proteomes

MDPI

Article

Quantitative Proteogenomic Characterization of Inflamed Murine Colon Tissue Using an Integrated Discovery, Verification, and Validation Proteogenomic Workflow

Andrew T. Rajczewski [1], Qiyuan Han [1], Subina Mehta [1], Praveen Kumar [1], Pratik D. Jagtap [1], Charles G. Knutson [2], James G. Fox [2], Natalia Y. Tretyakova [3] and Timothy J. Griffin [1,*]

[1] Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN 55455, USA; rajcz001@umn.edu (A.T.R.); hanxx963@umn.edu (Q.H.); smehta@umn.edu (S.M.); pravs3683@gmail.com (P.K.); pjagtap@umn.edu (P.D.J.)
[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; charlie.knutson@novartis.com (C.G.K.); jgfox@mit.edu (J.G.F.)
[3] Department of Medicinal Chemistry, the Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, USA; trety001@umn.edu
* Correspondence: tgriffin@umn.edu

Abstract: Chronic inflammation of the colon causes genomic and/or transcriptomic events, which can lead to expression of non-canonical protein sequences contributing to oncogenesis. To better understand these mechanisms, *Rag2*$^{-/-}$*Il10*$^{-/-}$ mice were infected with *Helicobacter hepaticus* to induce chronic inflammation of the cecum and the colon. Transcriptomic data from harvested proximal colon samples were used to generate a customized FASTA database containing non-canonical protein sequences. Using a proteogenomic approach, mass spectrometry data for proximal colon proteins were searched against this custom FASTA database using the Galaxy for Proteomics (Galaxy-P) platform. In addition to the increased abundance in inflammatory response proteins, we also discovered several non-canonical peptide sequences derived from unique proteoforms. We confirmed the veracity of these novel sequences using an automated bioinformatics verification workflow with targeted MS-based assays for peptide validation. Our bioinformatics discovery workflow identified 235 putative non-canonical peptide sequences, of which 58 were verified with high confidence and 39 were validated in targeted proteomics assays. This study provides insights into challenges faced when identifying non-canonical peptides using a proteogenomics approach and demonstrates an integrated workflow addressing these challenges. Our bioinformatic discovery and verification workflow is publicly available and accessible via the Galaxy platform and should be valuable in non-canonical peptide identification using proteogenomics.

Keywords: inflammation; proteogenomics; bioinformatics; colon cancer

Citation: Rajczewski, A.T.; Han, Q.; Mehta, S.; Kumar, P.; Jagtap, P.D.; Knutson, C.G.; Fox, J.G.; Tretyakova, N.Y.; Griffin, T.J. Quantitative Proteogenomic Characterization of Inflamed Murine Colon Tissue Using an Integrated Discovery, Verification, and Validation Proteogenomic Workflow. *Proteomes* **2022**, *10*, 11. https://doi.org/10.3390/proteomes10020011

Academic Editor: Yehia Mechref

Received: 29 January 2022
Accepted: 7 April 2022
Published: 14 April 2022

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 [1,2]. During inflammation, continued release of regulatory cytokines which serve to mediate the immune response promotes tumorigenesis [3] and eventual metastasis [4]. 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 [5,6]. The full picture of molecular changes which occur during chronic colon inflammation is of interest to advance our understanding of colorectal cancer etiology [1], as well as to seek opportunities for its diagnosis [7] and identification of therapeutic targets for its treatment [8].

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 cancer [9,10]. However, RNA-Seq is only able to assess the state of the transcriptome, which often does not match the expressed proteins (the proteome) associated with a specific tissue or disease state [11]. By contrast, MS-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 treatment [12].

In conventional "bottom-up" proteomics, MS data are searched against a reference FASTA database containing protein sequences encoded in canonical gene sequences for the organism of interest, thereby 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 sequences [13]. These non-canonical sequences are captured in RNA-Seq analyses, which detect and sequence all transcripts including those that may give rise to novel protein products.

Proteogenomics is a multi-omics approach which combines the comprehensive nature of RNA-Seq with the ability of MS-based proteomics to directly confirm the translation of these products into expressed proteins with potential functional implications, creating a more complete molecular picture of phenotypes as compared with a single omics technology [14,15]. Proteogenomics uses RNA-Seq data to generate an expanded protein sequence FASTA database, which can be used to confirm the expression of proteoforms [16] containing both canonical and novel non-canonical peptide sequences. Although proteogenomics has been shown to be a powerful approach for studying cancer [15,17], potential false-positive matches to non-canonical sequences remains a concern [18], requiring methods to verify the accuracy of PSMs using bioinformatic and/or analytic approaches. To aid in analysis, these assorted bioinformatics processes can be combined into simple workflows for automated, streamlined proteogenomic analyses [19].

In this study, we developed and utilized novel proteogenomic workflows to analyze chronic inflammation in proximal colon tissues in a mouse model of inflammatory bowel disease (IBD). Genetically engineered *Rag2*$^{-/-}$*Il10*$^{-/-}$ mice have been used in previous studies as models of chronic inflammation [20,21], as animals with these mutations develop chronic colon inflammation when subjected to bacterial infection [22]. *Rag2*$^{-/-}$*Il10*$^{-/-}$ mice were subjected to infection with *Helicobacter hepaticus* and allowed to develop chronic colon inflammation as described previously in Mangerich et al. [5], after which proximal colon tissues were harvested and proteins were isolated based off previously established protocols [23] for LC–MS analysis. Using the Galaxy for Proteomics (Galaxy-P) software suite, [24] we utilized two automated computational workflows to generate and refine [25] 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.

2. Materials and Methods

2.1. Materials

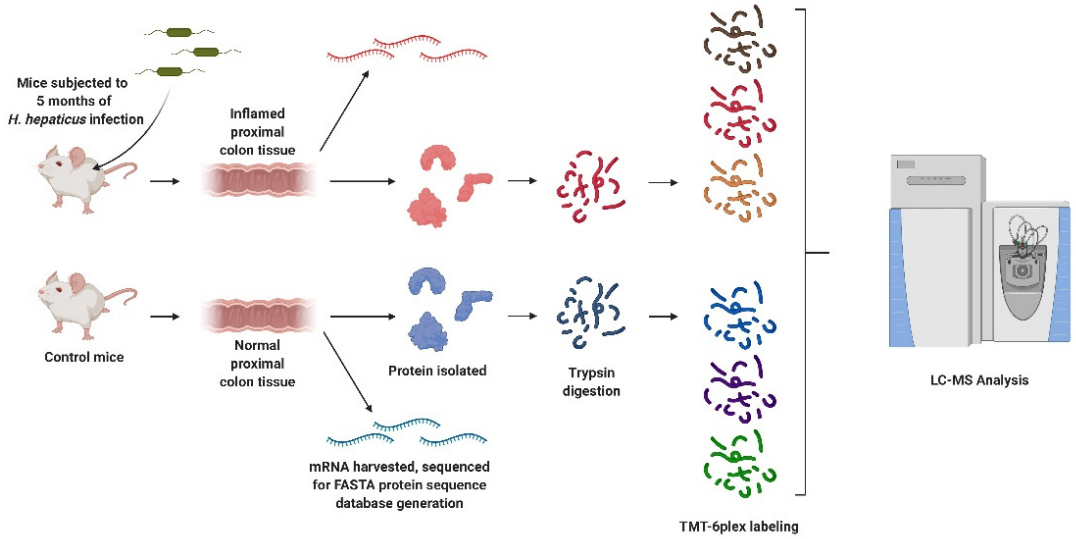
Proximal colon tissues were obtained from a previous study [5]. Triethylammonium bicarbonate (TEAB), urea, aprotinin, phenylmethanesulfonylfluoride (PMSF), 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), dithiothreitol (DTT) and iodoacetamide (IAA) were obtained from Millipore Sigma (Burlington, MA, USA). Trypsin was purchased from Promega Corporation (Madison, WI, USA). Formic acid was purchased from Honeywell Fluka (Mexico City, Mexico). Acetonitrile, water, and LTQ ESI Positive Ion Calibration Solution were obtained from Thermo Fisher Scientific (Waltham, MA, USA). Anhydrous acetonitrile was obtained from Glen Research (Sterling, VA, USA).

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

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

2.2. 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 (Scheme 1) [5].



Scheme 1. Outline of the experimental procedure. *Rag2*$^{-/-}$*Il10*$^{-/-}$ 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 (accessed on 10 August 2021).

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 S1). These samples were placed in individual Eppendorf tubes containing 100 µL of lysis buffer (25 mM TEAB, 8 M urea, 1 mM PMSF, and 2.5 µg/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 s over ice to lyse the cells; following lysis, samples were centrifuged at 15,000 rpm at 4 °C for 15 min, after which the protein content was measured via Pierce BCA Assay. From each sample, 100 µg aliquots of protein were added to individual Pall Nanosep 10 K spin columns. The lysis buffer was then removed via centrifugation at 14,000× *g* for 5 min, followed by the addition of 100 µL 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 µL of dilution buffer. The proteins were then reduced via the addition of 20 µL of DTT in the dilution buffer, followed by incubation at 55 °C for one hour. Samples were then alkylated with the addition of 10 µL of 375 mM IAA to the spin columns, followed by a 30-min incubation in the dark at room temperature. After alkylation, samples were then washed with a further three iterations of centrifugation and the addition of 100 µL of dilution buffer. Samples were finally reconstituted with 50 µL of dilution buffer, to which 4 µg of trypsin was added, and incubated at 37 °C overnight. Following incubation, peptide samples were isolated by spinning the samples through the column filters. A further 50 µL 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 a peptide colorimetric assay; 10 µg of peptides from each sample were then aliquoted into fresh vials and dried overnight under vacuum.

2.3. 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 100 mM 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 labeling reagent (Table S1). 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; 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. For each of six samples, eight HPLC fractions were collected, dried down under reduced vacuum, and reconstituted in 10 µL water containing 0.1% formic acid.

The eight fractionated peptide samples were analyzed on an Orbitrap Fusion Tribrid Mass Spectrometer interfaced with an Ultimate 3000 UHPLC. The Fusion LC–MS was calibrated in positive mode using LTQ ESI Positive Ion Calibration Solution. The UHPLC was run in nanoflow mode with a reverse-phase nanoLC column (35 cm × 250 µm) packed with 5 µm diameter Luna C18 resin. Samples were run on a 90-min 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 mode using a Top12 Full MS/dd-MS2 experiment with an expected chromatographic peak FWHM of 15 s. In the full scan mode, resolution was 70,000 with an AGC target of 1×10^6, 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 5×10^4, maximum IT of 50 ms, an isolation window of 2.0 m/z, an exclusion time of 30 s, and a normalized collision energy of 30. Data were collected in the centroid mode.

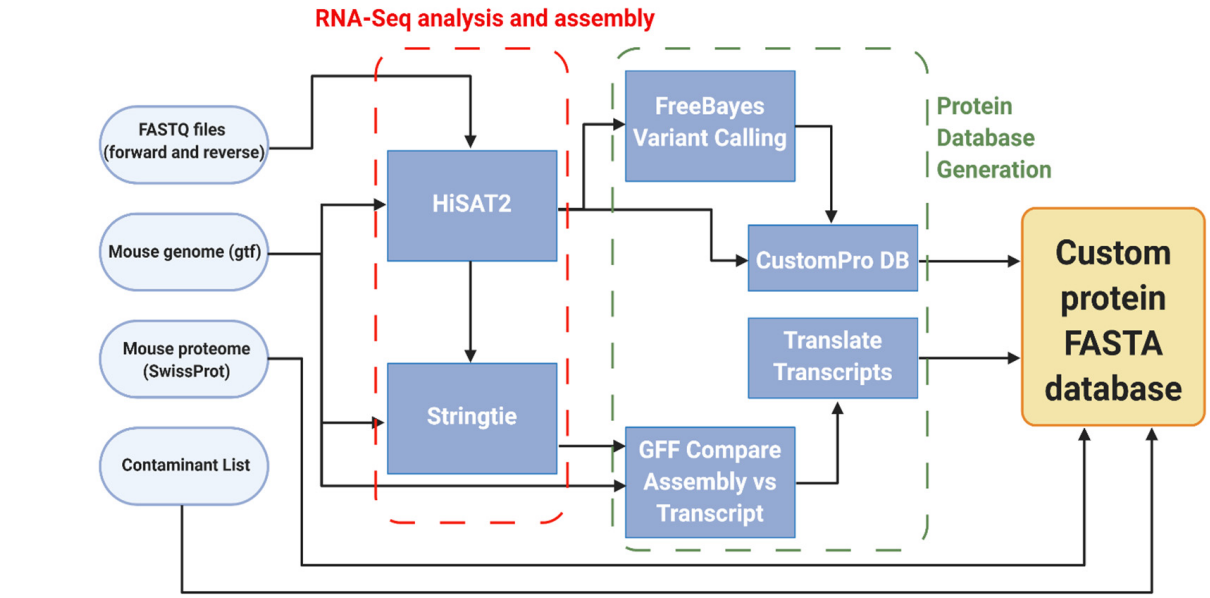
2.4. Database Construction

Computational work was performed using proteogenomics workflows and tools in the Galaxy for Proteomics (Galaxy-P) suite [26,27] as well as in Proteome Discoverer v2.2 (Thermo Fisher Scientific (Waltham, MA, USA)).

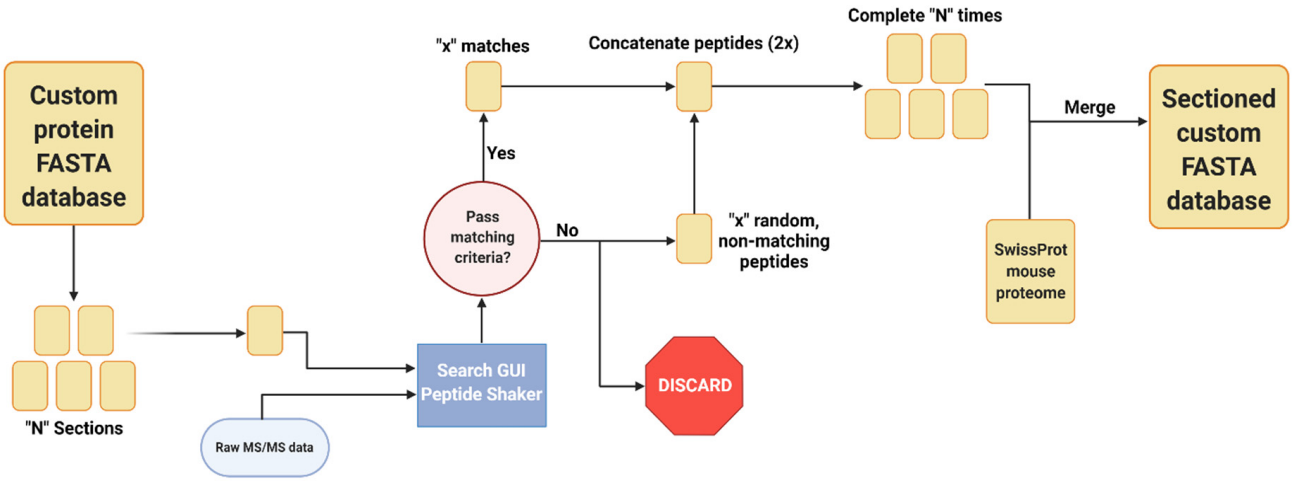
Raw RNA sequencing data were acquired from proximal colon samples of six additional mice from the colon inflammation study (QIYUAN), including three control and three inflamed samples (Figure 1a).

Sequencing data were collected at the University of Minnesota Genomics Center on an Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) sequencer run in high output mode using 50 bp paired end reads. These data were uploaded into Galaxy-P and used as an input for an integrated workflow [26] 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 HISAT2 [28] (v2.1.0, Kim lab, UT Southwestern, Dallas, TX, USA); this was then used to create a list of genetic variants using the Free Bayes (v1.1.0.46-0, Garrison lab, University of Tennessee Health Science Center, Memphis, TN, USA) Bayesian genetic variant detector [29]. This file was then utilized by the CustomProDB (v1.16.1.0, Zhang lab, Baylor College of Medicine, Houston, TX,

USA) tool [30] to create FASTA sequences of the mapped indel, single amino acid variants, and alternatively spliced sequences identified. These variants were then concatenated together with the canonical murine Uniprot FASTA database and a list of common mass spectrometry contaminants [31] as a custom RNA-Seq-based database. This workflow also used StringTie [32] (v1.3.3.1, Center for Computational Biology at Johns Hopkins University, Baltimore, MD, USA) to create an assembled gene transfer format file which was used to create a set of genomic coordinates complementary to the RNA-Seq FASTA database used in downstream applications [33], and effective for annotating other types of non-canonical transcripts not handled by CustomProDB.

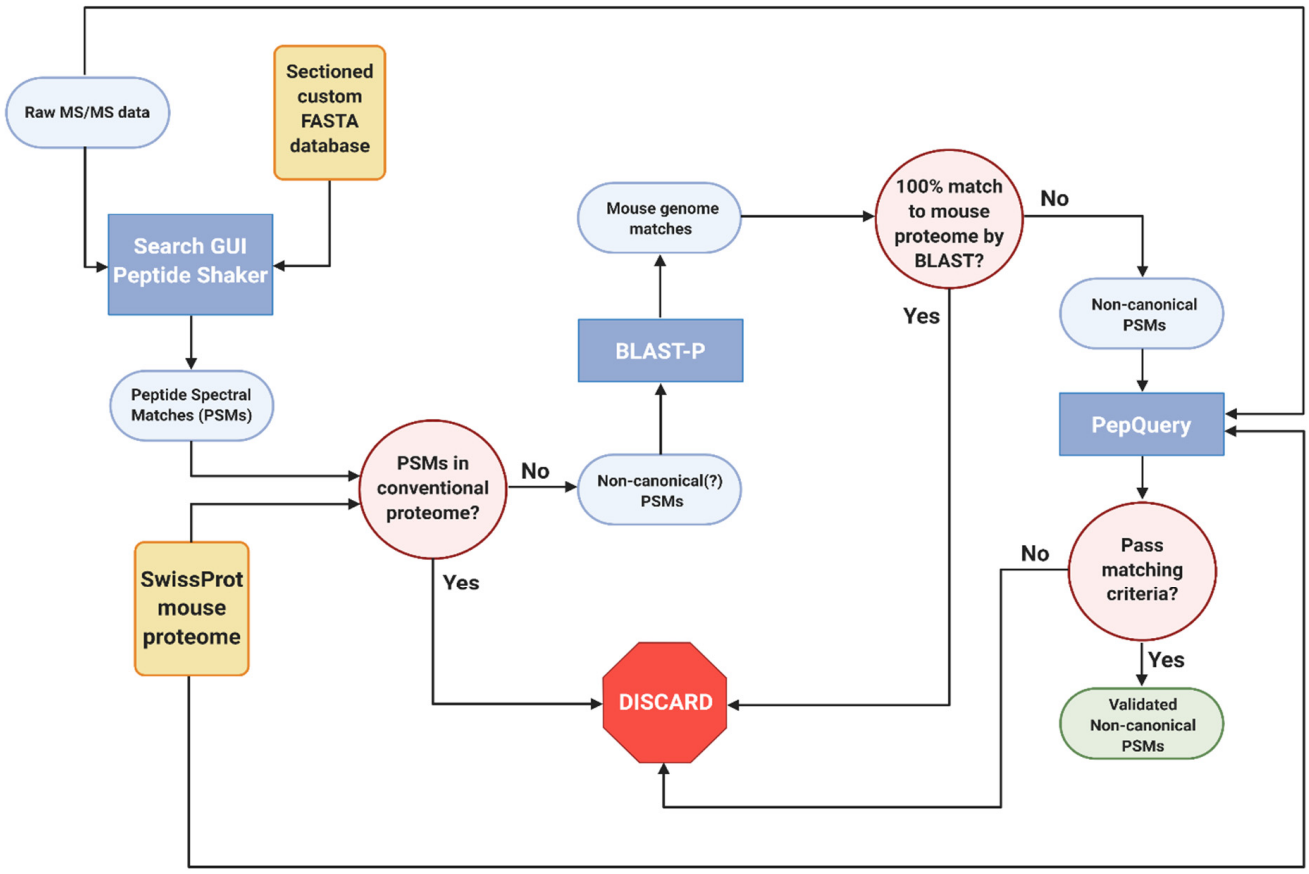


(**a**)



(**b**)

Figure 1. *Cont.*



(**c**)

Figure 1. 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.c.com, accessed on 10 August 2021.

2.5. Database Sectioning

The custom protein FASTA database was matched to MS/MS data to generate PSMs using a sectioning workflow created by Kumar et al. [25] (Figure 1b), 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 was used to search against the raw mass spectrometry data of the proximal colon samples using Search GUI [34] (v3.3.3.0, CompOmics, VIB-UGent Center for Medical Biotechnology at Ghent University, Ghent, Belgium), with N-terminal and lysine TMT-6 labeling, as well as cysteine carbamidomethylation being set as static modifications, while 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 PeptideShaker [35] (v1.16.4, CompOmics, VIB-UGent Center for Medical Biotechnology at Ghent University, Ghent, Belgium) to identify 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, which were then in turn concatenated together with the murine UniprotKB and contaminant sequence FASTA databases with redundant proteins removed.

2.6. Differential Abundance Proteomic and Proteogenomic Analysis

Raw mass spectrometry files were analyzed using Proteome Discoverer v2.2 (Thermo Fisher Scientific, Waltham, MA, USA) 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 MSstatsTMT (Vitek lab, Northeastern University, Boston, MA, USA) [36] using the "mstats" normalization algorithm. Gene ontology analyses were performed using the g-profiler package (Vilo lab, University of Tartu, Tartu, Estonia) [37], using an FDR cutoff of 0.05.

2.7. Identification, Verification and Validation of Non-Canonical Peptides

Given the large numbers of proteins, the annotation of non-canonical peptides is more efficiently performed using an automated workflow in the Galaxy-P platform (Figure 1c). 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 were 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) [38]; these results were filtered to look for those search results which had imperfect sequence alignments due to sequence substitutions or gaps in the sequence [27]. The genomic coordinates of these peptides were then determined using the PepPointer tool [39] 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 PepQuery [40] tool with unrestricted modification search mode and amino acid substitution mode engaged. 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 TMT-6 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 residues.

2.8. Validation and Quantitation of Non-Canonical Peptides

Peptides verified using PepQuery were further validated by targeted mass spectrometry analyses [41] 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 S2). 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 (15 cm × 250 µm). The Q-Exactive was calibrated in positive mode using LTQ ESI Positive Ion Calibration Solution. Samples were run on a 90-min 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. 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. In the full MS, resolution was 70,000 with

an AGC target of 3×10^6, a maximum IT of 200 ms, and a scan range of 400 to 1600 m/z. Parallel reaction monitoring experiments were conducted at a 35,000 resolution, an AGC target of 2×10^5, maximum IT of 100 ms, an isolation window of 4.0 m/z, an exclusion time of 30 s, and a normalized collision energy of 35. The resulting spectra were then analyzed in Skyline [42] against a spectral library of non-canonical peptides generated using Prosit [43]. 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.

For comparison of differential abundance levels of non-canonical peptides with their complementary mRNA levels, the original RNA-Seq data were run through a workflow in the Galaxy-P platform to perform differential transcriptomic analysis. Briefly, paired-end raw FASTQ files were cleaned up using Trimmomatic [44] to remove sequencing adaptors and aligned to the GRCm38 mm10 genome using HiSat2; the resulting BAM files were then assembled and quantified using Stringtie. The resulting transcript counts were then subjected to differential analysis using edgeR [45].

3. Results

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 Scheme 1) [21]. Each of these datasets 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 1a). 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 increased the likelihood of false positive PSMs while decreasing overall sensitivity for true positive PSMs [46], a sectioning workflow [25] was utilized to create a reduced RNA-Seq-based FASTA database (Figure 1b). 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.

3.2. 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 the 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 these identified protein groups corresponded to annotated proteins containing non-canonical sequences such as amino acid substitutions; 386 identified protein

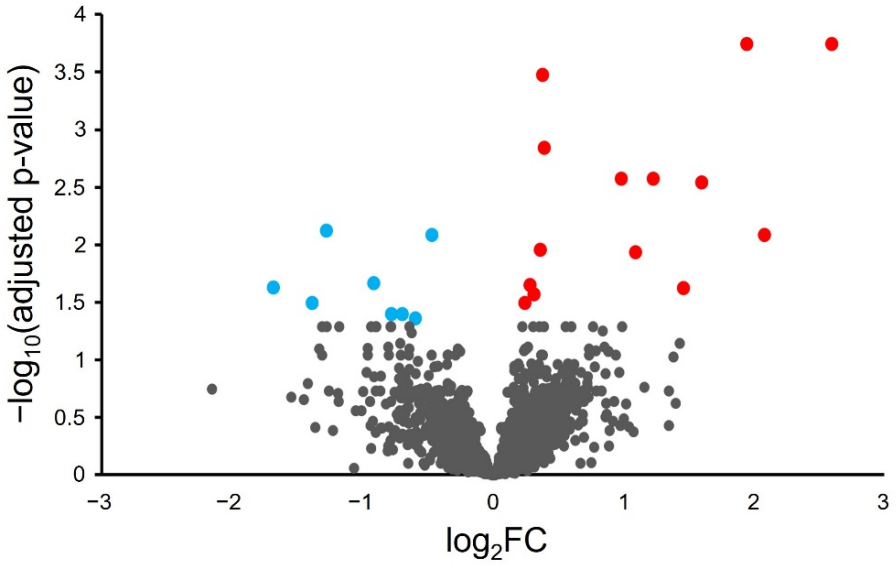
groups correspond to potentially novel proteins annotated by genomic coordinates (indicating novel truncations, proteins with retained introns/untranslated regions, previously untranslated regions of the genome, etc.), and 12 protein groups corresponded to known mass spectrometry contaminants. By contrast, the use of the conventional SwissProt FASTA database identified 8004 proteins organized into 4888 protein groups (data not provided).

Differential analysis was performed on the proteogenomics-derived results to associate proteome abundance changes with phenotypic changes in the inflamed tissue samples. A volcano plot of the $\log_2$ fold-change in protein abundance as a function of $-\log_{10}$ corrected *p*-value (Figure 2a) shows that most proteins do not show significant change with *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 1).

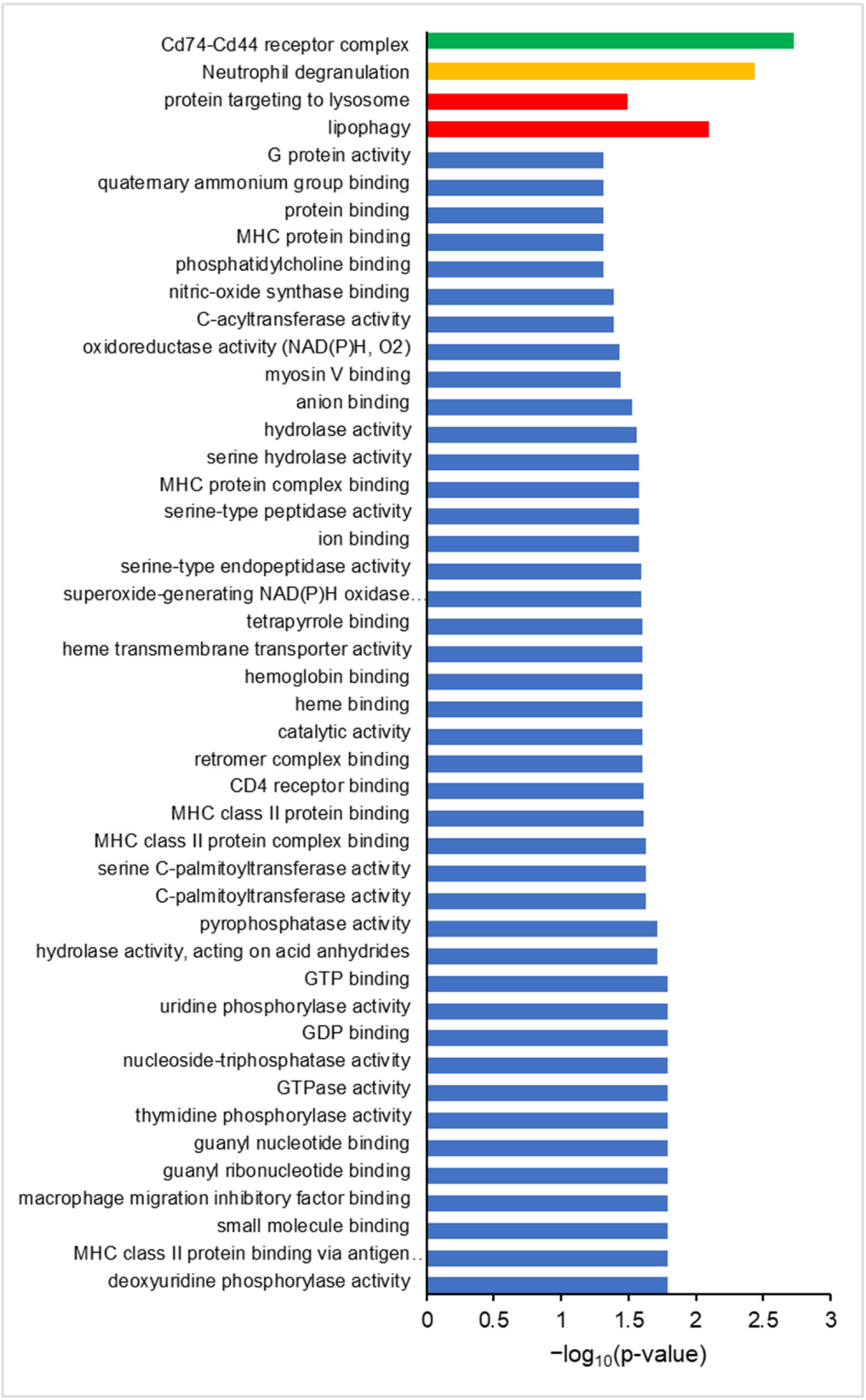
Gene ontology analysis of proteins with an increased abundance in inflamed colon tissue shows enriched GO terms consistent with an inflamed system, showing an enrichment of molecular function GO terms such as MHCI and MHCII complex binding, macrophage migration inhibition factor binding, and oxidoreductase activity, along with the Neutrophil Degranulation reactome and Cd74-Cd44 receptor complex CORUM term (Figure 2b). Proteins that are decreased in abundance in inflamed tissues show enriched GO terms corresponding to molecular functions such as fructose aldolase, the glycolysis/gluconeogenesis and proteosome degradation wikipathway terms, and the 20S proteosome CORUM term (Figure 2c).

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 (Large 1) (Figure S1a). 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 protein sequence generated from the RNA-Seq data (Figure S1b, Table 1).



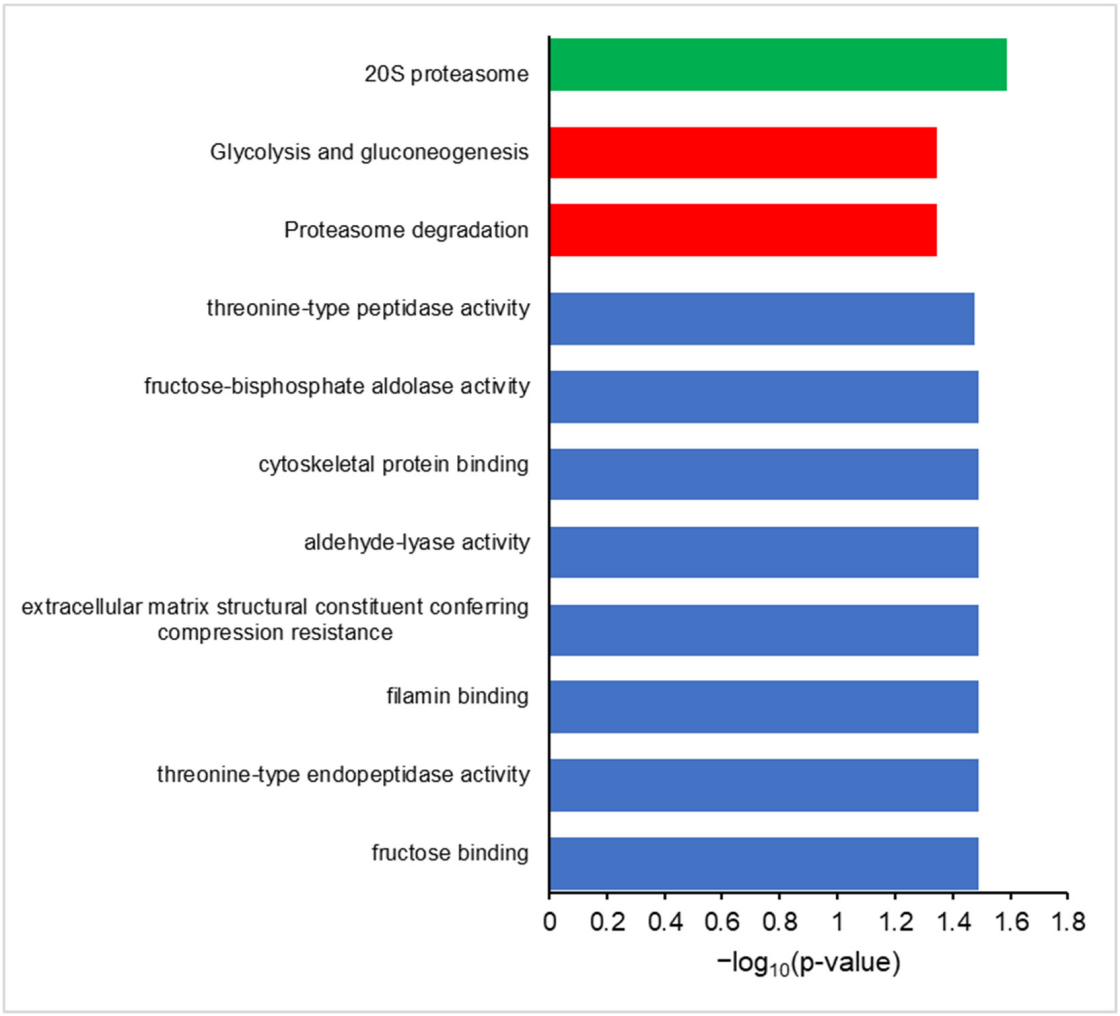
(**a**)

Figure 2. *Cont.*



(**b**)

Figure 2. *Cont.*



(**c**)

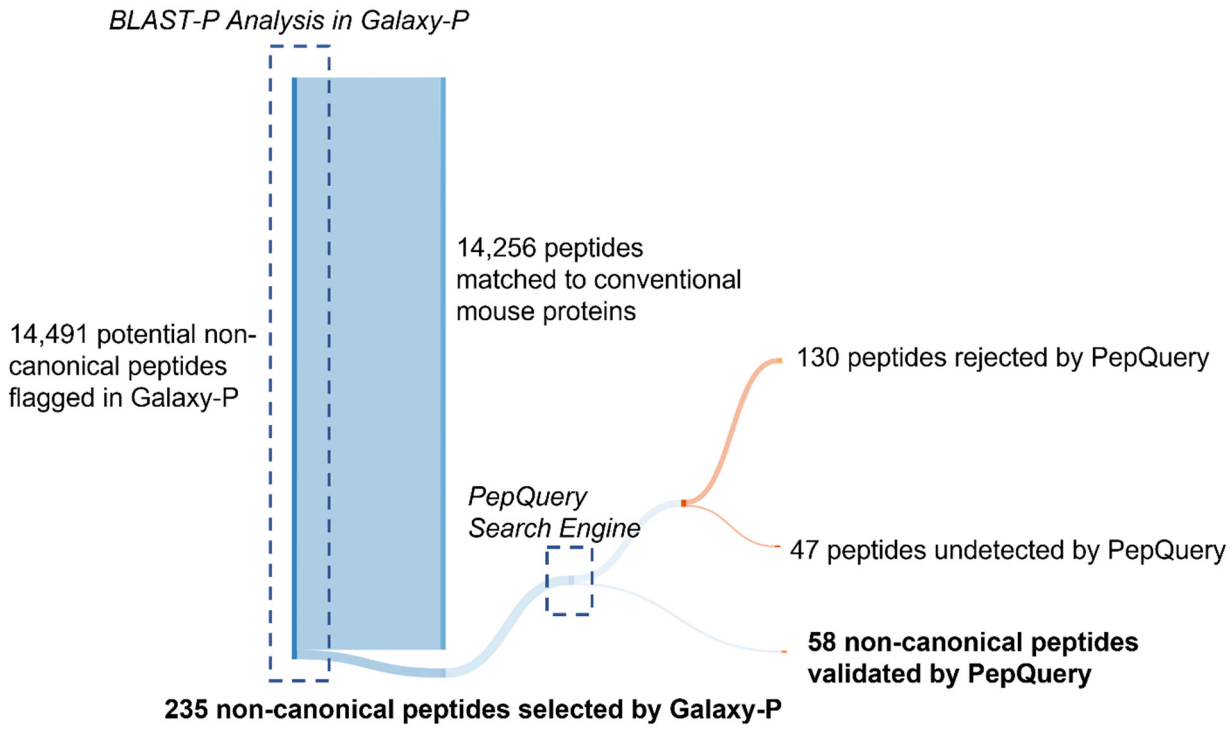
Figure 2. Differential proteogenomic analysis of inflamed proximal colon samples in comparison with untreated controls. (**a**) Enrichment of proteins in proximal colon tissue in response to chronic inflammation, as demonstrated via a volcano plot of log2 fold-change of protein abundance against -log10 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).

Table 1. 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.

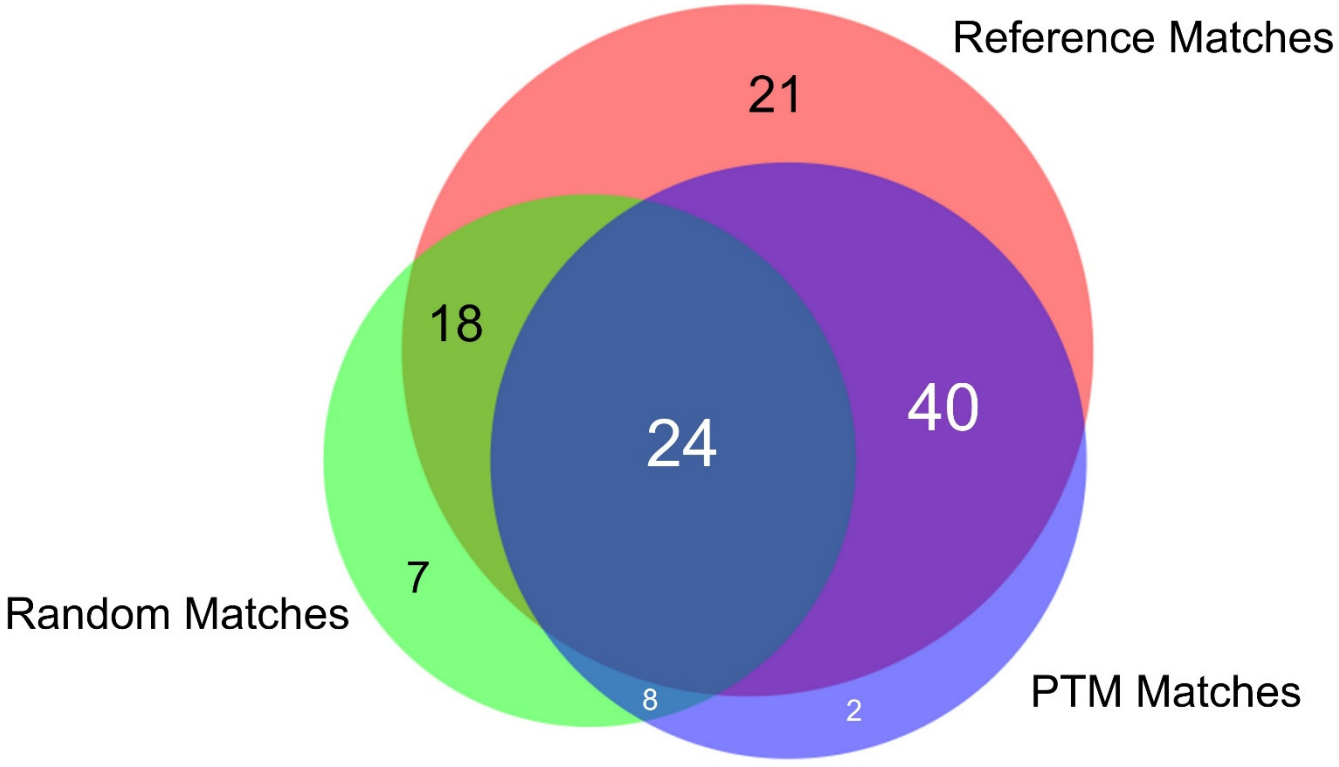
Accession	Description	Gene	Coverage (%)	No. Peptides	log_2FC	*p*-Value	q-Value
Q61646	Haptoglobin	Hp	37	12	2.60	1.27×10^{-7}	1.79×10^{-4}
P07361	Alpha-1-acid glycoprotein 2	Orm2	11	3	2.08	2.90×10^{-5}	8.13×10^{-3}
P07146	Anionic trypsin 2	Prss2	17	3	1.94	1.41×10^{-7}	1.79×10^{-4}
P52624	Uridine phosphorylase 1	Upp1	37	9	1.60	7.86×10^{-6}	2.85×10^{-3}
Q61093	Cytochrome b-245 heavy chain	Cybb	1	1	1.46	1.50×10^{-4}	2.37×10^{-2}
P04441	H-2 class II histocompatibility antigen gamma chain	Cd74	30	8	1.23	5.77×10^{-6}	2.66×10^{-3}
STRG.18707.1_i_2_260	chr8: 73261429–73261687+	-	7	1	1.09	5.45×10^{-5}	1.15×10^{-2}
Q91X72	Hemopexin	Hpx	43	18	0.98	6.29×10^{-6}	2.66×10^{-3}
O35704	Serine palmitoyltransferase 1	Sptlc1	15	6	0.39	2.25×10^{-6}	1.43×10^{-3}
Q9CPW4	Actin-related protein 2/3 complex subunit 5	Arpc5	48	7	0.38	3.94×10^{-7}	3.33×10^{-4}
O35114	Lysosome membrane protein 2	Scarb2	14	6	0.36	4.75×10^{-5}	1.10×10^{-2}
P51150	Ras-related protein Rab-7a	Rab7a	64	12	0.31	1.79×10^{-4}	2.67×10^{-2}
Q9WTL2	Ras-related protein Rab-25	Rab25	44	8	0.28	1.23×10^{-4}	2.24×10^{-2}
Q921J2	GTP-binding protein Rheb	Rheb	28	6	0.24	2.40×10^{-4}	3.20×10^{-2}
A6ZI44	Fructose-bisphosphate aldolase	Aldoa	63	23	−0.47	3.20×10^{-5}	8.13×10^{-3}
P57016	Ladinin-1	Lad1	17	8	−0.60	3.74×10^{-4}	4.31×10^{-2}
Q62000	Mimecan	Ogn	37	9	−0.70	3.28×10^{-4}	3.96×10^{-2}
P35385	Heat shock protein beta-7	Hspb7	33	4	−0.78	3.25×10^{-4}	3.96×10^{-2}
Q7TQD2	Tubulin polymerization -promoting protein	Tppp	17	3	−0.91	1.10×10^{-4}	2.16×10^{-2}
O55234	Proteasome subunit beta type-5	Psmb5	24	6	−1.28	2.36×10^{-5}	7.50×10^{-3}
Q99JI1	Musculoskeletal embryonic nuclear protein 1	Mustn1	18	1	−1.39	2.29×10^{-4}	3.20×10^{-2}
Q19LI2	Alpha-1B-glycoprotein	A1bg	2	1	−1.68	1.38×10^{-4}	2.33×10^{-2}

3.3. Galaxy-P Provides Peptide-Centric Discovery of Non-Canonical Sequences

The 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 with 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 3a). These peptides were hypothesized to correspond with novel proteoforms stemming from translation from unexpected genomic locations, splicing events, or non-synonymous coding sequence variants [27].

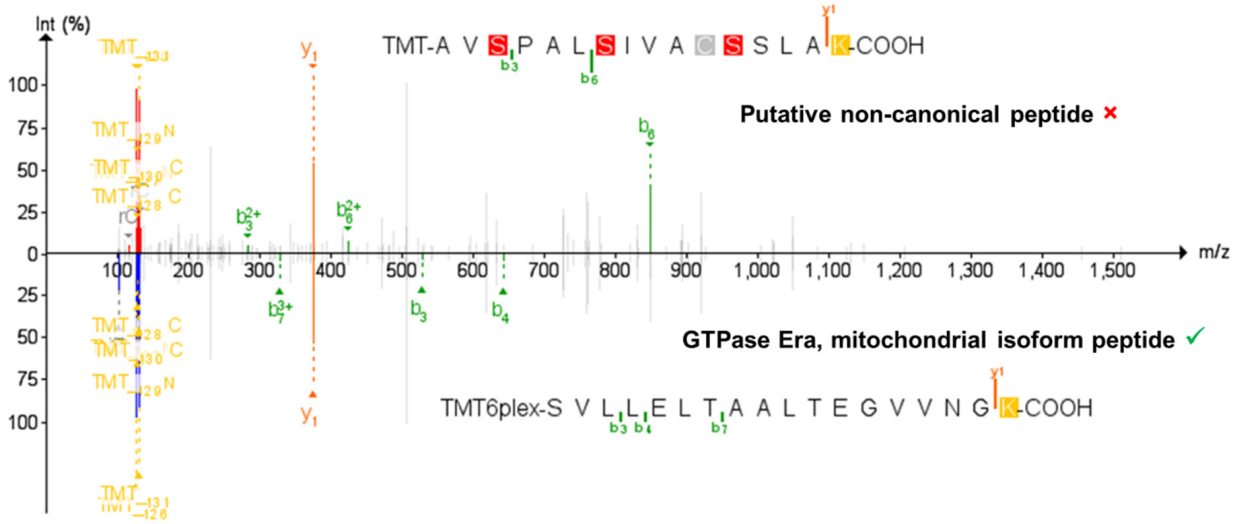


(**a**)

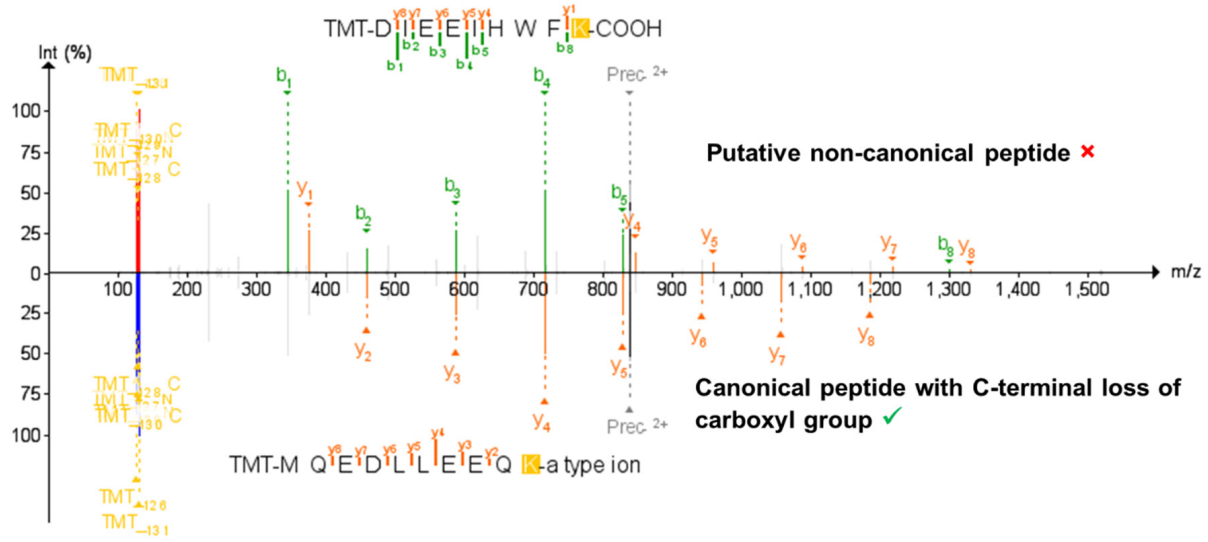


(**b**)

Figure 3. *Cont.*

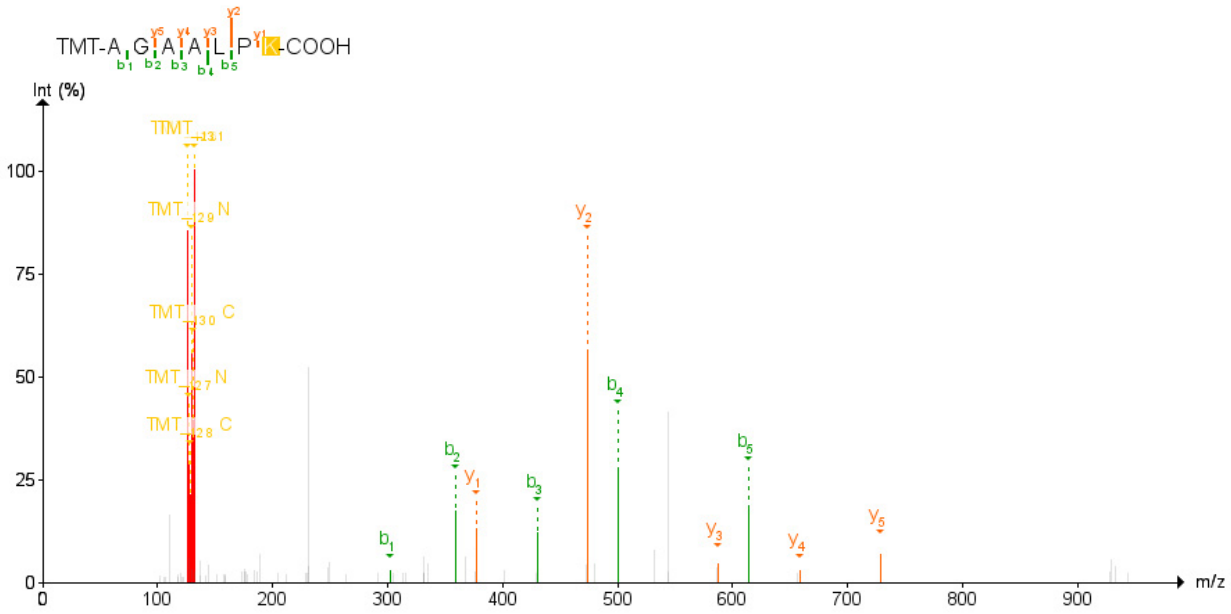


(**c**)



(**d**)

Figure 3. *Cont.*



(**e**)

Figure 3. 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. (**c**–**e**) generated using PDV accessed on 28 December 2021 [47].

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 v1.3 [40], 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 TMT-6-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 S3, Figure S2) in at least one of the fractionated samples. Of these 58 peptides, only eight were confirmed to be phosphorylated consistent with the original PSM and corresponding to peptides from translated intergenic regions and an assortment of genes (Table S3). These 58 peptides were largely unique to the Galaxy-P workflow, as none of these peptides was able to be detected in MSFragger with the custom FASTA database and only three

peptides—AAAAAAAAAAAAASHSVAK, IQSTNQILEAK, and WTSEFEASLINR—were able to be detected with MaxQuant using the custom FASTA database (Figure S3).

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 3a). 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 3b, 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 were tested against (Figure 3b). 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 3c), followed by those assigned high PepQuery-derived *p*-values (Figure 3e), with only two peptides being rejected exclusively for matching reference peptides with PTM modifications (Figure 3d).

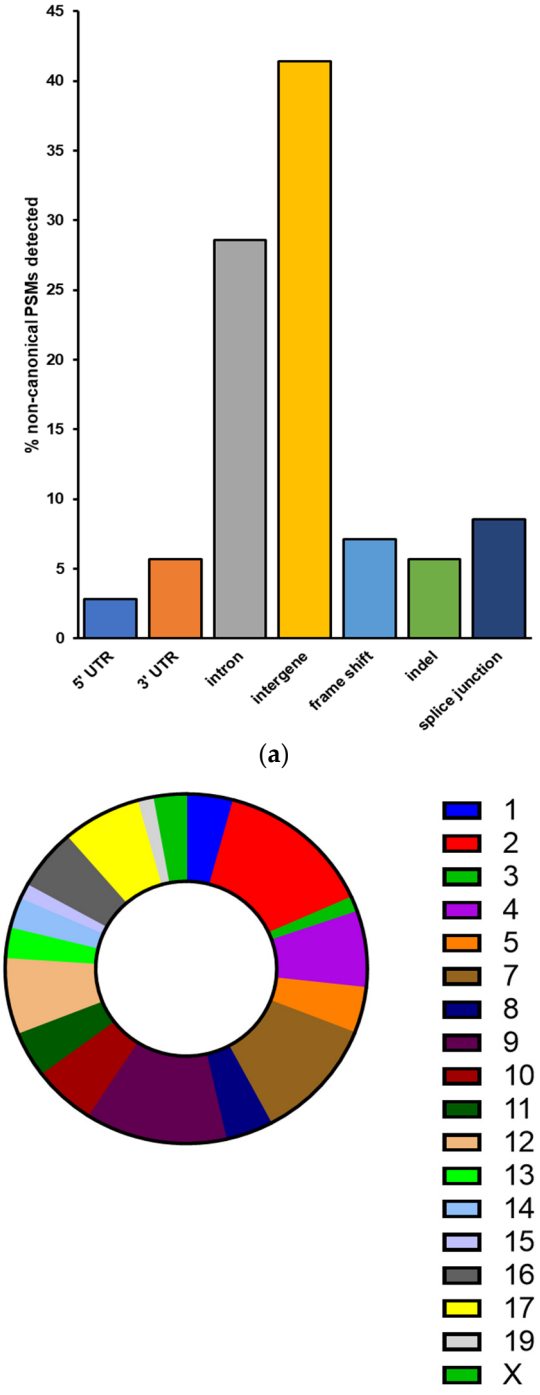
For the verified non-canonical peptides, the majority were found to be associated with intergenic regions not normally transcribed and translated into proteins (40.85%) as well as introns retained in the translated proteins (28.17%) (Figure 4a). The remaining variant peptides comprise indels, frameshifts, 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 4b). 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.

3.5. Targeted Proteomics Experiments Validate the Presence of Non-Canonical Peptides

The non-canonical peptides detected using search and verification workflows were found using 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/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 S3). Graphing the $\log_2$FC of 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 5a); this pattern was mirrored when comparing the change in peptide abundance with the $\log_2$FC of the RNA-Seq data of inflamed versus control samples, where there is a very weak correlation between the two (Figure 5b). 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

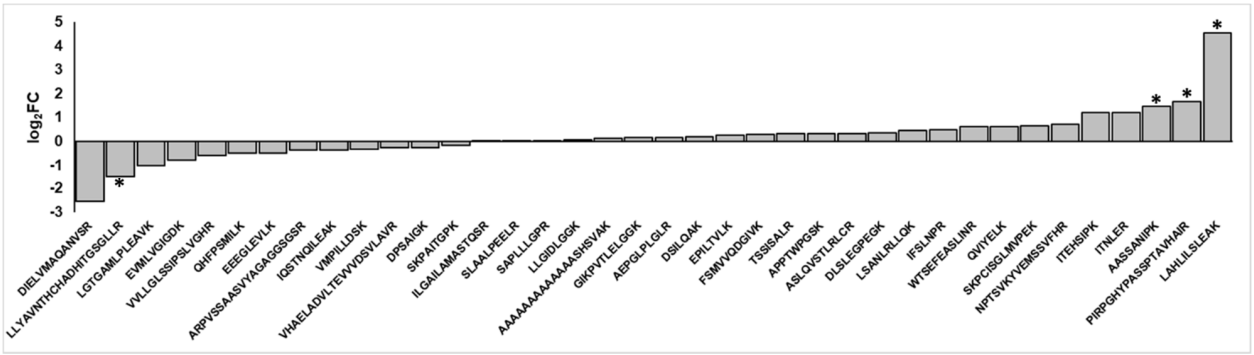
corresponding to a retained 3-UTR section in Sortilin-related receptor Sorl1 (AASSANIPK, Figure S4). In addition, a non-canonical peptide corresponding to an intergenic region on chromosome 19 was found to be depleted in the inflamed tissue samples relative to the control.



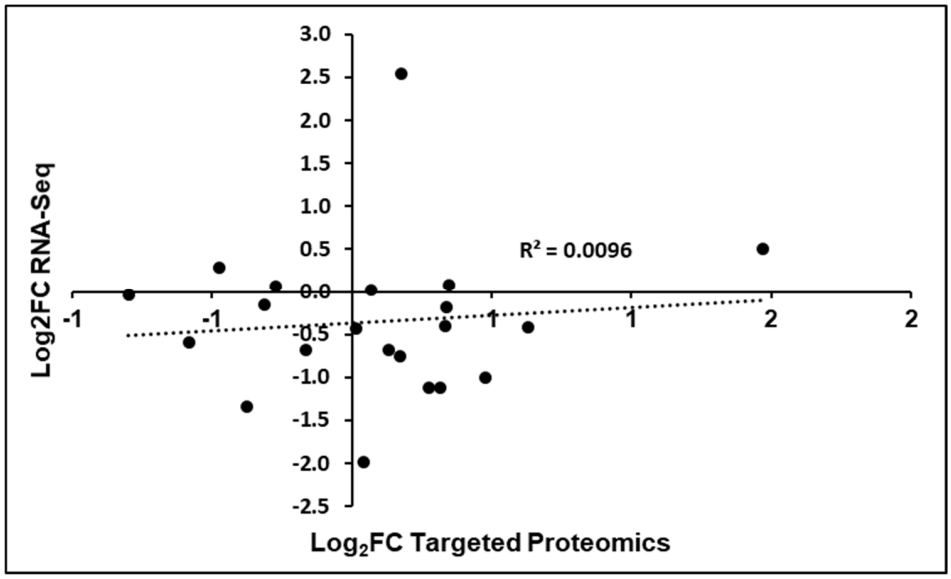
(**a**)

(**b**)

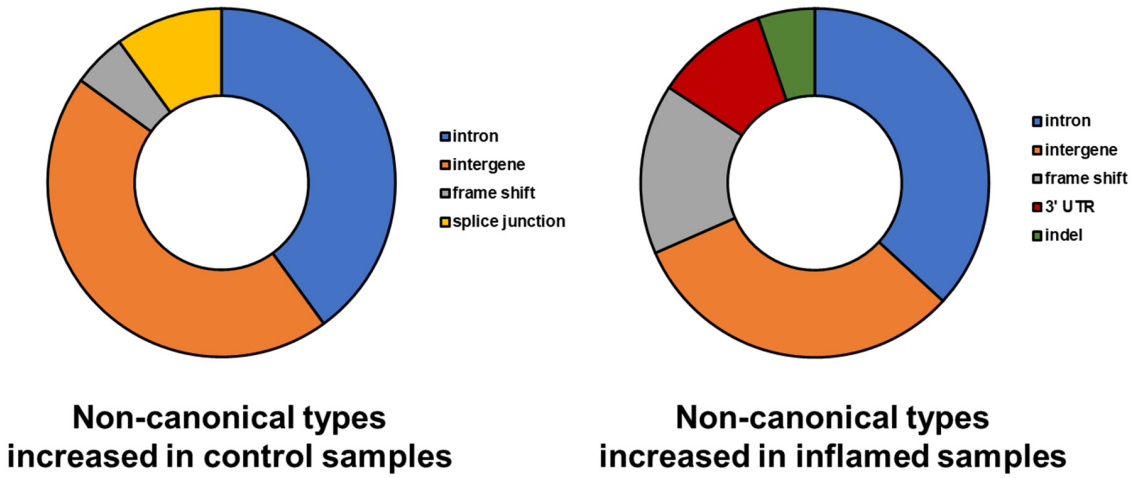
Figure 4. Characteristics of non-canonical sequences validated by PepQuery (**a**) Peptides with non-canonical sequences can be classified into several categories based on their altered sequence or location within a gene. (**b**) Chromosomal locations of non-canonical sequence peptides correspond to mouse chromosomes throughout the genome.



(**a**)



(**b**)



(**c**)

Figure 5. 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. Asterisks indicate a p-value < 0.05. (**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.

While the differences in abundances of 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 in increased abundance in the inflamed tissue or increased abundance in the control sample (Figure 5a). 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 5c). 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.

4. Discussion

In this study, high-resolution 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) Develop 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-*Rag2*^{tm1Fw}*Il10*$^{-/-}$ (*Rag2*$^{-/-}$ *Il10*$^{-/-}$), has been widely used to model inflammatory bowel disease in humans [20,21]. The double knockouts of Recombinase activating gene 2 (*Rag2*) and Interleukin-10 (*Il10*) gene 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 tissue.

The transition from chronic inflammation to oncogenesis is thought to be one of the subtle changes which occurs through a process of DNA damage accretion [48], epigenetic shifts [49], and eventual phenotypic alteration. This 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 with chronic inflammation, being identified in an earlier proteomics study of this model by Knutson et al. [50], indicating their utility as biomarkers for global inflammation; these proteins have also been seen to be increased in abundance in response to sepsis [51], chronic obstructive pulmonary disorder [52], and colorectal cancer [53]. The increased abundance in Prss2, a serine protease involved in the remodeling of the extracellular matrix [54], suggests that the inflamed proximal colon tissue can be considered to be in a chronically inflamed state [55] as increased abundance in Prss2 differentiates IBD patients from healthy patients [56], making the five-month exposure of these mice a suitable model for chronic inflammatory bowel disease. Other indications of chronic inflammation are the increased abundance in 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 cells [57]. Other increased proteins consistent

with an inflammatory phenotype include heavy-chain Cytochrome b-245 (Cybb), a key component of NADH oxidase in phagocytes needed to create superoxides as a part of the inflammatory response [58], serine palmitoyltransferase 1 (Sptlc1), the initial enzyme involved in sphingolipid synthesis [59] and GTP-binding protein Rheb (Rheb) which serves to activate mTOR1 and promote signal transduction [60]. Interestingly, the increase in abundance of Upp1 seen in the inflamed samples is consistent with the development of many cancers [61,62], indicating a degree of oncogenesis may have begun. 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 inflammation [63].

A major limitation when using TMT-labeling for quantitative proteogenomics is that TMT-based quantitation is protein-centric, inferring protein abundances 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 an advanced peptide-centric proteogenomic bioinformatic workflow 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 identifications [18]. Together, these two nodes of the workflow eliminate false positives of putative non-canonical peptides that are more effectively matched to canonical peptides or common contaminants. There are three ways in which the PepQuery search engine rejects potential non-canonical peptides, all of which were seen in our inflamed proximal colon data and are dependent upon the quality of the PSM within each fractionated mass spectrometry experiment (Figure 3b). 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 3c, top) as well or better to 44 peptides found within the canonical mouse proteome, including the GTPase Era, mitochondrial isoform peptide SVLLELTAALTEGVVNFK (Figure 3c, 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 3e). In this way, PepQuery can eliminate uncertain matches stemming from large mass errors by setting a minimal cutoff value of acceptable match confidences as expressed by *p*-values in the PepQuery outputs. Finally, including 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, removing the false positive of post-translational modifications to conventional peptides. This option resulted in the rejection of a PSM identifying the non-canonical sequence DIEEIHWFK in favor of a superior match to the canonical MQEQLLEEQK with an a-type ion on the C-terminus corresponding to the loss of part of the C-terminal lysine (Figure 3d). 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 the non-canonical 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 with the controls. These peptides encompass chromosomes throughout the murine genome and represent, principally, the translation of genomic sequences not normally translated, such as intergenic regions, introns, UTRs, etc., indicating potentially altered levels of epigenetic regulation and translational control during colon inflammation [64,65]. Parallel reaction monitoring allowed for the deeper sampling of detected peptides to enable more accurate quantitation as compared with 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, potential sample degradation prior to targeted analysis, or interference by co-eluting peaks. These questions make it difficult to determine conclusively whether these sequences were not actually present, or simply were not detectable by PRM. Future studies to answer such questions could include further optimization of a targeted methodology by including synthetic peptide standards, 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 LC and MS parameters via synthetic peptide standards prior to analysis.

The relevance of these 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 Browser [66]. The human gene sequences were then searched using the online PepQuery server against cancer-tissue derived mass spectrometry data from the Cancer Genome Atlas [30,31,33]. 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 S4). This demonstrates a potential for these peptides to serve as early biomarkers for human oncogenesis.

Beyond revealing differential protein abundances and sequence variations as a result of 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 sequences [67].

5. Conclusions

In this study, we examined the *Helicobacter hepaticus*-induced inflammation of proximal colon tissue in mice through mass spectrometry-based proteogenomics supplemented with RNA-Seq data. Our initial global proteomics analysis revealed an upregulation of proteins in our inflamed samples consistent with an inflammatory phenotype along with proteoforms that are undetectable using conventional bottom-up proteomics strategies. Through an automated, open access workflow in Galaxy-P, we were able to detect and validate non-canonical peptides across all samples, the majority of which could subsequently be validated using targeted mass spectrometry experiments. We believe this work to be significant in that the workflows presented here allow for the confident identifica-

tion of non-canonical peptides in mass spectrometry data stemming from insertions and deletions, amino acid substitutions, or alternative splicing events which would serve as invaluable biomarkers for the diagnosis and treatment of colon cancer. The open-source, user-friendly nature of the workflows used in this study allows for their ready uptake and use by non-bioinformaticians, expanding the use of proteogenomics to researchers beyond traditional mass spectrometrists and systems biologists. For future studies, we intend to further optimize the workflows detailed here, allowing for automated quantification of detected non-canonical peptide sequences, as well as automatic generation of parameters for targeted mass spectrometry analysis; in addition, we intend to expand our use of these tools to analyzing other tissues in mice subjected to *Helicobacter hepaticus* infection, such as the distal colon, cecum, and serum samples. In addition, future studies will utilize larger numbers of test animals to increase the statistical power of our analyses. Finally, targeted validation experiments will utilize exclusion lists and extended gradients to detect potential non-canonical peptides more effectively.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/proteomes10020011/s1, Table S1: Proximal colon samples used in this study. Table S2: Inclusion list for targeted detection of non-canonical peptides in proximal colon samples. Table S3: Non-canonical peptide sequences detected in the proximal colon tissues. Table S4: Human parallels to mouse non-canonical peptide sequences found in TCGA data. Figure S1: Genomic coordinates of proteins enriched in inflamed mouse proximal colon samples. Figure S2: Comparison of peptides detected using the workflows in Galaxy-P compared with the results of searching the data using MSFragger and MaxQuant. Figure S3: MS/MS spectra of non-canonical peptides passing PepQuery validation. Figure S4: Genomic coordinates of the non-canonical peptide sequence AASSANIPK.

Author Contributions: The manuscript was written through contributions of all authors. Conceptualization, A.T.R., P.D.J., N.Y.T. and T.J.G.; Methodology, A.T.R.; Software, S.M. and P.K.; Formal Analysis, A.T.R.; Resources, Q.H., C.G.K. and J.G.F.; Writing—Original Draft Preparation, A.T.R.; Writing—Review and Editing, Q.H., P.D.J., C.G.K., J.G.F., N.Y.T. and T.J.G.; Visualization, A.T.R.; Project Administration, P.D.J., N.Y.T. and T.J.G.; Funding Acquisition, N.Y.T. and T.J.G. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge funding for this work from the National Cancer Institute—Informatics Technology for Cancer Research (NCI-ITCR) grant 1U24CA199347 to T.J.G., as well as the National Institutes of Health (NIH) grants R01-CA100670, R01-CA095039 to N.Y.T. A.T.R. was supported by the National Institutes of Health Biotechnology Training Grant: NIH T32GM008347.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of material previously generated and published in Erdman et al. [21].

Informed Consent Statement: Not applicable.

Data Availability Statement: All data were deposited to the ProteomeXchange Consortium via the PRIDE partner repository (http://www.ebi.ac.uk/pride/archive/, accessed on 9 September 2021) and are publicly accessible with the data set identifier PXD028407.

Acknowledgments: The authors wish to acknowledge Luke Erber and Christopher Seiler for their advice in the prosecution of this research.

Conflicts of Interest: The authors declare no competing financial interests or conflicts of interest.

Abbreviations

AGC, automatic gain control; BCA, bicinchoninic acid assay; BLAST-P, Basic Local Alignment Search for Proteins; CORUM, comprehensive resource of mammalian protein complexes; DNA, deoxyribonucleic acid; DTT, dithiothreitol; IBD, Inflammatory Bowel Disease; IT, injection time; FDR, false discovery rate; FWHM, full width at half maximum; GO, gene ontology; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; HPLC, high-performance liquid chromatography; IAA, iodoacetamide; indel, insertion or deletion; LC–MS, liquid chromatography-mass spectrometry; MS, mass spectrometry; NCBI, National Center for Bioinformatic Information; PDV, Proteomics Data Viewer; PMSF, phenylmethylsulfonyl fluoride; PRM, parallel reaction monitoring; PSM, peptide

spectral match; PTM, post-translational modification; RNA-Seq, ribonucleic acid sequencing; RNS, reactive nitrogen species; ROS, reactive oxygen species; TCGA, The Cancer Genome Atlas; TEAB, triethylammonium bicarbonate; TMT, tandem mass tag; UCSC, University of California, Santa Cruz; UHPLC, ultra-high-performance liquid chromatography; UTR, untranslated region.

References

1. Chiba, T.; Marusawa, H.; Ushijima, T. Inflammation-Associated Cancer Development in Digestive Organs: Mechanisms and Roles for Genetic and Epigenetic Modulation. *Gastroenterology* **2012**, *143*, 550–563. [CrossRef] [PubMed]
2. 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. *Oncol. Lett.* **2015**, *9*, 1015–1026. [CrossRef] [PubMed]
3. Greten, F.R.; Eckmann, L.; Greten, T.F.; Park, J.M.; Li, Z.-W.; Egan, L.J.; Kagnoff, M.F.; Karin, M. IKKβ Links Inflammation and Tumorigenesis in a Mouse Model of Colitis-Associated Cancer. *Cell* **2004**, *118*, 285–296. [CrossRef] [PubMed]
4. Affara, N.I.; Coussens, L.M. IKKα at the Crossroads of Inflammation and Metastasis. *Cell* **2007**, *129*, 25–26. [CrossRef] [PubMed]
5. Mangerich, A.; Knutson, C.G.; Parry, N.M.; Muthupalani, S.; Ye, W.; Prestwich, E.; Cui, L.; McFaline, J.L.; Mobley, M.; Ge, Z.; et al. Infection-induced colitis in mice causes dynamic and tissue-specific changes in stress response and DNA damage leading to colon cancer. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1820–E1829. [CrossRef] [PubMed]
6. 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.; et al. DNA damage induced by chronic inflammation contributes to colon carcinogenesis in mice. *J. Clin. Investig.* **2008**, *118*, 2516–2525. [CrossRef]
7. Huang, Z.; Huang, D.; Ni, S.; Peng, Z.; Sheng, W.; Du, X. Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. *Int. J. Cancer* **2010**, *127*, 118–126. [CrossRef]
8. Alves Martins, B.A.; De Bulhões, G.F.; Cavalcanti, I.N.; Martins, M.M.; de Oliveira, P.G.; Martins, A.M.A. Biomarkers in colorectal cancer: The role of translational proteomics research. *Front. Oncol.* **2019**, *9*, 1284. [CrossRef]
9. 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*, 1962–1985. [CrossRef]
10. Jardim-Perassi, B.V.; Alexandre, P.A.; Sonehara, N.M.; De Paula-Junior, R.; Reis Júnior, O.; Fukumasu, H.; Chammas, R.; Coutinho, L.L.; Zuccari, D.A.P.D.C. RNA-Seq transcriptome analysis shows anti-tumor actions of melatonin in a breast cancer xenograft model. *Sci. Rep.* **2019**, *9*, 966. [CrossRef]
11. 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. *Front. Physiol.* **2018**, *9*, 812. [CrossRef] [PubMed]
12. Kisluk, J.; Ciborowski, M.; Niemira, M.; Kretowski, A.; Niklinski, J. Proteomics biomarkers for non-small cell lung cancer. *J. Pharm. Biomed. Anal.* **2014**, *101*, 40–49. [CrossRef] [PubMed]
13. Hegde, P.S.; White, I.R.; Debouck, C. Interplay of transcriptomics and proteomics. *Curr. Opin. Biotechnol.* **2003**, *14*, 647–651. [CrossRef] [PubMed]
14. Alfaro, J.A.; Sinha, A.; Kislinger, T.; Boutros, P.C. Onco-proteogenomics: Cancer proteomics joins forces with genomics. *Nat. Methods* **2014**, *11*, 1107–1113. [CrossRef]
15. Vasaikar, S.; Huang, C.; Wang, X.; Petyuk, V.A.; Savage, S.R.; Wen, B.; Dou, Y.; Zhang, Y.; Shi, Z.; Arshad, O.A.; et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **2019**, *177*, 1035–1049. [CrossRef]
16. Smith, L.M.; Kelleher, N.L.; The Consortium for Top Down Proteomics. Proteoform: A single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186–187. [CrossRef]
17. 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*, 755–765. [CrossRef]
18. Nesvizhskii, A. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114–1125. [CrossRef]
19. 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. [CrossRef]
20. 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. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 1027–1032. [CrossRef]
21. 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. *Int. J. Mol. Sci.* **2021**, *22*, 364. [CrossRef] [PubMed]
22. 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. *Am. J. Pathol.* **2003**, *162*, 691–702. [CrossRef]

23. Brennan, M.L.; Wu, W.; Fu, X.; Shen, Z.; Song, W.; Frost, H.; Vadseth, C.; Narine, L.; Lenkiewicz, E.; Borchers, M.T.; et al. A tale of two controversies: Defining both the role of peroxidases in nitrotyrosine formation in vivo using eosinophil peroxidase and myeloperoxidase-deficient mice, and the nature of peroxidase-generated reactive nitrogen species. *J. Biol. Chem.* **2002**, *277*, 17415–17427. [CrossRef]
24. 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. Multi-omic data analysis using Galaxy. *Nat. Biotechnol.* **2015**, *33*, 137–139. [CrossRef]
25. 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. *J. Proteome Res.* **2020**, *19*, 2772–2785. [CrossRef]
26. 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*, e43–e46. [CrossRef] [PubMed]
27. 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*, 5898–5908. [CrossRef]
28. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef]
29. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.
30. Wang, X.; Zhang, B. customProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29*, 3235–3237. [CrossRef]
31. Mellacheruvu, D.; Wright, Z.; Couzens, A.L.; Lambert, J.-P.; St-Denis, N.A.; Li, T.; Miteva, Y.V.; Hauri, S.; Sardiu, M.E.; Low, T.Y.; et al. The CRAPome: A contaminant repository for affinity purification–mass spectrometry data. *Nat. Methods* **2013**, *10*, 730–736. [CrossRef] [PubMed]
32. 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. *Nat. Biotechnol.* **2015**, *33*, 290–295. [CrossRef] [PubMed]
33. Mehta, S.; Griffin, T.J.; Jagtap, P.; Sajulga, R.; Johnson, J.; Kumar, P. Proteogenomics 1: Database Creation (Galaxy Training Materials). 2021. Available online: http:///training-material/topics/proteomics/tutorials/proteogenomics-dbcreation/tutorial.html (accessed on 29 April 2021).
34. 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*, 996–999. [CrossRef] [PubMed]
35. 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. *Nat. Biotechnol.* **2015**, *33*, 22–24. [CrossRef] [PubMed]
36. 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. *Mol. Cell. Proteom.* **2020**, *19*, 1706–1723. [CrossRef] [PubMed]
37. 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 Res.* **2019**, *47*, W191–W198. [CrossRef]
38. Gish, W.; States, D.J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **1993**, *3*, 266–272. [CrossRef]
39. Available online: https://github.com/galaxyproteomics/tools-galaxyp/tree/master/tools/pep_pointer (accessed on 6 April 2022).
40. Wen, B.; Wang, X.; Zhang, B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res.* **2019**, *29*, 485–493. [CrossRef]
41. 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. *Mol. Cell. Proteom.* **2012**, *11*, 1475–1488. [CrossRef]
42. 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*, 966–968. [CrossRef]
43. 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. *Nat. Methods* **2019**, *16*, 509–518. [CrossRef] [PubMed]
44. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
45. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
46. Kumar, D.; Yadav, A.K.; Dash, D. *Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data*; Springer: New York, NY, USA, 2017; Volume 1549, pp. 17–29.
47. Li, K.; Vaudel, M.; Zhang, B.; Ren, Y.; Wen, B. PDV: An integrative proteomics data viewer. *Bioinformatics* **2019**, *35*, 1249–1251. [CrossRef] [PubMed]

48. Farinati, F.; Cardin, R.; Degan, P.; Rugge, M.; Di Mario, F.; Bonvicini, P.; Naccarato, R. Oxidative DNA damage accumulation in gastric carcinogenesis. *Gut* **1998**, *42*, 351–356. [CrossRef] [PubMed]
49. Hatziapostolou, M.; Iliopoulos, D. Epigenetic aberrations during oncogenesis. *Cell. Mol. Life Sci.* **2011**, *68*, 1681–1702. [CrossRef] [PubMed]
50. 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. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E2332–E2341. [CrossRef]
51. Janz, D.R.; Bastarache, J.A.; Sills, G.; Wickersham, N.; May, A.K.; Bernard, G.R.; Ware, L.B. Association between haptoglobin, hemopexin and mortality in adults with sepsis. *Crit. Care* **2013**, *17*, R272. [CrossRef]
52. Winter, N.A.; Gibson, P.G.; Fricker, M.; Simpson, J.L.; Wark, P.A.; McDonald, V.M. Hemopexin: A Novel Anti-inflammatory Marker for Distinguishing COPD From Asthma. *Allergy, Asthma Immunol. Res.* **2021**, *13*, 450–467. [CrossRef]
53. Zhang, X.; Xiao, Z.; Liu, X.; Du, L.; Wang, L.; Wang, S.; Zheng, N.; Zheng, G.; Li, W.; Zhang, X.; et al. The Potential Role of ORM2 in the Development of Colorectal Cancer. *PLoS ONE* **2012**, *7*, e31868. [CrossRef]
54. 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. *Reprod. Biol. Endocrinol.* **2017**, *15*, 20. [CrossRef] [PubMed]
55. Wynn, T.A. Cellular and molecular mechanisms of fibrosis. *J. Pathol.* **2008**, *214*, 199–210. [CrossRef] [PubMed]
56. Wang, W.; Wu, L.; Wu, X.; Li, K.; Li, T.; Xu, B.; Liu, W. Combined analysis of serum SAP and PRSS2 for the differential diagnosis of CD and UC. *Clin. Chim. Acta* **2021**, *514*, 8–14. [CrossRef] [PubMed]
57. Waithman, J.; Moffat, J.M.; Patterson, N.L.; van Beek, A.E.; Mintern, J.D. Antigen Presentation. In *Reference Module in Biomedical Sciences*; Elsevier: Amsterdam, The Netherlands, 2014.
58. 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*, 1244–1258. [CrossRef] [PubMed]
59. 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 Dis.* **2018**, *9*, 173. [CrossRef]
60. 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 Dev.* **2008**, *22*, 2172–2177. [CrossRef]
61. Wang, J.; Xu, S.; Lv, W.; Shi, F.; Mei, S.; Shan, A.; Xu, J.; Yang, Y. Uridine phosphorylase 1 is a novel immune-related target and predicts worse survival in brain glioma. *Cancer Med.* **2020**, *9*, 5940–5947. [CrossRef]
62. Miyashita, H.; Takebayashi, Y.; Eliason, J.F.; Fujimori, F.; Nitta, Y.; Sato, A.; Morikawa, H.; Ohashi, A.; Motegi, K.; Fukumoto, M.; et al. Uridine phosphorylase is a potential prognostic factor in patients with oral squamous cell carcinoma. *Cancer* **2002**, *94*, 2959–2966. [CrossRef]
63. 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. *J. Crohn's Colitis* **2017**, *11*, 92–104. [CrossRef]
64. Hartnett, L.; Egan, L.J. Inflammation, DNA methylation and colitis-associated cancer. *Carcinogenesis* **2012**, *33*, 723–731. [CrossRef]
65. Janssen, W.J.; Danhorn, T.; Harris, C.; Mould, K.; Lee, F.F.-Y.; Hedin, B.R.; D'Alessandro, A.; Leach, S.M.; Alper, S. Inflammation-Induced Alternative Pre-mRNA Splicing in Mouse Alveolar Macrophages. *G3* **2020**, *10*, 555–567. [CrossRef] [PubMed]
66. 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*, 996–1006. [CrossRef] [PubMed]
67. 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*, 15. [CrossRef] [PubMed]