach has

been shown to provide more accurate quantitation, with more consistent detection of peptides across samples and from low-abundance proteins, with potential for high-throughput analysis^{80,81}. Such results can greatly benefit in the quantitative aspects of multi-omic analyses centered on MS-based proteomics data.

1.2.3 Improvements in MS instrumentation

Several innovations have also been made to MS instrumentation allowing for more sensitive peptide identification, and in turn, protein identification in bottom-up proteomics. One way of increasing the depth of proteome coverage is through pre-fractionation in the mass spectrometer itself using gas-phase fractionation (GPF), which is performed by multiple injections of an individual sample using variable isolation windows covering small (100-200) m/z ranges⁸²; by doing this, the mass spectrometer isolates and examines discrete mass ranges of precursor ions and reduces the amount of potential ion suppression by co-eluting peptides. This simple technique has been shown to be powerful enough to potentially eliminate the need for LC separations⁸³ and has been put to effective use in creating deep spectral libraries for DIA experiments⁷⁷.

A notable instrumental advance is the improved scan-rates. Faster acquisition of MS and MS/MS spectra enables more comprehensive sampling of peptides in complex mixtures, increasing the depth of detection while maintaining high mass resolution⁸⁴. For example, the scan rate of the Orbitrap-based family of mass spectrometers was doubled with the introduction of ultra-high-field orbital traps in the newer QExactive HF⁸⁵ and Fusion Tribrid⁸⁶ mass spectrometers. Increased sensitivity has enabled improved performance such as detection of the entire yeast proteome in as little as one hour⁸⁷.

In recent years, ion mobility spectrometry (IMS) has been integrated into many mass spectrometers, adding a new separation function for peptide mixtures through their collision cross section (CCS). This fractionation prior to MS detection further increases the number of peptides identified in complex mixtures^{88,89}. IMS can be accomplished by different platforms engineered for compatibility with MS instruments. One of these is high-field asymmetric waveform ion mobility spectrometry (FAIMS). In FAIMS, ions are drawn into a separation chamber via a carrier gas with an alternating RF signal and an applied counter voltage (CV)⁹⁰; by varying the CV, ions can be selectively separated by their CCS and fractionated before entering the mass spectrometer. Notably, FAIMS sources have been integrated into the latest generation of Orbitrap-based mass spectrometers and have shown their utility in improving the depth of coverage in protein sequencing in short gradient runs⁹¹ and detecting low-abundance peptides⁹² primarily due to the improved quality of MS/MS data generated. FAIMS coupled with Orbitrap instruments promises to increase the number of protein identifications, important when integrating quantitative proteomics data with NGS sequencing information in multi-omics studies.

Another IMS format is trapped ion mobility spectrometry (TIMs). Here, ions are drawn into a tunnel by a gas and held in place by an applied electric field; by incrementally lowering the applied field, ions are sequentially released into the mass spectrometer in order of decreasing CCS values⁹³. This scanning and fractionation process can be made even faster with a longer tunnel and two applied electric fields, in a process called parallel accumulation-serial fragmentation (PASEF). Here, ions can be continuously trapped and released into the mass spectrometer, greatly improving the sensitivity of TIMs⁹⁴. This

technology has been coupled with a fast-scanning time-of-flight (TOF) instrument in the Bruker timsTOF instrument, proving adept at sensitive and reproducible peptide detection from even material-limited samples⁹⁵. The TOF detection should benefit multi-omics analyses by offering extremely sensitive and reproducible quantitative proteomic results, with potential for analysis of larger sample cohorts in a high-throughput format, providing more complete results when integrated with other ‘omics data.

1.3. Bioinformatic integration of proteomics and other ‘omics

As detailed previously, the generation of deep MS-based proteomics and other ‘omics data (e.g., NGS data) has now become accessible for most research laboratories. However, the task of integrating different levels of ‘omic information with proteomic data is not necessarily trivial. Fortunately, significant innovations have occurred in the development of bioinformatic software tools and platforms that address this challenge. In this latter half of the review, we explore bioinformatic methodologies and software for proteomics-based multi-omics and their applications across multiple fields of research.

1.3.1 Common problems and strategies for data integration

When integrating proteomics with transcriptomic, genomic, metabolomic, or other data, there are several challenges that must be considered and addressed. Annotation of corresponding genes and their protein products is one such challenge; for example, unsynchronized annotations of proteomic and transcriptomic data make comparisons between coding regions and their expressed protein products difficult⁹⁶. As a possible solution, the Uniprot database⁹⁷ provides a well-curated repository of characterized

proteins from diverse organisms. Entries contain annotations for proteins including unique Uniprot identifiers cross-referenced with coding gene names, and other identifiers (e.g., RefSeq, Ensembl IDs, etc) useful for matching proteins to corresponding genomic or transcriptomic sequences⁹⁷. In addition, computational tools such as biomaRt can be used to automatically map protein sequences to common genome or transcriptome sequence coordinates⁹⁸.

Integrating proteomic and metabolomic data presents a different challenge. Unlike genes and their coding sequences, metabolites are not easily mapped directly to a protein's amino acid sequence; rather, the metabolites may be mapped to those enzymatically active proteins involved in their synthesis, accumulation, excretion, or degradation, as well as those proteins with which they have allosteric interactions⁹⁹. This can be done using metabolite databases such as the Human Metabolome Database¹⁰⁰, ConsensusPathDB¹⁰¹, PathBank¹⁰²102, etc., and is a functionality of many multi-omics software packages (see below).

Another important consideration for multi-omic analysis is the normalization of the quantitative data (e.g. protein and transcript abundance values), such that dynamic response at these different levels of 'omic information can be compared directly. Common strategies for normalization include logarithmic transformation, TMM normalization¹⁰³, or normalization relative to a standard in the data. These strategies can be implemented via one's own data manipulations or through specialized software such as NormalyzerDE¹⁰⁴ and pseudoQC¹⁰⁵.

For comparing large scale MS-based proteomic results with corresponding 'omic data (e.g., transcriptome data derived from RNA-Seq analysis, quantitative MS-based

metabolomics data), several different approaches exist. Commonly, researchers will conduct their analyses considering the intersection of expressed genes and/or identified metabolites and the corresponding proteins which were confidently identified and quantified. With this intersection of the ‘omics data, similarities and differences between proteins and corresponding transcripts/metabolites in response to stimuli can be compared. Methods such as component analyses^{106,107} and hierarchical clustering¹⁰⁸ examine the altered system responses that occur under a given condition. These comparative analyses provide insights into potential mechanisms of post-transcriptional or post-translational regulation, offering a unique look at molecular signatures underlying biological function and disease. When considering the union of the complete multi-omics data (e.g., all quantified proteins compared with all quantified transcripts or metabolites), enrichment analysis is often employed on each separate set of results, revealing information on biological pathways and molecular functionalities¹⁰⁹ that may be in common or different between ‘omic domains. In addition, functional relationships between ‘omic datasets can be examined using topographical network analyses to establish changes in the expression of known clusters of genes/gene products, discover new clusters of features, and examine common regulatory elements that may be of interest across datasets¹¹⁰. When ‘omics data is collected as a part of a time-course study, modelling software can be used to establish the dynamic patterns of biomolecule abundance by calculating their kinetic parameters and identifying elements (e.g., genes, proteins, and metabolites) with similar responses¹¹¹.

1.3.2 Current software applications for integrative analysis of multi-omics results

Although the computational methods for integrating MS-proteomic and other ‘omic data are known, implementing these different algorithms presents a daunting challenge for many researchers. Fortunately, computational biologists and bioinformaticians have developed accessible software to automate these tasks and generate useful readouts to interpret this data (Table 1.1). Given these developments, the challenge of 21st century systems biologists engaged in multi-omic analyses is not to find suitable software for their purposes, but to decide which software tools among many will most suit their purpose. To this end, in the section we offer some insights into software with high value for MS-based proteomics centered multi-omics (Table 1.1). While our listing of software is not exhaustive, those shown have been selected either through our own experience or through in-depth exploration of available tools, to select those with the most promise for multi-omic applications. We hope this serves as a starting point for researchers entering MS-based proteomics-centric multi-omic studies.

Table 1.1. A selection of bioinformatic tools for proteomic analyses.

Software	Data Types	Functionality	Language	Reference
PANTHER	gene features (data agnostic)	functional analysis	R	Mi et al. ¹¹²
gProfiler	gene features (data agnostic)	functional analysis	R	Raudvere et al. ¹¹³
reSTRING	gene features (data agnostic)	functional analysis	R	Manzini et al. ¹¹⁴
MOGSA	proteome, transcriptome	functional analysis	R	Meng et al. ¹¹⁵
WCGNA	proteome, transcriptome	network analysis	R	Langfelder et al. ¹¹⁶
STRINGdb	gene features (data agnostic)	network analysis	R	Szklarczyk et al. ¹¹⁷
MONGKIE	proteome, phosphoproteome, transcriptome	network analysis	Java	Jang et al. ¹¹⁸
moCluster	proteome, transcriptome	data clustering	R	Meng et al. ¹¹⁹
mixOmics	data agnostic	data clustering, data correlation, network analysis	R	Rohart et al. ¹²⁰
STATegRa	data agnostic	component analyses, functional analysis	R	Planell et al. ¹²¹
iOmicsPASS	genome, proteome, transcriptome	network analysis, functional analysis	C++, R	Koh et al. ¹²²

netOmics	metabolome, proteome, transcriptome	network analysis, functional analysis	R	Bodein et al. ¹²³
QuanTP	proteome, transcriptome (data agnostic)	heirarchical clustering, differential analysis, multivariate analysis	R	Kumar et al. ¹²⁴

Functional analyses, focused on revealing enriched biochemical processes indicated by ‘omics results, are a key aspect of multi-omics analyses. Several software tools have been created to perform this functionality, including gProfiler¹¹³, GOATOOLS¹²⁵, and reString¹¹⁴. Functional analysis tools like these are generally written with a single set of ‘omics data in mind (e.g., genes, proteins, metabolites), such that analysis is done for each separate ‘omic data set, with comparison of end results across the different levels of information; an exception to this is MOGSA, which was purpose-built to do gene-set analyses on multi-omics data¹¹⁵.

Topographic network analysis of multi-omics data can yield important information about clusters of molecular features that undergo systemic changes in response to stimuli, and for this reason many applications have been created for this purpose. The WGCNA¹¹⁶ package in R was designed to perform many aspects of weighted gene correlation network analysis on transcriptomic data, though it can also be used to analyze multiple sets of disparate ‘omics data¹²⁶. Another package that has been found to be useful is the MONGKIE package, providing visualization capabilities of complex multi-omics networks, enabling easier interpretation of results¹¹⁸.

Clustering MS-based proteomics data with other ‘omics data (most commonly quantitative transcriptomic data) illuminates potential mechanisms of regulation and response to stimuli. For such analysis, it is necessary to employ a clustering of clusters algorithm¹²⁷ which first clusters the individual ‘omics data, then clusters the clusters together to identify overarching patterns in multi-omics data. The package moCluster¹¹⁹ is an especially useful iteration of this strategy, as it is able to perform clustering analysis on multiple levels of ‘omics data in a fraction of the time of similar packages. Many of these

multi-omics software packages are in fact a suite of different algorithms packaged together into a single integrated tool. The mixOmics package¹²⁰ represents an exhaustive option for supervised multi-omic analyses, being capable of analyzing individual ‘omics datasets, multiple ‘omics datasets containing measurements of the same features using the DIABLO package¹²⁸, or meta-analyses of multiple instances of a single ‘omics analysis using the MINT package¹²⁹. Datasets in mixOmics are uploaded as pre-normalized matrices containing rows of features (e.g., genes, proteins etc.) and columns of conditional values with a categorical column containing meta-data on the system of interest. Up to three different datasets can be analyzed together, outputting clustering results, correlation analyses, and network analyses, among other possibilities. In addition, tutorials for this software are readily available at mixomics.org.

Another R package, STATegRa¹²¹, is wholly agnostic to the kind of ‘omics input and can accommodate multiple datasets. This package was developed through the STATegra consortium, an international effort to generate statistical analysis tools for ‘omics data¹³⁰. The input datasets for STATegRA also require pre-normalization as well as categorical metadata concerning their status as control or case experimental data. Each of the datasets is first subjected to quality control analyses, followed by joined component analyses of sets of two datasets to determine the ‘omics pairing that has the most significant relationship to the condition of interest. These two datasets are then subjected to nonparametric combination¹³¹ to increase their statistical power and determine the features of both datasets that have the most significant bearing on the condition of interest; these features are ultimately subjected to functional analysis via gene-set enrichment analysis.

A notable, recently described platform is iOmicsPASS¹²². This platform is unique in that it utilizes proteomics data, transcriptomics data, databases of transcription factor interactions and protein-protein interactions, and conditional metadata to determine the presence of subnetworks within the data which can then be scored with a pathway enrichment module for networks that are significantly enriched or depleted under varying conditions¹²². Ultimately, iOmicsPass yields both topographic networks of interacting biomolecules that are enriched and depleted, as well as functional analyses on these pathways to reveal the changes these networks are affecting in response to stimuli. A similar software package is netOmics, which was designed to process multiple ‘omics datasets over extended periods of time¹²³. Unlike other bioinformatics packages, netOmics uses raw data as inputs, which it can pre-process before analyses. Using the timeOmics algorithm¹³², netOmics selects models for each molecule detected to establish their changes over time, after which it creates networks to show linkages between them using protein-protein interaction networks and KEGG pathway databases. Ultimately, the researcher is left with multi-omic interaction networks as well as functional enrichment analysis results over the course of the experiment.

The multi-omic analyses performed depend largely on the background of the researcher, and the ‘omic data types gathered as a part of the experiment. The functional analysis and topographical analysis tools detailed in the initial portion of this section were developed for use with individual datasets of genes; aspiring bioinformaticians interested in using these tools for analyzing their proteomics data integrated with other ‘omics data can use these on the intersections of proteomics and other ‘omics datasets representing a relationship of interest (i.e., shared significant changes in abundance.) In addition, using

these tools as a part of a larger series of analyses may require a level of coding sophistication to input the desired parameters, submit the data and run the analysis; researchers who are less experienced in crafting and running scripts may then prefer the platforms with multiple functionalities, especially newOmics, as this platform allows for the input of raw data without a priori normalization or other processing on the part of the researcher. Other platforms that allow for automated queuing of tools as workflows may be of use to researchers with limited bioinformatic or programming experience (see below).

1.3.3 User-friendly multi-omics platforms for increased access and flexibility

Many software applications capable of MS-based proteomics-centered multi-omics analysis were developed as a stand-alone script or bundled package in R, Python, or C++ which are run through the command line or through an interpreter program. While this is not a problem for the skilled bioinformatician, many researchers who are less computationally-savvy are hindered by these software implementations. As such, many multi-omic software suites incorporate point-and-click graphical user interfaces (GUIs) that are user friendly and accessible to a wider range of researchers (**Table 1.2**). While there are some commercial options, such as Qiagen's Ingenuity Pathway Analysis (IPA)¹³³, there are a myriad of open-source options that are as powerful and simple-to-use as they are affordable.

Table 1.2. Multi-omics platforms with graphical user interfaces for ease of use.

Suite	Data Types	Functionality	Reference
Galaxy	data agnostic	function agnostic	Jalili et al. ¹³⁴
Perseus	proteome (data agnostic)	statistical analysis, functional analysis, network analysis	Tyanova et al. ¹³⁵
OpenOmics	genome, proteome, transcriptome, epigenome	network analysis, functional analysis	Tran et al. ¹³⁶
multiSLIDE	data agnostic	hierarchical clustering, differential analysis	Ghosh et al. ¹³⁷
MiBiOmics	data agnostic	network analysis	Zoppi et al. ¹³⁸

Although useful, stand-alone software has some limitations related to multi-omic data analysis. Scalability to handle the processing and memory requirements of large volume data and the ability to integrate disparate software for automated analysis of data from across 'omic domains are at the forefront of these limitations. To address these issues, bioinformatic workflow platforms have emerged. The Galaxy platform¹³⁴ is an open-source bioinformatics platform where bioinformatic tools can be integrated into automated workflows, implemented on powerful high-performance computing infrastructure, and accessed via a user-friendly GUI, designed for wet-bench researchers. Galaxy contains hundreds of open-source software developed for analysis of NGS sequencing data, along with numerous tools for interpretation of results such as DAVID¹³⁹, KEGGREST¹⁴⁰, and KOBAS¹⁴¹. Through collective work of our lab and a global network of others, the Galaxy for proteomics (Galaxy-P) project has implemented numerous tools for MS-based proteomics informatics into the platform, making it an ideal environment for multi-omic analysis. One example of a Galaxy-P tool is QuanTP, which can perform hierarchical clustering and differential analysis on quantitative proteomics and transcriptomics data, in addition to plotting the fold changes of features in the proteomic data against the transcriptomic data to examine the linear relationship between these results¹²⁴. QuanTP can also identify genes and the corresponding proteins that are discordant in their quantitative response, in addition to performing k-means clustering to determine clusters of discordant transcripts and proteins that may be regulated post-transcriptionally. Another multi-omics tool currently available in Galaxy is OpenOmics, a Python library and multi-omic workspace that interfaces with public 'omics databases and can accommodate proteomics, transcriptomics, genomics, and epigenomics data¹³⁶. Finally, Galaxy provides a suite of

metabolomics data analysis tools¹⁴², enabling analysis of metabolite data within the same environment and integration of results with other ‘omics data.

One platform which shows promise for multi-omic analysis is Perseus, the open-source matrix manipulation software developed for analysis of MS-based proteomics data¹³⁵. While developed for proteomics data, the software itself is data agnostic by design and has recently been updated to allow for R and Python scripts to be run within the software, and to enable access to Bioconductor, Conda, and other software repositories^{143,144}, making this a potential entrée for proteomics researchers into multi-omic analyses. For researchers who prefer heatmaps to other forms of data visualization, the multiSLIDE web application creates two heatmaps from raw tabular datasets and allows for a birds-eye visualization of both datasets simultaneously, as well as a direct comparison of a gene in two datasets using the lines that connect shared measurements between datasets¹³⁷. Finally, the MiBiOmics platform is a new web application that accepts up to three sets of ‘omics data and performs individual data processing steps on each dataset, individual data explorations in the form of component and network analyses and integrates the results together to give multi-omic networks, co-inertia plots, and hive plots to show relationships between the different datasets¹³⁸.

1.3.4 Prominent examples of proteomics-based multi-omics

In the context of describing the technologies that are enabling MS-based proteomics-centric multi-omics, it is worth pointing out some success stories in the application of this still maturing approach. We present here five exemplary studies, which

represent the potential of multi-omics centered on MS-based proteomics data to impact diverse fields of biological research (**Table 1.3**).

Table 1.3. Prominent examples of proteomics-centered multi-omics.

Study	'Omics technologies used	Application
Cavalli et al. ¹⁴⁵	proteomics, single-nuclei transcriptomics, chromosome conformation (epigenomics)	Regulatory networks for genes involved in hepatocellular carcinoma
Fornecker et al. ¹⁴⁶	proteomics, transcriptomics	Biomarkers for drug resistance in B-cell lymphoma
Alcazar et al. ¹⁴⁷	proteomics, transcriptomics, metabolomics, lipidomics	Biomarkers for the development of Type 1 Diabetes Mellitus
Lee et al. ¹⁴⁸	Proteomics, transcriptomics, metabolomics	Molecular mechanisms of PFOS toxicity
McLoughlin et al. ¹⁴⁹	Proteomics, transcriptomics, metabolomics	Nutrient stress reactions of maize

The use of multi-omics in biomedical research represents an especially ripe opportunity for multi-omic analysis, as the high levels of information provide a holistic picture of molecular underpinnings of health and disease. An excellent example of this is a study by Cavalli et al.¹⁴⁵, in which proteomics is integrated with single-nuclei transcriptomics and chromosomal conformation changes to demonstrate the regulatory networks at play during the onset of hepatocellular carcinoma (HCC). Multi-omic analyses are particularly useful in discerning biomarkers for diseases, as in Fornecker et al.¹⁴⁶ where drug resistance in B-cell lymphoma was investigated via multi-omics to reveal increased abundances of Hexokinase 3, S100 proteins, and others as drivers of this phenotype. Similarly, Alcazar et al. were able to determine through multi-omic integration of plasma sample data that inhibition of miRNA Let-7a-5p and increased activation of the inflammatory pathway proteins makes patients more prone to the development of Type 1 diabetes¹⁴⁷. The use of proteomics-based multi-omic analyses is not limited to biomedical research, having utility in ecotoxicological and agricultural studies. Integration of proteomics, transcriptomics, and metabolomics enabled Lee et al.¹⁴⁸ to show the molecular mechanisms of perfluorooctanesulfonic acid (PFOS) neurotoxicity in zebrafish, while McLoughlin et al. demonstrated the autophagic pathways that occur in maize in response to nutrient deprivation using multi-omic analysis¹⁴⁹.

1.4 Proteogenomics: genome- and transcriptome-driven proteomics

The nature of data analysis for bottom-up MS-based proteomics, coupled with the proliferation of NGS technologies for DNA and RNA sequencing, has given rise to proteogenomics - a multi-omics approach unique to the integration of data from these 'omic

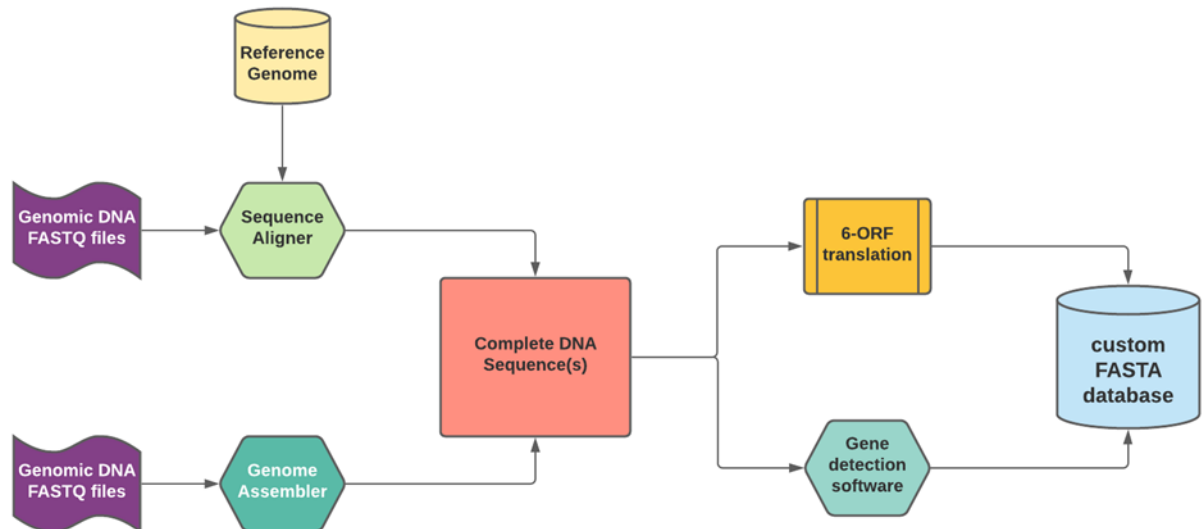
domains^{150,151}. In bottom-up proteomics, MS/MS peptide data is searched against a FASTA-formatted database containing sequences encompassing the proteome or proteomes of interest. In its traditional implementation, this approach relies on the a priori selection of a reference sequence database which may not include some sample-specific sequences of potential significance. For example, alternative splicing and amino acid substitutions are known to be the underlying etiology of many cancers¹⁵², and these sample-specific sequences may not be present in reference databases. The proteogenomics approach addresses these limitations by employing NGS sequencing of DNA or RNA within the biological sample of interest to generate a sample-specific sequence database which captures potentially translated, novel protein sequences derived from variant gene sequences and/or novel transcription and RNA processing events. MS/MS data are then searched against a combined database of both the reference and novel protein sequences of interest to gain conclusive evidence on the expression of unique proteins sequences that may play a key role in biology or disease.

Proteogenomics is generally conducted in some variations of the workflow presented in **Figure 1.2**. This workflow fuses algorithms traditionally used for specific 'omic domains (DNA/RNA sequencing and MS-based proteomics), also incorporating a number or customized tools necessary to integrate different datatypes and visualize outputs¹⁵³. A proteogenomics workflow begins with the alignment of DNA or RNA sequencing data for comparing it against a reference genome. This can be done using available open-source data or, ideally, from DNA or mRNA samples isolated from the same sample analyzed by MS-based proteomics. In the case of whole genome or exome sequencing data, the sequences can be either mapped against reference genomes using

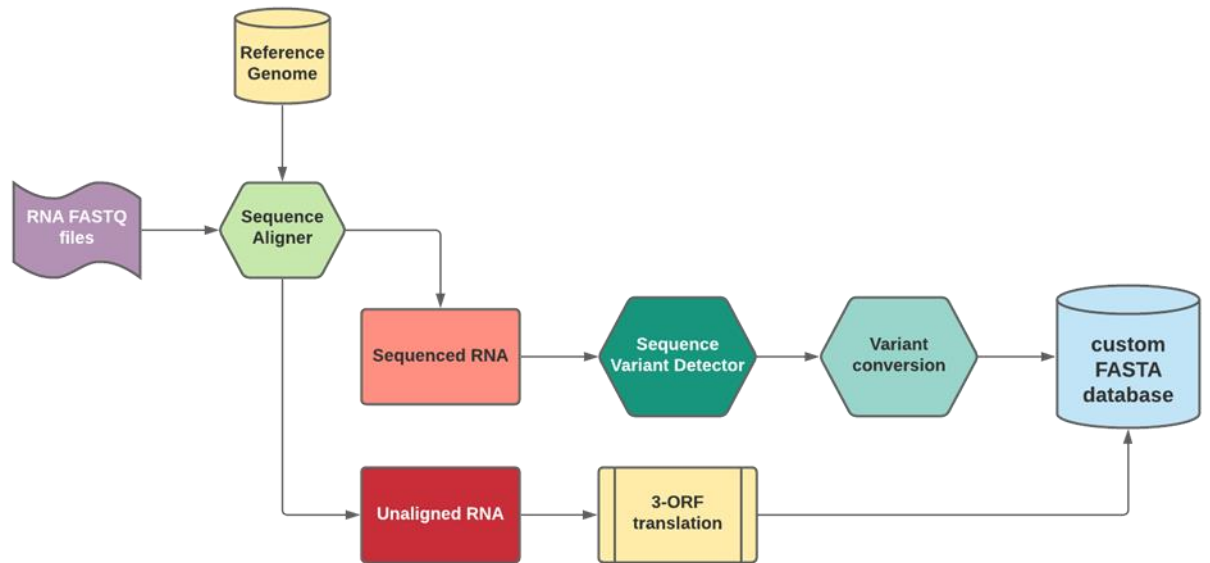
programs such as Bowtie¹⁵⁴, Minimap2¹⁵⁵, BWA¹⁵⁶, or can be assembled de novo into contigs and then whole genomes using tools such as Velvet¹⁵⁷, SGA¹⁵⁸, or others depending on the read length of the DNA. The sequenced genome can then be subjected to either 6-frame translation to potential proteins¹⁵⁹ or processed using protein prediction software such as Peptimapper¹⁶⁰ or getorf in EMBOSS¹⁶¹. For most, RNA-Seq data on expressed transcripts is a popular choice, as it provides a template of transcribed sequences that may give rise to the translated proteome. Here, sequencing data is aligned against a reference genome using programs like HiSat2¹⁶² or TopHat¹⁶³, followed by detection of variants and other novel transcripts using programs such as FreeBayes¹⁶⁴ or GATK¹⁶⁵. Novel RNA sequences can then be converted to protein sequences using programs such as CustomProDB¹⁶⁶. A recent alternative is the Spritz Database engine¹⁶⁷, which takes in raw FASTQ sequences and a reference proteome and generates a FASTA library containing non-canonical sequences.

Figure 1.2. Proteogenomics workflows. a) For generating a FASTA library from genomic sequencing data, the FASTQ files are either aligned against a reference genome or assembled into contigs and then a working genome. In either case, the resulting assembled sequencing data is either translated into proteins in six open reading frames or submitted to analysis using gene identifying software, the results of which are translated into proteins. b) For generating a FASTA library from RNA sequencing data or exome sequencing data, FASTQ files are aligned to a reference genome. The assembled data is then searched against a variant sequence detector, the results of which are then converted into a FASTA library. Unaligned sequences from UTR transcription or novel RNA processing events are subjected to three-frame translation. c) Raw proteomic data is searched against bespoke FASTA libraries to detect non-canonical peptide sequences. Flowcharts made using lucidchart.

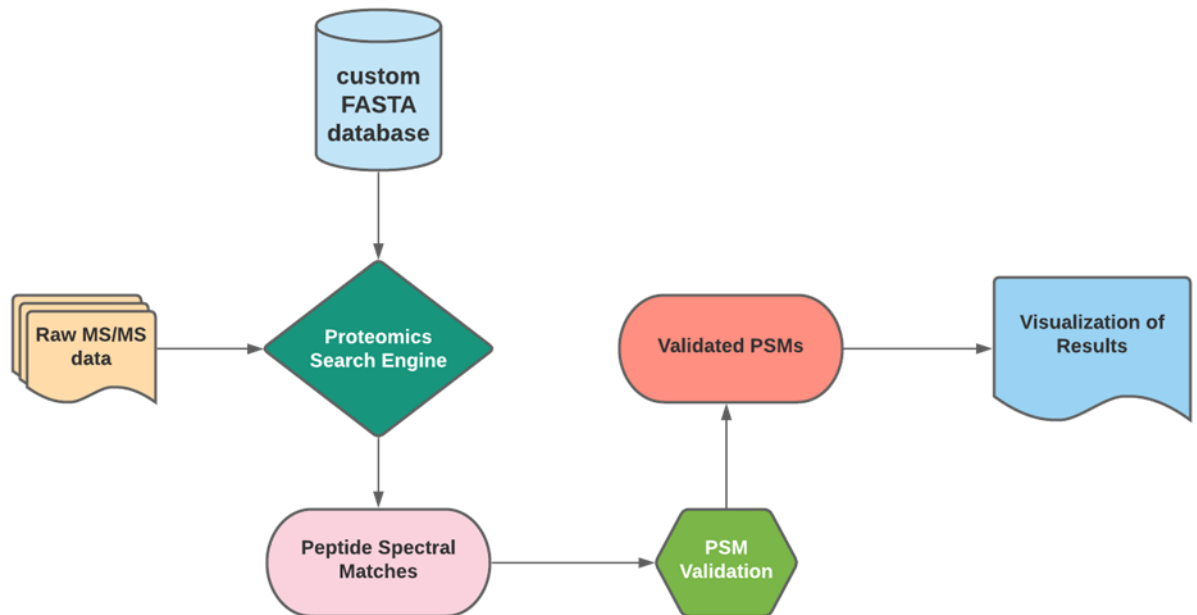
a)



b)



c)



Once a proteogenomics FASTA database is created, it can theoretically be used to query raw MS data using analysis software such as MaxQuant⁵⁹, SearchGUI¹⁶⁸, or many others. However, FASTA databases generated from genomic and transcriptomic data have the potential to be much larger than those of the conventional proteome, inviting the potential for increased false positive identifications¹⁶⁹. While this can be controlled for via more stringent false discovery rate (FDR) cutoffs during analysis, this in turn can result in decreased sensitivity as genuine identifications are removed along with false positives; this can be mitigated using strategies to decrease the database size such as two-step searching¹⁷⁰ as well as database sectioning and enrichment strategies¹⁷¹. Another concern is that potential non-canonical peptides matched to the proteogenomics database may be mismatches that correspond to normal peptides; tools such as BLAST-P¹⁷² and the PepQuery search engine¹⁷³ can be employed to ensure confidence in candidate novel peptide sequences identified via proteogenomics. When non-canonical peptides are identified in proteogenomics assays, it is useful to examine their differences to the canonical sequence by mapping them to the genome. This can be done using tools such as the Multi-omics Visualization Platform (MVP)¹⁷⁴ or the Proteogenomic Mapping Tool¹⁷⁵.

Due to the extensive number of tools necessary for proteogenomics analyses, conventional command line triggering of bioinformatics tools is a very cumbersome process. Multi-omics workflow platforms such as Galaxy, Peptimapper¹⁶⁰, or PANOPLY¹⁷⁶ allow for the automated generation of proteogenomics databases, searching mass spectrometry data against these databases, and statistical analysis of the results.

1.5. Projected future applications of proteomics-centered multi-omics

The value of multi-omics centered around MS-based proteomics data has been demonstrated in recent years. High profile studies via the Clinical Proteomics Tumor Analysis Consortium (CPTAC) have been notable examples¹⁷⁷, along with other biomedical studies^{178,179}. These approaches have also contributed significantly to research progress combating the ongoing SARS-CoV-2 pandemic^{180,181}. The technologies contributing to these multi-omic studies (NGS sequencing, high-resolution MS) have become ubiquitous and are now accessible to not only basic laboratory researchers, but also in translational and clinical settings^{182,183}. Thus, we are poised to usher in a new era of precision medicine which may bring together these multi-omic technologies to determine the best course of action for therapeutic interventions and increase the possibility of high value diagnostic and prognostic biomarkers. The outputs of these multi-omic studies (proteins and/or metabolites) nicely feed downstream clinical assays based on targeted MS methods¹⁸⁴, capable of sensitive, rapid, and accurate quantitative analysis across large patient cohorts. However, to realize the potential of MS-based proteomics centered multi-omics for clinical translation, advances are still needed. The processing and analysis of the raw data needs to be simplified, so that biologists and clinicians with minimal backgrounds in computer science, and limited time, can efficiently perform these analyses and generate reports with clear outcomes and suggested actions. Much of this review has discussed the bioinformatics suites with GUI interfaces which require no coding experience per se; continuing to develop such software platforms, with input from clinical partners, will be critical to making and keeping multi-omics a regular part of the lab and the clinic. Another challenge needing a solution is the incorporation of patient metadata

into the multi-omic workflows deployed for analysis of this data, although efforts and progress is being made on this front¹⁸⁵. Additionally, although the depth of MS-based proteomics has significantly improved in recent years, improvements in sequence coverage to identify novel proteoforms, possibly via “middle-down” approaches¹⁸⁶, could increase the value of information from proteogenomics. Lastly, single cell proteomics has lagged genomic and transcriptomic approaches for analyzing these highly valuable, material limited sample types, with high potential for advancing biomedicine. Promising methods by a few specialist labs^{187,188} offer hope, but these need to be proven reliable for use by the broader community.

Another emerging area that fits in the scope of MS-based proteomic-centered multi-omics is the field of metaproteomics¹⁸⁹. Metaproteomics incorporates metagenome information on microbial communities from a wide-variety of settings - from human host samples to complex samples (e.g., wastewater, soil) relevant to environmental studies. This multi-omic data can be used to create large protein sequence databases of potential microbe-derived proteins within these samples, which are then used for searching MS/MS data generated from these samples. When analyzed with specialized multi-omic tools^{190,191}, the results provide a unique snapshot of the functional proteins expressed by microbial communities which may drive host biology or regulate characteristics of complex ecological systems. These results can also help identify potential metabolic pathways and small molecules generated by the microbiota that play a role in interactions and regulatory mechanisms. Metaproteomics also expands the reach of proteomic-centered multi-omics to studying flora, fauna and microbial communities responding to environmental factors

(e.g., climate changes, pollution¹⁹², bioremediation¹⁹³) in addition to biomedical applications^{194,195}.

The continuous progress in proteome-centric multi-omics points to a promising future where this approach becomes routine. Portable mass spectrometers for deployment both in the field and in the clinic^{196,197}, coupled with automated and portable sample collection and processing devices^{198,199} could make sampling and MS-based proteomics analysis in the field and clinic a reliable option, complementing such approaches that are already emerging for NGS sequencing^{200,201}. Continued advances in multi-omic software platforms towards customized and automated pipelines would rapidly provide results from the generated data, aiding clinical decisions or guiding mitigation actions for environmental applications.

1.6 Concluding remarks and thesis goals

Bottom-up proteomics holds a valuable place within the hierarchy of ‘omics technologies, directly detecting the functional molecules that collectively drive biochemical mechanisms within a cell, tissue, or organism. While informative, proteome data is only one piece of the network of interconnected biomolecules responsible for cellular function and phenotypes. Integration with DNA or RNA sequencing information that may give rise to translated proteins, or metabolite information which indicates their biochemical state provides a more complete picture. Recent advances to bottom-up MS-based proteomics methodologies and instrumentation now makes deeper characterization of the proteome a reality, improving the value of integration with other ‘omic data (e.g., DNA/RNA sequencing results). At the same time, bioinformatics tools have emerged to

facilitate the analysis of large ‘omics data sets, including options for integration of MS-based proteomics data with other ‘omic levels of information. The integration of genomics and/or transcriptomics data with deep MS-based proteomics datasets has given rise to the area of proteogenomics, which offers promise in detecting previously unseen protein sequences belonging to proteoforms that may be key to biological processes and disease. As advances continue to make MS-based proteomics more cost-effective, sensitive and high-throughput, multi-omic analyses centered around this data have the potential to become a pillar of 21st century systems biology-based research -- impacting diverse fields from translational clinical applications to the study of complex environmental phenomena.

Given the great variety of proteomics-based multi-omics strategies available, we sought to apply these technologies to instances of disease, inflammation, and exposure to electrophilic contaminants. In using open-source mass spectrometry datasets from *in vitro* experiments and COVID19 patients as well as sophisticated bioinformatics workflows, in Chapter II of this Thesis we were able to determine the optimal targets for the detection of SARS-CoV-2 in nasopharyngeal swabs using mass spectrometry. Beyond using bioinformatics alone, in Chapter III we were interested in performing our own mass spectrometry analyses and applying RNA-Seq-based proteogenomics to a murine model of inflammatory bowel syndrome to catalog changes in protein abundance following long-term exposure to bacterial infection as well as detect and validate peptides with non-canonical sequences in these samples, as we hypothesized that these could serve as early biomarkers of oncogenesis. Building on this, in Chapter IV we exposed mice to lipopolysaccharide (LPS) or cigarette smoke (CS) for variable amounts of time before isolating their type II pneumocytes and subjecting them to proteomic analysis, with an aim

to characterizing the proteomic phenotypes of LPS-driven inflammation and comparing them with cigarette smoke exposure, eventually integrating these data with epigenomic and transcriptomic data from these samples. Finally, In Chapter V we expanded proteomics into the field of adductomics by validating a novel 4-hydroxybenzyl N-terminal adduct in hemoglobin and expanding its detection to the nucleophilic side chains within this protein, as well as comparing the ability of bottom-up proteomics to detect electrophilic adducts in hemoglobin with the FIRE protocol, a procedure in which N-terminal amino acids are isolated and analyzed via LC-MS with an aim towards characterizing the adducts formed at this site.

II. A RIGOROUS EVALUATION OF OPTIMAL PEPTIDE TARGETS FOR MS-BASED CLINICAL DIAGNOSTICS OF CORONAVIRUS DISEASE 2019 (COVID-19)

Adapted from:

Rajczewski AT, Mehta S, Nguyen DDA, et al. A rigorous evaluation of optimal peptide targets for MS-based clinical diagnostics of Coronavirus Disease 2019 (COVID-19). *Clin Proteomics*. 2021;18(1):15. Published 2021 May 10. doi:10.1186/s12014-021-09321-1.

This work was performed in collaboration with Subina Mehta, Dinh Duy An Nguyen, Dr. Björn A. Grüning, James E. Johnson, Dr. Thomas McGowan, and Dr. Timothy J. Griffin and under the direction of Dr. Pratik D. Jagtap. Andrew T. Rajczewski, Subina Mehta, Dinh Duy An Nguyen, and Dr. Pratik D. Jagtap performed constructed peptide libraries and validated targets bioinformatically and manually. Andrew T. Rajczewski performed bioinformatic comparisons of target peptides between SARS-CoV-2 and other coronaviruses. Dr. Björn A. Grüning, James E. Johnson, and Dr. Thomas McGowan installed and maintained the bioinformatic tools in Galaxy EU and Galaxy MSI. Andrew T. Rajczewski, Subina Mehta, and Dr. Pratik D. Jagtap constructed figures. Andrew T. Rajczewski wrote and edited the manuscript under the guidance of Drs. Timothy J. Griffin and Pratik D. Jagtap.

2.1. Introduction

In the latter half of 2019, a pneumonia-like disease arose in the Wuhan Province of China²⁰². Subsequent analysis showed the cause to be a betacoronavirus initially called 2019-novel coronavirus (2019-nCoV). This disease soon spread throughout the world and came to be known as coronavirus disease 2019 (COVID-19) with the clinical classification Sudden Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). As of the writing of this manuscript, there are over 616 million patients infected world-wide with COVID-19, with a current global death toll sitting at over 6.5 million people²⁰³. Patients report a litany of symptoms, ranging from fever, cough, and muscle aches in mild cases to acute respiratory distress syndrome (ARDS), multiple-organ failure, and death in the most severe cases^{204,205}.

While the development of therapeutic treatments for infected patients^{206,207} and the eventual development of vaccines against SARS-CoV-2^{208,209,210} are of great importance for the management of this disease, rapid and effective diagnosis of COVID-19 infection has been and continues to be of primary importance. Most testing strategies used in the diagnosis of active COVID-19 infections utilize quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) of viral RNA in samples collected from patients^{211,212}. Rapid COVID-19 testing is generally performed on readily accessible patient-derived samples with high viral loads, such as nasopharyngeal swabs and saliva. To improve turnover time and increase the volume of tests that can be performed, innovations in RNA-based testing have been introduced to cut down on the time required. Testing protocols have been developed that eschew the isolation of RNA from patient samples, allowing for much faster RT-qPCR analyses²¹³. In addition, techniques such as

Reverse Transcription Loop-mediated isothermal AMPLification (RT-LAMP)²¹⁴ and Specific High Sensitivity Enzymatic Reporter UnLOCKing (SHERLOCK)²¹⁵ diagnostics allow for rapid point-of-care detection of SARS-CoV-2 RNA without the need for sophisticated training in PCR.

While these techniques are generally fast and highly specific for viral RNA, improper sample collection, storage, or processing could result in the degradation of RNA yielding potential false negative tests. In addition, their reliance on sequence amplification using reverse transcriptases and DNA polymerases introduces the potential for false negatives through the inhibition of these enzymes by components of the sample^{216,217}. Due to the better chemical stability of proteins compared to RNA, as well as the lack of a need for intermediary enzymes and signal amplification via PCR, clinical proteomics has emerged as a potential supplemental test for the diagnosis of COVID-19 through direct detection of viral peptides via LC-MS²¹⁸. Specifically, targeted methods such as selected reaction monitoring (SRM) and parallel reaction monitoring (PRM) to detect peptides specific to the virus could be most useful in a clinical setting^{219,220}. However, not all the potential viral peptides derived from SARS-CoV-2 infection are equally suitable as targets, based on well-known limitations of targeted LC-MS methods for proteomics; some tryptic peptides of SARS-CoV-2 could have intrinsic physicochemical properties limiting their reproducible detection in a mass spectrometer, as well as co-elution from the LC with more abundant peptides that mask their presence in the sample. In addition, proteomics software can sometimes make putative peptide spectrum matches (PSMs) with spectra that are of poor quality, making for uncertain identification of peptides of interest^{221,222}. Additionally,

a key requirement for targeting peptides for virus detection is that these are specific to the SARS-CoV-2 virus, with no potential overlap with other coronaviruses or other organisms.

In order to evaluate the most robustly detectable SARS-CoV-2 peptides and make the detection of these viral peptides in human samples in a clinical setting all the more feasible, we set out to examine proteomic datasets from three cell culture-based studies^{223,224,225} and seven clinical studies^{226,227,228,229,230}. We utilized automated workflows implemented in the Galaxy platform and made accessible via the European Galaxy public instance to first identify as many SARS-CoV-2 peptides possible in all samples, creating a master list of SARS-CoV-2 peptides identified across the samples. We then interrogated these peptides using the PepQuery search engine¹⁷³ to confirm the quality of these PSMs and determine whether the matched sequences were unique to SARS-CoV-2 or could be better ascribed to the human proteome or that of another closely related coronavirus. Peptides and their associated PSMs which survived this rigorous filtering were then manually validated using the Multi-omics Visualization Platform¹⁷⁴ and further analyzed for specificity to the SARS-CoV-2 virus via BLAST-P¹⁷² and MetaTryp²³¹. Taken together, our analyses enable the construction of a high-confidence target peptide list that would form the basis of a targeted clinical proteomics assay for SARS-CoV-2 infection.

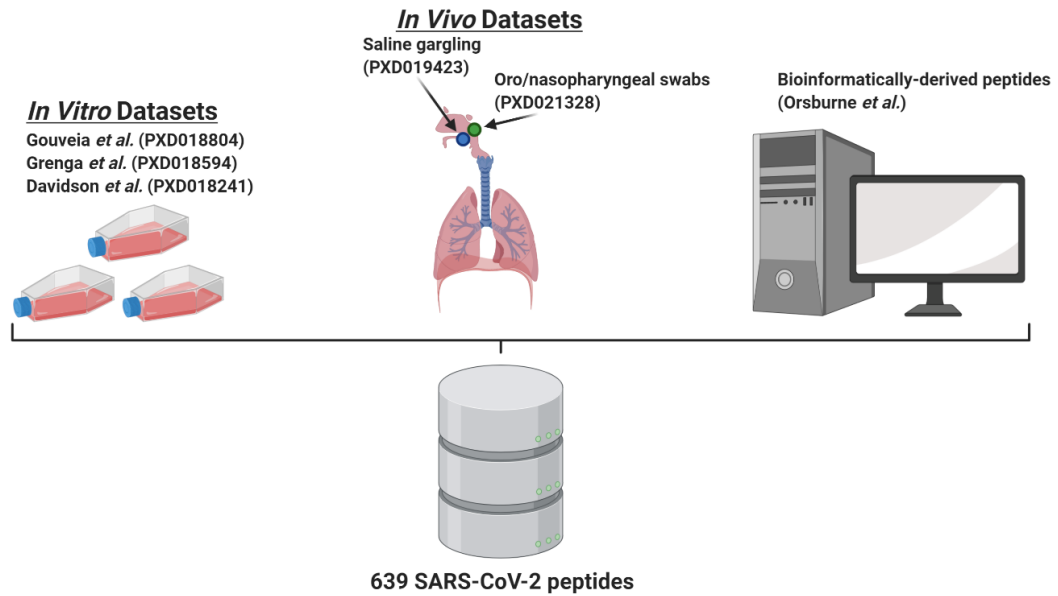
2.2. Materials and Methods

Case Studies

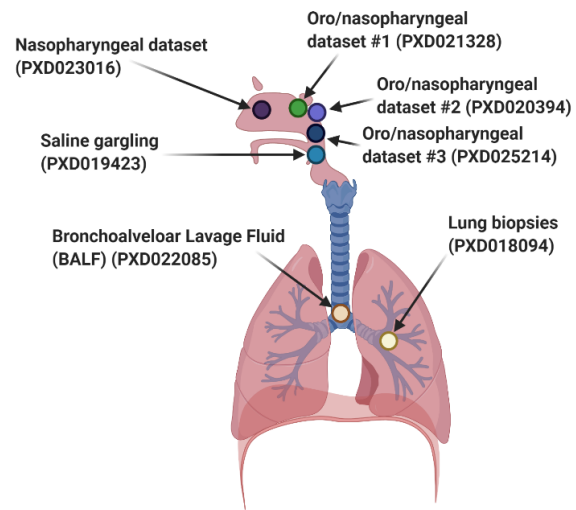
For establishing workflows to evaluate virus-specific peptides, three published cell culture datasets^{223,224,225} which used SARS-COV2 infected Vero cell lines were chosen (**Figure 2.1a**), along with five clinical datasets^{227,228,229,232,233} (**Figure 2.1b**).

Figure 2.1. MS/MS datasets used in the determination of optimal SARS-CoV-2 peptides for COVID-19 diagnosis. a) Cell culture, clinical, and bioinformatic datasets used to generate the SARS-CoV-2 peptide panel. b) Clinical datasets queried using the initially characterized peptide panel from a) to determine the feasibility of COVID-19 diagnosis via targeted proteomics as well as determine the optimal peptide targets for those assays. Figures were made using BioRender.

a)



b)



Cell Culture Datasets

Gouveia et al. published a dataset (PXD018804) with SARS-CoV-2 infected Vero cells from *Chlorocebus* primates to generate a high-resolution mass spectrometry dataset. The second dataset was published by Grenga et al. (PXD018594) wherein a seven-day time course shotgun proteomics study was performed on Vero E6 cells infected by Italy-INMI1 SARS-CoV-2 virus at two multiplicities of infection. The third cell culture dataset chosen was published by Davidson et al. (PXD018241), which also utilized Vero E6 cells to investigate the viral transcriptome and proteome.

Clinical Datasets

The first clinical dataset chosen was from the study by Cardozo et al. (PXD021328), wherein they collected bottom-up mass spectrometry (MS) data on combined oropharyngeal and nasopharyngeal samples from ten COVID-19 positive patient samples. A second clinical dataset was from the Ihling group (PXD019423) to detect SARS-CoV-2 virus proteins from saline gargle samples of COVID-19 infected patients. The third dataset was obtained from the Rivera group (PXD020394) comparative quantitative proteomic analysis from oro- and naso-pharyngeal swabs used for COVID-19 diagnosis was performed. Further, unanalyzed oro/nasopharyngeal data from Cardozo et al.²²⁶ (PXD025214) as well as a nasopharyngeal swab dataset from Bankar et al.²³⁴ (PXD023016) were interrogated for the presence of our proposed targets. Datasets derived from COVID-19 patient lung biopsies (PXD018094) and bronchoalveolar lavage fluid (BALF) (PXD022085) were analyzed to determine the utility of our workflow to identify SARS-CoV-2 in clinically relevant sample types.

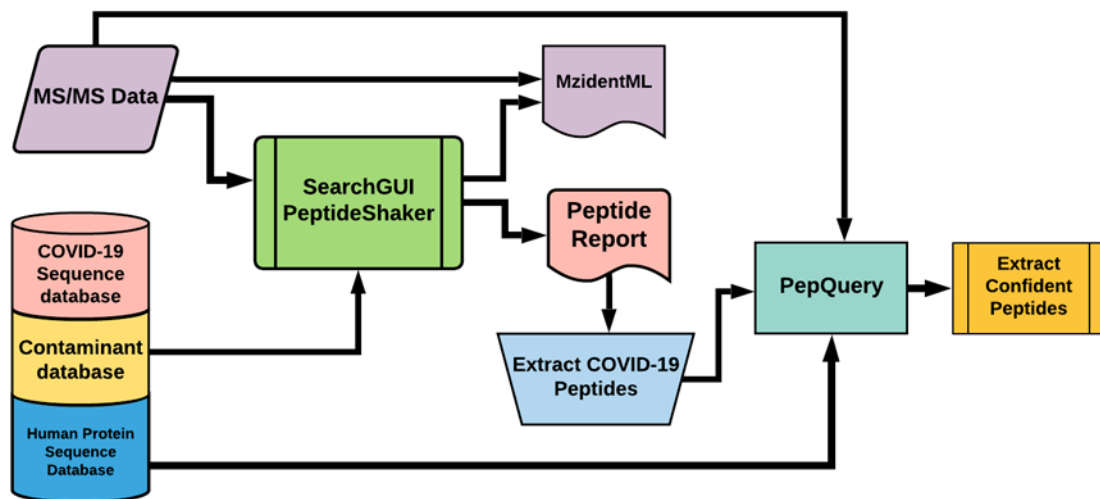
Sequence Database Searching

The Galaxy workflow for peptide identification (**Figure 2.2a**) includes conversion of RAW data to MGF and mzML format. In case of the cell culture study, the MGF files are searched against the combined database of *Chlorocebus* sequences, contaminant proteins (cRAP) and SARS-Cov-2 proteins. For the clinical database, the resultant MGF files were searched against the combined database of Human Uniprot proteome, contaminants, and SARS-Cov-2 proteins database.

For sequence database searching in the workflow, search algorithms - X! tandem, MSGF+, OMSSA were used within SearchGUI¹⁶⁸ to detect peptide spectral matches (PSMs), followed by False Discovery Rate (FDR) and protein grouping analysis using PeptideShaker²³⁵. The search parameters for digestion, modifications, tolerance, and FDR were chosen accordingly from the published papers for each of these datasets (**Table 2.1**). The peptide report generated using PeptideShaker was used to extract confident COVID-19 peptides. The peptides were validated using PepQuery analysis with MS tolerance of 10 ppm and MS/MS tolerance of 0.05 Da. The SARS-CoV-2 peptides detected from the three cell culture datasets and two clinical datasets were merged with the peptide list from *in silico* analysis of genomic sequences by Orsburn *et al.*²³⁶ to generate a peptide panel for interrogation of clinical data sets. The re-analysis of the dataset using the workflow is available online on the COVID-Galaxy website (<https://COVID-19.galaxyproject.org/proteomics>).

Figure 2.2. Workflows used in the interrogation of MS-data to identify and validate SARS-CoV-2 peptides a) Galaxy-based sequence database search workflow to detect and confirm SARS-CoV-2 peptides. MS/MS spectra from cell culture or clinical datasets were searched against appropriate protein sequence databases (protein sequences from COVID-19, contaminants, and Human Protein sequences) using SearchGUI/ Peptide Shaker. The peptide output was filtered to extract COVID-19 peptides, and the output was confirmed using PepQuery to extract confident peptides. mzidentML generated through this workflow was subsequently used for analysis in Lorikeet b) Workflow to verify detected SARS-CoV-2 peptides. A list of 639 Peptides (theoretical and validated peptides obtained from the cell-culture and clinical datasets) was subjected to PepQuery analysis of COVID-19 datasets to identify the presence of SARS-CoV-2 peptides. The quality of the peptide spectral matches (PSMs) was reviewed using Lorikeet visualization within the Multi-omics Visualization Platform for further validation. Peptides were also searched against NCBI-non redundant database and Unipept 4.3 for taxonomic annotation.

a)



b)

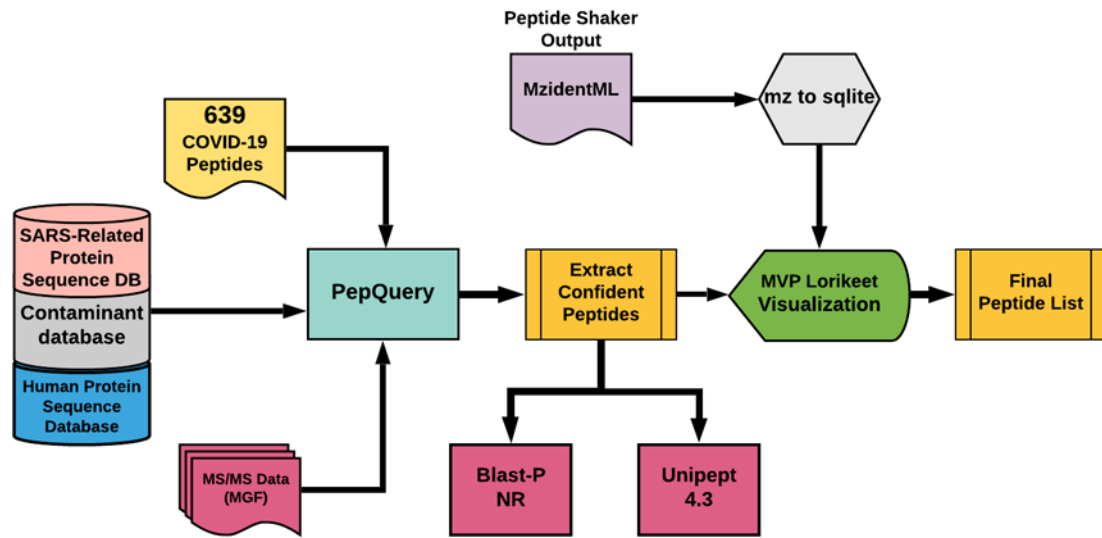


Table 2.1. Galaxy settings for SARS-CoV-2 peptide detection and validation. a) Search parameters for SearchGUI for the analysis of cell culture and clinical datasets in Galaxy. Parameters were based upon data analysis protocols detailed in the original publications. b) Parameters for the PepQuery search engine for the verification of clinical datasets in Galaxy.

a)

	Cell culture datasets			Clinical datasets		
<i>Search Parameters</i>	PXD018804	PXD018594	PXD018241	PXD021328	PXD019423	PXD020394
Algorithms	X!Tandem, MS-GF+, OMSSA, Comet	X!Tandem, MS-GF+, OMSSA	X!Tandem, MS-GF+, OMSSA	X!Tandem, MS-GF+, OMSSA	X!Tandem, MS-GF+, OMSSA	X!Tandem, MS-GF+, OMSSA
Digestion Enzymes	Trypsin	Trypsin	Trypsin	Trypsin	Trypsin	Trypsin
Missed cleavages	2	2	2	2	2	2
Precursor Ion Tolerance	5 ppm	5 ppm	10 ppm	10 ppm	10 ppm	10 ppm
Fragment Tolerance	0.02 Da	0.02 Da	0.6 Da	10 ppm	0.05 Da	0.05 Da
Minimum Charge	2	2	2	2	2	2

Maximum Charge	4	6	6	6	6	6
Fixed Modifications	Carbamidomethylation of C	Carbamidomethylation of C	Carbamidomethylation of C	-	Carbamidomethylation of C	Carbamidomethylation of C
Variable Modifications	Deamidation of N, Deamidation of Q, Oxidation of M	Deamidation of N, Deamidation of Q, Oxidation of M	Acetylation of protein N-term Oxidation of M	Acetylation of protein N-term, Oxidation of M	Deamidation of N, Oxidation of M	Oxidation of M
Minimum Peptide Length	6	8	8	8	8	8
Maximum Peptide Length	30	60	60	60	60	60
Maximum Precursor Error	10 ppm	5 ppm	10 ppm	10 ppm	10 ppm	10 ppm

b)

	Clinical datasets						
<i>PepQuery</i>	PXD0213	PXD0194	PXD0203	PXD0220	PXD0180	PXD0252	PXD0230
<i>Parameters</i>	28	23	94	85	94	14	16
Fixed modification(s)		Carbamidomethylation of C	Carbamidomethylation of C	Carbamidomethylation of C	Carbamidomethylation of C		
Variable modification(s)	Oxidation of M	Oxidation of M	Oxidation of M	Oxidation of M, N-term Acetylation	Oxidation of M, N-term Acetylation	Carbamidomethylation of C, Oxidation of M	Carbamidomethylation of C, Oxidation of M
Max Modifications	3	3	3	3	3	3	3
Unrestricted modification?	True	True	True	True	True	True	True
AA substitutions?	True	True	True	True	True	True	True

Precursor tolerance	10 ppm	10 ppm	10 ppm	10 ppm	10 ppm	10 ppm	10 ppm
Product tolerance	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da	0.05 Da
Digestion enzyme	Trypsin	Trypsin	Trypsin	Trypsin	Trypsin	Trypsin	Trypsin
Max missed cleavage	2	2	2	2	2	2	2
Fragmentation	CID/HC D	CID/HC D	CID/HC D	CID/HCD	CID/HCD	CID/HCD	CID/HCD
Scoring	HyperScore	HyperScore	HyperScore	HyperScore	HyperScore	HyperScore	HyperScore
Max charge	6	6	6	6	6	6	6
Min charge	2	2	2	2	2	2	2
Min peaks	10	10	10	10	10	10	10
Min score	12	12	12	12	12	12	12
Max length	45	45	45	45	45	45	45
# random peptides	1000	1000	1000	1000	1000	1000	1000

“Spectrum_ file” column?	True	True	True	True	True	True	True
--------------------------	------	------	------	------	------	------	------

Peptide Validation

This SARS-CoV-2 peptide panel was subjected to the Peptide Verification workflow (**Figure 2.2b**) against the clinical datasets specified above. The peptide validation workflow includes re-analysis by PepQuery as well as manual visualization and inspection in the Lorikeet application of Multi-omics Visualization Platform (MVP) to ascertain the quality of peptide sequences matched to MS/MS spectra. Unrestricted modification searching and amino acid substitutions were enabled in PepQuery to ensure the most rigorous search possible, with hypothetical post-translational modifications and amino acid substitutions applied to the reference peptides to examine every possible sequence match to the putative SARS-CoV-2 spectra. To rule out misidentification of host peptides and ensure the specificity of validated peptides for the SARS-CoV-2 virus, a reference proteome of human proteins as well as the proteomes of SARS-CoV, OC43, NL62, HKU1, 229E, SARS-MA15, SARS-WIV1, and MERS-CoV were used for this rigorous evaluation. The results from PepQuery were then filtered to remove any peptides which had matches to the reference proteomes, leaving only those peptides which aligned to the SARS-CoV-2 proteome. The spectra of the validated peptides were then manually annotated using the Multi-omics Visualization Platform (MVP)¹⁷⁴ or the Proteomics Data Viewer (PDV)²³⁷ to ensure the quality of the potential SARS-CoV-2 targets. The workflow also included additional, optional in-line characterization of these peptides by searching against NCBI-non redundant (nr) BLAST-P and Unipept¹⁹⁰ analysis. Further offline

analysis was performed using NCBI BLAST-P analysis as well as the MetaTRYP34 coronavirus database. The peptide validation workflow can be found at COVID Galaxy website (<https://COVID-19.galaxyproject.org/proteomics>).

2.3 Results

2.3.1. Sequence Database Searching Results

Sequence database searching to generate peptide spectral matches (PSMs) and to identify peptides from three cell culture datasets (**Figure 2.1a**) using the workflow shown in **Figure 2.2a** led to detection of 139 peptides, 99 peptides and 579 peptides, respectively. For the two clinical datasets analyzed using the workflow, we detected 76 and 8 peptides, respectively (**Table 2.2**). These peptides together represented 630 unique peptides corresponding to several proteins coded in the SARS-CoV-2 genome; to these we then added a further 9 unique peptides generated from *in silico* translated data by Orsburn et al.³⁸ to generate a list of 639 unique SARS CoV-2 peptides (**Supplemental Table 2.1**). This 639-peptide panel was further used to interrogate the clinical datasets and determine the reliability of their detection using untargeted MS-based proteomics. BLAST-P analysis of the 639-peptide panel showed that these peptides mapped to 27 proteins and open reading frames within the SARS-CoV-2 genome (**Figure 2.3**), with sequence coverage ranging from 4.7% coverage (Proofreading exoribonuclease Guanine-N7 methyltransferase protein) to 93.7% coverage (Nucleocapsid protein) (**Figure 2.4**).

Table 2.2. Peptides generated from MS datasets in the construction of the library and validation in the patient datasets

Dataset type	Manuscript (Proteome Xchange ID)	SARS-CoV-2 peptides detected using Database Search Workflow	Distinct Detected Peptides from DSW	SARS-CoV-2 peptides detected using Peptide Validation Workflow	Distinct Detected Peptides from PVW
Cell Culture	Gouveia <i>et al</i> (PXD018804)	139	630	-	-
	Grega <i>et al</i> (PXD018594)	99		-	-
	Davidson <i>et al</i> (PXD018241)	579		-	-
Clinical Datasets	Cardozo <i>et al</i> (PXD021328)	76	-	70	87
	Ihling <i>et al</i> (PXD019423)	8	-	21	
	Rivera <i>et al</i> (PXD020394)	-	-	10	
	Leng <i>et al</i> (PXD018094)	-	-	14	
	Zeng <i>et al</i> (PXD022085)	-	-	37	
	Cardozo <i>et al</i> (PXD025214)	-	-	39	
	Bankar <i>et al.</i> (PXD023016)	-	-	35	

Figure 2.3. Protein assignment of detected and validated SARS-CoV-2 peptides: Circos plot of peptides against SARS-CoV-2 proteins (outermost ring). Of the 639-peptide panel (2nd outermost ring), many peptides could be identified using our validation workflow in clinical and cell culture datasets (3rd outermost ring). Peptides derived from ORF9b, papain-like protease, Nsp4, Nsp10, uridylyate endoribonuclease (Nsp15) and certain spike protein peptides were only found in cell culture datasets (2nd innermost ring). Peptides chosen for targeted analysis are annotated in the innermost ring. Circos plot was generated in Galaxy²³⁸.

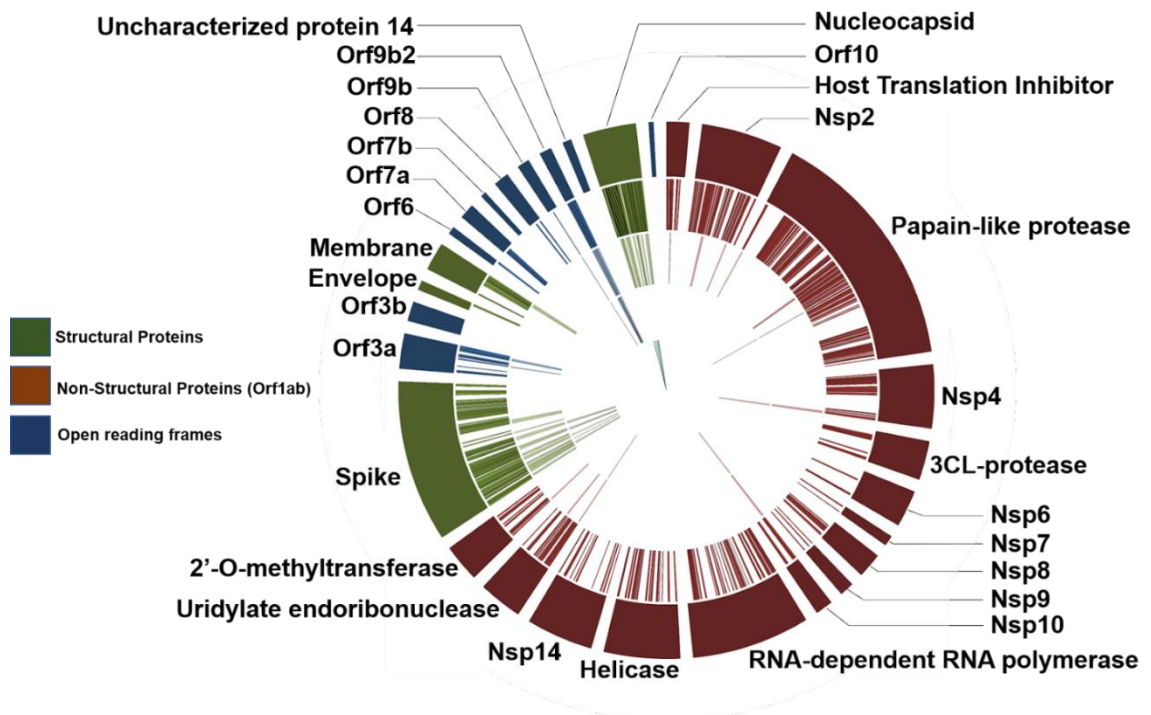
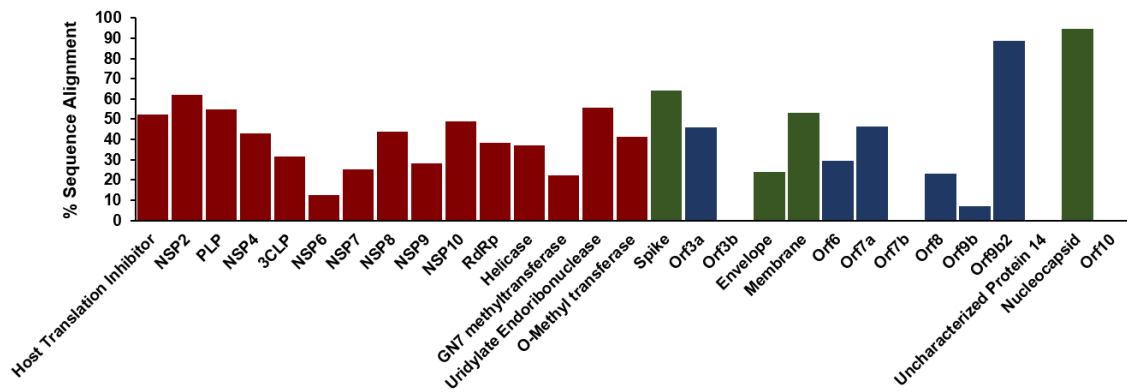


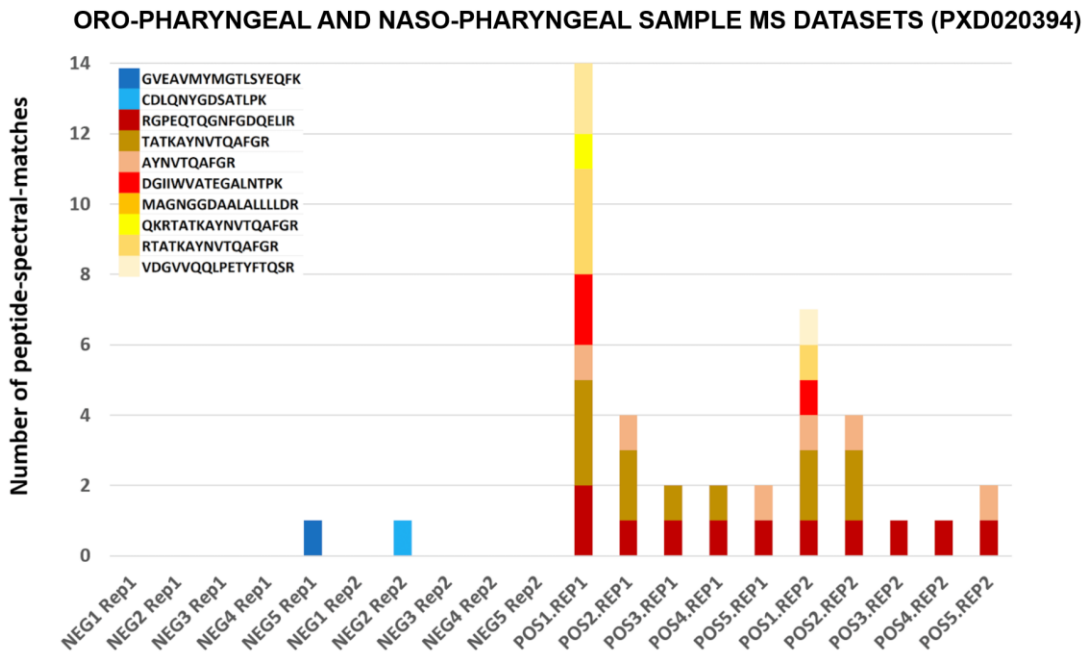
Figure 2.4. Alignment of the 639-peptide panel to viral proteins from SARS-CoV-2. Peptides detected from patient and cell culture datasets were aligned to the SARS-CoV-2 proteome. Proteins were colored in terms of their classification as structural proteins (green), non-structural proteins (maroon), or open reading frames (blue).



2.3.2. Peptide Validation Results

Having derived a comprehensive panel of 639 peptides detected across multiple COVID-19 datasets, we then utilized a validation workflow based around the PepQuery database to interrogate the dataset PXD020394, derived from oro- and nasopharyngeal swabs collected in the clinic from patients positive and negative for COVID-19. This resulted in detection of 10 SARS-CoV-2 peptides from our panel in these clinically relevant samples (**Figure 2.5**).

Figure 2.5. List of validated peptide spectral matches in the oro-pharyngeal and nasopharyngeal mass spectrometry dataset (PXD020394). The bar diagram above shows the peptide-spectral matches after running the validation workflow for 639 SARS-CoV-2 peptide-panel against the five COVID-19 positive patient samples (with replicates) and five COVID-19 negative patient samples (with replicates). Several SARS-CoV-2 peptides were detected in COVID-19 positive samples (See samples labeled ‘POS’”) and only two peptides were detected in two of the negative samples (NEG5 Rep1 and NEG2 Rep2). The SARS-CoV-2 peptides detected in COVID-19 negative samples did not meet the threshold of acceptable spectral quality in subsequent spectral validation.



We detected eight of the peptides in COVID-19 positive sample replicates - with the peptide RGPEQTQGNFGDQELIR being detected in all positive sample replicates, followed by TATKAYNVTQAFGR and AYNVTQAFGR detected in 6 out of 10 replicate samples (**Figure 2.5**). We also detected two peptides- GVEAVMYMGTLSEYQFK and CDLQNYGDSATLPK- from COVID-19 negative samples.

We also re-analyzed the clinical datasets used in the generation of the 639 panel (the second oro/nasopharyngeal dataset from Cardozo *et al.* as well as the saline gargling dataset), using our validation workflow. The validation workflow provides a complementary method to the initial sequence database searching method for confirming peptide spectrum matches, based primarily on the PepQuery tool. For the oro/nasopharyngeal dataset, we confirmed confident identification of 70 peptides using the peptide validation workflow (as compared to 76 detected using the initial sequence database searching workflow). For the saline gargling dataset, we confirmed the presence of 21 peptides using the peptide validation workflow (as compared to 8 peptides detected using the peptide search workflow). Considering all peptides detected in clinical samples using the peptide validation workflow, we detected 87 peptides with confidence (**Table 2.2**). These validated peptides were assigned to known proteins from the COVID-19 proteome. Most of the peptides detected in the upper respiratory tract were aligned to structural proteins making up the viral capsid such as nucleocapsid protein N, the viral matrix protein M, and the spike protein S; fewer peptides were aligned to proteins involved in viral replication such as papain-like protease, RNA-directed RNA polymerase, non-structural protein, 2'-O-methyltransferase and host translation inhibitor (**Figure 2.3**). The largest number peptides were identified in the oro/nasopharyngeal dataset that consisted of

combined oropharyngeal and nasopharyngeal swabs analyzed by Cardozo *et al.* By contrast, fewer peptides were identified from PXD019423 and PXD020493, which were derived from gargled saline samples and a second study of combined oropharyngeal and nasopharyngeal samples, respectively.

Based on the sample type from which they were derived (clinical samples versus *in vitro* cell culture experiments) and their source (empirically derived from MS/MS data versus theoretically determined based on genomic sequence data), we categorized the peptides as being present or absent in the various datasets based on their confident detection using our validation workflow. We found that the validated peptides clustered into distinct groups based on their source sample and dataset of origin, and how they were originally identified (**Supplemental Table 2.1**). Eleven peptides were found to be highly consistent across the upper respiratory clinical datasets as well as the *in vitro* cell culture datasets. In considering theoretical peptides proposed by the Orsburn *et al.*, eleven of those predicted peptides were observed in clinical samples and eight were detected in the *in vitro* cell culture samples. Twenty-two SARS-CoV-2 peptides that were not initially identified using the database search workflow were identified by matching to MS/MS spectra using the PepQuery-based validation workflow across multiple datasets.

Having established the presence of verified SARS-CoV-2 peptides in our initial clinical datasets, we then interrogated additional clinical datasets to further validate the utility of our methodology. Further patient datasets comprising oro/nasopharyngeal swabs (PXD025214) as well as nasopharyngeal datasets from COVID-19-positive patients (PXD023016) were analyzed using the PepQuery verification workflow against the 639-peptide panel. Analyses of these datasets revealed 39 and 35 verified peptides, respectively,

which showed considerable overlap with our initial analyses of oro/nasopharyngeal and gargling datasets. Clinical datasets from lung biopsies (PXD018094) and BALF (PXD022085) were also interrogated to determine the applicability of our approach in detecting SARS-CoV-2 within the deeper respiratory tract. Our validation workflow was able to confidently match MS/MS spectra to 15 peptides in the lung biopsy dataset and 37 peptides in the BALF dataset. In comparing the peptides found within the upper respiratory samples to those detected within the lung biopsy samples and the BALF samples, most of the peptides detected in the deep lung datasets are unique to the samples being analyzed, with no peptides in common with the upper respiratory tract samples (**Supplemental Table 2.1**). Despite this apparent disparity, BLAST-P analysis reveals an alignment of SARS-CoV-2 peptides identified in deep lung tissue corresponding to a similar complement of SARS-CoV-2 proteins as the upper respiratory tract datasets, including additional structural proteins such as the Spike protein and Membrane glycoprotein as well as other nonstructural and replication proteins such as RNA-directed RNA polymerase, Protease 3CL-PRO, etc. In addition, the lung biopsy and BALF datasets also included MS-data from patients negative for COVID-19. In contrast to the two SARS-CoV-2 PSMs identified in the oro/nasopharyngeal samples from COVID-19-negative patients, samples analyzed from lung biopsies of COVID-19-negative patients resulted in identification of 21 SARS-CoV-2 peptides using the verification workflow. Similarly, 37 peptides were detected in BALF samples isolated from patients that tested negative for COVID-19 using the verification workflow.

The last category of peptides evaluated in this study were detected from COVID-19 cell culture studies (**Supplemental Table 2.1**). These peptides were derived from

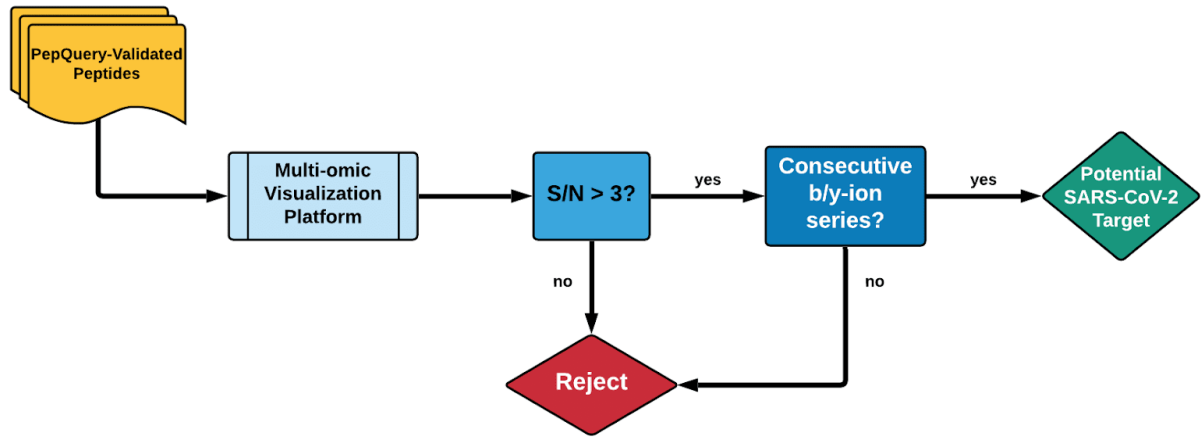
protein sequences that were not available in the initial Uniprot sequence databases but were subsequently added as more COVID19 strains were sequenced^{239,240}. We added these sequences to the sequence database to enable the detection of these COVID-19 proteoforms. Using this updated sequence database, we detected and validated twelve peptides from Accessory protein ORF9b from SARS-CoV-2 and two peptides from ORF1ab polyprotein from SARS-CoV-2. These peptides were observed only in cell culture data, and not in the clinical datasets (**Figure 2.3**).

2.3.3. Identifying Detected Peptides with the Highest Quality Spectra

As a quality check on our bioinformatic workflows, we utilized the Multi-Omics Visualization Platform and Proteomics Data Viewer to manually assess the spectral quality of the peptides that passed PepQuery validation, as well as elucidate the distribution of these peptides throughout the six datasets we analyzed. It is critical that the peptides used for targeted MS-based assays for detecting SARS-CoV-2 as targets have excellent spectral quality to ensure adequate reliability in detecting and quantifying these peptides across a variety of clinical samples. Here, we focused on four peptides (AYNVTQAFGR, MAGNNGDAALALLLDR, RGPEQTQGNFGDQELIR, DGIIWVATEGALNTPK) found in the SARS-CoV 19 positive patients from the second oro/nasopharyngeal dataset (PXD020934) that were also seen in the other clinical datasets as well as two peptides found in the negative patients (CDLQNYGDSATLPK, GVEAVMYMGTLSEYQFK) from the same oro/nasopharyngeal dataset as benchmark examples for manually validating our spectra. For these selected four peptides, from the virus-positive samples we found largely complete b- and/or y-ion series with at least three consecutive ions detected in either

series. In addition, we found that these fragment MS/MS ions showed intensities at least three-fold higher than the background noise level of the spectra. By contrast, the two peptides found in the negative samples had a very few fragment MS/MS ions detected which scarcely rose above the level of the background noise. Together, the MS/MS spectra of these six peptides were used to generate guidelines which were then used to manually interrogate the rest of the SARS-CoV-2 spectra as being genuine or misidentified by the bioinformatics software (**Figure 2.6**). Manual annotation of the MS/MS spectra found that 16 of the peptides validated in PepQuery had MS/MS spectra suitable for confident identification.

Figure 2.6. Guidelines for the manual validation of MS/MS spectra using the Multi-omic Visualization Platform (MVP). The MS/MS spectra of peptides that passed validation in PepQuery were manually annotated using MVP based on a test cohort of four peptides that passed validation in COVID-positive patient datasets and two peptides that passed validation in COVID-negative patient data. The signal-to-noise ratio of the product ions within MS/MS spectra was examined, and spectra containing product ions with at least a three-fold higher intensity than noise level were retained. Next, the degree of completeness of the b- and y-ion series was considered, with passing spectra determined to have at least three consecutive b- or y-ions in their series. Peptides with spectra that passed these criteria were considered valid peptide targets for the detection of SARS-CoV-2.



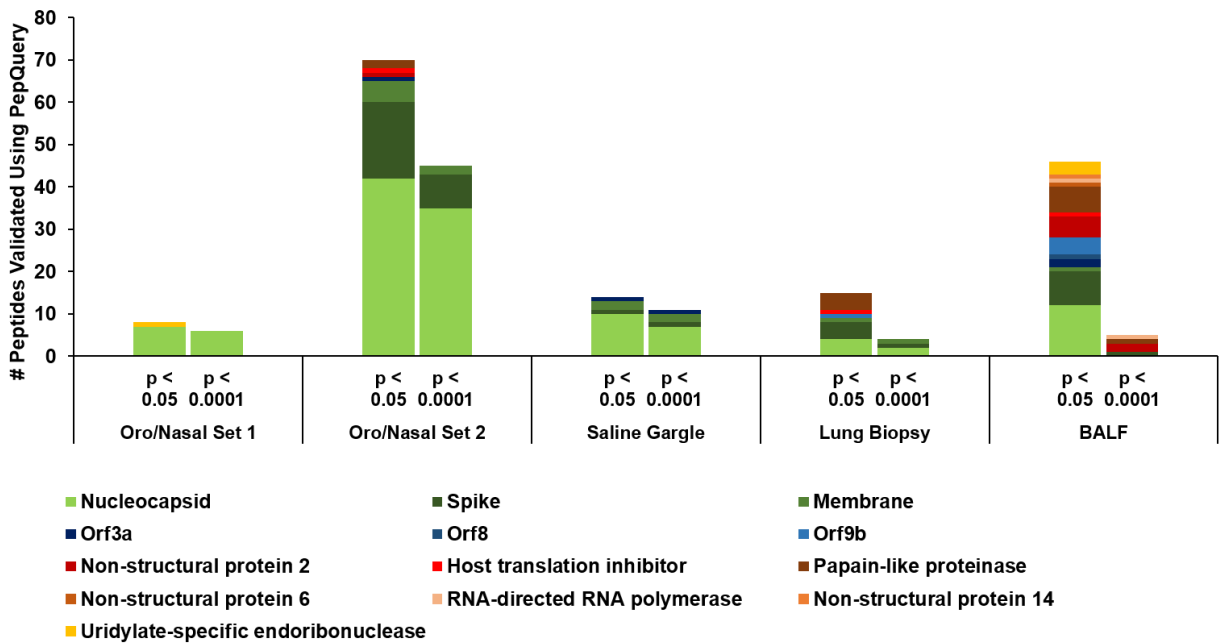
As a part of our investigation, we evaluated eight peptides that were reported by Orsburn *et al.*²³⁶ (**Supplemental Table 2.1**). However, Lorikeet visualization of the Peptide Spectral Match (PSM) quality detected only two peptides (with sequences ADETQALPQR and FDNVLPFNDGVYFASTEK) in the clinical sample PXD021328 dataset; of these the ADETQALPQR was also detected in all three cell cultures sample datasets while the FDNVLPFNDGVYFASTEK sequence peptide was detected in two of the three cell culture samples (**Supplemental Table 2.1**). All the eight peptides were found to have good quality of PSMs in the cell culture datasets by using manual validation. Out of these eight peptides, a peptide with sequence HTPINLVR was detected in all cell culture experimental datasets.

We were able to validate 22 peptides using PepQuery which were not detected in the database search workflow (**Supplemental Table 2.1**). Subsequent manual validation of these peptides determined only two peptides had good quality spectra. The peptide of sequence DGIIWVATEGALNTPKDHIGTR was validated by using PepQuery and manual visualization in the PXD019423 dataset along with another peptide with sequence FTALTQHGKEDLK from the PXD02132 dataset.

To determine the optimal candidates for the detection of SARS-CoV-2 using clinical MS-based assays, we resolved to focus on those peptides that passed PepQuery with the highest confidence, and subject these to manual inspection of spectral quality. We therefore sorted the results of our PepQuery analyses to include only those which had the highest confidence possible ($p\text{-value} < 0.0001$) to maximize the likelihood of passing our spectral annotation thresholds. In filtering the clinical datasets, we see a notable difference between the datasets derived from the upper respiratory tract (oro/nasopharyngeal datasets

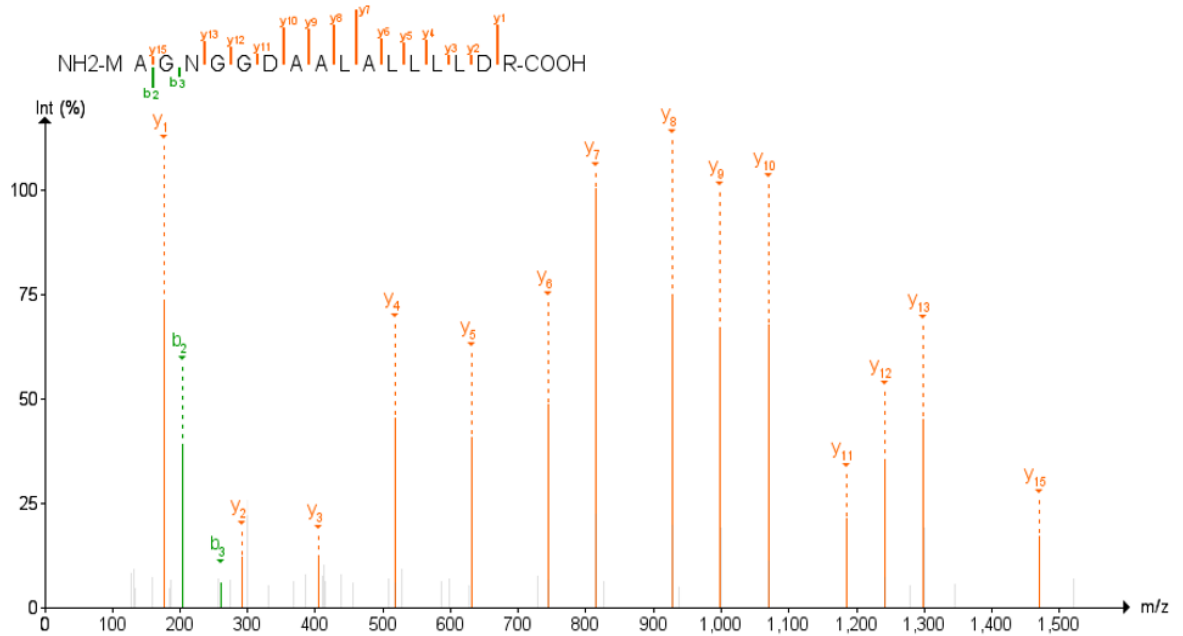
1 and 2 as well as the saline gargling dataset) and those derived from deep lung tissue (the lung biopsy and BALF datasets) (**Figure 2.7**). In filtering the PepQuery results from the upper respiratory tract datasets, we noted that the structural proteins that had the most identified peptides- the nucleocapsid, membrane protein, and spike proteins- show relatively little elimination of PSMs, while the proteins involved in viral replication are generally lost, indicating relatively high confidence in the PepQuery validation of the peptides of the viral structural proteins. By contrast, peptides found in all proteins in the lung biopsy and BALF datasets were filtered out at this step, yielding only 3 and 4 high-confidence peptides in each dataset, respectively, leaving single peptides of nucleocapsid, membrane protein, and spike protein in the lung biopsy samples and single peptides of the spike protein, papain-like protease, non-structural protein 2, and RNA-dependent RNA polymerase.

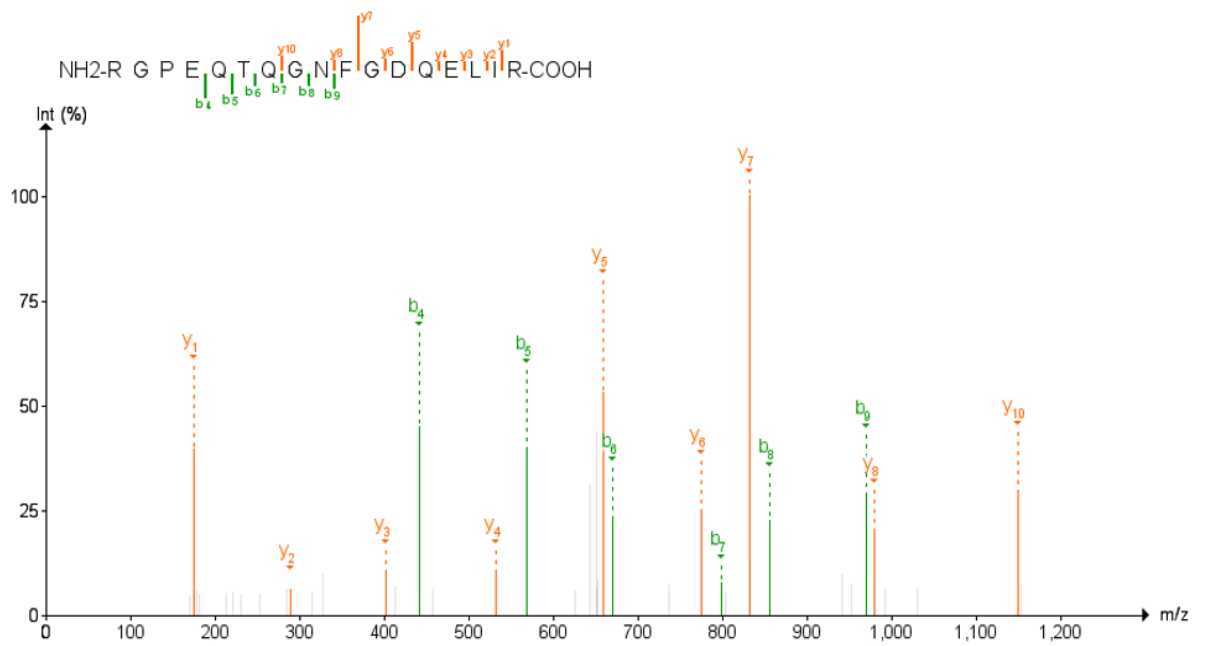
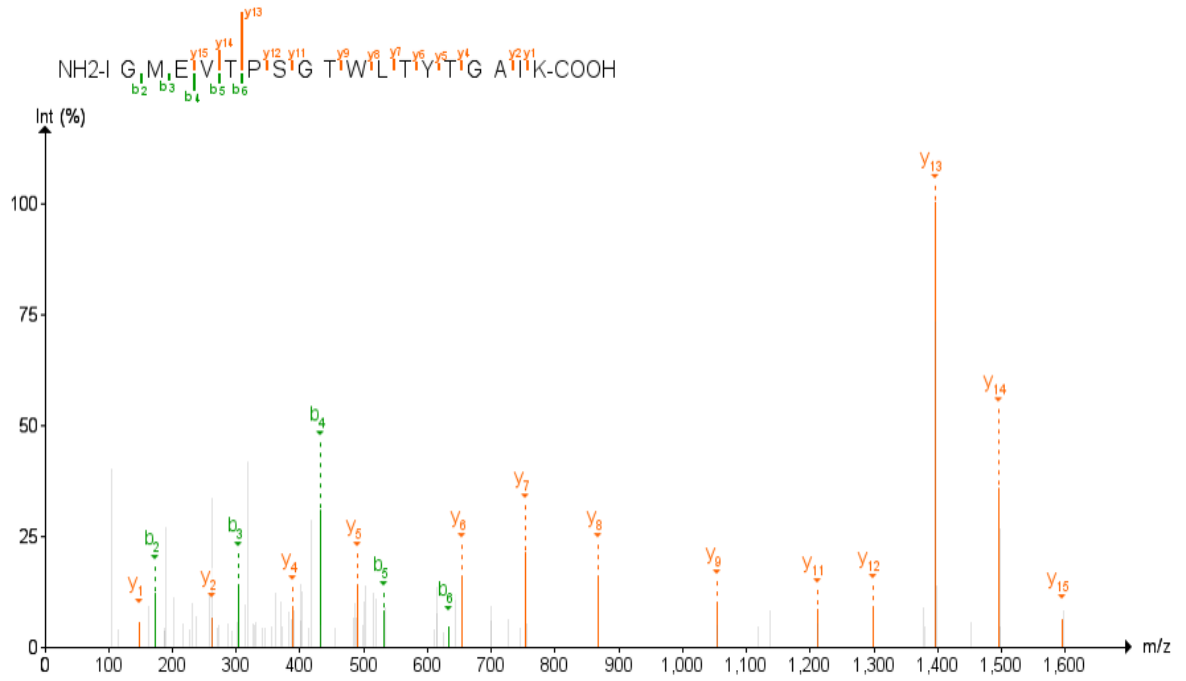
Figure 2.7. Peptide spectral matches (PSMs) of SARS-CoV-2 peptides in the upper respiratory clinical datasets are of higher confidence than deep lung datasets. PSMs validated in oro/nasopharyngeal datasets, saline gargling samples, lung biopsy samples, and bronchoalveolar lavage fluids (BALF) using PepQuery as grouped into the proteins they aligned to; columns correspond to those peptides that passed PepQuery validation with minimal required confidence (left) as well as those associated with higher confidence (right).



The spectra of the peptides identified with high confidence in the clinical datasets were then analyzed using MVP, which leverages the Lorikeet viewer for visualization of annotated peptide MS/MS spectra. Manual analysis of the high-confidence peptides detected in the lung biopsy and BALF datasets using our previously established guidelines showed only the single peptide FLALCADSIIGGAK, a component of Non-structural protein 2, in the BALF dataset as having a good quality spectrum, suggesting that the use of clinical samples collected using more invasive methods from deep within the lung may be unsuitable for detection of SARS-CoV-2 using a clinical proteomics strategy. In contrast, 11 peptides found in the upper respiratory tract datasets had high confidence and high-quality MS/MS-spectra. Of these, we then chose four peptides-MAGNGGDAALALLLDR, DGIWVATEGALNTPK, RGPEQTQGNFGDQELIR, and IGMEVTPSGTWLTYTGAIK, which were each identified in at least three of the five upper respiratory clinical datasets, determining these to be the most reliable peptides for proteomics-based detection of SARS-CoV-2 in clinical samples harvested from the upper respiratory tract (**Figure 2.8, Table 2.3**). We assert that these represent the best candidates for targeted proteomics screening for potential cases of COVID-19.

Figure 2.8. MS/MS spectra of SARS-CoV-2 peptides most confidently identified in PepQuery (p-value < 0.001) and across the most clinical samples. Spectral quality was interrogated using the Lorikeet viewer implemented within the Multi-Omics Visualization Platform (MVP); images for annotated PSMs for these peptides were created using the PDV platform from the Zhang lab.





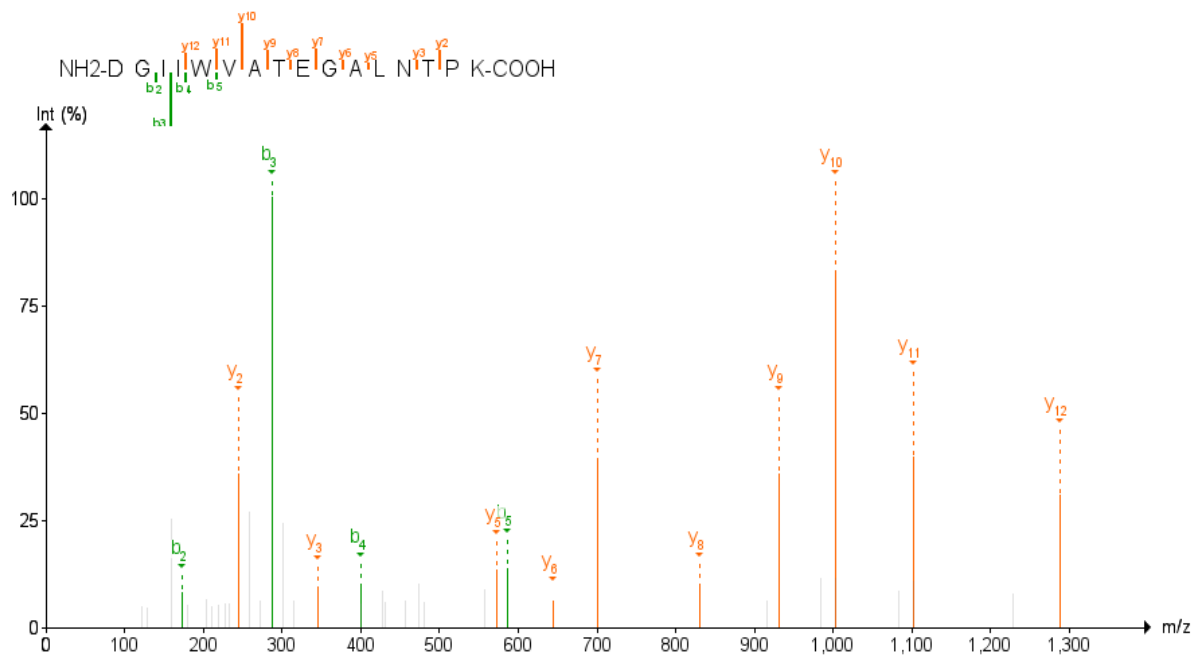


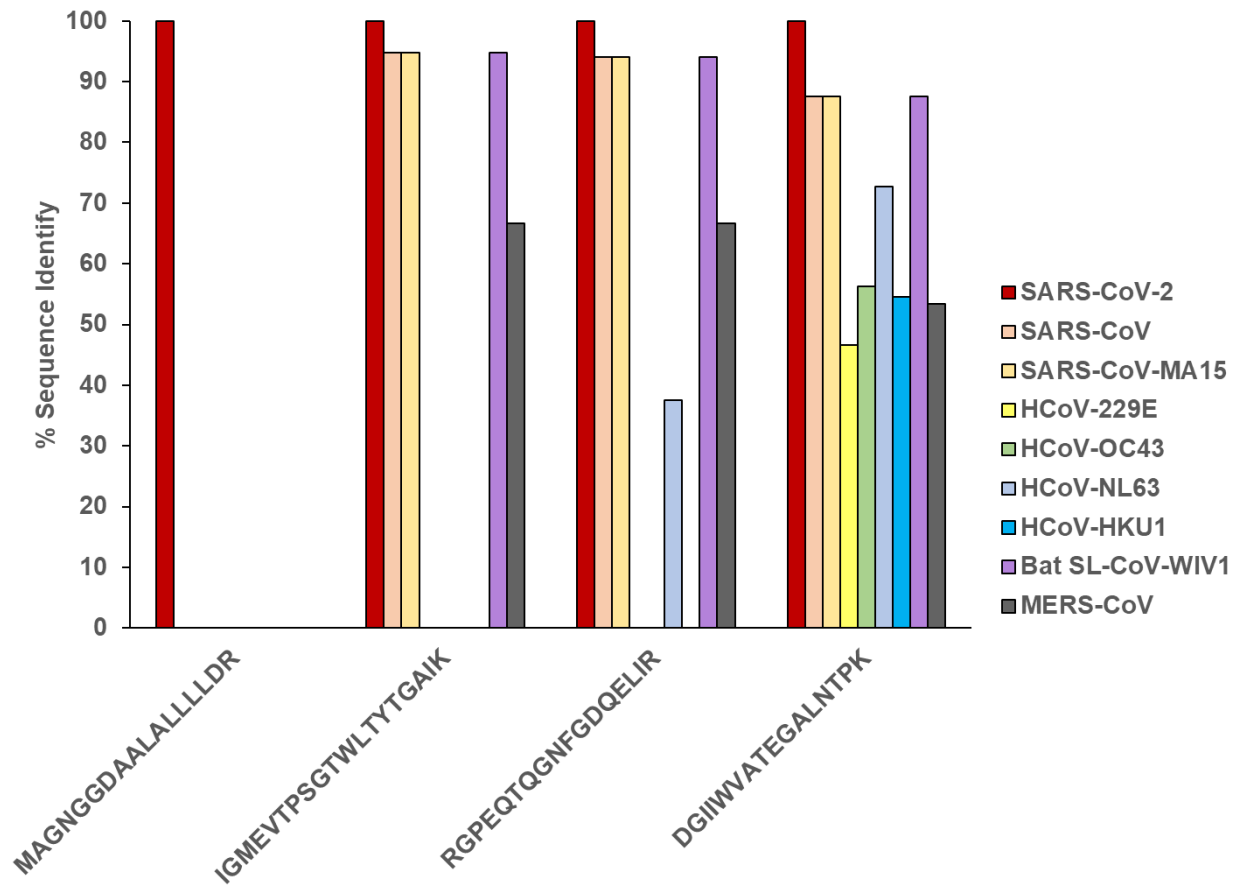
Table 2.3. The presence of the four optimal SARS-CoV-2 target peptides in clinical datasets. The presence of the peptides MAGNGGDAALALLLDR, DGIIWVATEGALNTPK, RGPEQTQGNFGDQELIR, and IGMEVTPSGTWLTYTGAIK in COVID-19 positive patients.

	MAGNGGDAALALL LDR	DGIIWVATEGALNTP K	RGPEQTQGNFGDQE LIR	IGMEVTPSGTWLTYT GAIK
<i>Upper Respiratory Tract Datasets</i>				
PXD020394	x	x	x	
PXD021328	x	x	x	x
PXD019432	x		x	x
PXD025214	x	x	x	x
PXD023016		x		
<i>Deep Lung Datasets</i>				
PXD018094				
PXD022085				

2.3.4 Viral Specificity of High-Quality Peptides Detected in SARS-CoV-2

We performed taxonomic analysis using MetaTryp to validate the specificity of the four highest-quality peptides detected in clinical samples to coronaviruses (**Figure 2.9a**). We found that these peptides mapped to proteomes of several coronaviruses, with each showing alignment SARS-CoV-2. To gauge the degree of specificity of these peptides for SARS-CoV-2 over other coronaviruses and their potential human host, we performed BLAST-P analysis of these peptides against proteomes for SARS-CoV-2, humans, and eight known pathogenic human coronaviruses. To interrogate all possible matches to the target organisms, a relatively lax E-value cutoff of 1 was used. In considering the sequence alignment of these peptides, the peptides examined found a high degree of alignment to the nucleocapsid protein (N-protein) of SARS-CoV-2 (**Figure 2.9b**). Each of the four distinct peptides that showed alignment to the N-protein also showed 100% sequence homology uniquely to SARS-CoV-2, with decreased sequence alignment in other closely related coronaviruses. One peptide sequence, MAGNGGDAALALLLDR, showed perfect alignment to the SARS-CoV-2 nucleocapsid protein with no alignment to the same protein in any other viruses. In all cases, no alignment to any human proteins was noted.

b)



2.4 Discussion

Clinical diagnostics using targeted MS-based proteomics has found considerable utility in recent years as a powerful tool for detecting peptide biomarkers characteristic of several diseases. Bottom-up proteomics has been used to characterize tumors in biopsied breast cancer tissues^{241,242}, to explore the phenotypic changes that occur with opportunistic fungal infections in HIV/AIDS patients²⁴³, and even differentiate between COVID-19 patients with differing WHO severity grades²⁴⁴. While these experiments effectively measure the phenotype of patients to infer a disease state, direct detection of proteins using targeted MS-based methods (SRM) from disease organisms can be used as a diagnostic assay for diseases. For these, it is critical that the most reliable peptides, specific to the protein of interest, are determined.

The pressing nature of the COVID-19 pandemic presents an opportunity for the use of targeted MS-based proteomics to supplement conventional RT-qPCR diagnostic procedures²¹² to mitigate the false negatives inherent in the detection of viral RNA²⁴⁵, along with other advantages of direct detection of peptides, such as chemical stability of the target molecules. Ideally, direct detection of diagnostic peptides would be achieved in samples easily collected in the clinic using non-invasive methods. While many labs have begun proteomic analysis of samples to identify SARS-CoV-2 infection in both in-vitro models and clinical samples, the development of targeted assays based on this work requires preliminary work to determine those peptides which are most reliably detected and most specific for unambiguous diagnosis of infection. To mitigate this and establish the best targets possible for a SARS-CoV-2 clinical proteomics assay, we identified detectable SARS-CoV-2 peptides using Galaxy-based workflows. To narrow this list down to the

most confident and reliably detected peptides, we then utilized a bioinformatics workflow built around the PepQuery search engine. Developed by Wen et al.¹⁷³, this search engine interrogates raw mass spectrometry data for spectral matches to pre-chosen peptide sequences of interest and compares these matched spectra to reference proteomes to see whether the peptide of interest is a better match to the data than any reference peptide, scoring the peptide match much faster and with much less processing power needed than a conventional sequence database search. By using PepQuery on peptides that have already been designated as potential matches, we can utilize the increased statistical power of using multiple peptide search engines²⁴⁶ common to many proteomics software suites on a much faster time scale. Using this as well as other tools available in the Galaxy platform, we were able to interrogate publicly available data to ascertain the most reliable peptides for detecting SARS-CoV-2.

In the two oro/nasopharyngeal datasets and gargled saline dataset we examined, we found 75 peptides within the original list of 639 detected peptides that showed a high-confidence match to SARS-CoV-2 proteins over human proteins or other coronavirus proteins, suggesting that the unambiguous detection of SARS-CoV-2 in patients using proteomics technology is theoretically possible. These peptides were found in proteins throughout the viral particle (**Figure 2.3**), with more structural protein peptides detected than replication proteins. It was observed that the datasets stemming from the clinical samples had noticeably fewer peptides validated in them compared to those from in vitro experiments; this is potentially due to larger amounts of material, the differential abundance of host proteins in clinical samples compared with cultured samples²⁴⁷, and the lack of viral clearance from cultured cells²⁴⁸. Of these, manual annotation found that 16

peptides could be truly said to have good quality MS/MS spectra, based on our thresholds for PSM quality and annotation.

From the 16 validated peptides with high-quality spectra, 11 peptides also were known to be high confidence matches in PepQuery. From these we chose four peptides that had high-confidence matches in PepQuery, were consistently seen in clinical samples, and were unique to SARS-CoV-2, making them the best candidates for diagnosis of COVID-19 using targeted MS-based methods. Given their high degree of specificity for SARS-CoV-2 and the high quality of their spectra, we postulate that the detection of any of these individual peptides in a clinical patient would warrant further clinical investigation of the patient's infection status. It is notable that these are all found within the nucleocapsid phosphoprotein, or N-protein. The nucleocapsid phosphoprotein is common to coronaviruses and serves to complex with and stabilize the viral RNA genome and package it into the viral particle^{249,250}. The viral ribonucleoprotein complex of N-protein and gRNA is localized beneath the matrix proteins (M-proteins) and spike proteins (S-proteins) that make up the capsid surface^{251,252}. As many copies of N-protein are needed to stabilize the viral gRNA, the N-protein is thought to be one of the most abundant proteins in the assembled SARS-CoV-2 viral particle²⁵³; analysis of SARS-CoV transcript levels in infected cells show the N-protein to be the most abundant RNA-based sub-genome within the cell²⁵⁴. Taken together, these phenomena explain the prominence of N-protein peptides across the proteomic datasets we examined. As the N-protein is a frequent amplification target for RT-qPCR assays as per FDA guidelines for diagnosis²⁵⁵, we believe that our results are complementary to current protocols in screening for and diagnosis of COVID-19.

In addition to upper respiratory tract clinical samples, we profiled datasets derived from deep within the respiratory tract, comprising a dataset derived from COVID-19 patient lung biopsies as well as a separate dataset of bronchoalveolar lavage fluid (BALF) samples from COVID-19 patients; we analyzed these MS-data against our 639-peptide panel to determine whether our methodology was suitable for SARS-CoV-2 detection in these samples. We found a lack of high-confidence peptides with high quality spectra in these samples, with only a single MS run from the PXD022085 sample yielding the peptide FLALCADSIHGGAK which was not found in the datasets derived from higher up in the respiratory tract. Our results would suggest that samples collected using invasive methods (biopsy, lung fluid extraction), in addition to being taxing on the patients to collect, demonstrate insufficient concentrations of viral particles to be robustly detected using MS-based methods and the workflows presented here. The complexity of the sample matrices may also affect the ability to detect SARS-CoV-2 peptides, as the upper respiratory tract dataset which showed the fewest proposed target peptides- PXD023016- was also the only upper respiratory tract dataset which utilized viral transport medium in the collection of patient samples. Viral transport medium contains added serum as a part of its formulation, adding to the protein background of the collected samples. The deep lung datasets were also noted for their complexity, being either homogenized bulk lung tissue (PXD018094) or protein- and lipid-rich bronchoalveolar lavage fluid (PXD02085). In addition, the deep lung datasets had more sample preparation steps than the upper respiratory tract datasets, providing more opportunities for adding confounding variables to the analysis. Our results suggest that samples collected using minimally invasive methods from the upper respiratory tract (oropharyngeal/nasopharyngeal swabs and gargling samples) and using

simplified, streamlined sample preparations would be most suitable for reliable detection of the SARS-CoV-2 virus targeting the high-confidence peptides we identify here – offering an optimal method for high-throughput diagnosis of infection.

While we believe the peptides presented here constitute promising targets for COVID-19 diagnosis, further experiments are required to establish targeted proteomics as a viable methodology for detection of SARS-CoV-2 infection. The limits of detection of these peptides need to be reliably established in larger numbers of human samples collected in the clinic to determine the minimal number of viral particles that can be detected. This could help determine the optimal sample type and procedure for collection to ensure reliable results. In addition, proteomic analysis of samples collected at different stages of SARS-CoV-2 infection should be performed to determine viability of targeted proteomics for detection during the full life cycle of infection. Finally, the sample processing that accompanies bottom-up proteomics²⁵⁶ should be optimized to be performed on a rapid time scale. Most conventional bottom-up proteomics experiments utilize trypsin digestions which occur overnight with incubation at 37°C, meaning a single sample would have to be processed and analyzed over the course of two days; this would have to be significantly reduced as the conventional 24–48-hour complete turnaround of RT-qPCR assays is being decreased using strategies such as direct RT-qPCR¹², RT-LAMP¹³, and CRISPR-based amplification strategies^{257,258,259}. The turnaround time of clinical proteomics can potentially be decreased for individual samples using modified or alternative protein digestion enzymes with higher rates of reactivity²⁶⁰; in addition, automation of clinical proteomics technology can provide reproducible, robust analyses of patient samples^{261,198}.

Beyond the peptides derived empirically from clinical and *in vitro* datasets, we also included theoretical SARS-CoV-2 peptides predicted bioinformatically by Orsburn et al. in our panel for validation; in doing so we were able to validate eight peptides in both clinical and *in vitro* datasets. It is worth noting, however that of these eight peptides, only two had good quality spectra in the clinical data, supporting the need for caution in accepting peptide identifications. The validation workflow presented here was also able to identify peptides in mass spectrometry data which conventional unbiased algorithms, such as our database search workflow presented in **Figure 2.2a**, are unable to identify; this may be of use in the analysis of complex patient and environmental mass spectrometry data collected for alternate purposes in the detection of SARS-CoV-2 under various conditions.

In conclusion, we interrogated multiple proteomic datasets from COVID-19 patients and *in vitro* experiments using bioinformatics workflows to determine which peptides from SARS-CoV-2 would make suitable targets for a clinical proteomics assay and which would make poor targets, potentially resulting in false negatives. Through our analyses, we found that of the 639 peptides that are readily detected across all samples, 87 of these were found to have a specific match to the SARS-CoV-2 proteome, rather than within the human proteome or other coronavirus proteomes. These peptides were narrowed down to 4 high-confidence peptides with excellent quality spectra found across most of the upper-respiratory tract clinical datasets analyzed in this study which we believe would be ideal candidates for diagnosis of COVID-19 via targeted proteomics. The workflows employed here for peptide identification and validation are well-documented, open-source, and hosted on the publicly accessible Galaxy Europe platform (usegalaxy.eu) where they can be edited, modified, or interfaced with other relevant bioinformatics tools to aid in

analysis of proteomics data. The workflows presented here were also used in subsequent analyses of patient samples to identify various SARS-CoV-2 strains as well as the microbial communities associated with COVID19 infection.

III. QUANTITATIVE PROTEOGENOMIC CHARACTERIZATION OF INFLAMED MURINE COLON TISSUE USING AN INTEGRATED DISCOVERY, VERIFICATION, AND VALIDATION PROTEOGENOMIC WORKFLOW

Adapted from:

Rajczewski AT, Han Q, Mehta S, et al. Quantitative Proteogenomic Characterization of Inflamed Murine Colon Tissue Using an Integrated Discovery, Verification, and Validation Proteogenomic Workflow. *Proteomes*. 2022;10(2):11. Published 2022 Apr 14. doi:10.3390/proteomes10020011

This is a collaborative project between Andrew T. Rajczewski, Dr. Qiyuan Han, Subina Mehta, Praveen Kumar, Dr. Charles G. Knutson, Dr. Pratik D. Jagtap, and Dr. James G. Fox under the direction of Drs. Natalia Y. Tretyakova and Timothy J. Griffin. Andrew T. Rajczewski processed samples acquired from Drs. Charles G. Knutson and James G. Fox. Andrew T. Rajczewski, Subina Mehta, and Praveen Kumar performed bioinformatic analyses using RNA-Seq data generated by Dr. Qiyuan Han. Andrew T. Rajczewski wrote the manuscript under the direction of Drs. Pratik D. Jagtap, Natalia Y. Tretyakova, and Timothy J. Griffin.

3.1 Introduction

Chronic inflammation has been linked to the development of many serious health problems, notably oncogenesis in several tissue types including those related to colorectal cancer^{262,263}. During inflammation, the continued release of regulatory cytokines which serve to mediate the immune response can promote tumorigenesis²⁶⁴ and metastasis²⁶⁵. In addition, chronic inflammation causes a burst of reactive oxygen species (ROS) and reactive nitrogen species (RNS) which can damage the host genome, contributing to oncogenesis via DNA damage and mutagenesis^{266,267}. The full picture of molecular changes which occur during chronic colon inflammation has the potential to advance our understanding of colorectal cancer etiology²⁶² as well as to seek opportunities for its diagnosis²⁶⁸ and identification of therapeutic targets for its treatment²⁶⁹.

Modern ‘omics technologies such as next-generation RNA sequencing (RNA-Seq) and mass spectrometry (MS)-based proteomics have allowed for marked advancements in studies of cancer¹⁷⁹. The use of transcriptomic data has allowed for characterization of gene expression within the microenvironments of tumors at various levels of development, providing a wealth of knowledge as to the specific disease biology associated with these conditions²⁷⁰. However, RNA-Seq is only able to assess the state of the transcriptome, which often does not match the gene products (the proteome) associated with a specific tissue or disease state²⁷¹. On the other hand, mass spectrometry-based proteomics can be used to quantitatively assess protein abundance in tumors relative to healthy tissue as well as to identify cancer biomarkers for early diagnosis and treatment²⁷².

In conventional “bottom-up” proteomics, MS data is searched against a reference FASTA database containing protein sequences encoded in canonical gene sequences for

the organism of interest, therefore excluding any proteins containing non-canonical sequences stemming from insertions, deletions, amino-acid substitutions, alternate splicing events, or any other atypical events leading to translation of proteins with unexpected amino acid sequences²⁷³. These non-canonical sequences are captured in RNA-Seq analyses, which detect all transcripts including those that may give rise to novel protein products.

Proteogenomics is a multi-omics approach which combines the completeness of RNA-Seq with the ability of MS-based proteomics to directly confirm the translation of these transcripts into expressed proteins with potential functional implications. Proteogenomics therefore provides a more complete molecular picture of inflammatory and cancer phenotypes as compared to approaches using a single omics technology^{274,275}. Proteogenomics uses RNA-Seq data to generate an expanded protein sequence FASTA database, which can be used to confirm the expression of proteoforms²³ containing both canonical and novel non-canonical peptide sequences. Although proteogenomics has been shown to be a powerful approach for studying cancer^{275,276}, the potential for false-positive matches to non-canonical sequences remains a concern¹⁵⁰, requiring methods to verify the accuracy of PSMs using bioinformatic and/or analytic approaches.

In this study, we developed and utilized novel proteogenomic workflows to analyze chronic inflammation in proximal colon tissues of a mouse model of IBD. Genetically engineered Rag2^{-/-}Il10^{-/-} mice have been used in previous studies as models of chronic inflammation^{277,278}, as animals with these mutations develop chronic colon inflammation when subjected to bacterial infection²⁷⁹. Rag2^{-/-}Il10^{-/-} mice were subjected to infection with *Helicobacter hepaticus* and allowed to develop chronic colon inflammation as

described in a previous study²⁶⁶. We first isolated and processed bulk proteins from proximal colon tissue for bottom-up proteogenomics and subjected them to LC-MS analysis. Using the Galaxy for Proteomics (Galaxy-P) software suite²⁸⁰, we utilized two automated computational workflows to generate and refine¹⁷¹ a transcriptome-derived FASTA database for proteogenomic analysis of the MS data. Finally, a rigorous bioinformatic workflow coupled with targeted MS methods was used to verify and validate non-canonical peptides. In total, our results provide unique insights into molecular signatures of inflammation in the colon and demonstrate a powerful proteogenomic pipeline for verification and validation of novel, non-canonical sequences derived from proteoforms underlying cancer-driving disease phenotypes.

3.2 Materials and methods

Materials

Proximal colon tissues were obtained from a previous study²⁶⁶. Triethylammonium bicarbonate (TEAB), urea, aprotinin, phenylmethanesulfonyl fluoride (PMSF), 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), dithiothreitol (DTT) and iodoacetamide (IAA) were obtained from Millipore Sigma (Burlington, MA). Trypsin was purchased from Promega Corporation (Madison, WI). Formic acid was purchased from Honeywell Fluka (Mexico City, Mexico). Acetonitrile and water were obtained from Thermo Fisher Scientific (Waltham, MA). Anhydrous acetonitrile was obtained from Glen Research (Sterling, VA).

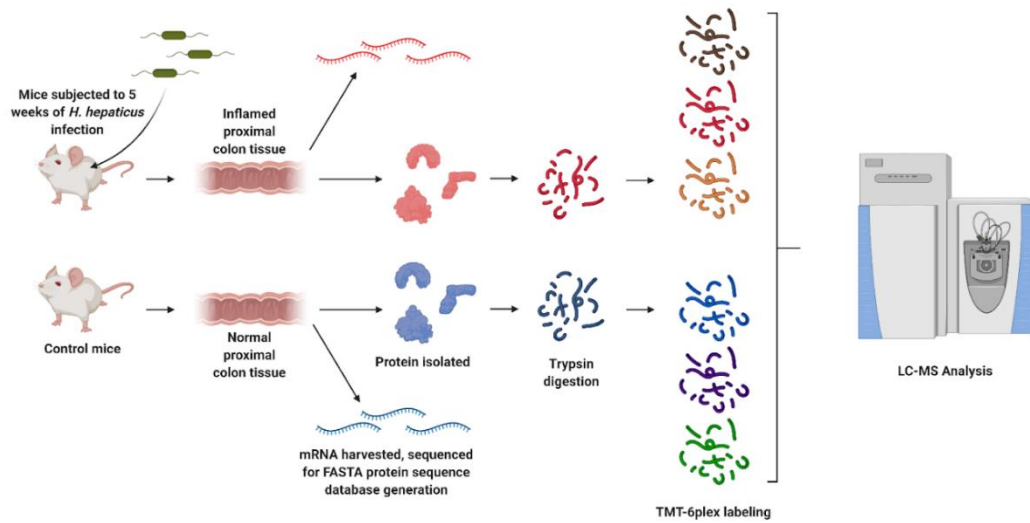
Kimble 1.5mL pestles were purchased from VWR International (Radnor, PA). Pall 10K Nanosep spin filters were utilized for digestion and were obtained from Millipore

Sigma (Burlington, MA). Pierce BCA Assay and Colorimetric Peptide Assay kits were obtained from Thermo Fisher Scientific (Waltham, MA). For isobaric labeling, TMTsixplex kit (lot #SH253249) was purchased from Thermo Fisher Scientific (Waltham, MA). For peptide desalting and fractionation, the Pierce High pH Fractionation kit was obtained from Thermo Fisher Scientific (Waltham, MA).

Treatment conditions, tissue and protein isolation and proteolytic digestion

Rag2^{-/-}Il10^{-/-} mice were subjected to three oral gavages over the course of one week with either saline (control) or *Helicobacter hepaticus* culture, after which the infected mice developed chronic colorectal inflammation (**Figure 3.1**)²⁶⁶.

Figure 3.1. Outline of the experimental procedure. Mice were infected with *Helicobacter hepaticus* to induce inflammation, and proximal colon proteins and mRNA were collected for proteogenomic analysis. Peptides were digested and labeled for differential proteomic analysis and variant discovery, with some unlabeled peptides reserved for quantitation of variants. Created with BioRender.com.



After 20 weeks, the mice were sacrificed, and colon tissues were collected for subsequent analysis. Our experiments utilized approximately 10 mg of proximal colon tissue harvested from control and infected mice (**Table 3.1**). These samples were placed in individual Eppendorf tubes containing 100 μ L of lysis buffer (25 mM TEAB, 8 M urea, 1 mM PMSF, and 2.5 μ g/mL aprotinin, pH = 8.5) and disrupted via grinding using 1.5 mL Kettle pestles. After homogenization, samples were subjected to probe sonication at 30% amplitude for 10 seconds over ice to lyse the cells; following lysis, samples were centrifuged at 15,000 rpm at 4°C for 15 minutes, after which the protein content was measured via Pierce BCA Assay. From each sample, 100 μ g aliquots of protein were added to individual Pall Nanosep 10K spin columns. The lysis buffer was then removed via centrifugation at 14,000g for 5 min, followed by the addition of 100 μ L of dilution buffer (25 mM TEAB, pH = 8.5). This was repeated twice more to remove the lysis buffer, with the proteins finally reconstituted in 100 μ L of dilution buffer. The proteins were then reduced via the addition of 20 μ L of DTT in the dilution buffer, followed by incubation at 55°C for one hour. Samples were alkylated with the addition of 10 μ L of 375 mM IAA to the spin columns, followed by a 30-minute incubation in the dark at room temperature. After alkylation, samples were washed with a further three iterations of centrifugation and the addition of 100 μ L of dilution buffer. Samples were finally reconstituted with 50 μ L of dilution buffer, to which 4 μ g of trypsin was added, and incubated at 37 °C overnight. After incubation, peptide samples were isolated by spinning the samples through the column filters. A further 50 μ L of digestion buffer was then added to the top of the spin columns and spun through via centrifugation. The peptide solution was then transferred to a fresh tube and the concentration determined through the Peptide Colorimetric Assay; 10 μ g of

peptides from each sample were then aliquoted into fresh vials and dried overnight under vacuum.

Table 3.1. Peptides generated from MS datasets in the construction of the library and validation in the patient datasets

Accession #	Mouse ID	Sample Type	TMT-6plex label
12-5632	5812	Control	126
12-5633	5813	Control	127
12-5634	5833	Control	128
12-5646	5874	H hepaticus infected	129
12-5647	5819	H hepaticus infected	130
12-5648	5820	H hepaticus infected	131

Peptide labeling, fractionation, and LC-MS/MS analysis

Peptides were labeled with TMT six-plex reagents for quantitative analysis. One dried-down aliquot of 10 μg from each sample was selected and reconstituted in 35 μL of 100mM HEPES, pH = 8.0. At the same time, TMT six-plex vials were brought to room temperature, after which the individual labels were reconstituted in 41 μL of anhydrous acetonitrile. Each peptide sample was then labeled via the addition of 10 μL of TMT labelling reagent (**Table 3.1**). The samples were then allowed to incubate for 2 h at room temperature, after which the reaction was terminated via the addition of 4 μL of 5% hydroxylamine and a further 15 min incubation.

Following incubation, the peptide concentrations of each labeled sample were measured using the Pierce Peptide Colorimetric Assay; thereafter, 5 μg of each of the six digested samples were concatenated into a single sample containing an equal amount of each of the labeled control and inflamed samples. The pooled sample was then desalted and fractionated using the Pierce High pH Fractionation Spin Columns using mobile phases containing 0.1% triethylamine and increasing amounts of acetonitrile into eight fractions (**Table 3.2**) which were collected, dried down under reduced vacuum, and reconstituted in 10 μL water containing 0.1% formic acid.

Table 3.2. Acetonitrile concentrations of eluted TMT-labeled peptides

Fraction #	Acetonitrile (%)
1	10
2	12.5
3	15
4	17.5
5	20
6	22.5
7	25
8	50

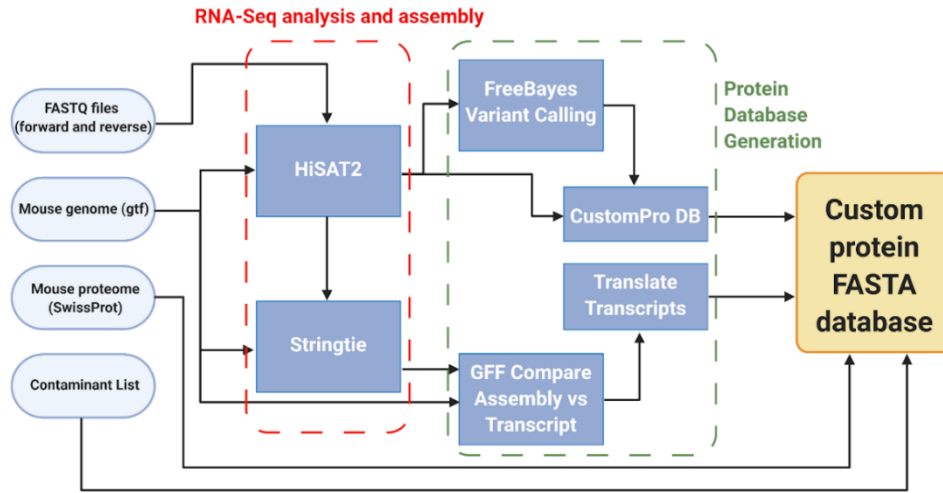
The eight fractionated peptide samples were analyzed on an Orbitrap Fusion Tribrid Mass Spectrometer interfaced with an Ultimate 3000 UHPLC. The UHPLC was run in nanoflow mode with a reverse-phase nanoLC column (35 cm x 250 μ m) packed with 5 μ m diameter Luna C18 resin. Samples were run on a 90-minute gradient with 5-22% buffer B (0.1% FA in acetonitrile) over 71 min, followed by 22-33% over 5 min, 33-90% over 5 min, a 90% buffer B wash for 4 min, and finally a 90-4% decrease in buffer B over 2 min followed by a 3 min equilibration at 4% buffer B. Samples were run at a flow rate of 300 nL/min. Peptides were analyzed in positive ion mode using a Top12 Full MS/dd-MS/MS experiment with an expected chromatographic peak FWHM of 15 seconds. In the full scan mode, resolution was 70,000 with an AGC target of 1e6, a maximum IT of 30 ms, and a scan range of 300 to 2000 m/z. Tandem mass spectrometry experiments were conducted at 17,500 resolution, AGC target of 5e4, maximum IT of 50 ms, an isolation window of 2.0 m/z and a normalized collision energy of 30. Data was collected in centroid mode.

Database construction

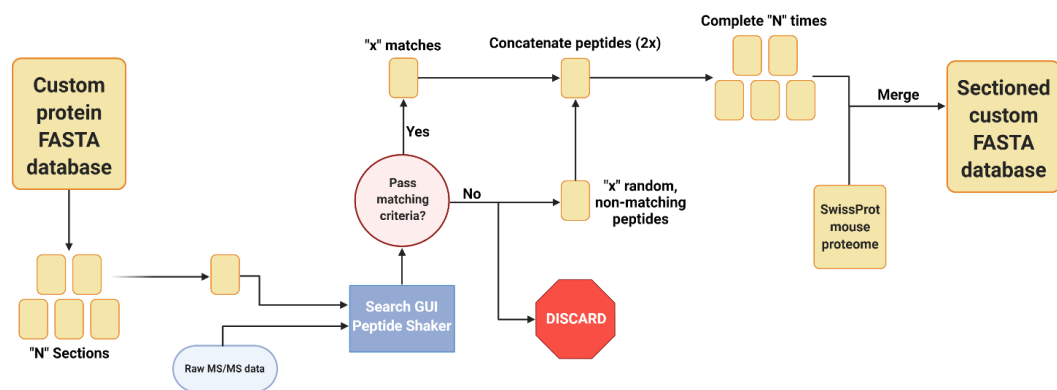
Computational work was performed using proteogenomics workflows and tools in the Galaxy for proteomics (Galaxy-P) suite^{281, 282} as well as in Proteome Discoverer v2.2 (Thermo Fisher Scientific (Waltham, MA)). Raw RNA sequencing data were acquired from proximal colon samples of six additional mice from the colon inflammation study, including three control and three treated samples (**Figure 3.2a**).

Figure 3.2. Galaxy-P-based bioinformatics workflows utilized in the study of inflamed colon proteogenomics a) Generation of RNA-Seq based custom protein FASTA database b) Sectioning workflow to reduce RNA-Seq FASTA database size c) Identification and verified of non-canonical variant peptides. All workflows created with BioRender.com

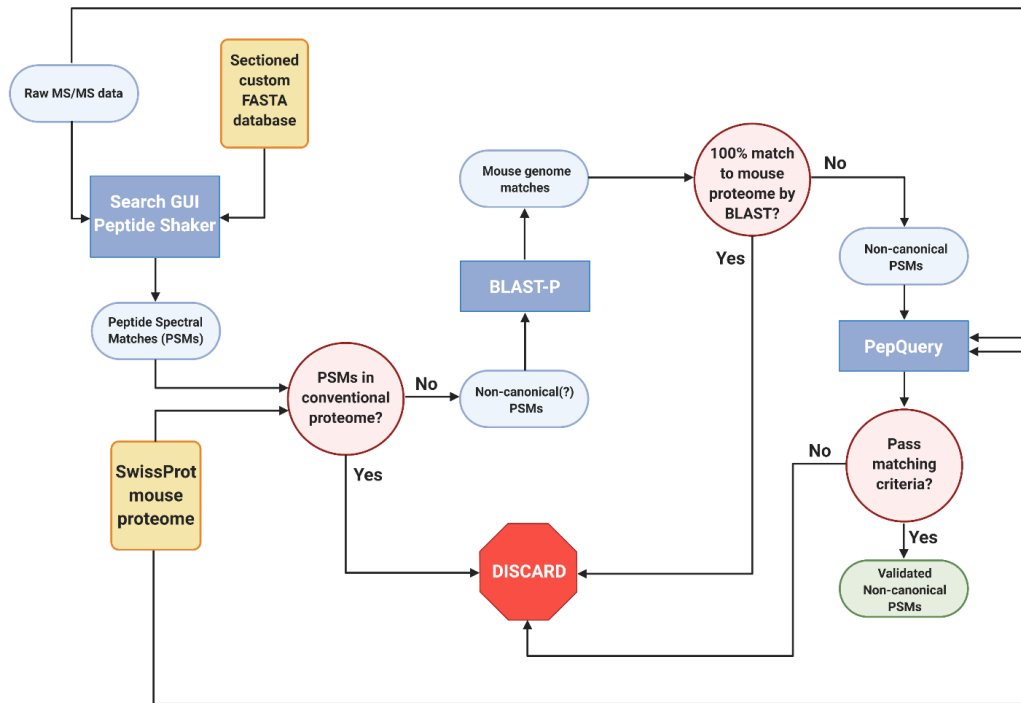
a)



b)



c)



Sequencing data was collected at the University of Minnesota Genomics Center on an Illumina HiSeq 2500 sequencer run in high output mode using 50bp paired end reads. These data were uploaded into Galaxy-P and used as an input for an integrated workflow²⁸¹ to generate a customized proteogenomic FASTA database. Briefly, the FASTQ files generated from these samples were paired with a murine genome annotation file and aligned via HISAT226 (v2.1.0); this is then used to create a list of genetic variants using the Free Bayes (v1.1.0.46-0) Bayesian genetic variant detector¹⁶⁴. This file is then utilized by the CustomProDB (v1.16.1.0) tool¹⁶⁶ to create FASTA sequences of the mapped indel, single amino acid variants, and alternatively spliced sequences identified. These variants are then concatenated together with the canonical Uniprot FASTA database and a list of common mass spectrometry contaminants²⁸³ as a custom RNA-Seq-based database. This workflow also uses StringTie²⁸⁴ (v1.3.3.1) to create an assembled gene transfer format file which is used to create a set of genomic coordinates complimentary to the RNA-Seq FASTA database used in downstream applications²⁸⁵, and effective for annotating other types of non-canonical transcripts not handled by CustomProDB.

Database sectioning

The resulting custom protein FASTA database was matched to MS/MS data to generate PSMs using a sectioning workflow created by Kumar et al.¹⁷¹ (**Figure 3.2b**), which provides increased sensitivity when working with large sequence databases while controlling false positives. The protein sequences in the database were randomly sorted into five smaller sections; each of these were used to search against the raw mass spectrometry data of the proximal colon samples using Search GUI¹⁶⁸ (v3.3.3.0), with N-

terminal and lysine TMTsixplex labeling as well as cysteine carbamidomethylation set as static modifications and methionine oxidation and phosphorylation at serine, threonine, and tyrosine were set as dynamic modifications. The X! Tandem search engine was used to identify peptides from the data against the individual batches. These results were then used by PeptideShaker²³⁵ (v1.16.4) to identify the proteins in the data against the individual batches. With the resulting PSM report, the proteins in each batch that were identified in the raw data with any level of confidence were retained, while the rest were discarded. For each protein in the batch that was retained, a discarded sequence was then selected at random and added back to the sectioned database. The five sections were then recombined back together to create a compact custom FASTA database enriched for protein sequences found in the inflamed colon samples.

Differential abundance proteomic and proteogenomic analyses

Raw mass spectrometry files were analyzed using Proteome Discoverer v2.2 in the TMT6 quantitation mode. The eight raw files were processed utilizing the basic Proteome Discoverer processing and consensus workflows designed for reporter ion quantitation. The murine SwissProt FASTA database was utilized for proteomics analysis, while the sectioned custom FASTA database with the RNA-Seq data-derived sequences was used for proteogenomics analysis. In all instances, carbamidomethylation at cysteine and TMT6 labeling at peptide N-termini and lysine residues were set as static modifications, while methionine oxidation and phosphorylation at serine, threonine, and tyrosine were set as dynamic modifications. Confidence for peptide identifications was set at an FDR cutoff of 0.01. The resulting PSM reports were used for quantitative analysis using MSstatsTMT²⁸⁶

using the “mstats” normalization algorithm. Gene ontology analyses were performed using the g-profiler package in R¹¹³, using an FDR cutoff of 0.05.

Identification, Verification and Validation of non-canonical peptides

Given the large numbers of proteins, the annotation of non-canonical peptides is more efficiently done using an automated workflow in Galaxy-P platform (**Figure 3.2c**). As with the sectioning workflow, the raw mass spectrometry data of the proximal colon tissue were searched against the custom protein FASTA database using SearchGUI and PeptideShaker. From the peptides that are identified, peptides from the murine reference and common contaminant reference proteomes were removed, leaving only potential non-canonical peptide sequences resulting from translation of unexpected genomic regions, novel splicing events, or amino acid coding sequence variants. These were then searched against the NCBI mouse proteome using Basic Local Alignment Search for Proteins (BLAST-P)²⁸⁷; these results were filtered to look for those search results which have imperfect sequence alignments due to sequence substitutions or gaps in the sequence²⁸². The genomic coordinates of these peptides are then determined using the PepPointer tool²⁸⁸ for further analysis and interrogation. Upon completion of the workflow, the identified non-canonical peptides were processed through an automated computational verification step using the PepQuery¹⁷³ tool in unrestricted modification search mode. Peptides were deemed to be valid if they had no matches to reference mouse or random peptides, had a p-value < 0.05, and no better scoring matches to any other peptides, such as reference peptides carrying a PTM. For PepQuery analysis, carbamidomethylation of cysteine residues as well as TMTsixplex labeling of N-termini and lysine residues were all set as

fixed modifications, while phosphorylation of serine, threonine, and tyrosine residues were set as variable modifications.

Validation and quantitation of non-canonical peptides

Peptides verified using PepQuery were further validated by targeted mass spectrometry analyses²⁸⁹ using 10 µg aliquots of unlabeled peptides reserved from the initial sample processing. The m/z values for molecular ions and MS/MS product ions of non-canonical peptides were determined from the original global analysis data and used to populate an inclusion list for use in targeted analyses (**Table 3.3**). For targeted analysis, samples were run on a Q-Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer interfaced with an Ultimate 3000 UHPLC run in nanoflow mode equipped with a nanocolumn packed with 5 µm diameter Luna C18 resin (15cm x 250µm). Samples were run on a 90-minute gradient with 5-22% buffer B (0.1% FA in acetonitrile) over 71 minutes, followed by 22-33% over 5 min, 33-90% over 5 min, a 90% buffer B wash for 4 min, and finally a 90-4% decrease in buffer B over 2 minutes, followed by a 3-minute equilibration at 4% buffer B. HPLC was conducted at a flow rate of 300 nL/min. The mass spectrometer was run in dual Full Scan and Parallel Reaction Monitoring mode; spectra were then analyzed in Skyline²⁹⁰ against a spectral library of non-canonical peptides generated using Prosit²⁹¹. Non-canonical peptides were identified by Skyline with at least three b- and/or y-ions, with peak areas of the detected product ions summed to represent the abundance of the peptide. The non-canonical peptide abundances were then tested for differential abundance using limma in R.

Table 3.3. Inclusion list for targeted detection of non-canonical peptides in proximal colon samples. Based on the initial global proteomics data, m/z values and charge states were determined for putative non-canonical peptides and used to create this inclusion list for targeted PRM analyses. Lower-case amino acid codes represent covalent modification by acetylation, oxidation, or phosphorylation.

Mass [m/z]	CS [z]	Polarity	Start [min]	End [min]	(N) CE	(N)CE type	Comment
403.90134	3	Positive	45.00	55.00	35	NCE	IQATLQLPQRR
404.58325	3	Positive	60.00	70.00	35	NCE	GIKPVTLELGK
405.17664	2	Positive	15.00	25.00	35	NCE	dPsAIGK
408.72577	2	Positive	15.00	25.00	35	NCE	dSILQAK
413.55450	3	Positive	35.00	45.00	35	NCE	FsmVVQDGIVK
417.73953	2	Positive	30.00	40.00	35	NCE	TSSISALR
418.72930	2	Positive	50.00	55.00	35	NCE	IGDYAGIK
422.73640	2	Positive	50.00	55.00	35	NCE	IGDYAGIK_H
426.24490	2	Positive	50.00	55.00	35	NCE	IGDYAGIK_2H
427.70746	2	Positive	40.00	45.00	35	NCE	DsILQAK
428.24850	2	Positive	50.00	55.00	35	NCE	IGDYAGIK_3H
429.73578	2	Positive	25.00	35.00	35	NCE	AASSANIPK
433.26210	2	Positive	50.00	55.00	35	NCE	IGDYAGIK_4H
434.20361	2	Positive	35.00	45.00	35	NCE	itNLER
434.88010	3	Positive	15.00	52.00	35	NCE	FsMVVQDGIVK
438.74319	2	Positive	15.00	25.00	35	NCE	tSSISALR
440.20490	3	Positive	35.00	45.00	35	NCE	FsmVVQDGIVK
443.27328	2	Positive	65.00	75.00	35	NCE	LLGIDLGGK
444.74811	2	Positive	45.00	55.00	35	NCE	iFSLNPR

446.76044	2	Positive	70.00	80.00	35	NCE	QVIYELK
448.70767	2	Positive	15.00	25.00	35	NCE	dsILQAK
449.77051	2	Positive	60.00	70.00	35	NCE	SKPAITGPK
453.24847	3	Positive	50.00	60.00	35	NCE	IGTGAmLPLEAVK
456.79178	2	Positive	70.00	80.00	35	NCE	EPILTVLK
462.28802	2	Positive	60.00	70.00	35	NCE	SAPLLLGPR
462.76016	2	Positive	30.00	35.00	35	NCE	ITEHSIPK
463.73135	2	Positive	15.00	25.00	35	NCE	IFsLNPR
469.72464	2	Positive	60.00	70.00	35	NCE	AAAsANIPK
470.74695	2	Positive	35.00	45.00	35	NCE	APPTWPGSK
471.75644	3	Positive	45.00	55.00	35	NCE	KANNINIQRR
478.72626	2	Positive	70.00	80.00	35	NCE	tsSISALR
486.74203	2	Positive	57.00	67.00	35	NCE	QVIyELK
487.91473	3	Positive	50.00	60.00	35	NCE	SKPcISGLmVPEK
496.77441	2	Positive	55.00	65.00	35	NCE	EPILtVLK
502.26819	2	Positive	40.00	50.00	35	NCE	sAPLLLGPR
507.74481	2	Positive	88.00	96.00	35	NCE	qVIyELK
509.92044	3	Positive	80.00	90.00	35	NCE	ILGAILAmAsTQSR
510.76004	2	Positive	35.00	45.00	35	NCE	sKPAItGPK
516.29022	2	Positive	35.00	45.00	35	NCE	VmPILLDSK
517.77618	2	Positive	45.00	55.00	35	NCE	ePILtVLK
522.76422	2	Positive	45.00	55.00	35	NCE	DLSLEGPEGK
523.26904	2	Positive	55.00	65.00	35	NCE	EEEGLEVLK
523.92719	3	Positive	50.00	60.00	35	NCE	iLGAILAmAsTQSR
525.56586	3	Positive	45.00	55.00	35	NCE	wTsEFEASLINR
530.79010	2	Positive	55.00	65.00	35	NCE	EVMLVGIGDK

531.95350	3	Positive	15.00	25.00	35	NCE	aAAAAAAAAAAAAASHSVA K
532.80383	2	Positive	75.00	85.00	35	NCE	aEPGLPLGLR
535.90820	3	Positive	45.00	55.00	35	NCE	aSLQVstLRLcR
537.29474	2	Positive	15.00	25.00	35	NCE	vmPILLDSK
543.76819	2	Positive	65.00	75.00	35	NCE	dLSLEGPEGK
544.28052	2	Positive	75.00	85.00	35	NCE	eEEGLEVLK
550.79913	2	Positive	65.00	75.00	35	NCE	QHFPsMILK
551.79053	2	Positive	70.00	80.00	35	NCE	eVMLVGIGDK
558.79291	2	Positive	95.00	105.00	35	NCE	QHFPsMILK
559.78656	2	Positive	45.00	50.00	35	NCE	eVmLVGIGDK
562.75055	2	Positive	25.00	35.00	35	NCE	DLsLEGPEGK
576.32092	3	Positive	125.00	135.00	35	NCE	VVLLGLsIPSLVGHR
578.35840	2	Positive	75.00	85.00	35	NCE	LSANLRLQK
579.79901	2	Positive	60.00	70.00	35	NCE	qHFPSMILK
589.78973	2	Positive	35.00	45.00	35	NCE	sLAALPEELR
590.77740	2	Positive	55.00	65.00	35	NCE	QHFPsMILK
602.97888	3	Positive	95.00	105.00	35	NCE	VVLLGLssIPSLVGHR
609.30970	2	Positive	35.00	40.00	35	NCE	GISNEGQNASIK
609.79706	4	Positive	115.00	125.00	35	NCE	vHAELADVLtEVVVDsVLA VR
611.82861	2	Positive	85.00	95.00	35	NCE	FSMVVQDGIVK
613.31680	2	Positive	35.00	40.00	35	NCE	GISNEGQNASIK_H
616.82530	2	Positive	35.00	40.00	35	NCE	GISNEGQNASIK_2H

618.82890	2	Positive	35.00	40.00	35	NCE	GISNEGQNASIK_3H
620.83240	2	Positive	35.00	40.00	35	NCE	GISNEGQNASIK_4H
622.84778	2	Positive	50.00	60.00	35	NCE	IQSTNQILEAK
629.95825	3	Positive	110.0 0	150.0 0	35	NCE	ARPVSSAASVYAGAGGsGS R
635.79163	2	Positive	60.00	70.00	35	NCE	YVEmSSVFHR
643.96631	3	Positive	80.00	90.00	35	NCE	aRPVSSAASVyAGAGGS R
644.35181	2	Positive	75.00	85.00	35	NCE	LAHLILsLEAK
645.33325	2	Positive	40.00	50.00	35	NCE	IQAtLQLPQRR
659.63855	3	Positive	90.00	100.0 0	35	NCE	NPTSVKYVEmsSVFHR
662.64783	3	Positive	117.0 0	127.0 0	35	NCE	NtPQLADIVATGFsvcGR
667.35437	2	Positive	85.00	95.00	35	NCE	gIKPVtLELGgK
683.83057	2	Positive	60.00	70.00	35	NCE	iQStNQILEAK
691.82530	2	Positive	62.00	67.00	35	NCE	TASEFDSAIAQDK
695.83240	2	Positive	62.00	67.00	35	NCE	TASEFDSAIAQDK_H
697.83600	2	Positive	62.00	67.00	35	NCE	TASEFDSAIAQDK_2H
701.34460	2	Positive	62.00	67.00	35	NCE	TASEFDSAIAQDK_3H
703.34810	2	Positive	62.00	67.00	35	NCE	TASEFDSAIAQDK_4H
723.37836	2	Positive	58.00	65.00	35	NCE	SKPcISGLMVPEK
726.85986	2	Positive	125.0 0	135.0 0	35	NCE	WTSEFEASLINR
731.36963	2	Positive	80.00	90.00	35	NCE	DIELVmAQANVSR
743.68970	3	Positive	105.0 0	115.0 0	35	NCE	pIRPGHyPASSPtAVHAIR

752.38068	2	Positive	100.0 0	110.0 0	35	NCE	sKPeISGLmVPEK
756.33984	3	Positive	105.0 0	115.0 0	35	NCE	PIRPGHyPASsPtAVHAIR
769.88850	2	Positive	79.00	89.00	35	NCE	ELGQSGVDTYLQTK
773.89560	2	Positive	79.00	89.00	35	NCE	ELGQSGVDTYLQTK_H
777.40420	2	Positive	79.00	89.00	35	NCE	ELGQSGVDTYLQTK_2H
780.35626	3	Positive	100.0 0	110.0 0	35	NCE	YVALDFEQEmAmAASSSSL EK
780.41110	2	Positive	79.00	89.00	35	NCE	ELGQSGVDTYLQTK_3H
783.40454	3	Positive	130.0 0	139.0 0	35	NCE	LLYAVNTHcHADHITGSGL LR
783.41480	2	Positive	79.00	89.00	35	NCE	ELGQSGVDTYLQTK_4H
786.07880	3	Positive	65.00	75.00	35	NCE	vHAELADVLTEVVVDsVLA VR
810.05829	3	Positive	95.00	105.0 0	35	NCE	LLYAVNtHcHADHITGSGLL R
810.34399	3	Positive	90.00	100.0 0	35	NCE	yVALDFEQEMAMAASSsSL EK

3.3 Results

3.3.1 Creation and sectioning of a custom RNA-Seq based FASTA database

Six sets of paired-end RNA-Seq data were obtained by sequencing RNA isolated from the proximal colons of Rag2^{-/-}Il10^{-/-} mice subjected to five months of *H. hepaticus*-induced inflammation along with matching controls (three animals per group, see **Figure 3.1**). Each of these sets was aligned and mapped to the mm10 mouse genome to create transcriptomic data for these samples; these individual sets of transcriptomic data were then converted to FASTA files representing the proteins that could potentially be translated from the sequencing data (**Figure 3.2a**). Concatenating these data together gave a combined RNA Seq-derived database that contained 1,402,947 sequences, corresponding to 1,348,407 protein sequences beyond the canonical mouse FASTA database.

As the large size of the RNA Seq-derived FASTA database would increase the likelihood of false positive PSMs while decreasing overall sensitivity for true positive PSMs¹⁶⁹, a sectioning workflow was utilized to create reduced RNA-Seq based FASTA database (**Figure 3.2b**). Use of the sectioning workflow reduced the RNA-Seq-derived FASTA database down to 423,071 protein sequences. Given that the workflow combines novel protein sequences detected in the raw data with an equivalent number of random sequences, the sectioned database corresponds to approximately 184,266 proteins containing non-canonical portions of their sequences derived from RNA sequences having PSMS in the proteomics data.

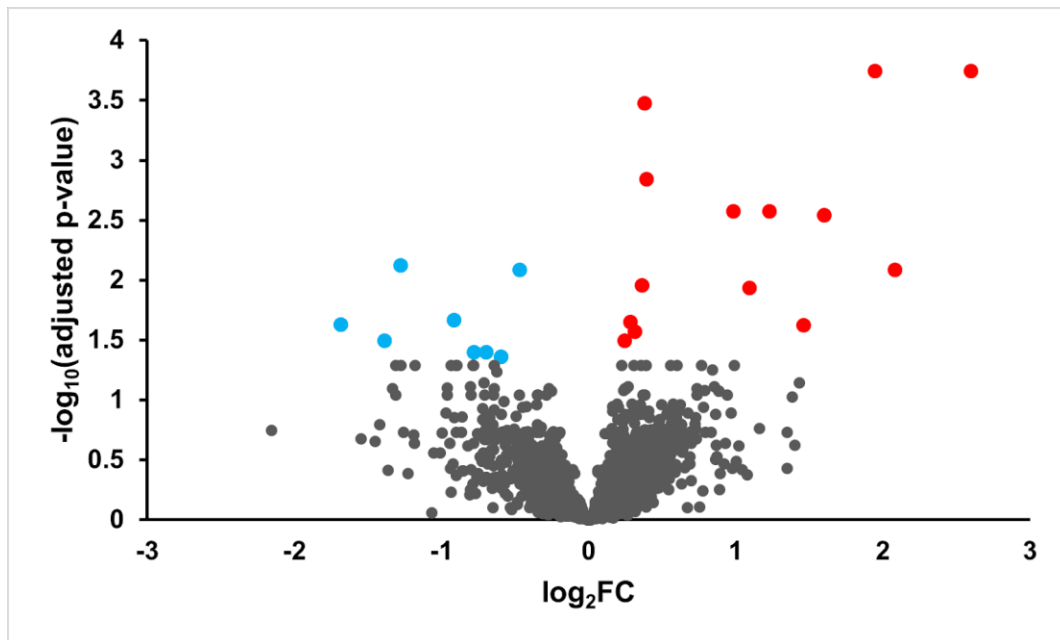
3.3.2 Global proteogenomic analysis reveals inflammation-driven changes in protein abundance

The reduced, sectioned proteogenomic FASTA database was merged with the reference mouse Uniprot database and a database of common MS contaminants, and the resulting merged database (proteogenomic database) was uploaded into Proteome Discoverer for global quantitative proteomic analysis of the inflamed proximal colon samples. For comparison, the mouse SwissProt FASTA database supplemented with common protein contaminants was also searched against the MS/MS data, offering a more conventional proteomic approach using a reference sequence database. Analysis of TMT-labeled peptides using the proteogenomic database identified 16,725 proteins in the proximal colon data grouped into 4865 protein groups. Of these protein groups, most were annotated proteins corresponding to entries within the mouse SwissProt FASTA database (91.7%). The rest of the identifications corresponded primarily to proteins containing non-canonical sequences generated in the database creation workflow in the Galaxy-P platform, with at least one peptide sequence identified as a part of the protein having a non-canonical sequence. Five of the identified protein groups corresponded to annotated proteins containing non-canonical sequences such as amino acid substitutions; 386 identified protein groups correspond to potentially novel proteins annotated solely by genomic coordinates (indicating novel truncations, proteins with retained introns/untranslated regions, previously untranslated regions of the genome, etc.), and 12 protein groups corresponded to known mass spectrometry contaminants. By contrast, the use of the conventional SwissProt FASTA database identified 8004 proteins organized into 4888 protein groups.

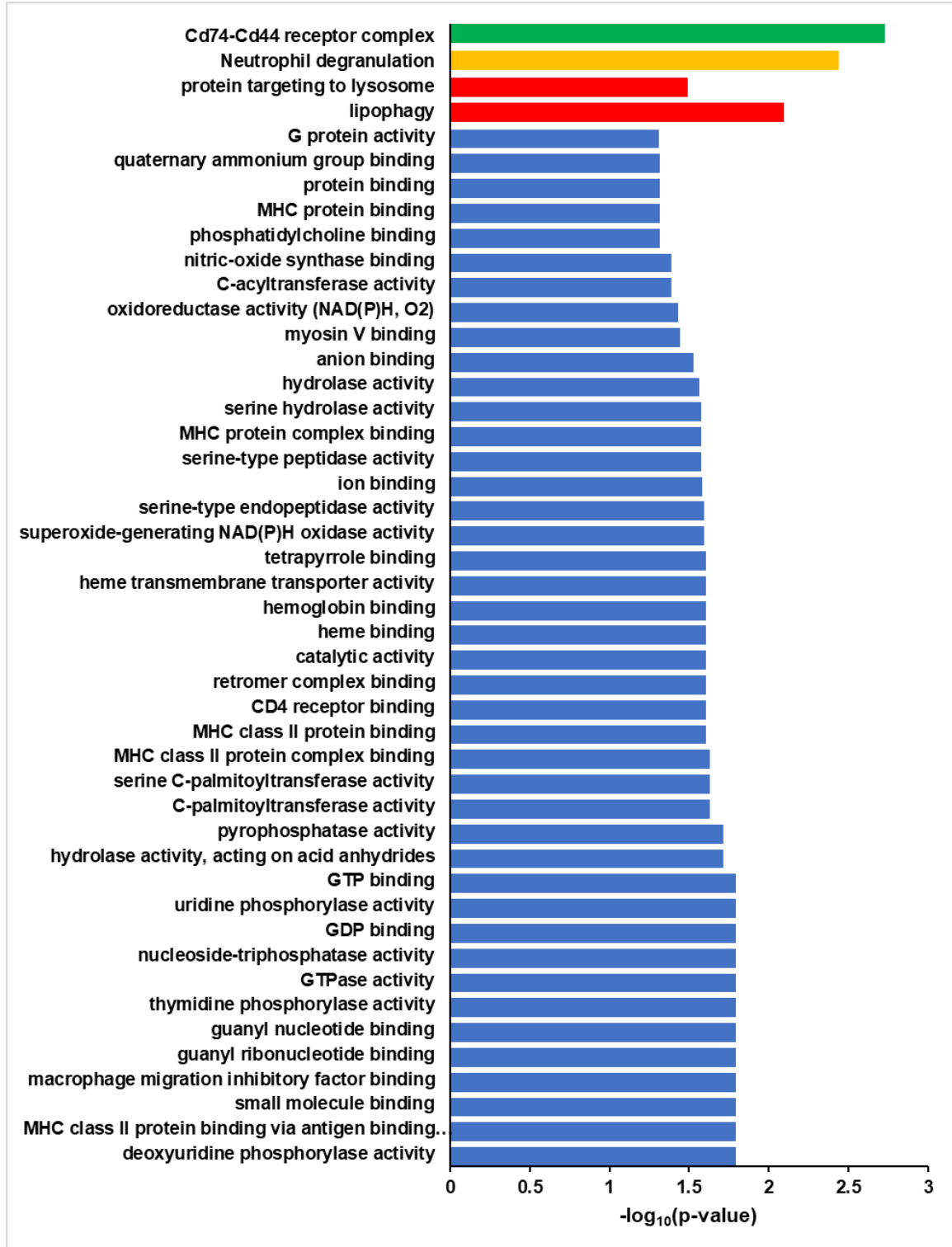
Differential analysis of the proteogenomics-derived results was performed to associate proteome abundance changes with phenotypic changes in the inflamed tissue samples. A volcano plot of the log₂ fold-change in protein abundance as a function of -log₁₀ corrected p-value (**Figure 3.3a**) shows that most proteins do not show significant change in abundance upon *H. hepaticus*-induced colon inflammation. Differential analysis shows a statistically significant (FDR < 0.05) increase in fourteen murine proteins and a decrease in eight murine proteins (**Table 3.4**).

Figure 3.3. Differential proteogenomic analysis of inflamed proximal colon samples. a) Enrichment of proteins in proximal colon tissue in response to chronic inflammation, as demonstrated via volcano plot of \log_2 fold-change of protein abundance against $-\log_{10}$ of corrected p-value. Proteins showing significant increases in abundance in inflamed tissues are highlighted in red, proteins showing decreased abundance in inflamed tissues are highlighted in blue. b) Gene Ontology analysis of increased abundance proteins in inflamed proximal colon samples shows enriched molecular functions (blue), biological pathways (red), reactomes (orange), and CORUM complexes (green). c) Gene Ontology analysis of decreased abundance proteins in inflamed proximal colon samples shows enriched molecular functions (blue), WikiPathways (brown), and CORUM complexes (green).

a)



b)



c)

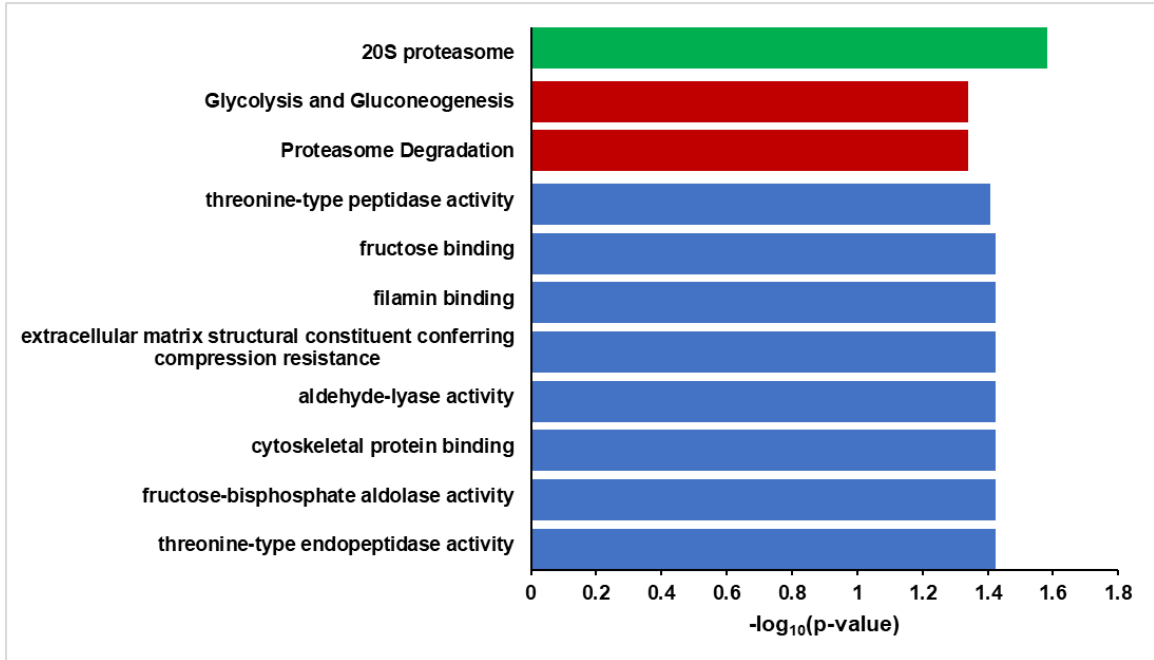


Table 3.4. Proteins identified as being increased in abundance in inflamed proximal colon tissue vs controls. Proteins shaded in red show increased abundance in inflamed proximal colon tissues, proteins shaded in blue show increased abundance in the control tissues.

Accession	Description	Gene	Coverage (%)	# Peptides	log ₂ F _C	p-value	q-value
Q61646	Haptoglobin	Hp	37	12	2.60	1.27E-07	1.79E-04
P07361	Alpha-1-acid glycoprotein 2	Orm2	11	3	2.08	2.90E-05	8.13E-03
P07146	Anionic trypsin 2	Prss2	17	3	1.94	1.41E-07	1.79E-04
P52624	Uridine phosphorylase 1	Upp1	37	9	1.60	7.86E-06	2.85E-03
Q61093	Cytochrome b-245 heavy chain	Cybb	1	1	1.46	1.50E-04	2.37E-02
P04441	H-2 class II histocompatibility antigen gamma chain	Cd74	30	8	1.23	5.77E-06	2.66E-03
STRG.18707.1_i_2_260	chr8: 73261429-73261687+	-	7	1	1.09	5.45E-05	1.15E-02
Q91X72	Hemopexin	Hpx	43	18	0.98	6.29E-06	2.66E-03
O35704	Serine palmitoyltransferase 1	Sptlc1	15	6	0.39	2.25E-06	1.43E-03

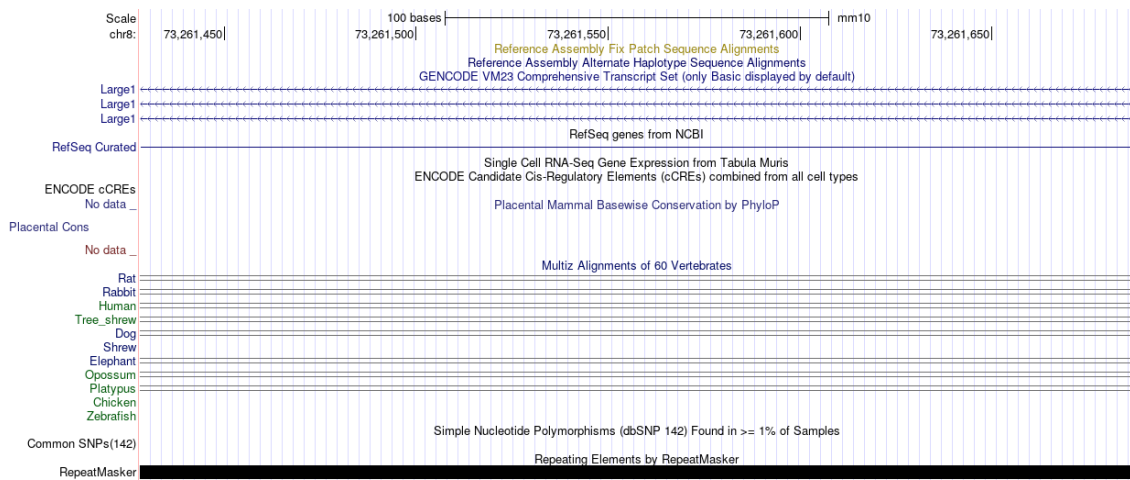
Q9CPW4	Actin-related protein 2/3 complex subunit 5	Arpc5	48	7	0.38	3.94E -07	3.33E -04
O35114	Lysosome membrane protein 2	Scarb2	14	6	0.36	4.75E -05	1.10E -02
P51150	Ras-related protein Rab-7a	Rab7a	64	12	0.31	1.79E -04	2.67E -02
Q9WTL2	Ras-related protein Rab-25	Rab25	44	8	0.28	1.23E -04	2.24E -02
Q921J2	GTP-binding protein Rheb	Rheb	28	6	0.24	2.40E -04	3.20E -02
A6ZI44	Fructose- bisphosphate aldolase	Aldoa	63	23	-0.47	3.20E -05	8.13E -03
P57016	Ladinin-1	Lad1	17	8	-0.60	3.74E -04	4.31E -02
Q62000	Mimecan	Ogn	37	9	-0.70	3.28E -04	3.96E -02
P35385	Heat shock protein beta-7	Hspb7	33	4	-0.78	3.25E -04	3.96E -02
Q7TQD2	Tubulin polymerization- promoting protein	Tppp	17	3	-0.91	1.10E -04	2.16E -02
O55234	Proteasome subunit beta type- 5	Psmb5	24	6	-1.28	2.36E -05	7.50E -03

Q99JI1	Musculoskeletal embryonic nuclear protein 1	Mustn 1	18	1	-1.39	2.29E -04	3.20E -02
Q19LI2	Alpha-1B- glycoprotein	A1bg	2	1	-1.68	1.38E -04	2.33E -02

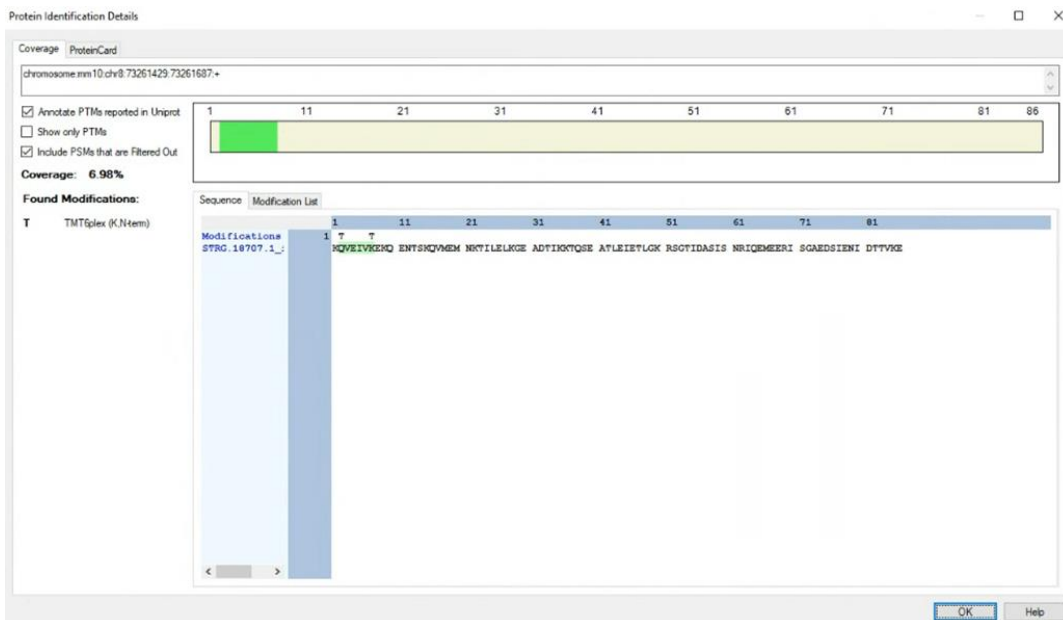
Gene ontology analysis of proteins with increased abundance in inflamed colon tissue shows enriched GO terms consistent with an inflamed system, showing an enrichment of molecular function GO terms such as MHCI and MHCII complex binding, macrophage migration inhibition factor binding, and oxidoreductase activity, along with the Neutrophil Degranulation reactome and Cd74-Cd44 receptor complex CORUM term (**Figure 3.3b**). Proteins that are decreased in abundance in inflamed tissues show enriched GO terms corresponding to molecular functions such as fructose aldolase activity, the glycolysis/gluconeogenesis and proteasome degradation wikipathway terms, and the 20S proteasome CORUM term (**Figure 3.3c**). Of the proteins found to be significantly increased in abundance in the inflamed proximal colon samples, one protein is unique to the proteogenomic FASTA database. This protein, STRG.18707.1_i_2_260, corresponds to mRNA translated from the (+) strand at chromosome 8, bases 73261429-73261687. This appears to be an untranslated region of the genome which complements the first intron of LARGE Xylosyl- And Glucuronyltransferase 1 (Large1) (Figure 3.4a). It should be noted that Proteome Discoverer only matched a single peptide QVEIVK at the N-terminus of the purported protein, comprising 7% of the entire sequence (Figure 3.4b, Table 3.4).

Figure 3.4. Non-canonical protein with differential abundance in the mouse model of colonic inflammation. The protein chr8: 73261429-73261687+ in the sectioned proteogenomic FASTA database was shown to be enriched in inflamed proximal colon samples. a) Genomic coordinates associated with chr8: 73261429-73261687+, visualized with the UCSC Genome Browser. b) Peptides associated with chr8: 73261429-73261687+ detected in Proteome Discoverer

a)



b)

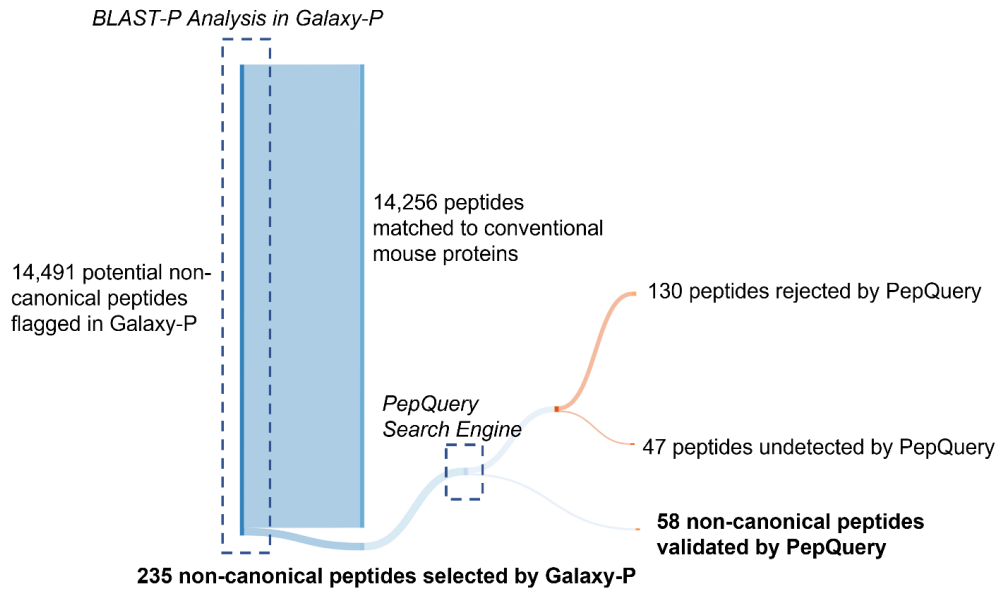


3.3.3 Galaxy-P provides peptide-centric discovery of non-canonical sequences

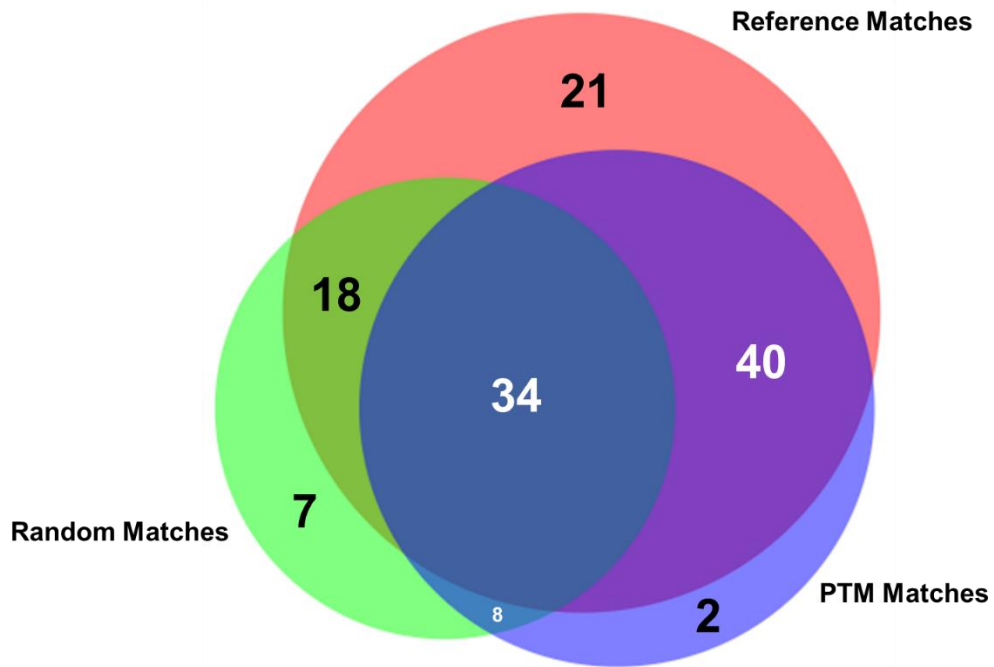
Isobaric quantitation strategy utilized in the global proteomics strategy is based on abundance measurements of proteins inferred from identified peptides which are labeled with the TMT-reagents; however, a peptide-level analysis is required to further verify and quantify non-canonical peptides belonging to unique proteoforms identified using the proteogenomic database. To this end, an additional workflow was utilized to identify non-canonical peptides in the inflamed proximal colon samples, which could be further verified and validated downstream. Analysis of the protein mass spectrometry data using Galaxy-P using the sectioned proteogenomic FASTA database revealed 14,491 peptides to protein sequences that had no direct sequence match in the canonical SwissProt mouse FASTA database. These peptides were then searched using BLAST-P to detect peptides mapping to the proteins, which carried non-canonical sequences. In filtering these results to remove any matches with 100% alignment to canonical sequences in the reference database, and matches with gaps of zero, the remaining peptide list was reduced to 235 peptides (**Figure 3.5a**). These peptides were hypothesized to correspond with novel proteoforms stemming from translation from unexpected genomic locations, splicing events, or non-synonymous coding sequence variants²⁸².

Figure 3.5. Validation of the non-canonical peptides results in the ultimate retention of 58 non-canonical peptides. a) The process of narrowing down the initial 14,491 non-canonical peptides using BLAST-P results in 235 peptides without matches to the conventional mouse proteome. Subsequent analysis by PepQuery results in 58 non-canonical proteins retained, with 130 peptides rejected by PepQuery. b) 130 non-canonical peptides rejected by PepQuery broken down along their reasons for failing PepQuery, specifically through finding a better match to a reference peptide, failing to pass the statistical barriers of the search engine, and/or matching to reference peptides with hypothetical post-translational modifications. c) Rejected non-canonical peptide spectral match (above) compared with a better scoring match to a reference proteome peptide (below). d) The use of the unrestrictive modification option demonstrates a superior match to a peptide with a modified sequence showing C-terminal a-type ionization, the loss of the alpha carbon and carboxyl group of the C-terminal lysine. e) PSM of a short rejected non-canonical peptide with repeated residues which can readily be matched to scrambled decoy peptides. Spectra generated using PDV²³⁷.

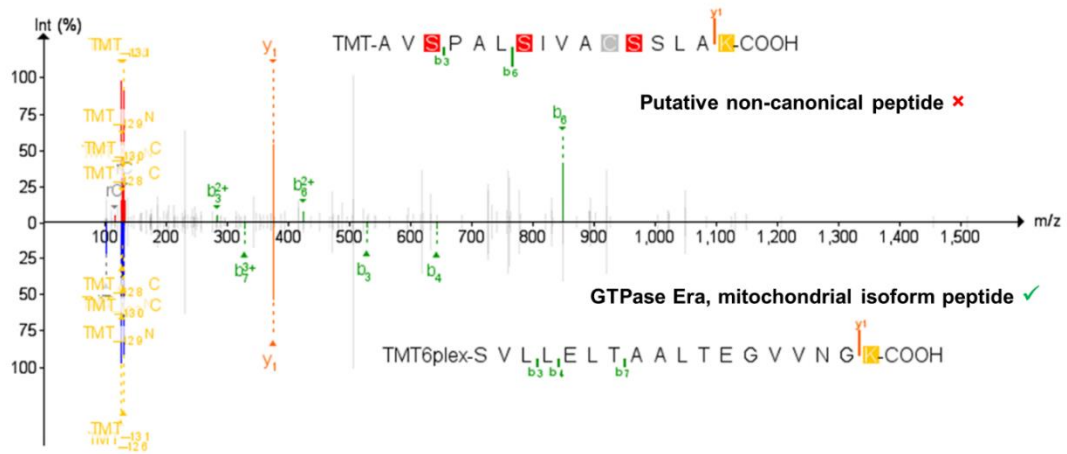
a)



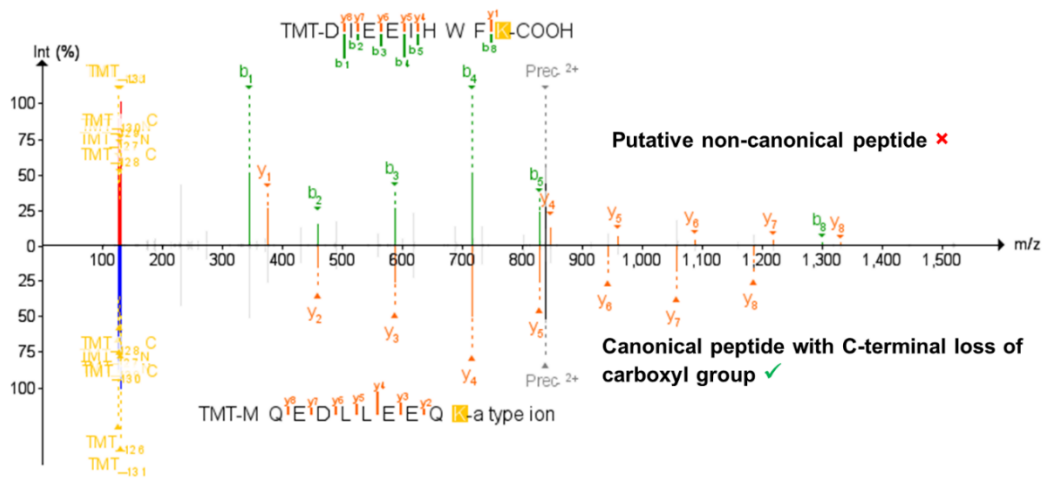
b)



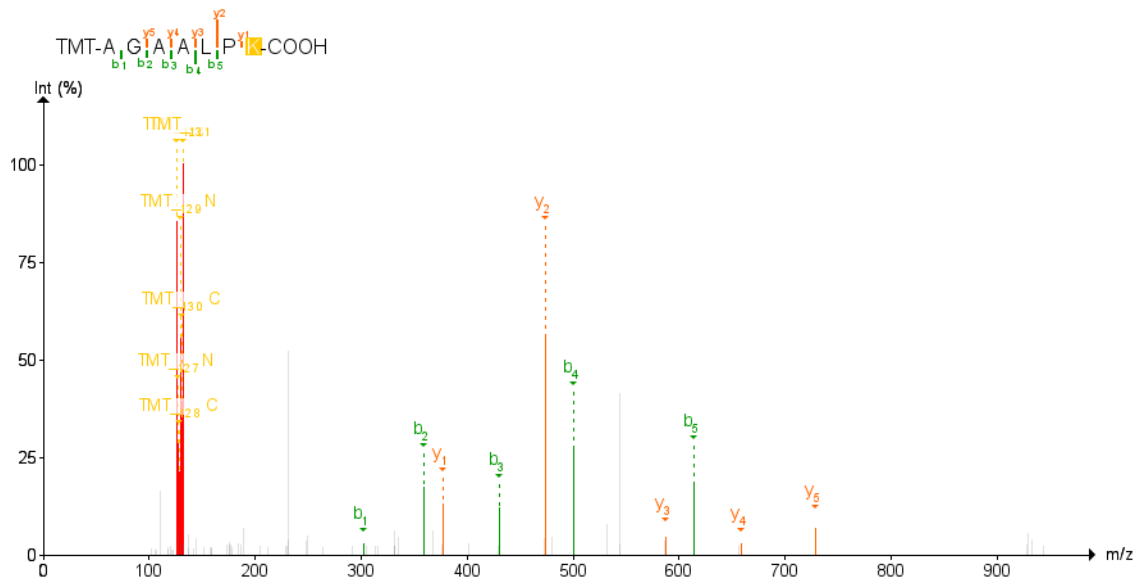
c)



d)



e)



3.3.4 PepQuery verifies the highest confidence non-canonical peptide candidates

To verify the variant peptides identified in inflamed proximal colon samples, we used PepQuery v1.338, implemented in Galaxy, on the 235 peptides identified in the discovery workflow. PepQuery provides a rigorous tool to evaluate the confidence of PSMs to non-canonical sequences, via testing for other possible matches (e.g., reference sequences, canonical sequences carrying PTMs) which may better match the MS/MS spectra in question. The list of 235 putative novel, non-canonical peptides was interrogated against the spectra of the TMTsixplex-labeled fractionated samples and compared to the canonical mouse Uniprot database. Unrestricted modification searching and single amino acid substitutions were performed as a part of the search to detect the strictest matches possible. To be considered passing matches, we used strict criteria where PepQuery had to deliver a p value of < 0.05 , rank = 1, and the number of unmodified PTM matches set to zero. Of the 235 non-canonical peptides, 58 were found to pass the strict verification criteria (**Table 3.5, Figure 3.6**) in at least one of the fractionated samples.

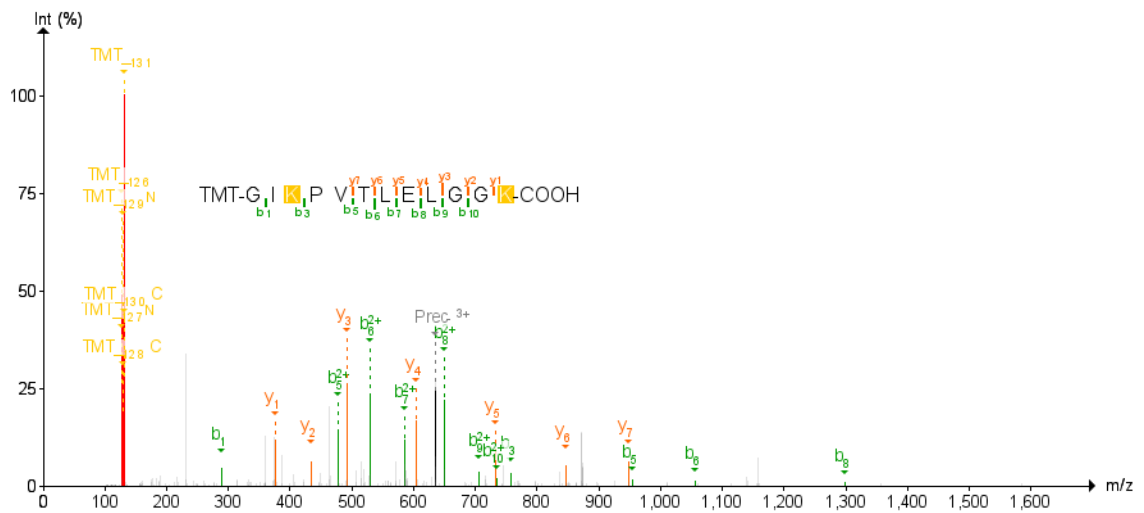
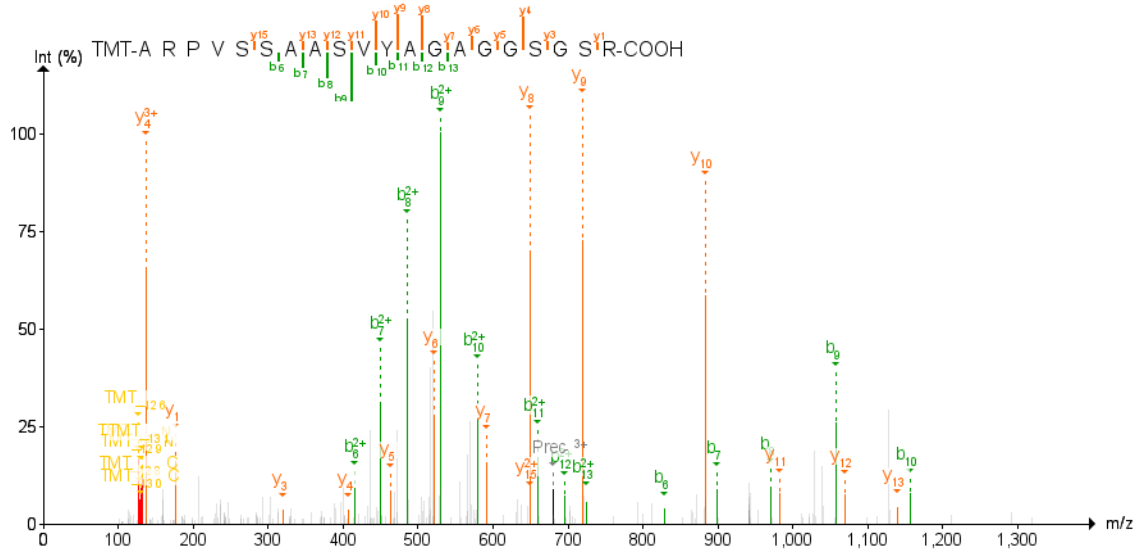
Table 3.5. Non-canonical sequence peptides identified, validated, and quantified in inflamed proximal colon tissues. Peptide precursors with at least three product ions (b- and/or y-ions) were detected in Skyline. A weighted contrast angle of the MS/MS spectra peaks against those of the reference library is reported in Skyline as the dot product, with a score of 1.0 representing a perfect match and 0.0 representing no match

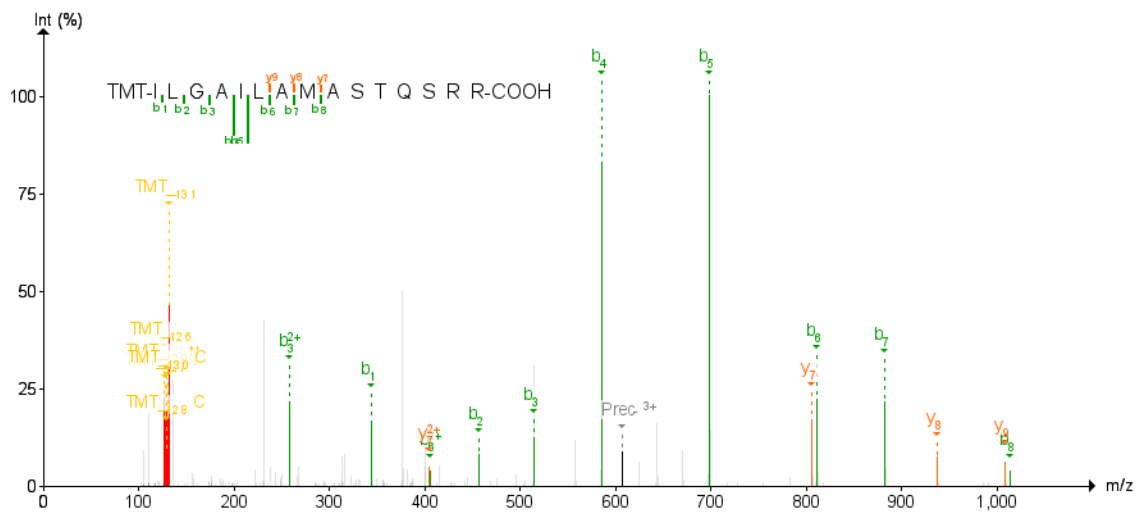
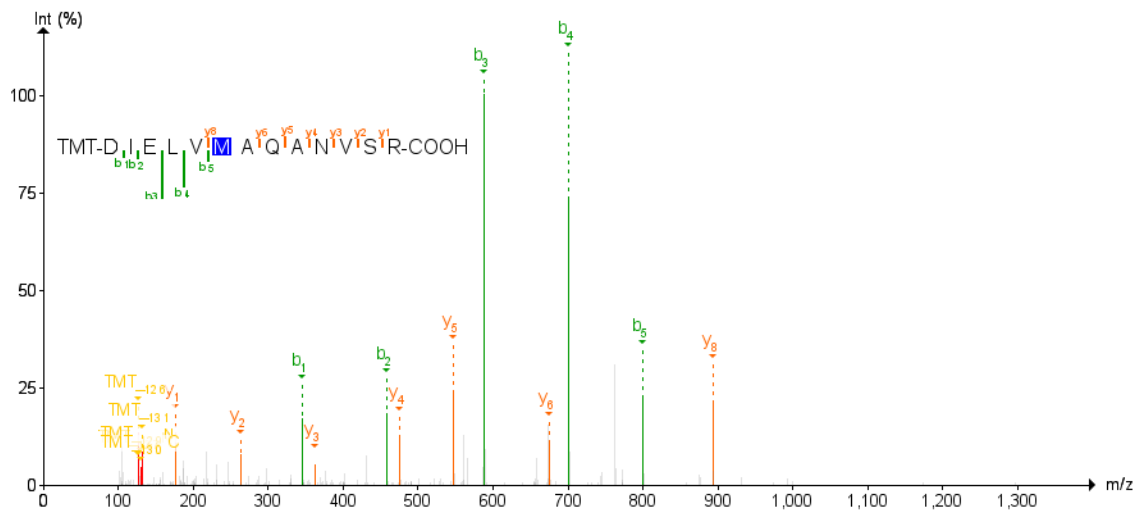
Peptide	Gene/Genetic Coordinates	Detected in Targeted	Skyline Dot Product
AAAAAAAAAAAAASHSVAK	Slc4a4	X	0.65
AASSANIPK	Sorl1	X	0.47
AEPGLPLGLR	Sec1	X	0.86
AGAVFLK	chr7:30972023-30972044		
AGPGRPAAAGGAAVRRR	Clstn1		
AMADELSEK	Nucb2		
APPTWPGSK	Slc1a5	X	0.41
ARPVSSAASVYAGAGGSGSR	Akap6	X	0.47
ASLQVSTLRLCR	Zfp219	X	0.71
CYVALDFEQEMAMVASSSLEK	chr2:130657397-130657463		
DIELVMAQANVSR	chr7:45080696-45080735	X	0.78
DIRQMINTESK	Morc3		
DLSLEGPEGK	Cenpv	X	0.95
DPSAIGK	chr4:109689395-109689416	X	0.61
DSILQAKL	chrX:5699255-5699279	X	0.78
EEEGLEVLK	Sft2d1	X	0.77
ELKEVIQR	chr9:107846694-107846718		

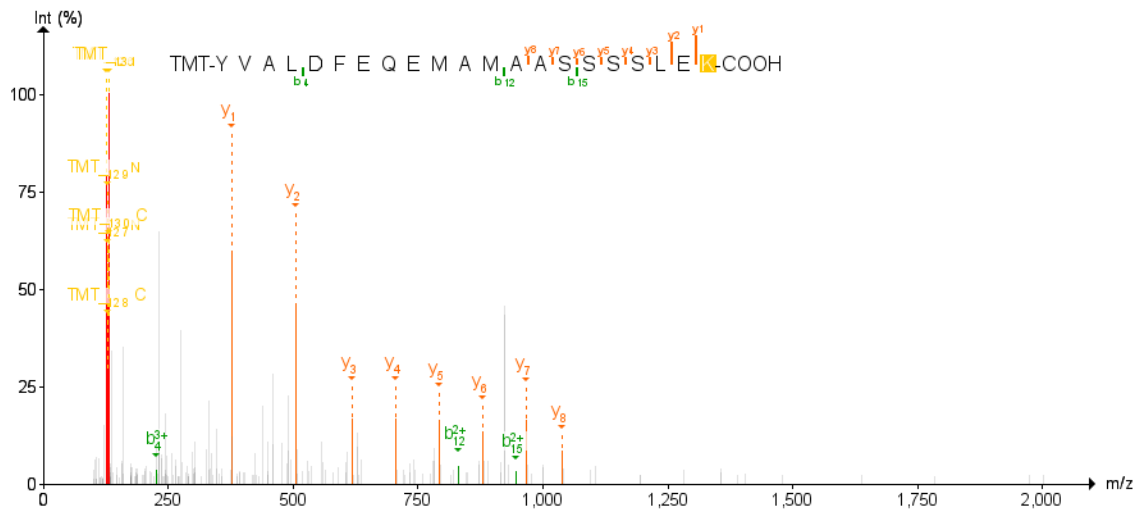
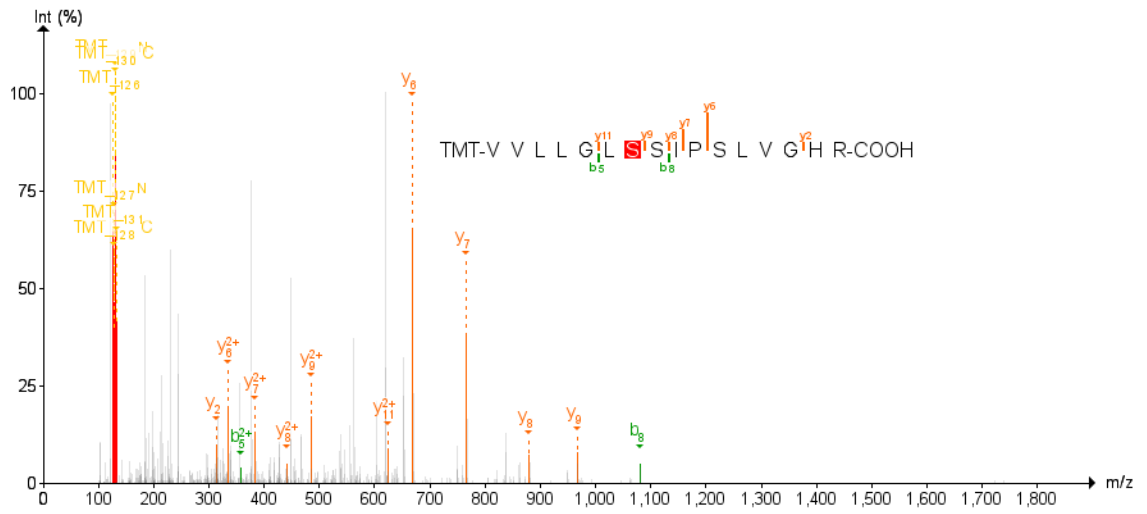
EPILTVLK	Lamb2	X	0.79
EVMLVGIGDK	Ppil4	X	0.56
FIVAIGK	Cpeb2		
FSMVVQDGIVK	chr11:64321586-64321619	X	0.61
GHARSSRMNAFPL	chr9:63755704-63755743		
GIKPVTLELGGK	Rcsd1	X	0.64
IFSLNPRSK	Cipc	X	0.79
ILGAILAMASTQSRR	chr12:55637554-55637599	X	0.78
IQSTNQILEAK	chr9:90120273-90120306	X	0.76
ITEHSIPK	Nol10	X	0.81
ITNLERGRER	chr17:66549926-66549956	X	0
KANNINIQRR	Scaf8		
KILQLVFR	chr16:46443630-46443654		
LAHLILSLEAK	chr15_KI270905v1_alt:3773278-31485159	X	0.43
LCYVALDFEQEMAMVASSSSLEK	chr2:130657394-130657463		
LGTGAMLPLEAVK	chr4:156226782-156226821	X	0.86
LLGIDLGGK	Huwe1	X	0.78
LLYAVNTHCHADHITGSGLLR	chr19:43526299-43526365	X	0.85
LPHLPSILEGRLK	Adprh12		
LQATLQLPQRR	Pm20d1		
LSANLRLQK	chr8:119530648-119530678	X	0.87
NPTSVKYVEMSSVFHR	Zfp729b	X	0.58
NTPQLADIVATGFSVCGRISIIRFPDVK	Gnb1l		
PIRPGHYPASSPTAVHAIR	chr2:149011574-149011631	X	0.79
QHFPSMILK	chr3:51956696-51956723	X	0.82

QVIYELK	Elmo1	X	0.75
RHQSAIVRR	Cdca4		
SAPLLLGPR	Cyp2s1	X	0.67
SFISLDRVTPR	chr4:16146372-16146405		
SKPAITGPK	Fendrr	X	0.76
SKPCISGLMVPEK	Glu1	X	0.62
SLAALPEELR	Fam214a	X	0.97
SSVRIGSGSWK	Adcy5		
TGDFQLHTNVNDGTEFGGSIYQK	Specc1l		
TSSISALR	Tgm7	X	0.78
VHAELADVLTEVVVDSVLAVR	1110038B12Rik	X	0.53
VMPILLDSK	Phkb	X	0.61
VLLGLSSIPSLVGHR	Fam107b	X	0.42
WTSEFEASLINR	chr14:57578727-57578763	X	0.81
YANNNSKY	Depdc5		
YVALDFEQEMAMAASSSSLEK	chr2:130463229-130463292		

Figure 3.6. Examples of MS/MS spectra for non-canonical peptides. Spectra of selected highly confident (p-value < 0.001) non-canonical peptides which passed PepQuery are presented here. Spectra were visualized using the Proteomics Data Viewer (PDV).







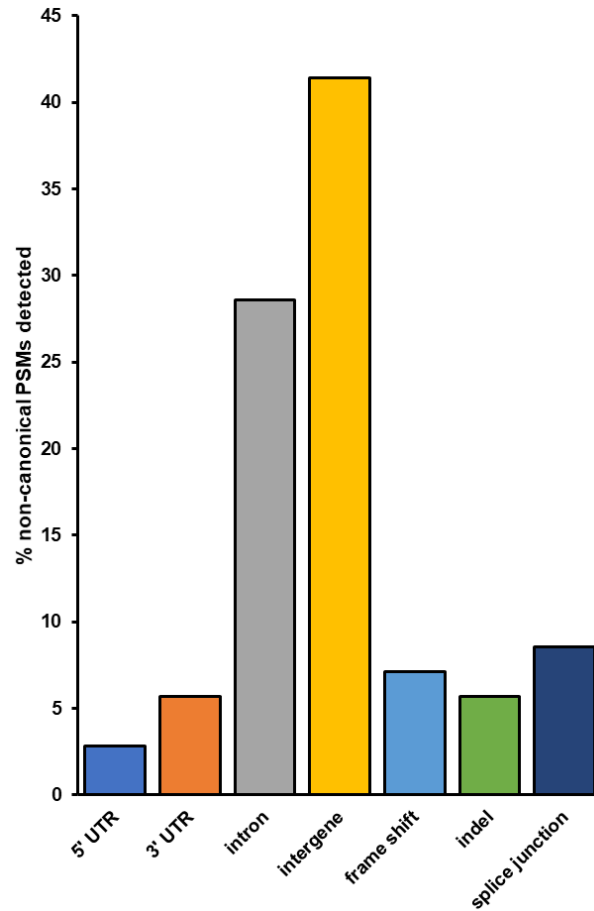
Among the 177 non-canonical peptides that did not pass PepQuery verification, 47 were unmatched by PepQuery to any spectra with sufficient quality scores and were not considered further (**Figure 3.5a**). The remaining 130 peptides had either superior matches to peptides in the reference FASTA database, an insufficient p-value matching the non-canonical sequence to pass statistical thresholds or matches to reference peptides containing potential PTMs. Interestingly, the non-canonical peptides which did not pass the PepQuery verification are not limited to each of these categories due to the possibility of matching an inputted peptide sequence to an MS/MS spectrum in any of the eight fractionated LC-MS runs in our data. As shown in **Figure 3.5b**, most of these non-canonical variants fail verification for multiple reasons, with 34 peptides failing for these three different reasons depending on the LC fraction-specific MS/MS files they are tested against (**Figure 3.5b**). Among non-canonical peptides which failed PepQuery verification for a single reason, the majority match to unmodified reference peptides with higher confidence than the non-canonical sequence (**Figure 3.5c**), followed by those assigned high PepQuery-derived p-values (**Figure 3.5e**), with only two peptides being rejected exclusively for matching reference peptides with PTM modifications (**Figure 3.5d**).

Among verified non-canonical peptides, the majority were found to be associated with intergenic regions not normally transcribed and translated into proteins (40.85%) as well as introns retained in the translated proteins (28.17%) (**Figure 3.7a**). The remaining variant peptides comprise indels, frameshifts, splice junctions, and sequences containing 5' and 3' untranslated regions. These peptides are derived from genes and intergenic regions found throughout the genome, excluding chromosomes 6, 18, and 20 (**Figure 3.7b**). Gene Ontology analysis of proteins corresponding to those non-canonical sequence peptides

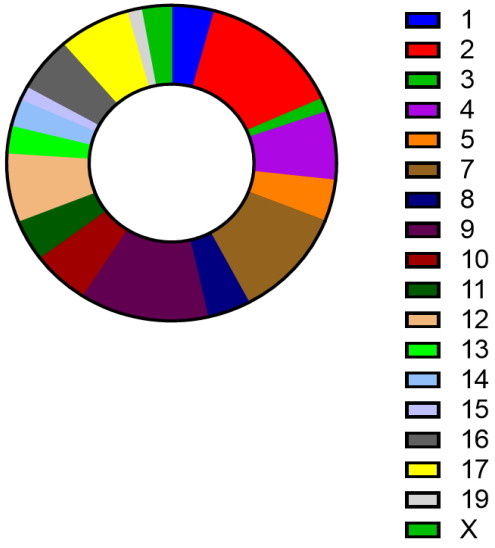
found within annotated genes showed no significantly enriched biological pathways common to this set of gene products.

Figure 3.7. Characteristics of non-canonical sequences validated by PepQuery a) Peptides with non-canonical sequences can be classified in several categories based on their altered sequence or location within a gene. b) Chromosomal locations of non-canonical sequence peptides correspond to mouse chromosomes throughout the genome.

a)



b)



3.3.5 Targeted proteomics experiments validate the presence of non-canonical peptides

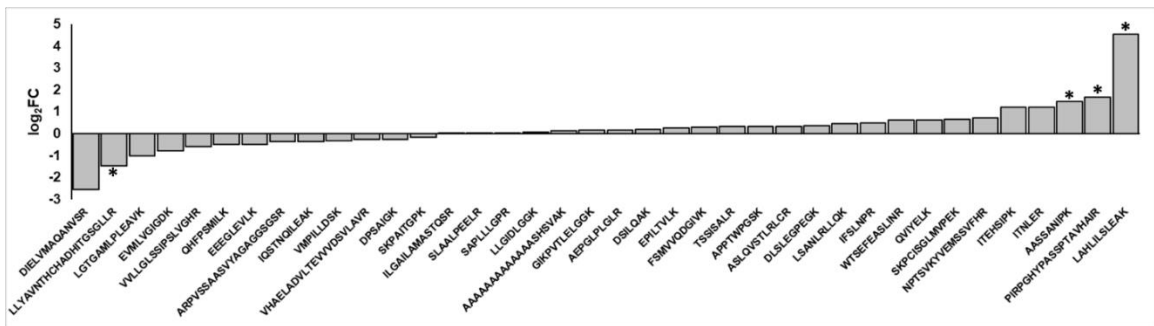
The non-canonical peptides were initially found using search and verification workflows mass spectrometry data for TMT-labeled, concatenated samples. Because TMT employs protein level-based quantification, we did not have a means to accurately quantify the non-canonical peptide sequences in the control and the inflamed colon samples. We therefore ran a separate set of targeted experiments to detect these novel peptide sequences from stored, unlabeled, and unfractionated samples. We used a targeted MS-based parallel reaction monitoring (PRM) assay based on empirically derived m/z and charge state values from the initial discovery-based analysis. The degree of variant abundance change in the inflamed samples was then expressed as the log₂ fold-change of inflamed versus controlled samples, for those peptides displaying confident PRM results (i.e., MS/MS spectra with at least three contiguous product ions in the b- or y-ion series).

Upon re-analyzing the samples, we found that of the 58 non-canonical peptides detected in the original TMT-labeled data, 38 were also detected in the targeted experiments with sufficient confidence (**Table 3.5**). Graphing the log₂FC of the reconstructed ion chromatograms for these peptides in inflamed versus control samples shows a general trend of half of the peptides being enriched upon inflammation and the other half being enriched in the control samples (**Figure 3.8a**); this pattern was mirrored in comparing the change in peptide abundance with the log₂FC of the RNA-Seq data of inflamed versus control samples, where there is a weak positive correlation between the two ($R^2 = 0.2392$) (**Figure 3.8b**). Ultimately, correcting for multiple hypothesis testing with limma in R found that the changes in abundance of these variants were not statistically

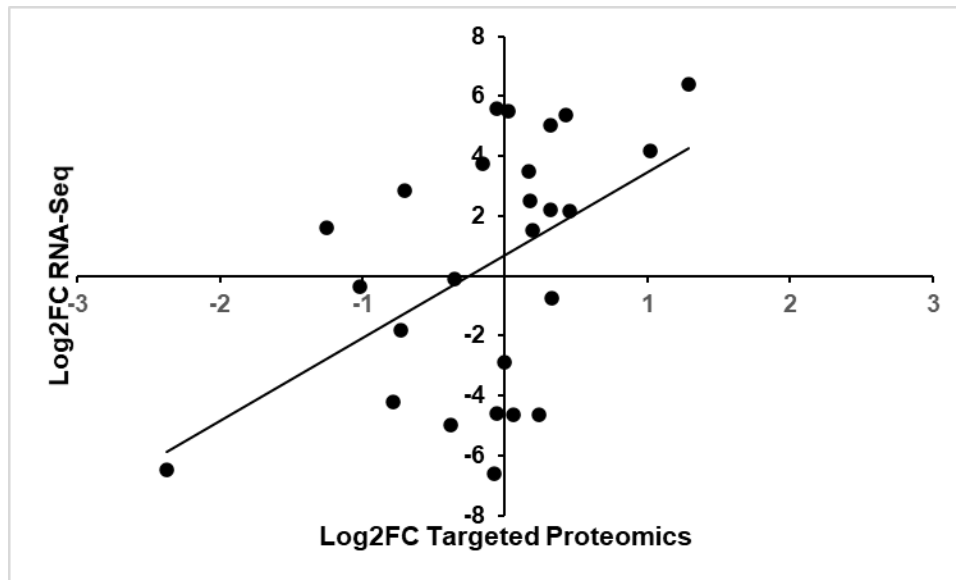
significant, though four peptides were found to have uncorrected p-values < 0.05 for enrichment or depletion upon inflammation. Of these, three non-canonical peptides showed an increased abundance in inflamed proximal colon samples; these corresponded to an intergenic peptide from chromosome 2 (PIRPGHYPASSPTAVHAIR), a peptide from chromosome 15 stemming from an alternative splicing event (LAHLILSLEAK), and a peptide corresponding to a retained 3-UTR section in Sortilin-related receptor Sorl1 (AASSANIPK, **Figure 3.9**). In addition, a non-canonical peptide corresponding to an intergenic region on chromosome 19 was found to be depleted in the inflamed tissue samples relative to the control.

Figure 3.8. Differential abundance analysis of non-canonical peptides detected in inflamed proximal colon samples. a) Fold-changes of variant peptides in the inflamed and control proximal colon samples, as measured via targeted mass spectrometry b) Comparison of RNA-Seq, proteomics-derived change in peptide abundances c) Categories of non-canonical peptides in peptides that show increased and decreased abundance in the inflamed proximal colon samples.

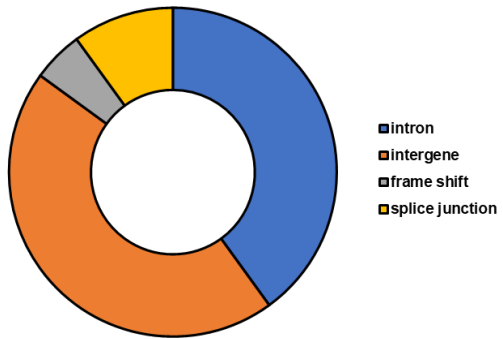
a)



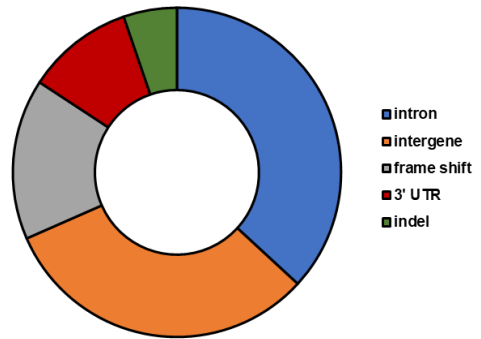
b)



c)

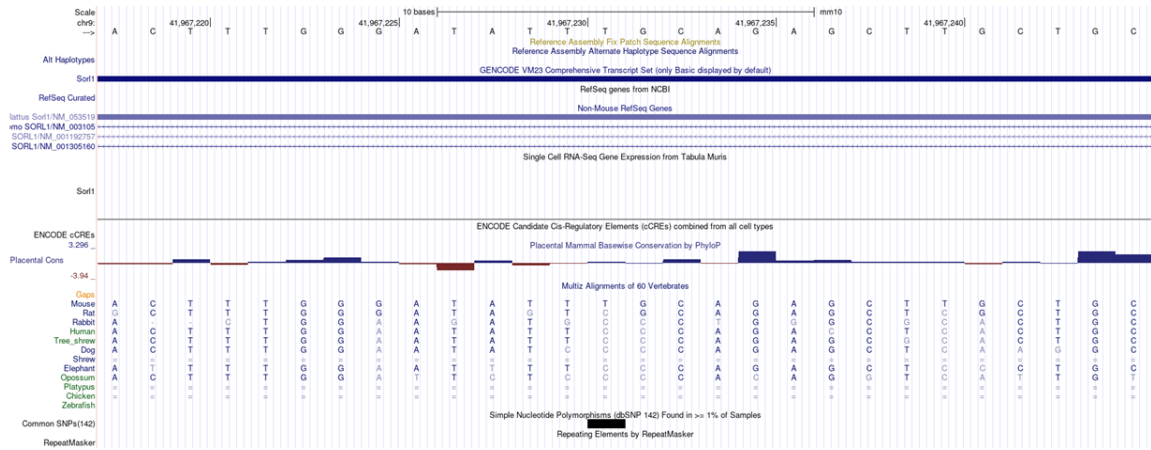


**Non-canonical types
increased in control samples**



**Non-canonical types
increased in inflamed samples**

Figure 3.9. The genomic coordinates of the peptide AASSANIPK, found in the 3' untranslated region of the Sorl1 gene. This peptide was found to have a slightly increased abundance in inflamed proximal colon tissue. Genomic coordinates determined via UCSC Genome Browser.



While the differences in abundances of the validated non-canonical peptides in inflamed samples and control tissues were not statistically significant, the variant peptides clustered into two groups that show a general trend of increased abundance in the inflamed tissue or increased abundance in the control sample (**Figure 3.8a**). There are notable differences between these two groups of peptides. In considering the type of variants present, intergenic regions and introns dominate both groups; however, the variant peptides that show increased abundance in the inflamed tissues are enriched for frameshifts, 3' UTRs, and indels (**Figure 3.8c**). In contrast, the variant peptides found to be decreased in abundance within the inflamed samples (and increased in the controls) contain splice junction variant peptides that are not seen at all in the group showing increased abundance.

3.4 Discussion

In this study, high-resolution protein mass spectrometry coupled with advanced proteogenomic analysis was utilized to characterize proteome dynamics of proximal colon tissue harvested from mice with chronic inflammation due to infection with *Helicobacter hepaticus*. The results were used to achieve several objectives: 1) Explore the quantitative changes of the proteome upon chronic colon inflammation, including expression levels of non-canonical protein sequences; 2) Demonstrate an integrated bioinformatic and targeted MS-based analytical workflow for verification and validation of non-canonical peptide sequences discovered via proteogenomics; 3) Utilize the knowledge from the verification and validation process as examples of pitfalls related to proteogenomic identification of non-canonical peptides that can inform more accurate studies using this multi-omic approach.

The mouse model utilized in our study, 129S6/SvEvTac-Rag2tm1FwIl10^{-/-} (Rag2^{-/-}Il10^{-/-}), has been widely used to model inflammatory bowel disease in humans^{277,278}. This mouse model was created via a knockout of the Recombinase activating gene 2(Rag2) and Interleukin-10 (Il10) gene. These double knockouts prevent the mice from forming mature T-cells or B-cells or in mitigating the development of chronic inflammation, respectively. As a result, Rag2^{-/-}Il10^{-/-} mice cannot resolve acute inflammation stages and will develop severe chronic inflammation, and eventually cancer, in their colon tissues.

The transition from chronic inflammation to oncogenesis occurs through a process of DNA damage accretion²⁹², epigenetic shifts²⁹³, and eventual phenotypic alteration. This process remains poorly understood and presents a rich landscape for research into biomarkers and therapy for early oncogenesis. In addition, while bottom-up proteomics has found great utility in the study of oncology, the use of conventional genome-derived FASTA databases results in non-canonical protein sequences being missed during data analysis. In this study, we explored the ability of proteogenomics approaches to identify novel protein variants, enabling a more complete characterization of protein dynamics in this model system.

Quantitative proteogenomics analysis utilizing isobaric peptide labeling with the TMT reagent detected several proteins showing increased abundance in the inflamed proximal colon samples. Three of these proteins: haptoglobin, hemopexin, and alpha-1-acid glycoprotein 2, were found to have increased abundance in the serum of Rag2^{-/-}Il10^{-/-} mice upon chronic inflammation. These three proteins were similarly identified in an earlier proteomics study of this model by Knutson et al.²⁹⁴, confirming the earlier findings and indicating their utility as biomarkers of global inflammation. The increased abundance

of Prss2, a serine protease involved in the remodeling of the extracellular matrix²⁹⁵, suggest that the inflamed proximal colon tissue can be considered to be well within the chronic inflammatory state²⁹⁶, making the five-month exposure of these mice a suitable model for chronic inflammatory bowel disease. This is further supported by the increased abundance of the H-2 class II histocompatibility antigen gamma chain Cd74 and the lysosome membrane protein 2 Scarb2, which are indicative of neoantigen generation and presentation to T cells²⁹⁷. Other proteins that increased in abundance are consistent with an inflammatory phenotype, including heavy-chain cytochrome b-245 (Cybb), a key component of NADH oxidase in phagocytes needed to create superoxides²⁹⁸, serine palmitoyltransferase 1 (Sptlc1), the initial enzyme involved in sphingolipid synthesis²⁹⁹, and GTP-binding protein Rheb (Rheb) which serves to activate mTOR1 and promote signal transduction³⁰⁰. These abundance changes to known factors of inflammation demonstrate the accuracy of the TMT-based quantitative proteomics strategy. The loss in abundance of muscle-specific proteins such as Aldoa (fructose-bisphosphate aldolase) and Mustn1 (musculoskeletal embryonic nuclear protein 1) may be due to alteration of the muscularis propria in the proximal colon in response to prolonged inflammation⁵⁴³⁰¹.

A major limitation when using TMT-labeling for quantitative proteogenomics is that this is a protein-centric method, which relies on quantitating inferred proteins from peptide sequence matches. When using proteogenomic approaches based on bottom-up MS-based proteomics, matches to non-canonical peptide sequences do not lend themselves to quantitation using this approach. Instead, more peptide-centric analysis is necessary to confirm the presence of these sequences and determine their potential abundance changes, which also reflects differential abundance of the proteoforms to which they belong.

To this end, we employed advanced peptide-centric proteogenomic bioinformatic workflows to identify non-canonical peptide sequences in an open discovery mode, followed by their verification using the PepQuery tool. The workflow first leverages BLAST-P to see whether putative non-canonical peptide sequences may instead match to other peptides in the conventional proteome; indeed, it was at this step that the STRG.18707.1_i_2_260 peptide QVEIVK was eliminated due to its perfect alignment somewhere else within the mouse proteome. PepQuery enables a rigorous verification of putative non-canonical sequences identified via upstream proteogenomic workflows, addressing a major challenge in proteogenomics to ensure confidence in these identifications¹⁵⁰. There are three ways in which the PepQuery search engine rejects potential non-canonical peptides, all of which were apparent on our inflamed proximal colon data and are dependent upon the quality of the PSM within each fractionated mass spectrometry experiment (**Figure 3.5b**). In the case of the putative non-canonical peptide AVSPALSIVACSSLAK identified in the first sample fraction, PepQuery can match the spectrum associated with this peptide (**Figure 3.5c, top**) as well or better to 44 peptides found within the canonical mouse proteome, including the GTPase Era, mitochondrial isoform peptide SVLLELTAALTEGVVNFK (**Figure 3.5c, bottom**), thus rejecting this PSM as identifying a canonical sequence. In another instance from the first fraction, spectra matched to peptides with several repeating residues such as in AGAALPK can potentially have their MS/MS matched to entries in randomized libraries generated in PepQuery, reducing the confidence in the PSM identification (**Figure 3.5e**). Finally, running additional stringent options in PepQuery such as unrestricted modification searching and/or amino acid substitution allows PepQuery to compare “non-canonical” PSMs with reference

proteome peptides containing PTMs or amino acid substitutions added in silico. This option resulted in the rejection of a PSM identifying the non-canonical sequence DIEEIHWFK in favor of a superior match to the canonical MQEQLLEEK with an a-type ion on the C-terminus corresponding to the loss of part of the C-terminal lysine (**Figure 3.5d**). Our results shown in this study provide a cautionary tale to others pursuing bottom-up proteogenomic studies, pointing to the need to carefully verify PSMs to putative non-canonical sequences.

During the final validation via targeted PRM mass spectrometry, 38 of these peptides could be detected and quantified by nanoLC-ESI-MS/MS, forming two similarly sized groups of peptides either showing abundance increase or decrease in the inflamed tissue compared to control. Parallel reaction monitoring allowed for deeper sampling of detected peptides to enable quantitation compared to the TMT-based discovery experiments, allowing us to explore the utility of these non-canonical peptides as quantitative indicators of inflammation, or potentially, early oncogenesis. Our inability to validate the remaining 21 of our peptide targets could be due to several factors, such as differences between the discovery and validation workflows (different instrument platforms, TMT-labeled peptides detected in the discovery versus non-labeled peptides in the validation, etc.), the lack of suitable peptide standards for targeted method construction or peptide quantitation, or potential issues with sample storage and degradation of the LC column used in the analyses. These questions make it difficult to know conclusively whether these sequences were not actually present, or simply were not detectable by PRM. Future studies would include the building of a targeted methodology using synthetic peptide analytes, reprocessing of desiccated protein digests that were saved from the initial

processing of the inflamed and control proximal colon samples using isotopically labeled internal standards for absolute quantification, in addition to initial optimization of the instrument with synthetic peptide standards prior to analysis.

The relevance of the non-canonical peptides detected in mouse proximal colon tissue to human inflammation and oncogenesis was examined via conversion of the mouse genome-coding coordinates for these peptides to analogous human genome coordinates via the LiftOver tool on the UCSC Genome Browser³⁰². Human gene sequences were then searched using the online PepQuery server against cancer-tissue derived mass spectrometry data from the Cancer Genome Atlas¹⁷⁷. While many non-canonical peptides did not have direct parallels within the human genome or breast, ovarian, and colon cancer datasets from the Cancer Genome Atlas, some sequences queried in the online PepQuery server did show evidence of human variant peptides that were from comparable genetic regions to the variants we observed in our analysis (**Table 3.6**). This demonstrates a potential for these peptides to serve as biomarkers for human oncogenesis.

Table 3.6. Human analogues of mouse non-canonical peptides. Human versions of murine non-canonical peptides found in TCGA datasets with PepQuery. Green-highlighted peptides show a decreased abundance in inflamed proximal colon samples while red-highlighted samples show an increased abundance in inflamed proximal colon samples.

Peptide	Lifted Human Coordinates	TCGA Colon Cancer	TCGA Breast Cancer	TCGA Breast Cancer PP	TCGA Ovarian	TCGA Ovarian PP	TCGA Ovarian GP	VU CC	PN NL CC	PN NL CC PP
AAAAA AAAAA AAASHS VAK	chr4:72053029-72053086			x						
ARPVSS AASVYA GAGGSG SR	chr14:33082441-33082494		x	x	x			x		
ASLQVS TLRLCR	chr14:21559259-21559295		x		x					
ELKEVI QR	chr3:4915974-49159798		x	x	x	x				

FSMVVQ DGIVK	chr17:1 366516 2- 136651 95		x		x			x		
LGTGAM LPLEAV K	chr1:90 3676- 903724	x			x					x
LLGIDL GGK	chrX:53 654138- 536542 05		x	x				x		
SKPAITG PK	chr16:8 653175 1- 865422 70	x	x	x	x	x	x	x	x	x
SKPCISG LMVPEK	chr1:18 235775 8- 182357 797	x	x	x					x	
VHAELA DVLTEV VVDSVL AVR	chr6:31 802600- 318030 93	x	x	x	x	x	x	x	x	x

VLLGL SSPSLV GHR	chr10:1									
	456282									
	4-		x	x	x	x	x		x	
	145628 72									

Beyond the results on protein response in colon inflammation and lessons learned in the verification and validation process, a significant deliverable of this work is a novel bioinformatic workflow for discovery and verification of non-canonical peptide sequences identified via proteogenomics. This easy-to-use, open-source and accessible Galaxy-based workflow allows researchers to avoid some of the pitfalls inherent to identifying novel non-canonical peptide sequences. As the workflow is currently focused on verifying novel PSMs, future iterations will incorporate tools for peptide-level quantitative analysis of non-canonical sequences³⁰³.

In summary, in this study we utilized proteogenomic analysis to characterize the protein composition of proximal colon tissue isolated from proximal colon tissues of Rag2^{-/-}Il10^{-/-} subjected to chronic inflammation via infection with *Helicobacter hepaticus*. Using RNA-Seq data from the same samples, we were able to detect differential abundance of several proteins known to be associated with inflammatory response. In addition, we were able to demonstrate the detection and quantitation of non-canonical peptides derived from novel proteoforms in the inflamed protein samples which, pending further study, may have potential as biomarkers for the early stages of colon cancer. We also provided insights into challenges involved in verifying and validating peptide sequences of interest identified in proteogenomic studies. Finally, we demonstrated the use of an accessible bioinformatic workflow for verifying non-canonical peptides discovered via proteogenomics, which illuminated pitfalls related to these identifications and should prove useful for others seeking to employ this approach in their research.

IV. MULTI-OMIC ANALYSES OF LPS- AND CIGARETTE SMOKE-DRIVEN INFLAMMATION IN TYPE II PNEUMOCYTES

This is a collaborative project between Andrew T. Rajczewski, Dr. Qiyuan Han, Dr. Jenna Fernandez, Nicholas Weirath, Alexander Lee, Donna Seabloom, Dr. Thomas Kono, Dr. Luke Erber, and Dr. Timothy Wiedmann under the direction of Drs. Natalia Y. Tretyakova and Timothy J. Griffin. Dr. Jenna Fernandez designed the cigarette smoke exposure chamber, and Dr. Jenna Fernandez and Andrew T. Rajczewski optimized exposure conditions under the guidance of Dr. Timothy Wiedmann. Andrew T. Rajczewski, Dr. Jenna Fernandez, Dr. Qiyuan Han, Alexander Lee, and Donna Seabloom conducted animal exposures and monitored cigarette smoke dosages. Andrew T. Rajczewski, Dr. Qiyuan Han, Dr. Jenna Fernandez, Nicholas Weirath, and Alexander Lee harvested and prepared lung samples for cell sorting and isolated biomolecules for analysis. Andrew T. Rajczewski and Dr. Luke Erber optimized the sample preparation protocols. Andrew T. Rajczewski performed protein digestion, peptide labelling, and fractionation as well as LC-MS and bioinformatic analysis of proteins from mouse samples. Drs. Qiyuan Han and Jenna Fernandez conducted RNA-seq and RRBS/oxRRBS with the University of Minnesota Genomics Center. Drs. Thomas Kono and Qiyuan Han conducted bioinformatics analysis for the RNA-seq and RRBS/oxo-RRBS data. Andrew T. Rajczewski wrote the manuscript under the direction of Drs. Natalia Y. Tretyakova, and Timothy J. Griffin.

4.1 Introduction

Lung cancer is responsible for the greatest number of cancer related deaths worldwide, representing the third highest cause of all projected new cancer cases and the highest cause of projected cancer deaths in the United States alone in 2022³⁰⁴. A major risk factor for lung cancer development is tobacco smoking, which remains popular with large portions of the global population despite years of public health advocacy, especially in East Asia and the Middle East³⁰⁵. While there is a strong correlation between lung cancer development and exposure to tobacco smoke, only some 10-15% of individuals who use tobacco products will eventually develop lung cancer³⁰⁶, indicating the presence of other confounding risk factors and genetic/epigenetic factors mediating the risk.

One of the important risk factors contributing to tobacco smoke-induced oncogenesis is inflammation. In a study by Chung-Han Ho et al.³⁰⁷, the incidence of lung cancer in cigarette smokers was shown to be significantly greater in those smokers that developed chronic obstructive pulmonary disorder (COPD). The alveoli and airways of COPD patients are damaged as a result of chronic inflammation, which generates reactive oxygen species and can lead to genetic and epigenetic changes³⁰⁸. In order to directly examine the impact of pulmonary inflammation on smoking-driven lung cancer, Melkamu et al.³⁰⁶ exposed A/J mice to 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), a prominent tobacco carcinogen, supplemented with lipopolysaccharide (LPS), a bacterial endotoxin known to trigger an inflammatory response³⁰⁶. Animals which were exposed to both chemicals developed a higher number of lung tumors of greater size than animals exposed to NNK alone, confirming the role of inflammation in lung oncogenesis.

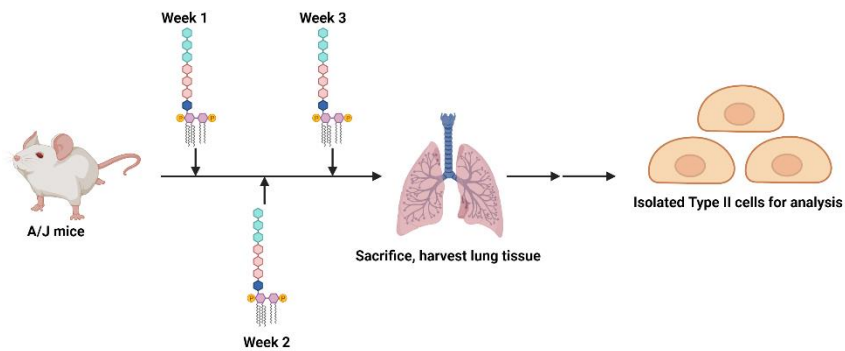
However, the exact mechanism by which inflammation contributes to smoking-induced cancer remains largely unknown.

In the present study, we applied a multi-omics approach to characterize epigenetic effects of cigarette smoke and inflammation in the lung using the A/J mouse model. This approach integrated proteomics, transcriptomics, and epigenomics to determine the molecular mechanisms by which inhalation exposure to cigarette smoke and lung inflammation alone contribute to oncogenesis. Our study focused on epigenetic changes in the target cells for smoking induced lung cancer, Type II pneumocytes. Type II cells are endothelial cells that constitute alveoli along with Type I pneumocytes and capillary endothelial cells. While their primary function is to modulate the permeability of the alveolus to gases via the secretion of mucus³⁰⁹, Type II pneumocytes are known to be the cells of origin for adenocarcinomas stemming from cigarette smoke exposure^{310,311} making them the ideal candidate for a multi-omic investigation into inflammation and potential oncogenesis stemming from cigarette smoke exposure.

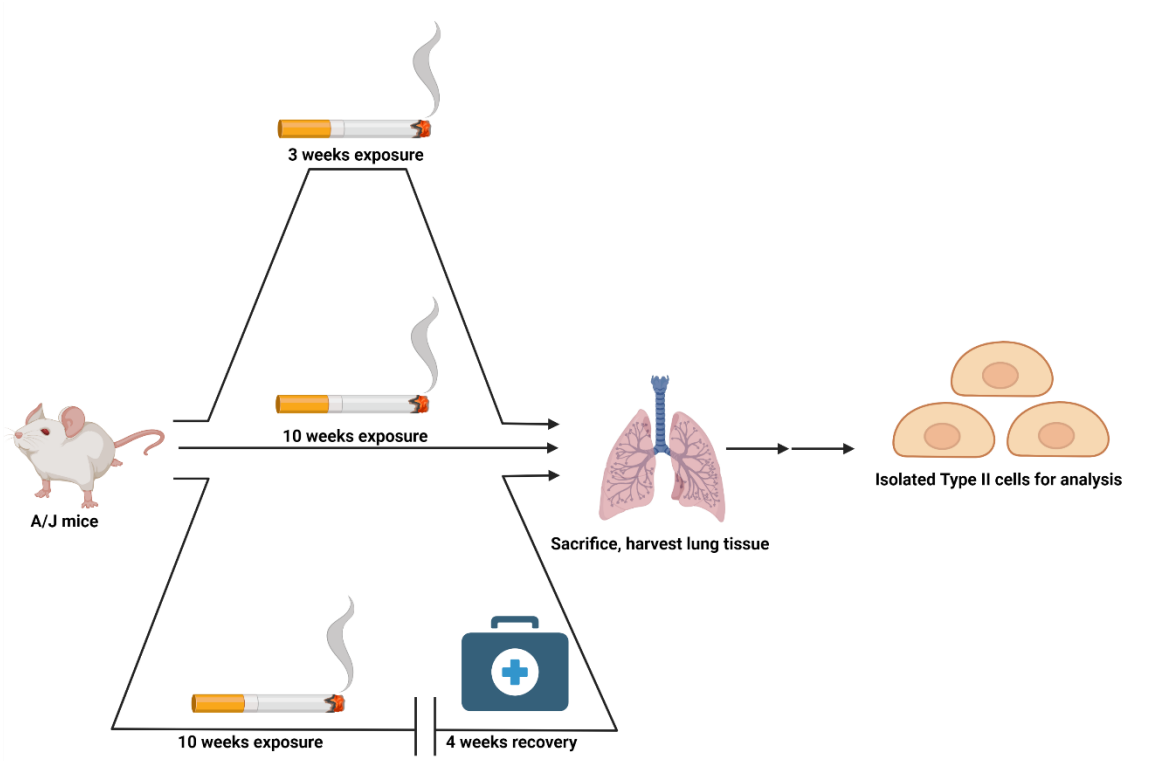
We hypothesized that exposure to cigarette smoke induces inflammation in the lungs and that chronic inflammation contributes to lung cancer etiology. To test this hypothesis, A/J mice were exposed to cigarette smoke or LPS for increased amounts of time (**Figure 4.1**). A subset of animals were allowed to recover for 4 weeks following exposure to examine the reversibility of the epigenetic changes observed in the lungs. Type II pneumocytes were isolated from bulk lung tissue by flow cytometry and subjected to bottom-up proteomics analyses using LC-MS analysis alongside with transcriptomic and epigenomic analyses (**Figure 4.1c**). We then compared our proteomics datasets to transcriptomic and epigenomic data gathered from the same datasets.

Figure 4.1. Experimental designs for multi-omic analyses of Type II pneumocytes from exposed and control mice. a) LPS exposure of A/J mice to induce pulmonary inflammation. b) Cigarette smoke exposure of A/J mice (3 weeks, 10 weeks, 10 weeks with 4-week recovery) c) Scheme for cell isolation and processing for proteomics, transcriptomics, and epigenomics.

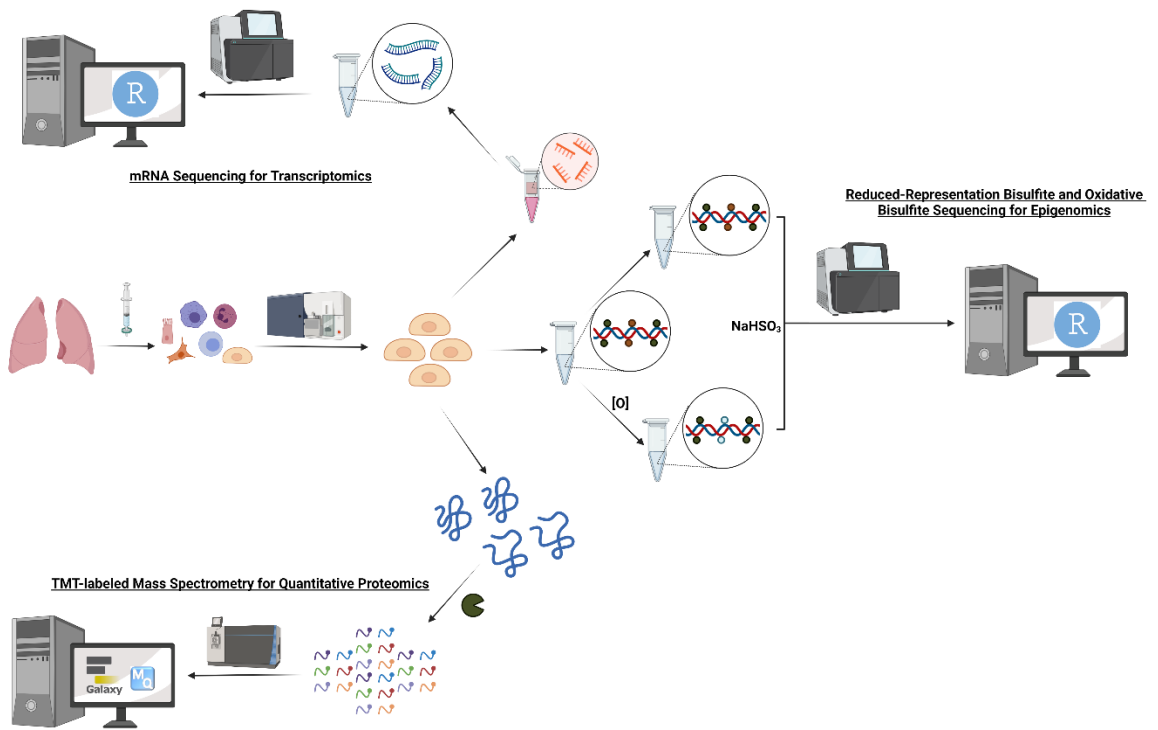
a)



b)



c)



4.2. Materials and methods

Materials

Allprotect tissue reagent was obtained from Qiagen (Hilden, DE). C18 Spin tips (84850), HeLa Protein Digest Standard (88328), Pierce Coomassie Bradford Protein Assay kit (23200), ReproSil-Pur C18-AQ solid phase (Fisher Scientific, NC0834952), TMT11plex label (A34808), and TMT sixplex label (90061) were purchased from Thermo Fisher Scientific (Waltham, MA, USA). Corning dispase (CLS354235), LPS (L2630), SpeedBead Magnetic Carboxylate beads (45152105050250 and 65152105050250), and triethylammonium bicarbonate (TEAB) (18597) were obtained from Millipore Sigma (Burlington, MA). C8 AttractSPE disks were purchased from Affiniseq.

Animal studies

Male and female A/J mice were obtained from the Jackson Laboratory (Bar Harbor, ME) and housed in specific-pathogen-free animal quarters at AeroCore Testing Services, University of Minnesota. All animal experiments were performed according to the U.S. National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals and was approved by the Institutional Animal Care and Use Committee, University of Minnesota.

Exposure protocols

For exposure (Figure 4.1), 6-week-old male and female A/J mice were segregated by sex and further subdivided into two groups of 14 mice per group. Mice in the treatment group were treated intranasally with LPS. LPS was dissolved in 50 μ l of PBS and given as

a divided dose of 25 μ l in each nostril. Mice were treated with 2.5 μ g on week 1, 5 μ g on week 2, and 7 μ g LPS on week 3. Mice in the control group were treated intranasally once a week for three weeks with 50 μ l of PBS, given as a divided dose of 25 μ l in each nostril. Mice were euthanized in a CO₂ chamber one day after their final treatment, after which the lungs were harvested for alveolar epithelial type II cells as described below. Immediately following animal sacrifice prior to harvesting the lungs, approximately 200 μ L of blood was collected from each animal via the iliac vein; following coagulation at room temperature for at least 30 minutes, blood samples were centrifuged at 1,000 x g for 15 minutes, after which the serum was decanted from the from the clot, each of which were then frozen separately at -80 °C. Brain, heart, and liver tissues were harvested following the isolation of the lungs and stored frozen at -80 °C in Allprotect tissue reagent pending further analyses.

With mice exposed to cigarette smoke, mice were again segregated by sex and placed into designated sides of plastic chambers for exposure to either air (control) or cigarette smoke. For those mice in the experimental groups, the smoking chamber was attached to a TE-10B smoking machine (Teague Enterprises, Woodland CA) equipped with pre-conditioned 1R6F cigarettes (University of Kentucky, Lexington KY). The chamber and machine were adjusted so that research subjects would be subjected to 100 mg/m³/hour of total particulate matter (TPM) given as 89% sidestream smoke and 11% mainstream smoke. Animals received 4 h of treatment per day, 5 days a week for variable numbers of weeks (**Figure 4.1b**). Carbon monoxide levels were monitored via an OM-EL-USB-CO USB data logger (Omega Engineering Inc., Norwalk, CT) to mitigate CO poisoning of test subjects.

Sample Isolation

Type II pneumocytes were isolated in accordance with published procedures³¹². In brief, after the mice were euthanized, the lungs were exposed and perfused with 10 mL of cold phosphate buffered saline (PBS). The lungs were then enzymatically digested by infusion of 2 mL dispase (Millipore Sigma) into the lungs through the trachea. The lungs were then removed and incubated in an additional 2 mL of dispase for one hour. Following their incubation, the lungs were manually disrupted with the resulting cell suspension labeled with antibodies specific for CD11c, CD11b, F4/80, CD19, CD45 and CD16/CD32. Samples from nine individual mice were pooled into triplicate sets, giving three sets of lungs per sample in a total of three samples. Type II pneumocytes were isolated by negative selection as the unlabeled cell population³¹². Type II pneumocytes were also gated as sideward scatter high (SSChigh) cell population which minimizes contamination with lymphoid cells by selecting cells with a higher granularity³¹². The cells were separated by fluorescence activated cell sorting (FACS) by the University Flow Cytometry Resource at the University of Minnesota using a BD FACSAria II P07800142 (BSL2) (BD Biosciences, San Jose, CA).

Following isolation via FACS, Type II pneumocytes were pelleted by centrifugation. The samples were centrifuged for 12 min at 200 x g and 4 °C and the supernatant was removed, retaining the final 1 mL which was transferred to a 1.7 mL Eppendorf tube. To this tube, 500 µL of PBS was added, and the tube was further centrifuged for 12 min at 200 x g at 4°C. The supernatant was removed, except for the bottom 100 µL. One mL of PBS was added again to the tube, and a portion of the cells (the

volumetric equivalent of 1.25×10^5 – 5×10^5 cells) were set aside to isolate protein. The remaining sample was centrifuged for 12 min at 800 x g and 4 °C. After this final centrifugation, all supernatant is removed, and the cell pellet was saved to isolate DNA and RNA.

Protein Extraction and Quantitation

To generate protein extract, the designated cells for each sample were transferred to 0.45 µm spin filters (Corning) and washed three times in cold PBS by suspending the cells in 500 µL buffer and centrifuging at 500 g to pellet the cells and pull the PBS through the filter, discarding the flow-through. Following the washes, the cells were lysed via the application of 50 µL lysis buffer (100 mM TEAB pH 8, 7 M urea, 2 M thiourea, 10% acetonitrile, and complete protease inhibitor tablets without EDTA) with vigorous pipetting, followed by centrifugation for 15,000 rpm for 15 min. The flow-through was collected, and the protein concentrations were determined via Qubit Fluorometer (Thermo Scientific, Waltham MA). Samples were stored at -80 °C until analysis.

Conventional TMT labeling of peptides

A methodology following conventional labeling protocols for 5 µg of peptides was used as a comparison to TMT labeling on commercially prepared C18 spin tips. Briefly, HeLa digest standards were reconstituted in 100 mM TEAB buffer, pH 8 to achieve the concentration of 0.1 mg/mL. To six individual Eppendorf tubes was added 5 µg of HeLa digest standard peptides, which were then evaporated to dryness. Samples were then reconstituted in 35 µL of TEAB buffer supplemented with 10 µL of anhydrous acetonitrile.

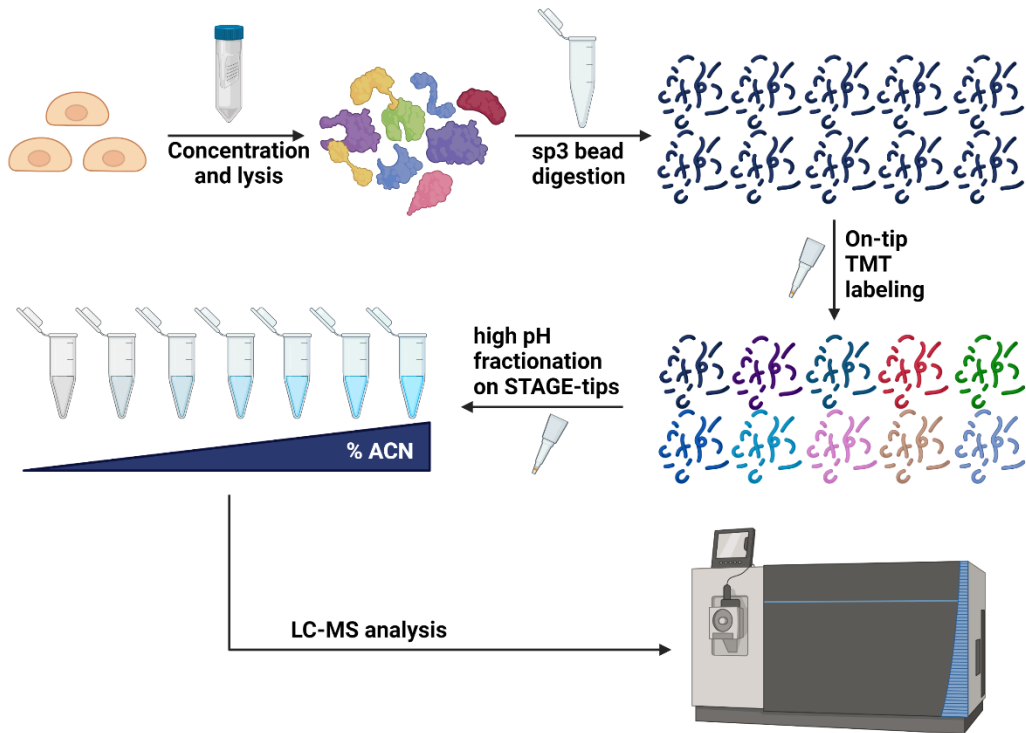
Next, 0.8 mg aliquots of TMTsixplex reagents were brought to room temperature and reconstituted in 41 μ L of anhydrous acetonitrile; each Eppendorf tube was supplemented with 5 μ L of one of the TMTsixplex reagents and were allowed to incubate for 2 hours at room temperature. TMT labeling is stopped after 2 hours with the addition of 5% hydroxylamine, after which the solutions are concatenated and evaporated to dryness. Peptides were then resuspended in 100 μ L 5% acetonitrile and 0.5% TFA and desalted on C18 spin columns prior to LC-MS analysis or high pH fractionation (see below).

Method development for on-tip TMT labeling of peptides

Due to the low levels of protein observed in the cell lysates, modified digestion, TMT-labeling, and fractionation protocols were utilized (**Figure 4.2**). A TMT-labeling strategy based on Myers et al.³¹³ was tested on HeLa extract digest with TMTzero label (Thermo Fisher Scientific, Waltham MA). One μ g of HeLa peptide standard was used for testing. Individual C18 spin columns were added to Eppendorf tubes for each sample to be processed, after which the columns were conditioned with 50 μ L of methanol spun at 1000 g for 1 minute. Following condition, columns were further conditioned with 50 μ L of 80% acetonitrile and 0.1% formic acid and spun at 1000 g for 1 minute, and finally equilibrated with 50 μ L of 0.1% formic acid spun at 1000 g for 1 minute. Dried-down 1 μ g sample aliquots were resuspended in 50 μ L 0.1% formic acid and run through the C18 columns twice at 1000g for 1 minute to immobilize peptides on the C18 columns. Following immobilization on the C18 columns, peptides were washed with two 50 μ L aliquots of 0.1% formic acid run through at 1000g for 1 minute. Prior to labeling, 0.8mg TMTsixplex reagent aliquots were brought to room temperature were reconstituted in 41 μ L of

anhydrous acetonitrile to create TMT stock solutions. To label peptides, 50 μL of working TMT labeling solutions (1 μL TMT stock solution in 49 μL 20 mM TEAB, pH = 8) were added to each C18 spin column containing peptides and centrifuged for 2 minutes at 300 g. Labeling was repeated three more times, with each sample corresponding to a specific TMT channel (Table S3). Following labeling events, excess TMT reagents were washed away with two 50 μL washes of 0.1% formic acid with centrifugation at 1000 g for 1 minute. Samples were eluted from the C18 spin tips into new Eppendorf tubes with an initial addition of 50 μL 80% acetonitrile and 0.1% formic acid with subsequent centrifugation followed by a second elution with 50 μL of 80% acetonitrile in 20 mM ammonium formate pH = 10. Samples are then concatenated together and evaporated to dryness via speed vac. For LC-MS analysis, samples were reconstituted in 10 μL of 0.1% formic acid in water.

Figure 4.2. Combined protocol for quantitative proteomics for samples with low amounts of total protein. Proteins are extracted from cells and immobilized on single-pot, solid-phase-enhanced sample preparation (sp3) beads for digestion, followed by TMT-labeling and high pH fractionation prior to LC-MS analysis.



Method development for STAGE-tip high pH fractionation of peptides

To fractionate low amounts of peptide at high pH on reverse-phase stage tips, we used a protocol adapted from Dimayacyac-Esleta *et al.*⁵⁵ and Kim *et al.*⁵⁶. Stage tips were prepared by packing single cutouts of C8 AttractSPE disk (Affinisep) into 200 μ L pipette tips, after which each was loaded with 100 μ L of 5 μ m ReproSil-Pur C18-AQ (Dr. Maisch GmbH, Ammerbuch DE) slurry (15 μ g/ μ L in 1:1 acetonitrile/100 mM ammonium formate, pH = 10) and further packed via 1500 g centrifugation for 2 minutes. Stage tips were then conditioned sequentially with the addition of 50 μ L methanol, 80% acetonitrile in 100 mM ammonium formate, pH = 10, and 20% acetonitrile in 100 mM ammonium formate, pH 10. Following conditioning, stage tips were transferred to fresh Eppendorf tubes with final equilibration of the columns done with 50 μ L of 100 mM ammonium formate, pH 10 (in each case, tips were centrifuged at 1500 g for 2 minutes). Six μ g of TMTsixplex-labelled HeLa peptide standards were reconstituted in 50 μ L of 100 mM ammonium formate, pH 10 and passed through stage tips twice with centrifugation at 1500 g for 2 min. With peptides immobilized on stage tips, the tips were transferred to fresh Eppendorf tubes and eluted with centrifugation at 1500g for 2-3 min using multiple 50 μ L aliquots of buffer containing 100 mM ammonium formate, pH 10 and increasing amounts of acetonitrile (**Supplemental Table 4.1**). After elution, the 17 fractionations were concatenated into 9 fractions in LC-MS vials as shown in **Supplemental Table 4.2** and evaporated to dryness via speed vac.

Sample processing

Recovered protein extracts were digested using single-pot solid-phase-enhanced sample preparation (SP3) beads⁴⁹. Briefly, sample volumes were brought up to 40 μ L using 100 mM TEAB pH 8 and reduced via the addition of DTT to 5 mM followed by incubation at 56 °C for 30 min. Following reduction, samples were alkylated via the addition of iodoacetamide to 8mM and incubated in the dark at room temperature for 30 min. While the protein samples were being reduced and alkylated, equal amounts of hydrophobic and hydrophilic SpeedBead Magnetic Carboxylate beads were mixed and washed three times in milli-q water. Following reduction and alkylation, protein samples are brought to a final volume of 48 μ L with phosphate-buffered saline, after which 2 μ L of washed bead mixture were added to each sample and the samples mixed via pipetting. Next, ethanol was added to each sample to a final ethanol concentration of 70%, after which the samples were mixed again and allowed to settle on the benchtop for 18 min. Samples were then added to a magnetic rack and the beads allowed to immobilize for 2 min. The supernatant of each sample was removed and discarded, after which the pelleted beads in each sample were washed three times in 80% ethanol. The washed beads were all sonicated for 1 min to break up the beads, after which the bead pellets in each sample were resuspended in 25 μ L of 20 mM TEAB (pH 8.5) supplemented with trypsin at a concentration of 1:25 enzyme to approximate protein abundance. The samples were then incubated overnight at 37 °C to digest the proteins immobilized on the beads. After the overnight digestion, samples were supplemented with an additional 25 μ L of trypsin solution and digested for a further 2 h at 37 °C. Following the second digestion, samples beads were added immobilized on a magnetic rack and the supernatant removed and retained. To extract the remaining peptides, beads were resuspended in 50 μ L of 0.1% formic acid and immobilized on a magnetic rack,

with the supernatants removed and pooled with the first round of supernatants. Peptide samples were then quantified with 280 nm absorbance on the nanodrop, with 1 μ g aliquots set aside and dried down in the speed vac. Dried samples were then labeled with TMT-11plex reagents (Thermo Fisher Scientific, Waltham MA) according to on-tip protocol established above using the scheme documented in **Figure 4.3**.

Figure 4.3. Grouping patterns for TMT-11plex labeling the LPS- and cigarette smoke-exposed samples. Each color represents an LC-MS experiment that was run separately, with the ¹³C channel representing pooled samples for normalization. “Con” = control, “Exp” = experimental (LPS or cigarette smoke).

Sample Group	male						Female					
LPS- 3wk	Con	Con	Con	Exp	Exp	Exp	Con	Con	Con	Exp	Exp	Exp
CS- 3wk	Con	Con	Con	Exp	Exp	Exp	Con	Con	Con	Exp	Exp	Exp
CS- 10wk	Con	Con	Con	Exp	Exp	Exp	Con	Con	Con	Exp	Exp	Exp
CS- 10wk + recovery	Con	Con	Con	Exp	Exp	Exp	Con	Con	Con	Exp	Exp	Exp

LC-MS conditions

Fractionated peptide samples were analyzed on an Orbitrap Fusion Tribrid Mass Spectrometer interfaced with an Ultimate 3000 UHPLC. The UHPLC was run in the nanoflow mode with a reverse-phase nanoLC column (15 cm x 250 μ m) packed with 5 μ m diameter Luna C18 resin. Samples were reconstituted in 10 μ L of buffer A (0.1% formic acid in water) prior to analysis. Samples were run on a 90-min gradient with a 5-22% increase in buffer B (0.1% formic acid in acetonitrile) over the first 71 min, followed by a 22-33% increase in buffer B over the next 5 min and rapid increase of 33-90% increase in buffer B over 5 min. The column was washed for 4 min at 90% buffer B, and the solvent composition was returned to 5% B over 2 min, followed by a 3 min equilibration at 5% buffer B. Samples were run at a flow rate of 300 nL/min. Peptides were analyzed in the positive ion mode using Top12 Full MS/dd-MS/MS mode with an expected chromatographic peak FWHM of 15 seconds. For the full MS scans, resolution was 70,000 with an AGC target of 1e6, a maximum IT of 30 ms, and a scan range of 300 to 2000 m/z. Tandem mass spectrometry (MS/MS) experiments were conducted at 30,000 resolution, AGC target of 5e4, maximum IT of 50 ms, an isolation window of 2.0 m/z and a normalized collision energy of 30. Data was collected in centroid mode

Data analysis

Raw mass spectrometry data were analyzed together in MaxQuant⁵⁹ using the Reporter Ion MS/MS quantification mode against the Uniprot Mus musculus proteome supplemented with the contaminants database. Carbamidomethylation of cysteine was included as a fixed modification, while oxidation of methionine, N-terminal acetylation,

and phosphorylation of serine, threonine, and tyrosine were included as variable modifications. Data were processed using the open source data manipulation platform Perseus¹³⁵ to generate volcano plots. Gene ontology (GO) analyses were conducted using gProfiler¹¹³. Interaction networks of proteins were generated using the STRING¹¹⁷ database. Proteomics data was compared with transcriptomic data generated from these same cells generated by using QuanTP¹²⁴ within the Galaxy MSI instance.

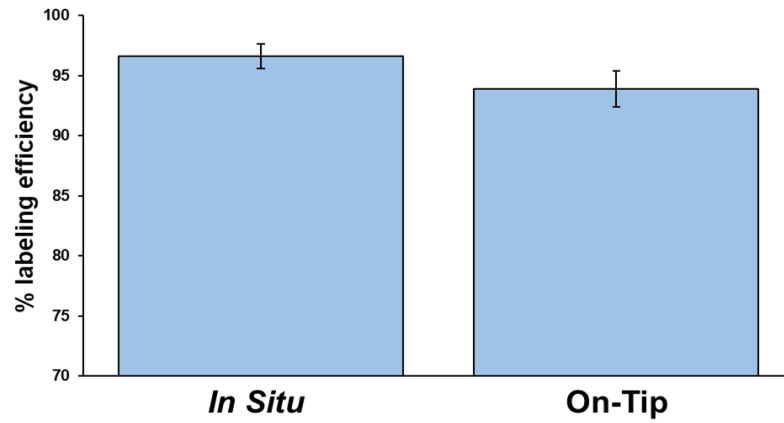
4.3 Results

4.3.1 Method validation of on-tip TMT labeling and STAGE-tip-based high pH fractionation

Following exposure of A/J mice to cigarette smoke and isolation of Type II pneumocytes from the lungs, total protein was extracted from cells for quantitative proteomics experiments. We found that extended treatment with inflammatory stimuli resulted in reduced numbers of Type II cells relative to the controls; this in turn resulted in low amounts of protein extracted in many of the samples (2.14-98.73 μg , **Supplemental Table 4.3**). Many of these amounts were insufficient for standard TMT labeling protocol, which require a minimum of 5 μg for TMT-labeling and 10 μg for high pH reverse-phase fractionation. C18-tip based TMT-labeling and STAGE-tip-based high pH fractionation protocols were initially tested on HeLa digest standards to ensure the efficacy of these procedures for this experiment and for future experiments in our laboratory. Based on the levels of proteins in each sample as reflected in **Supplemental Table 4.3**, we determined that 1 μg of peptides from each sample would be reliably available for TMT labeling; we therefore decided to use this amount of peptides to test the labeling efficiency of on-tip

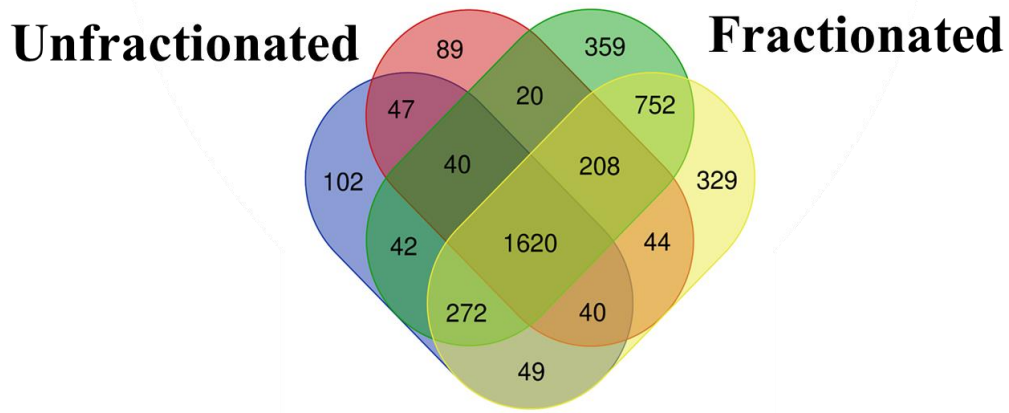
C18 labeling. In examining the degree of TMTsixplex labeling across all six channels, we found that the average labeling efficiency was approximately 94% across each of the TMT labels; while this was slightly below the desired level of 95% labeling efficiency, a t-test comparison against our in situ labeling strategy (97%) showed these numbers to not be significantly different to one another (Figure 4.4), indicating that the on-tip labeling strategy is an acceptable method for labeling low amounts of tryptic peptides.

Figure 4.4. Labeling efficiencies of *in situ* and on-tip strategies for TMT labeling. Average labeling efficiencies for TMT labels were determined for both labeling strategies using triplicate samples of HeLa digest. Statistical comparison of the labeling efficiency values yields a p-value greater than 0.05, indicating a lack of statistical difference between these methodologies



In addition to developing modified strategies for TMT-labeling, we also set out to test and establish a modified high pH fractionation protocol to improve the depth of protein coverage in our bottom-up proteomics. Using *in situ*-labeled commercially-available HeLa peptides from the on-tip TMT validation experiments, we fractionated two replicates of 6 μg of peptides immobilized on C18 STAGE tips into 17 fractions using alkaline buffer with increasing concentrations of acetonitrile (**Supplemental Table 4.1**) These were then concatenated into 9 fractions (**Supplemental Table 4.2**), which were individually analyzed via LC-MS and compared with two 1 μg injections of unfractionated 6 μg TMT-labeled commercial HeLa peptide samples. Using a short HPLC gradient (60 min), LC-MS analysis of two unfractionated samples yielded 2212 and 2018 protein IDs, respectively, with 1747 proteins shared between both runs. By contrast, global proteomics analysis of fractionated samples resulted in identification of 3313 and 3314 proteins, respectively, with 2852 proteins shared between these sample sets. This represents a 60% increase in proteins identified, with relatively few proteins unique to the unfractionated runs (**Figure 4.5**). Therefore, the fractionation protocol represents a demonstrable increase in protein IDs and was therefore chosen for use in the analysis of the animal samples.

Figure 4.5. Numbers of proteins identified in commercial HeLa peptides with and without high pH fractionation of TMT-labeled peptides. Two 6 µg aliquots of TMT-labeled HeLa digest were fractionated into seventeen fractions using increased acetonitrile, which were then concatenated into nine fractions and analyzed via LC-MS on a Fusion Tribrid Orbitrap Mass Spectrometer. Protein identifications from two injections of 1µg of labeled peptide were compared with the fractionated data.

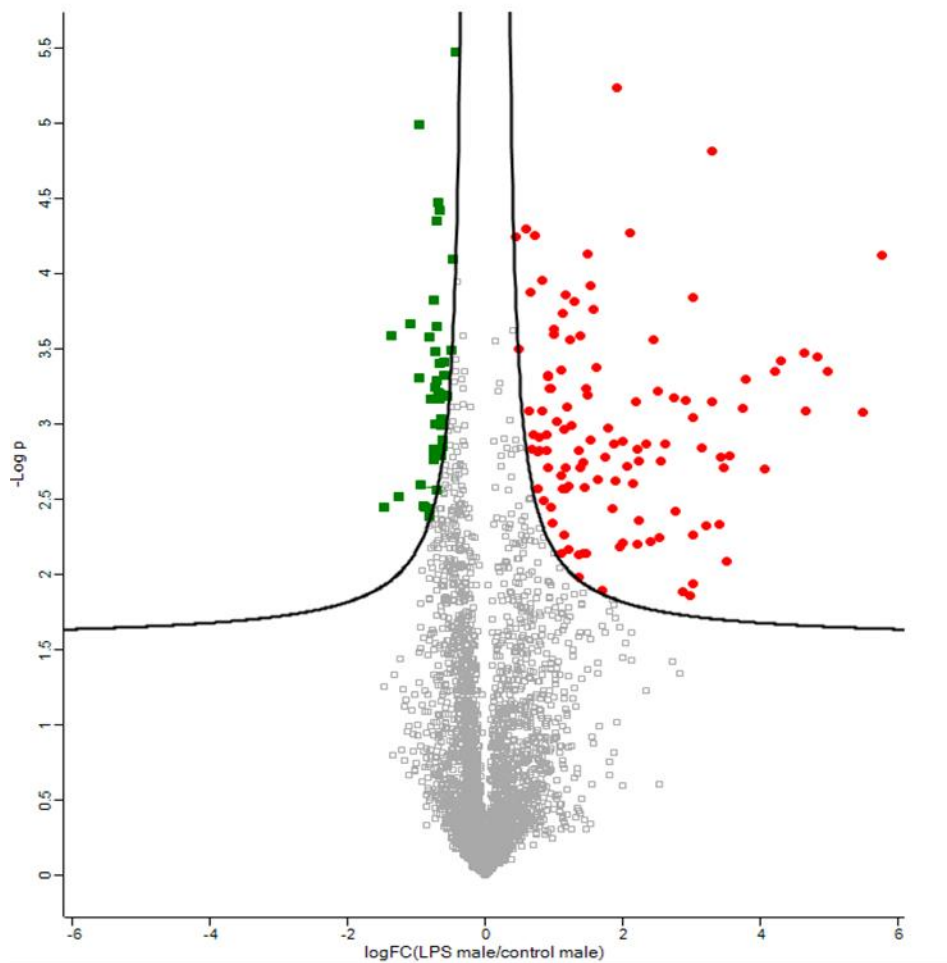


4.3.2 Bottom-up proteomics analyses of Type II cells of A/J mice intranasally treated with LPS demonstrates increased proteome changes consistent with inflammation

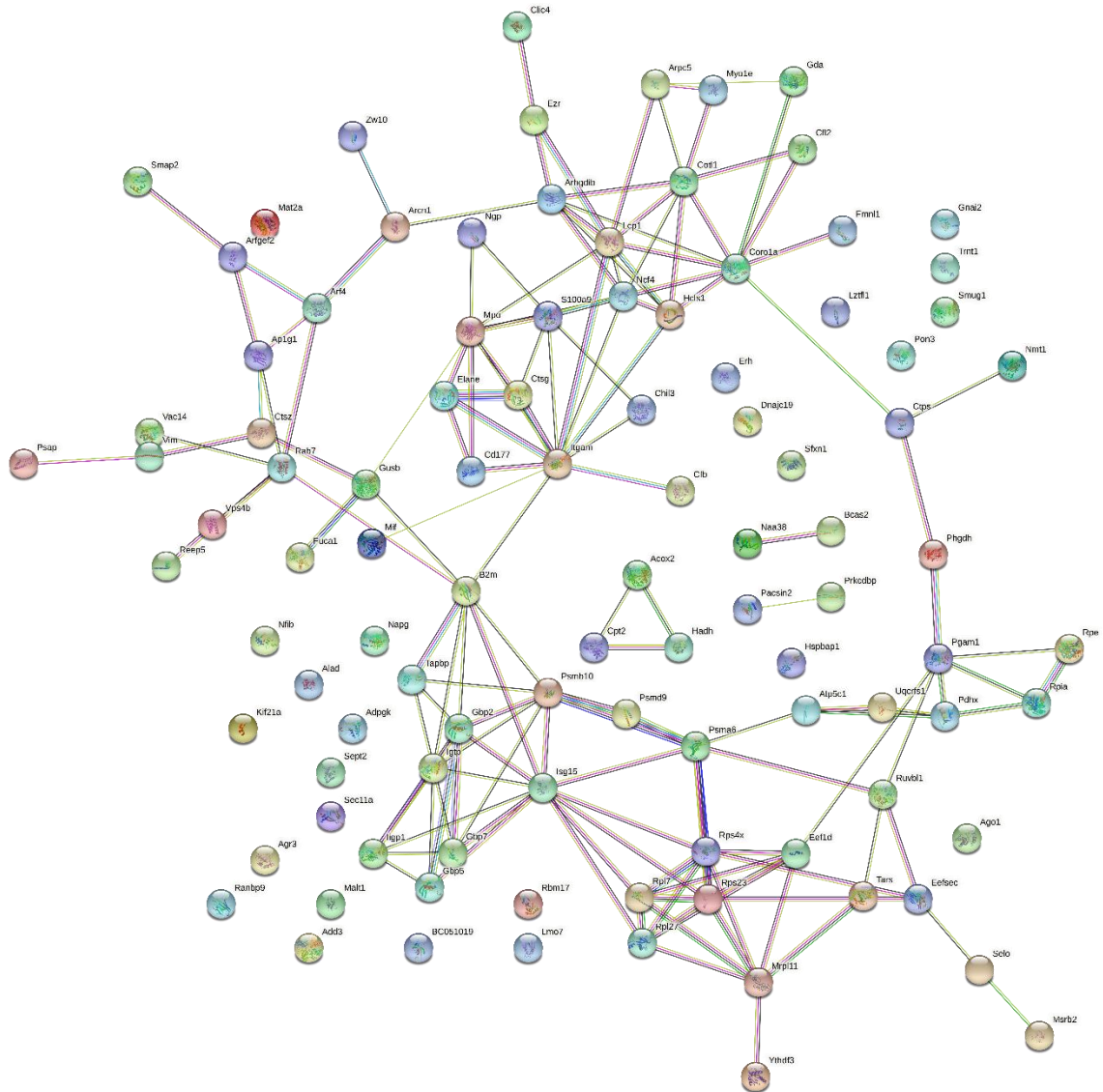
Quantitative global proteomics analysis of Type II cells harvested from LPS exposed A/J mice reliably detected 3352 protein groups across all samples. Quantitative analyses revealed that upon LPS exposure 114 proteins were increased in abundance and 43 proteins were decreased in abundance (see volcano plot in **Figure 4.6a**). Functional analysis of upregulated proteins using the STRING database suggests that these proteins are known to interact with one another, forming distinct clusters of co-expression and physical interaction (**Figure 4.6b**). Gene ontology analysis of the proteins increased in abundance upon LPS treatment (**Figure 4.6c**) shows enrichment for GO Reactome terms consistent with an inflammatory phenotype, such as “Antigen processing-Cross presentation”, “Neutrophil degranulation”, and “Innate Immune System”. These pathways are consistent with general inflammatory processes upon treatment with LPS.

Figure 4.6. Global proteomics analysis of LPS-treated Type II pneumocytes reveals characteristic changes to the protein content. a) Volcano plot of differentially abundant proteins in Type II pneumocytes following three weeks of LPS exposure. b) Protein-protein interactions of proteins increased in abundance. c) GO terms associated with proteins that were increased in abundance following LPS exposure.

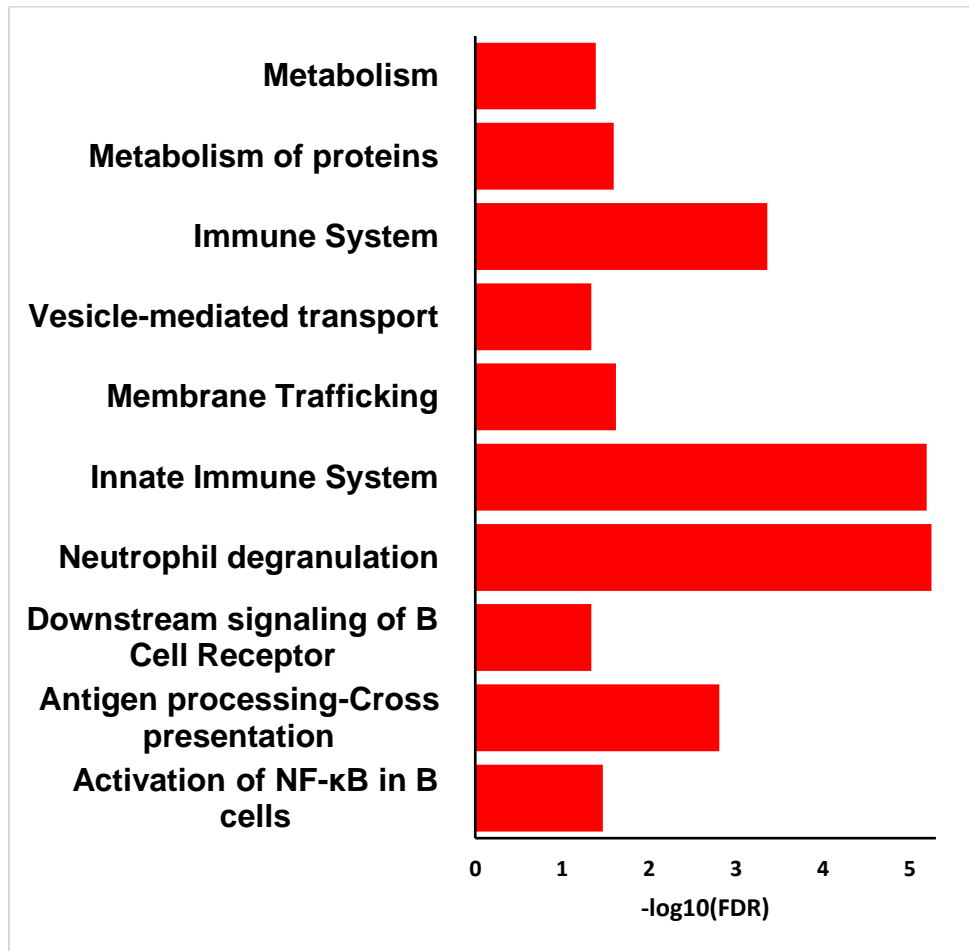
a)



b)



c)



4.3.3 Multi-omic analysis to examine the correlation between the transcriptome and the proteome responses to LPS

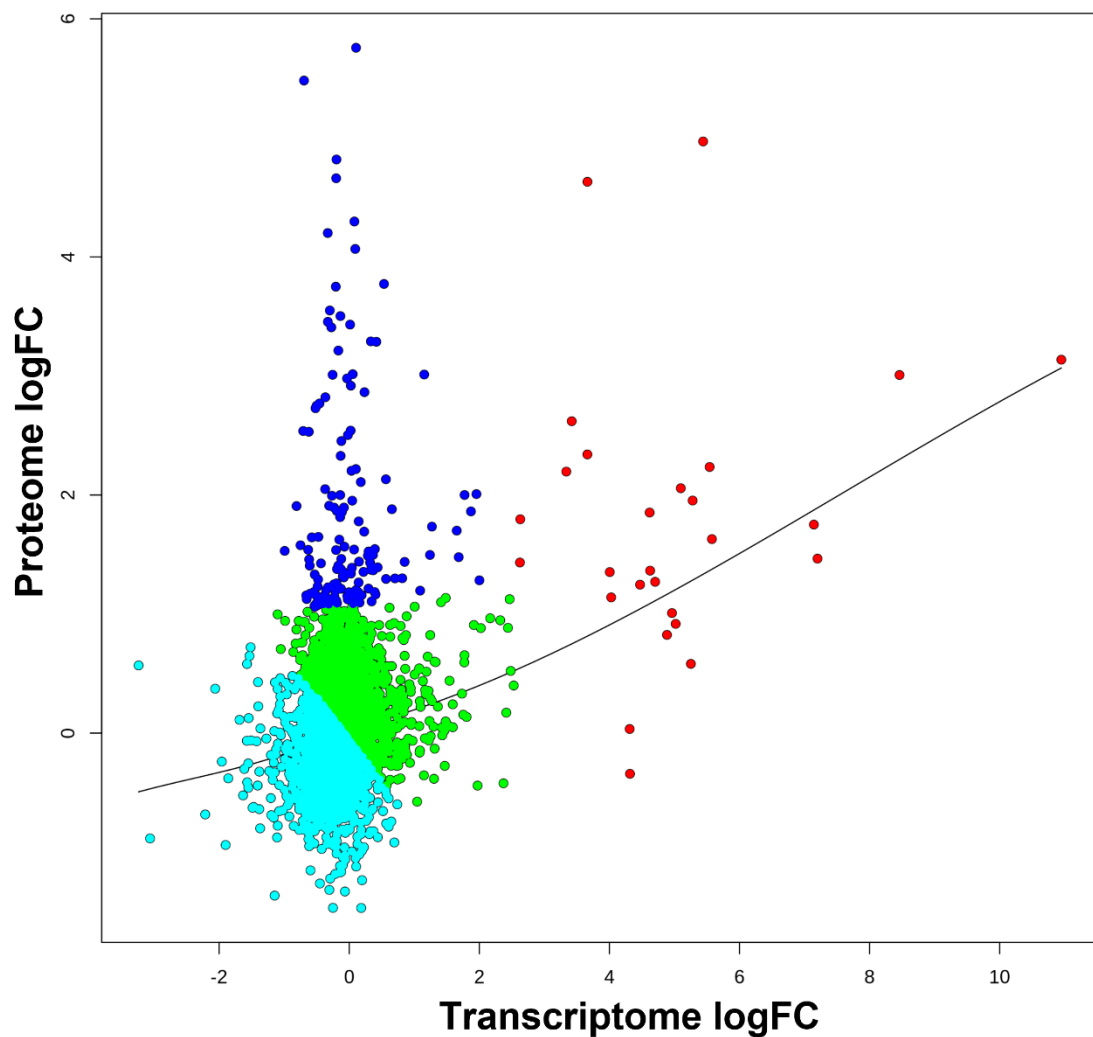
As initial separate GO analyses indicated a degree of homology between the transcriptomic and proteomic results, these were directly compared to one another using the QuanTP tool(12) within the Galaxy platform to examine the similarities and differences between RNA and protein changes in response to LPS. To compare the proteome and the transcriptome, the 3222 genes common to both data sets were identified and used to filter both datasets for QuanTP analysis; this resulted in considerable loss of information from the transcriptomic data (from 16,367 initial entries to 3222 genes for analysis), due in part to the limited depth of coverage of proteomics in comparison with transcriptomics. From QuanTP, we see that comparing the log₂FC of the transcriptome and proteome yields a plot that is nearly linear, with a Pearson correlation of 0.282 indicating a slight positive correlation between the levels of information (**Figure 4a**). In performing k-clustering, QuanTP generated four clusters corresponding to four arithmetic means. Three of these clusters- the green, cyan, and red clusters annotated in **Figure 4.7a**- fall roughly on a linear relationship between the transcriptome and the proteome. The fourth cluster, highlighted in blue, consists of 148 genes which have large increases in protein abundance in response to LPS treatment with little to no concomitant increase in mRNA expression, with the genes in this cluster contributing the most to the deviation of these data from an $R^2 = 1$ linear relationship. Gene ontology analysis of these genes shows a preponderance of terms associated with exocytosis and exosome secretion (**Figure 4.7b**), which is consistent with the release of cytokines seen in the inflammatory response. The lack of increased mRNA transcription of these genes has many potential causes, including discrepancies between

the half-lives of these transcripts and proteins, altered transcription/translation efficiencies, or potential posttranscriptional regulation; further experiments are required to investigate these hypotheses.

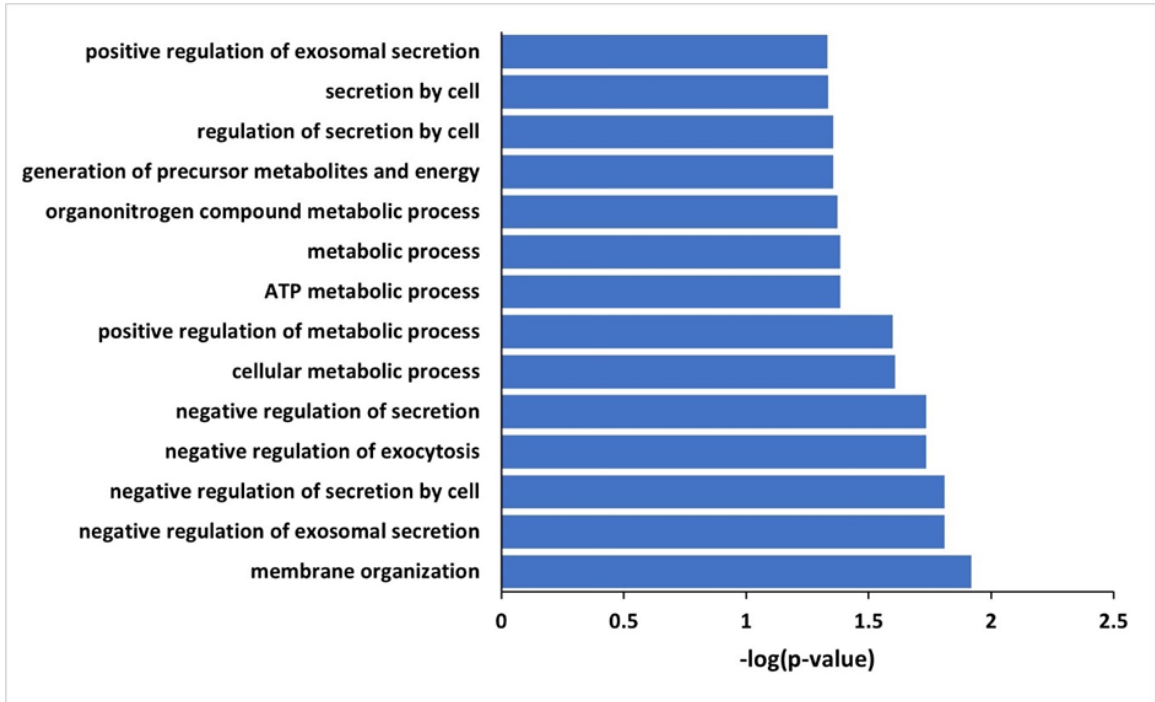
When those genes that show significant increases in protein abundances are compared with those genes that show significant increases in mRNA transcription, about 35% of the proteins seen to be significantly increased also show significant increases in mRNA transcription. Gene ontology analysis of these genes continues to correspond to inflammatory pathways, demonstrating the paramount nature of the inflammatory response in Type II cells of mice exposed to LPS (**Figure 4.7c**).

Figure 4.7. Multi-omic comparisons of proteomic and transcriptomic data of LPS-exposed Type II cells. a) Correlation plot between transcriptome and proteome generated in QuanTP b) GO Analysis of genes that show an increase in protein abundance without concurrent change in transcription (blue cluster in Figure 4.7a). c) GO Analysis of genes significantly increased at the transcriptomic and proteomic level.

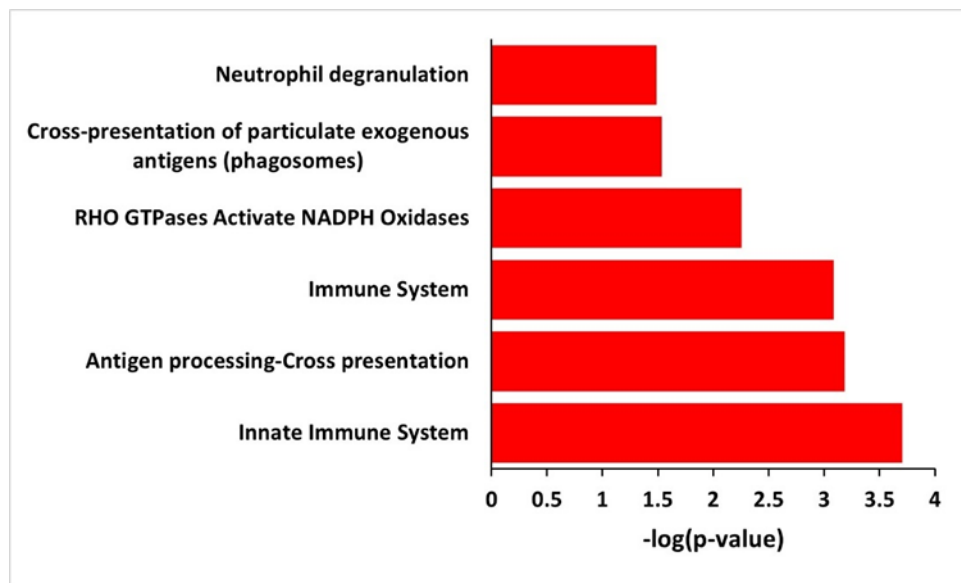
a)



b)



c)



4.3.4 Multi-omic comparison of proteomics, transcriptomics, and epigenomics after 3 weeks of LPS exposure

In addition to examining the dynamics between the proteomics and transcriptomics responses in Type II pneumocytes of LPS-exposed mice, we also sought to integrate these data with epigenomics data derived from the same cells. These epigenomics data, obtained by Dr. Qiyuan Han, consisted of differential methylation and differential hydroxymethylation data gathered through standard and oxidative reduced-representation bisulfite sequencing, respectively. In integrating the data, we noted that five genes had differential expression and protein abundance in response to significantly altered methylation across promoter regions, an intron, and exons (**Table 4.1**). By contrast, there were no genes observed to have differential expression and protein abundance that also showed differential hydroxymethylation in any regions of their genes.

Table 4.1. Multi-omic comparison of genes with significantly altered methylation, gene expression, and protein abundance.

ENSEMBL	Gene	Description	Region	$\Delta 5mC$	logFC (mRNA)	LogFC (Protein)
ENSMUSG00000001020	S100a4	S100 calcium binding protein A4	Promoter (<=1kb)	0.15825 3	3.04738926	2.19683937 2
ENSMUSG000000020183	Cpm	carboxypeptidase M	Intron	0.23869 8	- 1.35584488 2	- 1.36497210 5
			Promoter (2-3kb)	0.29342 5		
ENSMUSG000000021573	Tppp	tubulin polymerization promoting protein	Exon	0.28405 9	- 1.51882760 2	- 0.62715498 1
ENSMUSG000000026728	Vim	vimentin	Exon	0.18916 9	1.18467710 5	1.13518137 7
ENSMUSG000000055805	Fmn1l	formin-like 1	Exon	0.12821 4	3.73130922 8	1.14025973 3
			Exon	0.14311 4		

4.3.5 Bottom-up proteomics of Type II pneumocytes of mice exposed to cigarette smoke

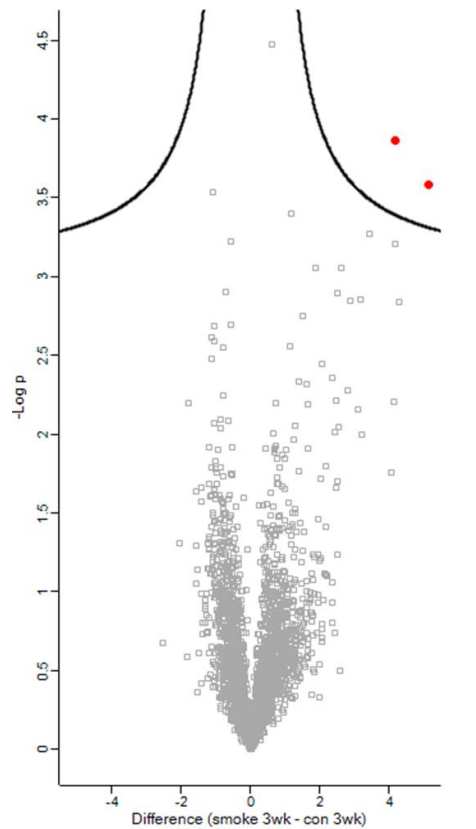
Changes of protein abundances in Type II pneumocytes in response to cigarette smoke exposure were also investigated using quantitative global proteomics. As these samples were TMT-labeled and analyzed with the LPS-exposed samples, the same 3352 proteins were detected across the cigarette smoke-exposed samples.

In Type II cells of mice that were exposed to three weeks of cigarette smoke, only two proteins- Rac family small GTPase 1 (Rac1) and Cytochrome b-c1 complex subunit Rieske, mitochondrial (Uqcrcf1)- were observed to have a significant degree of increased abundance (**Figure 4.8a**). By contrast, after 10 weeks of cigarette smoke exposure quantitative proteomics identified 22 proteins significantly increased in abundance upon exposure to ESC (**Figure 4.8b**). Gene ontology analysis of these upregulated genes revealed enriched Reactome terms for cell cycle regulation (e.g., “G2/M Checkpoints”, “The role of GTSE1 in G2/M progression after G2 checkpoint”, “Cyclin A:Cdk2-associated events at S phase entry”), hypoxia (“Cellular response to hypoxia”, “Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha”), DNA damage repair (“p53-Independent G1/S DNA damage checkpoint”, “p53-Independent DNA Damage Response”, “p53-Dependent G1/S DNA damage checkpoint”), etc. (**Figure 4.8c**). When mice were allowed to recover for 4 weeks after 10 weeks of cigarette smoke exposure, 74 proteins were significantly increased in abundance, and six proteins were significantly decreased in abundance (**Figure 4.8d**). These upregulated proteins, when submitted for GO analysis, were enriched for Biological Process terms for several metabolic processes (“Pyruvate metabolic process”, “Acetyl-CoA metabolic process”,

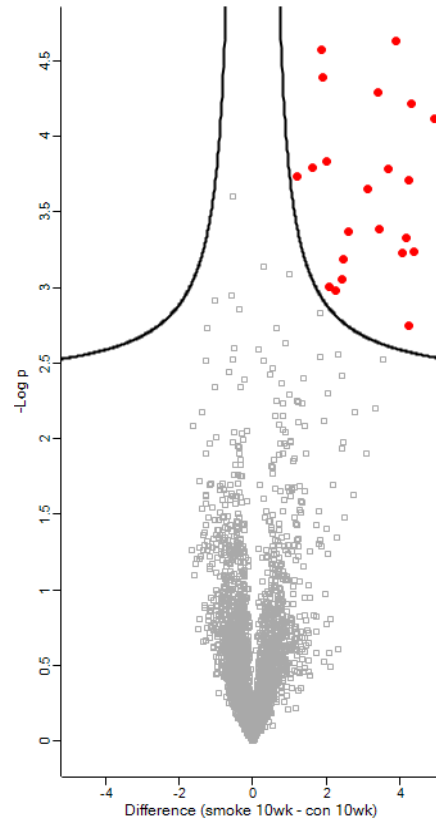
“Purine nucleotide metabolic process”), cellular polarity (“Establishment or maintenance of bipolar cell polarity”, “Establishment or maintenance of apical/basal cell polarity”) as well as the KEGG pathway for cancer carbon metabolism (**Figure 4.8e**).

Figure 4.8. Global proteomics analysis of Type II pneumocytes isolated from cigarette smoke-exposed A/J mice reveals exposure-dependent phenotypes. a) Volcano plot of differential proteome abundances after 3 weeks of cigarette smoke exposure b) Volcano plot of differential proteome abundances after 10 weeks of cigarette smoke exposure c) Selected Reactome GO terms enriched for proteins significantly increased after 10 weeks of cigarette smoke exposure d) Volcano plot of differential proteome abundances after 10 weeks of cigarette smoke exposure and 4 subsequent weeks of recovery in clean air e) GO terms enriched for proteins significantly increased after 10 weeks of cigarette smoke exposure with subsequent recovery.

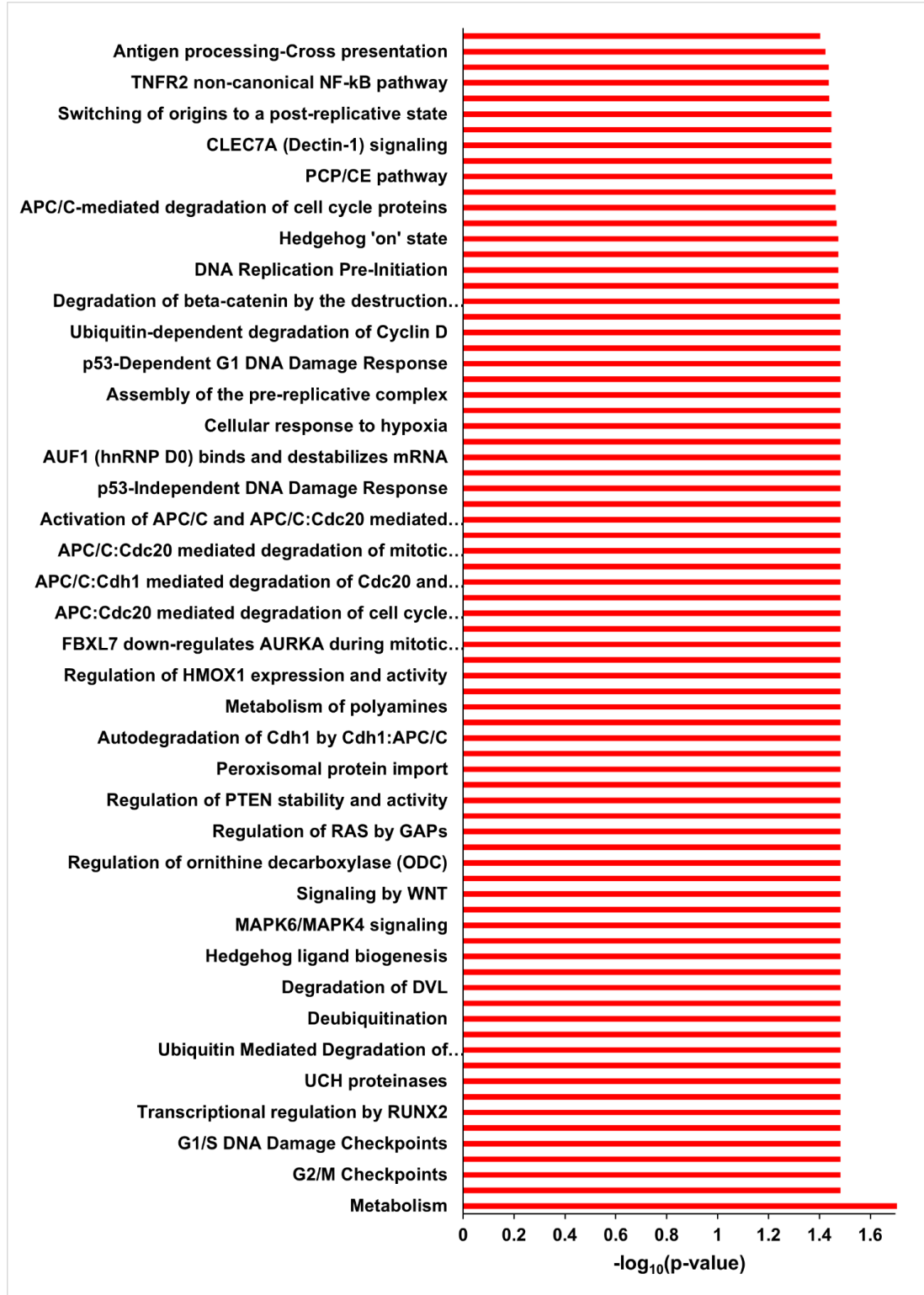
a)



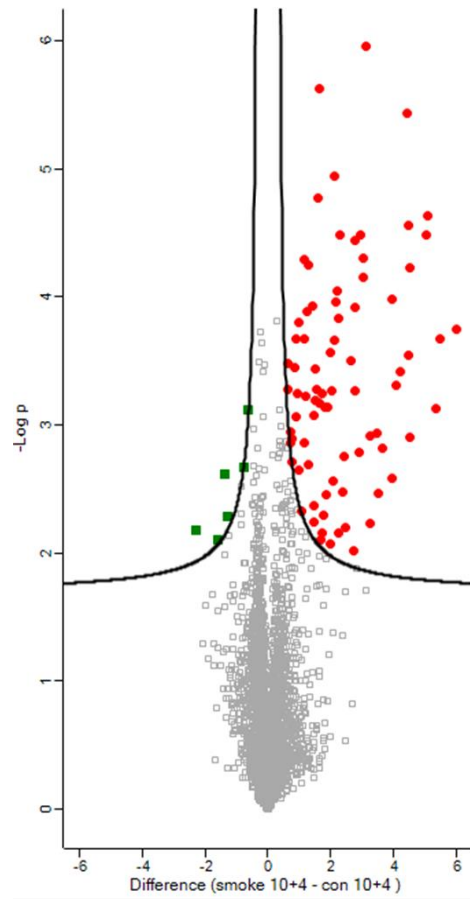
b)



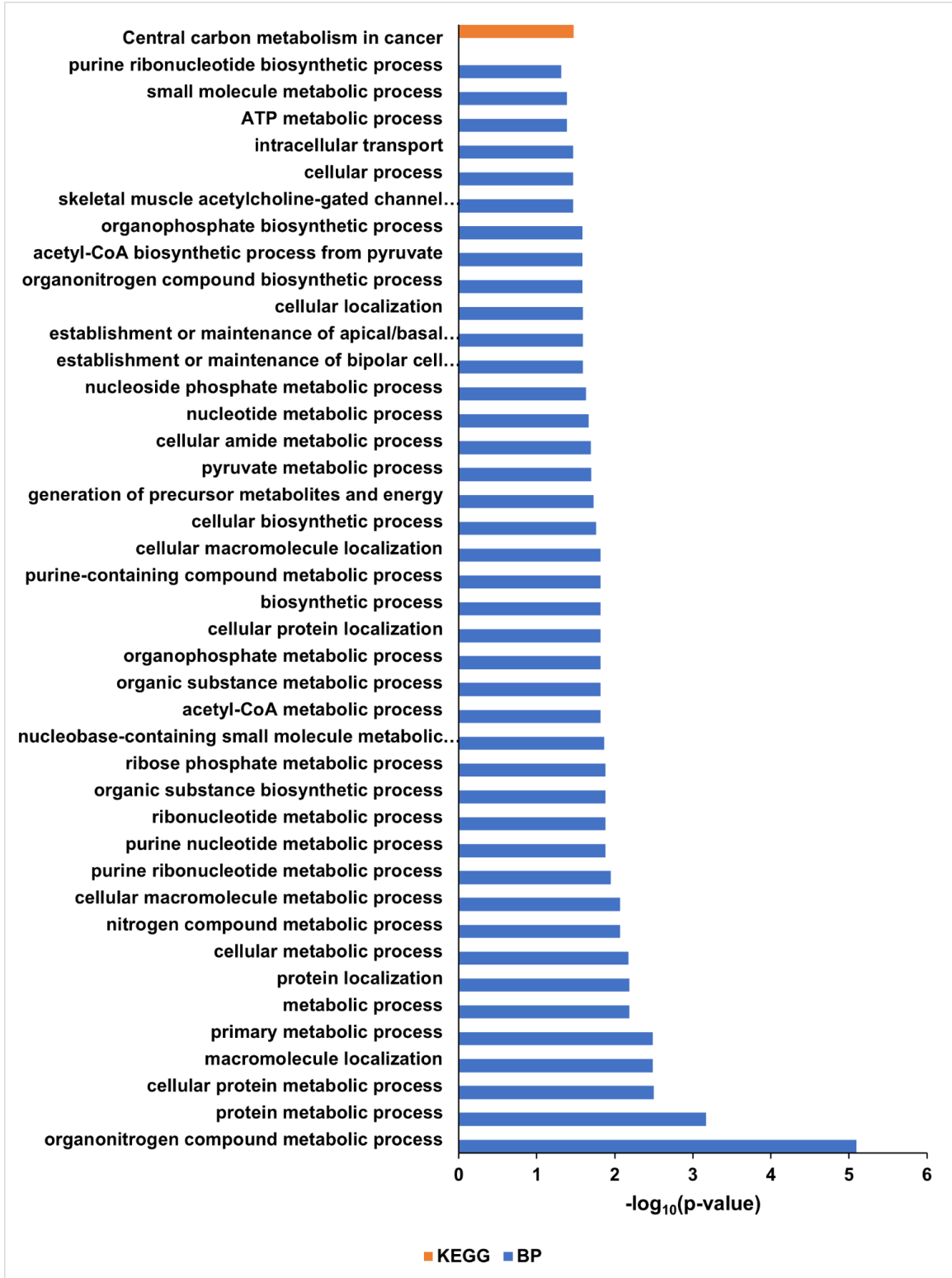
c)



d)



e)



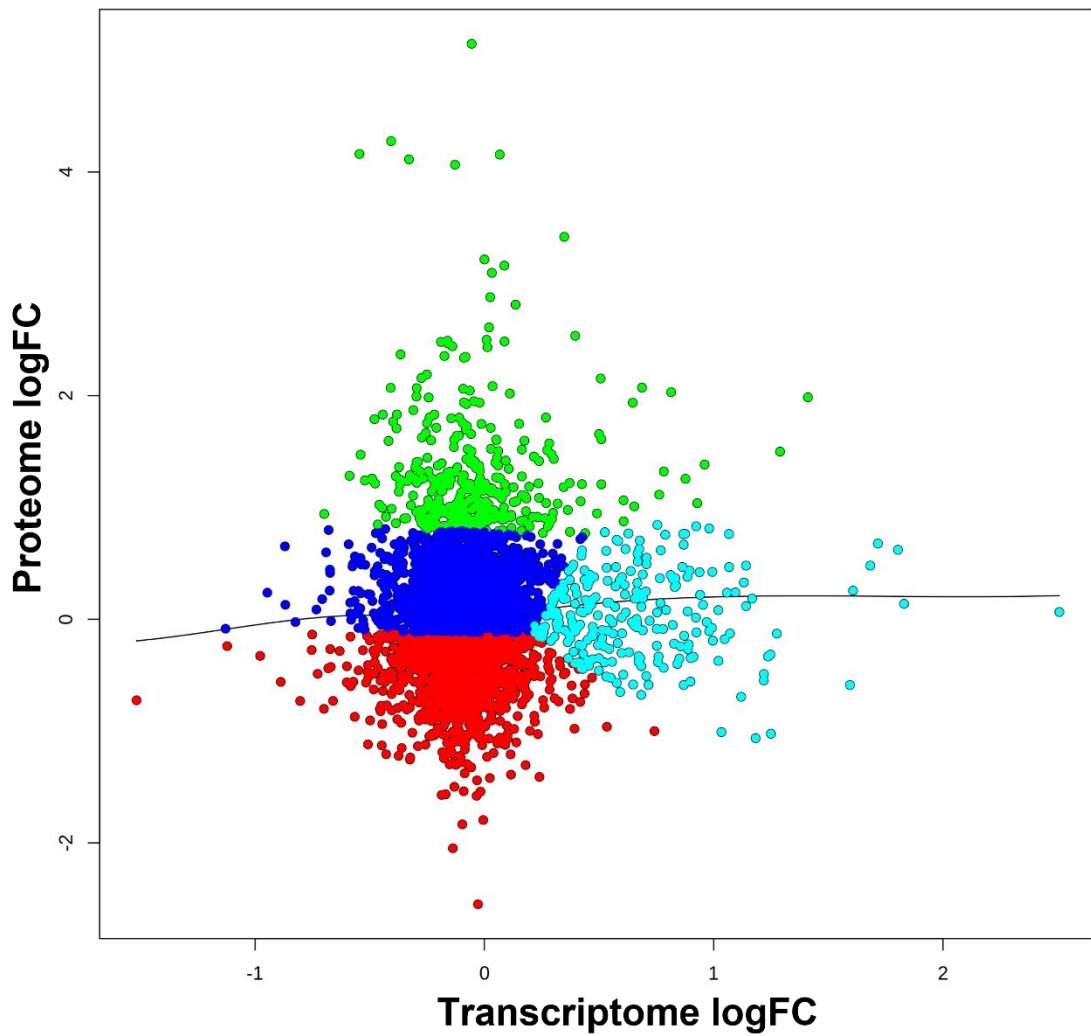
4.3.6 Multi-omic comparison of cigarette smoke-driven proteomic and transcriptomic changes in Type II cells of A/J mice treated with cigarette smoke

To examine the ways in which transcription and translation are altered by variable amounts of exposure to cigarette smoke, QuanTP was again used to compare the transcriptomic and proteomic data derived from cigarette smoke-exposed Type II pneumocytes.

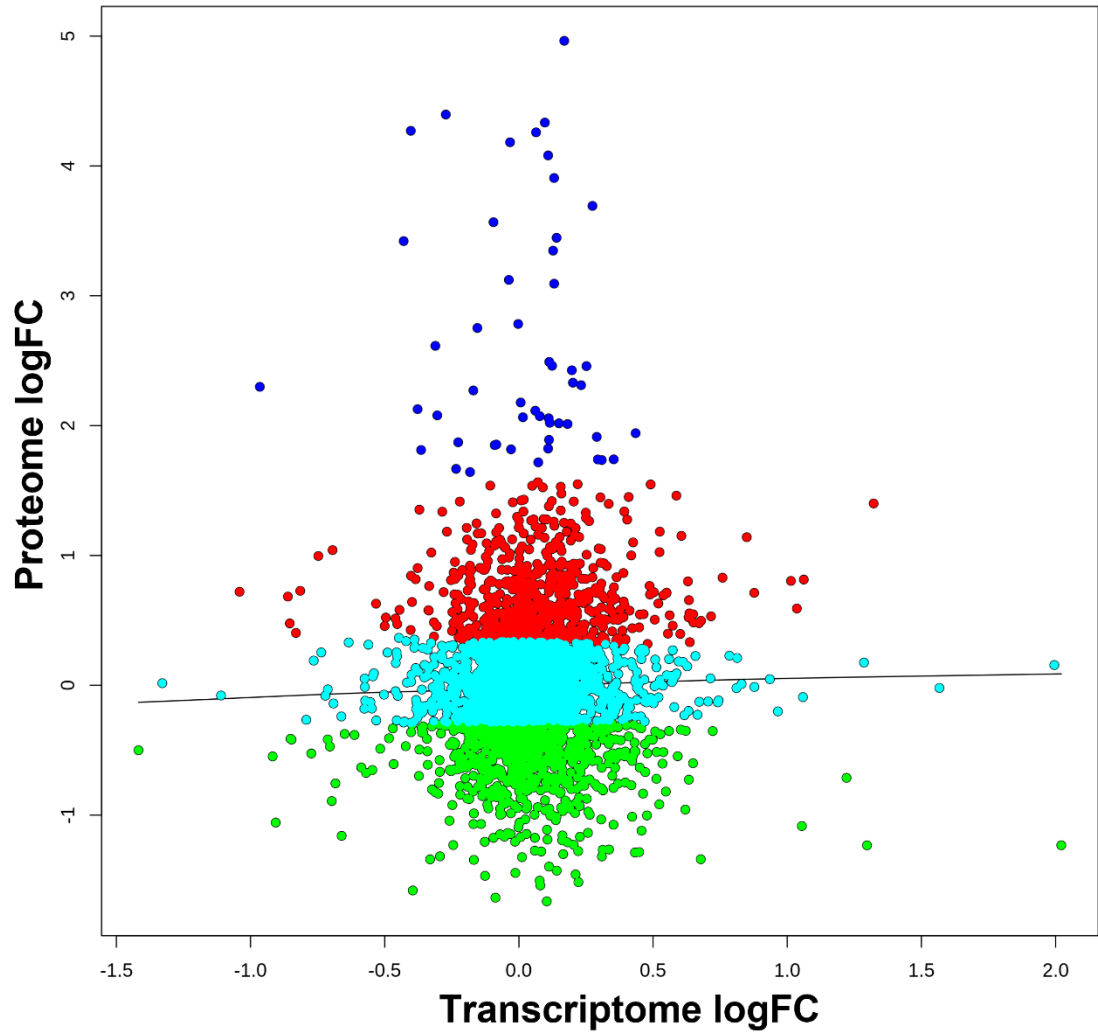
In contrast with the multi-omic comparison of transcriptomic data and proteomic data seen in LPS-exposed Type II cells, exposure to cigarette smoke seemed to result in a greater disjunction of the transcriptome and the proteome. Comparison of the 3-week exposure data in QuanTP (**Figure 9a**) shows that most data points have either considerable change in the transcriptome with little change in the proteome (cyan cluster) or vice versa (green cluster), with the rest of the data points showing only moderate levels of change (blue and red clusters). This pattern holds in the 10-week and 10-week with 4-week recovery samples (**Figures 9b and 9c**), though as the animals are allowed to grow older, they have more genes with larger increases in protein abundance without a corresponding change in mRNA abundance, which results in the shifting of the clusters into stacked groups along the y-axis.

Figure 4.9. Multi-omic comparisons of proteomic and transcriptomic data for cigarette smoke-exposed Type II pneumocytes using QuanTP a) 3 weeks of cigarette smoke b) 10 weeks of cigarette smoke c) 10 weeks of cigarette smoke with 4 weeks of post-exposure recovery

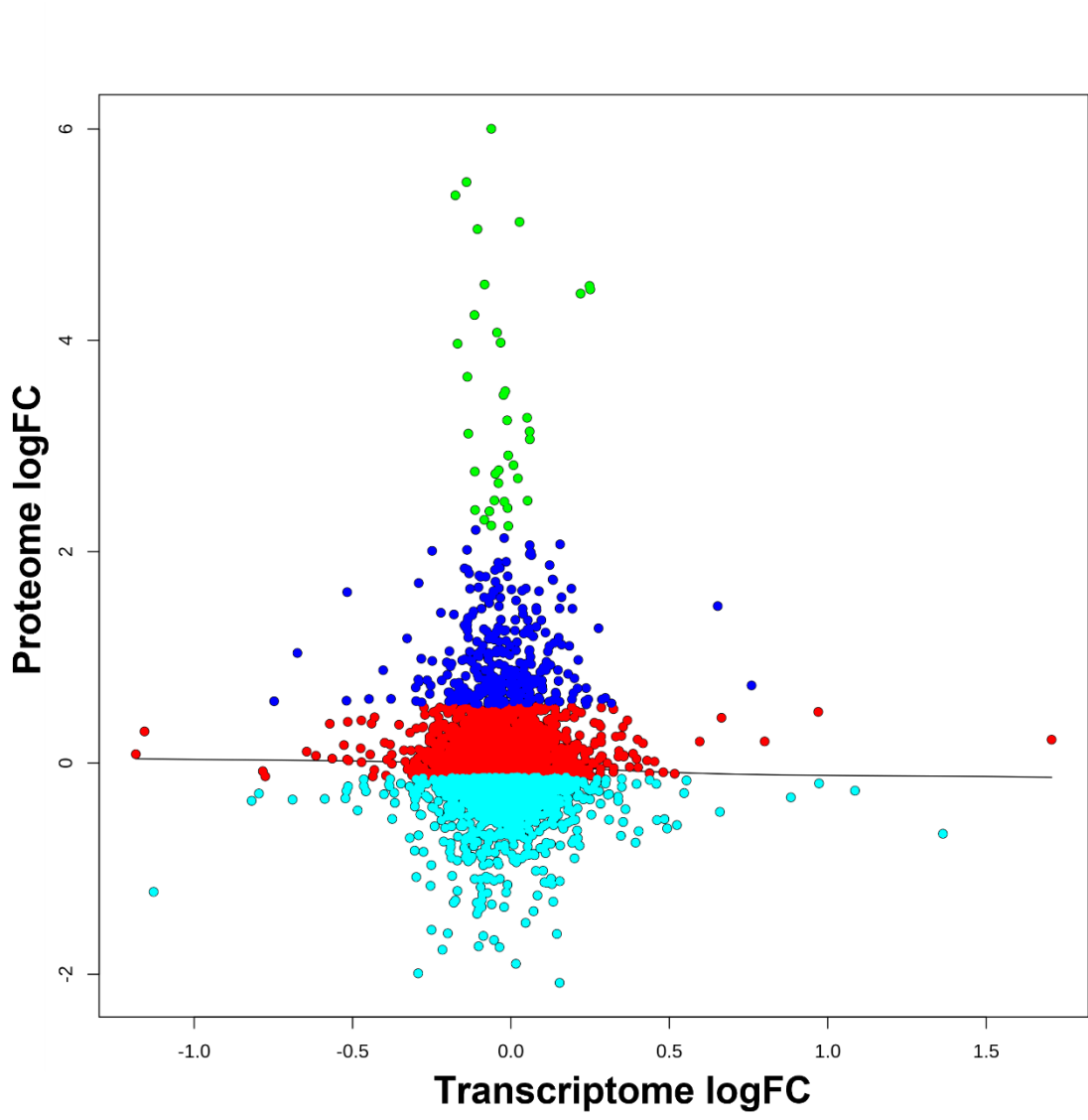
a)



b)



c)



When we narrowed our focus to those genes that showed significant changes in the transcriptomic and proteomic data following cigarette smoke exposure, most genes were discarded due to a lack of significant changes in one or the other of these levels of information. The datasets for Type II cells that experienced 3 weeks of cigarette smoke exposure and the Type II cells that experienced 10 weeks of cigarette smoke with 4 weeks of recovery had no genes with significant changes at the protein and mRNA levels.

Multi-omics analysis of Type II cells from A/J mice that were exposed cigarette smoke for 10 weeks identified four genes that showed significant changes in both mRNA and protein levels (**Table 4.2**). Interestingly, three of these genes- *Entpd1*, *Lgals3*, and *Hmgcl*- showed a significant decrease at the mRNA level with concurrent significant increases at the protein level. Only *Slc9a3r1*, Solute Carrier Family 9, Subfamily A (NHE3, Cation Proton Antiporter 3), Member 3 Regulator 1, showed significant increases in both datasets. When considering differential epigenomics data (methylation and hydroxymethylation) for these genes, only *Hmgcl* registers a change in the degree of hydroxymethylation, though it is to a low degree (-0.0238448) that is not generally considered significant.

Table 4.2. Genes showing significant changes at the transcriptome and proteome levels after 10 weeks of cigarette smoke exposure.

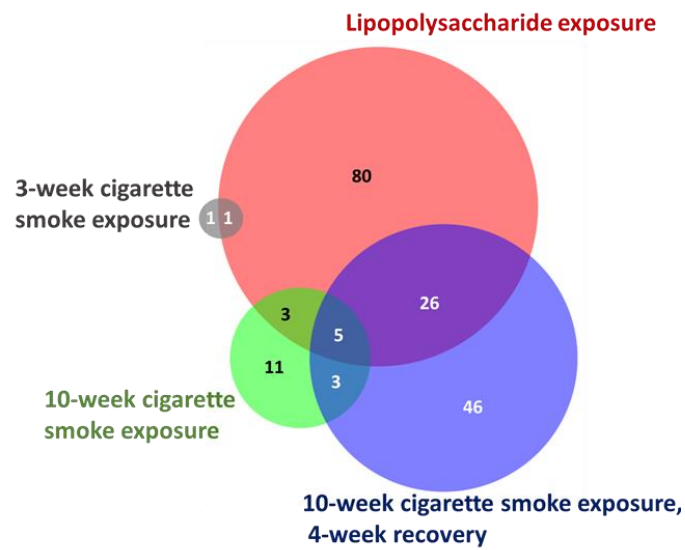
ENSEMBL	Gene	Description	mRNA logFC	protein logFC
ENSMUSG0000004812 0	Entpd1	Ectonucleoside triphosphate diphosphohydrolase 1	-0.429422114	3.421071062
ENSMUSG0000005033 5	Lgals3	Galectin	-0.402797711	4.270588617
ENSMUSG0000002867 2	Hmgcl	Hydroxymethylglutaryl-CoA lyase, mitochondrial	-0.272393928	4.395934294
ENSMUSG0000002073 3	Slc9a3r1	Solute Carrier Family 9, Subfamily A (NHE3, Cation Proton Antiporter 3), Member 3 Regulator 1	0.274244354	3.692081481

4.3.7 Direct comparison of LPS- and cigarette smoke-induced proteome changes shows distinct differences

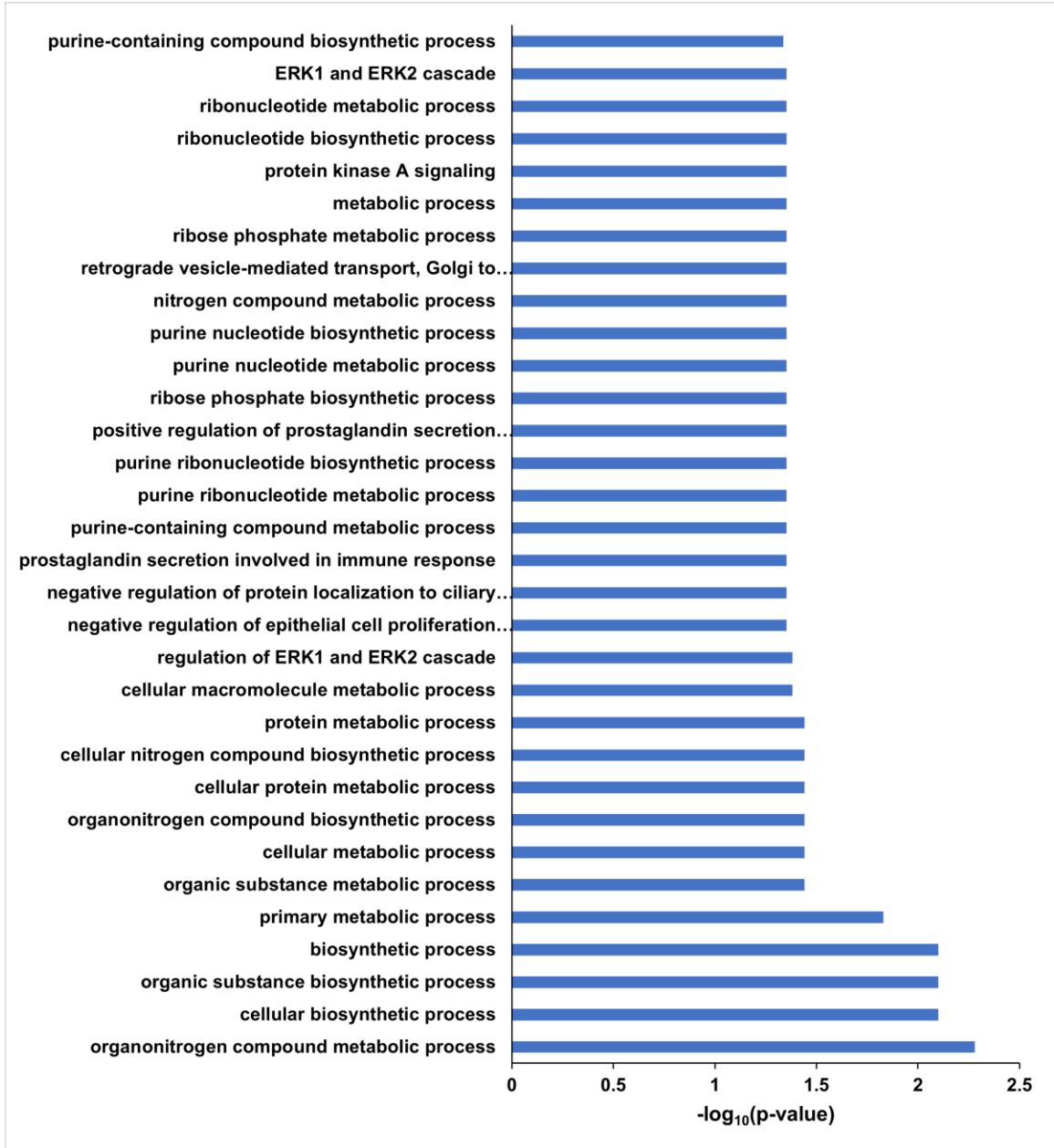
Changes in protein abundances for LPS-exposure and cigarette smoke-exposure were compared to one another to determine the degree of similarity to one another (**Figure 4.10**). Of the proteins that were increased in abundance after 3 weeks of cigarette smoke exposure, only Rbm17 was found to be increased following 3 weeks of LPS exposure. In the cases of Type II cells exposed to 3 weeks of LPS, 10 weeks of cigarette smoke, and 10 weeks of cigarette smoke with subsequent recovery, most proteins that were increased in each condition were unique to those conditions. Only five genes- Ruvbl1, Gnai2, Entpd1, Pdhx, and Psma6- were shared across LPS exposure as well as 10 weeks of cigarette smoke exposure and 10 weeks of cigarette smoke exposure with 4 weeks of recovery. The greatest similarities were noted between the proteomes of LPS exposure and 10 weeks of cigarette smoke exposure with recovery, with 31 proteins increased in abundance across both conditions. GO analysis of these genes shows enrichment of Biological Process terms associated with signal cascades and nucleic acid metabolism.

Figure 4.10. A comparison of proteins with significantly changed abundances following exposure. a) A Venn diagram comparing proteins increased in abundance following LPS exposure, 3 weeks of cigarette smoke exposure, 10 weeks of cigarette exposure, and 10 weeks of cigarette smoke exposure followed by a 4-week recovery period. b) GO analysis of proteins increased in abundance as a result of LPS exposure and 10 weeks of cigarette smoke exposure followed by 4 weeks of recovery.

a)



b)



4.3.9 A comparison of CPTAC lung cancer data shows potential therapeutic candidates of interest

We hypothesized that proteins which were significantly and sustainably increased following prolonged exposure to cigarette smoke may underpin tobacco smoke-driven oncogenesis with Type II cells and could serve as valuable biomarkers of this transformation or potential therapeutic targets. To identify proteins of interest in our data, publicly available proteomics datasets from the CPTAC3-Discovery lung adenocarcinoma (LUAD) were downloaded from the Clinical Proteomic Tumor Analysis Consortium (CPTAC)¹⁷⁷ and run through the Perseus data analysis suite to determine the proteins that were significantly increased and decreased in lung cancer. Due to the large number of samples in the cohort (117 tumor samples and 101 normal samples), many proteins were found to be significantly increased and decreased in these datasets (3578 proteins increased, 2789 proteins decreased).

Having observed the eight genes that were observed to be significantly increased after 10 weeks of cigarette smoke exposure with and without a post exposure recovery period, we then mined the increased proteins in LUAD data for these genes to determine potential oncogenes for study in smoking-driven cancer etiology. Of these eight genes, four genes- Pdhx, Psma6, Ruvbl1, and Ywhaq- were found to be significantly increased in lung adenocarcinoma and will be considered for further study (**Table 4.3**).

Table 4.3. Genes showing sustained increases in protein abundance after 10 weeks of cigarette smoke exposure also seen upregulated in CPTAC data. Highlighted cells represent significant values.

Gene	Description	logFC 3wks	logFC 10wks	logFC 10wks + 4wks recovery	CPTAC logFC
Pdhx	Pyruvate dehydrogenase protein X component, mitochondrial	0.2154	1.6422	1.1681	0.2177
Psma6	Proteasome subunit alpha type 6	-0.2788	4.0808	1.9039	0.1061
Ruvbl1	RuvB-like 1	3.4202	4.9637	5.3720	0.2067
Ywhaq	14-3-3 protein theta	-0.3169	4.2586	2.2421	0.2416

4.4 Discussion

In this study we sought to perform quantitative proteomics analyses of Type II pneumocytes which were isolated from the lungs of mice exposed to the inflammatory stimuli LPS and cigarette smoke. Due to their relatively low percentage of the makeup of lung tissue³¹⁴ in addition to the observed influx of immune cells in response to LPS or cigarette smoke exposure, many samples presented only a limited number of Type II cells and therefore a limited amount of protein for digestion (**Table 4.3**; the two female control samples from the 3 weeks of cigarette exposure represented extremely high amounts of protein as these represented the extraction of all the cells isolated from these samples, rather than a fraction as these were not needed for DNA and RNA extraction). Many of these samples fell below the lower limits of our in-house methods for TMT-labeling of peptides and fractionation of peptides prior to bottom-up proteomics, necessitating the developing and vetting of novel methodologies for our lab. These methods were quite successful, allowing for the detection and quantitation of 3352 proteins using conventional bottom-up proteomics procedures. These protocols will be of use in future studies of other samples with relatively low yields of protein.

In our analysis of LPS-exposed type II cells, we saw an increase in proteins that was consistent with an inflammatory response based on the GO analysis of these genes. In comparing these data with transcriptomic data acquired on the same data, we saw, barring genes associated with cell secretions which had a significant increase in protein abundance without a concurrent increase in gene transcription, a general agreement between these data, with a comparison of these datasets being somewhat linear and the significantly increased genes in both the proteome and transcriptome associated with the inflammatory response.

When also sought to integrate epigenomics data from the same Type II cells with our proteomic and transcriptomic data to obtain as complete a molecular-based phenotypic picture as was possible. In doing this, we saw five genes- S100a4, Cpm, Tppp, Vim, and Fmn11- which showed significant changes in methylation in addition to significant changes at the proteome and transcriptome levels. Each of these genes showed matching changes in protein abundance and gene expression. Interestingly, known oncogene S100a4^{315,316}, was shown to have increased expression while the promotor region was methylated, in contrast to the commonly accepted mode of gene silencing by promotor methylation³¹⁷. A significant decrease was seen in Cpm with promoter and intron methylation, suggesting a more complicated model of genetic control such as has been seen in evaluation of cancer gene expressions³¹⁸. Further analysis of the sequence context of the CpG sites that showed altered methylation is needed to evaluate the interplay between these three levels of biological information.

Our initial hypothesis in conducting this research was that cigarette smoke served to trigger oncogenesis at least partly through the inflammatory pathway; for this reason, we included LPS inflammation of the lungs as a positive control for inflammation of Type II cells. However, further analysis suggested that this was not the case. The proteins that saw increased abundance following LPS exposure and the various degrees of cigarette smoke exposure showed little commonality between them. In considering the multi-omic analyses of proteomic and transcriptomic data, the LPS data has a degree of linear correlation having a Pearson correlation of 0.282; by contrast, the QuanTP correlation plots of the cigarette smoke exposures show much lower degrees of correlation (0.0387, 0.0202,

and 1.96e-04 for 3 weeks, 10 weeks, and 10 weeks of exposure with a 4-week recovery period, respectively).

Despite these differences, LPS exposure and cigarette smoke exposure had some interesting commonalities. The conditions that most matched LPS exposure at the proteome level were the exposure of Type II cells to cigarette smoke for 10 weeks followed by a 4-week recovery period, with 31 proteins in common between them showing increased abundance. From GO analyses, these proteins appeared to be largely involved in signal transduction through the ERK cascade as well as nucleic acid metabolism, processes which are known to be involved in oncogenesis³¹⁹. By contrast, when considering the proteins that are increased in abundance after 10 weeks of cigarette exposure, they correspond with pathways consistent DNA repair, arresting of the cell cycle, hypoxia response, etc. (**Figure 4.8d**), suggesting a state of cancer prophylaxis that is overridden by unknown factors in the 4-week recovery period.

Given the well-established links between tobacco smoke consumption and the development of lung cancer, we hypothesized that extended exposure of Type II cells to cigarette smoke may result in phenotypic changes similar comparable to those seen in lung cancer patients. Eight proteins that are upregulated after 10 weeks of cigarette smoke exposure and remain upregulated after 4 further weeks of recovery in clean air; in comparing these with the proteins upregulated in the CPTAC lung adenocarcinoma dataset we see four homologous genes in common between the two- Pdhx, Psma6, Ruvbl1, and Ywhaq. Pyruvate dehydrogenase complex component X (Pdhx) is a constituent of the PDH complex needed for generation of acetyl coenzyme A is known to be unregulated in certain cancers³²⁰. Similarly, increased expression of Proteasome subunit alpha type-6

(Psm6), a component of the 20S proteasome, was found to be characteristic of lung cancer cell lines³²¹. Ruvb1 is known to be a component of histone acetylating complex and is dysregulated in some cancers³²². The fourth gene, Ywha, aka tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta, has a more nebulous role as a mediator of signal transduction through phosphorylated serine- or threonine-containing residues though has been shown to be increased in breast cancer³²³. Together, these represent potential bridges from inflammation to oncogenesis and may have therapeutic potential in the context of lung cancer.

In summary, we were able to perform quantitative proteomics on Type II cells isolated from mice exposed to LPS and to cigarette smoke for various periods of time, as well as integrate these data with transcriptomic and epigenomic data from the same samples. We found that LPS exposure produced an inflammatory phenotype with considerable correlation between the proteomic and transcriptomic data as well as a few genes that appeared to show significant changes at the methylome, transcriptome, and proteome levels. In addition, we found very few changes to the Type II pneumocyte proteome after 3 weeks of cigarette smoke exposure, though at 10 weeks of exposure and 14 weeks of exposure we saw considerably more changes; furthermore, we found four genes that showed sustained significant increases in protein abundance after 10 weeks of exposure and then a further four weeks of recovery that were found to be significantly increased in tumor samples of patients with lung adenocarcinomas. Future work will focus on elucidating the links between the altered methylation and upregulated gene expression with LPS exposure as well as examining the four genes selected in the cigarette smoke exposed samples for their suitability as therapeutic targets in lung cancer.

V. FURTHERING THE USE OF BOTTOM-UP PROTEOMICS FOR UNTARGETED ADDUCTOMICS OF HEMOGLOBIN

Adapted in part from:

Rajczewski AT, Ndreu L, Pujari SS, et al. Novel 4-Hydroxybenzyl Adducts in Human Hemoglobin: Structures and Mechanisms of Formation. *Chem Res Toxicol.* 2021;34(7):1769-1781. doi:10.1021/acs.chemrestox.1c00111

This work was performed in collaboration with Lorena Ndreu, Efstathios Vyronidis, and Dr. Suresh Pujari under the direction of Drs. Timothy J. Griffin, Margareta Å. Törnqvist, Isabella Karlsson, and Natalia Y. Tretyakova. Andrew T. Rajczewski, Efstathios Vyronidis, and Lorena Ndreu incubated and processed samples and performed global and targeted mass spectrometry. Dr. Suresh Pujari synthesized the quinone methide precursor used in these experiments. Andrew T. Rajczewski generated the figures and wrote and edited the manuscript under the guidance of Drs. Timothy J. Griffin, Isabella Karlsson, and Natalia Y. Tretyakova.

5.1 Introduction

Over their lifetimes, humans are exposed to large numbers of endogenous and exogenous electrophilic compounds, resulting in the formation of covalent adducts at nucleophilic sites of cellular biomolecules^{324,325}. These electrophiles range from products of normal cellular metabolism³²⁶ and dietary components^{327,328} to the metabolic byproducts of commensal microbiota³²⁹ and environmental contaminants^{330,331}. Characterization of the totality of protein and DNA adducts induced by endogenous and exogenous exposures in humans (adductomics) is of considerable importance for disclosing risk factors to human health as adducts are associated with an increased risk of cancer³³² and other chronic diseases^{333,334,335}.

Protein adducts are useful biomarkers of exposure to electrophiles due to their longevity as compared to DNA lesions, since the latter can be efficiently removed via cellular repair mechanisms³³⁶. Specifically, hemoglobin adducts are commonly used in human exposome studies due to the high abundance of hemoglobin in blood, ready blood sample availability, and the relatively long half-life of hemoglobin (120 days)³³⁷. Adducts to the N-terminal valine of hemoglobin are commonly used as biomarkers of human exposure due to the high solvent accessibility of this site and its low pKa value as compared to other potential nucleophilic sites within the protein^{338,339,340}. The Törnqvist group has developed an untargeted methodology for identifying novel N-terminal valine hemoglobin adducts (FIRE)³⁴¹. In this approach, the N-terminal valine residue of human hemoglobin is derivatized and cleaved via modified Edman degradation to yield a fluorescein isothiohydantoin-valine derivative (Figure 1a), which can subsequently be analyzed by HPLC-ESI-MS/MS^{342,343}. Studies using hemoglobin isolated from human subjects

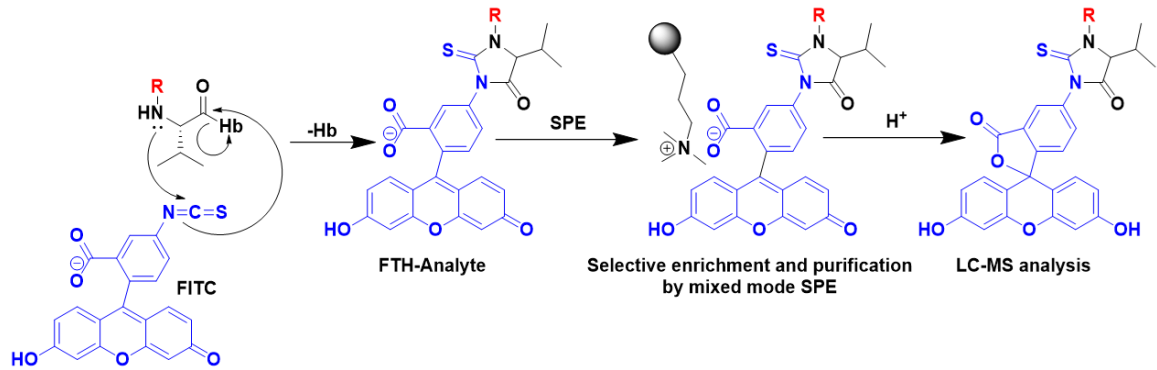
revealed a number of adducts at the N-terminal valine of hemoglobin resulting from methylation, carboxymethylation, and ethylation, as well as modification by ethyl vinyl ketone, acrylic acid, glyoxal, methylglyoxal, and 1-octen-3-one^{341,344}. Our recent work using the FIRE approach determined that the third-most abundant N-terminal valine adduct (106.042 Da, corresponding to the elemental formula C₇H₆O) corresponded to the addition of a 4-hydroxybenzyl group³⁴⁵. Although the exact source of this adduct in humans remained unknown, we proposed that they could be a result of para-quinone methide (para-QM)^{346,347,348} or 4-hydroxybenzaldehyde (HBA)³⁴⁹ reactions with the N-terminal valine of hemoglobin.

While the solvent exposed N-terminal valine residue of Hb is frequently modified upon exposure to electrophiles, side chains of many amino acids including cysteine, lysine, and histidine are inherently nucleophilic, rendering them as potential sites for adduct formation^{350,351}. For example, Cys34 of human serum albumin is a likely adduction site upon exposure to electrophiles^{352,353} and has been previously used to monitor human exposures³⁵⁴. We have shown that Cys145 of human *O*⁶-alkylguanine DNA alkyltransferase (AGT) is the preferred modification site upon exposure to 1,2,3,4-diepoxybutane³⁵⁵ and antitumor nitrogen mustards^{356,357} while cis-diamminedichloroplatinum (II) (cisplatin) targets several nucleophilic residues within the AGT protein including Glu110, Lys125, Cys145, His146, Arg147, and Cys150³⁵⁸. Sulfonamide adducts at β Cys93 of hemoglobin are associated with exposure to tobacco smoke components and grilled meat byproducts such as 2-amino-9H-pyrido[2,3-b]indole 4-aminobiphenyl (HONH-ABP)^{359,360}. Further, α,β -unsaturated aldehyde byproducts of

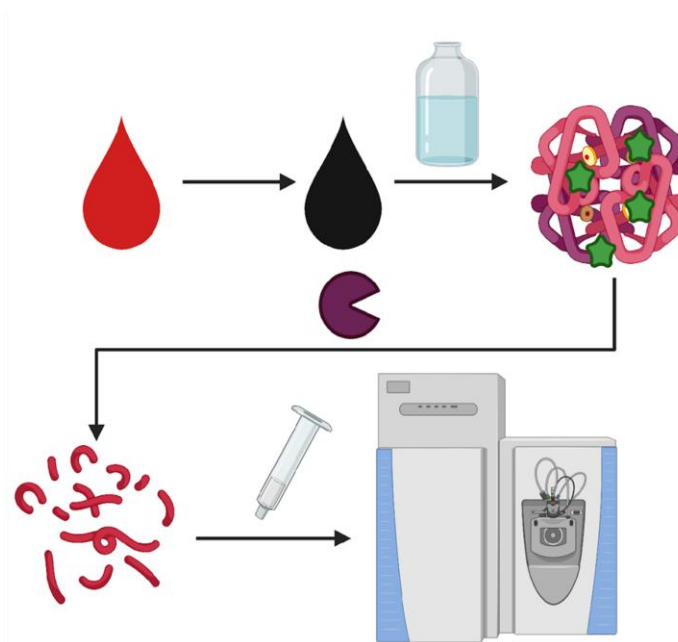
lipid peroxidation have been shown to form adducts at histidine side chains in hemoglobin³⁶¹.

Figure 5.1. Approaches for detection of hemoglobin adducts by mass spectrometry. a) Schematic representation of the use of fluorescein isothiocyanate, or FITC (FI) for the measurement of N-terminal protein adducts (R) via modified Edman (E) procedure (FIRE) b) Bottom-up proteomics approach for detection of hemoglobin adducts.

a)



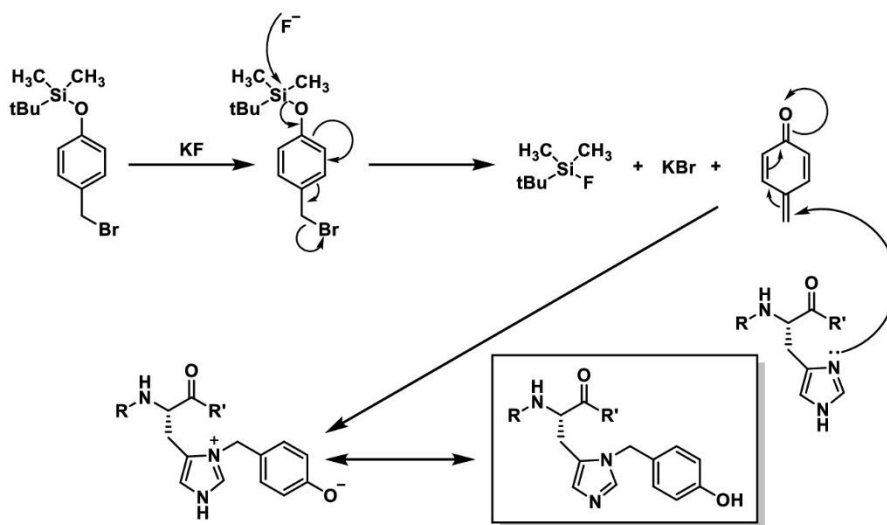
b)



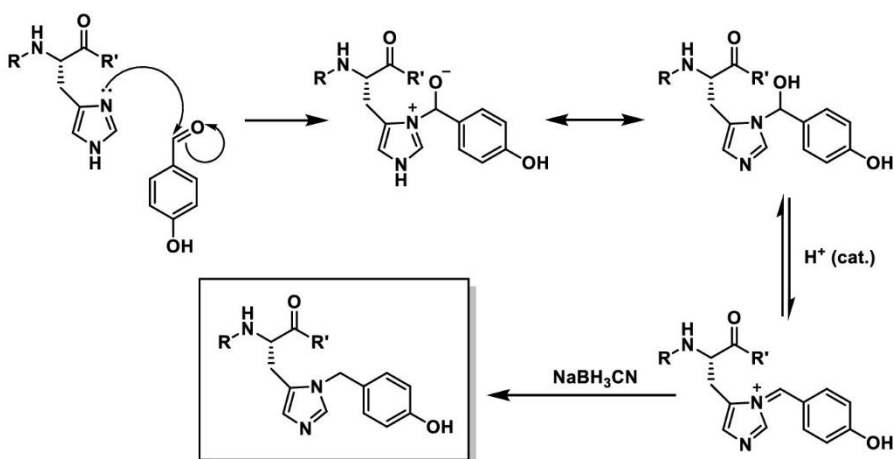
In the present study we sought to demonstrate the utility of bottom-up proteomics in the identification of hemoglobin adducts. First, we utilized bottom-up proteomics to characterize additional sites of 4-hydroxybenzyl adduct formation in hemoglobin and to help elucidate their potential source(s) in humans (Figure 1b). To help control the reaction of para-QM with hemoglobin, the molecule itself was eschewed in favor of a precursor that could be converted into the highly reactive para-QM when desired. A stable synthetic para-QM precursor^{362,363}, [4-(bromomethyl)phenoxy]-tert-butyl-dimethylsilane, was activated by potassium fluoride to form para-QM in situ (**Figure 5.2a**). Bottom-up global and targeted proteomics nanoLC-MS/MS workflows were used to identify the amino acid residues within protein containing 4-hydroxybenzyl adducts. Alternate sources of 4-hydroxybenzyl adducts in humans were also investigated through the reactions of hemoglobin with 4-hydroxybenzaldehyde (4-HBA) and through exposure of whole blood to UV light with and without addition of tyrosine, a possible precursor of para-QM (**Figures 5.2b, 5.2c**)³⁶⁴. Additionally, we sought to directly compare the FIRE protocol with the bottom-up proteomics approach, evaluating their utility for untargeted adductomics using a panel of electrophiles incubated with donor blood (**Figure 5.3**).

Figure 5.2. Proposed reaction mechanisms for the formation of 4-hydroxybenzyl adducts using histidine side chains as an example. a) Formation of para-QM from para-QM precursor with potassium fluoride and subsequent reaction. b) Reductive amination of 4-hydroxybenzaldehyde with sodium cyanoborohydride. c) Generation of para-QM upon reaction with ultraviolet light and subsequent adduct formation.

a)



b)



c)

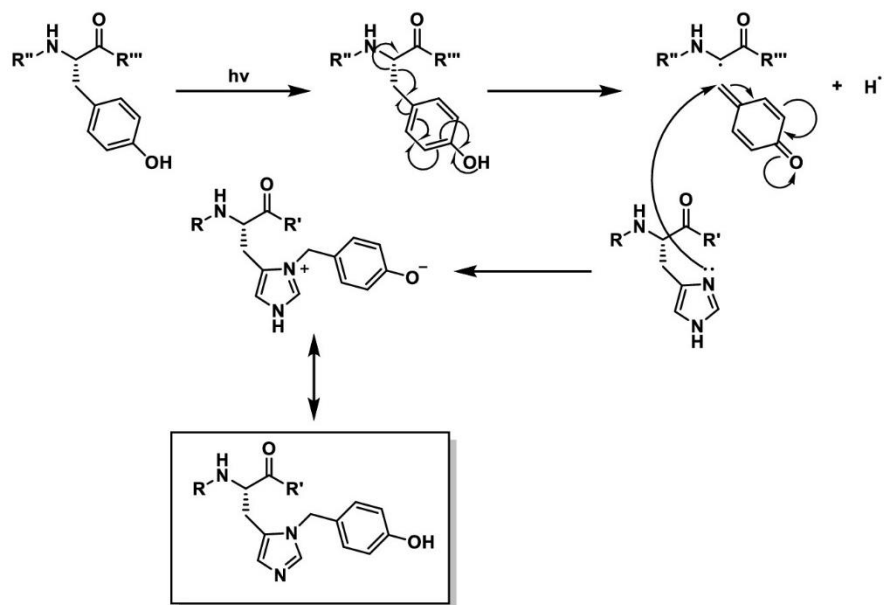
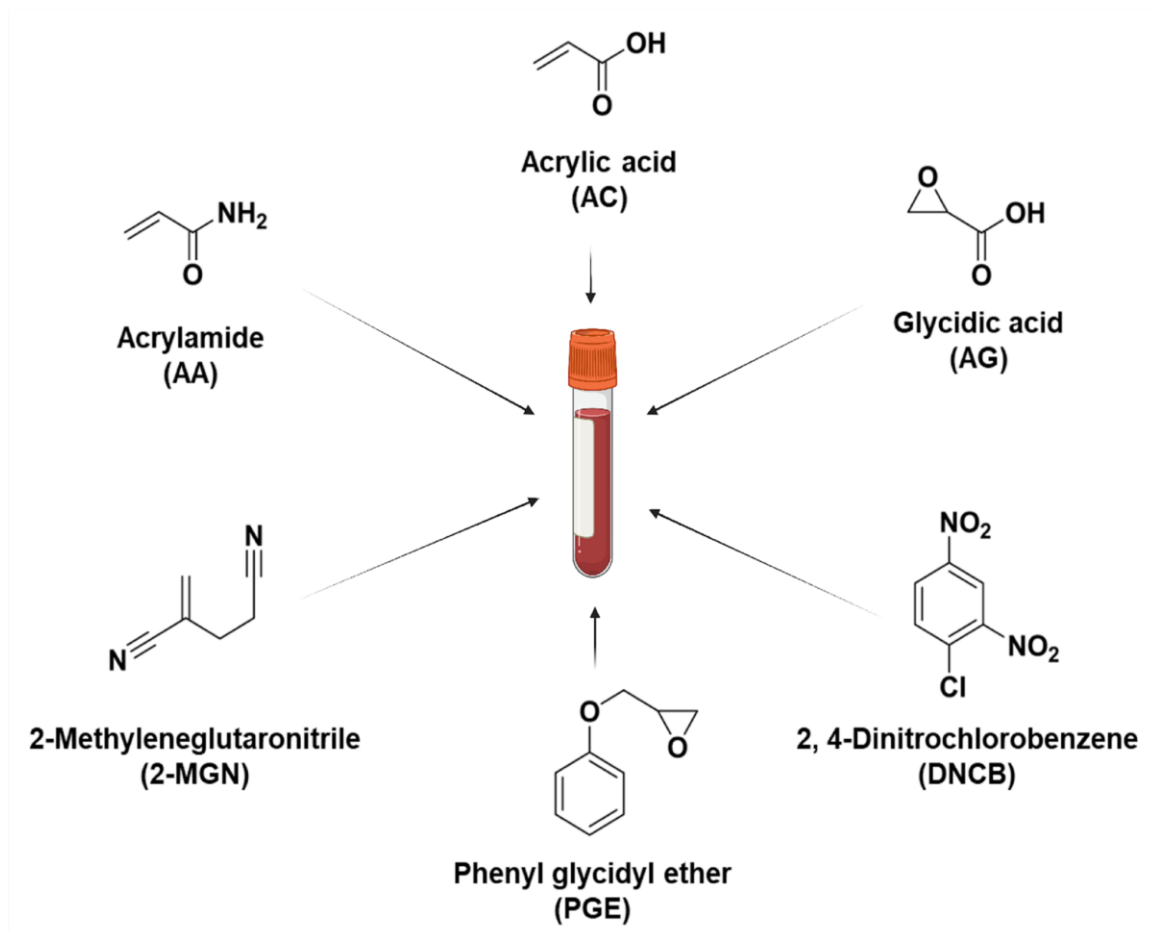


Figure 5.3. Study design to compare FIRE and bottom-up proteomics approaches for the detection of hemoglobin adducts. Donor blood is incubated with either acrylamide, acrylic acid, glycidic acid, 2-methyleneglutaronitrile (2-MGN), 2,3-epoxypropyl phenyl ether (PGE), or 1-chloro-2,4-dinitrobenzene (DNCEB) at varying concentrations prior to analysis using both methodologies and data analysis.



5.2. Materials and methods

Materials and instrumentation

For the analyses of 4-hydroxybenzyl adduct formation, human blood with added potassium EDTA was purchased from Biochemed Services (Winchester, VA). For bottom-up proteomics/FIRE comparisons, blood was acquired from the Karolinska Universitetslaboratoriet. [4-(Bromomethyl)phenoxy]-*tert*-butyl-dimethylsilane (para-QM precursor) was synthesized in our laboratory according to established protocols^{362,363}. 4-Hydroxybenzaldehyde, sodium cyanoborohydride, potassium chloride, potassium phosphate, trifluoroacetic acid, oxidized L-glutathione, acrylamide, acrylic acid, potassium oxirane-2-carboxylate, 1-chloro-2,4-dinitrobenzene, 2,3-epoxypropyl phenyl ether, and 2-methyleneglutaronitrile were purchased from Millipore Sigma (Burlington, MA). Potassium fluoride was obtained from Mallinkrodt (Staines-upon-Thames, UK). Sodium chloride and calcium chloride were procured from Thermo Fisher Scientific (Waltham, MA). Trypsin was purchased from Promega Corporation (Madison, WI). L-Tyrosine was provided by C. Eddington. Formic acid was purchased from Honeywell Fluka (Mexico City, MX). LC-MS grade acetonitrile and water were obtained from Thermo Fisher Scientific (Waltham, MA).

All nanoLC-MS/MS analyses were performed on a QExactive Orbitrap Mass Spectrometer interfaced to a Dionex Ultimate 3000 nanoLC System set to nanoflow mode. Liquid chromatography was carried out using a reverse-phase C18 nanospray column (250 μm x 34 cm) manually packed with Luna 5 μm C18 solid-phase media (Phenomenex). Photoreactions were performed in an RMR-600 Photochemical Reactor from Rayonet (Branford, CT).

Isolation of hemoglobin from blood

Hemoglobin was isolated from fresh human blood according to published protocols. Briefly, 10 mL of human blood was centrifuged at 800 g for 15 min at 4 °C to separate the erythrocytes from the plasma. The plasma was decanted from the pelleted erythrocytes, which were then subjected to three washes with an equal volume of cold Ringer's solution (250 mM NaCl, 10 mM KCl, 3 mM CaCl₂, pH = 7.4) followed by further centrifugation. To isolate hemoglobin, erythrocytes were suspended in an equal volume of distilled water and subjected to 5 min of sonication. Samples were then centrifuged at 15,000 rpm for 15 min to pellet cellular debris, reserving the hemoglobin-containing supernatant for further experiments. The concentration of HbO₂ in the supernatant was ascertained via absorption at $\lambda = 542 \text{ nm}$ ($\epsilon = 14.62 \text{ mM}^{-1}\text{cm}^{-1}$).

Hemoglobin reactions with para-QM precursor

Aliquots of human hemoglobin (64 mg) were diluted to 100 μL with 100 mM potassium phosphate buffer (pH= 7.4) containing 50 mM KF. The final hemoglobin concentration was 10 mM. Para-QM precursor was dissolved in DMSO, and aliquots were added to the hemoglobin samples ($n = 3$) to achieve the final concentrations of 1 mM, 5 mM, 10 mM, or 50 mM QM, followed by incubation overnight at 37 °C. Following treatment with para-QM, hemoglobin samples were added to Amicon Ultra centrifugal filters (Millipore-Sigma) with a 10 kDa cutoff and centrifuged at 14,000 g for 10 min. A solution of triethylammonium bicarbonate (TEAB, 25 mM pH = 8.0) was then added to the spin columns, followed by two additional rounds of buffer exchange.

Hemoglobin Reactions with 4-HBA

4-Hydroxybenzaldehyde (4-HBA) was dissolved in ethanol to obtain a 1 M stock solution. In two sets of three replicates, hemoglobin aliquots and 4-HBA stock were added to distilled water to give the final concentrations of 10 mM hemoglobin and 50 mM 4-HBA, respectively. Samples were incubated at 37 °C overnight. Following treatment, one set of samples was supplemented with sodium cyanoborohydride to a concentration of 10 mM and incubated at room temperature for 30 min. Samples were buffer exchanged into TEAB buffer three times using 10 K filters as described previously.

Exposure of human blood to ultraviolet radiation

Four quartz cuvettes were each filled with 300 µL of whole blood and subdivided into two groups of 2. One set of cuvettes was also supplemented with 250 µL of a saturated solution of L-tyrosine in Ringer's solution, while the other was supplemented with 250 µL of Ringer's solution. The solutions were exposed to ultraviolet light (254 nm) at room temperature for one hour. As a negative control, a set of cuvettes containing 250 µL blood and 250 µL Ringer's solution were incubated in the dark for one hour at room temperature. Following incubation, the samples were transferred to Eppendorf vials and centrifuged at 800 g for 10 min to pellet the erythrocytes in solution. The erythrocytes were then washed three times in Ringer's solution before being lysed in an equivalent volume of distilled water accompanied by 5 min of sonication. The samples were then centrifuged at 15,000 rpm for 15 min to pellet cellular debris, reserving the supernatant for mass spectrometry analysis.

Exposure of human blood to electrophile panel

To compare the ability of FIRE and bottom-up proteomics to detect hemoglobin adducts, 500 μL aliquots of donor blood were incubated with individual electrophiles at the electrophile concentrations and incubation times as documented in **Supplementary Table 5.1**. All incubations were conducted at 750 rpm at 37°C. Following exposure to electrophiles, blood samples were centrifuged at 2000 g for 5 minutes to pellet the erythrocytes, after which the plasma was removed and discarded, and the cells washed with 1 mL of cold 0.9% NaCl; this centrifugation and washing was repeated twice more. To lyse the cells, 0.7 mL of pure milli-q water were added to erythrocytes after which the samples were sonicated for 10 minutes and spun down at 15,000 rpm for 15 min to remove the cellular debris. The hemoglobin content of the samples was ascertained using the nanodrop as described above.

Sample processing for FIRE analysis

Following the isolation of hemoglobin, samples for FIRE were set up where 100-150 g/L hemoglobin solutions were added to 2.0 mL Eppendorf tubes to a final volume of 250 μL . To these were added 15 μL of fresh 1M KHCO_3 and 30 μL of FITC stock solution (5 mg solved in 30 μL DMF / sample). Samples were then incubated overnight at 37°C with 750 rpm rotations. Following overnight incubation, D_7 -labeled internal standards (2 pmol of acrylamide-Val- D_7 -FTH) was spiked into each sample. The remaining protein was precipitated via the addition of 1.5 mL acetonitrile followed by extensive vortexing and centrifugation at 3000 g for 5 minutes. The supernatants were decanted and supplemented with the addition of 25 μL of 1M NH_4OH .

To purify the FTH-analytes, Oasis MAX 1cc SPE cartridges were placed into an SPE manifold attached to a vacuum line. The columns were conditioned via the addition of 0.5 column volumes of 10 mM NH₄OH. Samples were then added to individual SPE cartridges and washed with 1 column volume of each acetonitrile and water. To protonate the FTH-analytes, 0.5 column volumes of 0.5% cyanoacetic acid in water were added to each cartridge, and air was blown through each cartridge to remove excess solvent. FTH-analytes were eluted with 1.1 mL of 0.25% cyanoacetic acid in 60% acetonitrile, after which the samples were dried overnight under reduced pressure in a speed vac apparatus.

Sample processing for proteomics analyses

Buffer-exchanged hemoglobin samples were treated with a 10-fold molar excess iodoacetamide in TEAB buffer in the dark at room temperature for 30 min. Following incubation, aliquots of protein (50 µg) were taken and buffer-exchanged three times into TEAB buffer. Each sample was then supplemented with trypsin at a ratio of 1:20 w/w and incubated overnight at 37 °C. Proteolytic digestion was terminated via the addition of formic acid to 10%, after which the samples were desalted via C18 spin columns and evaporated to dryness under vacuum.

Global nanoLC-MS/MS analyses

Tryptic digests were reconstituted in buffer A (0.1% FA in water) and analyzed on an Orbitrap QExactive Mass Spectrometer interfaced with a reverse-phase HPLC system operated in the nanoflow mode. The nanoLC column was eluted at a flow rate of 300 nL/min. Samples were analyzed using a gradient of 5 – 22 % buffer B (0.1% FA in

acetonitrile) over 71 min, followed by 22 - 33 % over 5 min, 33 - 90% over 5 minutes, a 90% buffer B wash for 4 min, and finally a 90 - 4% decrease in buffer B over 2 minutes followed by a 4 min equilibration at 4% B.

Peptides were analyzed in the positive ion mode using Full MS/dd-MS/MS. The instrument was operated at 70,000 resolution with an AGC target of 1 e^6 , a maximum IT of 30 ms, and a scan range of 300 to 2000 m/z. Tandem mass (MS/MS) spectra were captured at 17,500 resolution, AGC target of 5 e^4 , maximum IT of 50 ms, an isolation window of 2.0 m/z and a normalized collision energy of 30. Data were collected in the centroid mode. Control hemoglobin samples were used to create an exclusion list of unmodified hemoglobin peptides³⁶⁵.

Raw mass spectrometry data were analyzed using Proteome Discoverer 2.2. The data was searched against the SwissProt human proteome version 2017-10-25 containing the conventional hemoglobin FASTA sequence supplemented with hemoglobin alpha and beta subunit FASTA files with the N-terminal methionine residues removed to account for the N-terminal cleavage common in eukaryotic post-translational modification³⁶⁶. Variable modifications included oxidation at methionine, carbamidomethyl modification of cysteine, and a 106.042 Da modification at cysteine, histidine, lysine, arginine, serine, threonine, tyrosine, and N-termini of peptides corresponding to the putative 4-hydroxybenzyl group ($\text{C}_7\text{H}_6\text{O}$) with 10 ppm mass tolerance. In addition, a Percolator step³⁶⁷ was added with a strict target FDR of 0.01, a relaxed target FDR of 0.05, and a validation based on the q-value to ensure confident identification of peptide modifications.

Targeted nanoLC-MS/MS analyses

Peptides identified in the global proteomics study were used to create an inclusion list containing high-confidence peptides modified with para-QM (+106.042 Da) alongside the corresponding unmodified peptides. This inclusion list was used for targeted analysis of these peptides in tryptic digests using simultaneous PRM and full MS-SIM experiments. The instrument was operated in the positive ion mode, and samples were analyzed using the same nanoLC column and gradient as above. PRM experiments were conducted with mass resolution of 70,000, the AGC target of 2e5, the maximum IT of 250 ms, the isolation window of 1.0 m/z and a normalized collision energy of 25. The accompanying Full MS-SIM experiment was run at the resolution of 70,000, an AGC target of 3e6, a maximum IT of 200ms, and a scan range of 300 to 2000 m/z .

The resulting targeted mass spectrometry data was analyzed using Skyline³⁶⁸. The FASTA files of human hemoglobin alpha and beta subunits were imported into Skyline and used to build a curated target list of peptides of interest with variable 4-hydroxybenzyl modification at cysteine, histidine, lysine, serine, threonine, and tyrosine residues (added mass of 106.042 Da). Magellan storage files (MSFs) generated by Proteome Discoverer containing the identified MS/MS spectra of the hemoglobin-para-QM global data were used in Skyline to build a spectral library that the targeted data could be compared against. Peak areas of the three most abundant product ions of each precursor peptide were measured at each concentration of added para-QM and exported for analysis. Fraction of modified side chains following the addition of reactive species (para-QM precursor + KF, 4-HO-BA, or UV \pm Tyr) served as a proxy measurement for the reactivity of that side chain towards the added electrophiles. The extent of adduct formation at each site was expressed as a percentage of total peptide detected as follows:

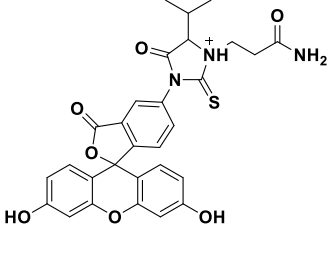
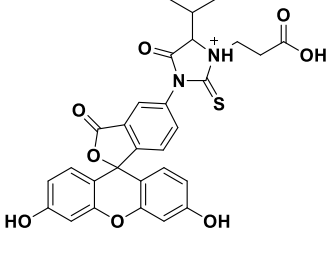
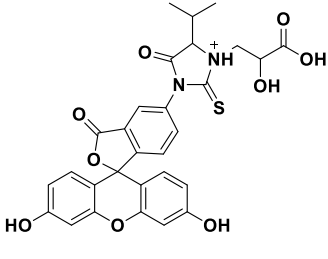
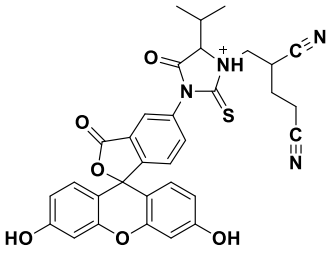
% Modified Side Chain

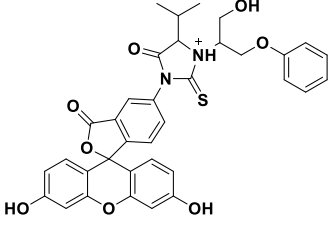
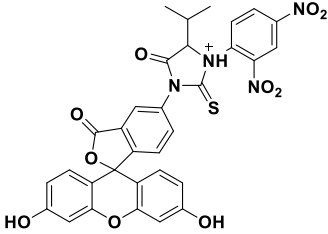
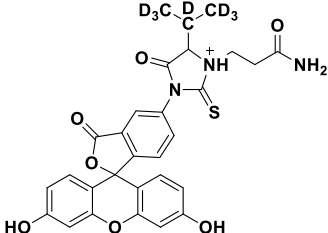
$$= \left(\frac{\Sigma \text{ Integrated Peak Areas of Peptides Containing Modified}}{\Sigma \text{ Integrated Peak Areas of All Peptides Containing Side Chain of Interest}} \right) \times 100\%$$

Targeted detection of N-terminal adducts isolated via FIRE

Following solid phase extraction of FTH derivatives of adducted N-terminal valines from blood, they were reconstituted in 40% ACN for LC-MS analysis. Samples were analyzed on a Q Exactive Quadrupole Orbitrap Hybrid Mass spectrometer interfaced with a Dionex UltiMate 3000 LC system containing a C18 HPLC column (3.0 μm , 2.1 mm \times 150 mm). Experiments were run at a flow rate of 350 $\mu\text{L}/\text{min}$ with solvent A as 0.1% formic acid in water and solvent B as 95% acetonitrile and 0.1% formic acid in water. The solvent composition was held at 25% B from 0 to 0.5 minutes, followed by a linear increase to 75% B by 8 min and further to 95% B by 10 min. The column was eluted with 95% B for 2 min as a washing step, followed by a decrease to 25% B by 12.01 minutes and column re-equilibration for 3 min. The mass spectrometer was run in the PRM mode with a 2 Da isolation window, with an inclusion list based on the m/z values listed in **Table 5.1**; analysis was performed using the $[\text{M} + \text{H} - 36]^+$ fragment peaks resulting from the loss of the isopropyl group in the analytes and the $[\text{M} + \text{H} - 43]^+$ fragment peak in the internal standard.

Table 5.1. Inclusion list for FTH-analytes in FIRE samples. The D₇-acrylamide adduct corresponds to the added internal standard.

Adduct	FTH-analyte formula	FTH-analyte structure	[M+H] ⁺ (<i>m/z</i>)
Acrylamide	C ₂₉ H ₂₆ N ₃ O ₇ S ⁺		560.14860
Acrylic Acid	C ₂₉ H ₂₅ N ₂ O ₈ S ⁺		561.13261
Glycidic Acid	C ₂₉ H ₂₅ N ₂ O ₉ S ⁺		577.12753
2-MGN	C ₃₂ H ₂₇ N ₄ O ₆ S ⁺		595.16458

PGE	$C_{35}H_{31}N_2O_8S^+$		639.17956
DNCB	$C_{32}H_{23}N_4O_{10}S^+$		655.11294
D7-Acrylamide	$C_{29}H_{19}D_7N_3O_7S^+$		567.19253

Predicted physicochemical properties of hemoglobin side chains

A consensus sequence for sites identified as having 4-hydroxybenzyl adducts was generated using ggseqlogo in R³⁶⁹. Briefly, a list of polypeptides containing the modified side chain with 7 residues before and after the adducted site were uploaded into this tool, after which the consensus sequence was generated.

To estimate the pKa values of the amino acid side chains in human hemoglobin, the H++ automated server (version 3.2)^{370,371,372} was used with a normal human hemoglobin NMR structure 2h35³⁷³ obtained from the RCSB Protein Data Bank³⁷⁴. In the determination of the side chain pKa values, the external dielectric constant was set to the default value of 80, and the external salinity to 0.15 M, and the internal dielectric constant to 4.

The relative accessible surface areas of hemoglobin side chains were obtained from the hemoglobin NMR structure 2h35 using the PISA webserver³⁷⁵. These values were subsequently normalized using the theoretical maximal allowed solvent accessibilities of amino acid side chains calculated in Tien *et al.*³⁷⁶

5.3 Results

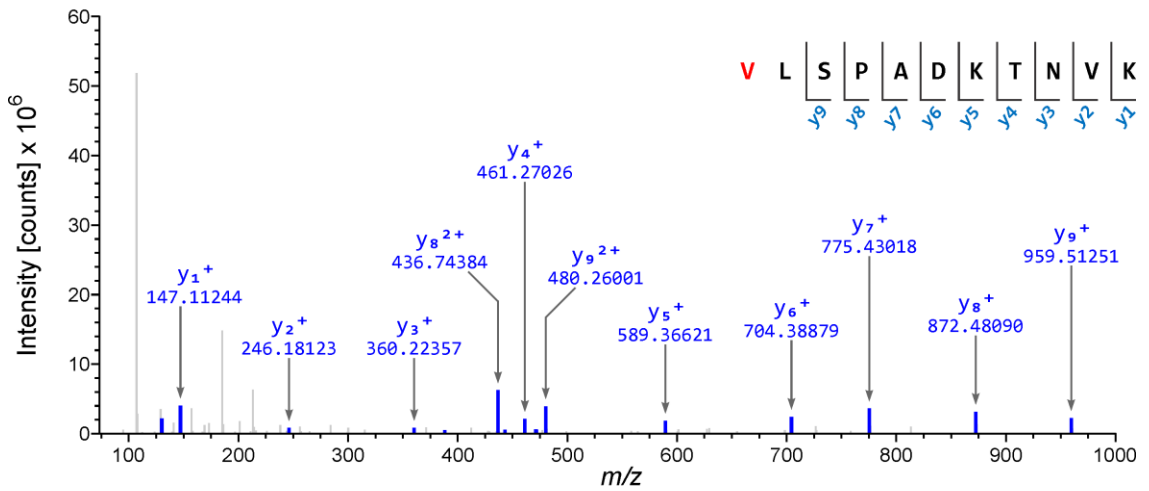
5.3.1 Global proteomic analysis of hemoglobin exposed to para-QM shows adduct formation at nucleophilic side chains

To identify amino acid residues within human hemoglobin that form adducts with para-QM, hemoglobin freshly isolated from human blood was incubated with a five-fold molar excess of para-QM precursor in the presence of potassium fluoride to release para-QM (**Figure 5.2a**). Para-QM treated hemoglobin was digested to peptides with trypsin, and

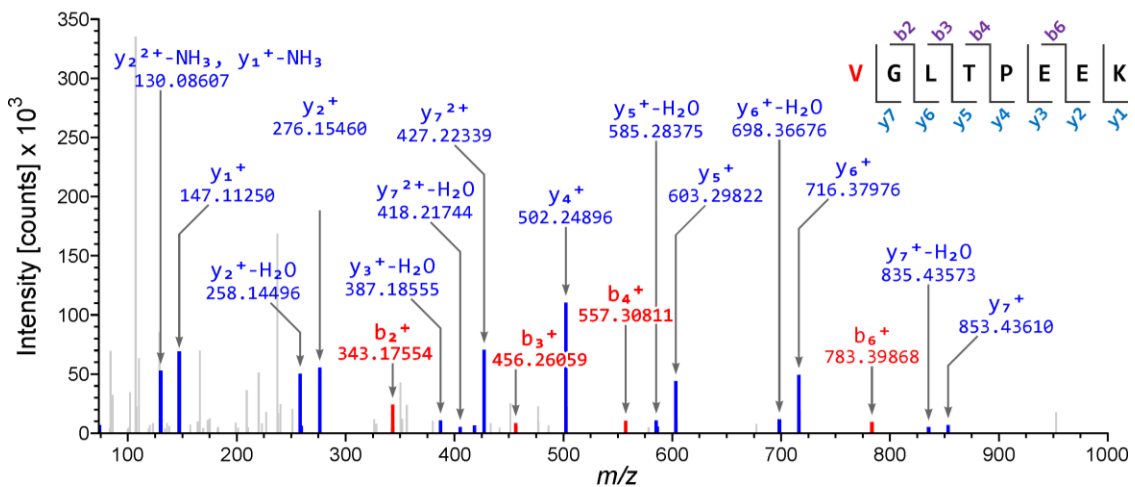
the resulting peptides were analyzed by global nanoLC-MS/MS as described in the Methods section. The resulting mass spectrometry data were processed using Proteome Discover. We utilized a variable modification of 106.042 Da (hydroxybenzyl group) and searched for the presence of this modification at various nucleophilic side chains as well as the N-termini of the protein. The hemoglobin molecule is a tetramer of two alpha and beta subunits; each of the two subunits has its own N-terminal valine residue which could accommodate a 4-hydroxybenzyl adduct. Analysis of the MS/MS spectra of tryptic peptides obtained from para-QM treated hemoglobin showed spectral matches consistent with 4-hydroxybenzyl adduct (106.042 Da) at the N-termini of both alpha and beta subunits of hemoglobin (**Figures 5.4a, 5.4b**). This agrees with our previous work which employed the FIRE procedure to detect 4-hydroxybenzyl adducts at the N-terminal valine of human hemoglobin, but which could not distinguish between subunits of the protein³⁴⁵.

Figure 5.4. Tandem mass spectra of a) Hemoglobin alpha subunit N-terminal peptide with N-terminal 4-hydroxybenzyl adduct, b) Hemoglobin beta subunit N-terminal peptide with N-terminal 4-hydroxybenzyl adduct, c) 4-hydroxybenzaldehyde adduct of histidine 45 in alpha subunit, d) 4-hydroxybenzaldehyde adduct of cysteine 93 in beta subunit. Spectra in a) and b) were sourced from Proteome Discoverer 2.2, c) and d) were taken from Skyline v20.2.

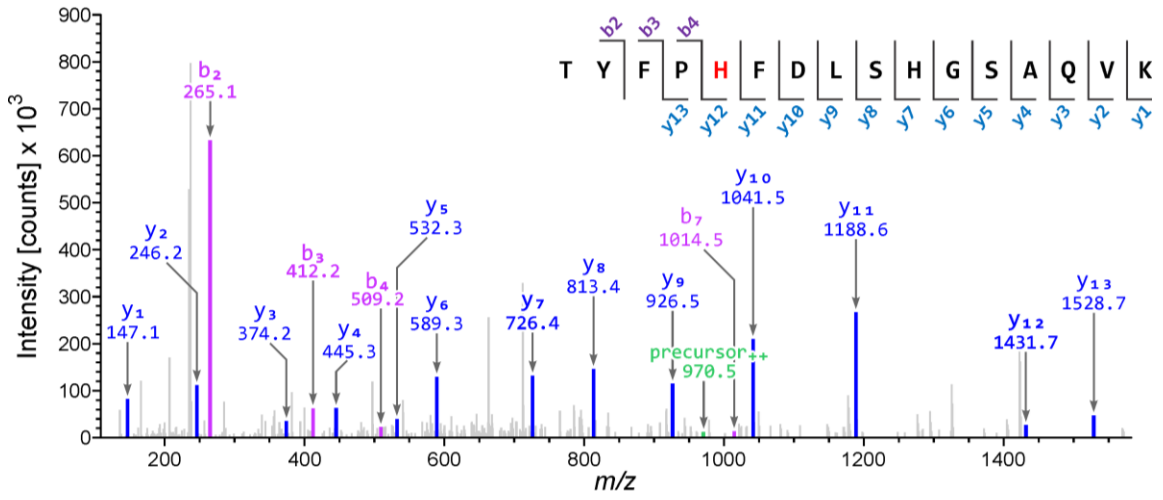
a)



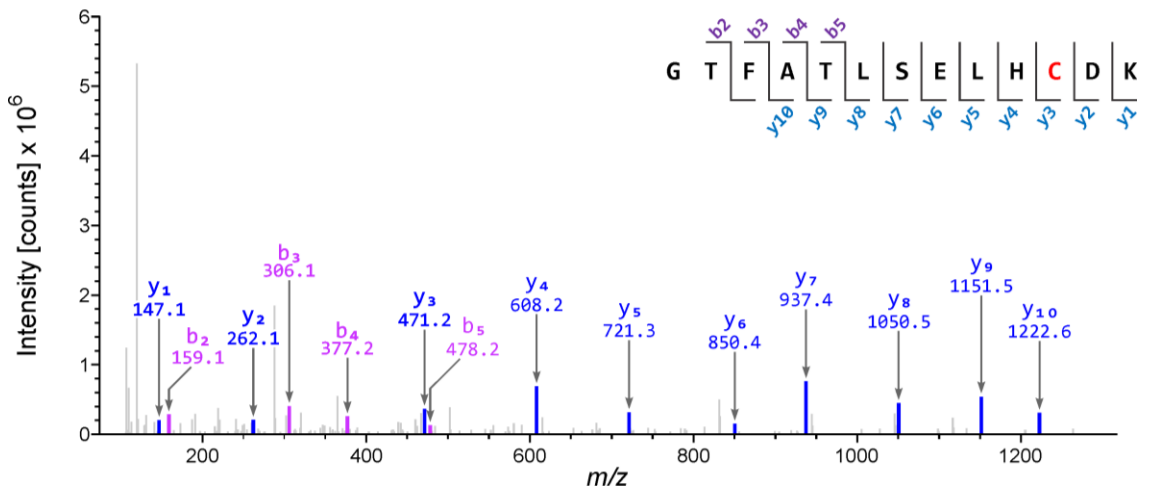
b)



c)



d)



Proteomics analyses allowed for 91% alpha subunit polypeptide sequence coverage and 95% beta subunit polypeptide sequence coverage of human hemoglobin (**Table 5.2**). In addition to 4-hydroxybenzyl adduct formation at N-terminal valines, global proteomics analyses led to preliminary identification of seventy-eight potentially adducted amino acid side chains. All MS/MS spectra were manually interrogated, and only those containing the requisite b- and y-ions for conclusive identification of the peptide and confident placement of the adduct were retained for further examination. This initial filtering resulted in twenty-nine potentially modified peptides (**Table 5.3**).

Table 5.2. Peptides detected in hemoglobin treated with excess para-QM.

Chain	Peptide	Theoretical [M+H]⁺ (Da)	Observed [M+H]⁺ (Da)	Charge
α	VLSPADK	729.41413	729.41197	2
	VGAHAGEYGAEALER	1529.7343	1529.72991	2
	VGAHAGEYGAEALERMFL SFPTTK	2582.27078	2582.26967	4
	MFLSFPTTK	1071.55433	1071.54949	2
	TYFPFDLSHGSAQVK	1833.89186	1833.88895	3
	KVADALTNVAHVDDMP NALSALSDLHAHK	3124.58441	3124.58387	5
	VADALTNVAHVDDMPN ALSALSDLHAHK	2996.48944	2996.48672	4
	LRVDPVNFK	1087.62585	1087.62476	3
	VDPVNFK	818.44068	818.43712	2
	LLSHCLLVTLAAHLPAEFT PAVHASLDK	3024.63392	3024.63735	4
	FLASVSTVLTSK	1252.71473	1252.71453	2
	FLASVSTVLTSKYR	1571.87917	1571.87967	3
	β	VHLTPEEK	952.50982	952.50859
VHLTPEEKSAVTALWGK		1866.01197	1866.00858	3
SAVTALWGK		932.51999	932.51885	2
VNVDEVGGEALGR		1314.66482	1314.66435	2
VNVDEVGGEALGRLLVVY PWTQR		2570.37255	2570.37741	3
LLVVYPWTQR		1274.72557	1274.72539	2

	FFESFGDLSTPDAVMGNP K	2058.94772	2058.94773	3
	VLGAFSDGLAHLNLK	1669.8908	1669.89054	3
	GTFATLSELHCDK	1478.6944	1478.6917	2
	LHVDPENFR	1126.56398	1126.56316	2
	LLGNVLVCVLAHHFGK	1776.99416	1776.99721	4
	EFTPPVQAAYQK	1378.70014	1378.69951	2
	VVAGVANALAHK	1149.67387	1149.67388	2

Table 5.3. Peptides chosen for targeted proteomic analysis.

Chain	Peptide	Modified Residue
α	AAWGKVGAGAGEYGAEALER	K16
	VGAHAGEYGAEALER	V17
	VGAHAGEYGAEALER	H20
	MFLSFPTTKTYFPHFDLSHGSAQ VK	M32
	MFLSFPTTKTYFPHFDLSHGSAQ VK	S35
	TYFPHFDLSHGSAQVK	T41
	TYFPHFDLSHGSAQVK	Y42
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	L100
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	S102
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	H103
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	C104
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	T108
	LLSHCLLVTLAAHPAEFTPAVHA SLDK	H112
β	SAVTALWGKVVNDEVGGEALG R	S9

SAVTALWGKVVNDEVGGEALG	T12
R	
FFESFGDLSTPDAVMGNPKVK	S44
KVLGAFSDGLAHLNLK	K66
VLGAFSDGLAHLNLK	V67
VLGAFSDGLAHLNLK	S72
GTFATLSELHCDK	G83
GTFATLSELHCDK	T84
GTFATLSELHCDK	T87
GTFATLSELHCDK	S89
GTFATLSELHCDK	H92
GTFATLSELHCDK	C93
LHVDPENFR	L96
LLGNVLCVLAHHFGK	C112
EFTPPVQAAYQK	T123
VVAGVANALAHK	V132

5.3.2 Targeted proteomic analysis verifies the presence of para-QM adducts at cysteine, histidine, lysine, serine, threonine, and tyrosine side chains

To verify the presence of 4-hydroxybenzyl adducts at sites of the protein initially identified in global proteomics experiments, targeted mass spectrometry analyses were conducted. Samples were re-analyzed in the parallel reaction mode (PRM) using an inclusion list of para-QM modified peptides obtained from global proteomics experiments (**Table 5.3**). Raw MS/MS data were searched in Skyline³⁶⁸ against a target list constructed from the peptides of the alpha and beta subunits of human hemoglobin with and without 4-hydroxybenzyl modification at the amino acid residues of interest. Assignments of peptides to the target list were confirmed using a spectral library constructed using MSF files generated in Proteome Discoverer during the global mass spectrometry analysis of hemoglobin exposed to excess para-QM. By using strict selection criteria where the exact position of the putative adduct can be localized using the b and y series in the MS/MS spectra, we were able to confirm 4-hydroxybenzyl modification of 14 amino acid side chains in the interior of the alpha and beta subunits of the protein. In samples treated with para-QM (**Table 5.5a**), mass spectrometry evidence was obtained for 4-hydroxybenzyl modification of cysteines (β Cys93, β Cys112), histidines (α His20, α His45, β His92, β His143), serines (β Ser44, β Ser72, β Ser89), threonines (β Thr84, β Thr87, β Thr123), and tyrosines (α Tyr24, α Tyr42) within the protein. Each of these modified peptides contains the 4-hydroxybenzyl adduct assigned to a specific amino acid residue using characteristic b- and y-ion series (**Figures 5.4c, 5.4d**). Hypothetical chemical structures of the adducts are shown in **Figure 5.5**. Interestingly, preliminary analyses suggested that para-QM

modified residues were localized in several discrete clusters within the alpha and beta subunits of the protein (**Figure 5.6**).

Figure 5.5. Putative structures of 4-hydroxybenzyl amino acid adducts.

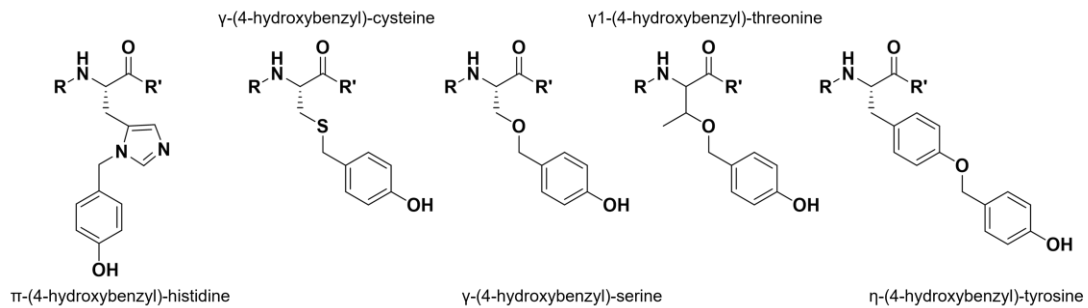


Figure 5.6. 4-Hydroxybenzyl adducted side chains identified in global proteomics analysis and confirmed via targeted mass spectrometry, highlighted in red. N-terminal valines are highlighted in green.

Subunit α VLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFP
 TTKTYFPHFDLSHGSAQVKGHGKKVADALTNVAHV
 DMPNALSALSDDLHAHKLRVDPVNFKLLSHCLLVTLAA
 HLPAEFTPAVHASLDFLASVSTVLTSKYR

Subunit β VHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPW
 TQRFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDG
 LAHLDNLLKGTFA~~TL~~SELHCDKLHVDPENFRLLGNVLV
 CVLAH~~H~~FGKEFTPPVQAAYQKVVAGVANALAHKYH

As expected, control samples of human blood not treated with para-QM contained background levels of 4-hydroxybenzyl adducts (**Table 5.4**). Adduct amounts in untreated samples ranged from 3.2×10^{-4} % modification of β His143 to 0.64 % modification of β Cys112. In our previous study that examined blood samples from smokers and nonsmokers (N = 6 per group), N-terminal valine 4-hydroxybenzyl adducts were present at levels of 380 ± 160 pmol/g of hemoglobin, which corresponds to approximately 3.55×10^{-4} % to 8.71×10^{-4} % modification of the N-terminal valine³⁴³. These results suggested that at least some of the internal amino acid side chains of hemoglobin were more reactive towards para-QM than the N-terminal valines. Although the number of 4-hydroxybenzyl adducts at β Ser44 was estimated to be as high as 1.2×10^{-1} %, adduct numbers did not increase following incubation with para-QM and were highly variable between the replicates (results not shown). Therefore 4-hydroxybenzyl adducts at β Ser44 may not be stable during tryptic digestion and other sample processing steps.

Table 5.4. Hemoglobin amino acid residues showing 4-hydroxybenzyl adduct formation before and after treatment with para-QM, 4-hydroxybenzaldehyde, and UV irradiation. a) Levels of adduct formation relative to control samples following incubation with potential 4-hydroxybenzyl adduct sources. Red highlighted values indicate increased adduct formation following treatment, green highlighted values indicate a loss of adduct following treatment. Residues annotated with an asterisk showed significant statistical differences to controls. The N-terminal peptides were not included in targeted experiments. b) Levels of endogenous adducts in the control hemoglobin samples.

a)

Subunit	Residue	% Adducted Side chain			
		para-QM (5-fold excess)	4-HBA (5-fold excess)	UV	UV + Y
α	His20	12.9 *	1.0	8.8E-01	1.3 *
	Tyr24	10.6 *	5.8E-01	1.2E-02	9.8E-01
	Tyr42	22.6	2.7	2.4E-02	5.8E-02
	His45	24.0	2.5	4.1E-02	3.6E-02
	Ser44	< 1E-04	2.9	2.1	2.2
	Ser72	24.6 *	< 1E-04	< 1E-04	< 1E-04
	Thr84	16.6 *	2.6	6.4E-02	5.2E-02
	Thr87	16.4 *	< 1E-04	< 1E-04	< 1E-04
β	Ser89	5.6	< 1E-04	1.5E-01	5.1E-02 *
	His92	27.1 *	1.3	< 1E-04	< 1E-04
	Cys93	32.8 *	< 1E-04	< 1E-04	< 1E-04
	Cys112	13.3 *	< 1E-04	< 1E-04	< 1E-04
	Thr123	1.8 *	< 1E-04	< 1E-04	< 1E-04
	His143	9.80 *	< 1E-04	< 1E-04	< 1E-04

* $p < 0.05$

b)

Subunit	Residue	Endogenous Adducted Side chains (%)	
		Mean	Std. Dev.
α	His20	8.7E-02	8.8E-02
	Tyr24	2.2E-01	3.0E-01
	Tyr42	1.8E-01	2.1E-01
	His45	1.2E-01	1.2E-01
β	Ser44	1.2E-01	2.7E-01
	Ser72	8.6E-02	8.0E-02
	Thr84	7.3E-03	8.8E-03
	Thr87	6.1E-03	8.5E-03
	Ser89	1.2E-02	1.5E-02
	His92	1.5E-02	1.7E-02
	Cys93	9.4E-03	1.1E-02
	Cys112	6.4E-01	9.2E-01
	Thr123	1.1E-03	1.4E-03
His143	3.2E-04	7.2E-04	

5.3.3 Differential reactivity of the hemoglobin side chains towards para-QM

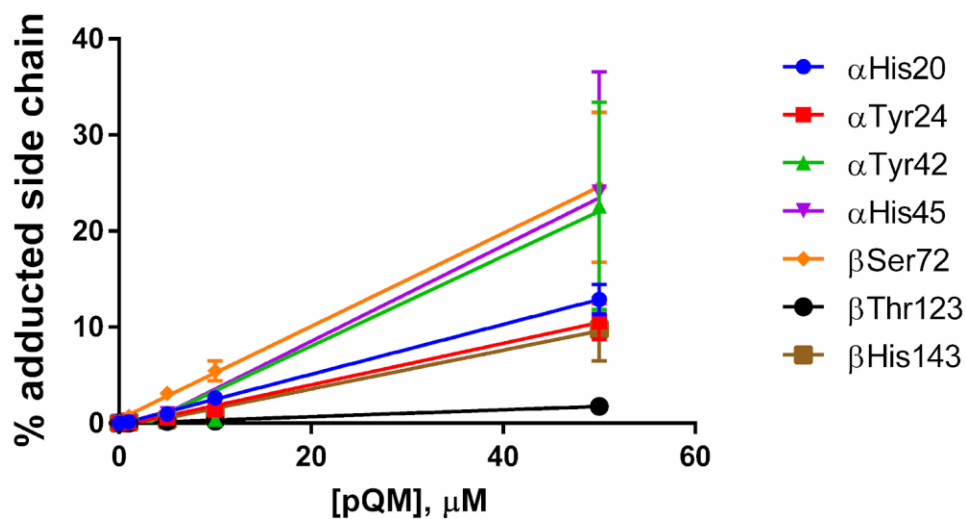
To compare the relative reactivity of various hemoglobin amino acid side chains toward para-QM, hemoglobin was treated with increasing amounts of para-QM, and adducted peptides were analyzed using the targeted PRM assay. Percent adduct formation at each site was quantified directly from HPLC-ESI-MS/MS peak areas corresponding to 4-hydroxybenzyl-modified and intact peptides as described in Materials and Methods.

Hemoglobin amino acid side chains showed concentration dependent formation of 4-hydroxybenzyl adduct and exhibited variable degrees of reactivity towards para-QM, with some residues showing a linear relationship between adduct formation and amount of para-QM added (**Figure 5.7a**) and other residues showing saturation of adduct levels at lower amounts of added para-QM (**Figure 5.7b**). On the alpha subunit of hemoglobin, the α His45 and α Tyr42 residues were less reactive than α His20 and α Tyr24 at lower exposure levels (0.1-0.5-fold excess para-QM), but ultimately had higher levels of adduct formation after treatment with a 5-fold excess of para-QM (23- 24%) (**Figure 5.7b**). These results suggest the contribution of both kinetic and thermodynamic factors to 4-hydroxybenzyl adduct yield at individual sites within the protein.

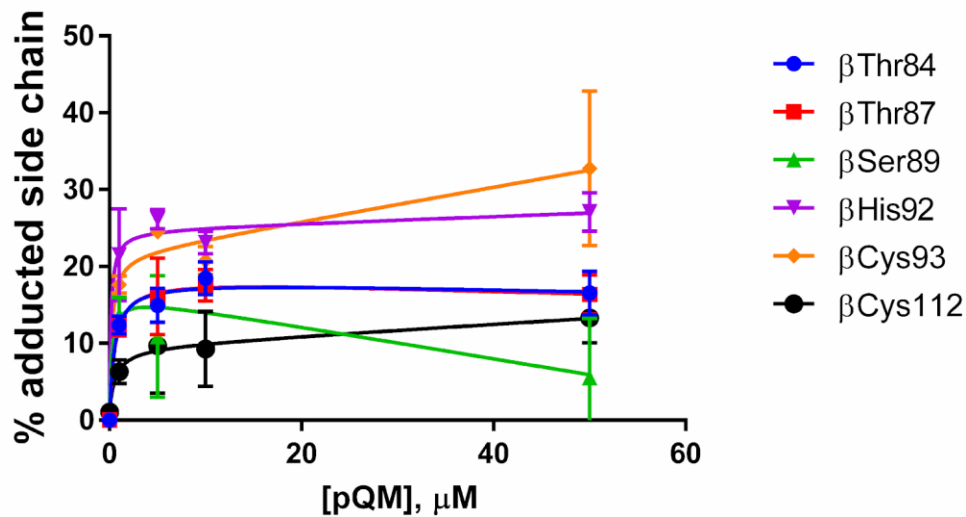
Figure 5.7. Side chain adduct occupancy of residues in hemoglobin with titration of increasing amounts of para-QM. Adduct formation on side chain residues follows either a

a) linear or b) saturation relationship with increasing addition of para-QM.

a)



b)



5.3.4 4-Hydroxybenzyl adducts form at side chains characterized by low pKa values and/or high degrees of solvent accessibility

Visualizing para-QM modified side chains within the three-dimensional structure of hemoglobin reveals that the 4-hydroxybenzylated sites preferentially occupy solvent exposed regions within the alpha helices of the protein (**Figure 5.8**). Relative reactivity of these sites correlates with the calculated relative solvent accessibilities of the side chain residues (**Figures 5.9a and 5.9b**), where 62.5% of highlighted residues show 50% solvent accessibility or greater. Interestingly, less than 50% of the dissociable side chains have favorable pKa values at physiological pH, suggesting that steric accessibility of these residues plays a more important role in their reactivity towards para-QM. In considering the protein structure further, it is apparent that these reactive side chains occur in discrete clusters around the protein molecule (**Figure 5.8**). Each cluster contains at least one residue with high affinity towards para-QM. For example, several amino acid side chains showing increased reactivity towards para-QM were found in and around β Cys93, the most reactive residue in the protein beta subunit (**Table 5.4a, Figure 5.8d**). This provides preliminary evidence for 4-hydroxybenzyl adduct migration along the protein as previously reported for ortho-QM adducts on DNA³⁷⁷.

Figure 5.8. a) NMR structure of human hemoglobin³⁷³ (PDB ID: 2h35) with 4-hydroxybenzyl adducted side chains shown in red. Alpha and beta subunits are gray, with the N-terminal valine residues shown in green and heme molecules in blue. Clusters of adduct sites in alpha subunits are shown in subfigures b) and c), clusters of adduct sites in beta subunits are shown in subfigures d) and e).

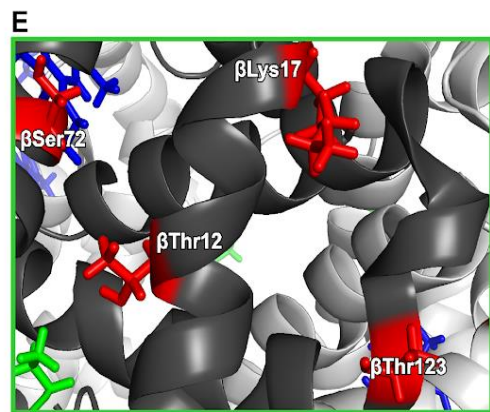
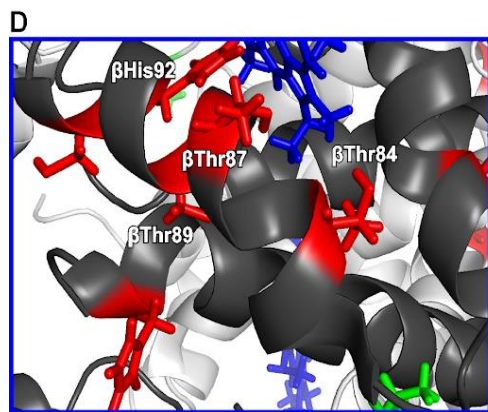
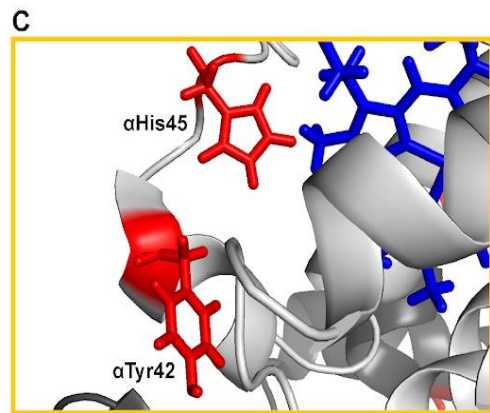
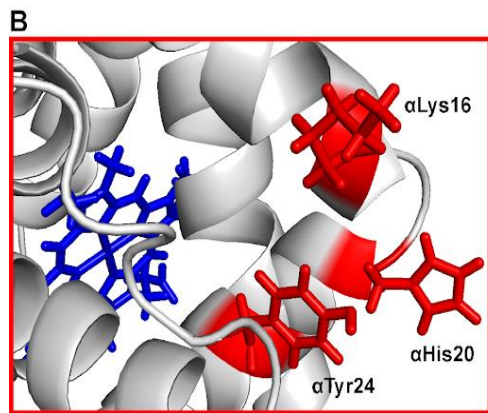
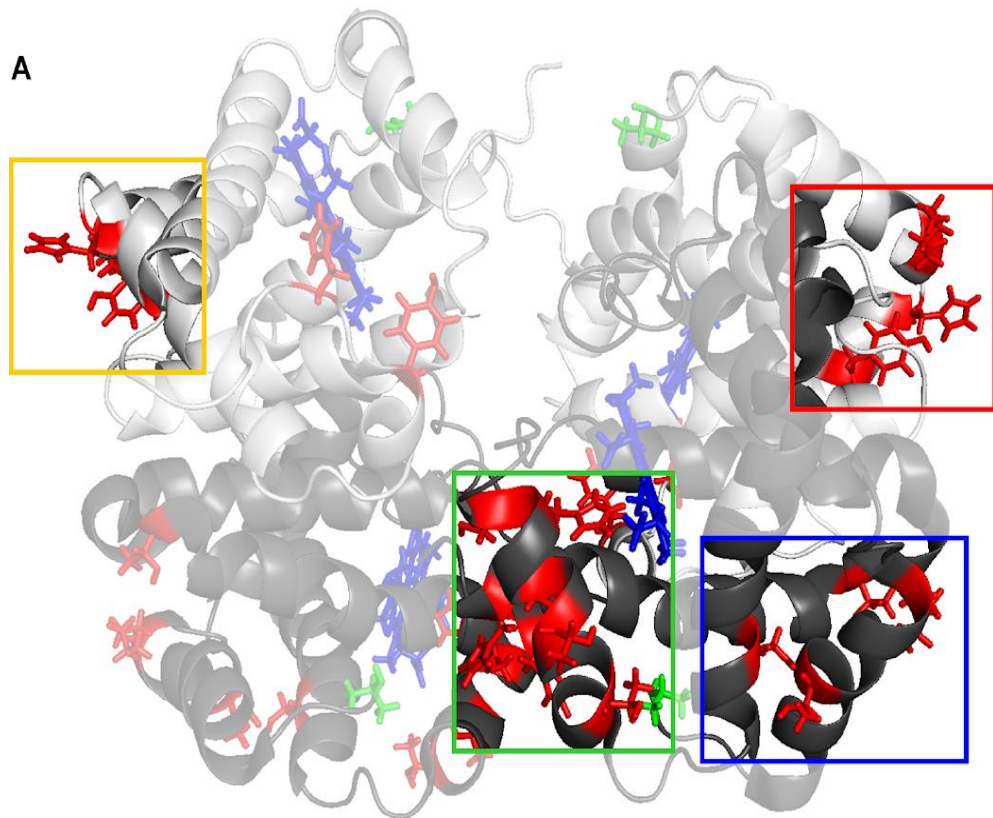
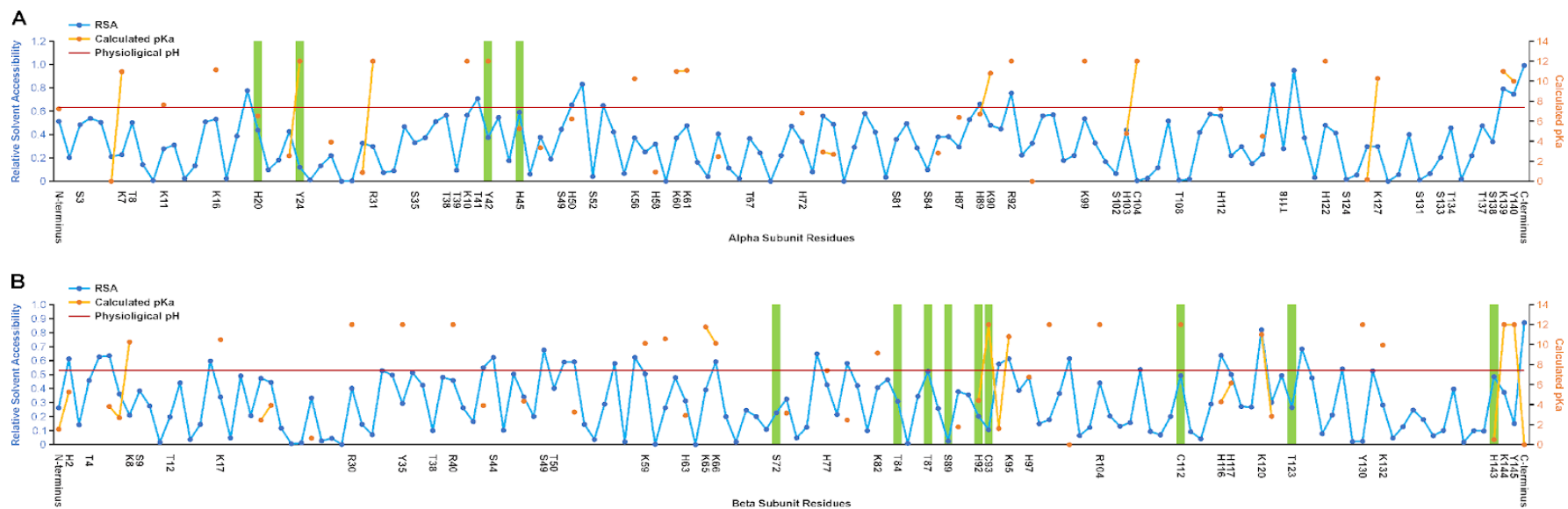


Figure 5.9. Biophysical properties of N-terminal valine and amino acid side chains within the a) alpha subunit and b) beta subunit. Labeled residues correspond to nucleophilic side chains. Relative solvent accessibilities are scaled relative to the theoretical values⁵². pKa values were not calculated for non-ionizable functional groups. Physiological pH is designated by the horizontal red line.

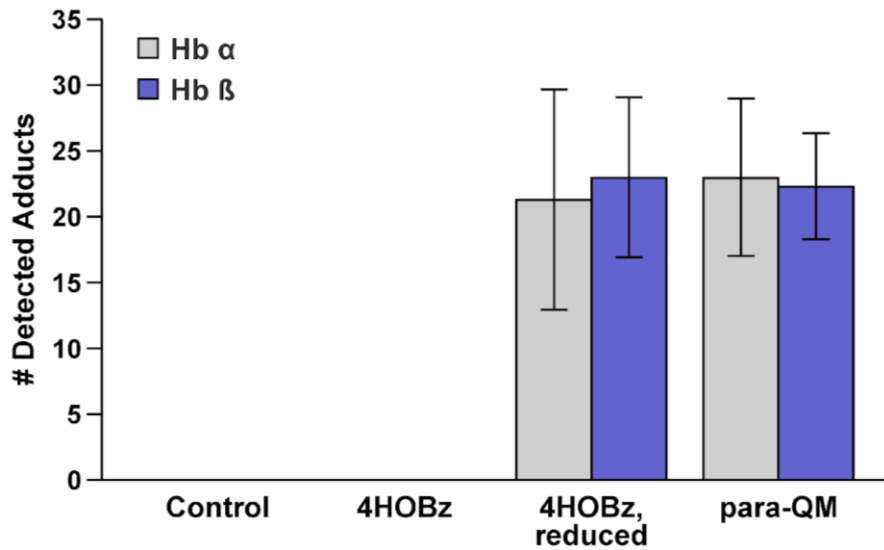


5.3.5 4-Hydroxybenzyl Adduct Formation in Hemoglobin Treated with 4-Hydroxybenzaldehyde

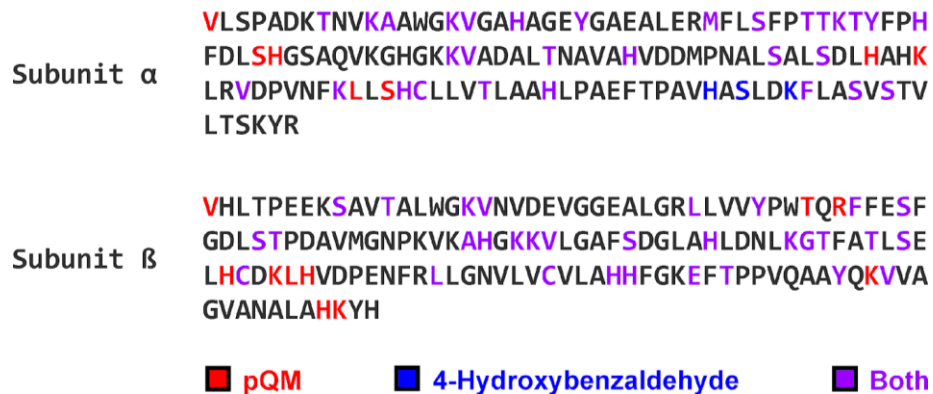
4-Hydroxybenzaldehyde (4-HBA) is found in the common food flavoring agent vanillin and is present in many common foods³⁷⁸. To determine whether 4-HBA can potentially serve as a source of 4-hydroxybenzyl adducts, human hemoglobin was reacted with five-fold molar excess of 4-HBA. Global proteomic analysis of tryptic digests revealed no adducts (results not shown). However, 4-hydroxybenzyl adduct formation was observed in both hemoglobin subunits when the 4-hydroxybenzaldehyde treated hemoglobin was subjected to reductive amination with sodium cyanoborohydride (**Figure 5.2b**). The sixty-four sites observed in 4-HBA/NaCNBH₃ treated hemoglobin using global proteomics analysis showed a high degree of overlap with those observed upon incubation with excess para-QM (**Figure 5.10a**). Most of the 4-hydroxybenzylated sites (77%) were shared between the two treatments (**Figure 5.10b**). However, adducts at α His122, α Ser124, α Lys127 were observed exclusively upon 4-hydroxybenzaldehyde treatment and were found near the C-terminus of the hemoglobin alpha subunit.

Figure 5.10. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA.

a)



b)



Peptides corresponding to 4-hydroxybenzaldehyde modification of α His122, α Ser124, and α Lys127 were added to the previously chosen sites in an inclusion list for targeted mass spectrometry analysis of hemoglobin treated with 4-hydroxybenzaldehyde and sodium cyanoborohydride. Targeted analysis of the hemoglobin treated with 4-hydroxybenzaldehyde/sodium cyanoborohydride provided evidence for 4-hydroxybenzyl modifications at α His20, α Tyr24, α Tyr42, α His45, β Ser44, β Thr84, and β His92 residues (**Table 5.4a**). While β Ser44 was not validated in the targeted analysis of para-QM-treated samples, the other six adducts associated with exposure to 4-hydroxybenzaldehyde/sodium cyanoborohydride were seen in both targeted experiments. Overall, 4-hydroxybenzyl adduct levels were higher in 4-hydroxybenzaldehyde treated samples as compared to controls, but the results did not reach statistical significance due to high variability between replicates (**Table 5.4a**). The efficiency of 4-hydroxybenzyl adduct formation in 4-hydroxybenzaldehyde treated hemoglobin was far lower than that seen in para-QM-treated samples (**Table 5.4a**), suggesting that 4-hydroxybenzaldehyde is less reactive towards nucleophilic side chains of hemoglobin as compared to para-QM.

5.3.6 4-Hydroxybenzyl adduct formation in hemoglobin exposed to UV light

A recent report documented the formation of para-QM upon UV irradiation of tyrosine.⁴¹ In this mechanism, para-QM is released from the side chain of Tyr via a free radical mechanism to leave behind a glycine residue (**Figure 5.2**). To determine whether UV light can lead to the formation of 4-hydroxybenzyl adducts on hemoglobin, human blood was exposed to ultraviolet radiation C (254 nm) with and without the addition of external L-tyrosine. Global proteomics experiments revealed 32 amino acid side chains

potentially carrying a 4-hydroxybenzyl modification following exposure to ultraviolet radiation, these adducts formed with or without the addition of tyrosine (**Figure 5.11a**). Of these potential 4-hydroxybenzylated sites, 84% were also detected in the global analysis of hemoglobin exposed to para-QM. Three of the five new adducts (α His122, α Ser124, α Lys127) detected in this experiment were also seen in the global analysis of the hemoglobin treated with 4-hydroxybenzaldehyde (**Table 5.4a, Figure 5.10b**).

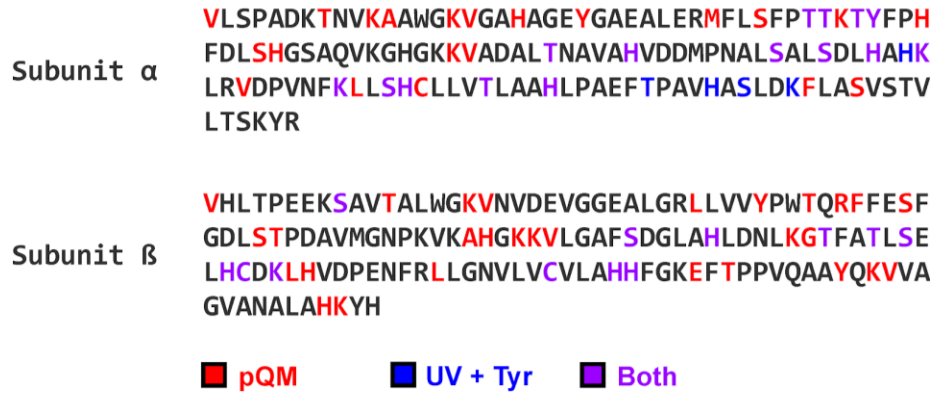
To validate our initial results, these 32 peptides were combined with the original list of peptides in **Table 5.3** to generate an inclusion list for validation via targeted mass spectrometry. Targeted MS experiments confirmed that UV irradiation induced 4-hydroxybenzyl adducts at multiple sites across alpha and beta subunits of hemoglobin including α His20, α Tyr24, α Tyr42, α His45, β Ser44, β Thr84, and β Ser89 (**Table 5.4a**). As was the case for 4-hydroxybenzaldehyde, the numbers of 4-hydroxybenzyl adducts in human blood exposed to ultraviolet radiation were much lower than in experiments with para-QM (**Table 5.4a**). α His20 and β Ser89 showed significant increases in abundance relative to their control samples.

Interestingly, the addition of free L-tyrosine did not significantly impact the formation of UV-induced 4-hydroxybenzyl adducts in hemoglobin, as the adduct levels with and without the addition of extra tyrosine were statistically equivalent (**Figure 5.11b**). These results suggest that tyrosine residues present within the protein, rather than free tyrosine, are the source of the observed 4-hydroxybenzyl adducts. To test this hypothesis, global proteomics data were searched against hemoglobin FASTA files in which tyrosines were replaced with glycines. As mentioned above, the release of para-QM from a side chain results in the formation of glycine at the same site (**Figure 5.2c**). In support of this

hypothesis, MS-based proteomics analyses revealed peptides in hemoglobin containing potential glycine “scars” in place of tyrosine (**Figure 5.12**). Two of the tyrosine residues that show glycine substitution are α Tyr24 and α Tyr42. Both Tyr residues had decreased amounts of 4-hydroxybenzyl adduct formation in UV-irradiated samples relative to their control samples and exhibited a slight increase upon the addition of external tyrosine to the reaction (**Table 5.4a**). Overall, this is consistent with a loss of tyrosine at these sites in hemoglobin alpha subunits and a local release of para-QM that is available for binding to neighboring amino acid residues. This effect is also seen at α His45, where UV exposure results in a decrease in adduct formation relative to the control samples. Adduct formation at α His45 stemming from para-QM release from α Tyr42 could be undetected in our analyses due to incomplete b- and y- ion series in the MS/MS spectrum. Of the residues identified as having increased 4-hydroxybenzyl adduct formation following UV exposure, β Ser44 and β Thr89 are in proximity to a tyrosine (11.4 and 7.2 angstroms away, respectively) (**Figure 5.13**). The remaining residues were found to be on the exterior of the protein. Further detailed investigations are needed to examine UV-mediated release of para-QM as a possible source of para-hydroxybenzyl adducts in hemoglobin.

Figure 5.11. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA.

a)



b)

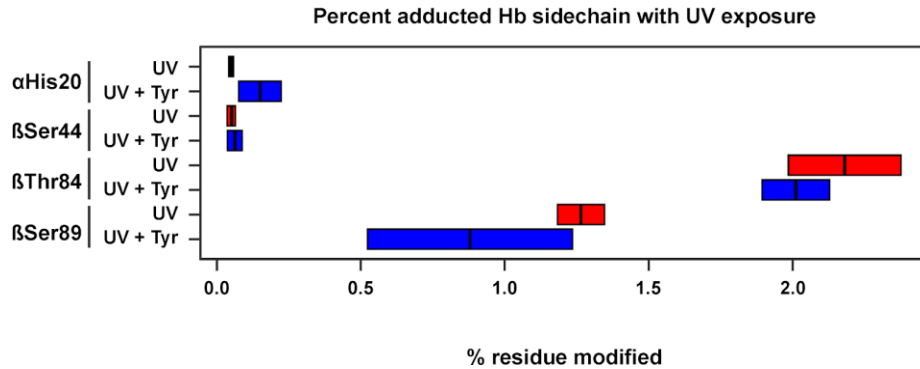
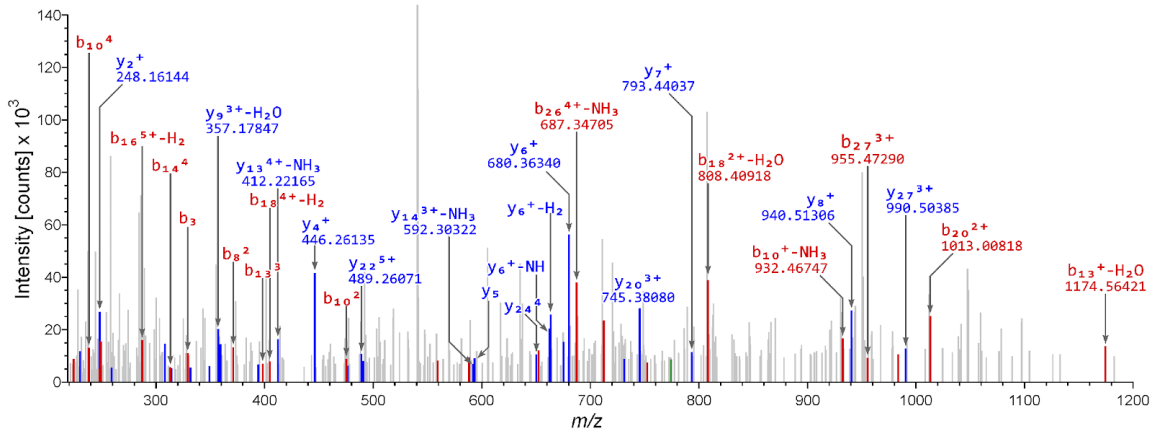


Figure 5.12. Characterization of side chain adducts produced with incubation with 4-hydroxybenzaldehyde. (4-HBA) a) Numbers of putative side chain adducts identified with incubation of hemoglobin in 4-hydroxybenzaldehyde, with and without reduction as compared to incubation in excess para-QM. b) 4-hydroxybenzyl adduct sites identified in global proteomics analysis following incubation with para-QM or 4-HBA.

a)



b)

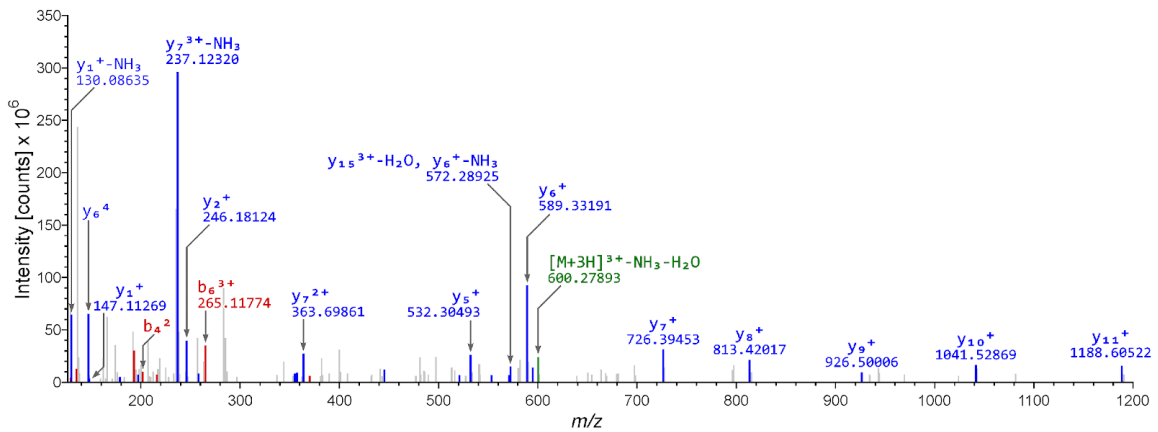
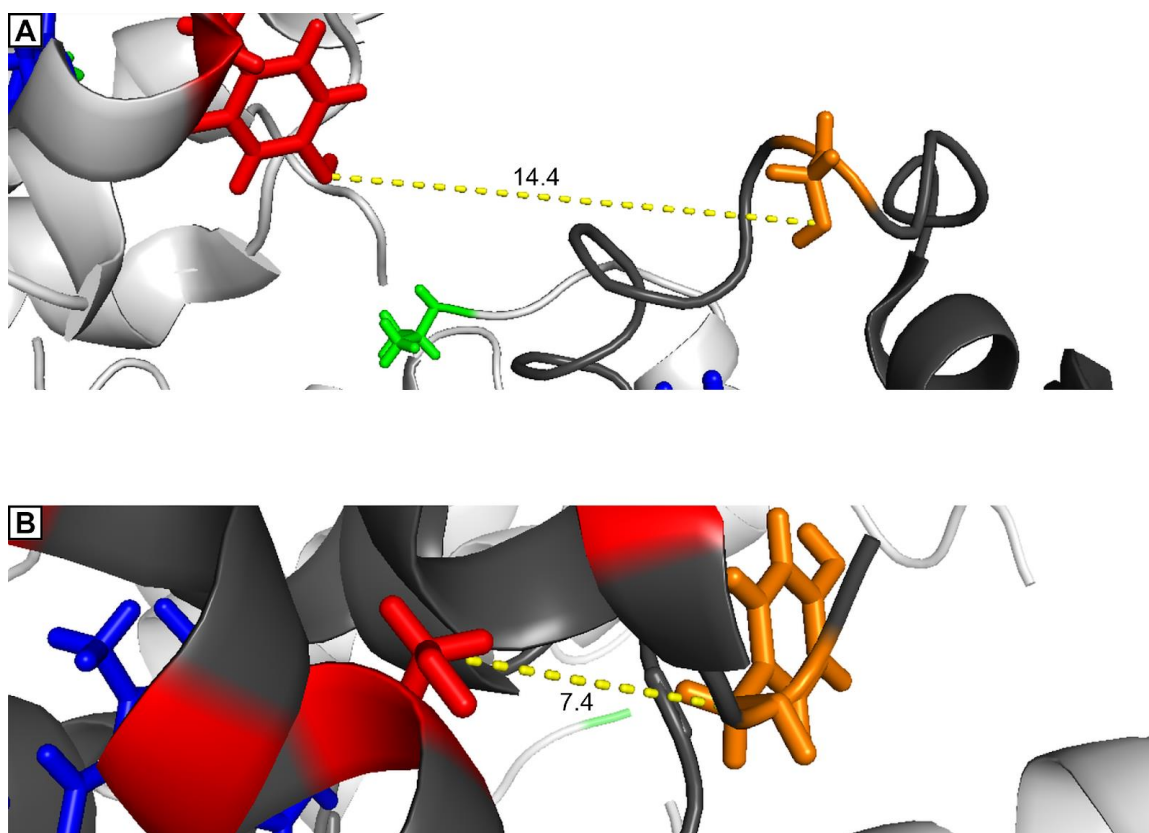


Figure 5.13. Distances between nearest tyrosine and a) α Ser44 and b) β Thr84 in following exposure to UV radiation.

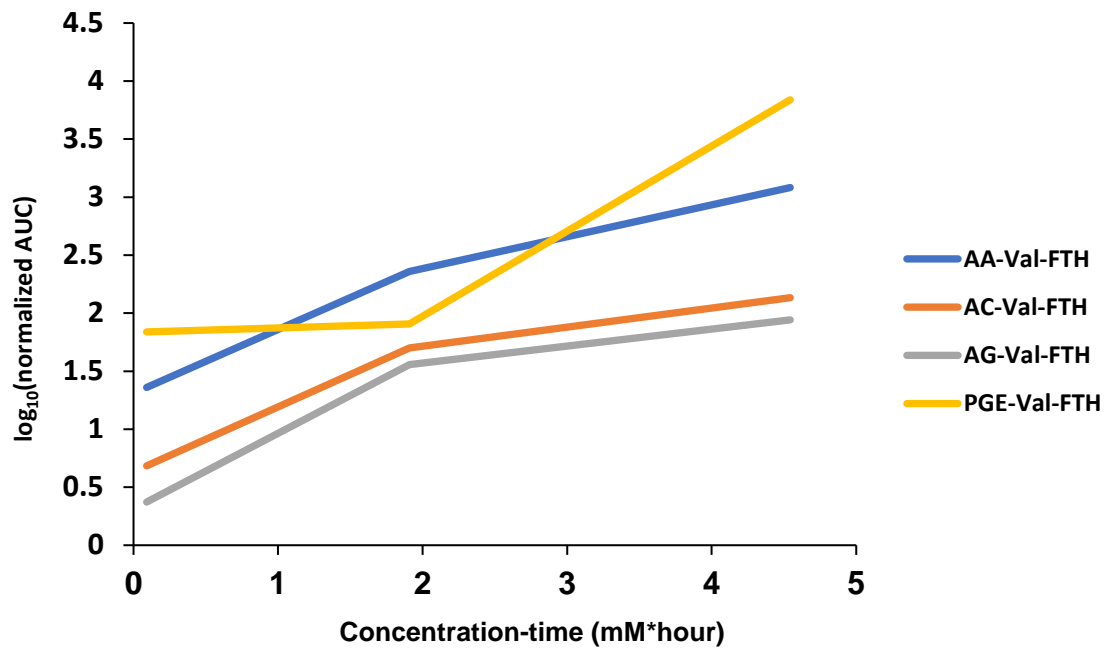


5.3.7 FIRE analysis of samples reveals linear dose-responses to most electrophiles

Having established the potential utility of bottom-up proteomics in identifying adducts at the N-termini and side chains of hemoglobin, we were eager to compare this approach directly against the FIRE methodology to determine which method, if any, would be better suited for future studies of exposed model organisms and patient samples. To this end, we employed a panel of six electrophiles (**Figure 5.3**) which we incubated individually with donor blood for various concentrations and times (**Supplemental Table 5.1**). Three of the panel members- acrylamide, acrylic acid, and glycidic acid- were chosen due to being previously observed forming adducts in hemoglobin^{379,341}; the remaining electrophiles 2-MGN, DNCB, and PGE were selected due to their previous use in studies of contact allergens³⁸⁰.

Following their incubation and cleanup in the FIRE methodology, blood samples exposed to various electrophiles were analyzed via LC-MS, where signals from each of the adducted FTH-valines were normalized against the D₇-acrylamide FTH-valine internal standard and plotted as a function of concentration-time, or the dosage of the applied electrophile multiplied by the length of time of exposure to the electrophile (**Figure 5.14**).

Figure 5.14. Dose-response curves of blood samples incubated with electrophiles and analyzed via the FIRE method.



In looking at the dose responses of the electrophile panel, four of the adducts (PGE, acrylamide, acrylic acid, and glycidic acid) demonstrated linear increases in signal intensity with greater exposure to electrophiles. Among these, PGE and acrylamide represent the most reactive electrophiles according to this method, having the greatest signals at 5 mM*hour. However, we see with FIRE very little signal in blood samples exposed to 2-MGN, only seeing signal at very high concentration*times (i.e., 105 mM*hour). N-terminal Hb adducts were not observed in any blood samples treated with DNCB regardless of concentration or incubation time. Since DNCB is a highly reactive molecule used in medical applications³⁸¹, this could be explained by the large size and aromaticity of DNCB decreasing the efficiency of the loss of the adducted N-terminal valine via FIRE.

5.3.8 Bottom-up proteomics allows for the detection of a greater variety of Hb adducts as compared to FIRE

The same exposed blood samples were analyzed via bottom-up proteomics. Proteome Discoverer software was able to detect 2-MGN adducts reliably at all concentration-times examined. In addition, N-terminal adducts of DNCB were detected via Proteome Discoverer and validated by manual inspection of the MS/MS spectra (**Figure 5.15**), demonstrating the ability of this method to detecting aromatic adducts beyond what FIRE can do. An additional level of information provided by bottom-up proteomics is the ability to assign N-terminal adducts to alpha and beta chains within hemoglobin (**Table 5.5**). In contrast, FIRE is unable to discern which chain within the hemoglobin these Val adducts are coming from.

Figure 5.15. MS/MS spectrum of N-terminal DNCB adduct.

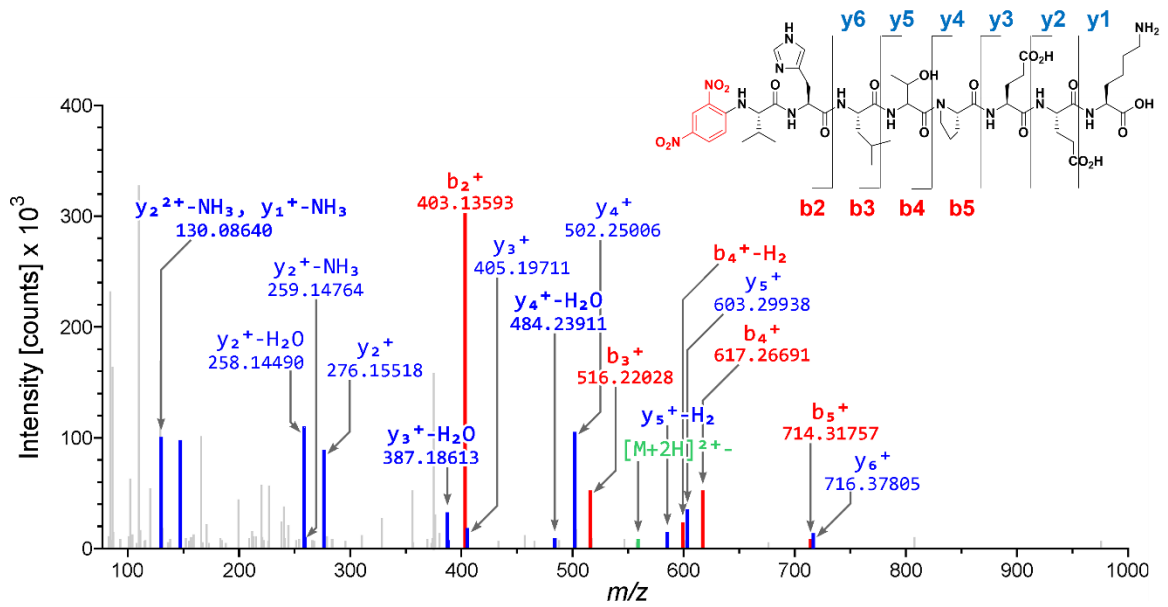
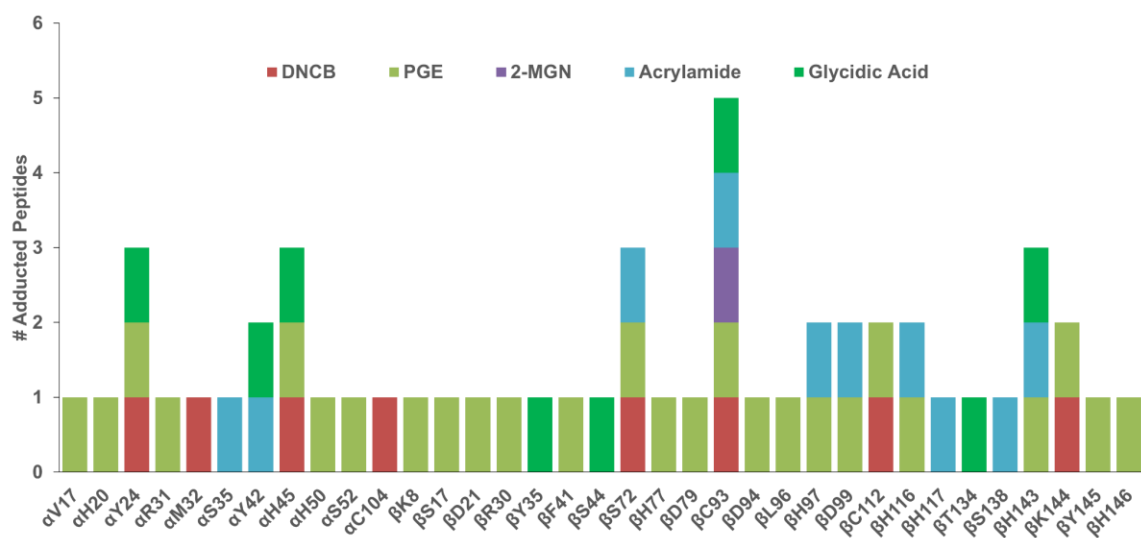


Table 5.5. N-terminal peptides detected by bottom-up proteomics following hemoglobin exposure to various electrophiles. Boxes highlighted in green represent manually validated peptides, red represent peptides that were detected but not validated, and grey represents peptides that were not detected.

	N-termini	
	Alpha	Beta
DNCB	Not Validated	Validated
PGE	Validated	Validated
2-MGN	No peptide	Validated
Acrylamide	Validated	Not Validated
Acrylic Acid	Validated	Validated
Glycidic Acid	Validated	Validated

The other advantage conferred by bottom-up proteomics, as established previously, is the ability to survey the nucleophilic sites on the side chains of the amino acids. Adducts from nearly all the electrophiles in our test panel were observed at several different nucleophilic side chains (**Figure 5.16**), including many of the side chains observed to form 4-hydroxybenzyl adducts in previous section (e.g., α Y24, α H45, β S72, β C93). The β C93 side chain adducts were observed most reliably, further demonstrating its reactivity and utility as a site for study in exposomics experiments.

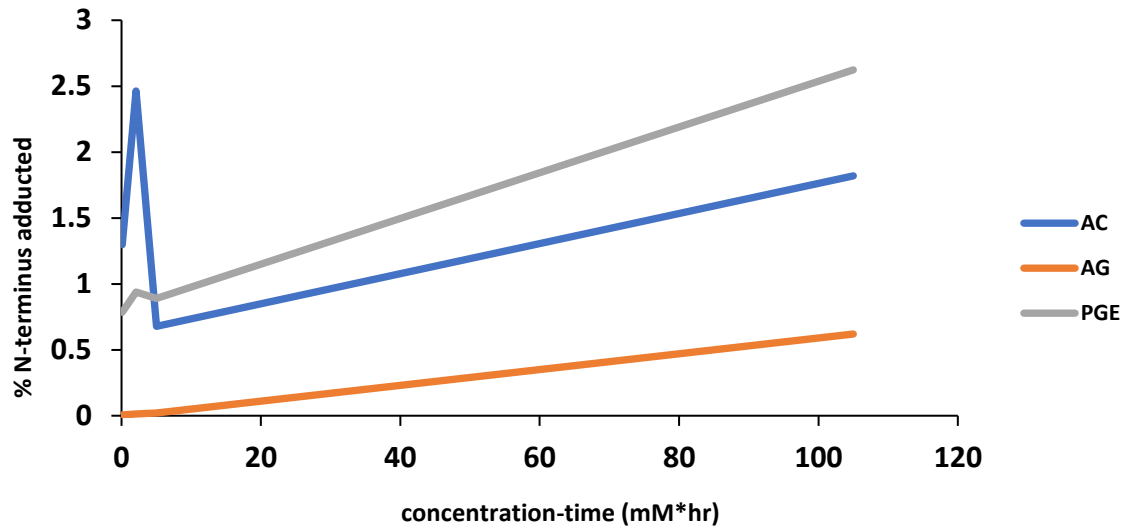
Figure 5.16. Adduct formation at nucleophilic side chains of human Hb following exposures of human blood to electrophiles.



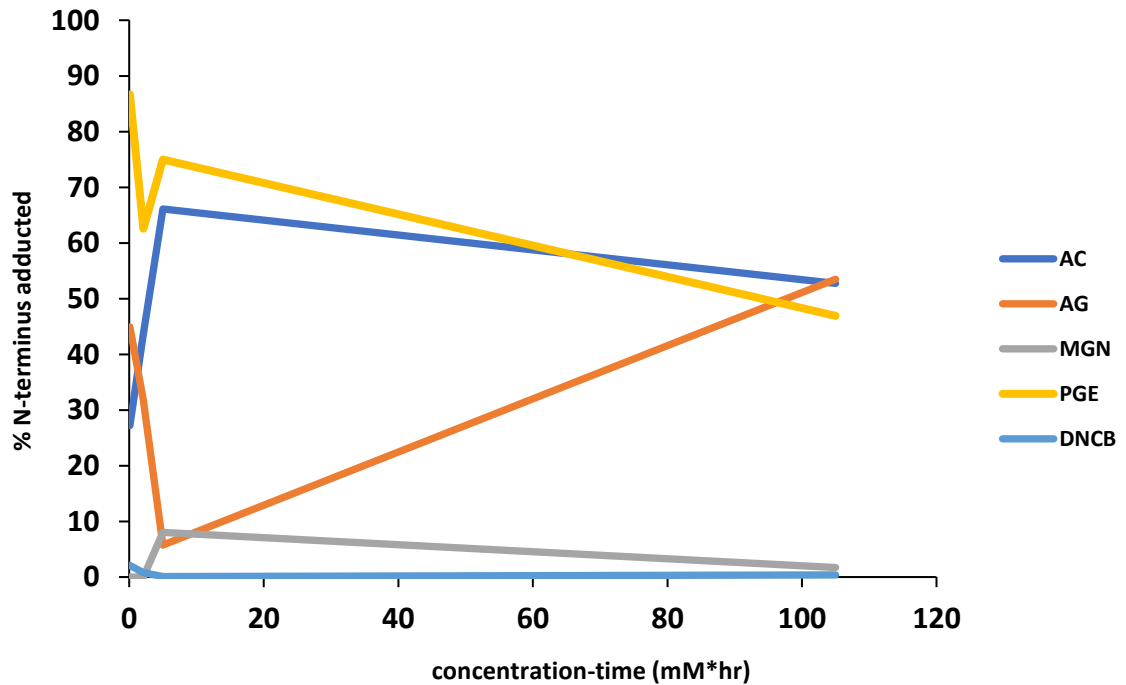
Using an isotopically labeled internal standard of acrylamide-Val-D₇-FTH, we were able to readily quantitate several of N-terminal adducts using the FIRE methodologies. We therefore sought to recapitulate the linear dose-response of adduct formation using a targeted mass spectrometry methodology for unmodified and modified N-terminal peptides. To normalize these data in the absence of internal standards, we divided the integrated signals of the modified peptides by the total signals of all peptides and expressed this number as a percentage. In looking at the N-terminal alpha chain, we see a slightly linear progression of increased adduct formation with increased incubation concentration-time (**Figure 17a**). By contrast, the N-terminus of the beta chain showed marked decreases in the percent of the peptide showing adduct formation, indicating either a loss of the adduct over time or a loss of signal of the unmodified N-terminal beta chain, indicating the need for an added internal standard for reliable quantitation (**Figure 17b**).

Figure 5.17. Dose-response curves derived from bottom-up proteomics. Adducted peptides were normalized to non-modified peptides at the a) alpha chain N-terminus and b) beta chain N-terminus

a)



b)



5.4 Discussion

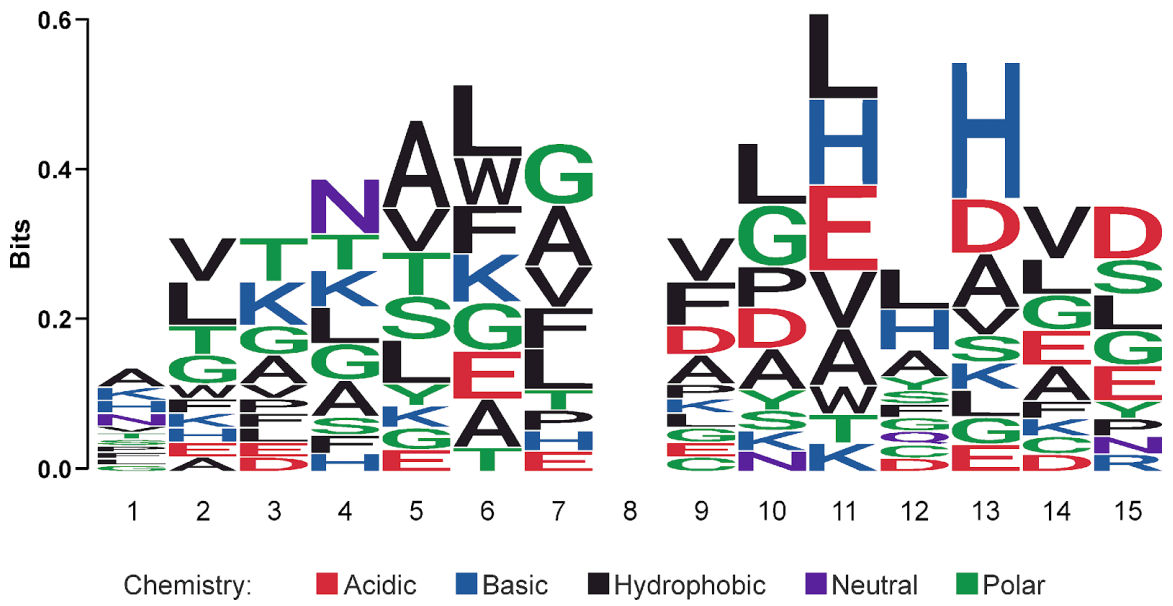
While covalent adducts formed at the N-terminal valine of hemoglobin are widely used for human exposome studies^{342,344} internal nucleophilic amino acid side chains are also reactive towards electrophiles^{350,359}. The first part of this study examined the formation of 4-hydroxybenzyl adducts at internal residues within human hemoglobin. We proposed that examination of the patterns of hemoglobin side chains susceptible to 4-hydroxybenzyl adduct formation may help identify the possible sources of such adducts (para-QM, para-benzaldehyde, UV radiation) in humans because these electrophiles form protein adducts by different mechanisms and may exhibit distinct site specificity within the protein^{382,383}.

Our MS based proteomics analyses revealed the formation of 4-hydroxybenzyl adducts at several side chains in hemoglobin following treatment with para-QM (**Figure 5.6**). Four of the preferentially adducted amino acid side chains are located within the alpha subunit of the protein (α His20, α Tyr24, α Tyr42, α His45), with additional ten adducts found on the beta subunit (β Ser44, β Ser72, β Thr84, β Thr87, β Ser89, β His92, β Cys93, β Cys112, β Thr123, β His143). Previous studies reported the preferential formation of hemoglobin adducts on cysteine residues of hemoglobin (Cys93 and Cys112) by other electrophiles^{382,384,385}. Although our experiments confirmed the formation of 4-hydroxybenzyl adducts at these two sites, they were not the only reactive side chains on the hemoglobin molecule. We found that 4-hydroxybenzyl adducts readily formed at histidine, serine, and tyrosine side chains of hemoglobin following treatment with para-QM (**Table 5.4a and Figure 5.2b**). Hemoglobin adducts at histidine and lysine side chains

have been previously reported e.g., upon reactions with 2-octenal and methylene diphenyl diisocyanate^{361,386}.

To help explain the observed site specificity for adduct formation at specific residues within the protein, local environment of the preferentially modified hemoglobin side chains was considered. Interrogating the consensus sequences³⁶⁹ of the adducted side chains within 7 residues of the reaction sites shows a very diverse set surrounding side chains which are unlikely to contribute to the increased nucleophilicity of the reactive side chains (**Figure 5.18**). This is reflected in the highly variable pKa of the adducted amino acid side chains (**Figures 5.9a and 5.9b**), with many residues showing high pKa values relative to the physiological pH (7.4). At the same time, many of the preferentially modified residues show increased solvent accessibility, which could increase their reactivity towards electrophiles³⁸⁷. In addition, the adducted residues are clustered into islands located in solvent exposed regions of the protein, with each cluster containing at least one side chain residue that is particularly reactive towards para-QM (e.g., α His45, β His92/ β Cys93) (**Figures 5.8b-5.8e**). While there is some evidence for migration of QM-nucleobase adducts along DNA chains³⁷⁷, it is likely that the clusters of adducts are the results of pre-association of the para-QM precursors with hemoglobin at especially accessible regions of the protein³⁸⁸.

Figure 5.18. Consensus sequence of side chains showing adduct formation (empty position 8).



The presence of 4-hydroxybenzyl adducted side chains in untreated hemoglobin demonstrates potential formation of these adducts at physiological conditions (**Table 5.4b**). Our prime motivation for treating hemoglobin with electrophiles was to determine which (if any) of these could be a source of 4-hydroxybenzyl side chain adducts in vivo.

Para-QM is an extremely reactive electrophilic species that is unlikely to survive long under physiological conditions^{389,390}. Therefore, para-QM itself is unlikely the source of 4-hydroxybenzyl adducts in humans unless released in situ from a suitable precursor. Therefore, we investigated other potential sources of 4-hydroxybenzyl adducts in humans. 4-Hydroxybenzaldehyde (4-HBA) was considered as a potential candidate due to its relative stability in an aqueous environment as compared to QMs and its ability to induce 4-hydroxybenzyl-protein adducts under reducing conditions^{346,391}. 4-HBA compound is found in many natural sources, most notably in vanilla³⁷⁸, a flavoring nearly ubiquitous in the modern world. 4-Hydroxybenzaldehyde is a major constituent of the aromatic compounds in vanilla extract³⁹² and is created during the curing of the vanilla pods. The addition of 4-hydroxybenzaldehyde to hemoglobin appears to result in the formation of 4-hydroxybenzyl adducts only following a subsequent reduction step (**Figure 5.2b**), implying that a reducing environment is necessary for adduct formation. This agrees with our previous work showing the formation of the 4-hydroxybenzyl adduct to valine after incubation of valine with 4-hydroxybenzaldehyde and a reducing agent³⁴⁵.

Exposure to hemoglobin to excess 4-hydroxybenzaldehyde and subsequent reduction resulted in a small number of side chain adducts relative to those formed upon exposure to para-QM (**Table 5.5a**), including histidine adducts formed through Schiff base formation and reduction (**Figure 5.6**) as well as adduct formation on the hydroxyl groups

of serine, threonine, and tyrosine. We speculate that this may occur through an acid-catalyzed loss of water with subsequent reduction (**Figure 5.2b**), though further work is needed to explore the nature of this adduct. While the levels of 4-hydroxybenzaldehyde-hemoglobin adducts increased in treated samples relative to controls, these differences were not statistically significant.

Another possible route for 4-hydroxybenzyl adduct formation via para-QM involves in situ formation of para-QM in blood via reactions of free tyrosine or tyrosine side chains of proteins with ultraviolet radiation.⁴¹ However, global proteomics analysis of hemoglobin peptides obtained upon UV irradiation of human blood in the presence or the absence of free tyrosine showed relatively few reproducible sites corresponding with those seen in the endogenous samples (**Table 5.4a**). Future studies should explore adduct formation via oxidation of excess p-cresol in the liver to para-QM via p450 or peroxidase enzymes³⁹³. Recent studies revealed a pathway by which gut microbiota high in Clostridia can metabolize excess L-tyrosine in the gut into p-cresol³⁹⁴, which could give rise to p-QM and 4-hydroxybenzyl adducts upon disruption of gut microbiota.

Despite the utility of the FIRE method in detecting adducts at the N-terminal valine of, there are some drawbacks to its use, principally the concern that larger, bulkier electrophilic adducts may prevent efficient removal from hemoglobin by modified Edman degradation and may go undetected in this assay^{395,396}. We therefore sought to compare FIRE and bottom-up proteomics directly and see which method might be better suited for the detection of hemoglobin adducts in humans. To do this, we utilized a panel of six electrophiles consisting of known adduct-forming electrophiles (acrylamide, acrylic acid, and glycidic acid) and known contact allergens (2-MGN, PGE, and DNCB). These were

then incubated individually with samples of donor blood at various times and concentrations (Supplementary Table 5.1) before the exposed samples were divided in two, the hemoglobin extracted via hypotonic lysis of the red blood cells, and FIRE as well as bottom-up proteomics performed on the resulting samples.

Among the two methods, bottom-up proteomics can detect a greater variety of adducts as compared to FIRE. N-terminal Val adducts of all six electrophiles were reliably detected using proteomics as opposed to the five seen with FIRE. In addition, proteomics can detect additional side chain adducts not limited to N-terminal Val, expanding the available sites that can be assayed for adduct formation. In contrast, the FIRE method is limited to the N-termini of the protein. Taken together, our results indicate that bottom-up proteomics has more utility in the detection of novel protein adducts in hemoglobin, though the quantitation of these adducts using bottom-up proteomics in our laboratory is still under development. Normalizing against the unlabeled peptide as in our 4-hydroxybenzyl adduct study, we attempted to demonstrate the linear dose response in our proteomic data as we did for our FIRE data but found that these dose responses were not recapitulated (**Figure 5.17**), suggesting the need for internal standards for normalization of the adduct peptide signal; whether a single isotopically labeled N-terminal peptide will suffice or whether adducted standards for each adduct of interest are needed is currently under investigation.

In summary, this investigation provides compelling evidence for the utility of bottom-up proteomics in the identification of exposure-driven adducts in hemoglobin. Our analyses have provided the first evidence that exposure to para-QM leads to 4-hydroxybenzyl adduct formation on the thiols, hydroxyl groups, amino, and imidazole amino acid side chains of human hemoglobin (**Table 5.4, Figure 5.8**). In addition, 4-

hydroxybenzyl adducts at these side chains are also observed with exposure to 4-hydroxybenzaldehyde under reducing conditions. Our new results extend the knowledge on reactive sites for different electrophiles in hemoglobin and expands the utility of hemoglobin as a record-keeping molecule of the human exposome, allowing for the identification of adducts throughout the protein. We also demonstrate that mapping the adduction sites within hemoglobin via mass spectrometry-based proteomics can help identify possible electrophilic precursors of known N-terminal valine adducts because each electrophile produces a characteristic pattern of amino acid side chain adducts within the protein. Finally, we maintain that bottom-up proteomics may be the superior method for the untargeted detection of novel hemoglobin adducts, as it is able to observe a larger number of adducts formed including larger aromatic adducts.

VI. Summary and Conclusions

Bottom-up proteomics employs mass spectrometry and bioinformatic workflows to characterize and potentially quantitate the proteins in a system of interest. This technology can provide a more accurate snapshot of the phenotype of a given system than next-generation sequencing technologies, as the changes that occur at the transcriptomic level in response to stimuli are not always translated into proteins. However, bottom-up proteomics is too limited in terms of the number of peptides it can reliably detect and confidently identify. Along with instrumental sensitivity, one contributing factor to these limitations is the inability to detect unexpected, non-canonical sequences within the protein sequence database used to identify peptides from tandem mass spectra (MS/MS) obtained.

Some of these limitations can be mitigated through the integration of proteomics analysis with other forms of 'omics data to achieve a more complete molecular picture of the system and related phenotypes of interest. Using customized bioinformatic tools and RNA-Seq data, non-canonical peptides can be detected and validated that would be invisible to normal proteomics. By combining proteomics data directly with other forms of sequencing data, such as quantitative RNA-Seq, changes in the regulation of gene expression and protein translation can be ascertained that normally go undetected. These approaches were used to characterize systems that were potentially undergoing multiple levels of dysregulation in gene product expression in response to infection, inflammation, or exposure.

Chapter II of the thesis characterized the best peptides for detection of SARS-CoV-2 in human samples. Through a series of publicly available mass spectrometry datasets

(**Figure 2.1**) we were able to generate a 639-peptide panel of potential viral peptides. In searching human patient datasets against this panel, we demonstrated that structural proteins of the virus were most readily detected in human patients (**Figure 2.3**). Through automated and manual validation, we ultimately found that four peptides from the nucleocapsid (**Figure 2.8**) - unique to SARS-CoV-2 virus - were most readily detected in human patients and could be excellent candidates for targeted analysis of clinical samples. Workflows developed in this study have been useful for co-infection analysis during COVID-19 infection¹⁸¹, strain detection during pandemic waves³⁹⁷, and other ongoing clinical proteomics studies.

Chapter III of this thesis characterized the quantitative proteomic and proteogenomic changes that occurred in a model of inflammatory bowel disease. Having acquired RNA-Seq data from the proximal colon tissues generated by Dr. Qiyuan Han, we were able to generate an expanded FASTA library containing unique sequences using workflows in the Galaxy-P platform (**Figures 3.2A, 3.2B**). Using this FASTA library in our data analyses allowed us to note the differential abundance of proteins consistent with an inflammatory phenotype (**Figure 3.3A, Table 3.3**). Analysis of our results in a workflow to annotate non-canonical peptides (**Figure 3.2C**) resulted in the putative identification of some 235 non-canonical peptides of which 58 peptides were validated bioinformatically (**Figure 3.5a**) and 39 validated using targeted mass spectrometry (**Table 3.4**). This work demonstrated the importance of bioinformatic and analytical validation in characterizing non-canonical peptides initially identified using untargeted discovery-based proteogenomic workflows.

Chapter IV of this thesis was devoted to the multi-omic characterization of murine Type II cells exposed to LPS and cigarette smoke for variable lengths of time to induce inflammation. As a part of this analysis, we validated the use of C18 spin columns to perform isobaric labelling (**Figure 4.4**) on low amounts of peptide and in-house stage-tip high pH fractionation (**Figure 4.5**) to ensure the maximum amount of proteins were detected in our samples. We were able to determine that LPS exposure resulted in the increased abundance of several proteins involved in the inflammatory process (**Figures 4.6A and 4.6C**); integrated analysis of proteomics data and transcriptomics data showed a general agreement between the responses of most genes, especially those involved in the inflammatory process (**Figures 4.7A, 4.7C**). Finally, we were able to show that significant changes in protein abundance with cigarette smoke exposure only occurred after 10 weeks of exposure (**Figures 4.8B-4.8E**), had significantly greater disjunction with their matching RNA-Seq data (**Figure 4.10**), and showed relatively little in common with LPS exposure (**Figure 4.11A**).

In chapter V we examined the ability of bottom-up proteomics as a detection strategy for untargeted adductomics in hemoglobin. With bottom-up proteomics we were able to validate the formation of N-terminal 4-hydroxybenzyl adducts in hemoglobin (**Figure 5.4A**) which were first proposed by analysis using the FIRE method. We were also able to demonstrate the formation of these adducts at side chains (**Figure 5.6**) and demonstrate 4-quinone methide as a likely source of these adducts (**Table 5.5A**). Finally, we showed that bottom-up proteomics was able to detect more adducts than the FIRE method (**Figure 5.14 and Table 5.5**), suggesting its greater suitability for untargeted adductomics.

In summary, this thesis identified peptides for targeted detection of SARS-CoV-2 (Chapter II), demonstrated quantitative changes to the proteome as well as identified and validated non-canonical peptides present in inflamed proximal colon tissue (Chapter III), characterized the proteomic changes that occur in response to LPS and cigarette smoke exposure and integrated them with other ‘omics data (Chapter IV), and showed the utility of bottom-up proteomics in detecting adducts in hemoglobin (Chapter V). Together these discoveries provide important foundations for the early detection of infectious agents, biomarkers of severe inflammation, and exposure to potentially hazardous substances.

VII. Future Directions

7.1 Targeted detection and absolute quantitation of SARS-CoV-2 peptides in patient samples

Detection of SARS-CoV-2 infection via LC-MS represents a potential alternative to qRT-PCR detection, particularly in instances where RNA samples cannot be adequately preserved before analysis. In Chapter II of this Thesis, we established that four nucleocapsid peptides can be readily detected in patient samples and were unique to SARS-CoV-2 over other coronaviruses. However, as we were working exclusively with publicly available datasets, we have yet to perform reliable quantitative analyses on these targets to conclusively determine their true utility as biomarkers of SARS-CoV-2 infection and whether their level of abundance in human samples can be correlated with disease severity. Therefore, one future avenue of investigation would be to collaborate with clinical laboratories in acquiring SARS-CoV-2 nasopharyngeal swabs and to add isotopically labeled standards of our four peptides at known concentrations, digesting them and performing targeted analysis on these peptides. This will validate that our proposed peptides are the most reliably detected, and allow us to correlate peptide abundance levels with infection status, disease severity, therapeutic efficacy, etc. We also anticipate the use of the workflows that we have developed and peptide targets that we have identified for monitoring of wastewater systems to assess the prevalence of COVID-19 or other pathogens in the population.

7.2 Targeted detection and absolute quantitation of non-canonical peptides in proximal colon tissue at different stages of inflammation and oncogenesis

In Chapter III, we detected and validated the presence of 39 non-canonical peptides from genes across the entire murine genome in infected and control proximal colon tissue. While these experiments were important in establishing the utility of this approach for biomarker discovery and indicated their increased abundance in the tissues isolated from mice undergoing inflammation, some of our results lacked statistical significance. This may be due to a combination of factors, such as possible sample degradation and a small sample size ($n = 3$) for each group. A future study would focus on repeating this experiment with larger numbers of tissues in both groups to improve the statistical power of these measurements and utilize the addition of isotopically labeled internal standard peptides for each of these non-canonical peptides to achieve more accurate quantitation of these peptides in our test subjects. In addition, we are interested in probing other cell types, tissues, and biological samples for the presence of these peptides to aid in their use as potential biomarkers that are readily attainable from plasma, urine, feces, etc. without intrusive sampling of the proximal colons of patients.

7.3 Evaluation of the roles of Pdhx, Psma6, Ruvbl1, and Ywhaq in cell proliferation and smoking-induced lung cancer

In chapter IV, we found that the proteins Pdhx, Psma6, Ruvbl1, and Ywhaq were significantly increased in abundance in Type II cells of mice subjected to 10 weeks of cigarette smoke exposure. These proteins were still increased in abundance after a 4-week period of recovery in clean air and also significantly increased in the lung adenocarcinoma

using data available at CPTAC³⁹⁸. Taken together, this suggest a role for these genes in linking cigarette smoke exposure, inflammation, and oncogenesis. To test this hypothesis, we intend to perform gene knockdowns of these proteins individually using siRNA and assess the impact on their proliferation via MTT assay, in which cell viability is measured through cleavage of MTT (3-(4,5-dimethylthazol-2-yl)-2,5-diphenyl tetrazolium bromide) by dehydrogenases in the mitochondria of live cells and spectrophotometric measurement at 570 nm as a proxy for the number of live cells³⁹⁹. In future studies, cell culture models should be exposed to cigarette smoke, followed by protein-protein crosslinking and affinity purification experiments to determine the interaction networks of these proteins in response to cigarette smoke exposure.

7.4 Automation of untargeted adductomics in Galaxy

In Chapter V of this Thesis, we demonstrated the utility of bottom-up proteomics in detecting hemoglobin adducts. We detected N-terminal Val adducts of hemoglobin in human blood samples exposed to a panel of six electrophiles, demonstrating potential advantages of this approach as compared to the established FIRE method for use in untargeted adductomics. An important caveat of our experiments is that we treated human blood with known electrophiles and could predict the exact structures and masses of Hb adducts to look for in the LC-MS/MS bottom-up analysis. For future untargeted adductomics experiments with bottom-up proteomics, MS² spectra need to be assayed for relevant b- and y- ions to identify modified N-terminal peptides from the alpha and beta chains in hemoglobin. To simplify this, we are now assembling an automated workflow in the Galaxy bioinformatics suite to determine m/z differences between the peptides of

interest, calculate the potential chemical structures of this adduct, and search a database for known instances of this adduct in DNA and proteins. This approach could also be complemented using “open modification” searching algorithms in contemporary sequence database searching programs, which also seek to characterize new modifications of unknown structure and mass.

BIBLIOGRAPHY

- (1) Chifman, J.; Laubenbacher, R.; Torti, S. V. A systems biology approach to iron metabolism. *A Systems Biology Approach to Blood* **2014**, 201-225.
- (2) Kreeger, P. K.; Lauffenburger, D. A. Cancer systems biology: a network modeling perspective. *Carcinogenesis* **2009**, *31* (1), 2-8. DOI: 10.1093/carcin/bgp261 (accessed 12/8/2021).
- (3) Men, A. E.; Wilson, P.; Siemering, K.; Forrest, S. Sanger DNA sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine* **2008**, 1-11.
- (4) Costa, V.; Angelini, C.; De Feis, I.; Ciccodicola, A. Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology* **2010**, 2010.
- (5) Ren, S.; Peng, Z.; Mao, J.-H.; Yu, Y.; Yin, C.; Gao, X.; Cui, Z.; Zhang, J.; Yi, K.; Xu, W. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell research* **2012**, *22* (5), 806-821.
- (6) Beane, J.; Vick, J.; Schembri, F.; Anderlind, C.; Gower, A.; Campbell, J.; Luo, L.; Zhang, X. H.; Xiao, J.; Alekseyev, Y. O. Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer prevention research* **2011**, *4* (6), 803-817.
- (7) Leimena, M. M.; Ramiro-Garcia, J.; Davids, M.; van den Bogert, B.; Smidt, H.; Smid, E. J.; Boekhorst, J.; Zoetendal, E. G.; Schaap, P. J.; Kleerebezem, M. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC genomics* **2013**, *14* (1), 1-14.
- (8) Mager, S.; Schönberger, B.; Ludewig, U. The transcriptome of zinc deficient maize roots and its relationship to DNA methylation loss. *BMC Plant Biology* **2018**, *18* (1), 372. DOI: 10.1186/s12870-018-1603-z.
- (9) Balmer, N. V.; Klima, S.; Rempel, E.; Ivanova, V. N.; Kolde, R.; Weng, M. K.; Meganathan, K.; Henry, M.; Sachinidis, A.; Berthold, M. R. From transient transcriptome responses to disturbed neurodevelopment: role of histone acetylation and methylation as epigenetic switch between reversible and irreversible drug effects. *Archives of toxicology* **2014**, *88* (7), 1451-1468.
- (10) Hall, J.; Taylor, J.; Valentine, H. R.; Irlam, J. J.; Eustace, A.; Hoskin, P.; Miller, C. J.; West, C. M. Enhanced stability of microRNA expression facilitates classification of FFPE tumour samples exhibiting near total mRNA degradation. *British journal of cancer* **2012**, *107* (4), 684-694.
- (11) Valencia-Sanchez, M. A.; Liu, J.; Hannon, G. J.; Parker, R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development* **2006**, *20* (5), 515-524.
- (12) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* **2013**, *113* (4), 2343-2394. DOI: 10.1021/cr3003533 PubMed.

- (13) Hecht, E. S.; Scigelova, M.; Eliuk, S.; Makarov, A. Fundamentals and advances of orbitrap mass spectrometry. *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation* **2006**, 1-40.
- (14) Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wiegand, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & cellular proteomics* **2011**, *10* (9).
- (15) Beck, S.; Michalski, A.; Raether, O.; Lubeck, M.; Kaspar, S.; Goedecke, N.; Baessmann, C.; Hornburg, D.; Meier, F.; Paron, I.; et al. The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics *. *Molecular & Cellular Proteomics* **2015**, *14* (7), 2014-2029. DOI: 10.1074/mcp.M114.047407 (accessed 2021/12/09).
- (16) Welton, J. L.; Khanna, S.; Giles, P. J.; Brennan, P.; Brewis, I. A.; Staffurth, J.; Mason, M. D.; Clayton, A. Proteomics analysis of bladder cancer exosomes. *Molecular & cellular proteomics* **2010**, *9* (6), 1324-1338.
- (17) Van Vliet, K.; Mohamed, M. R.; Zhang, L.; Villa, N. Y.; Werden, S. J.; Liu, J.; McFadden, G. Poxvirus proteomics and virus-host protein interactions. *Microbiology and Molecular Biology Reviews* **2009**, *73* (4), 730-749.
- (18) Rešetar, D.; Martinović, T.; Pavelić, S. K.; Andjelković, U.; Josić, D. Proteomics and Peptidomics as Tools for Detection of Food Contamination by Bacteria. In *Advances in Food Diagnostics*, 2017; pp 97-137.
- (19) Han, M.-J.; Lee, S. Y.; Koh, S.-T.; Noh, S.-G.; Han, W. H. Biotechnological applications of microbial proteomes. *Journal of Biotechnology* **2010**, *145* (4), 341-349. DOI: <https://doi.org/10.1016/j.jbiotec.2009.12.018>.
- (20) Mueller, R. S.; Deneff, V. J.; Kalnejais, L. H.; Suttle, K. B.; Thomas, B. C.; Wilmes, P.; Smith, R. L.; Nordstrom, D. K.; McCleskey, R. B.; Shah, M. B. Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular systems biology* **2010**, *6* (1), 374.
- (21) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. The quantitative proteome of a human cell line. *Molecular Systems Biology* **2011**, *7* (1), 549, <https://doi.org/10.1038/msb.2011.82>. DOI: <https://doi.org/10.1038/msb.2011.82> (accessed 2021/12/09).
- (22) Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* **2013**, *35* (12), 1050-1055, <https://doi.org/10.1002/bies.201300066>. DOI: <https://doi.org/10.1002/bies.201300066> (accessed 2021/12/09).
- (23) Smith, L. M.; Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nature methods* **2013**, *10* (3), 186-187.
- (24) Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* **2001**, *17* (2), 100-107. DOI: [https://doi.org/10.1016/S0168-9525\(00\)02176-4](https://doi.org/10.1016/S0168-9525(00)02176-4).
- (25) Rabiller, M.; Getlik, M.; Klüter, S.; Richters, A.; Tückmantel, S.; Simard, J. R.; Rauh, D. Proteus in the World of Proteins: Conformational Changes in Protein Kinases. *Archiv der Pharmazie* **2010**, *343* (4), 193-206. DOI: <https://doi.org/10.1002/ardp.201000028>.

- (26) Bu, Z.; Callaway, D. J. E. Chapter 5 - Proteins MOVE! Protein dynamics and long-range allostery in cell signaling. In *Advances in Protein Chemistry and Structural Biology*, Donev, R. Ed.; Vol. 83; Academic Press, 2011; pp 163-221.
- (27) Patti, G. J. Separation strategies for untargeted metabolomics. *Journal of separation science* **2011**, *34* (24), 3460-3469.
- (28) Maiorano, F.; Ambrosino, L.; Guarracino, M. R. The MetaboX library: building metabolic networks from KEGG database. In *International Conference on Bioinformatics and Biomedical Engineering*, 2015; Springer: pp 565-576.
- (29) Sampson, J. N.; Boca, S. M.; Shu, X. O.; Stolzenberg-Solomon, R. Z.; Matthews, C. E.; Hsing, A. W.; Tan, Y. T.; Ji, B.-T.; Chow, W.-H.; Cai, Q. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiology and Prevention Biomarkers* **2013**, *22* (4), 631-640.
- (30) Pinu, F. R.; Beale, D. J.; Paten, A. M.; Kouremenos, K.; Swarup, S.; Schirra, H. J.; Wishart, D. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites* **2019**, *9* (4), 76.
- (31) Huang, S.; Chaudhary, K.; Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics* **2017**, *8* (84), Mini Review. DOI: 10.3389/fgene.2017.00084.
- (32) Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **2020**, *14*, 1177932219899051. DOI: 10.1177/1177932219899051 From NLM.
- (33) Ghazalpour, A.; Bennett, B.; Petyuk, V. A.; Orozco, L.; Hagopian, R.; Mungrue, I. N.; Farber, C. R.; Sinsheimer, J.; Kang, H. M.; Furlotte, N.; et al. Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLOS Genetics* **2011**, *7* (6), e1001393. DOI: 10.1371/journal.pgen.1001393.
- (34) Smith, L. M.; Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **2018**, *359* (6380), 1106-1107.
- (35) Wither, M. J.; Hansen, K. C.; Reisz, J. A. Mass Spectrometry-Based Bottom-Up Proteomics: Sample Preparation, LC-MS/MS Analysis, and Database Query Strategies. *Current Protocols in Protein Science* **2016**, *86* (1), 16.14.11-16.14.20, <https://doi.org/10.1002/cpps.18>. DOI: <https://doi.org/10.1002/cpps.18> (accessed 2021/12/16).
- (36) Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **2002**, *1* (11), 845-867. DOI: 10.1074/mcp.r200007-mcp200 From NLM.
- (37) Polaskova, V.; Kapur, A.; Khan, A.; Molloy, M. P.; Baker, M. S. High-abundance protein depletion: Comparison of methods for human plasma biomarker discovery. *ELECTROPHORESIS* **2010**, *31* (3), 471-482. DOI: <https://doi.org/10.1002/elps.200900286>.
- (38) Jankovska, E.; Lipcseyova, D.; Svrlikova, M.; Pavelcova, M.; Kubala Havrdova, E.; Holada, K.; Petrak, J. Quantitative proteomic analysis of cerebrospinal fluid of women newly diagnosed with multiple sclerosis. *International Journal of Neuroscience* **2020**, 1-11. DOI: 10.1080/00207454.2020.1837801.
- (39) Duangkumpha, K.; Stoll, T.; Phetcharaburanin, J.; Yongvanit, P.; Thanan, R.; Techasen, A.; Namwat, N.; Khuntikeo, N.; Chamadol, N.; Roytrakul, S.; et al. Urine proteomics study

reveals potential biomarkers for the differential diagnosis of cholangiocarcinoma and periductal fibrosis. *PLOS ONE* **2019**, *14* (8), e0221024. DOI: 10.1371/journal.pone.0221024.

(40) Prieto, D. A.; Chan, K. C.; Johann, D. J.; Ye, X.; Whitely, G.; Blonder, J. Preparation and Immunoaffinity Depletion of Fresh Frozen Tissue Homogenates for Mass Spectrometry-Based Proteomics in the Context of Drug Target/Biomarker Discovery. In *Proteomics for Drug Discovery: Methods and Protocols*, Lazar, I. M., Kontoyianni, M., Lazar, A. C. Eds.; Springer New York, 2017; pp 71-90.

(41) Blume, J. E.; Manning, W. C.; Troiano, G.; Hornburg, D.; Figa, M.; Hesterberg, L.; Platt, T. L.; Zhao, X.; Cuaresma, R. A.; Everley, P. A.; et al. Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nature Communications* **2020**, *11* (1), 3662. DOI: 10.1038/s41467-020-17033-7.

(42) Kaur, G.; Poljak, A.; Ali, S. A.; Zhong, L.; Raftery, M. J.; Sachdev, P. Extending the Depth of Human Plasma Proteome Coverage Using Simple Fractionation Techniques. *J Proteome Res* **2021**, *20* (2), 1261-1279. DOI: 10.1021/acs.jproteome.0c00670 From NLM.

(43) Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **2009**, *6* (5), 359-362. DOI: 10.1038/nmeth.1322.

(44) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods* **2014**, *11* (3), 319-324. DOI: 10.1038/nmeth.2834.

(45) Chen, W.; Wang, S.; Adhikari, S.; Deng, Z.; Wang, L.; Chen, L.; Ke, M.; Yang, P.; Tian, R. Simple and Integrated Spintip-Based Technology Applied for Deep Proteome Profiling. *Analytical Chemistry* **2016**, *88* (9), 4864-4871. DOI: 10.1021/acs.analchem.6b00631.

(46) HaileMariam, M.; Eguez, R. V.; Singh, H.; Bekele, S.; Ameni, G.; Pieper, R.; Yu, Y. S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *Journal of Proteome Research* **2018**, *17* (9), 2917-2924. DOI: 10.1021/acs.jproteome.8b00505.

(47) Yuan, H.; Zhang, S.; Zhao, B.; Weng, Y.; Zhu, X.; Li, S.; Zhang, L.; Zhang, Y. Enzymatic Reactor with Trypsin Immobilized on Graphene Oxide Modified Polymer Microspheres To Achieve Automated Proteome Quantification. *Analytical Chemistry* **2017**, *89* (12), 6324-6329. DOI: 10.1021/acs.analchem.7b00682.

(48) Hughes, C. S.; Foehr, S.; Garfield, D. A.; Furlong, E. E.; Steinmetz, L. M.; Krijgsveld, J. Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular Systems Biology* **2014**, *10* (10), 757, <https://doi.org/10.15252/msb.20145625>. DOI: <https://doi.org/10.15252/msb.20145625> (accessed 2021/12/14).

(49) Johnston, H. E.; Yadav, K.; Kirkpatrick, J. M.; Biggs, G. S.; Oxley, D.; Kramer, H. B.; Samant, R. S. Solvent Precipitation SP3 (SP4) enhances recovery for proteomics sample preparation without magnetic beads. *bioRxiv* **2021**, 2021.2009.2024.461247. DOI: 10.1101/2021.09.24.461247.

(50) Yates, J. R.; Carmack, E.; Hays, L.; Link, A. J.; Eng, J. K. Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry. *2-D Proteome Analysis Protocols* **1999**, 553-569.

(51) Wang, Z.; Ma, H.; Smith, K.; Wu, S. Two-dimensional separation using high-pH and low-pH reversed phase liquid chromatography for top-down proteomics. *International*

Journal of Mass Spectrometry **2018**, *427*, 43-51. DOI: <https://doi.org/10.1016/j.ijms.2017.09.001>.

(52) Chan, K. C.; Issaq, H. J. Fractionation of peptides by strong cation-exchange liquid chromatography. *Methods Mol Biol* **2013**, *1002*, 311-315. DOI: 10.1007/978-1-62703-360-2_23 From NLM.

(53) Yu, P.; Petzoldt, S.; Wilhelm, M.; Zolg, D. P.; Zheng, R.; Sun, X.; Liu, X.; Schneider, G.; Huhmer, A.; Kuster, B. Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. *Analytical Chemistry* **2017**, *89* (17), 8884-8891. DOI: 10.1021/acs.analchem.7b01356.

(54) Grassetti, A. V.; Hards, R.; Gerber, S. A. Offline pentafluorophenyl (PFP)-RP prefractionation as an alternative to high-pH RP for comprehensive LC-MS/MS proteomics and phosphoproteomics. *Analytical and Bioanalytical Chemistry* **2017**, *409* (19), 4615-4625. DOI: 10.1007/s00216-017-0407-6.

(55) Dimayacyac-Esleta, B. R.; Tsai, C. F.; Kitata, R. B.; Lin, P. Y.; Choong, W. K.; Lin, T. D.; Wang, Y. T.; Weng, S. H.; Yang, P. C.; Arco, S. D.; et al. Rapid High-pH Reverse Phase StageTip for Sensitive Small-Scale Membrane Proteomic Profiling. *Anal Chem* **2015**, *87* (24), 12016-12023. DOI: 10.1021/acs.analchem.5b03639 From NLM.

(56) Kim, H.; Dan, K.; Shin, H.; Lee, J.; Wang, J. I.; Han, D. An efficient method for high-pH peptide fractionation based on C18 StageTips for in-depth proteome profiling. *Analytical Methods* **2019**, *11* (36), 4693-4698, 10.1039/C9AY01269A. DOI: 10.1039/C9AY01269A.

(57) Lee, H.-J.; Kim, H.-J.; Liebler, D. C. Efficient Microscale Basic Reverse Phase Peptide Fractionation for Global and Targeted Proteomics. *Journal of proteome research* **2016**, *15* (7), 2346-2354. DOI: 10.1021/acs.jproteome.6b00102 PubMed.

(58) Bubis, J. A.; Levitsky, L. I.; Ivanov, M. V.; Tarasova, I. A.; Gorshkov, M. V. Comparative evaluation of label-free quantification methods for shotgun proteomics. *Rapid Commun Mass Spectrom* **2017**, *31* (7), 606-612. DOI: 10.1002/rcm.7829 From NLM.

(59) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **2016**, *11* (12), 2301-2319. DOI: 10.1038/nprot.2016.136.

(60) Orsburn, B. C. Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **2021**, *9* (1), 15.

(61) Stead, D. A.; Paton, N. W.; Missier, P.; Embury, S. M.; Hedeler, C.; Jin, B.; Brown, A. J. P.; Preece, A. Information quality in proteomics. *Briefings in Bioinformatics* **2008**, *9* (2), 174-188. DOI: 10.1093/bib/bbn004 (accessed 1/10/2022).

(62) Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res* **2016**, *15* (4), 1116-1125. DOI: 10.1021/acs.jproteome.5b00981 From NLM.

(63) Yu, F.; Haynes, S. E.; Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics* **2021**, *20*, 100077. DOI: <https://doi.org/10.1016/j.mcpro.2021.100077>.

(64) Högberg, A.; von Stechow, L.; Bekker-Jensen, D. B.; Weinert, B. T.; Kelstrup, C. D.; Olsen, J. V. Benchmarking common quantification strategies for large-scale

phosphoproteomics. *Nature Communications* **2018**, *9* (1), 1045. DOI: 10.1038/s41467-018-03309-6.

(65) Chen, X.; Wei, S.; Ji, Y.; Guo, X.; Yang, F. Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics* **2015**, *15* (18), 3175-3192. DOI: 10.1002/pmic.201500108 From NLM.

(66) Krüger, M.; Moser, M.; Ussar, S.; Thievensen, I.; Lubert, C. A.; Forner, F.; Schmidt, S.; Zanivan, S.; Fässler, R.; Mann, M. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **2008**, *134* (2), 353-364. DOI: 10.1016/j.cell.2008.05.033 From NLM.

(67) Deng, J.; Erdjument-Bromage, H.; Neubert, T. A. Quantitative Comparison of Proteomes Using SILAC. *Curr Protoc Protein Sci* **2019**, *95* (1), e74. DOI: 10.1002/cpp.74 From NLM.

(68) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003**, *75* (8), 1895-1904. DOI: 10.1021/ac0262560 From NLM.

(69) Pottiez, G.; Wiederin, J.; Fox, H. S.; Ciborowski, P. Comparison of 4-plex to 8-plex iTRAQ Quantitative Measurements of Proteins in Human Plasma Samples. *Journal of Proteome Research* **2012**, *11* (7), 3774-3781. DOI: 10.1021/pr300414z.

(70) Li, J.; Van Vranken, J. G.; Pontano Vaites, L.; Schweppe, D. K.; Huttlin, E. L.; Etienne, C.; Nandhikonda, P.; Viner, R.; Robitaille, A. M.; Thompson, A. H.; et al. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nature Methods* **2020**, *17* (4), 399-404. DOI: 10.1038/s41592-020-0781-4.

(71) Li, J.; Cai, Z.; Bomgarden, R. D.; Pike, I.; Kuhn, K.; Rogers, J. C.; Roberts, T. M.; Gygi, S. P.; Paulo, J. A. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *Journal of Proteome Research* **2021**, *20* (5), 2964-2972. DOI: 10.1021/acs.jproteome.1c00168.

(72) Jiang, H.; Zhang, L.; Zhang, Y.; Xie, L.; Wang, Y.; Lu, H. HST-MRM-MS: a novel high-sample-throughput multiple reaction monitoring mass spectrometric method for multiplex absolute quantitation of hepatocellular carcinoma serum biomarker. *Journal of proteome research* **2018**, *18* (1), 469-477.

(73) Han, X.; Aslanian, A.; Yates, J. R., 3rd. Mass spectrometry for proteomics. *Current opinion in chemical biology* **2008**, *12* (5), 483-490. DOI: 10.1016/j.cbpa.2008.07.024 PubMed.

(74) Meyer, J. G. Fast Proteome Identification and Quantification from Data-Dependent Acquisition–Tandem Mass Spectrometry (DDA MS/MS) Using Free Software Tools. *Methods and protocols* **2019**, *2* (1), 8.

(75) Chapman, J. D.; Goodlett, D. R.; Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrometry Reviews* **2014**, *33* (6), 452-470. DOI: <https://doi.org/10.1002/mas.21400>.

(76) Govaert, E.; Van Steendam, K.; Willems, S.; Vossaert, L.; Dhaenens, M.; Deforce, D. Comparison of fractionation proteomics for local SWATH library building. *PROTEOMICS*

- 2017**, 17 (15-16), 1700052, <https://doi.org/10.1002/pmic.201700052>. DOI: <https://doi.org/10.1002/pmic.201700052> (accessed 2021/12/16).
- (77) Pino, L. K.; Just, S. C.; MacCoss, M. J.; Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Molecular & Cellular Proteomics* **2020**, 19 (7), 1088-1103. DOI: 10.1074/mcp.P119.001913 (accessed 2021/12/16).
- (78) Sinitcyn, P.; Hamzeiy, H.; Salinas Soto, F.; Itzhak, D.; McCarthy, F.; Wichmann, C.; Steger, M.; Ohmayer, U.; Distler, U.; Kaspar-Schoenefeld, S.; et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology* **2021**, 39 (12), 1563-1573. DOI: 10.1038/s41587-021-00968-7.
- (79) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods* **2019**, 16 (6), 509-518.
- (80) Zhou, Y.; Tan, Z.; Xue, P.; Wang, Y.; Li, X.; Guan, F. High-throughput, in-depth and estimated absolute quantification of plasma proteome using data-independent acquisition/mass spectrometry ("HIAP-DIA"). *Proteomics* **2021**, 21 (5), e2000264. DOI: 10.1002/pmic.202000264 From NLM.
- (81) Messner, C. B.; Demichev, V.; Wendisch, D.; Michalick, L.; White, M.; Freiwald, A.; Textoris-Taube, K.; Vernardis, S. I.; Egger, A.-S.; Kreidl, M. Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell systems* **2020**, 11 (1), 11-24. e14.
- (82) Kennedy, J.; Yi, E. C. Use of gas-phase fractionation to increase protein identifications: application to the peroxisome. *Methods Mol Biol* **2008**, 432, 217-228. DOI: 10.1007/978-1-59745-028-7_15 From NLM.
- (83) Meyer, J. G.; Niemi, N. M.; Pagliarini, D. J.; Coon, J. J. Quantitative shotgun proteome analysis by direct infusion. *Nature methods* **2020**, 17 (12), 1222-1228. DOI: 10.1038/s41592-020-00999-z PubMed.
- (84) Trujillo, E. A.; Hebert, A. S.; Brademan, D. R.; Coon, J. J. Maximizing Tandem Mass Spectrometry Acquisition Rates for Shotgun Proteomics. *Analytical Chemistry* **2019**, 91 (20), 12625-12629. DOI: 10.1021/acs.analchem.9b02979.
- (85) Kelstrup, C. D.; Jersie-Christensen, R. R.; Batth, T. S.; Arrey, T. N.; Kuehn, A.; Kellmann, M.; Olsen, J. V. Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *Journal of Proteome Research* **2014**, 13 (12), 6187-6195. DOI: 10.1021/pr500985w.
- (86) Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; et al. Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. *Anal Chem* **2013**, 85 (24), 11710-11714. DOI: 10.1021/ac403115c From NLM.
- (87) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The one hour yeast proteome. *Mol Cell Proteomics* **2014**, 13 (1), 339-347. DOI: 10.1074/mcp.M113.034769 From NLM.
- (88) Distler, U.; Kuharev, J.; Navarro, P.; Tenzer, S. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. *Nat Protoc* **2016**, 11 (4), 795-812. DOI: 10.1038/nprot.2016.042 From NLM.

- (89) May, J. C.; McLean, J. A. Ion Mobility-Mass Spectrometry: Time-Dispersive Instrumentation. *Analytical Chemistry* **2015**, *87* (3), 1422-1436. DOI: 10.1021/ac504720m.
- (90) Swearingen, K. E.; Moritz, R. L. High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. *Expert review of proteomics* **2012**, *9* (5), 505-517. DOI: 10.1586/epr.12.50 PubMed.
- (91) Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; Rütger, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients*. *Molecular & Cellular Proteomics* **2020**, *19* (4), 716-729. DOI: <https://doi.org/10.1074/mcp.TIR119.001906>.
- (92) Klaeger, S.; Apffel, A.; Clauser, K. R.; Sarkizova, S.; Oliveira, G.; Rachimi, S.; Le, P. M.; Tarren, A.; Chea, V.; Abelin, J. G.; et al. Optimized Liquid and Gas Phase Fractionation Increases HLA-Peptidome Coverage for Primary Cell and Tissue Samples. *Molecular & Cellular Proteomics* **2021**, *20*. DOI: 10.1016/j.mcpro.2021.100133 (accessed 2022/01/09).
- (93) Ridgeway, M. E.; Lubeck, M.; Jordens, J.; Mann, M.; Park, M. A. Trapped ion mobility spectrometry: A short review. *International Journal of Mass Spectrometry* **2018**, *425*, 22-35. DOI: <https://doi.org/10.1016/j.ijms.2018.01.006>.
- (94) Meier, F.; Brunner, A. D.; Koch, S.; Koch, H.; Lubeck, M.; Krause, M.; Goedecke, N.; Decker, J.; Kosinski, T.; Park, M. A.; et al. Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* **2018**, *17* (12), 2534-2545. DOI: 10.1074/mcp.TIR118.000900 From NLM.
- (95) Aballo, T. J.; Roberts, D. S.; Melby, J. A.; Buck, K. M.; Brown, K. A.; Ge, Y. Ultrafast and Reproducible Proteomics from Small Amounts of Heart Tissue Enabled by Azo and timsTOF Pro. *Journal of Proteome Research* **2021**, *20* (8), 4203-4211. DOI: 10.1021/acs.jproteome.1c00446.
- (96) Reeves, G. A.; Talavera, D.; Thornton, J. M. Genome and proteome annotation: organization, interpretation and integration. *Journal of The Royal Society Interface* **2009**, *6* (31), 129-147. DOI: doi:10.1098/rsif.2008.0341.
- (97) The UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **2021**, *49* (D1), D480-D489. DOI: 10.1093/nar/gkaa1100 (accessed 12/20/2021).
- (98) Smedley, D.; Haider, S.; Ballester, B.; Holland, R.; London, D.; Thorisson, G.; Kasprzyk, A. BioMart – biological queries made easy. *BMC Genomics* **2009**, *10* (1), 22. DOI: 10.1186/1471-2164-10-22.
- (99) Luzarowski, M.; Vicente, R.; Kiselev, A.; Wagner, M.; Schlossarek, D.; Erban, A.; de Souza, L. P.; Childs, D.; Wojciechowska, I.; Luzarowska, U.; et al. Global mapping of protein–metabolite interactions in *Saccharomyces cerevisiae* reveals that Ser-Leu dipeptide regulates phosphoglycerate kinase activity. *Communications Biology* **2021**, *4* (1), 181. DOI: 10.1038/s42003-021-01684-3.
- (100) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: the Human Metabolome Database. *Nucleic acids research* **2007**, *35* (Database issue), D521-D526. DOI: 10.1093/nar/gkl923 PubMed.
- (101) Kamburov, A.; Stelzl, U.; Lehrach, H.; Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research* **2013**, *41* (D1), D793-D800. DOI: 10.1093/nar/gks1055 (accessed 12/20/2021).

- (102) Wishart, D. S.; Li, C.; Marcu, A.; Badran, H.; Pon, A.; Budinski, Z.; Patron, J.; Lipton, D.; Cao, X.; Oler, E.; et al. PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res* **2020**, *48* (D1), D470-d478. DOI: 10.1093/nar/gkz861 From NLM.
- (103) Robinson, M. D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **2010**, *11* (3), R25. DOI: 10.1186/gb-2010-11-3-r25.
- (104) Willforss, J.; Chawade, A.; Levander, F. NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *Journal of Proteome Research* **2019**, *18* (2), 732-740. DOI: 10.1021/acs.jproteome.8b00523.
- (105) Wang, S.; Yang, H. pseudoQC: A Regression-Based Simulation Software for Correction and Normalization of Complex Metabolomics and Proteomics Datasets. *PROTEOMICS* **2019**, *19* (19), 1900264. DOI: <https://doi.org/10.1002/pmic.201900264>.
- (106) Hyvärinen, A. Independent component analysis: recent advances. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **2012**, *371* (1984), 20110534-20110534. DOI: 10.1098/rsta.2011.0534 PubMed.
- (107) Tzeng, D.-Y.; Berns, R. S. A review of principal component analysis and its applications to color technology. *Color Research & Application* **2005**, *30* (2), 84-98, <https://doi.org/10.1002/col.20086>. DOI: <https://doi.org/10.1002/col.20086> (accessed 2021/12/20).
- (108) Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery* **2012**, *2* (1), 86-97, <https://doi.org/10.1002/widm.53>. DOI: <https://doi.org/10.1002/widm.53> (accessed 2021/12/20).
- (109) Mooney, M. A.; Wilmot, B. Gene set analysis: A step-by-step guide. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **2015**, *168* (7), 517-527. DOI: 10.1002/ajmg.b.32328 PubMed.
- (110) Santamaria, C.; Garcia-Mora, B.; Rubio, G.; Falcó, A. Topographic representation of cancer data using Boolean Networks. *Modelling for Engineering & Human Behaviour* **2019**, 180.
- (111) Hou, J.; Acharya, L.; Zhu, D.; Cheng, J. An overview of bioinformatics methods for modeling biological pathways in yeast. *Briefings in functional genomics* **2016**, *15* (2), 95-108. DOI: 10.1093/bfgp/elv040 PubMed.
- (112) Mi, H.; Ebert, D.; Muruganujan, A.; Mills, C.; Albu, L.-P.; Mushayamaha, T.; Thomas, P. D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* **2020**, *49* (D1), D394-D403. DOI: 10.1093/nar/gkaa1106 (accessed 3/3/2022).
- (113) Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **2019**, *47* (W1), W191-W198. DOI: 10.1093/nar/gkz369 (accessed 12/20/2021).

- (114) Manzini, S.; Busnelli, M.; Colombo, A.; Franchi, E.; Grossano, P.; Chiesa, G. reString: an open-source Python software to perform automatic functional enrichment retrieval, results aggregation and data visualization. *Scientific Reports* **2021**, *11* (1), 23458. DOI: 10.1038/s41598-021-02528-0.
- (115) Meng, C.; Basunia, A.; Peters, B.; Gholami, A. M.; Kuster, B.; Culhane, A. C. MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data. *Mol Cell Proteomics* **2019**, *18* (8 suppl 1), S153-s168. DOI: 10.1074/mcp.TIR118.001251 From NLM.
- (116) Langfelder, P.; Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **2008**, *9* (1), 559. DOI: 10.1186/1471-2105-9-559.
- (117) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **2018**, *47* (D1), D607-D613. DOI: 10.1093/nar/gky1131 (accessed 3/3/2022).
- (118) Jang, Y.; Yu, N.; Seo, J.; Kim, S.; Lee, S. MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. *Biology direct* **2016**, *11* (1), 10-10. DOI: 10.1186/s13062-016-0112-y PubMed.
- (119) Meng, C.; Helm, D.; Frejno, M.; Kuster, B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *J Proteome Res* **2016**, *15* (3), 755-765. DOI: 10.1021/acs.jproteome.5b00824 From NLM.
- (120) Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology* **2017**, *13* (11), e1005752. DOI: 10.1371/journal.pcbi.1005752.
- (121) Planell, N.; Lagani, V.; Sebastian-Leon, P.; van der Kloet, F.; Ewing, E.; Karathanasis, N.; Urdangarin, A.; Arozarena, I.; Jagodic, M.; Tsamardinos, I.; et al. STATegra: Multi-Omics Data Integration – A Conceptual Scheme With a Bioinformatics Pipeline. *Frontiers in Genetics* **2021**, *12* (143), Technology and Code. DOI: 10.3389/fgene.2021.620453.
- (122) Koh, H. W. L.; Fermin, D.; Vogel, C.; Choi, K. P.; Ewing, R. M.; Choi, H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *npj Systems Biology and Applications* **2019**, *5* (1), 22. DOI: 10.1038/s41540-019-0099-y.
- (123) Bodein, A.; Scott-Boyer, M. P.; Perin, O.; KA, L. C.; Droit, A. Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Res* **2021**. DOI: 10.1093/nar/gkab1200 From NLM.
- (124) Kumar, P.; Panigrahi, P.; Johnson, J.; Weber, W. J.; Mehta, S.; Sajulga, R.; Easterly, C.; Crooker, B. A.; Heydarian, M.; Anamika, K.; et al. QuanTP: A Software Resource for Quantitative Proteo-Transcriptomic Comparative Data Analysis and Informatics. *Journal of Proteome Research* **2019**, *18* (2), 782-790. DOI: 10.1021/acs.jproteome.8b00727.
- (125) Klopfenstein, D. V.; Zhang, L.; Pedersen, B. S.; Ramírez, F.; Warwick Vesztrocy, A.; Naldi, A.; Mungall, C. J.; Yunes, J. M.; Botvinnik, O.; Weigel, M.; et al. GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports* **2018**, *8* (1), 10872. DOI: 10.1038/s41598-018-28948-z.
- (126) Kelly, R. S.; Chawes, B. L.; Blighe, K.; Virkud, Y. V.; Croteau-Chonka, D. C.; McGeachie, M. J.; Clish, C. B.; Bullock, K.; Celedón, J. C.; Weiss, S. T.; et al. An Integrative

Transcriptomic and Metabolomic Study of Lung Function in Children With Asthma. *Chest* **2018**, *154* (2), 335-348. DOI: 10.1016/j.chest.2018.05.038 From NLM.

(127) Hoadley, K. A.; Yau, C.; Wolf, D. M.; Cherniack, A. D.; Tamborero, D.; Ng, S.; Leiserson, M. D. M.; Niu, B.; McLellan, M. D.; Uzunangelov, V.; et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **2014**, *158* (4), 929-944. DOI: 10.1016/j.cell.2014.06.049 From NLM.

(128) Singh, A.; Gautier, B.; Shannon, C. P.; Vacher, M.; Rohart, F.; Tebbutt, S. J.; Le Cao, K.-A. DIABLO—an integrative, multi-omics, multivariate method for multi-group classification. *BioRxiv* **2016**, 067611.

(129) Rohart, F.; Eslami, A.; Matigian, N.; Bougeard, S.; Le Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC bioinformatics* **2017**, *18* (1), 1-13.

(130) Conesa, A. The STATegra project: new statistical tools for analysis and integration of diverse omics data. *EMBnet. journal* **2014**, *20* (A), 768.

(131) Karathanasis, N.; Tsamardinos, I.; Lagani, V. OmicsNPC: applying the non-parametric combination methodology to the integrative analysis of heterogeneous omics data. *PLoS one* **2016**, *11* (11), e0165545.

(132) Bodein, A.; Scott-Boyer, M. P.; Perin, O.; KA, L. C.; Droit, A. timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics* **2021**. DOI: 10.1093/bioinformatics/btab664 From NLM.

(133) Krämer, A.; Green, J.; Pollard, J., Jr.; Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014**, *30* (4), 523-530. DOI: 10.1093/bioinformatics/btt703 From NLM.

(134) Jalili, V.; Afgan, E.; Gu, Q.; Clements, D.; Blankenberg, D.; Goecks, J.; Taylor, J.; Nekrutenko, A. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research* **2020**, *48* (W1), W395-W402. DOI: 10.1093/nar/gkaa434 (accessed 1/10/2022).

(135) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **2016**, *13* (9), 731-740. DOI: 10.1038/nmeth.3901.

(136) Tran, N. C.; Gao, J. X. OpenOmics: A bioinformatics API to integrate multi-omics datasets and interface with public databases. *Journal of Open Source Software* **2021**, *6* (61), 3249.

(137) Ghosh, S.; Datta, A.; Choi, H. multiSLIDE is a web server for exploring connected elements of biological pathways in multi-omics data. *Nature Communications* **2021**, *12* (1), 2279. DOI: 10.1038/s41467-021-22650-x.

(138) Zoppi, J.; Guillaume, J.-F.; Neunlist, M.; Chaffron, S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinformatics* **2021**, *22* (1), 6. DOI: 10.1186/s12859-020-03921-8.

(139) Jiao, X.; Sherman, B. T.; Huang, D. W.; Stephens, R.; Baseler, M. W.; Lane, H. C.; Lempicki, R. A. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics (Oxford, England)* **2012**, *28* (13), 1805-1806. DOI: 10.1093/bioinformatics/bts251 PubMed.

- (140) Tenenbaum, D.; RUnit, S.; Maintainer, M. B. P.; Carlson, M.; ThirdPartyClient, K. Package 'KEGGREST'. *R Foundation for Statistical Computing: Vienna, Austria* **2019**.
- (141) Bu, D.; Luo, H.; Huo, P.; Wang, Z.; Zhang, S.; He, Z.; Wu, Y.; Zhao, L.; Liu, J.; Guo, J.; et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Research* **2021**, *49* (W1), W317-W325. DOI: 10.1093/nar/gkab447 (accessed 7/16/2021).
- (142) Giacomoni, F.; Le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J. F.; Jacob, D.; et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015**, *31* (9), 1493-1495. DOI: 10.1093/bioinformatics/btu813 From NLM.
- (143) Rudolph, J. D.; Cox, J. A Network Module for the Perseus Software for Computational Proteomics Facilitates Proteome Interaction Graph Analysis. *Journal of Proteome Research* **2019**, *18* (5), 2052-2064. DOI: 10.1021/acs.jproteome.8b00927.
- (144) Yu, S. H.; Ferretti, D.; Schessner, J. P.; Rudolph, J. D.; Borner, G. H. H.; Cox, J. Expanding the Perseus Software for Omics Data Analysis With Custom Plugins. *Curr Protoc Bioinformatics* **2020**, *71* (1), e105. DOI: 10.1002/cpbi.105 From NLM.
- (145) Cavalli, M.; Diamanti, K.; Pan, G.; Spalinskas, R.; Kumar, C.; Deshmukh, A. S.; Mann, M.; Sahlén, P.; Komorowski, J.; Wadelius, C. A Multi-Omics Approach to Liver Diseases: Integration of Single Nuclei Transcriptomics with Proteomics and HiCap Bulk Data in Human Liver. *OMICS: A Journal of Integrative Biology* **2020**, *24* (4), 180-194. DOI: 10.1089/omi.2019.0215 (accessed 2022/01/13).
- (146) Fornecker, L.-M.; Muller, L.; Bertrand, F.; Paul, N.; Pichot, A.; Herbrecht, R.; Chenard, M.-P.; Mauvieux, L.; Vallat, L.; Bahram, S.; et al. Multi-omics dataset to decipher the complexity of drug resistance in diffuse large B-cell lymphoma. *Scientific Reports* **2019**, *9* (1), 895. DOI: 10.1038/s41598-018-37273-4.
- (147) Alcazar, O.; Hernandez, L. F.; Nakayasu, E. S.; Nicora, C. D.; Ansong, C.; Muehlbauer, M. J.; Bain, J. R.; Myer, C. J.; Bhattacharya, S. K.; Buchwald, P.; et al. Parallel Multi-Omics in High-Risk Subjects for the Identification of Integrated Biomarker Signatures of Type 1 Diabetes. *Biomolecules* **2021**, *11* (3), 383.
- (148) Lee, H.; Sung, E. J.; Seo, S.; Min, E. K.; Lee, J.-Y.; Shim, I.; Kim, P.; Kim, T.-Y.; Lee, S.; Kim, K.-T. Integrated multi-omics analysis reveals the underlying molecular mechanism for developmental neurotoxicity of perfluorooctanesulfonic acid in zebrafish. *Environment International* **2021**, *157*, 106802. DOI: <https://doi.org/10.1016/j.envint.2021.106802>.
- (149) McLoughlin, F.; Augustine, R. C.; Marshall, R. S.; Li, F.; Kirkpatrick, L. D.; Otegui, M. S.; Vierstra, R. D. Maize multi-omics reveal roles for autophagic recycling in proteome remodelling and lipid turnover. *Nature Plants* **2018**, *4* (12), 1056-1070. DOI: 10.1038/s41477-018-0299-2.
- (150) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014**, *11* (11), 1114-1125. DOI: 10.1038/nmeth.3144 From NLM.
- (151) Zhang, B.; Whiteaker, J. R.; Hoofnagle, A. N.; Baird, G. S.; Rodland, K. D.; Paulovich, A. G. Clinical potential of mass spectrometry-based proteogenomics. *Nature Reviews Clinical Oncology* **2019**, *16* (4), 256-268. DOI: 10.1038/s41571-018-0135-7.

- (152) Bonnal, S. C.; López-Oreja, I.; Valcárcel, J. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology* **2020**, *17* (8), 457-474. DOI: 10.1038/s41571-020-0350-x.
- (153) Tariq, M. U.; Haseeb, M.; Aledhari, M.; Razzak, R.; Parizi, R. M.; Saeed, F. Methods for Proteogenomics Data Analysis, Challenges, and Scalability Bottlenecks: A Survey. *IEEE Access* **2021**, *9*, 5497-5516. DOI: 10.1109/access.2020.3047588 From NLM.
- (154) Langmead, B. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics* **2010**, *Chapter 11*, Unit-11.17. DOI: 10.1002/0471250953.bi1107s32 PubMed.
- (155) Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34* (18), 3094-3100. DOI: 10.1093/bioinformatics/bty191 (accessed 12/20/2021).
- (156) Houtgast, E. J.; Sima, V.-M.; Bertels, K.; Al-Ars, Z. GPU-accelerated BWA-MEM genomic mapping algorithm using adaptive load balancing. In *International conference on architecture of computing systems*, 2016; Springer: pp 130-142.
- (157) Zerbino, D. R. Using the velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics* **2010**, *31* (1), 11.15. 11-11.15. 12.
- (158) Simpson, J. T.; Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome research* **2012**, *22* (3), 549-556.
- (159) Morgenstern, B.; Dress, A.; Werner, T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences* **1996**, *93* (22), 12098-12103.
- (160) Guillot, L.; Delage, L.; Viari, A.; Vandenbrouck, Y.; Com, E.; Ritter, A.; Lavigne, R.; Marie, D.; Peterlongo, P.; Potin, P. Peptimapper: proteogenomics workflow for the expert annotation of eukaryotic genomes. *BMC genomics* **2019**, *20* (1), 1-19.
- (161) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in genetics* **2000**, *16* (6), 276-277.
- (162) Kim, D.; Langmead, B.; Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **2015**, *12* (4), 357-360.
- (163) Trapnell, C.; Pachter, L.; Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25* (9), 1105-1111.
- (164) Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* **2012**.
- (165) DePristo, M. A.; Banks, E.; Poplin, R.; Garimella, K. V.; Maguire, J. R.; Hartl, C.; Philippakis, A. A.; del Angel, G.; Rivas, M. A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **2011**, *43* (5), 491-498. DOI: 10.1038/ng.806.
- (166) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29* (24), 3235-3237.
- (167) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *Journal of Proteome Research* **2021**, *20* (4), 1826-1834. DOI: 10.1021/acs.jproteome.0c00407.

- (168) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X! Tandem searches. *Proteomics* **2011**, *11* (5), 996-999.
- (169) Kumar, D.; Yadav, A. K.; Dash, D. Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. Springer New York, 2017; pp 17-29.
- (170) Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *PROTEOMICS* **2013**, *13* (8), 1352-1357. DOI: <https://doi.org/10.1002/pmic.201200352>.
- (171) Kumar, P.; Johnson, J. E.; Easterly, C.; Mehta, S.; Sajulga, R.; Nunn, B.; Jagtap, P. D.; Griffin, T. J. A sectioning and database enrichment approach for improved peptide spectrum matching in large, genome-guided protein sequence databases. *Journal of Proteome Research* **2020**.
- (172) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **2009**, *10*.
- (173) Wen, B.; Wang, X.; Zhang, B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Research* **2019**, *29* (3), 485-493. DOI: 10.1101/gr.235028.118.
- (174) McGowan, T.; Johnson, J. E.; Kumar, P.; Sajulga, R.; Mehta, S.; Jagtap, P. D.; Griffin, T. J. Multi-omics Visualization Platform: An extensible Galaxy plug-in for multi-omics data visualization and exploration. *GigaScience* **2020**, *9* (4). DOI: 10.1093/gigascience/giaa025.
- (175) Sanders, W. S.; Wang, N.; Bridges, S. M.; Malone, B. M.; Dandass, Y. S.; McCarthy, F. M.; Nanduri, B.; Lawrence, M. L.; Burgess, S. C. The Proteogenomic Mapping Tool. *BMC Bioinformatics* **2011**, *12* (1), 115. DOI: 10.1186/1471-2105-12-115.
- (176) Mani, D. R.; Maynard, M.; Kothadia, R.; Krug, K.; Christianson, K. E.; Heiman, D.; Clauser, K. R.; Birger, C.; Getz, G.; Carr, S. A. PANOPLY: a cloud-based platform for automated and reproducible proteogenomic data analysis. *Nature Methods* **2021**, *18* (6), 580-582. DOI: 10.1038/s41592-021-01176-6.
- (177) Rudnick, P. A.; Markey, S. P.; Roth, J.; Mirokhin, Y.; Yan, X.; Tchekhovskoi, D. V.; Edwards, N. J.; Thangudu, R. R.; Ketchum, K. A.; Kinsinger, C. R.; et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J Proteome Res* **2016**, *15* (3), 1023-1032. DOI: 10.1021/acs.jproteome.5b01091 From NLM.
- (178) Zhang, H.; Liu, T.; Zhang, Z.; Payne, S. H.; Zhang, B.; McDermott, J. E.; Zhou, J. Y.; Petyuk, V. A.; Chen, L.; Ray, D.; et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166* (3), 755-765. DOI: 10.1016/j.cell.2016.05.069 From Nlm.
- (179) Petralia, F.; Tignor, N.; Reva, B.; Koptyra, M.; Chowdhury, S.; Rykunov, D.; Krek, A.; Ma, W.; Zhu, Y.; Ji, J.; et al. Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell* **2020**, *183* (7), 1962-1985.e1931. DOI: 10.1016/j.cell.2020.10.044 From NLM.
- (180) Suvarna, K.; Salkar, A.; Palanivel, V.; Bankar, R.; Banerjee, N.; Gayathri, J. P. M.; Srivastava, A.; Singh, A.; Khatri, H.; Agrawal, S.; et al. A Multi-omics Longitudinal Study Reveals Alteration of the Leukocyte Activation Pathway in COVID-19 Patients. *J Proteome Res* **2021**, *20* (10), 4667-4680. DOI: 10.1021/acs.jproteome.1c00215 From NLM.

- (181) Thuy-Boun, P. S.; Mehta, S.; Gruening, B.; McGowan, T.; Nguyen, A.; Rajczewski, A. T.; Johnson, J. E.; Griffin, T. J.; Wolan, D. W.; Jagtap, P. D. Metaproteomics Analysis of SARS-CoV-2-Infected Patient Samples Reveals Presence of Potential Coinfecting Microorganisms. *Journal of Proteome Research* **2021**, *20* (2), 1451-1454. DOI: 10.1021/acs.jproteome.0c00822.
- (182) Docking, T. R.; Parker, J. D. K.; Jädersten, M.; Duns, G.; Chang, L.; Jiang, J.; Pilsworth, J. A.; Swanson, L. A.; Chan, S. K.; Chiu, R.; et al. A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nature Communications* **2021**, *12* (1), 2474. DOI: 10.1038/s41467-021-22625-y.
- (183) Chauvin, A.; Boisvert, F.-M. Clinical Proteomics in Colorectal Cancer, a Promising Tool for Improving Personalised Medicine. *Proteomes* **2018**, *6* (4), 49. DOI: 10.3390/proteomes6040049 PubMed.
- (184) Meyer, J. G.; Schilling, B. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert review of proteomics* **2017**, *14* (5), 419-429. DOI: 10.1080/14789450.2017.1322904 PubMed.
- (185) Thomas, J. P.; Modos, D.; Korcsmaros, T.; Brooks-Warburton, J. Network Biology Approaches to Achieve Precision Medicine in Inflammatory Bowel Disease. *Front Genet* **2021**, *12*, 760501. DOI: 10.3389/fgene.2021.760501 From NLM.
- (186) Pandeswari, P. B.; Sabareesh, V. Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Advances* **2019**, *9* (1), 313-344, 10.1039/C8RA07200K. DOI: 10.1039/C8RA07200K.
- (187) Budnik, B.; Levy, E.; Harmange, G.; Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology* **2018**, *19* (1), 161. DOI: 10.1186/s13059-018-1547-5.
- (188) Cong, Y.; Motamedchaboki, K.; Misal, S. A.; Liang, Y.; Guise, A. J.; Truong, T.; Huguet, R.; Plowey, E. D.; Zhu, Y.; Lopez-Ferrer, D.; et al. Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per mammalian cell. *Chemical Science* **2021**, *12* (3), 1001-1006, 10.1039/D0SC03636F. DOI: 10.1039/D0SC03636F.
- (189) Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems* **2019**, *4* (3), e00115-00119. DOI: doi:10.1128/mSystems.00115-19.
- (190) Gurdeep Singh, R.; Tanca, A.; Palomba, A.; Van Der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept 4.0: Functional Analysis of Metaproteome Data. *Journal of Proteome Research* **2019**, *18* (2), 606-615. DOI: 10.1021/acs.jproteome.8b00716.
- (191) Easterly, C. W.; Sajulga, R.; Mehta, S.; Johnson, J.; Kumar, P.; Hubler, S.; Mesuere, B.; Rudney, J.; Griffin, T. J.; Jagtap, P. D. metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes. *Mol Cell Proteomics* **2019**, *18* (8 suppl 1), S82-s91. DOI: 10.1074/mcp.RA118.001240 From NLM.
- (192) Matallana-Surget, S.; Jagtap, P. D.; Griffin, T. J.; Beraud, M.; Wattiez, R. Chapter 17 - Comparative Metaproteomics to Study Environmental Changes. In *Metagenomics*, Nagarajan, M. Ed.; Academic Press, 2018; pp 327-363.

- (193) Bargiela, R.; Herbst, F.-A.; Martínez-Martínez, M.; Seifert, J.; Rojo, D.; Cappello, S.; Genovese, M.; Crisafi, F.; Denaro, R.; Chernikova, T. N.; et al. Metaproteomics and metabolomics analyses of chronically petroleum-polluted sites reveal the importance of general anaerobic processes uncoupled with degradation. *Proteomics* **2015**, *15* (20), 3508-3520. DOI: 10.1002/pmic.201400614 PubMed.
- (194) Wang, Y.; Zhou, Y.; Xiao, X.; Zheng, J.; Zhou, H. Metaproteomics: A strategy to study the taxonomy and functionality of the gut microbiota. *Journal of Proteomics* **2020**, *219*, 103737. DOI: <https://doi.org/10.1016/j.jprot.2020.103737>.
- (195) Jagtap, P. D.; Viken, K. J.; Johnson, J.; McGowan, T.; Pendleton, K. M.; Griffin, T. J.; Hunter, R. C.; Rudney, J. D.; Bhargava, M. BAL Fluid Metaproteome in Acute Respiratory Failure. *American journal of respiratory cell and molecular biology* **2018**, *59* (5), 648-652. DOI: 10.1165/rcmb.2018-0068LE PubMed.
- (196) Devereaux, Z. J.; Reynolds, C. A.; Fischer, J. L.; Foley, C. D.; DeLeeuw, J. L.; Wager-Miller, J.; Narayan, S. B.; Mackie, K.; Trimpin, S. Matrix-Assisted Ionization on a Portable Mass Spectrometer: Analysis Directly from Biological and Synthetic Materials. *Analytical Chemistry* **2016**, *88* (22), 10831-10836. DOI: 10.1021/acs.analchem.6b00304.
- (197) Costanzo, M. T.; Boock, J. J.; Kemperman, R. H.; Wei, M. S.; Beekman, C. R.; Yost, R. A. Portable FAIMS: Applications and future perspectives. *International journal of mass spectrometry* **2017**, *422*, 188-196.
- (198) Müller, T.; Kalxdorf, M.; Longuespée, R.; Kazdal, D. N.; Stenzinger, A.; Krijgsveld, J. Automated sample preparation with SP 3 for low-input clinical proteomics. *Molecular systems biology* **2020**, *16* (1), e9111.
- (199) Fu, Q.; Kowalski, M. P.; Mastali, M.; Parker, S. J.; Sobhani, K.; van den Broek, I.; Hunter, C. L.; Van Eyk, J. E. Highly Reproducible Automated Proteomics Sample Preparation Workflow for Quantitative Mass Spectrometry. *Journal of proteome research* **2018**, *17* (1), 420-428. DOI: 10.1021/acs.jproteome.7b00623 PubMed.
- (200) Jain, M.; Olsen, H. E.; Paten, B.; Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology* **2016**, *17* (1), 1-11.
- (201) Cervantes, J.; Perry, C.; Wang, M. C. Teaching next-generation sequencing to medical students with a portable sequencing device. *Perspectives on medical education* **2021**, *10* (4), 252-255.
- (202) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine* **2020**, *382* (8), 727-733. DOI: 10.1056/NEJMoa2001017.
- (203) Nuzzo, J.; Moss, W.; Kahn, J.; Rutgow, L.; al., e. *Have states flattened the curve?* Johns Hopkins University, 2020. <https://coronavirus.jhu.edu/> (accessed November 15th, 2020).
- (204) Sanyaolu, A.; Okorie, C.; Marinkovic, A.; Patidar, R.; Younis, K.; Desai, P.; Hosein, Z.; Padda, I.; Mangat, J.; Altaf, M. Comorbidity and its Impact on Patients with COVID-19. *SN Comprehensive Clinical Medicine* **2020**, *2* (8), 1069-1076. DOI: 10.1007/s42399-020-00363-4.
- (205) Ye, Q.; Wang, B.; Mao, J. The pathogenesis and treatment of the 'Cytokine Storm' in COVID-19. *The Journal of infection* **2020**, *80* (6), 607-613. DOI: 10.1016/j.jinf.2020.03.037 PubMed.

- (206) Beigel, J. H.; Tomashek, K. M.; Dodd, L. E.; Mehta, A. K.; Zingman, B. S.; Kalil, A. C.; Hohmann, E.; Chu, H. Y.; Luetkemeyer, A.; Kline, S.; et al. Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal of Medicine* **2020**, *383* (19), 1813-1826. DOI: 10.1056/NEJMoa2007764.
- (207) Riva, L.; Yuan, S.; Yin, X.; Martin-Sancho, L.; Matsunaga, N.; Pache, L.; Burgstaller-Muehlbacher, S.; De Jesus, P. D.; Teriete, P.; Hull, M. V. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **2020**, *586* (7827), 113-119.
- (208) Poland, G. A.; Ovsyannikova, I. G.; Crooke, S. N.; Kennedy, R. B. SARS-CoV-2 Vaccine Development: Current Status. *Mayo Clinic Proceedings* **2020**, *95* (10), 2172 - 2188. DOI: <https://doi.org/10.1016/j.mayocp.2020.07.021>.
- (209) Dagotto, G.; Yu, J.; Barouch, D. H. Approaches and challenges in SARS-CoV-2 vaccine development. *Cell host & microbe* **2020**.
- (210) Jackson, L. A.; Anderson, E. J.; Roupael, N. G.; Roberts, P. C.; Makhene, M.; Coler, R. N.; McCullough, M. P.; Chappell, J. D.; Denison, M. R.; Stevens, L. J.; et al. An mRNA Vaccine against SARS-CoV-2 — Preliminary Report. *New England Journal of Medicine* **2020**, *383* (20), 1920-1931. DOI: 10.1056/NEJMoa2022483.
- (211) Nagura-Ikeda, M.; Imai, K.; Tabata, S.; Miyoshi, K.; Murahara, N.; Mizuno, T.; Horiuchi, M.; Kato, K.; Imoto, Y.; Iwata, M.; et al. Clinical Evaluation of Self-Collected Saliva by Quantitative Reverse Transcription-PCR (RT-qPCR), Direct RT-qPCR, Reverse Transcription–Loop-Mediated Isothermal Amplification, and a Rapid Antigen Test To Diagnose COVID-19. *Journal of Clinical Microbiology* **2020**, *58* (9), e01438-01420. DOI: 10.1128/jcm.01438-20.
- (212) Corman, V. M.; Landt, O.; Kaiser, M.; Molenkamp, R.; Meijer, A.; Chu, D. K.; Bleicker, T.; Brünink, S.; Schneider, J.; Schmidt, M. L.; et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **2020**, *25* (3), 2000045. DOI: <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.
- (213) E., K.; R., F.; P., K. Direct-RT-qPCR Detection of SARS-CoV-2 without RNA Extraction as Part of a COVID-19 Testing Strategy: From Sample to Result in One Hour. *Diagnostics* **2020**, *10* (8).
- (214) Lu, R.; Wu, X.; Wan, Z.; Li, Y.; Jin, X.; Zhang, C. A Novel Reverse Transcription Loop-Mediated Isothermal Amplification Method for Rapid Detection of SARS-CoV-2. *International Journal of Molecular Sciences* **2020**, *21* (8).
- (215) Joung, J.; Ladha, A.; Saito, M.; Kim, N.-G.; Woolley, A. E.; Segel, M.; Barretto, R. P. J.; Ranu, A.; Macrae, R. K.; Faure, G.; et al. Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing. *New England Journal of Medicine* **2020**, *383* (15), 1492-1494. DOI: 10.1056/NEJMc2026172 (accessed 2021/04/20).
- (216) Tichopad, A.; Didier, A.; Pfaffl, M. W. Inhibition of real-time RT–PCR quantification due to tissue-specific contaminants. *Molecular and Cellular Probes* **2004**, *18* (1), 45-50. DOI: <https://doi.org/10.1016/j.mcp.2003.09.001>.
- (217) Schrader, C.; Schielke, A.; Ellerbroek, L.; Johne, R. PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology* **2012**, *113* (5), 1014-1026, <https://doi.org/10.1111/j.1365-2672.2012.05384.x>. DOI: <https://doi.org/10.1111/j.1365-2672.2012.05384.x> (accessed 2021/01/15).

- (218) Foster, M. W.; Gerhardt, G.; Robitaille, L.; Plante, P.-L.; Boivin, G.; Corbeil, J.; Moseley, M. A. Targeted proteomics of human metapneumovirus in clinical samples and viral cultures. *Analytical chemistry* **2015**, *87* (20), 10247-10254.
- (219) Wolf-Yadlin, A.; Hautaniemi, S.; Lauffenburger, D. A.; White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proceedings of the National Academy of Sciences* **2007**, *104* (14), 5860-5865. DOI: 10.1073/pnas.0608638104.
- (220) Guerin, M.; Gonçalves, A.; Toiron, Y.; Baudalet, E.; Pophillat, M.; Granjeaud, S.; Fourquet, P.; Jacot, W.; Tarpin, C.; Sabatier, R.; et al. Development of parallel reaction monitoring (PRM)-based quantitative proteomics applied to HER2-Positive breast cancer. *Oncotarget* **2018**, *9* (73), 33762-33777. DOI: 10.18632/oncotarget.26031.
- (221) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; et al. Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics. *Analytical Chemistry* **2004**, *76* (13), 3556-3568. DOI: 10.1021/ac035229m.
- (222) Wu, F.-X.; Gagné, P.; Droit, A.; Poirier, G. G. Quality assessment of peptide tandem mass spectra. *BMC Bioinformatics* **2008**, *9* (S6), S13. DOI: 10.1186/1471-2105-9-s6-s13.
- (223) Gouveia, D.; Grenga, L.; Gaillard, J. C.; Gallais, F.; Bellanger, L.; Pible, O.; Armengaud, J. Shortlisting SARS-CoV-2 Peptides for Targeted Studies from Experimental Data-Dependent Acquisition Tandem Mass Spectrometry Data. *Proteomics* **2020**, *20* (14), e2000107. DOI: 10.1002/pmic.202000107 From NLM.
- (224) L., G.; F., G.; O., P.; *al., e.* Shotgun proteomics analysis of SARS-CoV-2-infected cells and how it can optimize whole viral particle antigen production for vaccines. *Emerging Microbes & Infections* **2020**, *9* (1), 1712-1721. DOI: DOI: 10.1080/22221751.2020.1791737.
- (225) Davidson, A. D.; Williamson, M. K.; Lewis, S.; *al., e.* Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Biology* **2020**, *12* (68), - 68. DOI: - 10.1186/s13073-020-00763-0.
- (226) Cardozo, K. H. M.; Lebkuchen, A.; Okai, G. G.; Schuch, R. A.; Viana, L. G.; Olive, A. N.; Lazari, C. D. S.; Fraga, A. M.; Granato, C. F. H.; Pintão, M. C. T.; et al. Establishing a mass spectrometry-based system for rapid detection of SARS-CoV-2 in large clinical sample cohorts. *Nature communications* **2020**, *11* (1), 6201-6201. DOI: 10.1038/s41467-020-19925-0 PubMed.
- (227) Ihling, C.; Tänzler, D.; Hagemann, S.; Kehlen, A.; Hüttelmaier, S.; Arlt, C.; Sinz, A. Mass Spectrometric Identification of SARS-CoV-2 Proteins from Gargle Solution Samples of COVID-19 Patients. *J Proteome Res* **2020**, *19* (11), 4389-4392. DOI: 10.1021/acs.jproteome.0c00280 From NLM.
- (228) Rivera, B.; Leyva, A.; Portela, M. M.; Moratorio, G.; Moreno, P.; Durán, R.; Lima, A. Quantitative proteomic dataset from oro- and naso-pharyngeal swabs used for COVID-19 diagnosis: Detection of viral proteins and host's biological processes altered by the infection. *Elsevier- Data in brief* **2020**, *32*, 106121. DOI: 10.1016/j.dib.2020.106121 L2 - PubMed.

- (229) Zeng, H. L.; Chen, D.; Yan, J.; Yang, Q.; Han, Q. Q.; Li, S. S.; Cheng, L. Proteomic characteristics of bronchoalveolar lavage fluid in critical COVID-19 patients. *The FEBS Journal* **2020**. DOI: 10.1111/febs.15609.
- (230) Leng, L.; Cao, R.; Ma, J.; Mou, D.; Zhu, Y.; Li, W.; Lv, L.; Gao, D.; Zhang, S.; Gong, F.; et al. Pathological features of COVID-19-associated lung injury: a preliminary proteomics report based on clinical samples. *Signal Transduct Target Ther* **2020**, 5 (1), 240. DOI: 10.1038/s41392-020-00355-9 From NLM.
- (231) Saunders, J. K.; Gaylord, D. A.; Held, N. A.; Symmonds, N.; Dupont, C. L.; Shepherd, A.; Kinkade, D. B.; Saito, M. A. METATRYP v 2.0: Metaproteomic Least Common Ancestor Analysis for Taxonomic Inference Using Specialized Sequence Assemblies—Standalone Software and Web Servers for Marine Microorganisms and Coronaviruses. *Journal of Proteome Research* **2020**, 19 (11), 4718-4729. DOI: 10.1021/acs.jproteome.0c00385.
- (232) Leng, L.; Cao, R.; Ma, J.; Mou, D.; Zhu, Y.; Li, W.; Lv, L.; Gao, D.; Zhang, S.; Gong, F.; et al. Pathological features of COVID-19-associated lung injury: a preliminary proteomics report based on clinical samples. *Signal Transduction and Targeted Therapy* **2020**, 5 (1). DOI: 10.1038/s41392-020-00355-9.
- (233) Cardozo, K. H. M.; Lebkuchen, A.; Okai, G. G.; *al., e.* Fast and low-cost detection of SARS-CoV-2 peptides by tandem mass spectrometry in clinical samples. *Research Square* **2020**. DOI: DOI: 10.21203/rs.3.rs-28883/v1.
- (234) Bankar, R.; Suvarna, K.; Ghantasala, S.; Banerjee, A.; Biswas, D.; Choudhury, M.; Palanivel, V.; Salkar, A.; Verma, A.; Singh, A.; et al. Proteomic investigation reveals dominant alterations of neutrophil degranulation and mRNA translation pathways in patients with COVID-19. *iScience* **2021**, 24 (3), 102135. DOI: 10.1016/j.isci.2021.102135 PubMed.
- (235) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology* **2015**, 33 (1), 22-24. DOI: 10.1038/nbt.3109.
- (236) Orsburn, B. C.; Jenkins, C.; Miller, S. M.; Neely, B. A.; Bumpus, N. N. In silico approach toward the identification of unique peptides from viral protein infection: Application to COVID-19. Cold Spring Harbor Laboratory: 2020.
- (237) Li, K.; Vaudel, M.; Zhang, B.; Ren, Y.; Wen, B. PDV: an integrative proteomics data viewer. *Bioinformatics* **2019**, 35 (7), 1249-1251. DOI: 10.1093/bioinformatics/bty770.
- (238) Rasche, H.; Hiltmann, S. Galactic Circos: User-friendly Circos plots within the Galaxy platform. *GigaScience* **2020**, 9 (6). DOI: 10.1093/gigascience/giaa065.
- (239) Capobianchi, M. R.; Rueca, M.; Messina, F.; Giombini, E.; Carletti, F.; Colavita, F.; Castilletti, C.; Lalle, E.; Bordi, L.; Vairo, F.; et al. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clinical Microbiology and Infection* **2020**, 26 (7), 954-956. DOI: 10.1016/j.cmi.2020.03.025.
- (240) Colavita, F.; Lapa, D.; Carletti, F.; Lalle, E.; Messina, F.; Rueca, M.; Matusali, G.; Meschi, S.; Bordi, L.; Marsella, P. Virological Characterization of the First 2 COVID-19 Patients Diagnosed in Italy: Phylogenetic Analysis, Virus Shedding Profile From Different Body Sites, and Antibody Response Kinetics. In *Open Forum Infectious Diseases*, 2020; Oxford University Press US: Vol. 7, p ofaa403.

- (241) Pozniak, Y.; Balint-Lahat, N.; Rudolph, J. D.; Lindskog, C.; Katzir, R.; Avivi, C.; Pontén, F.; Ruppin, E.; Barshack, I.; Geiger, T. System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. *Cell systems* **2016**, *2* (3), 172-184.
- (242) Yanovich, G.; Agmon, H.; Harel, M.; Sonnenblick, A.; Peretz, T.; Geiger, T. Clinical proteomics of breast cancer reveals a novel layer of breast cancer classification. *Cancer research* **2018**, *78* (20), 6001-6010.
- (243) Chen, Y.; Huang, A.; Ao, W.; Wang, Z.; Yuan, J.; Song, Q.; Wei, D.; Ye, H. Proteomic analysis of serum proteins from HIV/AIDS patients with *Talaromyces marneffe* infection by TMT labeling-based quantitative proteomics. *Clinical proteomics* **2018**, *15* (1), 40.
- (244) Messner, C. B.; Demichev, V.; Wendisch, D.; Michalick, L.; White, M.; Freiwald, A.; Textoris-Taube, K.; Vernardis, S. I.; Egger, A.-S.; Kreidl, M.; et al. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Systems* **2020**, *11* (1), 11-24.e14. DOI: <https://doi.org/10.1016/j.cels.2020.05.012>.
- (245) Wikramaratna, P. S.; Paton, R. S.; Ghafari, M.; Lourenço, J. Estimating the false-negative test probability of SARS-CoV-2 by RT-PCR. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **2020**, *25* (50), 2000568. DOI: 10.2807/1560-7917.ES.2020.25.50.2000568 PubMed.
- (246) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics : MCP* **2013**, *12* (9), 2383-2393. DOI: 10.1074/mcp.R113.027797 PubMed.
- (247) Liu, W.-K.; Xu, D.; Xu, Y.; Qiu, S.-Y.; Zhang, L.; Wu, H.-K.; Zhou, R. Protein profile of well-differentiated versus un-differentiated human bronchial/tracheal epithelial cells. *Heliyon* **2020**, *6* (6), e04243. DOI: 10.1016/j.heliyon.2020.e04243.
- (248) Marcus-Sekura, C. Process changes and their effect on process evaluation for viral clearance. *Dev Biol Stand* **1996**, *88*, 125-130. From NLM.
- (249) Chang, C.-k.; Sue, S.-C.; Yu, T.-h.; Hsieh, C.-M.; Tsai, C.-K.; Chiang, Y.-C.; Lee, S.-j.; Hsiao, H.-h.; Wu, W.-J.; Chang, W.-L. Modular organization of SARS coronavirus nucleocapsid protein. *Journal of biomedical science* **2006**, *13* (1), 59-72.
- (250) Chang, C.-k.; Hou, M.-H.; Chang, C.-F.; Hsiao, C.-D.; Huang, T.-h. The SARS coronavirus nucleocapsid protein—forms and functions. *Antiviral research* **2014**, *103*, 39-50.
- (251) de Haan, C. A.; Rottier, P. J. Molecular interactions in the assembly of coronaviruses. *Advances in virus research* **2005**, *64*, 165-230.
- (252) Mortola, E.; Roy, P. Efficient assembly and release of SARS coronavirus-like particles by a heterologous expression system. *FEBS letters* **2004**, *576* (1-2), 174-178.
- (253) Bar-On, Y. M.; Flamholz, A.; Phillips, R.; Milo, R. Science Forum: SARS-CoV-2 (COVID-19) by the numbers. *Elife* **2020**, *9*, e57309.
- (254) Hiscox, J. A.; Wurm, T.; Wilson, L.; Britton, P.; Cavanagh, D.; Brooks, G. The coronavirus infectious bronchitis virus nucleoprotein localizes to the nucleolus. *Journal of Virology* **2001**, *75* (1), 506-512.
- (255) Ravi, N.; Cortade, D. L.; Ng, E.; Wang, S. X. Diagnostics for SARS-CoV-2 detection: A comprehensive review of the FDA-EUA COVID-19 testing landscape. *Biosensors and Bioelectronics* **2020**, *165*, 112454.

- (256) Gundry, R. L.; White, M. Y.; Murray, C. I.; Kane, L. A.; Fu, Q.; Stanley, B. A.; Van Eyk, J. E. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Current protocols in molecular biology* **2010**, *90* (1), 10.25. 11-10.25. 23.
- (257) Joung, J.; Ladha, A.; Saito, M.; Segel, M.; Bruneau, R.; Huang, M.-l. W.; Kim, N.-G.; Yu, X.; Li, J.; Walker, B. D.; et al. Point-of-care testing for COVID-19 using SHERLOCK diagnostics. *medRxiv* **2020**, 2020.2005.2004.20091231. DOI: 10.1101/2020.05.04.20091231.
- (258) Dara, M.; Talebzadeh, M. CRISPR/Cas as a Potential Diagnosis Technique for COVID-19. *Avicenna journal of medical biotechnology* **2020**, *12* (3), 201-202.
- (259) Zhang, F.; Abudayyeh, O. O.; Gootenberg, J. S. A protocol for detection of COVID-19 using CRISPR diagnostics. *A protocol for detection of COVID-19 using CRISPR diagnostics* **2020**, *8*.
- (260) Gutierrez, D. B.; Gant-Branum, R. L.; Romer, C. E.; Farrow, M. A.; Allen, J. L.; Dahal, N.; Nei, Y.-W.; Codreanu, S. G.; Jordan, A. T.; Palmer, L. D. An integrated, high-throughput strategy for multiomic systems level analysis. *Journal of proteome research* **2018**, *17* (10), 3396-3408.
- (261) Lee, J.; Kim, H.; Sohn, A.; Yeo, I.; Kim, Y. Cost-Effective Automated Preparation of Serum Samples for Reproducible Quantitative Clinical Proteomics. *Journal of proteome research* **2019**, *18* (5), 2337-2345.
- (262) Chiba, T.; Marusawa, H.; Ushijima, T. Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. *Gastroenterology* **2012**, *143* (3), 550-563.
- (263) Fernandes, J. V.; Fernandes, T. A. A. d. M.; De Azevedo, J. C. V.; Cobucci, R. N. O.; De Carvalho, M. G. F.; Andrade, V. S.; De Araujo, J. M. G. Link between chronic inflammation and human papillomavirus-induced carcinogenesis. *Oncology letters* **2015**, *9* (3), 1015-1026.
- (264) Greten, F. R.; Eckmann, L.; Greten, T. F.; Park, J. M.; Li, Z.-W.; Egan, L. J.; Kagnoff, M. F.; Karin, M. IKK β links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell* **2004**, *118* (3), 285-296.
- (265) Affara, N. I.; Coussens, L. M. IKK α at the crossroads of inflammation and metastasis. *Cell* **2007**, *129* (1), 25-26.
- (266) Mangerich, A.; Knutson, C. G.; Parry, N. M.; Muthupalani, S.; Ye, W.; Prestwich, E.; Cui, L.; McFaline, J. L.; Mobley, M.; Ge, Z. Infection-induced colitis in mice causes dynamic and tissue-specific changes in stress response and DNA damage leading to colon cancer. *Proceedings of the National Academy of Sciences* **2012**, *109* (27), E1820-E1829.
- (267) Meira, L. B.; Bugni, J. M.; Green, S. L.; Lee, C.-W.; Pang, B.; Borenshtein, D.; Rickman, B. H.; Rogers, A. B.; Moroski-Erkul, C. A.; McFaline, J. L. DNA damage induced by chronic inflammation contributes to colon carcinogenesis in mice. *The Journal of clinical investigation* **2008**, *118* (7), 2516-2525.
- (268) Huang, Z.; Huang, D.; Ni, S.; Peng, Z.; Sheng, W.; Du, X. Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. *International journal of cancer* **2010**, *127* (1), 118-126.

- (269) Brown, J. R.; DuBois, R. N. COX-2: a molecular target for colorectal cancer prevention. *Journal of Clinical Oncology* **2005**, *23* (12), 2840-2855.
- (270) Jardim-Perassi, B. V.; Alexandre, P. A.; Sonehara, N. M.; de Paula-Junior, R.; Júnior, O. R.; Fukumasu, H.; Chammas, R.; Coutinho, L. L.; de Campos Zuccari, D. A. P. RNA-Seq transcriptome analysis shows anti-tumor actions of melatonin in a breast cancer xenograft model. *Scientific reports* **2019**, *9* (1), 1-13.
- (271) Jia, J.; Liu, X.; Li, L.; Lei, C.; Dong, Y.; Wu, G.; Hu, G. Transcriptional and translational relationship in environmental stress: RNAseq and ITRAQ proteomic analysis between sexually reproducing and parthenogenetic females in *Moina micrura*. *Frontiers in physiology* **2018**, *9*, 812.
- (272) Kisluk, J.; Ciborowski, M.; Niemira, M.; Kretowski, A.; Niklinski, J. Proteomics biomarkers for non-small cell lung cancer. *Journal of pharmaceutical and biomedical analysis* **2014**, *101*, 40-49.
- (273) Hegde, P. S.; White, I. R.; Debouck, C. Interplay of transcriptomics and proteomics. *Current opinion in biotechnology* **2003**, *14* (6), 647-651.
- (274) Alfaro, J. A.; Sinha, A.; Kislinger, T.; Boutros, P. C. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature methods* **2014**, *11* (11), 1107-1113.
- (275) Vasaikar, S.; Huang, C.; Wang, X.; Petyuk, V. A.; Savage, S. R.; Wen, B.; Dou, Y.; Zhang, Y.; Shi, Z.; Arshad, O. A. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **2019**, *177* (4), 1035-1049. e1019.
- (276) Wang, J.; Mouradov, D.; Wang, X.; Jorissen, R. N.; Chambers, M. C.; Zimmerman, L. J.; Vasaikar, S.; Love, C. G.; Li, S.; Lowes, K.; et al. Colorectal Cancer Cell Line Proteomes Are Representative of Primary Tumors and Predict Drug Sensitivity. *Gastroenterology* **2017**, *153* (4), 1082-1095. DOI: 10.1053/j.gastro.2017.06.008.
- (277) Erdman, S. E.; Rao, V. P.; Poutahidis, T.; Rogers, A. B.; Taylor, C. L.; Jackson, E. A.; Ge, Z.; Lee, C. W.; Schauer, D. B.; Wogan, G. N.; et al. Nitric oxide and TNF- α trigger colonic inflammation and carcinogenesis in *Helicobacter hepaticus*-infected, *Rag2*-deficient mice. *Proceedings of the National Academy of Sciences* **2009**, *106* (4), 1027. DOI: 10.1073/pnas.0812347106.
- (278) Han, Q.; Kono, T. J. Y.; Knutson, C. G.; Parry, N. M.; Seiler, C. L.; Fox, J. G.; Tannenbaum, S. R.; Tretyakova, N. Y. Multi-Omics Characterization of Inflammatory Bowel Disease-Induced Hyperplasia/Dysplasia in the *Rag2*^{-/-}/*Il10*^{-/-} Mouse Model. *International Journal of Molecular Sciences* **2021**, *22* (1). DOI: 10.3390/ijms22010364.
- (279) Erdman, S. E.; Poutahidis, T.; Tomczak, M.; Rogers, A. B.; Cormier, K.; Plank, B.; Horwitz, B. H.; Fox, J. G. CD4⁺ CD25⁺ regulatory T lymphocytes inhibit microbially induced colon cancer in *Rag2*-deficient mice. *The American journal of pathology* **2003**, *162* (2), 691-702.
- (280) Boekel, J.; Chilton, J. M.; Cooke, I. R.; Horvatovich, P. L.; Jagtap, P. D.; Käll, L.; Lehtiö, J.; Lukasse, P.; Moerland, P. D.; Griffin, T. J. Multi-omic data analysis using Galaxy. *Nat Biotechnol* **2015**, *33* (2), 137-139. DOI: 10.1038/nbt.3134 From NLM.
- (281) Chambers, M. C.; Jagtap, P. D.; Johnson, J. E.; McGowan, T.; Kumar, P.; Onsongo, G.; Guerrero, C. R.; Barsnes, H.; Vaudel, M.; Martens, L.; et al. An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res* **2017**, *77* (21), e43-e46. DOI: 10.1158/0008-5472.can-17-0331 From NLM.

- (282) Jagtap, P. D.; Johnson, J. E.; Onsongo, G.; Sadler, F. W.; Murray, K.; Wang, Y.; Shenykman, G. M.; Bandhakavi, S.; Smith, L. M.; Griffin, T. J. Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *J Proteome Res* **2014**, *13* (12), 5898-5908. DOI: 10.1021/pr500812t From NLM.
- (283) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardi, M. E.; Low, T. Y. The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nature methods* **2013**, *10* (8), 730-736.
- (284) Pertea, M.; Pertea, G. M.; Antonescu, C. M.; Chang, T.-C.; Mendell, J. T.; Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **2015**, *33* (3), 290-295.
- (285) Mehta, S.; Griffin, T. J.; Jagtap, P.; Sajulga, R.; Johnson, J.; Kumar, P. *Proteogenomics 1: Database Creation (Galaxy Training Materials)*. 2021. /training-material/topics/proteomics/tutorials/proteogenomics-dbcreation/tutorial.html (accessed 2021 Thu Apr 29).
- (286) Huang, T.; Choi, M.; Tzouros, M.; Golling, S.; Pandya, N. J.; Banfai, B.; Dunkley, T.; Vitek, O. MSstatsTMT: Statistical detection of differentially abundant proteins in experiments with isobaric labeling and multiple mixtures. *Molecular & Cellular Proteomics* **2020**, *19* (10), 1706-1723.
- (287) Gish, W.; States, D. J. Identification of protein coding regions by database similarity search. *Nature genetics* **1993**, *3* (3), 266-272.
- (288) Kumar, P. PepPointer. 2017.
- (289) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & Cellular Proteomics* **2012**, *11* (11), 1475-1488.
- (290) Maclean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; Maccoss, M. J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26* (7), 966-968. DOI: 10.1093/bioinformatics/btq054.
- (291) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **2019**, *16* (6), 509-518. DOI: 10.1038/s41592-019-0426-7.
- (292) Farinati, F.; Cardin, R.; Degan, P.; Rugge, M.; Mario, F. D.; Bonvicini, P.; Naccarato, R. Oxidative DNA damage accumulation in gastric carcinogenesis. *Gut* **1998**, *42* (3), 351-356. DOI: 10.1136/gut.42.3.351 From Nlm.
- (293) Hatziaepostolou, M.; Iliopoulos, D. Epigenetic aberrations during oncogenesis. *Cell Mol Life Sci* **2011**, *68* (10), 1681-1702. DOI: 10.1007/s00018-010-0624-z From Nlm.
- (294) Knutson, C. G.; Mangerich, A.; Zeng, Y.; Raczynski, A. R.; Liberman, R. G.; Kang, P.; Ye, W.; Prestwich, E. G.; Lu, K.; Wishnok, J. S.; et al. Chemical and cytokine features of innate immunity characterize serum and tissue profiles in inflammatory bowel disease. *Proceedings of the National Academy of Sciences* **2013**, *110* (26), E2332-E2341. DOI: 10.1073/pnas.1222669110.
- (295) Hayashi, K.-G.; Hosoe, M.; Kizaki, K.; Fujii, S.; Kanahara, H.; Takahashi, T.; Sakumoto, R. Differential gene expression profiling of endometrium during the mid-luteal phase of

- the estrous cycle between a repeat breeder (RB) and non-RB cows. *Reproductive Biology and Endocrinology* **2017**, *15* (1). DOI: 10.1186/s12958-017-0237-6.
- (296) Wynn, T. A. Cellular and molecular mechanisms of fibrosis. *J Pathol* **2008**, *214* (2), 199-210. DOI: 10.1002/path.2277 From Nlm.
- (297) Waithman, J.; Moffat, J. M.; Patterson, N. L.; van Beek, A. E.; Mintern, J. D. Antigen Presentation. In *Reference Module in Biomedical Sciences*, Elsevier, 2014.
- (298) Keller, C. W.; Kotur, M. B.; Mundt, S.; Dokalis, N.; Ligeon, L.-A.; Shah, A. M.; Prinz, M.; Becher, B.; Münz, C.; Lünemann, J. D. CYBB/NOX2 in conventional DCs controls T cell encephalitogenicity during neuroinflammation. *Autophagy* **2021**, *17* (5), 1244-1258. DOI: 10.1080/15548627.2020.1756678 PubMed.
- (299) Li, Z.; Kabir, I.; Tietelman, G.; Huan, C.; Fan, J.; Worgall, T.; Jiang, X.-C. Sphingolipid de novo biosynthesis is essential for intestine cell survival and barrier function. *Cell Death & Disease* **2018**, *9* (2), 173. DOI: 10.1038/s41419-017-0214-1.
- (300) Nardella, C.; Chen, Z.; Salmena, L.; Carracedo, A.; Alimonti, A.; Egia, A.; Carver, B.; Gerald, W.; Cordon-Cardo, C.; Pandolfi, P. P. Aberrant Rheb-mediated mTORC1 activation and Pten haploinsufficiency are cooperative oncogenic events. *Genes & Development* **2008**, *22* (16), 2172-2177. DOI: 10.1101/gad.1699608.
- (301) Chen, W.; Lu, C.; Hirota, C.; Iacucci, M.; Ghosh, S.; Gui, X. Smooth Muscle Hyperplasia/Hypertrophy is the Most Prominent Histological Change in Crohn's Fibrostenosing Bowel Strictures: A Semiquantitative Analysis by Using a Novel Histological Grading Scheme. *Journal of Crohn's and Colitis* **2017**, *11* (1), 92-104. DOI: 10.1093/ecco-jcc/jjw126 (accessed 12/30/2021).
- (302) Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. The human genome browser at UCSC. *Genome Res* **2002**, *12* (6), 996-1006. DOI: 10.1101/gr.229102 From Nlm.
- (303) Mehta, S.; Easterly, C. W.; Sajulga, R.; Millikin, R. J.; Argentini, A.; Eguinoa, I.; Martens, L.; Shortreed, M. R.; Smith, L. M.; McGowan, T.; et al. Precursor Intensity-Based Label-Free Quantification Software Tools for Proteomic and Multi-Omic Analysis within the Galaxy Platform. *Proteomes* **2020**, *8* (3), 15.
- (304) Siegel, R. L.; Miller, K. D.; Fuchs, H. E.; Jemal, A. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians* **2022**, *72* (1), 7-33, <https://doi.org/10.3322/caac.21708>. DOI: <https://doi.org/10.3322/caac.21708> (accessed 2022/09/18).
- (305) Reitsma, M. B.; Kendrick, P. J.; Ababneh, E.; Abbafati, C.; Abbasi-Kangevari, M.; Abdoli, A.; Abedi, A.; Abhilash, E. S.; Abila, D. B.; Aboyans, V.; et al. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *The Lancet* **2021**, *397* (10292), 2337-2360. DOI: [https://doi.org/10.1016/S0140-6736\(21\)01169-7](https://doi.org/10.1016/S0140-6736(21)01169-7).
- (306) Melkamu, T.; Qian, X.; Upadhyaya, P.; O'Sullivan, M. G.; Kassie, F. Lipopolysaccharide Enhances Mouse Lung Tumorigenesis: A Model for Inflammation-Driven Lung Cancer. *Veterinary Pathology* **2013**, *50* (5), 895-902. DOI: 10.1177/0300985813476061 (accessed 2022/09/15).

- (307) Ho, C.-H.; Chen, Y.-C.; Wang, J.-J.; Liao, K.-M. Incidence and relative risk for developing cancer among patients with COPD: a nationwide cohort study in Taiwan. *BMJ Open* **2017**, *7* (3), e013195. DOI: 10.1136/bmjopen-2016-013195.
- (308) Durham, A. L.; Adcock, I. M. The relationship between COPD and lung cancer. *Lung cancer (Amsterdam, Netherlands)* **2015**, *90* (2), 121-127. DOI: 10.1016/j.lungcan.2015.08.017 PubMed.
- (309) Otto, W. Lung epithelial stem cells. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **2002**, *197* (4), 527-535.
- (310) Lin, C.; Song, H.; Huang, C.; Yao, E.; Gacayan, R.; Xu, S.-M.; Chuang, P.-T. Alveolar Type II Cells Possess the Capability of Initiating Lung Tumor Development. *PLoS ONE* **2012**, *7* (12), e53817. DOI: 10.1371/journal.pone.0053817.
- (311) Xu, X.; Rock, J. R.; Lu, Y.; Futtner, C.; Schwab, B.; Guinney, J.; Hogan, B. L. M.; Onaitis, M. W. Evidence for type II cells as cells of origin of K-Ras–induced distal lung adenocarcinoma. *Proceedings of the National Academy of Sciences* **2012**, *109* (13), 4910-4915. DOI: doi:10.1073/pnas.1112499109.
- (312) Gereke, M.; Autengruber, A.; Gröbe, L.; Jeron, A.; Bruder, D.; Stegemann-Koniszewski, S. Flow cytometric isolation of primary murine type II alveolar epithelial cells for functional and molecular studies. *J Vis Exp* **2012**, (70). DOI: 10.3791/4322 From NLM.
- (313) Myers, S. A.; Rhoads, A.; Cocco, A. R.; Peckner, R.; Haber, A. L.; Schweitzer, L. D.; Krug, K.; Mani, D. R.; Clauser, K. R.; Rozenblatt-Rosen, O.; et al. Streamlined Protocol for Deep Proteomic Profiling of FAC-sorted Cells and Its Application to Freshly Isolated Murine Immune Cells. *Mol Cell Proteomics* **2019**, *18* (5), 995-1009. DOI: 10.1074/mcp.RA118.001259 From NLM.
- (314) Crapo, J. D.; Barry, B. E.; Gehr, P.; Bachofen, M.; Weibel, E. R. Cell number and cell characteristics of the normal human lung. *Am Rev Respir Dis* **1982**, *126* (2), 332-337. DOI: 10.1164/arrd.1982.126.2.332 From NLM.
- (315) Fei, F.; Qu, J.; Zhang, M.; Li, Y.; Zhang, S. S100A4 in cancer progression and metastasis: A systematic review. *Oncotarget* **2017**, *8* (42), 73219.
- (316) Ambartsumian, N.; Klingelhöfer, J.; Grigorian, M. The multifaceted S100A4 protein in cancer and inflammation. *Calcium-binding proteins of the EF-hand superfamily* **2019**, 339-365.
- (317) Newell-Price, J.; Clark, A. J.; King, P. DNA methylation and silencing of gene expression. *Trends Endocrinol Metab* **2000**, *11* (4), 142-148. DOI: 10.1016/s1043-2760(00)00248-4 From NLM.
- (318) Moarii, M.; Boeva, V.; Vert, J.-P.; Reyat, F. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **2015**, *16* (1), 873. DOI: 10.1186/s12864-015-1994-2.
- (319) Fey, D.; Matallanas, D.; Rauch, J.; Rukhlenko, O. S.; Kholodenko, B. N. The complexities and versatility of the RAS-to-ERK signalling system in normal and cancer cells. *Seminars in Cell & Developmental Biology* **2016**, *58*, 96-107. DOI: <https://doi.org/10.1016/j.semcdb.2016.06.011>.
- (320) Inoue, J.; Kishikawa, M.; Tsuda, H.; Nakajima, Y.; Asakage, T.; Inazawa, J. Identification of PDHX as a metabolic target for esophageal squamous cell carcinoma. *Cancer science* **2021**, *112* (7), 2792-2802. DOI: 10.1111/cas.14938 PubMed.

- (321) Kakumu, T.; Sato, M.; Goto, D.; Kato, T.; Yogo, N.; Hase, T.; Morise, M.; Fukui, T.; Yokoi, K.; Sekido, Y. Identification of proteasomal catalytic subunit PSMA 6 as a therapeutic target for lung cancer. *Cancer science* **2017**, *108* (4), 732-743.
- (322) Zhang, R.; Cheung, C. Y.; Seo, S.-U.; Liu, H.; Pardeshi, L.; Wong, K. H.; Chow, L. M. C.; Chau, M. P.; Wang, Y.; Lee, A. R.; et al. RUVBL1/2 Complex Regulates Pro-Inflammatory Responses in Macrophages via Regulating Histone H3K4 Trimethylation. *Frontiers in Immunology* **2021**, *12*, Original Research. DOI: 10.3389/fimmu.2021.679184.
- (323) Chen, D.-B.; Zhao, Y.-J.; Wang, X.-Y.; Liao, W.-J.; Chen, P.; Deng, K.-J.; Cong, X.; Fei, R.; Wu, X.; Shao, Q.-X.; et al. Regulatory factor X5 promotes hepatocellular carcinoma progression by transactivating tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta and suppressing apoptosis. *Chinese medical journal* **2019**, *132* (13), 1572-1581. DOI: 10.1097/CM9.0000000000000296 PubMed.
- (324) Farmer, P. B. DNA and protein adducts as markers of genotoxicity. In *Toxicol Lett*, Vol. 149; 2004; pp 3-9.
- (325) Törnqvist, M.; Fred, C.; Haglund, J.; Helleberg, H.; Paulsson, B.; Rydberg, P. Protein adducts: quantitative and qualitative aspects of their formation, analysis and applications. *J Chromatogr B Analyt Technol Biomed Life Sci* **2002**, *778* (1-2), 279-308. DOI: 10.1016/s1570-0232(02)00172-1 From NLM.
- (326) Niemelä, O.; Parkkila, S.; Ylä-Herttuala, S.; Halsted, C.; Witztum, J. L.; Lanca, A.; Israel, Y. Covalent protein adducts in the liver as a result of ethanol metabolism and lipid peroxidation. *Lab Invest* **1994**, *70* (4), 537-546. From NLM.
- (327) Zhu, Y.; Duan, X.; Qin, N.; Lv, J.; Wu, G.; Wei, F. Health risk from dietary exposure to polycyclic aromatic hydrocarbons (PAHs) in a typical high cancer incidence area in southwest China. *Science of The Total Environment* **2019**, *649*, 731-738. DOI: <https://doi.org/10.1016/j.scitotenv.2018.08.157>.
- (328) Polachova, A.; Gramblicka, T.; Parizek, O.; Sram, R. J.; Stupak, M.; Hajslova, J.; Pulkrabova, J. Estimation of human exposure to polycyclic aromatic hydrocarbons (PAHs) based on the dietary and outdoor atmospheric monitoring in the Czech Republic. *Environmental Research* **2020**, *182*, 108977. DOI: <https://doi.org/10.1016/j.envres.2019.108977>.
- (329) Kautiainen, A.; Midtvedt, T.; Törnqvist, M. Intestinal bacteria and endogenous production of malonaldehyde and alkylators in mice. *Carcinogenesis* **1993**, *14* (12), 2633-2636. DOI: 10.1093/carcin/14.12.2633.
- (330) Blondin, O.; Viau, C. Benzo(a)pyrene-blood protein adducts in wild woodchucks used as biological sentinels of environmental polycyclic aromatic hydrocarbons contamination. *Archives of Environmental Contamination and Toxicology* **1992**, *23* (3), 310-315. DOI: 10.1007/BF00216239.
- (331) Hemminki, K.; Grzybowska, E.; Chorazy, M.; Twardowska-Sauchka, K.; Sroczynski, J. W.; Putman, K. L.; Randerath, K.; Phillips, D. H.; Hewer, A.; Santella, R. M.; et al. DNA adducts in humans environmentally exposed to aromatic compounds in an industrial area of Poland. *Carcinogenesis* **1990**, *11* (7), 1229-1231. DOI: 10.1093/carcin/11.7.1229 (accessed 1/7/2021).

- (332) Bartsch, H.; Nair, J. Chronic inflammation and oxidative stress in the genesis and perpetuation of cancer: role of lipid peroxidation, DNA damage, and repair. *Langenbecks Arch Surg* **2006**, *391* (5), 499-510. DOI: 10.1007/s00423-006-0073-1 From NLM.
- (333) De Flora, S.; Izzotti, A.; Randerath, K.; Randerath, E.; Bartsch, H.; Nair, J.; Balansky, R.; van Schooten, F.; Degan, P.; Fronza, G.; et al. DNA adducts and chronic degenerative disease. Pathogenetic relevance and implications in preventive medicine. *Mutat Res* **1996**, *366* (3), 197-238. From NLM.
- (334) Nair, J.; Gansauge, F.; Beger, H.; Dolara, P.; Winde, G.; Bartsch, H. Increased etheno-DNA adducts in affected tissues of patients suffering from Crohn's disease, ulcerative colitis, and chronic pancreatitis. *Antioxid Redox Signal* **2006**, *8* (5-6), 1003-1010. DOI: 10.1089/ars.2006.8.1003 From NLM.
- (335) Setshedi, M.; Wands, J. R.; Monte, S. M. Acetaldehyde adducts in alcoholic liver disease. *Oxid Med Cell Longev* **2010**, *3* (3), 178-185. DOI: 10.4161/oxim.3.3.12288 From NLM.
- (336) Skipper, P. L.; Peng, X.; Soohoo, C. K.; Tannenbaum, S. R. Protein adducts as biomarkers of human carcinogen exposure. *Drug Metab Rev* **1994**, *26* (1-2), 111-124. DOI: 10.3109/03602539409029787 From NLM.
- (337) Troester, M. A.; Lindstrom, A. B.; Kupper, L. L.; Waidyanatha, S.; Rappaport, S. M. Stability of Hemoglobin and Albumin Adducts of Benzene Oxide and 1,4-Benzoquinone after Administration of Benzene to F344 Rats. *Toxicological Sciences* **2000**, *54* (1), 88-94. DOI: 10.1093/toxsci/54.1.88 (accessed 1/11/2021).
- (338) Rosen, C. B.; Francis, M. B. Targeting the N terminus for site-selective protein modification. *Nat Chem Biol* **2017**, *13* (7), 697-705. DOI: 10.1038/nchembio.2416 From NLM.
- (339) Sereda, T. J.; Mant, C. T.; Quinn, A. M.; Hodges, R. S. Effect of the alpha-amino group on peptide retention behaviour in reversed-phase chromatography. Determination of the pK(a) values of the alpha-amino group of 19 different N-terminal amino acid residues. *J Chromatogr* **1993**, *646* (1), 17-30. DOI: 10.1016/s0021-9673(99)87003-4 From NLM.
- (340) Carlsson, H.; Törnqvist, M. An Adductomic Approach to Identify Electrophiles In Vivo. *Basic Clin Pharmacol Toxicol* **2017**, *121 Suppl 3*, 44-54. DOI: 10.1111/bcpt.12715 From NLM.
- (341) Carlsson, H.; Törnqvist, M. Strategy for identifying unknown hemoglobin adducts using adductome LC-MS/MS data: Identification of adducts corresponding to acrylic acid, glyoxal, methylglyoxal, and 1-octen-3-one. *Food and Chemical Toxicology* **2016**, *92*, 94-103. DOI: <https://doi.org/10.1016/j.fct.2016.03.028>.
- (342) Carlsson, H.; von Stedingk, H.; Nilsson, U.; Törnqvist, M. LC-MS/MS Screening Strategy for Unknown Adducts to N-Terminal Valine in Hemoglobin Applied to Smokers and Nonsmokers. *Chemical Research in Toxicology* **2014**, *27* (12), 2062-2070. DOI: 10.1021/tx5002749.
- (343) von Stedingk, H.; Rydberg, P.; Törnqvist, M. A new modified Edman procedure for analysis of N-terminal valine adducts in hemoglobin by LC-MS/MS. *J Chromatogr B Analyt Technol Biomed Life Sci* **2010**, *878* (27), 2483-2490. DOI: 10.1016/j.jchromb.2010.03.034 From NLM.

- (344) Carlsson, H.; Motwani, H. V.; Osterman Golkar, S.; Törnqvist, M. Characterization of a Hemoglobin Adduct from Ethyl Vinyl Ketone Detected in Human Blood Samples. *Chemical research in toxicology* **2015**, *28* (11), 2120-2129. DOI: 10.1021/acs.chemrestox.5b00287 PubMed.
- (345) Degner, A.; Carlsson, H.; Karlsson, I.; Eriksson, J.; Pujari, S. S.; Tretyakova, N. Y.; Törnqvist, M. Discovery of Novel N-(4-Hydroxybenzyl)valine Hemoglobin Adducts in Human Blood. *Chem Res Toxicol* **2018**, *31* (12), 1305-1314. DOI: 10.1021/acs.chemrestox.8b00173 From NLM.
- (346) Bolton, J. L. Quinone Methide Bioactivation Pathway: Contribution to Toxicity and/or Cytoprotection? *Current organic chemistry* **2014**, *18* (1), 61-69. DOI: 10.2174/138527281801140121123046 PubMed.
- (347) Elgawish, M. S.; Kishikawa, N.; Helal, M. A.; Ohyama, K.; Kuroda, N. Molecular modeling and spectroscopic study of quinone–protein adducts: insight into toxicity, selectivity, and reversibility. *Toxicology Research* **2015**, *4* (4), 843-847, 10.1039/C5TX00098J. DOI: 10.1039/C5TX00098J.
- (348) Gaikwad, N. W.; Bodell, W. J. Formation of DNA adducts by microsomal and peroxidase activation of p-cresol: role of quinone methide in DNA adduct formation. *Chem Biol Interact* **2001**, *138* (3), 217-229. DOI: 10.1016/S0009-2797(01)00274-5 From NLM.
- (349) Matthews, R. G.; Massey, V.; Sweeley, C. C. Identification of p-hydroxybenzaldehyde as the ligand in the green form of old yellow enzyme. *The Journal of biological chemistry* **1975**, *250* (24), 9294-9298. PubMed.
- (350) Isom, A. L.; Barnes, S.; Wilson, L.; Kirk, M.; Coward, L.; Darley-Usmar, V. Modification of Cytochrome c by 4-hydroxy- 2-nonenal: Evidence for histidine, lysine, and arginine-aldehyde adducts. *Journal of the American Society for Mass Spectrometry* **2004**, *15* (8), 1136-1147. DOI: 10.1016/j.jasms.2004.03.013.
- (351) Segerbäck, D. Alkylation of DNA and hemoglobin in the mouse following exposure to ethene and ethene oxide. *Chem Biol Interact* **1983**, *45* (2), 139-151. DOI: 10.1016/0009-2797(83)90064-9 From NLM.
- (352) Grigoryan, H.; Edmands, W.; Lu, S. S.; Yano, Y.; Regazzoni, L.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Analytical chemistry* **2016**, *88* (21), 10504-10512. DOI: 10.1021/acs.analchem.6b02553 PubMed.
- (353) Lu, S. S.; Grigoryan, H.; Edmands, W. M.; Hu, W.; Iavarone, A. T.; Hubbard, A.; Rothman, N.; Vermeulen, R.; Lan, Q.; Rappaport, S. M. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environ Sci Technol* **2017**, *51* (1), 46-57. DOI: 10.1021/acs.est.6b03955 From NLM.
- (354) Rappaport, S. M.; Li, H.; Grigoryan, H.; Funk, W. E.; Williams, E. R. Adductomics: characterizing exposures to reactive electrophiles. *Toxicol Lett* **2012**, *213* (1), 83-90. DOI: 10.1016/j.toxlet.2011.04.002 From NLM.
- (355) Michaelson-Richie, E. D.; Loeber, R. L.; Codreanu, S. G.; Ming, X.; Liebler, D. C.; Campbell, C.; Tretyakova, N. Y. DNA-protein cross-linking by 1,2,3,4-diepoxybutane. *Journal of proteome research* **2010**, *9* (9), 4356-4367. DOI: 10.1021/pr1000835 PubMed.

- (356) Loeber, R. L.; Michaelson-Richie, E. D.; Codreanu, S. G.; Liebler, D. C.; Campbell, C. R.; Tretyakova, N. Y. Proteomic analysis of DNA-protein cross-linking by antitumor nitrogen mustards. *Chem Res Toxicol* **2009**, *22* (6), 1151-1162. DOI: 10.1021/tx900078y From NLM.
- (357) Groehler, A. t.; Villalta, P. W.; Campbell, C.; Tretyakova, N. Covalent DNA-Protein Cross-Linking by Phosphoramidate Mustard and Nornitrogen Mustard in Human Cells. *Chem Res Toxicol* **2016**, *29* (2), 190-202. DOI: 10.1021/acs.chemrestox.5b00430 From NLM.
- (358) Ming, X.; Groehler, A.; Michaelson-Richie, E. D.; Villalta, P. W.; Campbell, C.; Tretyakova, N. Y. Mass Spectrometry Based Proteomics Study of Cisplatin-Induced DNA-Protein Cross-Linking in Human Fibrosarcoma (HT1080) Cells. *Chemical research in toxicology* **2017**, *30* (4), 980-995. DOI: 10.1021/acs.chemrestox.6b00389 PubMed.
- (359) Cai, T.; Bellamri, M.; Ming, X.; Koh, W.-P.; Yu, M. C.; Turesky, R. J. Quantification of Hemoglobin and White Blood Cell DNA Adducts of the Tobacco Carcinogens 2-Amino-9H-pyrido[2,3-b]indole and 4-Aminobiphenyl Formed in Humans by Nanoflow Liquid Chromatography/Ion Trap Multistage Mass Spectrometry. *Chemical Research in Toxicology* **2017**, *30* (6), 1333-1343. DOI: 10.1021/acs.chemrestox.7b00072.
- (360) Pathak, K. V.; Chiu, T. L.; Amin, E. A.; Turesky, R. J. Methemoglobin Formation and Characterization of Hemoglobin Adducts of Carcinogenic Aromatic Amines and Heterocyclic Aromatic Amines. *Chem Res Toxicol* **2016**, *29* (3), 255-269. DOI: 10.1021/acs.chemrestox.5b00418 From NLM.
- (361) Yoshitake, J.; Shibata, T.; Shimayama, C.; Uchida, K. 2-Alkenal modification of hemoglobin: Identification of a novel hemoglobin-specific alkanolic acid-histidine adduct. *Redox Biol* **2019**, *23*, 101115. DOI: 10.1016/j.redox.2019.101115 From NLM.
- (362) Huang, C.; Liu, Y.; Rokita, S. E. Targeting duplex DNA with the reversible reactivity of quinone methides. *Signal Transduction and Targeted Therapy* **2016**, *1* (1), 16009. DOI: 10.1038/sigtrans.2016.9.
- (363) Rokita, S. E.; Yang, J.; Pande, P.; Greenberg, W. A. Quinone Methide Alkylation of Deoxycytidine. *The Journal of Organic Chemistry* **1997**, *62* (9), 3010-3012. DOI: 10.1021/jo9700336.
- (364) Kang, H.; Tolbert, T. J.; Schöneich, C. Photoinduced Tyrosine Side Chain Fragmentation in IgG4-Fc: Mechanisms and Solvent Isotope Effects. *Mol Pharm* **2019**, *16* (1), 258-272. DOI: 10.1021/acs.molpharmaceut.8b00979 From NLM.
- (365) Hodge, K.; Have, S. T.; Hutton, L.; Lamond, A. I. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *Journal of Proteomics* **2013**, *88*, 92-103. DOI: <https://doi.org/10.1016/j.jprot.2013.02.023>.
- (366) Wingfield, P. T. N-Terminal Methionine Processing. *Current protocols in protein science* **2017**, *88*, 6.14.11-16.14.13. DOI: 10.1002/cpps.29 PubMed.
- (367) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom* **2016**, *27* (11), 1719-1727. DOI: 10.1007/s13361-016-1460-7 From NLM.
- (368) Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; MacLean, B.; MacCoss, M. J. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom Rev* **2020**, *39* (3), 229-244. DOI: 10.1002/mas.21540 From NLM.

- (369) Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **2017**, *33* (22), 3645-3647. DOI: 10.1093/bioinformatics/btx469 From NLM.
- (370) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research* **2012**, *40* (Web Server issue), W537-W541. DOI: 10.1093/nar/gks375 PubMed.
- (371) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* **2005**, *33* (Web Server issue), W368-371. DOI: 10.1093/nar/gki464 From NLM.
- (372) Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins* **2006**, *63* (4), 928-938. DOI: 10.1002/prot.20922 From NLM.
- (373) Xu, Y.; Zheng, Y.; Fan, J. S.; Yang, D. A new strategy for structure determination of large proteins in solution without deuteration. *Nat Methods* **2006**, *3* (11), 931-937. DOI: 10.1038/nmeth938 From NLM.
- (374) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. In *Nat Struct Biol*, Vol. 10; 2003; p 980.
- (375) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **2007**, *372* (3), 774-797. DOI: 10.1016/j.jmb.2007.05.022 From NLM.
- (376) Tien, M. Z.; Meyer, A. G.; Sydykova, D. K.; Spielman, S. J.; Wilke, C. O. Maximum allowed solvent accessibilities of residues in proteins. *PloS one* **2013**, *8* (11), e80635.
- (377) Deeyaa, B. D.; Rokita, S. E. Migratory ability of quinone methide-generating acridine conjugates in DNA. *Organic & Biomolecular Chemistry* **2020**, *18* (8), 1671-1678.
- (378) Gallage, N. J.; Møller, B. L. Vanilla: the most popular flavour. In *Biotechnology of natural products*, Springer, 2018; pp 3-24.
- (379) von Stedingk, H.; Vikström, A. C.; Rydberg, P.; Pedersen, M.; Nielsen, J. K. S.; Segerbäck, D.; Knudsen, L. E.; Törnqvist, M. Analysis of Hemoglobin Adducts from Acrylamide, Glycidamide, and Ethylene Oxide in Paired Mother/Cord Blood Samples from Denmark. *Chemical Research in Toxicology* **2011**, *24* (11), 1957-1965. DOI: 10.1021/tx200284u.
- (380) Ndreu, L.; Erber, L. N.; Törnqvist, M.; Tretyakova, N. Y.; Karlsson, I. Characterizing Adduct Formation of Electrophilic Skin Allergens with Human Serum Albumin and Hemoglobin. *Chemical Research in Toxicology* **2020**, *33* (10), 2623-2636. DOI: 10.1021/acs.chemrestox.0c00271.
- (381) de Prost, Y.; Paquez, F.; Touraine, R. Dinitrochlorobenzene treatment of alopecia areata. *Arch Dermatol* **1982**, *118* (8), 542-545. DOI: 10.1001/archderm.118.8.542 From NLM.
- (382) Aasa, J.; Abramsson-Zetterberg, L.; Carlsson, H.; Törnqvist, M. The genotoxic potency of glycidol established from micronucleus frequency and hemoglobin adduct levels in mice. *Food and Chemical Toxicology* **2017**, *100*, 168-174.
- (383) Favinha, A. G.; Barreiro, D. S.; Martins, J. N.; O'Toole, P.; Pauleta, S. R. Acrylamide-hemoglobin adduct: A spectroscopic study. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2020**, *241*, 118644.

- (384) Kassa, T.; Strader, M. B.; Nakagawa, A.; Zapol, W. M.; Alayash, A. I. Targeting β Cys93 in hemoglobin S with an antisickling agent possessing dual allosteric and antioxidant effects. *Metallomics* **2017**, *9* (9), 1260-1270.
- (385) Miranda, J. Highly reactive cysteine residues in rodent hemoglobins. *Biochemical and biophysical research communications* **2000**, *275* (2), 517-523.
- (386) Gries, W.; Leng, G. Analytical determination of specific 4, 4'-methylene diphenyl diisocyanate hemoglobin adducts in human blood. *Analytical and bioanalytical chemistry* **2013**, *405* (23), 7205-7213.
- (387) Gould, N. S.; Evans, P.; Martínez-Acedo, P.; Marino, S. M.; Gladyshev, V. N.; Carroll, K. S.; Ischiropoulos, H. Site-specific proteomic mapping identifies selectively modified regulatory cysteine residues in functionally distinct protein networks. *Chemistry & biology* **2015**, *22* (7), 965-975.
- (388) Montero, E. I.; Benedetti, B. T.; Mangrum, J. B.; Oehlsen, M. J.; Qu, Y.; Farrell, N. P. Pre-association of polynuclear platinum anticancer agents on a protein, human serum albumin. Implications for drug design. *Dalton Transactions* **2007**, (43), 4938-4942, 10.1039/B708433C. DOI: 10.1039/B708433C.
- (389) Peter, M. G. Chemical modifications of biopolymers by quinones and quinone methides. *Angewandte Chemie International Edition in English* **1989**, *28* (5), 555-570.
- (390) Thompson, D. C.; Thompson, J. A.; Sugumaran, M.; Moldéus, P. Biological and toxicological consequences of quinone methide formation. *Chemico-biological interactions* **1993**, *86* (2), 129-162.
- (391) Toteva, M. M.; Moran, M.; Amyes, T. L.; Richard, J. P. Substituent Effects on Carbocation Stability: The p KR for p-Quinone Methide. *Journal of the American Chemical Society* **2003**, *125* (29), 8814-8819.
- (392) Podstolski, A.; Havkin-Frenkel, D.; Malinowski, J.; Blount, J. W.; Kourteva, G.; Dixon, R. A. Unusual 4-hydroxybenzaldehyde synthase activity from tissue cultures of the vanilla orchid *Vanilla planifolia*. *Phytochemistry* **2002**, *61* (6), 611-620.
- (393) Yan, Z.; Zhong, H. M.; Maher, N.; Torres, R.; Leo, G. C.; Caldwell, G. W.; Huebert, N. Bioactivation of 4-methylphenol (p-cresol) via cytochrome P450-mediated aromatic oxidation in human liver microsomes. *Drug metabolism and disposition* **2005**, *33* (12), 1867-1876.
- (394) Rajakovich, L. J.; Balskus, E. P. Metabolic functions of the human gut microbiota: the role of metalloenzymes. *Natural product reports* **2019**, *36* (4), 593-625.
- (395) Ross, F. E.; Zamborelli, T.; Herman, A. C.; Yeh, C.-H.; Tedeschi, N. I.; Luedke, E. S. Detection of acetylated lysine residues using sequencing by Edman degradation and mass spectrometry. In *Techniques in protein chemistry*, Vol. 7; Elsevier, 1996; pp 201-208.
- (396) Steinke, L.; Cook, R. G. Identification of phosphorylation sites by Edman degradation. *Methods Mol Biol* **2003**, *211*, 301-307. DOI: 10.1385/1-59259-342-9:301 From NLM.
- (397) Mehta, S.; Carvalho, V. M.; Rajczewski, A. T.; Pible, O.; Grüning, B. A.; Johnson, J. E.; Wagner, R.; Armengaud, J.; Griffin, T. J.; Jagtap, P. D. Catching the Wave: Detecting Strain-Specific SARS-CoV-2 Peptides in Clinical Samples Collected during Infection Waves from Diverse Geographical Locations. *Viruses* **2022**, *14* (10), 2205.
- (398) Chen, Y. J.; Roumeliotis, T. I.; Chang, Y. H.; Chen, C. T.; Han, C. L.; Lin, M. H.; Chen, H. W.; Chang, G. C.; Chang, Y. L.; Wu, C. T.; et al. Proteogenomics of Non-smoking Lung

Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell* **2020**, *182* (1), 226-244.e217. DOI: 10.1016/j.cell.2020.06.012 From NLM.
(399) van Meerloo, J.; Kaspers, G. J.; Cloos, J. Cell sensitivity assays: the MTT assay. *Methods Mol Biol* **2011**, *731*, 237-245. DOI: 10.1007/978-1-61779-080-5_20 From NLM.