UNIVERSITY OF CALIFORNIA SAN DIEGO

Decoding Phenotypes via Transcriptomics and Proteomics: Cancer and beyond

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Miin Sophia Lin

Committee in charge:

Professor Vineet Bafna, Chair
Professor David Gonzalez, Co-Chair
Professor Nuno Bandeira
Professor Rob Knight
Professor Pavel Pevzner

2022

The Dissertation of Miin Sophia Lin is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

Table of Contents

List of Supplemental Files

Lin_Chapter_1_Supplemental_Materials.zip

Lin_Chapter_1_Supplementary_Tables.xlsx

Lin_Chapter_2_Supplemental_Materials.zip

Lin_Chapter_2_Supplementary_Tables.xlsx

Lin_Chapter_3_Supplemental_Materials.zip

Lin_Table_3.1.xlsx

Lin_Chapter_3_Supplementary_Tables.xlsx

List of Figures

Acknowledgements

Chapter 1 includes content of material being prepared for submission by Lin, Miin S.; Lie, Kai Kristoffer; Varunjikar, Madhushri Shrikant; Søfteland, Liv; Dellafiora, Luca; Dorne, Jean-Lou Christian Michel; Ørnsrud, Robin; Sanden, Monica; Berntssen, Marc; Bafna, Vineet; Rasinger, Josef D. in a manuscript with a tentative title of "Proteogenomic analysis of multi-tissue data for *in vivo – in vitro* extrapolations of the Atlantic Salmon (*Salmo salar*) chemical defensome." The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Cell Systems 2018. Beyter, Doruk; Lin, Miin S.; Yu, Yanbao; Pieper, Rembert; Bafna, Vineet. (2018). "ProteoStorm: An Ultrafast Metaproteomics Database Search Framework." Cell Systems 7(4): 563-467. The dissertation author was the primary investigator and author of this paper.

Chapter 3 includes content of material being prepared for submission by Lin, Miin S. and Bafna, Vineet in a manuscript with the tentative title "Genes involved in the maintenance of extrachromosomal DNA." The dissertation author was the primary investigator and author of this paper.

Vita

2012 Bachelor of Arts in Biology and Molecular Biology & Biochemistry, certificate in Informatics and Modeling (Integrative Genomic Sciences), Wesleyan University

2013 Master of Arts in Biology, Wesleyan University

2022 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

Publications

Beyter, Doruk [1]; Lin, Miin S.[1]; Yu, Yanbao; Pieper, Rembert; Bafna, Vineet. (2018). "ProteoStorm: An Ultrafast Metaproteomics Database Search Framework." *Cell Systems* 7(4): 563-467.

Lin, M.S., Cherny, J.J., Fournier, C.T., Roth, S.J., Krizanc, D. and Weir, M.P. (2014). "Assessment of MS/MS search algorithms with parent-protein profiling." *J. Proteome Res.* 13(4):1823-1832.

Fournier, C.T., Cherny, J.J., Truncali, K., Robbins-Pianka, A., Lin, M.S., Krizanc, D. and Weir, M.P. (2012). "Amino termini of many yeast proteins map to downstream start codons." *J. Proteome Res.* 11(12):5712-5719.

[1] These authors contributed equally to this work.

ABSTRACT OF THE DISSERTATION

Decoding Phenotypes via Transcriptomics and Proteomics: Cancer and beyond

by

Miin Sophia Lin

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Vineet Bafna, Chair
Professor David Gonzalez, Co-Chair

While genomics approaches are important in studying host phenotype alterations in response to environmental changes or disease, proteomics approaches offer a complementary perspective by providing a direct readout of expressed functional pathways. Proteogenomic strategies utilizing RNA-sequencing data to construct splice graph databases have been used in a variety of applications to identify novel splice junctions and mutated peptides. The work in this dissertation begins with the integration of splice databases into a proteogenomic pipeline for the validation of the recently released annotation of the Atlantic salmon genome, and the validation of primary hepatocytes as *in vitro* models for salmon toxicity studies. Searching in-house generated LC-MS/MS datasets against splice databases

constructed from publicly available and in-house-generated salmon transcriptomics data, our proteogenomic pipeline identified 183 events in support of 71 transcript predictions. These included novel genes, corrections to current annotations, and support for Ensembl transcripts. In addition to host-expressed proteins, microbial-expressed proteins can also alter host phenotype. In the absence of prior taxonomic information, tandem mass spectra would be searched against large pan-microbial databases, requiring heavy computational workload and reducing sensitivity. Using both software and algorithmic methods, we developed ProteoStorm, an efficient database search framework for large-scale metaproteomics studies, that significantly reduced runtime from 22 weeks to 9.7 hours while retaining 96% of peptide identifications when compared to MSGF+. A reanalysis of a urinary tract infection dataset revealed a complex pattern of polymicrobial expression, including previously identified microbes. In the final chapter, we used transcriptomics data from TCGA to identify a set of genes that may be involved in the maintenance of ecDNA amplicons in cancer. Specifically, we applied the Boruta algorithm, which incorporates the Random Forest classifier for feature selection, to a gene expression matrix of samples classified as ecDNA positive or negative, and selected 408 Core genes predictive of ecDNA status. We further extended the list with 235 highly co-expressed genes using hierarchical clustering, resulting in a total of 643 CorEx genes. Subsequent gene set enrichment analysis revealed an up-regulation of biological processes involved in cell cycle, cell division, and DNA damage response, and a down-regulation of immune system processes.

Chapter 1: Proteogenomic Analysis of Multi-tissue Data for in vivo – in vitro Extrapolations of the Atlantic Salmon (*Salmo salar*) Chemical Defensome

1.1: Abstract

In this study, we describe an integrative transcriptomics and proteomics analysis workflow for the efficient validation and revision of complex genomes and the creation of proteogenomics expression matrices to facilitate omics data integration and visualization in non-model species *in vivo* and *in vitro*. Using the recently annotated Atlantic salmon (*Salmo salar*) genome as example, we constructed proteogenomic databases from publicly available transcriptomic data of twelve salmon tissues, and in-house generated transcriptomic data of salmon liver (n=44) and primary hepatocyte cultures (n=60). Searching these data against all corresponding in-house generated LC-MS/MS datasets (n=127) allowed for the identification of translated peptides. We identified nearly 80,000 peptides providing direct evidence of translation for over 40,000 RefSeq structures. The identifications included 183 co-located peptide groups ("proteogenomic events") that supported a single transcript each, and in each case, either corrected a previous annotation, supported an Ensembl annotation not in Refseq, or identified a novel (previously un-annotated) gene. Further analysis of the identified proteins revealed distinct proteomic profiles for each tissue type analyzed, and a separation of liver and primary hepatocyte samples based on inherent differences between *in vivo* and *in vitro* experiments. Focusing on proteins involved in the defense against xenobiotics, distinct and overlapping expression patterns were detected across all salmon tissues, and a large degree of homology between *in vivo* and *in vitro* liver systems was observed. Similar to previous publications on the zebrafish (*Danio rerio*), Our data also corroborates the role of proteogenomic analyses in extending our understanding of complex

genomes. Specifically, we show that 3D primary hepatocytes are a suitable *in vitro* model system for studying effects in Atlantic salmon, and ultimately, demonstrate the usability of proteogenomics in mechanistic toxicology studies in non-model species.

1.2: Introduction

Fish physiology and toxicology are moving towards the concomitant analysis of genomic, proteomic, and metabolomic data sets (Martyniuk and Denslow, 2009). An integrative analysis of global changes across the omics hierarchy provides distinct advantages in the elucidation of the functional state of the cell, offers insights into the mechanistic toxicology of xenobiotics, and facilitates the discovery of novel biomarkers relevant to toxicant exposure (Waters and Fostel, 2004). In well-studied model-organisms such as rodents or zebrafish, proteomics and multi-level omics analyses have been frequently employed in mechanistic toxicity research (Bernhard et al., 2018; Mellingen et al., 2022, 2021; Rasinger et al., 2018, 2017, 2014). In non-model organisms such as (farmed) Atlantic salmon (*Salmo salar*), few studies include omics-based approaches when accessing and describing the toxicity of xenobiotics, mainly due to technical bioinformatics challenges, (Bernhard et al., 2019; Hampel et al., 2015; Søderstrøm et al., 2022).

However, this is likely to change in the future, since advancements in genome sequencing, assembly, and annotation have been facilitating mechanistic research in non-model organisms in many key-areas, including agricultural research, veterinary medicine, zootechnics, ecology, toxicology, and food-safety (Dellafiora and Dall'Asta, 2017; Heck and Neely, 2020). Aquatic non-model species also saw increased sequencing efforts in recent years which yielded both draft and high-quality genome assemblies for a wide range of fish-

species (Malmstrøm et al., 2017). The increasing availability of transcriptomic and proteomic data for these species, provides multi-level omics information that can be used to validate and correct previous genome annotations (Prasad et al., 2017), which often are based on gene prediction models (Ruggles et al., 2017), and allow for the prediction of new genes (Ansong et al., 2008). Such proteogenomic analyses have been demonstrated for many organisms (Castellana and Bafna, 2010; Nesvizhskii, 2014) including, zebrafish (*Danio rerio*), where integrative analyses of transcriptome and proteome data was demonstrated to advance the understanding even of well annotated fish genomes (Kelkar et al., 2014).

The release of the Atlantic salmon genome (Davidson et al., 2010), and a comprehensive annotation of its genes (Lien et al., 2016) has significantly impacted omics research on wild and domesticated salmonids including toxicological research using transcriptomics and proteomics approaches (Bernhard et al., 2019; Søderstrøm et al., 2022). For marine aquaculture and ecosystem research, mechanistic toxicity assessments are needed when marine fish species such as Atlantic salmon are exposed to legacy (Lundebye et al., 2017; Nøstbakken et al., 2021) or emerging contaminants (Merel et al., 2019; Regueiro et al., 2017), and information on contaminant-specific metabolism and toxic mode of action (MOA) are lacking (Rasinger et al., 2022). Genomic assessment of genes involved in the metabolism of foreign compounds in several model teleost fish species (including zebrafish, medaka (*Oryzias latipes*), Atlantic killifsh (*Fundulus heteroclitus*), Atlantic cod, and three-spined stickleback) has identified integrative networks comprising key xenobiotic receptors, transcription factors, biotransformation enzymes, transporters, antioxidants, and metal- and heat-responsive genes (Eide et al., 2021) which collectively are known as the chemical defensome (Goldstone et al., 2006). Chemical defense proteins such as cytochromes P450

(Cyp) for example, catalyze the metabolic conversion of xenobiotics in the liver (Sprenger et al., 2022) and are involved in the excretion of metabolites by the hepatocytes (Hammer et al., 2021). Inter-species differences in the expression, abundances and activities of orthologs of different Cyp genes related to xenobiotic metabolism including Cyp1, Cyp2, Cyp3, and Cyp4 were reported for both fish (Eide et al., 2021), and mammals and their respective cell model systems (Hammer et al., 2021). The use of omics and cell model systems as integral components of New Approach Method (NAM) batteries are increasingly being implemented in human health, animal health and ecological risk assessment (Astuto et al., 2022; Krewski et al., 2020). For Atlantic salmon, an in-depth omics-based analysis of the toxicant induced chemical defensome in vivo and in vitro has yet to be performed.

In this study, we apply proteogenomic strategies (Woo et al., 2015, 2014a, 2014b) to the Atlantic salmon in order to validate and revise the current reference genome annotation, and to provide a framework for applying proteogenomic approaches to other fish species of commercial and environmental interest. We provide both RNA-seq and LC-MS/MS datasets for salmon liver and 3D primary liver hepatocyte cultures and shotgun proteomics mass spectrometry (LC-MS/MS) datasets for twelve salmon tissues. Given the lack of mass-spectrometry-based proteomic data and the common use of primary liver hepatocytes as an *in vitro* system in salmon research (Olsvik and Søfteland, 2020; Søderstrøm et al., 2022; Olsvik et al., 2017; Søderstrøm et al., 2022; Søfteland et al., 2014), these datasets (127 LC-MS/MS experiments) are valuable for the scientific community to support the rapidly advancing development of omics tools for mechanistic research in non-model species and their respective cell models. Here, as a proof of concept, we use proteogenomics expression matrices to describe similarities and differences between Atlantic salmon liver tissue and 3D

primary hepatocyte samples, and to visualize the chemical the Atlantic salmon defensome to facilitate the use of cell models for in vitro – in vivo extrapolation (IVIVE) for target tissue toxicity assessments.

1.3: Materials and Methods

Samples

Atlantic salmon (*Salmo salar L.*) tissue samples of both genders (SalmoBreed strain) were obtained from a previously published trial (Berntssen et al., 2021). For proteogenomic analyses, 22 tissues samples comprising of liver, brain, eye, gill, gut, pyloric caecum, head kidney, heart, muscle, ovary, skin, and spleen were collected from two randomly chosen fish. In addition, liver samples were collected from 44 randomly selected fish and a total of 60 3D salmon primary hepatocyte samples were included. In vitro samples were prepared as previously described (Olsvik and Søfteland, 2020; Søfteland et al., 2009). The hepatocytes used in this study had cell viability between 82-98%. The hepatocytes were plated on 2 $\mu g/cm^2$ laminin (Sigma-Aldrich, Oslo, Norway) coated 3D Alvetex Scaffolds (200 µm cross-linked polystyrene membranes, 42 µm mean void size, REPROCELL, Glasgow, United Kingdom) in 12 well culture plates (FALCON, Corning, VWR, Bergen, Norway). 2.9×10^6 cells in 2.5 ml complete L-15 medium were used per well.

RNAseq analyses

Total RNA was extracted from liver and primary hepatocyte samples using BioRobot® EZ1 and RNA Tissue Mini Kit (Qiagen, Hilden, Germany) including DNAase treatment as instructed in the RNA Tissue Mini Kit manual (Qiagen, Hilden, Germany).

NanoDrop ND-1000 UV–vis Spectrophotometer (NanoDrop Technologies, Wilmington, USA) was used to measure RNA quality/purity. RNA integrity was analyzed using an Agilent 2100 Bioanalyzer and the RNA 6000 Nano LabChip kit (Agilent Technologies, Palo Alto, USA). Samples had 260/280 and 260/230 ratios between 2.0 – 2.1 and between 2.0 - 2.2 respectively. Library preparation were performed by the Norwegian Sequencing Centre (www.sequencing.uio.no). DNA libraries from individual samples were constructed using TruSeq Stranded mRNA Library Prep Kit (Illumina) and standard Illumina adaptors for multiplexing as described by the manufacturer (Illumina). Individual libraries were sequenced using the NextSeq Illumina platform according to manufacturer instructions, generating single end 75bp read libraries. TrimGalore 0.4.2 tool (https://github.com/FelixKrueger/TrimGalore) was applied for removing adaptors and for quality trimming using default parameters. Sequence quality for each sample was investigated using FastQC imbedded in TrimGalore. SAM files were generated by mapping to the above mentioned salmon genome using the Hisat2 short read aligner version 2.0.4. Transcript abundance for the individual libraries was estimated using FeatureCounts (42) of the Subread package (http://subread.sourceforge.net/).

Proteomics mass spectrometry

Protein extraction from livers and tryptic protein digestions were performed according to in-house standardized protocols. In short, liver tissue samples were lysed by sonication in 10µl lysis buffer/mg tissue (4% SDS, 0.1M Tris-HCl pH 7.6), using an ultrasonication rod (Q55 Sonicator, Qsonica, CT, USA) at 30% amplitude for 30 sec, or until tissue was dissolved. 3D hepatocytes cultured in Alvetex scaffolds were washed with PBS.

The hepatocytes were incubated with 600 µL 2x DIGE lysis solution (Rasinger et al., 2014) for 30 min on a rotating platform (100 rpm) at room temperature. The lysate was homogenized 5 times with a 20-gauge needle. The cell suspension was transferred to a centrifugation tube and centrifuged at 12000 rpm for 30 min at 4 °C. The supernatant was transferred to a new centrifugation tube and flash frozen and stored at −80 °C. Supernatants were collected and protein concentration were determined PierceTM BCA Protein assay kit (Thermo Scientific). 1M dithiotreitol was added to the lysates, to obtain a final concentration of 0.1M, and the mix was incubated at 95ºC for 5 min. The samples were further processed using a Filter Aided Sample Preparation (FASP) protocol with trypsin digestion as described by (Wisniewski et al., 2009).

Mass spectrometry analyses were performed following standardized protocols at the Proteomics Unit at the University of Bergen, Norway (PROBE). In short, for Orbitrap Elite data-dependent-acquisition (DDA) analysis, between 0.5 and 1 ug protein, dissolved as tryptic peptides in 2% acetonitrile (ACN), 0.1% formic acid (FA), were injected into an Ultimate 3000 RSLC system (Thermo Scientific, Sunnyvale, California, USA) connected online to a linear quadrupole ion trap-orbitrap (LTQ-Orbitrap Elite) mass spectrometer (Thermo Scientific, Bremen, Germany), which in turn was equipped with a nanospray Flex ion source (Thermo Scientific). Samples were loaded and desalted on a pre-column (Acclaim PepMap 100, 2cm x 75µm ID nanoViper column, packed with 3µm C18 beads) at a flow rate of 5µl/min for 5 minutes (min) with 0.1% trifluoroacetic acid (TFA). Peptides were separated during a biphasic ACN gradient from two nanoflow UPLC pumps (flow rate of 270 nl /min) on a 50 cm analytical column (Acclaim PepMap 100, 50 cm x 75µm ID nanoViper column, packed with 3 µm C18 beads). Solvent A and B was 0.1% TFA (vol/vol)

in water and 100% ACN, respectively. The gradient composition was 5% B during trapping (5min) followed by 5-7% B (over 1min), 7-21% B (134min), 21-34% B (45min), and 34-80% B (10min). Elution of very hydrophobic peptides and conditioning of the column were performed during 20 minutes isocratic elution with 80% B and 20 minutes isocratic elution with 5% B, respectively.

The eluting peptides from the LC-column were ionized in the electrospray and analyzed by the LTQ-Orbitrap Elite. The mass spectrometer was operated in DDA-mode to automatically switch between full scan MS and MS/MS acquisition. Instrument control was through Tune 2.7.0, and Xcalibur 2.2. Survey full scan MS spectra (from m/z 300 to 2,000) were acquired in the Orbitrap with a resolution of R = 240,000 at m/z 400 (after accumulation to a target value of 1e6 in the linear ion trap with maximum allowed ion accumulation time of 300ms). The 12 most intense eluting peptides above an ion threshold value of 3000 counts, and charge states of two or higher, were sequentially isolated to a target value of 1e4 and fragmented in the high-pressure linear ion trap by low-energy collision-induced-dissociation (CID) with normalized collision energy of 35% and wideband-activation enabled. The maximum allowed accumulation time for CID was 150 ms, the isolation with maintained at 2 Da, activation q = 0.25, and activation time of 10 ms. The resulting fragment ions were scanned out in the low-pressure ion trap at normal scan rate, and recorded with the secondary electron multipliers. One MS/MS spectrum of a precursor mass was allowed before dynamic exclusion for 40s. Lock-mass internal calibration was not enabled.

Proteogenomics analyses

To construct tissue-specific splice graph databases, raw RNA-seq reads were retrieved from BioProjects PRJNA72713 and PRJNA260929, trimmed using TrimGalore (v.0.5.0), merged if from BioProject PRJNA260929, and mapped to the *Salmo salar* RefSeq genome (GCF_000233375.1_ICSASG_v2) using STAR (v.2.7.0.f) in 2-pass mode. Genome indexes for GCF_000233375.1_ICSASG_v2_genomic.fna were generated for reads from Illumina MiSeq and Illumina HiSeq 2000 instruments (genome index creation: hiseq_2x100: --sjdbOverhang 99 --sjdbGTFtagExonParentTranscript Parent --genomeChrBinNbits 14, miseq_2x250: --sjdbOverhang 249 --sjdbGTFtagExonParentTranscript Parent --genomeChrBinNbits 14). RNA-seq read alignments in SAM format were then used to construct a splice graph database using the SpliceDB method (minimum number of reads: 3) (Woo et al., 2014). Similarly, a splice graph database was constructed for the liver and 3D primary hepatocyte dataset using matched RNA-seq data (minimum number of reads: 10).

The proteogenomics workflow searches MS/MS spectra against reference protein and proteogenomic databases using a multi-stage FDR approach similar to that described by Woo et al. (2014). Briefly, MS/MS spectra were first searched against the *Salmo salar* RefSeq proteome (GCF_000233375.1_ICSASG_v2_protein.faa) to identify reference proteins and peptides. Unidentified spectra not passing a 1% PSM-level FDR are subsequently searched against a proteogenomic-based splice graph protein database using MS-GF+ (Kim and Pevzner, 2014). A final search against an enhanced database containing RefSeq proteins, Augustus predictions, and Ensembl proteins supported by proteogenomic events was conducted using MS-GF+ for downstream analyses. For the *in vitro* and *in vivo*

dataset, the enhanced database contained 97,555 RefSeq protein sequences, 71 Ensembl protein sequences, and 1 Augustus sequence. For the tissue dataset, the enhanced database contained 97,555 RefSeq protein sequences, 107 Ensembl protein sequences, and 8 Augustus sequences. In addition to sequences corresponding to the 3 novel Augustus gene predictions, additional Augustus sequences were included if proteogenomic evidence pointed to corrections in RefSeq or Ensembl annotations.

Genomic regions flanking identified proteogenomic events were submitted to Blastx, where high-scoring segment pairs (HSPs; *e*-value ≤ 1e-10) to proteins in the non-redundant (nr) database provided protein coding support for the genomic region. Specifically, sub-sequences of the *Salmo salar* genome (GCF_000233375.1_ICSASG_v2), defined as the nucleotides spanning the genomic coordinates of identified proteogenomic events and 350bp of flanking sequences were submitted to Blastx (online server) for a six-frame translation and comparison against protein sequences in the nr database. HSPs were included as hints to the gene prediction software Augustus. For predicted genes and transcripts, corresponding protein sequences were submitted to Blastp for comparison against protein sequences in the nr database. Blastp hits with significant alignments (*e*-value ≤ 1e-10, alignment ≥10 amino acids) were considered as orthologs in support of the predicted gene structure.

Along with identified proteogenomic events, HSPs from Blastx, and annotations from the *Salmo salar* genome (i.e., RefSeq GCF_000233375.1_ICSASG_v2 and Ensembl release-99), identified peptides from a MSGF+ search against the RefSeq *Salmo salar* proteome along with Ensembl proteins supported by proteogenomic events were provided as hints to Augustus (v 3.3.3) for gene-prediction (parameters: --codingseq=on, --species= zebrafish, --alternatives-from-evidence=true, --allow_hinted_splicesites=atac, --

extrinsicCfgFile= extrinsic.M.RM.E.W.P_revised_930.cfg, --softmasking=on). Hints were limited to those within 60,000 bp of a given proteogenomic event group. Predicted genes with evidence from at least one proteomic hint were retained for further analysis.

To compute a protein-level FDR, proteins are scored based on precursors (combination of peptide sequence and charge) passing a 1% level FDR, where the score of a precursor is defined as the SpecEValue of its best scoring PSM. Specifically, the score of a protein is initialized as the sum of the negative log of the SpecEValue of its precursors. After choosing the highest scoring protein, precursors to that protein are removed and protein scores are recalculated. This process is iterated until no precursors are left. A protein false discovery rate is then computed as the number of decoy proteins/target proteins.

Protein expression and chemical defensome analyses

MS/MS spectra were searched against the RefSeq *Salmo salar* proteome with additional Ensembl and Augustus predicted proteins supported by proteogenomic and comparative genomic evidence (Tissue dataset: 107 Ensembl and 8 Augustus predictions; Liver and Hepatocyte dataset: 71 Ensembl and 1 Augustus predictions). Ensembl proteins are defined as translations resulting from Ensembl gene predictions (Salmo_salar.ICSASG_v2.pep.all.fa). Raw spectral counts, S_p, for identified proteins (1% protein-level FDR) in each sample are based on spectra passing a 1% PSM-level and 1% precursor-level FDR, where c_p is the number of PSMs representing peptide, p, mapping to a protein, P, and N(p) is the number of proteins mapping to a peptide p.

$$S_p = \sum_{p \in P}^{P} \frac{c_p}{N(p)}$$

A normalized spectral count, PSK_p, is implemented to account for protein length.

$$SPK_p = \frac{S_p}{|P| \times 0.01}$$

$$SPK_{allP} = \sum_{P \in all\ P}^{all\ P} SPK_p$$

$$PSK_p = \frac{SPK_p}{\frac{SPK_{all\ P}}{1000}}$$

To generate a heatmap of the expression matrix, hierarchical clustering is performed separately on the spectrum files (columns) and proteins (rows) (ward.D2, Euclidean) of the normalized PSK expression matrix, and the matrix further log transformed for visualization purposes.

A list of genes included in the chemical defensome of five model teleost fish was created based on the supplementary material provided in (Eide et al., 2021). For sanity checking, files and codes created were first used to find chemical defensome related genes in the Zebra fish (*Danio rerio*) reference proteome (UP000000437, March 2022) and the results obtained were compared with data presented in (Eide et al., 2021). Following the successful completion of the sanity check (data not shown), the defensome gene list from (Eide et al., 2021) was matched against the Atlantic Salmon reference proteomes (UP000087266; March 2022) and the protein expression matrix obtained in the present study using a proteogenomic strategy. For improving dentification rates of chemical defensome proteins in salmon, using PAW_BLAST db_to_db_blaster.py (https://github.com/pwilmart/PAW_BLAST) multi-tissue and liver specific proteogenomic expression matrices were matched against zebra fish reference proteome (UP000000437, March 2022). Outputs were recoded using tidyverse (v.1.3.0) functions (Wickham et al.,

2019); plots were created using the packages ggplot2 (v.3.3.2), UpSetR (v.1.4.0) and ComplexHeatmap (v.2.4.3).

1.4: Results

Proteogenomic support for novel gene and Ensembl assembly annotations

For a systems-wide representation of the Atlantic salmon, samples from twelve salmon tissues, as also salmon primary hepatocyte cultures, were subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS). Spectra were searched against the reference salmon proteome (RefSeq assembly GCF_000233375.1) followed by a custom proteogenomic splice database using a multi-stage FDR approach (Woo et al., 2015) (**Fig. 1.1A**, panel 1). 79,498 peptides matched the 40,590 RefSeq gene structures, and included 19,132 spliced peptides, providing strong proteomic validation of these transcripts. However, 274 peptides representing 183 distinct proteogenomic events matched genomic loci where transcripts had been observed but were not accompanied by a RefSeq gene annotation. We next checked if these events were supported by other annotations of the genome, and found that 86 of 274 identified proteogenomic event peptides supported gene annotations in the Ensembl assembly, suggesting that reference proteomes based solely on the RefSeq assembly may not be comprehensive.

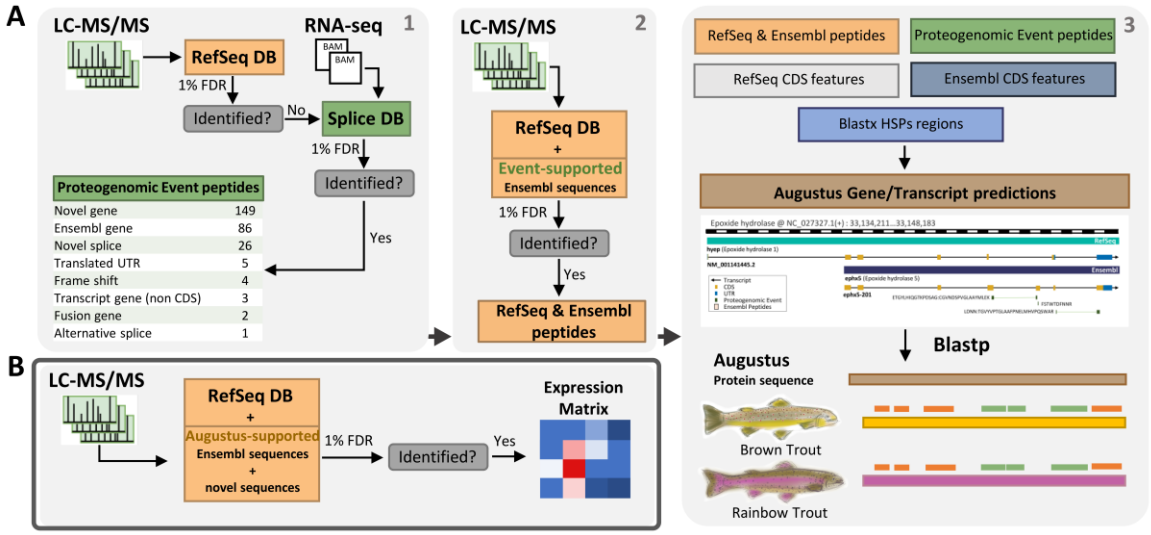
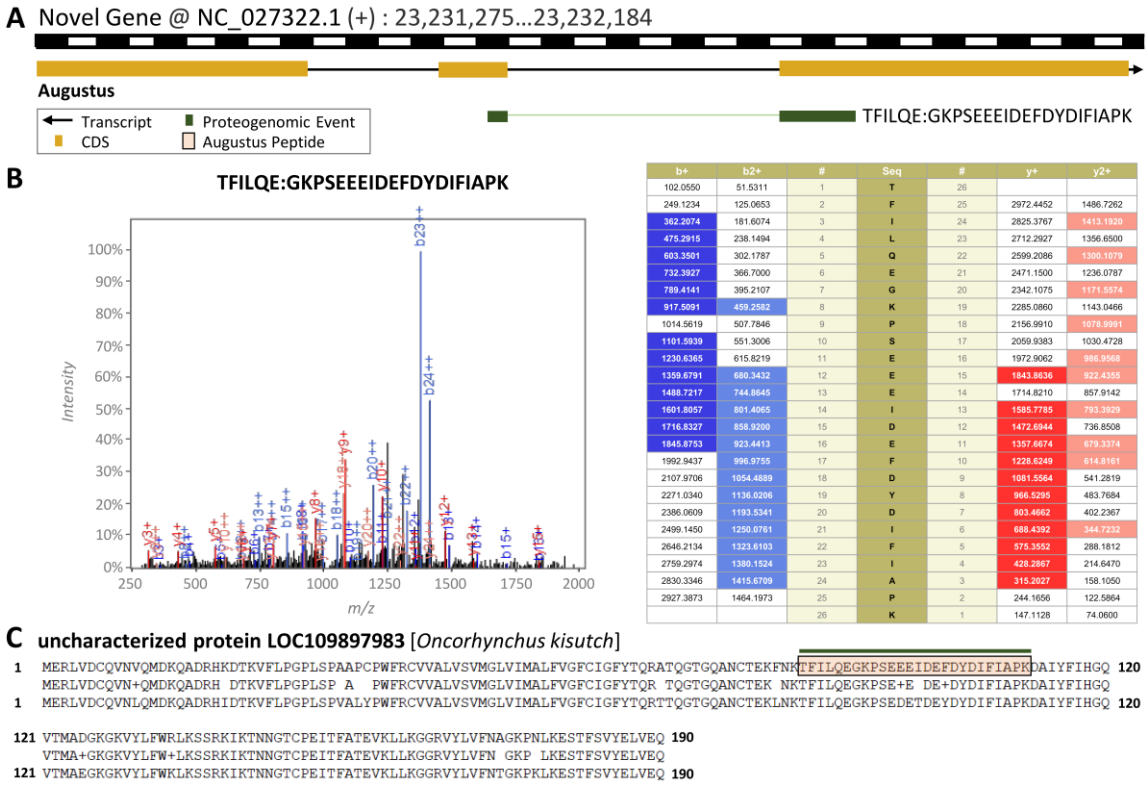


Figure 1.1: Proteogenomics workflow for gene prediction and validation of reference Salmon annotations. (A) MS/MS spectra were searched against the reference *Salmo salar* proteome and a proteogenomic splice database in a multi-stage FDR approach (1). Proteogenomic databases were constructed using either proteomic sample-matched RNA-seq reads or publicly available transcriptomic data. Of the 276 proteogenomic events identified in (1), 86 support gene annotations in the Ensembl assembly. Spectra were therefore searched against the reference proteome appended with proteogenomic supported Ensembl proteins to provide additional hints for gene prediction (2). Identified peptides from (1) and (2), RefSeq and Ensembl annotated coding regions (CDS), and Blastx high-scoring pair (HSP) regions were submitted as hints to Augustus for gene prediction (3). Corresponding protein sequences were subsequently submitted to Blastp for identification of homologs, providing additional comparative genomic support for predictions. **(B)** For downstream analyses, MS/MS spectra were searched against an enhanced reference proteome with additional novel proteins and Ensembl annotated proteins supported by Augustus predictions.

To detect novel gene structures and reference annotation corrections, we provided the gene prediction software Augustus (Stanke et al., 2004) with hints to coding sequence regions, including genomic coordinates for identified reference peptides and proteogenomic events, reference annotation features, and genomic regions that are also known to be protein-coding in other organisms (**Fig. 1.1A**, panel 3; methods). As additional hints from proteomic sources may increase the accuracy of Augustus gene predictions, we also included peptides that were identified from searching spectra a second time against the reference proteome

with additional proteogenomic supported Ensembl proteins (**Fig. 1.1A**, panel 2). This resulted in Augustus predictions for 71 transcripts (**Table S1.1**) supported by 681 peptides. The protein sequences of predicted transcripts were then submitted to Blast for the identification of homologs as further support for the prediction (**Fig. 1.1A**, panel 3).

While many transcripts matched Ensembl annotations, others provided corrections to RefSeq annotations. Of the 71 predicted transcripts, 62 were highly matched to Ensembl annotations, with 20 of the 62 without a RefSeq annotation in the region. We identified 3 novel genes in previously unannotated regions. One such example is a novel gene on chromosome 23 (NC_027322.1) that is supported by a proteogenomic peptide TFILQE:GKPSEEEIDEFDYDIFIAPK spanning the junction between exons 2 and 3 of the predicted transcript (Augustus_ID #65 in **Table S1.1**; **Fig. 1.2A**). The peptide annotated multiple peaks corresponding to b/y ions and explaining 17 of the 50 highest intensity peaks (**Fig. 1.2B**). A Blastp search of the predicted protein sequence against the non-redundant (nr) database resulted in a hit to an uncharacterized protein (LOC109897983; identity: 92.6%; e-value: 7.33e-129) in the coho salmon (*Oncorhynchus kisutch*), a closely related salmonid species (**Fig. 1.2C**). Together, these results suggest strong proteogenomic and comparative genomic evidence in support of the novel gene.



b+	b2+	#	Seq	#	y+	y2+
102.0550	51.5311	1	T	26		
249.1234	125.0653	2	F	25	2972.4452	1486.7262
362.2074	181.6074	3	I	24	2825.3767	1413.1920
475.2915	238.1494	4	L	23	2712.2927	1356.6500
603.3501	302.1787	5	Q	22	2599.2086	1300.1079
732.3927	366.7000	6	E	21	2471.1500	1236.0787
789.4141	395.2107	7	G	20	2342.1075	1171.5574
917.5091	459.2582	8	K	19	2285.0860	1143.0466
1014.5619	507.7846	9	P	18	2156.9910	1078.9991
1101.5939	551.3006	10	S	17	2059.9383	1030.4728
1230.6365	615.8219	11	E	16	1972.9062	986.9568
1359.6791	680.3432	12	E	15	1843.8636	922.4355
1488.7217	744.8645	13	E	14	1714.8210	857.9142
1601.8057	801.4065	14	I	13	1585.7785	793.3929
1716.8327	858.9200	15	D	12	1472.6944	736.8508
1845.8753	923.4413	16	E	11	1357.6674	679.3374
1992.9437	996.9755	17	F	10	1228.6249	614.8161
2107.9706	1054.4889	18	D	9	1081.5564	541.2819
2271.0340	1136.0206	19	Y	8	966.5295	483.7684
2386.0609	1193.5341	20	D	7	803.4662	402.2367
2499.1450	1250.0761	21	I	6	688.4392	344.7232
2646.2134	1323.6103	22	F	5	575.3552	288.1812
2759.2974	1380.1524	23	I	4	428.2867	214.6470
2830.3346	1415.6709	24	A	3	315.2027	158.1050
2927.3873	1464.1973	25	P	2	244.1656	122.5864
		26	K	1	147.1128	74.0600

Figure 1.2: Proteogenomics-supported novel gene prediction. (A) Augustus predicted a gene on chromosome 23 (NC_027322.1) in a region previously unannotated by RefSeq and Ensembl. A peptide identified in the proteogenomic database search (green) maps to CDS annotations in the predicted transcript. **(B)** Peptide-spectrum match (PSM) of the proteogenomic event peptide TFILQE:GKPSEEEIDEFDYDIFIAPK identified in the *Salmo salar* ovary tissue showing b-ion (blue) and y-ion (red) matches of charge +1 and +2 (fragment mass tolerance 0.4 Da). **(C)** A Blastp search of the predicted protein against the non-redundant database resulted in a hit to an uncharacterized protein (LOC109897983) in the salmonid *Oncorhynchus kisutch* (Coho salmon) (XP_020348278.1; e-value: 7.33e-129; identity: 176/190, 92.6%).

Of the 41 Augustus predictions supporting corrections to reference annotations, the myosin heavy chain 7 (myh7) gene on the negative strand of chromosome three (NC_027302.1) is an example where the RefSeq gene prediction is truncated and missing coding regions 3' downstream of the annotated stop codon (Augustus_ID #1; **Fig. 1.3A**). Primarily expressed in the cardiac muscle, the myh7 gene may also be expressed in other skeletal muscles. Indeed, seventeen proteogenomic event peptides were identified in the

salmon heart and eye tissues, with four being expressed in both. Mapping to CDS regions in the Ensembl gene (ENSSSAT00000137950.1), many of these peptides also spanned exon-exon junction sites (**Fig. 1.3A; Table S1.2**).

To rule out the possibility that the Ensembl gene was misannotated and has been accounted for elsewhere by the reference annotation, the corresponding protein (ENSSSAP00000103327.1) was submitted to Blastp for a search against the nr database. This resulted in a hit to a myosin-like protein on chromosome nineteen (XP_014015175.1) of the reference annotation. Despite an overall high-scoring alignment between the two protein sequences (identity: 88.4%; e-value: 0.0), in areas with poorer alignment, the Ensembl protein is supported by high quality proteogenomic event peptides (**Fig. 1.3B**). For example, the proteogenomic peptides LLTVLFQNYAGTDA:ADDSK and KKLESDTNQLQTEVEEAVQECR at positions 615 and 1733, respectively, of the Ensembl protein have high-quality peptide-spectrum matches (PSMs) with a high coverage of b- and y-ions (**Fig. S1.1**). Furthermore, a proteomic search against an enhanced reference database (methods) identified 198 peptides that span the length of the Ensembl protein (**Fig. 1.3B**), suggesting that a correction should be made to the current annotation.

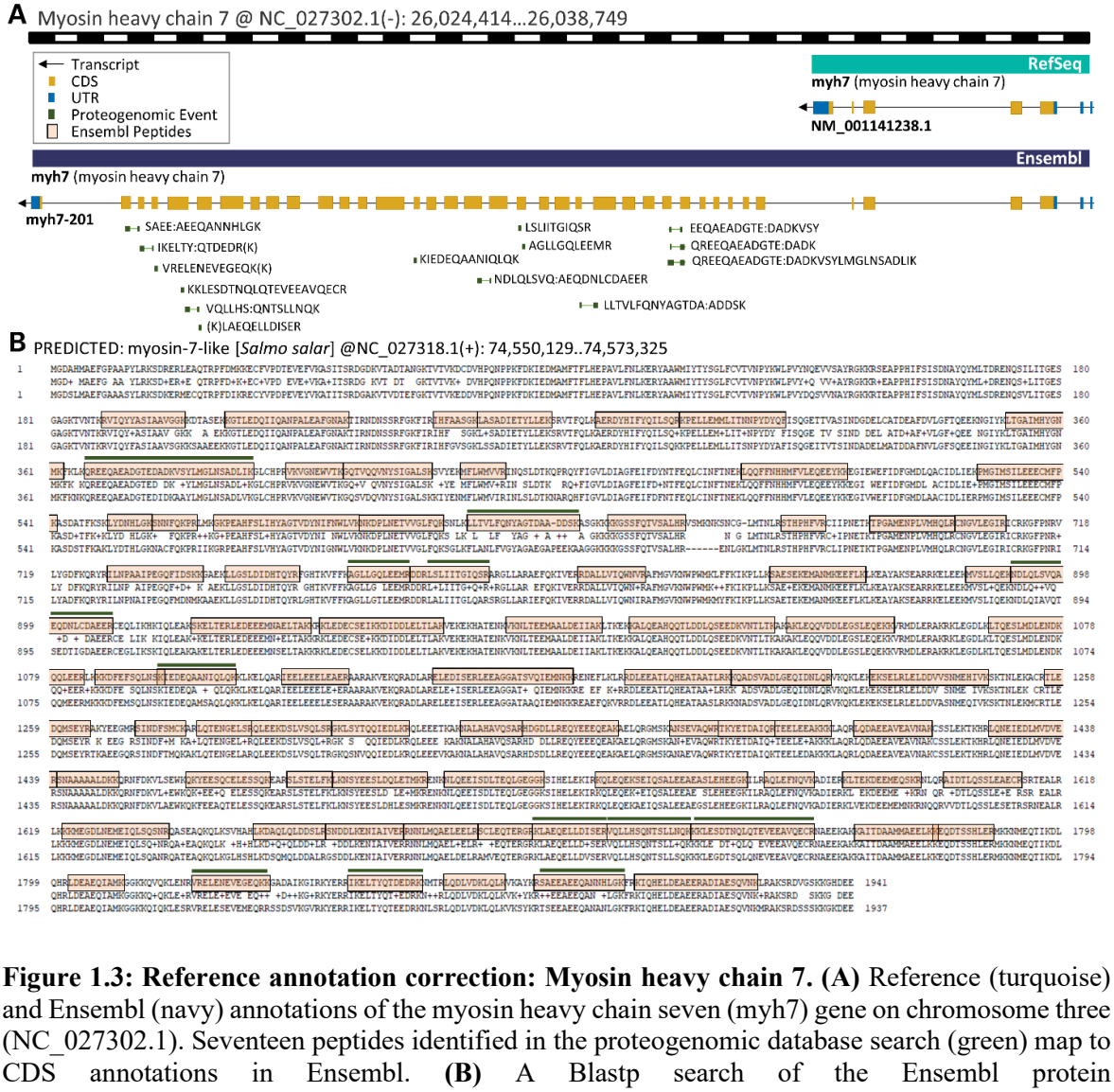


Figure 1.3: Reference annotation correction: Myosin heavy chain 7. (A) Reference (turquoise) and Ensembl (navy) annotations of the myosin heavy chain seven (myh7) gene on chromosome three (NC_027302.1). Seventeen peptides identified in the proteogenomic database search (green) map to CDS annotations in Ensembl. **(B)** A Blastp search of the Ensembl protein (ENSSSAP00000103327.1) against the non-redundant database resulted in a hit to a predicted myosin-7-like protein (XP_014015175.1) on chromosome 19 (NC_027318.1) of the *Salmo salar* genome (e-value: 0; identity: 1718/1943, 88.4%). A proteomic search against an enhanced reference database identified 198 (orange) peptides mapping to ENSSSAP00000103327.1, including the previously identified proteogenomic event peptides (green lines).

As another example, the Ensembl epoxide hydrolase gene (ENSSSAT00000145609.1) on the positive strand of chromosome 28 (NC_027327.1) represents either a correction to the reference annotation or an additional isoform (Augustus_ID #68; **Fig. 1.4A**). Specifically, we identified 20 peptides mapping to the

Ensembl protein (ENSSSAP00000110682.1), including three proteogenomic event peptides supporting the partial translation of a 3' untranslated region (UTR) and the addition of novel splice junctions. As before, we submitted a Blastp search of the Ensembl protein against the nr database to verify that the predicted transcript has not been mapped elsewhere in the reference annotation. While there was a hit to the salmon epoxide hydrolase 1-like protein (XP_014000050.1) on chromosome 15, there was poor overall alignment (identity: 58.2%; e-value: 4e-176), as also presence of high-quality proteogenomic peptides at regions with poorer alignment (**Fig. 1.4B**). These peptides, FSTWTDFNNR, LDNN:TGVYVPTGLAAFPNELMHVPQSWAR, and ETGYLHIQGTKPDSAG:CGVNDSPVGLAAYMLEK at positions 287, 344, and 254, respectively, of the Ensembl protein, have high-quality peptide-spectrum matches (PSMs) with a high coverage of b- and y-ions (**Fig. S1.2**). In contrast, the Ensembl protein has a higher scoring alignment (identity: 98.7%; e-value: 0.0) to the epoxide hydrolase 1-like isoform X1 (XP_029621434.1) from the brown trout (*Salmo trutta*), suggesting that the Ensembl annotation is supported by conserved coding regions (**Fig. 1.4B**).

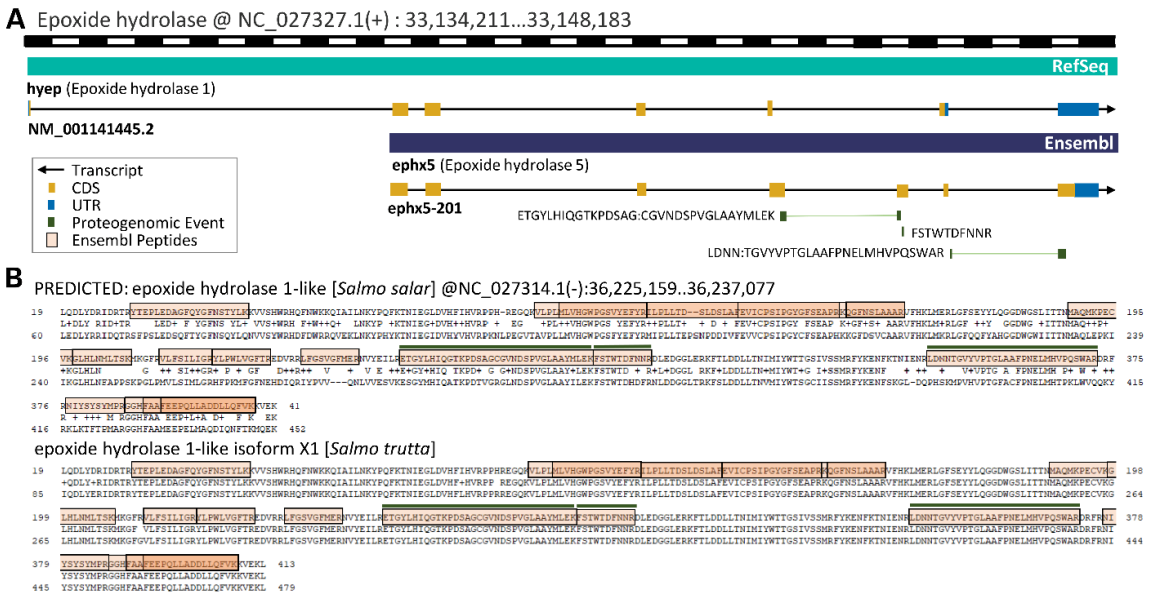


Figure 1.4: Additional isoform: Epoxide hydrolase. (A) Reference (turquoise) and Ensembl (navy) annotations of the epoxide hydrolase gene on chromosome 28 (NC_027327.1). Three peptides identified in the proteogenomic database search (green) map to CDS annotations in Ensembl. **(B)** A Blastp search of the Ensembl protein (ENSSSAP00000110682.1) against the non-redundant database resulted in hits to *Salmo salar* proteins, including XP_014000050.1 on chromosome 15 (NC_027314.1) (e-value: 4e-176; identity: 231/397, 58.2%), and other salmonids proteins, including XP_029621434.1 from Salmo trutta (e-value: 0; identity: 390/395, 98.7%). A proteomic search against an enhanced reference database identified 20 (orange) peptides mapping to ENSSSAP00000110682.1, including the previously identified proteogenomic event peptides (green lines).

Multi-tissue LC-MS/MS dataset: Tissue-specific protein profiles

The availability of tissue specific proteomic data provides an opportunity to study protein expression directly across different tissues. We generated 23 LC-MS/MS salmon tissue specific datasets as a resource for the community (MassIVE ID: MSV000089139 and **Table S1.3**). Given the strong proteogenomic evidence in support of Ensembl and Augustus predicted transcripts, we searched spectra again; this time against an enhanced reference database containing RefSeq proteins, Augustus predictions, and Ensembl proteins (**Fig. 1.1B**; methods). Given the strong proteogenomic evidence in support of Ensembl and

Augustus predicted transcripts, we searched spectra a final time against an enhanced reference database containing RefSeq proteins, Augustus predictions, and Ensembl proteins (**Fig. 1.1B**; methods). We identified a total of 9,392 proteins (59,899 peptides) across all tissues, with brain, ovary, and liver samples expressing the highest number of proteins uniquely identified in a specific tissue (**Fig. 1.5A**). Principal component analysis (PCA) of the normalized expression count values of the 600 most highly expressed proteins (methods) across all twenty-three LC-MS/MS samples showed strong clustering of samples based on the twelve tissue types, with principal component 1 and 2 explaining for 29.2% and 15.0% of variance, respectively (**Fig. 1.5B**).

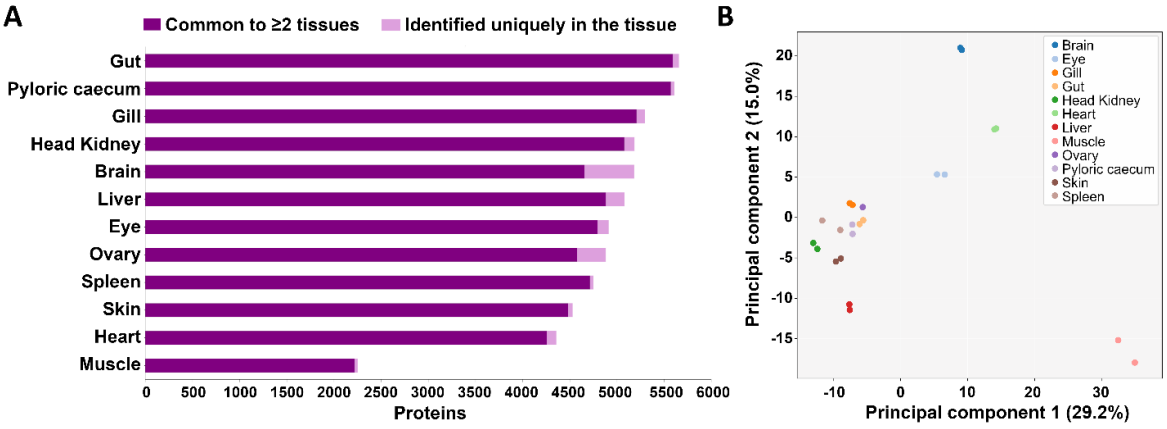


Figure 1.5: Distinct Proteomic Profiles of Salmon Tissues. (A) A total of 9,392 proteins (59,899 peptides) were identified in twelve tissue types. The highest number of proteins were identified in the Gut (5,675 proteins; 43,795 peptides), while the highest number of tissue-unique proteins were identified in the Brain (526 proteins**;** 1,467 peptides). **(B)** Principal component analysis (PCA) of the top 600 proteins across all twenty-three LC-MS/MS samples reveals grouping of samples based on the twelve tissue types, with principal component 1 and 2 explaining for 29.2% and 15.0% of variance, respectively.

Hierarchical clustering of the expression matrix (top 600 proteins) further revealed tissue-specific protein signatures (**Fig. S1.3**; **Table S1.4**). Cluster #1 was characterized by proteins found across all tissues, including those related to oxygen transportation in blood (**Fig. S1.4A**). The gill, spleen, head kidney, and heart showed a particularly high expression of these proteins, consistent with the role that these organs play in the vascular system. Cluster #2 contained proteins related to maintaining the integrity and function of the intestine, and fittingly, were also expressed in the pyloric caecum, which is adjacent to the gut (**Fig. S1.4B**). Cluster #3 contained proteins related to muscle anatomy and function, including actin, myosin, and those involved in the glycolytic process - these proteins are highly expressed in muscle tissue and not in other tissue types (**Fig. S1.4D**). Cluster #4 consisted of proteins related to heart function and structure, including those involved in calcium ion binding, ventricular cardiac myofibril assembly, heart contraction, and oxygen transportation (**Fig. S1.4C**). Cluster #6 is dominated by proteins found in the integrity of skin, including keratin and collagen related proteins (**Fig. S1.4E**). Together, these results validate the robustness of generating a matrix of protein expression values using a proteogenomic strategy.

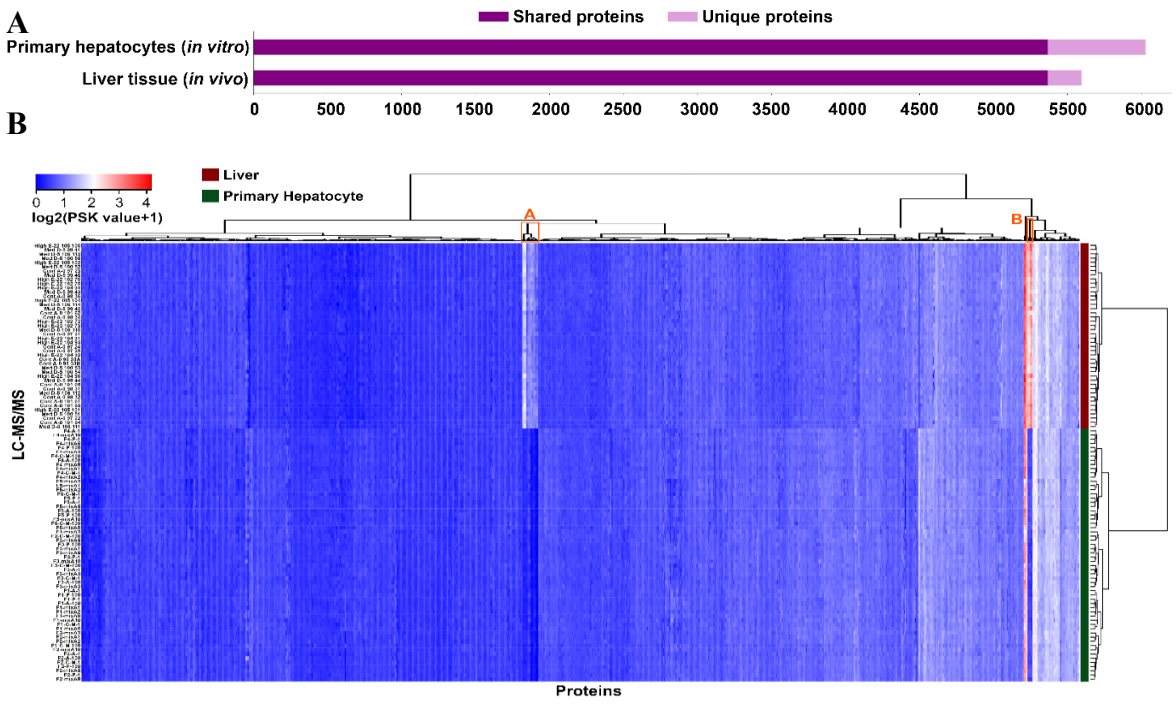


Figure 1.6: Salmon primary hepatocytes and liver tissue are strongly correlated. (A) Of the 6,280 identified proteins (39,676 peptides), 5,389 proteins (37,872 peptides) were identified in both primary hepatocytes and liver tissue, while 661 proteins (1,169 peptides) and 230 proteins (635 peptides) were identified uniquely, respectively. **(B)** Hierarchical clustering of the PSK normalized expression matrix (104 spectrum files, top 1,000 proteins) reveals a separation of Salmon primary hepatocyte (green) from Salmon liver tissue (red) samples. Proteins highly expressed in liver tissue but not primary hepatocyte samples include hemoglobin alpha and beta subunit proteins (orange rectangles).

Salmon primary hepatocytes as in vitro model for Atlantic salmon liver

LC-MS/MS analysis of liver samples collected from 44 randomly selected fish and a total of 60 3D salmon primary hepatocyte samples were performed and both datasets were made available on MassIVE (ID: MSV000089141) and **Table S1.5**. Of the 6,280 identified proteins (39,676 peptides), 5,389 proteins (37,872 peptides) were identified in both primary hepatocytes (*in vitro*) and liver tissue (*in vivo*) samples. Of the remaining 891 proteins, 661 proteins (1,169 peptides) were uniquely identified *in vitro*, and 230 proteins (635 peptides) were uniquely identified *in vivo* (**Fig. 1.6A**).

These results show a strong correlation between *in vitro* and *in vivo* expression. A fewer number of spectra acquired for *in vivo* samples may account for the difference seen in the number of identified peptides and proteins (**Fig. S1.6**). Hierarchical clustering of the expression matrix (104 spectrum files, top 1,000 proteins) revealed a separation of *in vitro* samples from *in vivo* samples (**Fig. 1.6B**). Proteins highly expressed *in vivo* but not *in vitro* (Cluster A, B) include proteins with functional and structural roles in blood, including hemoglobin alpha and beta subunit proteins and serum albumin (**Table S1.6**; **Fig. S1.5**).

The Atlantic salmon chemical defensome

Protein expression matrices of multi-tissue, liver and primary hepatocyte samples were matched against the full complement of the putative chemical defensome of model teleost fish species (**Table S1.7**) adapted from the supplementary material of (Eide et al., 2021). Using the salmon-specific annotation of the multi-tissue and the liver/primary hepatocyte protein matrices, a total of 68 (**Table S1.3**) and 59, (**Table S1.5**), respectively of chemical defensome specific proteins were identified. Following a PAW_BLAST search for zebra fish orthologs, the identification of chemical defensome specific proteins improved; a total of 306 and 248 proteins of the multi-tissue and liver/primary hepatocyte data, respectively were successfully mapped to gene patterns and ontology terms linked to proteins involved in the detoxification and clearance of xenobiotic compounds in Atlantic salmon (**Fig. 1.7A; Fig. 1.7B**). Hierarchical clustering based on the twelve tissue types of the 306 multi-tissue defensome proteins showed tissue-specific proteins signatures and clustering of samples based on tissue type (**Fig. S1.7**). Distinct and overlapping defensome protein expression were observed across all salmon tissues analyzed (**Fig. 1.7C**), and a large

degree of homology between *in vivo* and *in vitro* liver systems was observed; of 248 liver-specific defensome proteins present in the proteogenomics expression matrix; only two proteins were detected exclusively in the *in vitro* system (**Fig. 1.7D**).

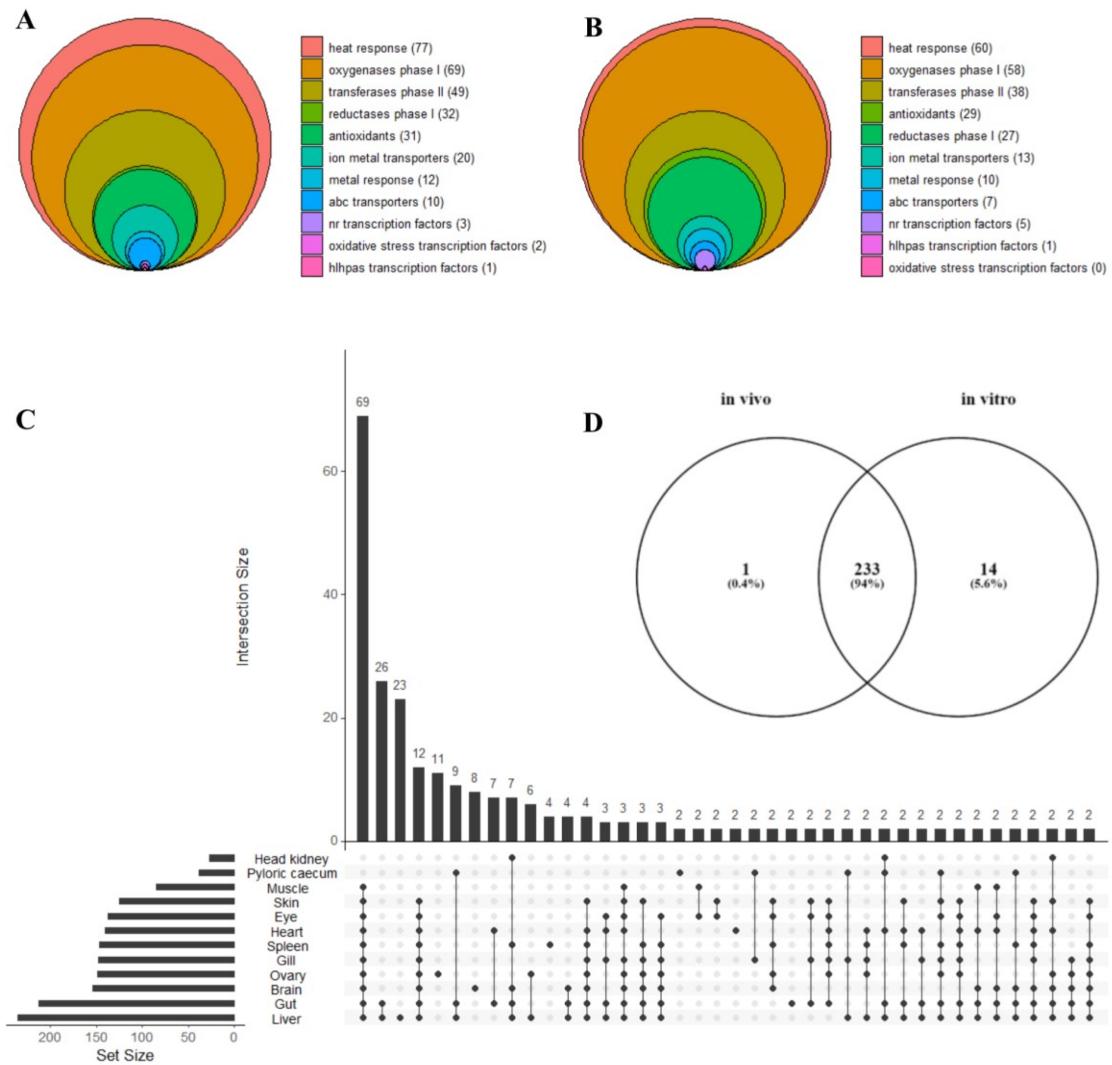


Figure 1.7: Atlantic salmon chemical defensome. Bubble plot of chemical defensome specific proteins of multi-tissue **(A)** and liver/hepatocyte **(B)** proteogenomics expression matrices **(C)** UpSet plot visualizing unique and overlapping defensome protein expression across all salmon tissues analyzed **(D)** Venn diagram displaying distinct and overlapping defensome proteins detected in vivo and in vitro liver systems.

1.5: Discussion

In this manuscript, we provide a comprehensive proteogenomic characterization of the Atlantic salmon (*Salmo salar*) genome by acquiring tandem spectra from multiple salmon tissues, salmon liver (Berntssen et al., 2021), and cultured hepatocytes commonly used for in vitro toxicity testing (Olsvik and Søfteland, 2020; Søderstrøm et al., 2022; Søfteland et al., 2014). We identified nearly 80,000 peptides providing direct evidence of translation for over 40,000 RefSeq structures. In addition, our data highlighted differences between the two annotation platforms RefSeq and Ensembl. Our data indicated that salmon reference proteomes based solely on the RefSeq assembly may not be as comprehensive as the ones based on Ensembl annotations. In the context of RNAseq mapping, it was found previously that Ensembl annotates more genes than RefSeq does (Zhao and Zhang, 2015). With functional annotation results potentially having a strong influence on the conclusions to be drawn (McCarthy et al., 2014), the selection of one annotation over another can have a strong effect on the overall outcome of omics studies (Frankish et al., 2015).

For transcriptomic-focused studies, it was suggested previously to choose a less complex genome annotation such as RefSeq, for reproducible and robust gene expression estimates; for exploratory research, a more complex genome annotation such as Ensembl, was recommended (Wu et al., 2013). However, until experimentally validated, genome annotations remain hypothetical (Prasad et al., 2017); several proteogenomics strategies to validate and correct previous annotations have been developed (Castellana and Bafna, 2010; Nesvizhskii, 2014; Ruggles et al., 2017) and have been applied on several different species (Prasad et al., 2017; Ruggles et al., 2017) including zebrafish (Kelkar et al., 2014). Like zebrafish, also salmon is a well-studied organism (Lien et al., 2016); however, a

proteogenomic analysis has yet to be performed on this important aquaculture species of high economic and cultural value (Fedoroff et al., 2022). Indeed, as was shown previously for the zebrafish (Kelkar et al., 2014), our integrative transcriptomics and proteomics analysis suggested corrections to hundreds of annotations of the Atlantic salmon genome and identified several novel genes that had been missed by other annotation efforts. Our results are consistent with earlier proteogenomic studies, suggesting that proteomic evidence can be used to confirm translation and as such be an important part of annotation pipelines (Castellana and Bafna, 2010; Nesvizhskii, 2014). In contrast with genomic annotation, but similar to transcript-based annotation, proteogenomic annotation provides a dynamic picture of the proteome that displays species- and tissue-specific expression patterns of proteins which can be analyzed and measured using standard shotgun proteomics approaches (Rasinger et al., 2016). The analysis of multiple salmon tissue confirmed this. Based on one single expression matrix, paired-samples from the same tissue clustered together, and separated from other tissues. For the generation of this expression matrix, our proteogenomic pipeline utilized the union of all peptides to support the annotation effort. We also developed a normalized spectral count approach estimating protein expression levels that corrected for protein length and total number of spectra acquired in the samples. This in part tackles some of the challenges associated with proteogenomics analyses, namely the integration and visualization of data sets comprising quantitative measurements across multiple data modalities (Ruggles et al., 2017). The simple protocol and automated expression matrix generation presented here, allows for our pipeline to be used to both perform mechanistic research in newly or only scarcely annotated non-model organisms (and their respective in vitro NAM counterparts) at the core of agricultural, veterinary and ecosystem research

(Heck and Neely, 2020), and to improve availability and design of peptide markers for use in food and feed forensics such as the detection of non-permitted protein material of animal origin in aquaculture feed (Belghit et al., 2021; Rasinger et al., 2016; Varunjikar et al., 2022a, 2022b).

Here, as a preliminary proof of concept for the applicability of our proteogenomic annotation and analysis pipeline for mechanistic research in modern-day marine aquaculture research, we compared the cultured primary hepatocyte proteome against liver tissue proteome using a single expression matrix. Except for the expression of blood related proteins, which were missing in the cultured cells probably as a result of cell extraction method and culturing conditions, we found a very high degree of consistency in expression profiles with a strong correlation between in vitro and in vivo expression. This is in line with previous studies on human samples in which also a high degree of similarity was observed when comparing proteomic expression data of human liver tissue and isolated hepatocytes (Vildhede et al., 2015) or genomic RNA-Seq data of liver models with in vivo human liver (Gupta et al., 2020). In addition to the global comparison of protein expression between liver tissue and hepatocytes, we specifically looked for the expression of chemical defensome related proteins *in vivo* and *in vitro*. The data revealed that based on proteogenomics analyses, a large degree of homology between *in vivo* and *in vitro* liver systems was observed indicating that primary salmon hepatocytes seem to be a suitable model system for mechanistic toxicity studies in Atlantic salmon liver.

1.6: Conclusion

Our results suggest that as demonstrated for zebrafish (Kelkar et al., 2014), an integrated transcriptomic and proteomic analysis (i.e., a proteogenomic characterization) of the Atlantic salmon improves upon current annotation, provides a window into tissue specific protein expression, and offers a road map for future, hypothesis driven investigations on the assessment of toxicodynamic consequences of multiple toxicant exposures in other species. The present study also provides evidence that for mechanistic studies of the chemical defensome extrapolations of in vitro toxicant studies into to whole tissue or organism responses is feasible. This may eventually allow for a better molecular understanding of how living organisms sense and react to the environment and its stressors supporting a more informed analysis and interpretation of food-safety and (eco)toxicology studies in non-model species.

Acknowledgements

Chapter 1 includes content of material being prepared for submission by Lin, Miin S.; Lie, Kai Kristoffer; Varunjikar, Madhushri Shrikant; Søfteland, Liv; Dellafiora, Luca; Dorne, Jean-Lou Christian Michel; Ørnsrud, Robin; Sanden, Monica; Berntssen, Marc; Bafna, Vineet; Rasinger, Josef D. in a manuscript with a tentative title of "Proteogenomic analysis of multi-tissue data for *in vivo – in vitro* extrapolations of the Atlantic Salmon (*Salmo salar*) chemical defensome." The dissertation author was the primary investigator and author of this paper.

1.7: References

Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S., Smith, R.D. (2008). Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief. Funct. Genomic. Proteomic. 7, 50–62.

Astuto, M.C., Di Nicola, M.R., Tarazona, J.V., Rortais, A., Devos, Y., Liem, A.K.D., Kass, G.E.N., Bastaki, M., Schoonjans, R., Maggiore, A., Charles, S., Ratier, A., Lopes, C., Gestin, O., Robinson, T., Williams, A., Kramer, N., Carnesecchi, E., Dorne, J.-L.C.M. (2022). In Silico Methods for Environmental Risk Assessment: Principles, Tiered Approaches, Applications, and Future Perspectives, in: Benfenati, E. (Ed.), In Silico Methods for Predicting Drug Toxicity. Springer US, New York, NY, pp. 589–636.

Belghit, I., Varunjikar, M., Lecrenier, M.-C., Steinhilber, A.E., Niedzwiecka, A., Wang, Y.V., Dieu, M., Azzollini, D., Lie, K., Lock, E.-J., Berntssen, M.H.G., Renard, P., Zagon, J., Fumière, O., van Loon, J.J.A., Larsen, T., Poetz, O., Braeuning, A., Palmblad, M., Rasinger, J.D. (2021). Future feed control – Tracing banned bovine material in insect meal. Food Control 108183.

Bernhard, A., Rasinger, J.D., Betancor, M.B., Caballero, M.J., Berntssen, M.H.G., Lundebye, A.-K., Ørnsrud, R. (2019). Tolerance and dose-response assessment of subchronic dietary ethoxyquin exposure in Atlantic salmon (Salmo salar L.). PLoS One 14, e0211128.

Bernhard, A., Rasinger, J.D., Wisløff, H., Kolbjørnsen, Ø., Secher Myrmel, L., Berntssen, M.H.G., Lundebye, A.-K., Ørnsrud, R., Madsen, L. (2018). Subchronic dietary exposure to ethoxyquin dimer induces microvesicular steatosis in male BALB/c mice. Food Chem. Toxicol. 118, 608–625.

Berntssen, M.H.G., Rosenlund, G., Garlito, B., Amlund, H., Sissener, N.H., Bernhard, A., Sanden, M. (2021). Sensitivity of Atlantic salmon to the pesticide pirimiphos-methyl, present in plant-based feeds. Aquaculture 531, 735825.

Castellana, N., Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: a computational perspective. J. Proteomics 73, 2124–2135.

Davidson, W.S., Koop, B.F., Jones, S.J.M., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., Omholt, S.W. (2010). Sequencing the genome of the Atlantic salmon (Salmo salar). Genome Biol. 11, 403.

Dellafiora, L., Dall'Asta, C. (2017). Forthcoming Challenges in Mycotoxins Toxicology Research for Safer Food-A Need for Multi-Omics Approach. Toxins 9.

Eide, M., Zhang, X., Karlsen, O.A., Goldstone, J.V., Stegeman, J., Jonassen, I., Goksøyr, A. (2021). The chemical defensome of five model teleost fish. Sci. Rep. 11, 1–13.

Fedoroff, N., Benfey, T., Giddings, L.V., Jackson, J., Lichatowich, J., Lovejoy, T., Stanford, J., Thurow, R.F., Williams, R.N. (2022). Biotechnology can help us save the genetic heritage of salmon and other aquatic species. Proc. Natl. Acad. Sci. U. S. A. 119, e2202184119.

Frankish, A., Uszczynska, B., Ritchie, G.R.S., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R., Harrow, J. (2015). Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics 16 Suppl 8, S2.

Goldstone, J.V., Hamdoun, A., Cole, B.J., Howard-Ashby, M., Nebert, D.W., Scally, M., Dean, M., Epel, D., Hahn, M.E., Stegeman, J.J. (2006). The chemical defensome: environmental sensing and response genes in the Strongylocentrotus purpuratus genome. Dev. Biol. 300, 366–384.

Gupta, R., Schrooders, Y., Hauser, D., van Herwijnen, M., Albrecht, W., Ter Braak, B., Brecklinghaus, T., Castell, J.V., Elenschneider, L., Escher, S., Guye, P., Hengstler, J.G., Ghallab, A., Hansen, T., Leist, M., Maclennan, R., Moritz, W., Tolosa, L., Tricot, T., Verfaillie, C., Walker, P., van de Water, B., Kleinjans, J., Caiment, F. (2020). Comparing in vitro human liver models to in vivo human liver using RNA-Seq. Arch. Toxicol.

Hammer, H., Schmidt, F., Marx-Stoelting, P., Pötz, O., Braeuning, A. (2021). Cross-species analysis of hepatic cytochrome P450 and transport protein expression. Arch. Toxicol. 95, 117–133.

Hampel, M., Alonso, E., Aparicio, I., Santos, J.L., Leaver, M. (2015). Hepatic proteome analysis of Atlantic salmon (Salmo salar) after exposure to environmental concentrations of human pharmaceuticals. Mol. Cell. Proteomics 14, 371–381.

Heck, M., Neely, B.A. (2020). Proteomics in Non-model Organisms: A New Analytical Frontier. J. Proteome Res. 19, 3595–3606.

Kelkar, D.S., Provost, E., Chaerkady, R., Muthusamy, B., Manda, S.S., Subbannayya, T., Selvan, L.D.N., Wang, C.-H., Datta, K.K., Woo, S., Dwivedi, S.B., Renuse, S., Getnet, D., Huang, T.-C., Kim, M.-S., Pinto, S.M., Mitchell, C.J., Madugundu, A.K., Kumar, P., Sharma, J., Advani, J., Dey, G., Balakrishnan, L., Syed, N., Nanjappa, V., Subbannayya, Y., Goel, R., Prasad, T.S.K., Bafna, V., Sirdeshmukh, R., Gowda, H., Wang, C., Leach, S.D., Pandey, A. (2014). Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. Mol. Cell. Proteomics 13, 3184–3198.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5, 5277.

Krewski, D., Andersen, M.E., Tyshenko, M.G., Krishnan, K., Hartung, T., Boekelheide, K., Wambaugh, J.F., Jones, D., Whelan, M., Thomas, R., Yauk, C., Barton-Maclaren, T., Cote, I. (2020). Toxicity testing in the 21st century: progress in the past decade and future perspectives. Arch. Toxicol. 94, 1–58.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Våge, D.I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., Davidson, W.S. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205.

Lundebye, A.-K., Lock, E.-J., Rasinger, J.D., Nøstbakken, O.J., Hannisdal, R., Karlsbakk, E., Wennevik, V., Madhun, A.S., Madsen, L., Graff, I.E., Ørnsrud, R. (2017). Lower levels of Persistent Organic Pollutants, metals and the marine omega 3-fatty acid DHA in farmed compared to wild Atlantic salmon (*Salmo salar*). Environ. Res. 155, 49–59.

Malmstrøm, M., Matschiner, M., Tørresen, O.K., Jakobsen, K.S., Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Sci Data 4, 160132.

Martyniuk, C.J., Denslow, N.D. (2009). Towards functional genomics in fish using quantitative proteomics. Gen. Comp. Endocrinol. 164, 135–141.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B., Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6, 26.

Mellingen, R.M., Myrmel, L.S., Lie, K.K., Rasinger, J.D., Madsen, L., Nøstbakken, O.J. (2021). RNA sequencing and proteomic profiling reveal different alterations by dietary methylmercury in the hippocampal transcriptome and proteome in BALB/c mice. Metallomics 13.

Mellingen, R.M., Myrmel, L.S., Rasinger, J.D., Lie, K.K., Bernhard, A., Madsen, L., Nøstbakken, O.J. (2022). Dietary Selenomethionine Reduce Mercury Tissue Levels and Modulate Methylmercury Induced Proteomic and Transcriptomic Alterations in Hippocampi of Adolescent BALB/c Mice. Int. J. Mol. Sci. 23, 12242.

Merel, S., Regueiro, J., Berntssen, M.H.G., Hannisdal, R., Ørnsrud, R., Negreira, N. (2019). Identification of ethoxyquin and its transformation products in salmon after controlled dietary exposure via fish feed. Food Chem. 289, 259–268.

Nesvizhskii, A.I. (2014). Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125.

Nøstbakken, O.J., Rasinger, J.D., Hannisdal, R., Sanden, M., Frøyland, L., Duinker, A., Frantzen, S., Dahl, L.M., Lundebye, A.-K., Madsen, L. (2021). Levels of omega 3 fatty acids, vitamin D, dioxins and dioxin-like PCBs in oily fish; a new perspective on the

reporting of nutrient and contaminant data for risk-benefit assessments of oily seafood. Environ. Int. 147, 106322.

Olsvik, P.A., Berntssen, M.H.G., Søfteland, L. (2017). In vitro toxicity of pirimiphos-methyl in Atlantic salmon hepatocytes. Toxicol. In Vitro 39, 1–14.

Olsvik, P.A., Søfteland, L. (2020). Mixture toxicity of chlorpyrifos-methyl, pirimiphos-methyl, and nonylphenol in Atlantic salmon (*Salmo salar*) hepatocytes. Toxicol Rep 7, 547–558.

Prasad, T.S.K., Mohanty, A.K., Kumar, M., Sreenivasamurthy, S.K., Dey, G., Nirujogi, R.S., Pinto, S.M., Madugundu, A.K., Patil, A.H., Advani, J., Manda, S.S., Gupta, M.K., Dwivedi, S.B., Kelkar, D.S., Hall, B., Jiang, X., Peery, A., Rajagopalan, P., Yelamanchi, S.D., Solanki, H.S., Raja, R., Sathe, G.J., Chavan, S., Verma, R., Patel, K.M., Jain, A.P., Syed, N., Datta, K.K., Khan, A.A., Dammalli, M., Jayaram, S., Radhakrishnan, A., Mitchell, C.J., Na, C.-H., Kumar, N., Sinnis, P., Sharakhov, I.V., Wang, C., Gowda, H., Tu, Z., Kumar, A., Pandey, A. (2017). Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res. 27, 133–144.

Rasinger, J.D., Carroll, T.S., Lundebye, A.K., Hogstrand, C. (2014). Cross-omics gene and protein expression profiling in juvenile female mice highlights disruption of calcium and zinc signalling in the brain following dietary exposure to CB-153, BDE-47, HBCD or TCDD. Toxicology 321, 1–12. https://doi.org/10.1016/j.tox.2014.03.006

Rasinger, J.D., Carroll, T.S., Maranghi, F., Tassinari, R., Moracci, G., Altieri, I., Mantovani, A., Lundebye, A.-K., Hogstrand, C. (2018). Low dose exposure to HBCD, CB-153 or TCDD induces histopathological and hormonal effects and changes in brain protein and gene expression in juvenile female BALB/c mice. Reprod. Toxicol. 80, 105–116.

Rasinger, J.D., Frenzel, F., Braeuning, A., Bernhard, A., Ørnsrud, R., Merel, S., Berntssen, M.H.G. (2022). Use of (Q)SAR genotoxicity predictions and fuzzy multicriteria decision-making for priority ranking of ethoxyquin transformation products. Environ. Int. 158, 106875.

Rasinger, J.D., Lundebye, A.-K., Penglase, S.J., Ellingsen, S., Amlund, H. (2017). Methylmercury Induced Neurotoxicity and the Influence of Selenium in the Brains of Adult Zebrafish (Danio rerio). Int. J. Mol. Sci. 18, 725.

Rasinger, J.D., Marbaix, H., Dieu, M., Fumière, O., Mauro, S., Palmblad, M., Raes, M., Berntssen, M.H.G. (2016). Species and tissues specific differentiation of processed animal proteins in aquafeeds using proteomics tools. J. Proteomics 147, 125–131.

Regueiro, J., Negreira, N., Hannisdal, R., Berntssen, M.H.G. (2017). Targeted approach for qualitative screening of pesticides in salmon feed by liquid chromatography coupled to traveling-wave ion mobility/quadrupole time-of-flight mass spectrometry. Food Control 78, 116–125.

Ruggles, K.V., Krug, K., Wang, X., Clauser, K.R., Wang, J., Payne, S.H., Fenyö, D., Zhang, B., Mani, D.R. (2017). Methods, Tools and Current Perspectives in Proteogenomics. Mol. Cell. Proteomics 16, 959–981.

Søderstrøm, S., Lie, K.K., Lundebye, A.-K., Søfteland, L. (2022). Beauvericin (BEA) and enniatin B (ENNB)-induced impairment of mitochondria and lysosomes - Potential sources of intracellular reactive iron triggering ferroptosis in Atlantic salmon primary hepatocytes. Food Chem. Toxicol. 161, 112819.

Søfteland, L., Eide, I., Olsvik, P.A. (2009). Factorial design applied for multiple endpoint toxicity evaluation in Atlantic salmon (Salmo salar L.) hepatocytes. Toxicol. In Vitro 23, 1455–1464.

Søfteland, L., Kirwan, J.A., Hori, T.S.F., Størseth, T.R., Sommer, U., Berntssen, M.H.G., Viant, M.R., Rise, M.L., Waagbø, R., Torstensen, B.E., Booman, M., Olsvik, P.A. (2014). Toxicological effect of single contaminants and contaminant mixtures associated with plant ingredients in novel salmon feeds. Food Chem. Toxicol. 73, 157–174.

Sprenger, H., Rasinger, J.D., Hammer, H., Naboulsi, W., Zabinsky, E., Planatscher, H., Schwarz, M., Poetz, O., Braeuning, A. (2022). Proteomic analysis of hepatic effects of phenobarbital in mice with humanized liver. Arch. Toxicol. 96, 2739–2754.

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32, W309-12.

Varunjikar, M.S., Belghit, I., Gjerde, J., Palmblad, M., Oveland, E., Rasinger, J.D. (2022a). Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species. Food Control 137, 108888.

Varunjikar, M.S., Moreno-Ibarguen, C., Andrade-Martinez, J.S., Tung, H.-S., Belghit, I., Palmblad, M., Olsvik, P.A., Reyes, A., Rasinger, J.D., Lie, K.K. (2022b). Comparing novel shotgun DNA sequencing and state-of-the-art proteomics approaches for authentication of fish species in mixed samples. Food Control 131, 108417.

Vildhede, A., Wiśniewski, J.R., Norén, A., Karlgren, M., Artursson, P. (2015). Comparative Proteomic Analysis of Human Liver Tissue and Isolated Hepatocytes with a Focus on Proteins Determining Drug Exposure. J. Proteome Res. 14, 3305–3314.

Waters, M.D., Fostel, J.M. (2004). Toxicogenomics and systems toxicology: aims and prospects. Nat. Rev. Genet. 5, 936–948.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software.

Wisniewski, J.R., Zougman, A., Nagaraj, N., Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods 6, 359.

Woo, S., Cha, S.W., Bonissone, S., Na, S., Tabb, D.L., Pevzner, P.A., Bafna, V. (2015). Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. J. Proteome Res. 14, 3555–3567.

Woo, S., Cha, S.W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., Bafna, V. (2014a). Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. 13, 21–28.

Woo, S., Cha, S.W., Na, S., Guest, C., Liu, T., Smith, R.D., Rodland, K.D., Payne, S., Bafna, V. (2014b). Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics 14, 2719–2730.

Wu, P.-Y., Phan, J.H., Wang, M.D. (2013). Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics 14 Suppl 11, S8.

Zhao, S., Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16, 97.

Chapter 2: ProteoStorm – An Ultrafast Metaproteomics Database Search Tool

2.1: Abstract

Shotgun metaproteomics has potential to reveal the functional landscape of microbial communities, but lacks appropriate methods for complex samples with unknown compositions. In the absence of prior taxonomic information, tandem mass spectra would be searched against large pan-microbial databases, which requires heavy computational workload and reduces sensitivity. We present ProteoStorm, an efficient database search framework for large-scale metaproteomics studies, which identifies high-confidence peptide-spectrum matches (PSMs) while achieving a two to three orders-of-magnitude speedup over popular tools. A reanalysis of a urinary tract infection (UTI) dataset of 110 individuals revealed a complex pattern of polymicrobial expression, including sub-types of urinary tract infections, cases of bacterial vaginosis, and evidence of no underlying disease. Importantly, compared to the initial UTI study that restricted the search database to a manually-curated list of 20 genera, ProteoStorm identified additional genera that were previously unreported, including a case of infection with the rare pathogen *Propionimicrobium*.

2.2: Main Text

Metaproteomics is an emerging field that identifies expressed proteins using tandem mass spectrometry (MS/MS) to decipher the taxonomic and functional realm of complex microbial environments. Without prior knowledge of the active organisms in a sample, downstream analyses are highly dependent on the accurate assignment of MS/MS spectra to peptide sequences from massive (~10Gb+) pan-microbial protein databases. While iterative

search methods utilizing conventional database search tools (Jagtap, et al., 2013; Tang, et al., 2016; Zhang, et al., 2016) increase identifications by reducing the initial search space, and methods such as ComPIL (Chatterjee, et al., 2016) address heavy computational workload by distributing data across multiple servers, searching large databases in reasonable time frames, without the requirement of high computational capacity, remains a major impediment. We present ProteoStorm, an ultrafast metaproteomics database search framework that utilizes a multi-staged, efficient filtering of peptide-spectrum matches (PSMs). On large pan-microbial databases and spectral datasets, ProteoStorm achieved a speedup by two to three orders-of-magnitude over popular database search tools while maintaining high sensitivity in terms of peptides identified.

There are many reasons why conventional database search tools do not scale well in metaproteomics studies. For a large database, D, a common practice is to split D into k arbitrary small files d_i, where each of m spectral files is searched against each d_i. While memory efficient, this practice results in the inefficient and redundant search of spectra against peptides that would never result in a viable match. ProteoStorm assumes that for each detectable protein in a microbial sample, there exists at least one fully-tryptic peptide with no variable modification, identifiable with a sufficiently low p-value PSM. This motivates a multi-stage strategy (**Fig. 2.1a**), where a fast, fully-tryptic search in the first stage is followed by a semi-tryptic search in the second stage that is limited to proteins identified in the first stage. Each stage of ProteoStorm is composed of three core modules: i) database and spectra partitioning, ii) efficient and sensitive peptide filtering, and iii) PSM p-value computation (Methods).

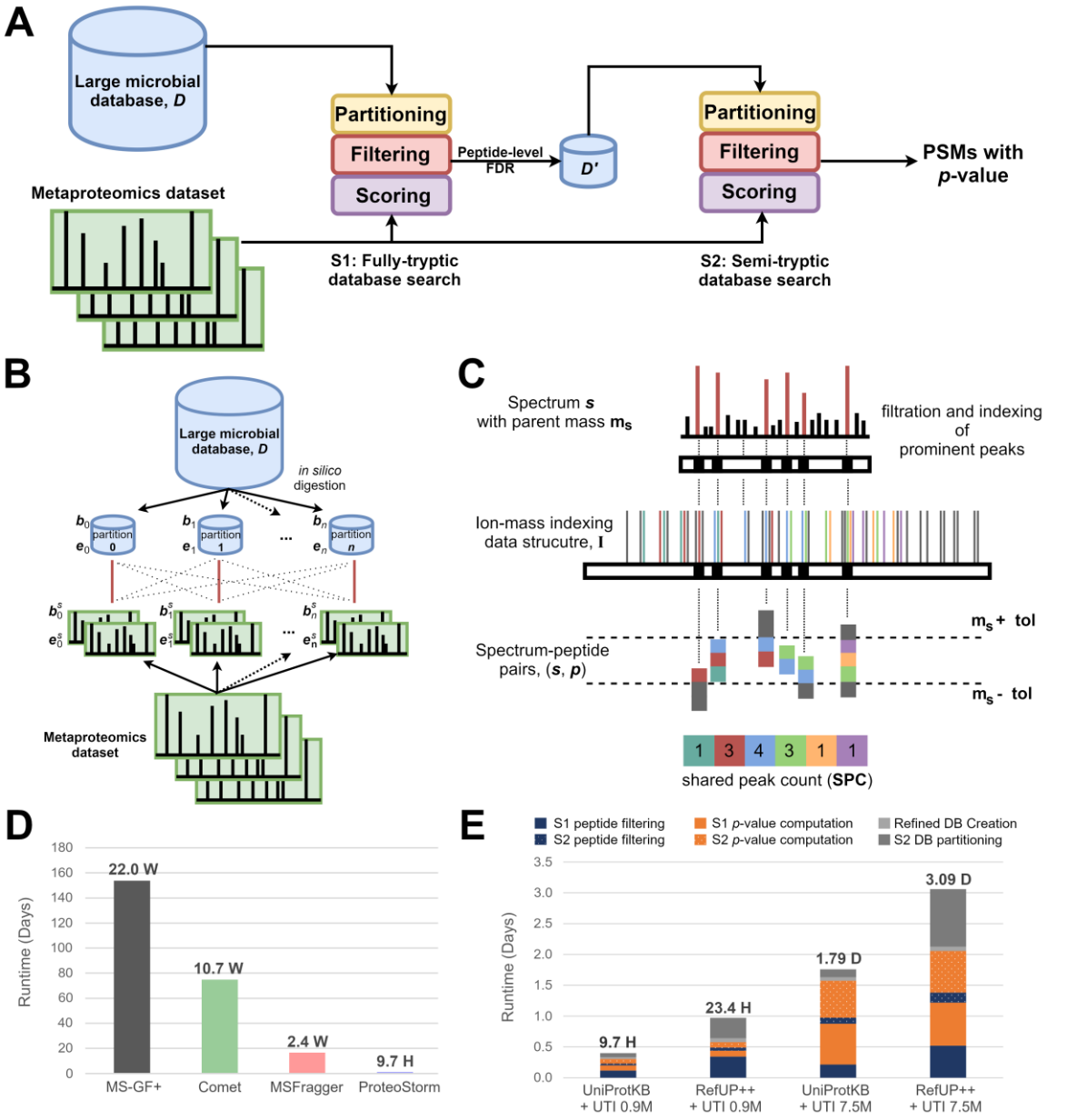


Figure 2.1: ProteoStorm search framework: performance and scalability. (a) Each stage of ProteoStorm is composed of three modules: i) data partitioning, ii) peptide filtering, and iii) p-value computation. Identifications from a fully-tryptic search are used to construct a refined protein database for a semi-tryptic search. PSMs with p-values are reported. **(b)** In the first module, database and spectra partitioning dramatically reduces search space from all peptides (black dotted lines) to peptides within spectra parent mass ranges (red vertical lines). **(c)** In the second module, spectrum-peptide pairs are filtered based on shared counts of prominent spectral peaks (red) to b-/y-ions of peptides (numbers within blocks) using an ion-mass indexing data structure. Each colored block represents a unique peptide within the parent mass tolerance of a given spectrum. **(d)** 946,845 spectra were searched against the UniProtKB database using ProteoStorm, MSGF+, Comet, and MSFragger. ProteoStorm required 9.7 CPU-hours, while other tools required CPU-weeks to complete. **(e)** Breakdown of ProteoStorm runtime by module. S1 and S2 represent the two different stages of ProteoStorm. See also Figure S2.3.

In the first module, ProteoStorm bins the unique set of *in silico* digested tryptic peptides from a pan-microbial database, *D*, into *n* database partitions with pre-defined mass windows in Daltons (**Fig. 2.1b**; Methods). Separately, spectra are organized into *n* matching partitions, based on their parent masses. Each spectral partition is searched only against its matching database partition (instead of *k* database files mentioned above), reducing the number of computations dramatically. Moreover, the partitioning of the entire pan-microbial database is a one-time operation, and the partitions can be used for any metaproteomics experiment.

To overcome the inefficiency of scoring spectra against peptides with poor match of b-/yions, the second module of ProteoStorm filters spectrum-peptide pairs with insufficient number of shared peaks using an ion-mass indexing data structure, *I* (**Fig. 2.1c**; Methods). Similar to previous studies (Rinner, et al., 2008; Stein, 1995; Burke, et al., 2017; Kong, et al., 2017), ProteoStorm utilizes the fact that theoretical ions from different peptides may share the same m/z value given a charge and fragment mass tolerance. For each database partition, ProteoStorm constructs *I*, where ion-mass index i references a mass-sorted list of peptides containing a b-/y-ion within the mass tolerance of *i*. For each prominent peak in a spectrum *s*, its corresponding ion-mass index $i \in I$ is accessed to retrieve the collection of peptides, *P*, containing a matching ion. For all peptides $p \in P$ that are within the parent mass tolerance of the spectrum, the shared peak count (SPC) of the spectrum-peptide pair (*s,p*) is incremented by one. Spectrum-peptide pairs with a sufficient SPC are retained for further analysis in the third module. This ion-mass indexing-based peptide filtering enables ultra-fast querying of a spectrum against a peptide database as it bypasses all peptide ions with no

matching spectrum peak, and matches prominent spectral peaks against all peptide ions simultaneously.

In the third module, ProteoStorm utilizes the MS-GF+ generating function (Kim & Pevzner, 2014) to compute *p*-values for (*s,p*) pairs with the highest MS-GF+ raw score for a given spectrum *s*. Accurate estimation of *p*-values allows for the subsequent computation of a peptide-level FDR. Peptides identified in the first stage are used to construct a refined protein database for a semi-tryptic second-stage search (Methods).

We evaluated the performance of ProteoStorm using 7.5 million LC-MS/MS spectra obtained from urinary pellets (UP) of 110 suspected urinary tract infection (UTI) cases and five healthy individuals from two related studies (Yu, et al., 2015; Yu, et al., 2017). When searching the full UTI data set against a database of 18.8 million microbial sequences from UniProtKB (Release 2017_07), ProteoStorm completed in 1.79 CPUdays – a 485x speedup compared to the estimated 123.8 CPU-weeks for MS-GF+. Given these encouraging speedups and noting the extensive time requirements of other tools, we used a subset of 17 individuals (0.9 million spectra) to perform a systematic comparison of identified peptides and runtimes against MS-GF+, Comet (Eng, et al., 2013), and MSFragger (Kong, et al., 2017).

At 1% peptide-level FDR, MS-GF+ identified 12,139 peptides in an estimated CPU-22 weeks, Comet identified 9,341 peptides in an estimated CPU-10.7 weeks, and MSFragger identified 11,530 peptides in an estimated CPU-2.4 weeks. In contrast, ProteoStorm identified 13,550 peptides in 9.7 CPU-hours, a speedup by 382x, 186x, and 41x, respectively (**Fig. 2.1d**). ProteoStorm identified 96% (11,657) of the MS-GF+ peptides, as also 95.9% (10,331) of the peptides identified by at least two of the three tools (**Fig. S2.3b**). On the other

hand, Comet identified 64.3% and 78.1% of peptides respectively, and MSFragger identified 75.9% and 93.5% of peptides respectively.

All tools had comparable identification rates, with ProteoStorm slightly higher at 5.98% in contrast to the range of 5.02% to 5.42% for other tools. The low identification rates were likely due to missing sequences in the initial pan-microbial database. To test this, we also searched a small wastewater metaproteomics dataset (150,216 spectra), where a specialized database (2.47Gb) had been constructed using the Graph2Pep/Graph2Prot approach (Tang, et al., 2016) on metagenomics data. ProteoStorm identified 10,633 peptides, or 95.9% (4,790) of the MS-GF+ peptides, in 52.6 minutes, while the Graph2Pep/Graph2Prot approach identified 8,740 peptides, or 83.6% (4,175) of the MSGF+ peptides, and took over 3.8 CPU hours (**Fig. S2.1b**). Graph2Pep/Graph2Prot reduced the initial search space by 124x, while ProteoStorm by 184x (**Fig. S2.1a**). Compared to searching against the UniProtKB database, searching against the metagenome derived database improved identification rate from 7.2% to 15.5%.

The speedup provided by ProteoStorm is further amplified by searching the full UTI data set against a comprehensive database (RefUP++; Methods) containing 95 million sequences. When increasing both the number of spectra from 0.9M to 7.5M and the database size from UniProtKB (7.8 Gb) to RefUP++ (34.1Gb), ProteoStorm speedup over MS-GF+ increased from an estimated 382x to 953x (**Fig. S2.3c**). The reduction of the initial search space ranged from 34x to 76x depending on the dataset and database searched (**Fig. S2.3a**). Breaking-down ProteoStorm runtime by module shows that the increase is mostly due to a prolonged computation of *p*-values as the number of spectra increases (**Fig. 2.1e**). While

there is also an increase in runtime when solely increasing the initial database size, this is mainly due to the stage-two database partitioning.

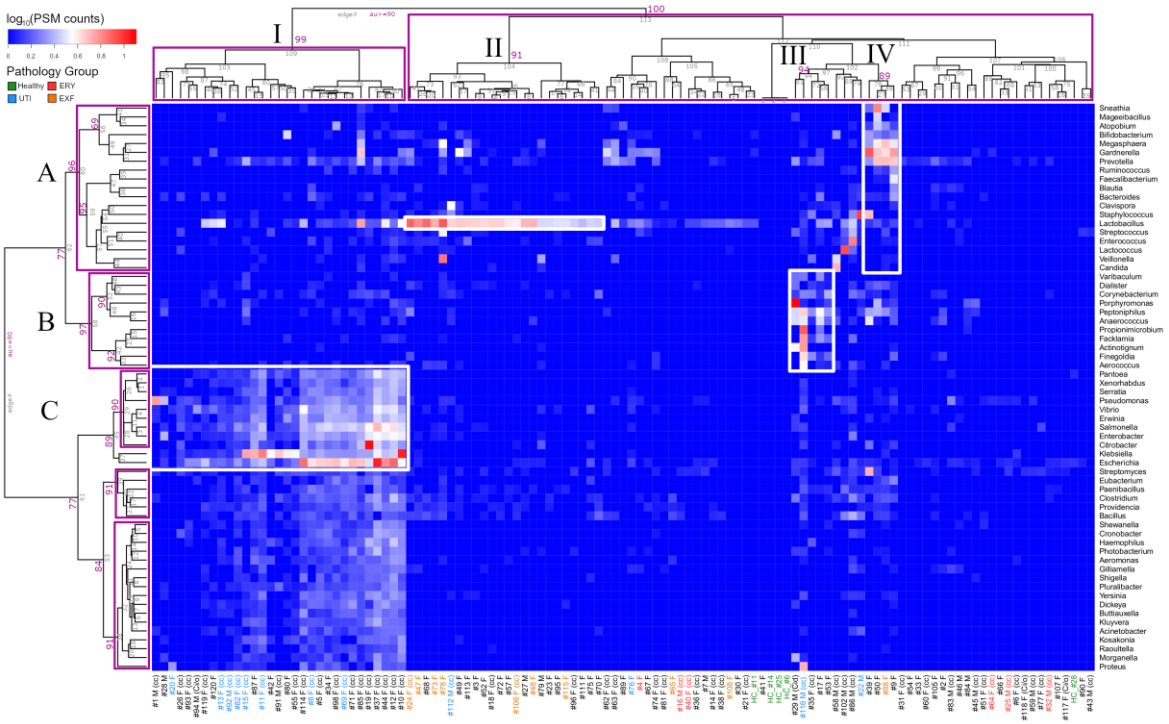


Figure 2.2: ProteoStorm identifies bi-clusters of individuals with similar microbial compositions. Searching the full UTI dataset against the RefUP++ database (2,259 genera) using a genera-restriction approach, ProteoStorm identified 64 genera. Out of 73,092 peptides, 28.5% (20,833) mapped uniquely to a single genus. Four bi-clusters (white boxes) were inferred from clusters with an approximately unbiased (au) p-value greater than 0.90 (magenta boxes), indicating a complex pattern of polymicrobial expression, including sub-types of urinary tract infections, cases of bacterial vaginosis, and evidence of no underlying disease. Pathology groups: Healthy, ERY (erythrocyte/vascular injury), EXF (exfoliation of squamous epithelial and urothelial cells), and UTI (urinary tract infection). See also Table S2.1 and Table S2.2.

The initial UTI study (Yu, et al., 2015) restricted the search database to a manually curated list of 20 genera, identifying 15 as being expressed in 110 individuals. Using an unbiased search of the full UTI dataset against the RefUP++ database (2,259 genera) and a genera-restriction approach (Methods), we identified 64 genera, including the previous 15

(**Table S2.1**). Out of 73,092 peptides, 28.5% (20,833) mapped uniquely to a single genus. Moreover, in 98 out of the 110 individuals, we found that every individual had either the same or a superset of the genera previously reported. Overall, we observed a complex pattern of polymicrobial expressions (**Fig. 2.2**). Using clusters supported by multiscale bootstrap resampling (**Table S2.2**), we could observe at least four bi-clusters. The most common uropathogen in UTI is *Escherichia coli*, followed by other gramnegative microbes, including *Klebsiella*, *Citrobacter*, *Pseudomonas*, and *Enterobacter* (Ronald, 2002). Individuals infected by these dominant uropathogens are present in clusters C-I (genera: 89au; individuals: 99au), including eight of the twelve individuals who were also clinically diagnosed with UTI (Yu, et al., 2017). Six of the eight individuals were aged 61 to 84, and exhibited polymicrobial UTI (p-UTI) with either *Escherichia* or *Klebsiella* dominant over the other (**Table S2.1**). Indeed, p-UTI is common among the elderly, and *Escherichia coli* and *Klebsiella* frequently co-occur in p-UTI (Laudisio, et al., 2015).

Individuals infected by obligate or facultative anaerobic pathogens that are frequently undetected using typical urine culture diagnostic methods are present in cluster B-III (genera: 97au; individuals: 94au) (Imirzalioglu, et al., 2008). Of note, individual #116 was clinically diagnosed with UTI, and while a previous study reported *Proteus mirabilis* as the major UTI causing pathogen, ProteoStorm identified the rare UTI pathogen *Propionimicrobium* (1,129 out of 3,531 PSMs) (Ikeda, et al., 2017) as the most abundant genera followed by *Facklamia* (544 PSMs), *Actinotignum* (538 PSMs), *Proteus* (424 PSMs), and other gram-positive bacteria (**Table S2.1**).

As the female urine specimens were clean catch samples, we cannot distinguish between true urethral/bladder colonization and possible contamination by microbes that are

part of the vaginal and vulvar microbiota. In cluster A-II (individuals: 91au), which is dominated by *Lactobacillus*, we observed seven of twelve individuals who previously expressed high abundances of human epithelia cell proteins (Yu, et al., 2017). The dominance of Lactobacilli together with evidence of shed squamous epithelial cells could suggest a lack of an underlying pathology. In contrast, individuals in cluster A-IV (genera: 96a; individuals: 89au) revealed higher abundances of microbes known to contribute to bacterial vaginosis (BV), notably *Gardnerella*, *Prevotella*, *Atopobium*, *Sneathia* (Ma, et al., 2012; Onderdonka, et al., 2016). This data is consistent with BV as an underlying pathology related to these specimens.

As a UTI dataset is considered less complex, we searched a human infant gut metaproteomics dataset (2.4M spectra) against the RefUP++ database using the generarestriction approach (Methods). ProteoStorm identified 19 genera (**Table S2.3**) from 31,606 peptides mapping uniquely to a single genus (54.9% of all identifications), and detected similar shifts in microbial abundance across day of life 21, 34, and 50 when compared to the previous study (Xiong, et al., 2017) (**Fig. S2.2a**). Searching against the matched metagenome database, ProteoStorm identified 61,364 peptides, or 94.5% (53,847) of the MS-GF+ peptides (**Fig. S2.2b**) at an identification rate of 24.2%.

Metaproteomic studies based on database search approaches depend on the effective and efficient analysis of data. In clinical cases, the accurate identification of pathogens affects treatment options, and failure to detect certain microbes due to non-inclusiveness in the search database may lead to suboptimal treatment or misuse of antibiotics. The use of reference protein repositories is advantageous, but as they grow with increasing sequencing data, so does the need for efficient computational tools that analyze complex, multi-species

communities. ProteoStorm is an ultrafast and highly sensitive tool for the metaproteomics community, and is available through GitHub (https://github.com/proteostorm/ProteoStorm).

2.3: Methods

Data sources

Pan-microbial databases. The UniProtKB database (7.8Gb) in this study consists of 18,755,274 protein sequences from reference proteomes of 6,521 bacterial species (UniProt Release 2017.07). A comprehensive database (RefUP++; 34.1Gb; 2,259 genera) was constructed using 94,916,719 bacterial and fungal protein sequences from RefSeq (Release 85; ftp://ftp.ncbi.nlm.nih.gov/refseq/release/), and UniProt Pan Proteomes (Release 2017.12) and Reference Proteomes (Release 2017.07) (ftp://ftp.uniprot.org/pub/databases/uniprot/). We provide the RefUP++ database as a set of 58 files, accessible via Github. The fasta header information has been processed to include the name of the protein entry followed by the genus name of the organism and the taxonomy ID (see Genera-restriction approach).

Urinary Tract Infection (UTI) MS/MS dataset. To evaluate the performance of ProteoStorm, we used a dataset previously published in two related studies (Yu, et al., 2015; Yu, et al., 2017). LC-MS/MS Raw files of urine pellet (UP) samples from 110 suspected urinary tract infection (UTI) cases and five healthy individuals were provided by researchers at The J. Craig Venter Institute (JCVI) or downloaded from the PRIDE Database (PXD004713). As described previously (Yu, et al., 2017), the samples were collected for diagnostic purposes, considered medical waste, and de-identified prior to transfer to JCVI.

Samples were run in replicate on an Ultimate 3000-nano LC and Q Exactive mass spectrometer system coupled via a FLEX nano-electrospray ion source (Thermo Scientific). RAW files were subjected to peak-picking (Vendor) and converted to Mascot generic format (MGF) format using MSConvert (ProteoWizard v3.0) (Chambers, et al., 2012).

Given the focus of ProteoStorm on metaproteomic data analysis, we performed an initial search of the full UTI dataset against the Human reference proteome (UP000005640; Release 2017.07) using MS-GF+ (variable modifications: protein N-terminal acetylation, and methionine oxidation; static modifications: cysteine carboxyamidomethylation), and excluded the 3,913,142 spectra that were identified as high-confidence human-matching spectra (1% PSM-level FDR per MS/MS experiment) from the spectra partitioning step, resulting in 7,552,992 spectra for the full UTI dataset.

Human Infant Gut MS/MS dataset. To evaluate the performance of ProteoStorm on a more complex dataset, we searched a subset of a human infant gut metaproteomics dataset (MassIVE ID: MSV000080565). Spectra representing stool samples collected on day of life 21, 34, and 50 from infant 23 (healthy preterm infant with mild lung disease) were chosen based on the higher complexity of the samples as determined in the previous study (Xiong, et al., 2017). High-confidence human-matching spectra (135,380) were excluded from the spectra partitioning step, resulting in 2,360,544 spectra.

Wastewater MS/MS dataset. To evaluate the performance of ProteoStorm when searching against a matched metagenome database, we searched a dataset (150,216 spectra) representing oleaginous mixed microbial communities sampled from the surface of a

biological wastewater treatment plant on Oct 12, 2011 (SD6; PeptideAtlas ID: PASS00577) (Muller, et al., 2014).

ProteoStorm core module 1: Data partitioning.

ProteoStorm performs a fast, fully-tryptic search in the first stage followed by a semi-tryptic search in the second stage that is limited to proteins identified in the first stage. Each stage consists of three core modules: i) database and spectra partitioning, ii) ion-mass index-based peptide filtering, and iii) PSM *p*-value computation.

In the first module, *in silico* digested peptides and spectra are sorted and binned into *n* matching partitions based on parent mass. The theoretical mass of a peptide, *p*, with cysteine carboxyamidomethylation as a static modification, m_p, is defined as the sum of the monoisotopic masses of residues in *p*, a hydrogen atom, a hydroxyl group, and a carbamidomethyl group for each cysteine in *p*. Given a distribution of theoretical masses for *N* peptides, we define partition *i* by a pair of indices (b_i, e_i), where the number of peptides satisfying $b_i \leq m_p < e_i$ is *N*/*n*. We provide mass distribution files for both the UniProtKB and RefUP++ databases, accessible via GitHub.

For spectra, to allow for at most one isotopic error of δ (i.e., 1.003355), and a precursor ion mass tolerance of ε (e.g., 10ppm), partition *i* is defined slightly differently as

$$b_i^s = b_i\left(1 - \left(\frac{\varepsilon}{10^6}\right)\right)$$

$$e_i^s = e_i\left(1 + \left(\frac{\varepsilon}{10^6}\right)\right) + \delta$$

Spectra partitioning. Given the interval pairs (b_i^s, e_i^s) for all i, we assign spectra to n bins based on parent mass. The mass of a spectrum, s, is defined as $m_s = M \cdot z - p^+ \cdot z$, where M is the precursor ion mass (in units of m/z), z is the charge state of the precursor ion, and p^+ is the mass of a proton. We assign spectrum s to each bin i if $b_i^s \leq m_s \leq e_i^s$.

To bin spectra efficiently, we sort them by their mass and use a merge operation to assign them to the appropriate bins. Spectra with less than 10 peaks are not included in the spectral partitioning process. As the distribution of spectra based on parent mass does not necessarily follow that of theoretical peptides, a maximum spectral partition size is enforced, creating multiple files for spectra belonging to bin i.

Database partitioning. ProteoStorm takes as input a set of database files, D, and assigns an index j to each file, where $0 \leq j < |D|$. Given D, files are read in batches of 1Gb, and protein sequences are stored in memory as a single string, T, where the character "$" indicates the start of a protein (an additional "$" character is appended at the end of the string to denote the end of the string). Each string T is subjected to *in silico* tryptic digestion, where peptides are limited to the standard 20 amino acids, isoleucine transformed (i.e., leucine replaced with isoleucine), and allowed at most one missed cleavage site. If a peptide at the protein N-terminus, cleavage of its n-terminal methionine is allowed.

Each peptide, p, is stored in a string representation of its corresponding database partition i, and written to file when a buffer size of 600MB is reached. To map peptides back to proteins after stage one, we store additional information along with the sequence for each peptide p, including the index, j, of the originating database file, the protein index h within database file d_j, and the pre- and post-amino acids. For a target-decoy approach (Elias &

Gygi, 2007), protein sequences are reversed and concatenated in a similar manner to form string *F*. Origin from a decoy protein is indicated by prefix 'd_' for protein index *h*. For a semi-tryptic *in silico* digest, either the n-terminus or c-terminus of a fully-tryptic peptide is anchored and amino acids are removed from the opposite terminus to form semi-tryptic peptides.

As the same peptide, *p*, may result from the *in silico* digestion of one or more strings *T* (or *F*), ProteoStorm combines multiple mapping information from duplicate peptide entries in each partition *i*, and writes the mass of an unique peptide *p*, the peptide sequence, and the combined mapping information of *p* to create the final database partition. Additionally, ProteoStorm assigns a category to peptide *p* in the following order of preference: fully-tryptic/target, fully-tryptic/decoy, semi-tryptic/target, or semi-tryptic/decoy. The reason for this is twofold. First, a given spectrum *s* is always assigned the peptide sequence from its best scoring (*s*,*p*) pair, and given that fully-tryptic peptides score higher than semi-tryptic peptides during *p*-value computation, fully-tryptic peptides are preferred over semi-tryptic peptides. Second, in the case that a (*s*,*p*) pair with a peptide from a decoy sequence scores equally as well as one with a peptide from a target sequence, the latter is always preferred.

As an example, the following peptides have a mass within [b_0, e_0), and are written to database partition 0 in a tab separated format. The peptide sequence "AGGGGGGG" can be mapped to two proteins, specifically, the first occurrence is in the 2,936th entry of database file d_{4076}, or UP000061569_69.fasta, and the 1,632nd entry in database file d_{2526}, or UP000030518_1300345.fasta. The pre- and post-amino acid "R-" indicates that the peptide is a fully-tryptic protein c-term peptide.

Mass (Da)	Peptide Sequence	Peptide Index	Mapping Information
488.1979	R.AGGGGGGG.-	0	4076|2936|R-;2526|1632|R-
502.2136	R.GAGGAGGG.-	XXX_1	4640|d_2272|R-
502.2136	R.GGGGGGAA.-	2	6530|897|R-

In the second stage of ProteoStorm, the refined protein database (see Refined database creation) is provided as two files, a target sequence database and a decoy sequence database. A semi-tryptic *in silico* digest is performed, and database partitions are created as described previously, with the exception that sequences from the target database are not reversed to form string *F*. Instead, sequences from the decoy database are concatenated to form string *F*.

ProteoStorm core module 2: Peptide filtering.

The second module of ProteoStorm computes the shared peak counts (SPC) between each spectrum and candidate peptides within its parent mass tolerance using an efficient ion-mass indexed data structure. For each database partition, peptide ions (b-/y-ions) with a maximum fragment ion charge of 2+ are used to construct an ion-mass indexing data structure, *I*, where each ion-mass index, $i \in I$, references a mass-sorted list of peptides containing a b-/y-ion within the mass tolerance of *i* (**Fig. 2.1c**). For each spectrum, *s*, peaks with an intensity less than $1.0 \cdot Q$, where Q is the mean of the intensities of peaks at the bottom 25%, are considered noise and removed. Prominent peaks are defined as the top seven intense peaks in a window of 75 Daltons. For each prominent peak in a spectrum *s*, its corresponding ion-mass index i is accessed to retrieve a mass-sorted list of peptides, *P*, containing a matching ion. For all candidate peptides $p \in P$ that are within the parent mass

tolerance of the spectrum and at most one isotopic error away from the spectrum parent mass, the shared peak count (SPC) of the spectrum-peptide pair (*s*,*p*) is incremented by one.

To increase the efficiency of database search (filtering as many peptides as possible) without sacrificing sensitivity (retaining true positives), for each spectrum *s*, any candidate peptide with a SPC less than max(M_{min}, $M_{max} - 1$) is filtered, where

$$M_{max} = \max_{p \in P}\left(\mathrm{SPC}(s,p)\right).$$

We chose $M_{min} = 7$ for fully-tryptic peptides and $M_{min} = 7$ for semi-tryptic peptides, which works well for both the HCD fragmentation-based UTI dataset, the CID fragmentation-based human infant gut dataset, and the HCD fragmentation-based wastewater dataset. Determining M_{min} based on the length and mass of a peptide, the instrument type, and fragmentation method is optimal but requires further investigation. Retained spectrum-peptide pairs are subjected to further analysis in core module 3.

ProteoStorm core module 3: *p*-value computation.

To accurately estimate the *p*-value of spectrum-peptide pairs and allow for computation of a false discovery rate, we created a modified version of MS-GF+ (v2018.04.09) (Kim & Pevzner, 2014) that directly calls its raw score computation and generating function approach. To identify high-confidence PSMs using the target-decoy approach, the peptide that achieves the highest MS-GF+ raw score for each spectrum *s* among the retained (*s*,*p*) pairs from core module 2 is selected for *p*-value computation using the generating function approach. Briefly, given an SPC score of *k* for a spectrum-peptide pair (*s*,*p*), the generating function computes the probability of a randomly generated peptide

(all residues independently and identically distributed) obtaining an SPC score of k or higher against spectrum s.

Refined database creation.

Define a peptide group as an unmodified peptide sequence and the sum of its modifications (e.g., 'PEPTIDES+57.021464'). The p-value of a peptide group is equal to that of its best scoring (s, p) pair. For any p-value threshold, θ, let $f(\theta)$ denote the number of decoy groups whose p-value is $\leq \theta$. Similarly, let $t(\theta)$ denote the number of target groups whose p-value is $\leq \theta$. A peptide-level false discovery rate (FDR) is thus defined as $\mathrm{FDR}^{(\theta)} = \frac{f(\theta)}{t(\theta)}$.

To determine high-confidence peptides, L, in the first stage of ProteoStorm, a script identifies peptide groups across all MS/MS experiments, and determines the p-value threshold for a peptide-level FDR of 5%. Peptide groups that pass the threshold are retained as high-confidence peptides $p \in L$. Subsequently, all proteins containing a peptide $p \in L$ are included in a refined protein database, D'. Using the protein mapping information retained during database partitioning, a script iterates through target peptides in L, writing both the protein sequence and its reverse sequence to D'. Redundant protein sequences, including those that only differ by protein name, are not allowed. A similar procedure is applied to high-confidence decoy peptides, with the exception that only decoy protein sequences are written to D'. Proposed by Bern and Kil (Bern & Kil, 2011), this method of generating a decoy database for a second stage search includes a conservative bias by including more decoy sequences than target. The more sophisticated version of this method, where the reversed sequences of target proteins are added to the decoy protein sequences until the total

number of decoy sequences is equal to the number of target sequences was validated by Jeong et al. (Jeong, et al., 2012). The refined protein database is provided in the S1_OutputFiles directory as a combined target decoy database "RefinedProteinDB.fasta."

ProteoStorm Output.

After completion of stage two, the identified PSMs (no FDR applied) are reported in the S2_OutputFiles directory as file "ProteoStorm_output.txt." For all analyses in this manuscript, we applied a 1% peptide-level FDR (pooled across all MS/MS experiments) for peptide identifications, and a 1% PSM-level FDR (per MS/MS experiment) for PSM identifications, where PSMs representing high-confidence peptides are reported.

ProteoStorm is designed to be flexible with regards to user needs. As an example, we used peptides from stage two of ProteoStorm to analyze metaproteomic samples at the genera level. Details on using ProteoStorm for other taxonomies and post-translational modifications are accessible via GitHub (Alternative configurations for ProteoStorm).

Genera-restriction approach.

To identify microbial communities and infection patterns in the UTI dataset, we implemented a genera-restriction approach for ProteoStorm, where the pan-microbial database, D, includes only proteomes with available taxonomic information at the genus level, and the refined protein database, D', is based on high-confidence peptides identified in stage-one that uniquely map to a single genus among the restricted genera (see *Refined database creation* below). The three core modules of ProteoStorm are as described previously.

Database preprocessing. To determine the genus to which a proteome or protein entry belongs, a script parses NCBI taxonomy .xml files downloaded using the NCBI E-utilities API, and rewrites the fasta header information to include the name of the protein entry followed by the genus name of the organism and the taxonomy ID. For UniProt proteomes, file names already include a proteome identifier (UPID) that can be queried using the UniProt API to obtain its corresponding taxonomy ID. For RefSeq entries, taxonomy IDs are found in the RefSeq-release#.catalog file. Given that the UTI dataset represents human microbiome samples, we limited proteins to sequences belonging to genera from bacteria and fungi.

Refined database creation. To construct a genera-restricted protein database, confidence peptides (1% peptide-level FDR), *C*, identified in stage-one of ProteoStorm are used to infer a set of genera, G_c, with at least one peptide $p \in C$ uniquely mapping to each genus *g*. Specifically, each genus, *g*, has a relative abundance score,

$$A_g = \frac{S_g}{\sum_i^G S_g},$$

where S_g is the number of peptides in *C* that map uniquely to proteins belonging to genus *g*, and *G* is the set of all available genera. Target peptides in *C* contribute to the score of target genera, while decoy peptides in *C* contribute to the score of decoy genera. Genera with a higher A_g score than the most abundant decoy genera are included in the set of high-confidence genera, G_c. As abundance thresholds are largely dependent on the complexity of a microbial sample, we assumed that genera with a higher abundance than the most-abundant decoy genera are more likely to be present in the samples.

Given the set of high-confidence genera, G_c, and the list of high-confidence peptides (1% peptide-level FDR), C, a refined protein database is constructed in a similar manner to the procedure described previously. The additional criterion is that all proteins mapping to a peptide $p \in C$ must also belong to a genus $g \in G_c$.

To construct a refined protein database based on a different level of taxonomy, users can use the taxonomy ID provided for each protein entry in the RefUP++ database to query the NCBI taxonomy database (NCBI E-utilities API or taxonomy .xml files accessible via GitHub). A list of high-confidence taxa would be based on peptides mapping uniquely to that taxa (see GitHub: Alternative configurations for ProteoStorm).

Comparisons against conventional tools.

For timing and identification comparisons of ProteoStorm against conventional database search tools, we required all processes to use a single core and less than 8Gb of RAM on an Intel® Core™ i7-6700K CPU (4 cores, 8 threads) with an installation of Windows 10 Pro. The UniProtKB database (7.8Gb) and 12+5 UTI dataset were used in the comparison. Given the engineering differences of MS-GF+ (v.20161026) (Kim & Pevzner, 2014), Comet (2016.01 rev. 0) (Eng, et al., 2013), and MSFragger (20170103_v2) (Kong, et al., 2017), different approaches were implemented to satisfy the criteria mentioned above. Additionally, given the extensive time requirements of these tools, we estimated runtimes and obtained results through parallel runs on a Linux server or the local machine. In all searches, cysteine carboxyamidomethylation was included as a static modification. A 1% peptide-level FDR, as described previously, was applied using SpecEvalue for MS-GF+, e-value for Comet, and probabilities from running PeptideProphet

(parameters: --clevel 0 --combine --decoyprobs --expectscore --ppm -accmass --nonparam) for MSFragger.

MS-GF+. Spectra were arbitrarily split into six files of size 800Mb and *D* was arbitrarily split into *k* database files of size 60Mb (132 database files for the UniProtKB database; 573 database files for the RefUP++ database). Spectra file #3 contained the closest number of spectra to the average number of spectra across all six files (157,945 spectra), and was searched against six database files with the highest number of fully-tryptic peptides. Runtime was estimated as the average runtime of the six database searches multiplied by the total number of spectra files and database files. Search parameters: -t 10ppm -tda 1 -m 3 -inst 3 -e 1 -ntt 1 -minLength 8 -maxLength 40 -thread 1. PSMs with peptides having greater than one miscleavage were removed after computing a 1% peptide-level and 1% PSM-level FDR.

Comet. A spectrum_batch_size of 25,000 was specified in the parameter file, and a combined spectra file was searched against the same six database files used to estimate runtime for MS-GF+. Runtime was estimated as the average of the six database searches, multiplied by the total number of database files. Default search parameters (comet.params.high-high) were used with the following exceptions:

decoy_search = 1

num_threads = 1

peptide_mass_tolerance = 10.00

num_enzyme_termini = 1

allowed_missed_cleavage = 1 no

variable mods, activation_method =

HCD digest_mass_range = 488.0

5700.0 clip_nterm_methionine = 1

spectrum_batch_size = 25000

decoy_prefix = XXX_

MSFragger. Spectra were arbitrarily split into five files of size 1000Mb and *D* was arbitrarily split into 1,583 target-decoy concatenated database files of size 10Mb. Spectra file #1 contained the closest number of spectra to the average number of spectra across all five files (191,219 spectra). Runtime was estimated as the average of the database searches, multiplied by the total number of spectra files and database files. Default search parameters were used with the following exceptions:

num_threads = 1

precursor_mass_tolerance = 10.00

precursor_mass_units = 1

precursor_true_tolerance = 10.00

isotope_error = 0/1

num_enzyme_termini = 1

no variable modifications digest_min_length = 8

digest_max_length = 40

digest_mass_range = 488.0 5700.0

max_fragment_charge = 3

minimum_peaks = 10

min_matched_fragments = 4

#recommended for narrow window search minimum_ratio = 0.00

Comparison against Graph2Pep/Graph2Prot approach.

To provide a fair comparison of ProteoStorm against MS-GF+ and the Graph2Pep/Graph2Prot approach (Tang, et al., 2016), we reconstructed the initial database (MG_SD6.500.faa and MG_SD6.contig-pep.fixedKR.fasta) used for the SD6 metaproteomics dataset by following the README file provided in the Graph2Pro GitHub repository (https://github.com/COL-IU/Graph2Pro). The reanalysis was limited to the SD6 dataset as the assembled contig file (SRR1544596-SD6MG-s2-63mer-k31d1.contig) and edge file from graph (SRR1544596-SD6MG-s2-63mer-k31d1.updated.edge) were only available for SD6 (http://darwin.informatics.indiana.edu/Dbgraph/example/). The following parameters were used for all searches utilizing MS-GF+: -t 15ppm -tda 0 -m 3 -inst 1 -e 1 -ntt 1 minLength 8 -maxLength 40, with the exception of step 9 (second stage search) for the Graph2Pep/Graph2Prot approach, where -tda 1 was used. ProteoStorm parameters: -PrecursorMassTolerance 15 --FragmentMassTolerance 0.015 --InstrumentID 1 -FragmentMethodID 3.

For the Graph2Pep/Graph2Prot approach, we followed all steps until the eighth step, where an error was encountered in the Graph2Pro software. As such, to obtain final results for the Graph2Pep/Graph2Prot approach, we ran MS-GF+ on the provided second stage database (SD6_hybrid_DBGraphPep2Pro_5.fasta) using the parameters above. For the default MS-GF+ run, the SD6.mgf file was searched against a combined target-decoy

database (MG_SD6.contig-pep.fixedKR.fasta, MG_SD6.500.faa, and reversed protein sequences from MG_SD6.500.faa; 2.47 GB in size; 43,308,680 entries). Following the previous study (Tang, et al., 2016), we applied a 1% PSM-level FDR to obtain the final identifications.

Comparison against human infant gut dataset.

To evaluate ProteoStorm on a more complex dataset, we searched the infant gut dataset (infant 23, day of life 21, 34, 50) against a matched metagenome database and compared to MS-GF+ results. As we already searched the dataset against a Human reference proteome (see Data sources) with contaminant sequences to remove high-confidence human-matching spectra, we removed Human and contaminant sequences from the metagenome database provided by the previous study (Xiong, et al., 2017) (Infant23_Human_Ref2011_IgA_contams.fasta), resulting in a microbial sequence database of 84,562 entries (27.5Mb). Parameters for MS-GF+ : -t 10ppm -m 1 -inst 1 -e 1 -ntt 1. Parameters for ProteoStorm: --PrecursorMassTolerance 10 --FragmentMassTolerance 0.6 --InstrumentID 1 --FragmentMethodID 1. A 1% peptidelevel FDR was applied as described previously.

To identify the taxonomic composition of infant 23's gut microbiome, we searched the dataset against the RefUP++ database using the genera-restriction approach in ProteoStorm. A normalized genera matrix was constructed as previously described, with the following exception: Normalizing was based on a set of human proteins, H, that were expressed in all samples collected from infant 23 (i.e., day 21 run 1, day 21 run 2, day 34 run 1, day 34 run 2, day 50 run 1, day 50 run2).

Genera matrix.

After the second stage of ProteoStorm, a genera matrix is constructed, where each row is a genus, *g*, each column is an individual, and values are normalized relative abundances in spectral counts, averaged across technical replicates. While there isn't an optimal way to normalize across experiments, we normalized based on spectral counts for a set of human proteins, H, that were expressed in all healthy cohort samples (representative of a baseline expression level for each MS/MS experiment). To identify genera, we performed exact matching of high-confidence peptides (1% peptide-level FDR), C', to protein sequences in the pan-microbial database, *D*, using the PeptideMatchCMD_1.0.jar tool (Chen, et al., 2013). The normalized relative abundance for a genus g in a MS/MS experiment, *E*, is defined as

$$(G_a)_N = \frac{G_a}{\alpha_E},$$

where G_a is the number of PSMs passing a 1% PSM-level FDR that belong to a peptide $p \in C'$ that maps uniquely to genus g. The normalization factor, α_E, is the sum of PSMs passing a 1% PSM-level and 1% peptide-level FDR that map to a protein in H, divided by the number of proteins in H expressed in experiment *E*. If a peptide matches to more than one protein, the number of PSMs is divided by the number of matched proteins.

Hierarchical clustering.

As suggested by literature (Clarke & Green, 1988), the pseudo count-adjusted ($+1x10^{-30}$) genera matrix was square root transformed prior to computing Bray-Curtis dissimilarity matrices (vegdist function in R package vegan v2.4.6) for both the genera

(rows) and individuals (columns). Unsupervised hierarchical clustering of the dissimilarity matrices was performed separately (R function hclust, method: ward.D2). For better visualization, the square root transformed genera matrix was further $\log_{10}$ transformed prior to heatmap construction (heatmap.2 function in R package gplots v3.0.1).

Clusters were evaluated using an edited version of the R package pvclust (v2.0.0) (Table S2). The source code was downloaded from https://github.com/cran/pvclust, and line 382 in pvclust-internal.R was changed from "dist(t(x),method)" to "vegdist(t(x), method = "bray", binary=FALSE)", where vegdist is the function from R package vegan (v2.4.6) that computes a dissimilarity index (Bray-Curtis) and returns a distance object. The bootstrap sample size, nboot, was set at 100,000. Clusters with a probability value (approximately unbiased (AU) *p*-value) greater than 0.9 were considered strongly supported by multiscale bootstrap resampling.

2.4: Data and Software Availability

ProteoStorm is a collection of scripts written in python (v2.7), that calls an executable compiled from C++ scripts for peptide filtering (core module 2) as well as a java .jar executable from a modified version of MS-GF+ for raw score and *p*-value computations (core module 3). ProteoStorm is available at https://github.com/proteostorm/ProteoStorm. The LC-MS/MS urinary pellet dataset is available on MassIVE (MSV000082031) in RAW file format and peak-picked MGF file format.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Cell Systems 2018. Beyter, Doruk; Lin, Miin S.; Yu, Yanbao; Pieper, Rembert; Bafna, Vineet. (2018). "ProteoStorm: An Ultrafast Metaproteomics Database Search Framework." *Cell Systems* 7(4): 563-467. The dissertation author was the primary investigator and author of this paper.

2.5: References

Bern, M, and Kil Y.J. (2011). Comment on "Unbiased Statistical Analysis for MultiStage Proteomic Search Strategies". J Proteome Res *10*, 2123–2127.

Burke, M.C., Mirokhin, Y.A., Tchekhovskoi, D.V., Markey, S.P., Heidbrink Thompson, J., Larkin, C., and Stein, S.E. (2017). The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. J Proteome Res *16*, 19241935.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol *30*, 918-920.

Chatterjee, S., Stupp, G.S., Park, S.K., Ducom, J.C., Yates, J.R., 3rd, Su, A.I., and Wolan, D.W. (2016). A comprehensive and scalable database search system for metaproteomics. BMC Genomics *17*, 642.

Chen, C., Li, Z., Huang, H., Suzek, B.E., Wu, C.H., and UniProt, C. (2013). A fast Peptide Match service for UniProt Knowledgebase. Bioinformatics *29*, 2808-2809.
Clarke, K.R., and Green, R.H. (1988). Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series *46*, 213-226.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods *4*, 207-214.
Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics *13*, 22-24.

M. Ikeda, T. Kobayashi, T. Suzuki, Y. Wakabayashi, Y. Ohama, S. Maekawa, S. Takahashi, Y. Homma, K. Tatsuno, T. Sato, S. Okugawa, K. Moriya, H. Yotsuyanagi. (2017). Propionimicrobium lymphophilum and Actinotignum schaalii bacteraemia: a case report. New Microbes New Infect *18*, 18-21.

Imirzalioglu, C., Hain, T., Chakraborty, T., and Domann, E. (2008). Hidden pathogens uncovered: metagenomic analysis of urinary tract infections. Andrologia *40*, 66-71.

Jagtap, P., Goslinga, J., Kooren, J.A., McGowan, T., Wroblewski, M.S., Seymour, S.L., and Griffin, T.J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics *13*, 1352-1357.

Jeong, K., Kim, S., and Bandeira, N. (2012). False discovery rates in spectral identification. BMC Bioinformatics. 13 Suppl 16, S2.

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun *5*, 5277.

Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods *14*, 513-520.

Laudisio, A., Marinosci, F., Fontana, D., Gemma, A., Zizzo, A., Coppola, A., Rodano, L., and Antonelli Incalzi, R. (2015). The burden of comorbidity is associated with symptomatic polymicrobial urinary tract infection among institutionalized elderly. Aging Clin Exp Res *27*, 805-812.

Ma, B., Forney, L.J., and Ravel, J. (2012). Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol *66*, 371-389.

Muller, E.E.L., Pinel,N., Laczny, C.C.,Hoopmann, M.R., Narayanasamy, S., Lebrun, L.A., Roume, H., Lin, J., May, P., Hicks, N.D., Heintz-Buschart, A., Wampach, L., Liu, C.M.; Price, L.B., Gillece, J.D., Guignard, C., Schupp, J.M., Vlassis, N., Baliga, N.S., Moritz, R.L., Keim, P.S., and Wilmes, P. (2014) Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. Nat Commun *5*, 5603.

Onderdonk, A.B., Delaney, M.L., and Fichorova, R.N. (2016). The Human Microbiome during Bacterial Vaginosis. Clin Microbiol Rev *29*, 223-238.

Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L.N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008). Identification of cross-linked peptides from large sequence databases. Nat Methods *5*, 315-318.

Ronald, A. (2002). The etiology of urinary tract infection: traditional and emerging pathogens. Am J Med 113 Suppl 1A, 14S-19S.

Stein, S.E. (1995). Chemical substructure identification by mass spectral library searching. J Am Soc Mass Spectrom *6*, 644-655.

Tang H., Li, S., and Ye, Y. (2016). A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. PLoS Comput Biol *12*, e1005224.

Xiong, W., Brown, C.T., Morowitz, M.J., Banfield, J.F., and Hettich, R.L. (2017). Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life. Microbiome. *5*, 72.

Yu, Y., Sikorski, P., Bowman-Gholston, C., Cacciabeve, N., Nelson, K.E., and Pieper, R. (2015). Diagnosing inflammation and infection in the urinary system via proteomics. J Transl Med *13*, 111.

Yu, Y., Sikorski, P., Smith, M., Bowman-Gholston, C., Cacciabeve, N., Nelson, K.E., and Pieper, R. (2017). Comprehensive Metaproteomic Analyses of Urine in the Presence and Absence of Neutrophil-Associated Inflammation in the Urinary Tract. Theranostics *7*, 238-252.

Xu Zhang, Zhibin Ning, Janice Mayne, Jasmine I. Moore, Jennifer Li, James Butcher, Shelley Ann Deeke, Rui Chen, Cheng-Kang Chiang, Ming Wen, David Mack, Alain Stintzi, Daniel Figeys. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome *4*, 31.

Chapter 3: Genes Involved in the Maintenance of Extrachromosomal DNA (ecDNA) via Transcriptomics

3.1: Main Text

While extrachromosomal DNA (ecDNA) has been associated with oncogene amplification and poor outcome across multiple cancers, the processes involved in the maintenance of ecDNA remains elusive. EcDNA are frequently found in a variety of tumors subtypes. A recent analysis of 1,921 tumors from The Cancer Genome Atlas (TCGA) suggested that ecDNA prevalence ranges from 0% to 59.6% across multiple tumor subtypes (Kim et al., 2020). The analysis classified tumor samples into five sub-types: ecDNA(+), BFB, heavily rearranged, linear, and no-amplification. Here, we treated the latter two categories as ecDNA(-). In order to understand the transcriptional programs active in generating and maintaining ecDNA, we selected 870 samples from 14 tumor types with at least three ecDNA(+) samples, and compared the gene expression data of 234 ecDNA(+) and 636 ecDNA(-) samples (**Table S3.1**).

A conventional gene expression analysis may not differentiate between a gene that is highly expressed in only a subset of ecDNA(+) samples versus a gene that has a small but significant higher expression in all ecDNA(+) samples. For example, suppose there were 100 ecDNA(+) and 100 ecDNA(-) samples. Consider gene A, where the normalized expression values in ecDNA(+) samples are 11 for 10 samples and 1 for 90 samples, and the normalized expression values in ecDNA(-) samples are 1 for 100 samples. Also consider gene B, where the normalized expression values in ecDNA(+) samples are 2.5 for 50 samples and 1.5 for 50 samples, and the values in ecDNA(-) samples are 1 for 100 samples. In a conventional gene expression analysis, gene A is perhaps more significantly over-expressed, but it may or may not be the type of gene we are looking for. For example, it could reside

on an amplicon in 10 samples and thereby be over-expressed but may not play a general pan-cancer role in ecDNA maintenance. Instead, we decided to look for a collection of genes that were most predictive of ecDNA status, with the assumption that the genes with the most predictive values are universally critical to ecDNA maintenance.

3.1.1: Machine learning of ecDNA status identifies candidate genes for ecDNA maintenance.

To identify a minimal set of genes whose expression values were predictive of the sample being ecDNA(+), we used an automated feature selection algorithm, Boruta (**Fig. 3.1a**; Kursa and Rudnicki, 2010). Given a gene expression matrix, *M*, of dimension *r* x c, where *r* is the number of samples and *c* is the number of genes, Boruta generates *c* shadow feature vectors by random shuffling of gene vectors in matrix *M*, generating a new matrix *M'* of dimension *r* x 2*c*. A random forest algorithm is then used to quantify the importance of each gene in separating ecDNA(+) from ecDNA(-) samples. In a given Boruta run, for each iteration $i \in n$, where $i = \{1,2,3, \dots, n\}$, there are two possible outcomes for a gene. If the gene scores higher than the best scoring shadow feature, the gene is considered a "hit". If the gene scores lower than the best scoring shadow feature, the gene is considered a "non-hit". The number of hits, k_g, for each gene, *g*, after *i* iterations is stored in a list, L_i. Similar to a coin toss event of *n* trials, the probability that a gene is a "hit" or "non-hit" in each iteration is 0.5, and follows the binomial distribution.

After *i* iterations, the probability that a gene has *k* or more hits, $P(X > x)$, where $x = k - 1$, can be computed using the binomial distribution survival function. This probability is computed for each gene, *g*, with k_g number of hits and stored in list P_i. As Boruta involves testing multiple genes per iteration and testing the same genes repeatedly over *i* iterations,

two multiple hypothesis testing corrections are applied to list P_i. First, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is applied to control the False Discovery Rate (FDR) at 5%, and second, the Bonferroni correction is applied to control for the family-wise error rate at $\alpha = 0.05$. If the probability of a gene satisfies both corrections, the gene is accepted as a Boruta gene. Separately, the probability that a gene, g, has k_g or lower hits, $P(X \leq k_g)$, can be computed using the binomial cumulative distribution function. For genes that are not accepted, if the probability that a gene has k_g or lower hits satisfies both corrections, the gene is rejected as a Boruta gene.

Given the unequal representation of ecDNA(+) and ecDNA(-) samples within each of the 14 tumor types, we performed Boruta on 200 datasets, each consisting of a random selection of 80% of the 870 samples, and chose a cutoff of at least 10 of the 200 runs as the criteria for a gene to be considered an important feature for downstream analysis. The Boruta analysis identified a set of 408 Core genes.

3.1.2: Extending the core set with co-expressed genes

We note that the core set is not a comprehensive list. Suppose gene C has identical expression values to gene B, a Core gene, described above. The Boruta analysis will pick either gene B or gene C, but not both, because adding gene C does not increase the predictive value. However, both genes may play an important role in ecDNA maintenance. To correct this, we ran pvclust (Suzuki, R. et al., 2006) to cluster all gene expression values, and identify stable clusters using multiscale bootstrap resampling (**Fig. 3.1b**; Methods). We used an approximately unbiased (AU) *p*-value threshold of 0.95 to select the most highly co-expressed gene clusters. An AU *p*-value greater than 0.95 represents the rejection of the null

hypothesis that a group of genes fail to form a stable cluster at a significance level of 0.05. Recomputing the number of Boruta runs that members of a cluster were selected in, we selected clusters that appeared in at least 10 of the 200 Boruta runs (Methods). This resulted in the selection of 354 clusters (**Table S3.2**).

Notably, among the 354 clusters, only 2 clusters (with 14 genes) did not contain any Core genes, and most (344/354) contained 1 or 2 Core genes (**Fig. 3.1c**). When selecting clusters that contained at least 1 Core and 1 co-expressed gene, 53/71 clusters contained 1 to 3 Core genes (**Fig. S3.1a**), confirming that a few genes per co-expressed cluster provide sufficient predictive value, but that other co-expressed genes might still play an important role in ecDNA maintenance. This is true for clusters of various sizes, including the 2-member cluster #1940 and the 21-member cluster #1509. In cluster #1940, CSTF1 has similar expression values to the Core gene RAE1 (**Fig. S3.1b; Table S3.3**). While not necessarily increasing the predictive value, CSTF1 is involved in aberrant alternative splicing events along with other splicing factors including RAE1 (Wang et al., 2019), which is a mitotic checkpoint regulator that promotes tumor growth in cancer (Kobayashi, Y. et al., 2021). In cluster #1509, 12 genes were highly co-expressed with 9 Core genes (**Fig. S3.1c; Table S3.3**), and were enriched in cell-cycle related biological processes (Methods).

Summarizing, the 354 clusters contained 643 genes, which included 408 Core genes and 235 additional genes (**Fig. 3.1b**). Together, we define these genes as the CorEx (Core+co-expressed) genes (**Table 3.1**). Importantly, the total number of CorEx genes per cluster was also small (**Fig. 3.1d**), with only 7 clusters carrying more than 10 genes, indicating that most of the CorEx genes have a distinct role in ecDNA(+) samples. We present our investigations on the biological functions of clusters, later in the paper.

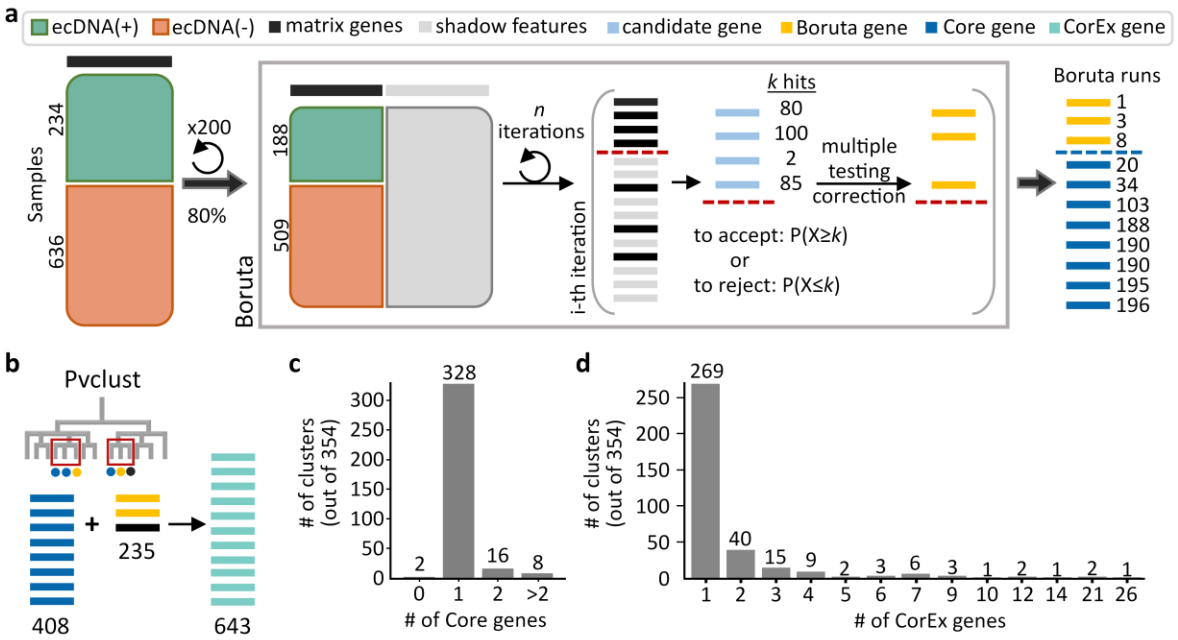


Figure 3.1: Genes predictive of ecDNA status. (a) Using the feature selection algorithm, Boruta, 408 Core genes were selected. **(b)** Identification of highly co-expressed stable gene clusters added an additional 235 genes to the final list of 643 CorEx genes. **(c)** Out of 354 clusters, the majority (344) of clusters contained 1 or 2 Core genes. **(d)** Out of 354 clusters, most clusters are small clusters, with only 7 clusters containing more than 10 members.

3.1.3: CorEx gene expression correlates strongly with ecDNA occurrence

We validated the relevance of CorEx genes in ecDNA maintenance by running cross-validation experiments (**Fig. 3.2a**; Methods) to test the predictive power of CorEx gene expression in determining the ecDNA status of the sample. The expression of the core-set predicted the ecDNA status of samples with precision ranging from 0.452 to 0.903, and recall from 0.304 to 0.717. Expectedly, the precision and recall values did not change when switching between Core genes and CorEx genes, because each of the non-core gene in the CorEx list had an expression pattern similar to at least one Core gene (**Fig. 3.2b**). While the precision recall numbers are not high enough for the gene expression to be used directly in predicting the ecDNA status, they were significantly better than a random subset of genes.

For precision values at least 0.7, the mean recall using CorEx genes was 0.49. In contrast, the mean recall for randomly selected genes was 0.23 (*p*-value 1.8e-35; Mann-Whitney U test).

To test the persistence of CorEx genes across tumor types, we re-computed Cliff's delta values for each of the 11 TCGA tumor types that had at least 10 ecDNA(+) and at least 10 ecDNA(-) samples. The directionality of gene expression patterns was significantly similar to TCGA in each tissue type with one exception (**Fig. 3.2c; Table S3.4**). Specifically, the *p*-values against a null hypothesis of no match to the direction in TCGA ranged from 4.2e-12 to 3.5e-85 (Fisher's exact test) for the significant associations (Methods). The results were similar if we tested using only Core genes (**Fig. S3.2a**). The sole exception was the tumor type of Sarcoma (SARC). It is notable that genome rearrangements in sarcomas are different from other solid tumors (Garsed et al., 2009). In addition to ecDNA, these samples are known to have extensively rearranged structures indicative of chromothripsis and neo-chromosome formation (Garsed et al., 2014), and it is also possible that the detection of ecDNA in these highly rearranged chromosomes was flawed. Overall, the results are consistent with a pan cancer role for the 643 CorEx genes in ecDNA maintenance, with a similar number of genes being up, or down regulated.

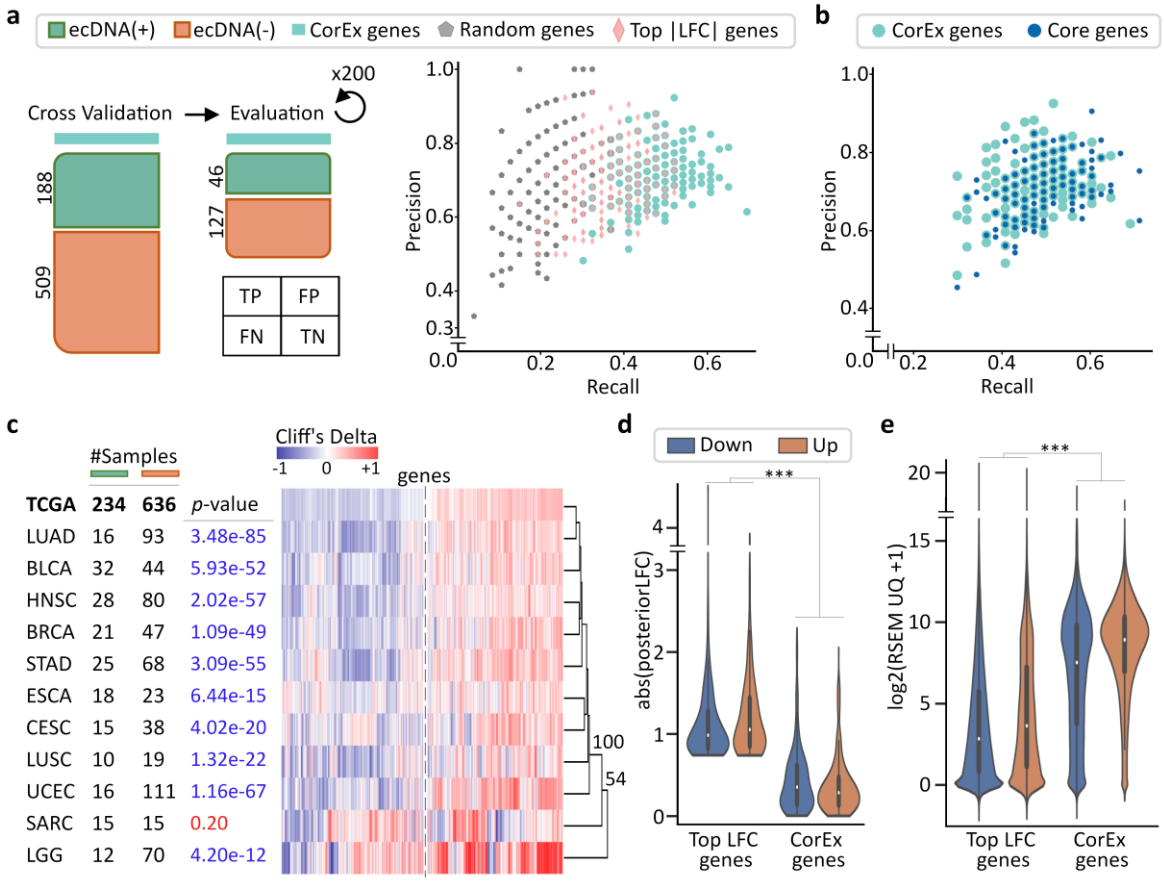


Figure 3.2: Validation of CorEx genes (a) Cross-validation experiments validating the predictive value of CorEx genes. CorEx genes have higher recall rates compared to 643 randomly selected genes and the top 643 genes (DE genes) based on logarithmic fold changes from a DESeq2 analysis. **(b)** CorEx genes have similar predictive value to Core genes alone. **(c)** CorEx genes are consistently up- or down-regulated in ecDNA(+) samples across tumor types, with the exception of SARC. **(d)** Of the 643 DE genes, 240 are up-regulated while 403 are down-regulated in ecDNA(+) samples. Of the CorEx genes, 325 are up-regulated while 318 are down-regulated. The absolute posterior LFC values of the Top LFC gene set is significantly greater than that of the CorEx genes (p-value 1.83e-158). **(e)** The normalized gene values of the CorEx genes are significantly higher than that of the Top LFC gene set (p-value 0.0). ***p < 0.001.

3.1.4: CorEx genes are very different from differentially expressed genes

We asked if the CorEx gene list was significantly different from a list of differentially expressed genes. We performed a differential expression analysis using DESeq2 (Love et al., 2014; Methods) and picked the 643 most significantly differentially expressed (DE) genes in terms of the absolute value of their posterior log-fold change (LFC) values. Using

the sign of the posterior LFC value as the determinant for directionality, 240 of these genes were up-regulated, while 403 were down regulated. Notably, this list was quite different with only 86 genes overlapping with the CorEx gene set (**Table S3.5**; **Fig. S3.2b**). As discussed in the introduction, the DE genes were also not as predictive of ecDNA status as the CorEx genes. For precision values at least 0.7, the recall ranged from 0.217 to 0.522, significantly less than the recall of CorEx genes analysis (Mann-Whitney U, *p*-value 4.8e-21; **Fig. 3.2a**).

The DE genes are also different from the CorEx genes by other metrics. Not surprisingly, the log-fold change (LFC) values of the DE genes were higher than the LFC values of the CorEx genes (**Fig. 3.2d,** MWU *p*-value 1.83e-158). However, much of the LFC change was due to the very low expression of the DE genes in ecDNA(+) or ecDNA(-). In fact, the CorEx genes had very high expression in general compared to the DE genes (**Fig. 3.2e,** MWU *p*-value 0.0). While the fold-change in expression between ecDNA(+) and ecDNA(-) samples was not that high, it was persistent across all samples.

As one example, the DE gene PNMT has the eighth highest posterior LFC value at 2.89, and most ecDNA(+) samples (223/234) have a normalized expression value less than 8.5, with 11/234 ecDNA(+) samples having an expression value of at least 11. Of the 11 samples, 5 samples contain PNMT on an ecDNA amplicon (**Fig. S3.2c**). Similarly, the DE gene ITLN1 has the second highest posterior LFC value of 3.92. However, most ecDNA(+) samples (210/234) have a normalized expression value less than 8. Of the 24/234 ecDNA(+) samples having an expression value of at least 8, 3 samples contain ITLN1 on a non-cyclic amplicon (**Fig. S3.2d**).

The results confirm our intuition that differential expression can arise due to multiple reasons, including low expression of the gene in a majority of samples, or the copy number

amplification of a gene in a few samples. In contrast, the CorEx genes are selected based on persistent over- or under-expression in ecDNA(+) samples.

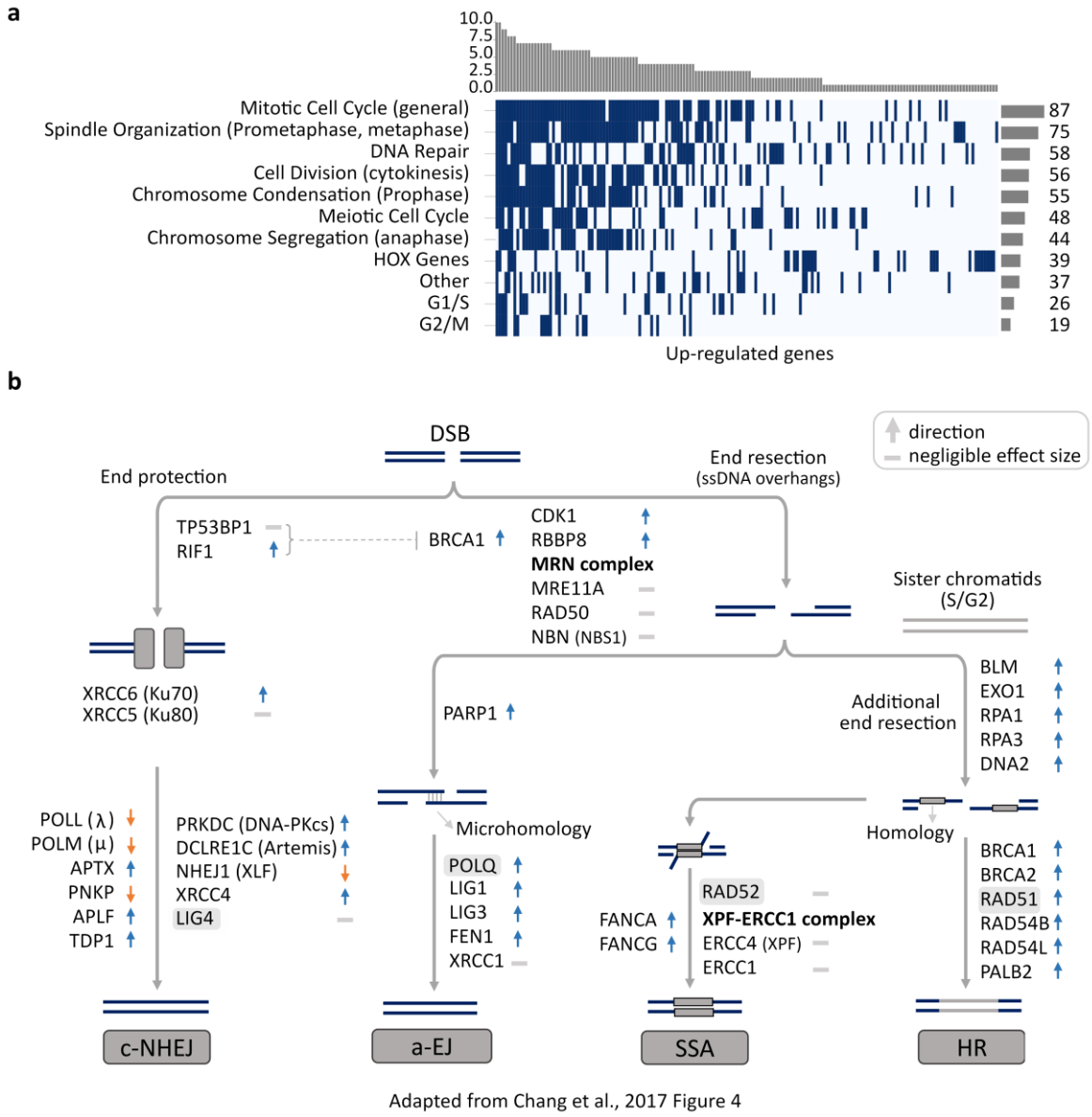


Figure 3.3: Up-regulated CorEx genes. (a) GO Biological processes enriched in up-regulated genes. **(b)** Genes up- or down-regulated in genes involved in major DNA damage repair pathways.

3.1.5: CorEx genes primarily up-regulate only three biological processes: Cell Cycle, Cell division, and DNA Damage Response

To identify enriched biological processes specific to either up-regulated or down-regulated genes in ecDNA(+) samples, we combined two metrics of effect size, Cliff's delta (Cliff, N., 1993, 1996), and posterior logarithmic fold change (Love et al., 2014) to determine the directionality of CorEx genes (Methods). The two effect size metrics were mostly in agreement in terms of directionality, with only 14 out of 7,288 genes passing the negligible effect size thresholds not in concordance. This more stringent approach, in comparison to a simple directionality based on the sign of a single effect size value, was applied given that gene set enrichment is dependent on not only the number of up- or down-regulated genes but also the degree of overlap with genes under a specific biological process term (Methods). Genes that do not pass a negligible effect size magnitude should not be considered up- or down-regulated in ecDNA(+) samples. Using this approach, among the 643 CorEx genes, 262 genes were found to be up-regulated in ecDNA(+), while 271 were found to be down-regulated (**Table 3.1**; Methods). 110 genes did not make the effect size cut-off. The numbers were similar for the 408 core genes, with 190 up-regulated, 196 down-regulated, and 22 genes not making the cut-off.

We performed a gene set enrichment analysis to identify the Gene Ontology (GO) biological processes that are enriched in CorEx genes (Methods). Briefly, we applied a one-sided Fisher's exact test using gene sets from MSigDB, using a false discovery rate of 5% (Benjamini-Hochberg procedure). The UP-regulated genes were enriched for 187 Biological processes (**Table S3.6**). Note that the GO-BP terms are not independent, because of their hierarchical organization, and sharing of genes across different GO terms. A visual

inspection revealed that the terms corresponded to only a few broad categories. Therefore, we used an approach similar to that used in DAVID (Huang et al., 2007) to cluster the biological processes enriched by the UP-regulated genes into 11 broad categories (**Table S3.7**; **Fig. S3.3**; Methods). The 11 categories were assigned a name upon manual inspection of the member terms, or called "Other." We next constructed a waterfall plot to understand the genes covered by these 11 categories (**Fig. 3.3a**).

Together, the 10 biological process categories explain 165 of the 262 up-regulated CorEx genes, and provide a very clear picture of the pathways involved. Specifically, the up-regulated genes largely belong to (a) cell-cycle regulation (MITOTIC/MEIOTIC CELL CYCLE, CHROMOSOME CONDENSATION, G1/S, G2/S) (b) cell-division process (SPINDLE ORGANIZATION, CELL DIVISION, CHROMOSOME SEGREGATION), (c) DNA Damage response (DNA_REPAIR), or (d) the HOX Gene cluster (**Table S3.7**). As mentioned in the co-expression results above, one of the largest clusters containing 21 members, cluster #1509 contained 9 Core genes and 12 co-expressed genes, and is enriched in GOBP terms related to the cell cycle (**Table S3.8**).

While Meiotic cell-cycle is also enriched, only 7 genes are allocated specifically to the group of enriched terms: SEPP1, SNRPA1, TMEM203, RNF114, TAF4, TNFAIP6, and PTX3 (**Table S3.9**). Though these genes have roles in the meiotic cell cycle, they have also been implicated in cancer and other inflammatory diseases. The spliceosomal protein SNRPA1 is a pro-metastatic splicing enhancer (Fish et al., 2022). TNFAIP6, together with PTX3, activates the Wnt/β-catenin pathway to promote gastric carcinoma cell invasion (Cui et al., 2022). TMEM203 is a STING-centered signaling regulator implicated in inflammatory diseases (Li et al., 2019). RNF114 is a zinc-binding protein whose over-expression is an

indicator of epithelial inflammation and implicated in various tumors (Feng et al., 2022). TAF4, a transcription initiation factor, when overexpressed, is implicated in ovarian cancer by playing a role in dedifferentiation that promotes metastasis and chemoresistance (Ribeiro et al., 2014).

Only 4 genes assigned to the "Other" category did not enrich a biological process from the 10 categories mentioned above: FADD, SHCBP1, NRAS, and TSPAN31. These 4 genes are implicated in tumor proliferation. TSPAN31 and SCHBP1 are associated with CDK4, a cell cycle regulator (Dong et al., 2019; Wang et al., 2017). TSPAN31 is co-amplified with CDK4, MARCH9, CYP27B1, METTL1, METTLE21B, TSFM, AVIL, and CTDSP2, and associated with poor prognosis (Martinez-Monleon et al., 2022). Indeed, we also observed that TSPAN31 is co-amplified with these genes on ecDNA amplicons in 13 TCGA samples (**Table S3.10**). FADD amplification is associated with cancers including head and neck squamous cell carcinoma, and amplified with other oncogenes including CTTN (Chien et al., 2016). We also observed that FADD and CTTN are co-expressed in cluster #6306 and seen amplified together on ecDNA in 4 samples (**Table S3.2**; **Table S3.10**). In summary, the UP-regulated CorEx genes primarily impact 3 biological processes: Cell cycle process, cell division, and DNA damage repair.

3.1.6: CorEx genes upregulate specific double-strand break repair pathways

The enriched DNA damage response GO terms contained terms "double-strand break repair," and "recombination", but none contained the terms "single-strand," "nucleotide-excision," or "mismatch-repair." (**Table S3.7**), suggesting that the CorEx genes largely up-regulate pathways involved in double-strand break (DSB) repair and homologous recombination-based repair. In normal cells, the most efficient repair mechanism for DSB

repair is classical non-homologous end-joining (c-NHEJ). End resection at the breaks is used to generate 5′ or 3′ overhangs to generate small regions of micro-homology, which are then ligated back. When c-NHEJ is compromised, other end joining pathways are activated, all of which involved extensive resection. These included error free homology directed repair, which requires a template, such as the sister chromatid when the cell is in G2 phase. Other choices include the more error-prone alternative end-joining (Alt-EJ), or single-strand annealing (SSA) pathways (Chang et al., 2017).

The choice of these varied DSB repair mechanisms for ecDNA maintenance is not well understood. We hand-curated a list of 129 genes involved in DSB repair and marked them for their role in one or more of these 4 pathways (**Table S3.11**; References). Of these genes, a high number (82) were up-regulated in ecDNA(+), while a smaller number (17) were down-regulated, relative to ecDNA(-) samples. For this analysis, we recomputed Cliff's delta and posterior LFC effect size values using 1,440 samples representing 24 tumor subtypes (**Table S3.12**; **Table S3.15**). This contrasts with all genes, where near identical number of genes (5,256, and 5,251) were up- and down-regulated in ecDNA(+) samples (*p*-val< 0.0001; Fisher's exact test) confirming that DDR genes are significantly up-regulated relative to all differentially expressed genes. When broken down to the roles of genes in individual DSB repair pathways, we find that while Alt-EJ with 11 up-regulated and 1 down-regulated genes each (*p*-val 0.0063), SSA (11 up, 1 down), and HR (46 up, 8 down; *p*-val<0.00001) are all up-regulated, classical NHEJ (14 up, 7 down; *p*-val: 0.19) is not significantly up-regulated in ecDNA(+) samples relative to ecDNA(-) samples (Methods).

The expression of key genes in these pathways is consistent with this observation (**Fig. 3.3b**). P53-binding protein 1 (TP53BP1) is a key mediator of pathway choice, and

along with RIF1, blocks end resection to promote the c-NHEJ pathway choice (Daley and Sung, 2014). The binding of the Ku70–Ku80 heterodimer to a DSB then acts as a recruitment platform for the subsequent binding of other c-NHEJ proteins. While XRCC4–DNA ligase IV (LIG4) is sufficient to repair blunt ends, depending on the degree of microhomology between the ends, different combinations of nucleases and polymerases make up various subpathways involved in end-processing prior to ligation (Chang et al., 2017). While we found that most of the major c-NHEJ genes (i.e., Ku70, DNA-PKcs, Artemis, XRCC4, TDP1) were up-regulated in ecDNA(+) samples relative to ecDNA(-) samples, genes related to the subpathways for incompatible 5' and 3' ends were down-regulated in ecDNA(+) samples, including XLF, Polymerase λ (POLL), and μ (POLM). Specifically, in the case of incompatible 5' overhangs, Artemis-dependent resection is followed by XLF stabilization of the Ku70-Ku80 heterodimer prior to ligation (Chang et al., 2017). In the case of incompatible 3' overhangs, POLL and POLM are required to generate regions of microhomology prior to ligation. The difference between the two being that POLM promotes the ligation of incompatible 3′ overhangs in a template independent manner, while POLL promotes the ligation of terminally compatible overhangs requiring fill-in synthesis (Ghosh et al., 2021). Furthermore, the polynucleotide kinase (PNKP) necessary for the processing of 5' ends lacking a phosphate (Chang et al., 2017) is also down-regulated in ecDNA(+) samples. These results suggest that a subset of classical NHEJ subpathways are not significantly more active in DSB repair in ecDNA(+) samples relative to ecDNA(-) samples.

The factor TP53BP1 is key to blocking resection and promoting the c-NHEJ pathway choice, but is displaced by BRCA1 and the MRN complex to initiate resection in the broken strands (Daley et al., 2014; Fouquin et al., 2017). BRCA1 was significantly up-regulated in

ecDNA(+) samples (**Fig. 3.3b**). Among the alternative repair pathways, key genes in the Alt-EJ pathway, including Pol-θ, LIG1, LIG3, FEN1 are all significantly up-regulated. Similarly, PARP1, which also participates in Alt-EJ when Ku proteins are impaired, is significantly up-regulated. The genes BLM, EXO1, RPA1,3 which promote additional resection, as well as BRCA1, BRCA2, RAD51 and others that support Homology directed repair are significantly up-regulated. Thus, while DSB repair genes enriched among the up-regulated CorEx genes, the specific pathway choice is more subtle. It is dominated by alt-EJ and homology directed repair pathways, while c-NHEJ is repressed. The alt-EJ pathway likely has an outsized role in DSB repair for ecDNA, particularly during G1, because homology directed repair may not be possible in the absence of a template.

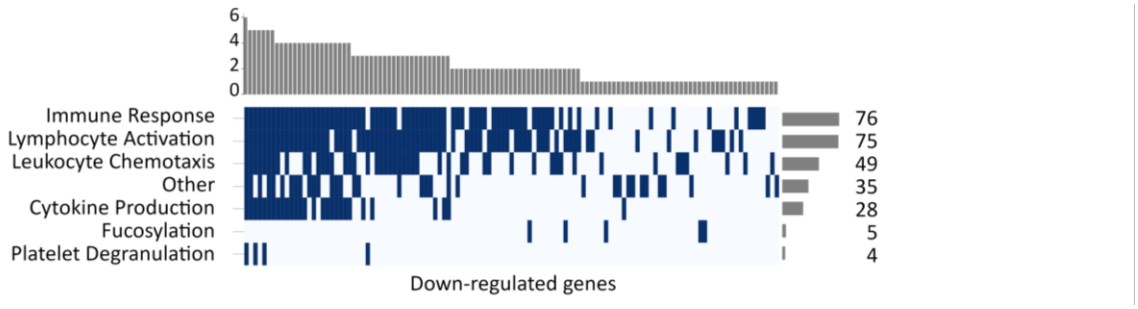


Figure 3.4: Down-regulated CorEx genes. GO biological processes enriched in down-regulated genes.

3.1.7: CorEx genes primarily down regulate immune system processes

Using a process similar to that used for up-regulated genes, the down-regulated genes enriching 73 GO terms (**Table S3.13**) could be clustered into 7 broad categories, including "Other." (**Fig. 3.4**; **Table S3.14**; **Fig. S3.4a).** Surprisingly, almost all categories were immunomodulatory, and related to some facet of the cancer immune cycle (Chen and Mellman, 2013). The most enriched broad category contained 75 genes relating to Leukocyte

activation pathways, including genes enriching T-cell activation (28 CorEx genes; *p*-val 2.58E-05), and regulation of cell-cell adhesion (18 CorEx genes; *p*-val 7.24e-03). Cytokine activation was down-regulated (especially for genes in the IL-12 (6 genes, *p*-val 3.10e-02), TNF super-family (12 CorEx genes, *p*-val 7.24e-03)) and Toll like receptor 2 signaling (4 CorEx genes, *p*-val 4.48e-03), which suggest a reduction of the proinflammatory environment needed for cancer antigen presentation. Finally, the broad category of Leukocyte chemotaxis is also enriched among the down-regulated genes. The chemotaxis genes include many chemokines and their receptors involved in trafficking of T cells to the site of the tumor. Other downregulated genes include fucosyl-transferases and genes involved in NF-kb regulation (14 CorEx genes, *p*-val 2.28e-02). Thus, an intriguing part of the ecDNA(+) samples is the downregulation of many genes involved in mounting a tumor response.

As we looked at bulk sequencing, the sampled cells included tumor cells and cells from the tumor microenvironment. A recent study analyzing the tumor microenvironment (TME) of ecDNA vs. non-ecDNA in 7 tumor subtypes revealed an association of ecDNA presence with immune evasion (Wu et al., 2022). The immune typing of TCGA samples using the Thorsson et al. classifications further revealed an up-regulation of lymphocyte depletion as also a down-regulation of inflammatory and TGF-β dominant qualities in tumors with ecDNA. Briefly, Thorsson et al. identified 6 immune subtypes via consensus clustering of 160 immune expression signatures to define modules that were then used for model-based clustering of 30 cancer types (Thorsson et al., 2018). More than 10,000 TCGA samples spanning 33 cancer types were then categorized into the following 6 immune subtypes: wound healing (C1), IFN-γ dominant (C2), inflammatory (C3), lymphocyte

depleted (C4), immunologically quiet (C5), and TGF-β dominant (C6). Our results, which have slightly improved classification of ecDNA(+) cells were broadly consistent with those from the Wu et al study (**Fig. S3.4b**).

3.2: Methods

TCGA sample ecDNA status classification

Amplicon Classifier (version 049) classified amplicons detected in 1,921 TCGA samples into five sub-types: ecDNA, BFB, heavily rearranged (complex non-cyclic), linear, and no-amplification. When classifying a sample with multiple amplicons, the order of preference is as follows: ecDNA, BFB, heavily rearranged, linear, and no-amplification. We treated the latter two categories as ecDNA(-). Of the 1,921 samples, 1,535 samples classified as ecDNA(+) and ecDNA(-) had RNA-seq data, including 1,406 primary solid tumor samples, 95 metastatic samples, and 34 primary blood derived cancer – peripheral blood samples. Removing metastatic samples results in a total of 1,440 samples, including 243 ecDNA(+) and 1,197 ecDNA(-) samples. While the set of 1,440 samples represented 24 tumor types, ten of these tumor types had insufficient numbers of ecDNA(+) samples, including four tumor types with no ecDNA(+) samples. To prevent the 561 ecDNA(-) samples representing these tumors from skewing the analysis, we removed 570 samples representing tumor types with less than three ecDNA(+) samples. This resulted in a total of 870 samples representing 14 tumor types, of which 234 were classified as ecDNA(+) and 636 were classified as ecDNA(-).

Gene expression Datasets

Gene expression data for 32 studies part of the TCGA Pan-cancer Atlas was downloaded from cBioPortal (01.05.2021) (https://www.cbioportal.org/). The cBioPortal "data_RNA_Seq_v2_expression_median.txt" data is sourced from the file "EB++AdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv" (synapse id:

syn4976363). Briefly, the matrices contain batch corrected values of the upper-quartile (UQ) normalized RSEM estimated counts data from Broad firehose (tumor.uncv2.mRNAseq_RSEM_all.txt). Missing values due to the batch effect correction process were imputed using *K*-nearest neighbors (KNN). For each tumor type, values were imputed based on gene vectors under the assumption that genes are similarly expressed between samples of the same tumor type. For genes with less than 60% of samples with missing values, values were imputed using the logarithmic (base 2) of the gene expression value plus one, and subsequently back-transformed when writing the imputed matrices to file. The resulting gene expression matrix used for the Boruta analysis described below consisted of 870 TCGA samples and 16,309 protein-coding genes (based on "hgnc_complete_set.txt" downloaded from HGNC on 7.24.2018). To generate a RSEM raw counts matrix for the DESeq2 analysis described below, mRNAseq_Preprocess.Level_3 data was downloaded from Broad Firehose (tumor.uncv2.mRNAseq_raw_counts.txt).

Boruta Analysis

To identify a minimal set of genes whose expression values were predictive of the sample being ecDNA(+), we used an automated feature selection algorithm, Boruta **(**Kursa and Rudnicki, 2010). Given a gene expression matrix, *M*, of dimension *r* x c, where *r* is the number of samples and *c* is the number of genes, Boruta generates *c* shadow feature vectors by random shuffling of gene vectors in matrix *M*, generating a new matrix *M'* of dimension *r* x 2*c*. A random forest algorithm is then used to quantify the importance of each gene in separating ecDNA(+) from ecDNA(-) samples. In a given Boruta run, for each iteration $i \in n$, where $i = \{1,2,3, \dots, n\}$, there are two possible outcomes for a gene: 1) if the gene scores

higher than the best scoring shadow feature, the gene is considered a "hit", and 2) if the gene scores lower than the best scoring shadow feature, the gene is considered a "non-hit". The number of hits, k_g, for each gene, g, after i iterations is stored in a list, L_i. Similar to a coin toss event of n trials, the probability that a gene is a "hit" or "non-hit" in each iteration is 0.5, and follows the binomial distribution.

After i iterations, the probability that a gene has k or more hits, $P(X > x)$, where $x = k - 1$, can be computed using the binomial distribution survival function. This probability is computed for each gene, g, with k_g number of hits and stored in list P_i. As Boruta involves testing multiple genes per iteration and testing the same genes repeatedly over i iterations, two multiple hypothesis testing corrections are applied to list P_i. First, the Benjamini-Hochberg procedure is applied to control the False Discovery Rate (FDR) at 5%, and second, the Bonferroni correction is applied to control for the family-wise error rate at $\alpha = 0.05$. If the probability of a gene satisfies both corrections, the gene is accepted as a Boruta gene. Separately, the probability that a gene, g, has k_g or lower hits, $P(X \leq k_g)$, can be computed using the binomial cumulative distribution function. For genes that are not accepted, if the probability satisfies both corrections, the gene is rejected as a Boruta gene.

The Boruta script used for this analysis is a revised version of the BorutaPy python package (6.21.2021; https://github.com/scikit-learn-contrib/boruta_py) with the addition of early stop parameters (≤50 tentative feature counts stagnant for 5 iterations) and output of hits per iteration. We randomly split the 870 samples into training (80%) and testing (20%) data to enable the evaluation of selected features. Given the unequal representation of ecDNA(+) and ecDNA(-) samples within each of the tumor types, we opted to perform the following procedure 200 times. For each of the 200 runs, Boruta is performed on the training

data. In short, the Boruta method takes as input the original expression matrix, *M*, appends shadow features that are randomized versions of gene features in *M*, performs random forest classification on the data (class_weight=balanced_subsample, max_depth=7), selects the highest feature importance score, *s*, of shadow features as a cut off, and returns features with a higher score than *s*. Of the 941 genes identified in the Boruta analysis, 408 genes were selected in at least 10 of the 200 Boruta runs, and subsequently defined as the Core set of genes in downstream analyses.

Highly Co-expressed genes

To identify genes co-expressed with the cores set of Boruta genes, hierarchical clustering of the 16,309 genes was performed using the R package pvclust (ver. 2.2-0; dist.method=correlation, method=ward.D2, nboot=1000). A total of 843 significant clusters (AU > 0.95) with at least 1 Boruta gene were selected, consisting of 1,375 genes. To obtain the final list of CorEx genes, we apply a minimal count of 10 runs for the gene or 10 runs for the cluster of genes seen in 200 Boruta runs. A cluster is determined to be seen in a Boruta run if at least one of its members is selected in the run. This results in 354 clusters, with a total number of 643 genes, of which 408 are Core genes.

Evaluation of CorEx genes

To evaluate the 408 Core genes as predictive of ecDNA presence in tumor samples, a random forest classifier was trained on the 80% data sets and tested on the 20% data sets defined previously. Hyper-parameters are tuned using RandomizedSearchCV followed by GridSearchCV (k_fold =5, f1 score). Furthermore, we performed a similar evaluation using

408 randomly selected genes instead of the Core genes for each of the 200 training and testing data sets. A similar evaluation was performed using 643 randomly selected genes instead of the CorEx genes. We further evaluated a set of 643 most differentially expressed genes based on the absolute posterior LFC effect size values from a conventional DE analysis using DESeq2 (Love et al., 2014) as described below.

Default DE analysis

In a typical differential expression workflow, a package such as DESeq2 (Love et al., 2014) takes as input a matrix of raw read counts and performs the following steps: i) Given K_{ij}, the raw read count for gene *i* in sample *j*, a model is built for each gene assuming that read counts follow certain distributions (i.e., negative binomial (NB) where the mean ≠ variance): $K_{ij} \sim NB(u_{ij}, \alpha_i)$, where u_{ij} is the mean and α_i the dispersion, ii) parameters for the NB model (e.g., dispersion) are estimated, iii) a NB-generalized linear model (GLM) is fitted for each gene, iv) a Wald test is performed on each of the model coefficients (i.e., log fold change) from step iii to extract *p*-values, and a multiple-testing correction, such as Benjamini and Hochberg procedure, is performed to obtain adjusted *p*-values, v) differentially expressed genes are then selected using an adjusted *p*-value<0.05 and a user-defined biologically relevant LFC threshold.

In such workflows, parameter estimation often involves assumptions based on pooled information of genes. For example, during the empirical Bayes shrinkage dispersion estimation in step 2, DESeq2 assumes that genes of similar average expression strength have similar dispersion. After a gene-wise dispersion is estimated using maximum likelihood for each individual gene (maximum-likelihood estimate, MLE), a smooth curve is fitted to these

values, and gene-wise MLEs are shrunk towards the smooth curve to produce the maximum *a posteriori* (MAP) as the final dispersion estimate. Furthermore, to correct for a strong variance of LFC estimates for genes with low read counts, DESeq2 shrinks LFC estimates towards zero using an empirical Bayes approach, resulting in the shrunken LFC metric for effect size (i.e., posteriorLFC).

In our analyses, we performed a default DESeq2 (R package, ver. 1.36.0) analysis for two reasons: i) to obtain posteriorLFC effect size values per gene for integration with its Cliff's delta effect size value when determining if a gene is up- or down-regulated in ecDNA(+) samples, and ii) to obtain a list of *n* top-ranked genes by absolute value of the posteriorLFC (with application of an adjusted *p*-value <0.05 cutoff and LFC threshold of $log_2(1.1) = 0.13$) for use in comparison against genes selected as important in the prediction of ecDNA in samples. For comparisons against Core genes, *n* is set to 408, and for comparisons against CorEx genes, *n*, is set to 643.

To obtain the posterior LFC effect size metric between ecDNA(+) vs. ecDNA(-) samples for each gene's expression, we fed as input to DESeq2 a matrix of raw RSEM estimated counts (matrix construction described above). To take into account batch effects, we included the center and platform information of samples, downloaded from synapse id syn4976363 (EB++GeneExpAnnotation.tsv), in the design of the DESeq object:

```
DESeq_object <- DESeqDataSetFromMatrix(countData =
AC_0_4_9_TCGA_matrix, colData = coldata, design = ~batch+condition)
```

To compute results, the lfcThreshold was set to $log_2(1.1)$ for an accurate computation of *p*-values and the contrast set to c("condition", "ecDNA(+)", "ecDNA(-)") to obtain the logarithmic fold change of the form $\log_2\left(\frac{\text{ecDNA}(+)}{\text{ecDNA}(-)}\right)$. By setting the lfcThreshold,

the null hypothesis tested is that $|LFC| \leq \theta$, where $\theta = \log_2(1.1)$, and the alternative hypothesis is that $|LFC| > \theta$. A $log_2(1.1)$ value is chosen as the minimal value/negligible effect size threshold as it represents a 10% fold-change, and anything below this fold-change would likely not be of biological interest (McCarthy and Smyth, 2014). The specific commands run are as follows:

```
DESeq_object$condition <- relevel(DESeq_object$condition, ref =
"Non_ecDNA")

DESeq_object <- DESeq(DESeq_object)

results <- results(DESeq_object, lfcThreshold= log2(1.1),
contrast=c("condition","ecDNA","Non_ecDNA"))
```

To obtain the shrunken LFC (i.e., posteriorLFC) effect sizes as described above, we used the lfcShrink function provided in DESeq2, and used the default apeglm method for the empirical Bayes shrinkage procedure (Zhu, Ibrahim, and Love 2018):

```
lfcShrink(DESeq_object, lfcThreshold= log2(1.1),
coef="condition_ecDNA_vs_Non_ecDNA", type="apeglm")
```

Up- or Down-regulated genes in ecDNA(+) samples

To categorize genes as "up-" or "down-" regulated in ecDNA(+) samples, we integrated two effect size metrics, Cliff's delta (*d*; Cliff 1993, 1996) and the DESeq2 posterior log2 fold change (posteriorLFC; Love et al., 2014). Effect size is a measure of the magnitude of deviation from the null hypothesis, and unlike *p*-values, has the advantage of not being impacted by sample size (Romano, 2006). This property is especially useful when comparing effect size values of a gene between tests where sample sizes differ. Comparing *p*-values between such tests would be invalid.

Cliff's delta, d, is a non-parametric measure of the separation between two distributions and ranges from -1 to 1. Given two distributions, $X = \{i_1, i_2, \dots, i_m\}$ and $Y = \{j_1, j_2, \dots, j_n\}$, comparisons are made between each of m values in X and n values in Y. Cliff's delta is computed as $d = \frac{\#(i>j)-\#(i<j)}{mn}$, where $\#(i > j)$ is the number of times a member of X is greater than a member of Y, and $\#(i < j)$ is the number of times a member of X is less than a member of Y (Cliff, 1993). A negative d indicates that values in Y tend to be higher than X, while a positive d indicates that values in X tend to be higher than Y. The magnitude of the effect size of Cliff's delta can be separated into four levels: $|d| < 0.147$ for negligible effects, $0.147 \leq |d| < 0.33$ for small effects, $0.33 \leq |d| < 0.474$ for medium effects, and $|d| \geq 0.474$ for large effects (Romano, 2006). The python package used to compute Cliff's delta values can be accessed at https://github.com/neilernst/cliffsDelta. The input values used to compute Cliff's delta are log-transformed normalized gene expression values plus one as described in the "Gene expression Datasets" section of methods.

The DESeq2 posterior LFC is computed as described above. To allow integration of the Cliff's delta effect size with the posterior LFC, we also separated the LFC values into four levels: $|\text{LFC}| < log_2(1.1)$ for negligible effects, $log_2(1.1) \leq |\text{LFC}| < log_2(1.5)$ for small effects, $log_2(1.5) \leq |\text{LFC}| < log_2(2)$ or medium effects, and $|\text{LFC}| \geq log_2(2)$ for large effects.

For each gene, g, whether the gene is up-regulated or down-regulated in ecDNA(+) samples is determined by the signage and magnitude of its effect sizes d_g and LFC_g. The initial criteria for d_g and LFC_g to be used as a determinant in the direction of a gene is for it to have a magnitude larger than that of a negligible effect. If the signage of d_g and LFC_g are both positive, gene g is considered up-regulated in ecDNA(+). If only a single value has a

magnitude larger than the negligible effect threshold (e.g, $d_g > 0$ and $LFC_g = -0.1$), gene g is considered up-regulated in ecDNA(+). In the case of conflicting signages between the two values, the effect size with a larger magnitude takes precedence. For example, if $d_g = -0.2$ and $LFC_g = 0.847$, given that d_g has a small negative effect and LFC_g has a large positive effect, gene g is considered up-regulated in ecDNA(+) samples.

Tumor heatmap

A Cliff's delta effect size matrix representing 643 CorEx genes was generated to compare TCGA with tumor expression patterns. For each of the 11 tumor types with at least 10 ecDNA(+) and at least 10 ecDNA(-) samples, we re-computed Cliff's delta. Using a Fisher's exact test (fisher_exact function from the scipy.stats python package; alternative hypothesis: two-sided), we tested the null-hypothesis of whether the up- and down-directionality of CorEx genes in TCGA vs. each tumor were independent of each other. The contingency table is as below. The directionality of a gene (up or down) is based solely on the signage of the gene's Cliff's delta effect size value.

		Tumor	
		UP	DOWN
TCGA	UP	a	b
	DOWN	c	d

Gene Set Enrichment Analysis

To identify gene ontology biological processes that are enriched in our CorEx set of genes, we applied a hypergeometric test (one-sided Fisher's exact test; fisher_exact function from the scipy.stats python package) using gene set to gene mappings from msigdb (c5.go.bp.v7.5.1.entrez) and a universe of N=16k for the number of genes measured in the

RNAseq data. The false discovery rate was controlled at 5% and adjusted *p*-values were computed using the Benjamini-Hochberg procedure (fdrcorrection from python statsmodels package). A final set of GOBP terms with adjusted *p*-value<0.05 was used for downstream analysis. The contingency table is as follows, where N is the total number of genes in the universe, n is the number of DE genes (either up- or down-regulated in ecDNA(+)), m is the number of genes in the gene set, and k is the number of DE genes in the gene set.

	DE	Non-DE
Inside gene set	k	m-k
Outside gene set	n-k	N+k-n-m

Grouping Gene Sets

To group enriched gene sets into categories for visualization purposes, Cohen's kappa coefficient (python sklearn cohen_kappa_score) was used to determine term-term "connectivity" (agreement of term-term pairs) – an approach described in the DAVID paper (Huang et al., 2007). Given a $r \times c$ binary matrix, where enriched GOBP terms are rows, CorEx genes are columns, and values are 1 if a CorEx gene is part of the GOBP term or 0 otherwise, kappa scores were computed between each pair of terms, where term $t_i \in t$ and $i = \{1,2,3, \ldots . n\}$: $Kappa\_Score(t_x, t_y)$, where $x \in i$ and y $\in i$.

Each term, t_i, formed an initial seeding group, g_i, where a term t_x is part of g_i if $Kappa\_Score(t_i, t_x) \geq score\ threshold$. If at least 50% of term-term pairs in g_i have a $Kappa\_Score \geq score\ threshold$, the initial seeding group g_i is retained for the next step. The second criteria ensures that terms within the same seeding group have strong interconnectivity. An iterative merging of seeding groups then follows: groups sharing p% or more members are merged. Representative term determined by highest interconnectivity

score with other members of group. After the automatic grouping process, a manual inspection leads to the merging of outliers or smaller groups into representative groups. We used a kappa score threshold of 0.5, and condition of p≥25% of shared members when merging for the down-regulated genes, and a kappa score threshold of 0.6 and condition of p≥50% of shared members when merging for the up-regulated genes.

DDR pathway genes

We hand-curated 88 genes for double-stranded break DNA damage repair pathways (a-EJ, HR, c-NHEJ, SSA) via an extensive literature search, and added an additional 51 genes from the following MSigDB (c5.go.bp.v2022.1.Hs.entrez.gmt) GO biological process terms: GO:0097680 (DOUBLE STRAND BREAK REPAIR VIA CLASSICAL NONHOMOLOGOUS END JOINING), GO:0097681 (DOUBLE STRAND BREAK REPAIR VIA ALTERNATIVE NONHOMOLOGOUS END JOINING), GO:1905168 (POSITIVE REGULATION OF DOUBLE STRAND BREAK REPAIR VIA HOMOLOGOUS RECOMBINATION), and GO:0045002 (DOUBLE STRAND BREAK REPAIR VIA SINGLE STRAND ANNEALING). This resulted in a final list of 129 genes. The directionality of the genes, as either up- or down-regulated in ecDNA(+) samples, is based on the full set of 1,440 samples, consisting of 243 ecDNA(+) and 1,197 ecDNA(-) samples representing 24 tumor types (**Table S3.12**; **Table S3.15**).

To test whether the number of genes passing our effect size thresholds for all genes 5,256 (UP) and 5,251 (DOWN) was significantly different from the up-/down-regulated genes implicated in each of the 4 pathways, we performed a Fisher's exact test (fisher_exact

function from the scipy.stats python package) on the contingency table below, where c and d are the up- and down-regulated genes for each of the pathways tested.

Contingency table:

	UP	DOWN
All genes	5,256	5,251
Pathway	c	d

Pathway values:

	c	d
c-NHEJ	14	7
Alt-EJ	11	1
HR	46	8
SSA	11	1

Physical presence on amplicons

To determine physical presence of a gene on an amplicon, gene coordinates listed in the gene annotations file GRCh37/human_hg19_september_2011/ Genes_July_2010_hg19.gff, downloaded from the AA repo on 3/21/2022, were mapped to amplicon genomic intervals (bed files). A gene is determined to be physically present on an ecDNA amplicon if its genomic coordinates are encompassed within the ecDNA amplicon genomic intervals.

Ranking of CorEx genes

To rank CorEx genes by importance, we computed harmonic mean rank values based on three categories: a) the average MDI (feature importance) values extracted from the trained random forest models on 200 training sets during the evaluation method described above, b) the number of Boruta runs (out of 200) that a gene is selected in, rounded to 2-

digits, and c) the number of Boruta runs (out of 200) that a gene is selected in when counting by cluster, rounded to 2-digits. The ranks for each category are adjusted separately so that genes with the same value share the same rank value. For example, if using the Boruta run count as the rank value:

	Run count	round(run_count/10.0)	Rank	Adjusted rank
Gene A	200	20	1	1
Gene B	200	20	2	1
Gene C	200	20	3	1
Gene D	187	19	4	4

The harmonic mean rank is defined as:

$$\text{harmonic mean rank} = \frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$$

Tumor immune subtype

We classified ecDNA(+) and ecDNA(-) samples into the 6 immune subtype categories (Thorsson et al., 2018) provided in Table S6 from Bagaev et al., 2021.

Acknowledgements

Chapter 3 includes content of material being prepared for submission by Lin, Miin S. and Bafna, Vineet in a manuscript with the tentative title “Genes involved in the maintenance of extrachromosomal DNA.” The dissertation author was the primary investigator and author of this paper.

3.3: References

Bagaev A, Kotlov N, Nomie K, Svekolkin V, Gafurov A, Isaeva O, Osokin N, Kozlov I, Frenkel F, Gancharova O, Almog N, Tsiper M, Ataullakhanov R, Fowler N. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy *Cancer Cell.* 39(6):845-865.e7

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, 57(1), 289-300.*

Chang, H., Pannunzio, N., Adachi, N., & Lieber, M. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol, 18(8), 495-506.*

Chen, D., & Mellman, I. (2013). Oncology meets immunology: the cancer-immunity cycle. *Immunity, 39(1), 1-10.*

Chien, H., Cheng, S., Chuang, W., Liao, C., Wang, H., & Huang, S. (2016). Clinical Implications of FADD Gene Amplification and Protein Overexpression in Taiwanese Oral Cavity Squamous Cell Carcinomas. *PLoS One, 11(10), e0164870.*

Cliff, N. (1996). Answering Ordinal Questions with Ordinal Data Using Ordinal Statistics. *Multivariate Behav Res, 31(3), 331-50.*

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114(3), 494–509.*

Cui, H., Zhang, L., Chen, B., Zhang, F., Xu, H., Ma, G., Cao, L., & Wang, M. (2022). TNFAIP6 Promotes Gastric Carcinoma Cell Invasion via Upregulating PTX3 and Activating the Wnt/. *Contrast Media Mol Imaging, 2022, 5697034.*

Daley, J., & Sung, P. (2014). 53BP1, BRCA1, and the choice between recombination and end joining at DNA double-strand breaks. *Mol Cell Biol, 34(8), 1380-8.*

Dong, Y., Yuan, Y., Yu, H., Tian, G., & Li, D. (2019). SHCBP1 is a novel target and exhibits tumor-promoting effects in gastric cancer. *Oncol Rep, 41(3), 1649-1657.*

Feng, Z., Li, L., Zeng, Q., Zhang, Y., Tu, Y., Chen, W., Shu, X., Wu, A., Xiong, J., Cao, Y., & Li, Z. (2022). Silencing Inhibits the Proliferation and Metastasis of Gastric Cancer. *J Cancer, 13(2), 565-578.*

Fish, L., Khoroshkin, M., Navickas, A., Garcia, K., Culbertson, B., Hänisch, B., Zhang, S., Nguyen, H., Soto, L., Dermit, M., Mardakheh, F., Molina, H., Alarcón, C., Najafabadi, H., & Goodarzi, H. (2021). A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science, 372(6543).*

Fouquin, A., Guirouilh-Barbat, J., Lopez, B., Hall, J., Amor-Guéret, M., & Pennaneach, V. (2017). PARP2 controls double-strand break repair pathway choice by limiting 53BP1 accumulation at DNA damage sites and promoting end-resection. *Nucleic Acids Res, 45(21), 12325-12339.*

Garsed, D., Holloway, A.J., Thomas, D.M. (2009). Cancer-associated neochromosomes: a novel mechanism of oncogenesis. *Bioessays*, 31(11):1191-200.

Garsed, D., Marshall, O., Corbin, V., Hsu, A., Di Stefano, L., Schröder, J., Li, J., Feng, Z., Kim, B., Kowarsky, M., Lansdell, B., Brookwell, R., Myklebost, O., Meza-Zepeda, L., Holloway, A., Pedeutour, F., Choo, K., Damore, M., Deans, A., Papenfuss, A., & Thomas, D. (2014). The architecture and evolution of cancer neochromosomes. *Cancer Cell, 26(5), 653-67.*

Ghosh, D., & Raghavan, S. (2021). 20 years of DNA Polymerase μ, the polymerase that still surprises. *FEBS J, 288(24), 7230-7242.*

Huang, D., Sherman, B., Tan, Q., Collins, J., Alvord, W., Roayaei, J., Stephens, R., Baseler, M., Lane, H., & Lempicki, R. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol, 8(9), R183.*

Kim, H., Nguyen, N., Turner, K., Wu, S., Gujar, A., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., Amin, S., Yi, E., Menghi, F., Schulte, J., Henssen, A., Chang, H., Beck, C., Mischel, P., Bafna, V., & Verhaak, R. (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet, 52(9), 891-897.*

Kobayashi, Y., Masuda, T., Fujii, A., Shimizu, D., Sato, K., Kitagawa, A., Tobo, T., Ozato, Y., Saito, H., Kuramitsu, S., Noda, M., Otsu, H., Mizushima, T., Doki, Y., Eguchi, H., Mori, M., & Mimori, K. (2021). Mitotic checkpoint regulator RAE1 promotes tumor growth in colorectal cancer. *Cancer Sci, 112(8), 3173-3189.*

Kursa, M., & Rudnicki, W. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36(11), 1-13.*

Li, Y., James, S., Wyllie, D., Wynne, C., Czibula, A., Bukhari, A., Pye, K., Bte Mustafah, S., Fajka-Boja, R., Szabo, E., Angyal, A., Hegedus, Z., Kovacs, L., Hill, A., Jefferies, C., Wilson, H., Yongliang, Z., & Kiss-Toth, E. (2019). TMEM203 is a binding partner and regulator of STING-mediated inflammatory signaling in macrophages. *Proc Natl Acad Sci U S A, 116(33), 16479-16488.*

Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15(12), 550.*

Martinez-Monleon, A., Kryh Öberg, H., Gaarder, J., Berbegall, A., Javanmardi, N., Djos, A., Ussowicz, M., Taschner-Mandl, S., Ambros, I., Øra, I., Sandstedt, B., Beiske, K., Ladenstein, R., Noguera, R., Ambros, P., Gordon Murkes, L., Ljungman, G., Kogner, P., Fransson, S., & Martinsson, T. (2022). Amplification of CDK4 and MDM2: a detailed study of a high-risk neuroblastoma subgroup. *Sci Rep, 12(1), 12420.*

McCarthy, D., & Smyth, G. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics, 25(6), 765-71.*

Ribeiro, J., Lovasco, L., Vanderhyden, B., & Freiman, R. (2014). Targeting TBP-Associated Factors in Ovarian Cancer. *Front Oncol, 4, 45.*

Romano, J., Kromrey, J., Coraggio, J., & Skowronek, J.. (2006). Appropriate statistics for ordinal level data: should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics, 22(12), 1540-2.*

Thorsson, V., Gibbs, D., Brown, S., Wolf, D., Bortone, D., Ou Yang, T., Porta-Pardo, E., Gao, G., Plaisier, C., Eddy, J., Ziv, E., Culhane, A., Paull, E., Sivakumar, I., Gentles, A., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J., Mose, L., Vo, N., Liu, J., Liu, Y., Rader, J., Dhankani, V., Reynolds, S., Bowlby, R., Califano, A., Cherniack, A., Anastassiou, D., Bedognetti, D., Mokrab, Y., Newman, A., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T., Noushmehr, H., Pedamallu, C., Bullman, S., Ojesina, A., Lamb, A., Zhou, W., Shen, H., Choueiri, T., Weinstein, J., Guinney, J., Saltz, J., Holt, R., Rabkin, C., Lazar, A., Serody, J., Demicco, E., Disis, M., Vincent, B., Shmulevich, I., & Network, C. (2018). The Immune Landscape of Cancer. *Immunity, 48(4), 812-830.e14.*

Wang, J., Zhou, Y., Li, D., Sun, X., Deng, Y., & Zhao, Q. (2017). TSPAN31 is a critical regulator on transduction of survival and apoptotic signals in hepatocellular carcinoma cells. *FEBS Lett, 591(18), 2905-2918.*

Wang, Q., Xu, T., Tong, Y., Wu, J., Zhu, W., Lu, Z., & Ying, J. (2019). Prognostic Potential of Alternative Splicing Markers in Endometrial Cancer. *Mol Ther Nucleic Acids, 18, 1039-1048.*

Wu, T., Wu, C., Zhao, X., Wang, G., Ning, W., Tao, Z., Chen, F., & Liu, X. (2022). Extrachromosomal DNA formation enables tumor immune escape potentially through regulating antigen presentation gene expression. *Sci Rep, 12(1), 3590.*

Zhu, A., Ibrahim, J., & Love, M. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics, 35(12), 2084-2092.*

Additional references for hand-curated DSB repair pathways

Aceytuno, R., Piett, C., Havali-Shahriari, Z., Edwards, R., Rey, M., Ye, R., Javed, F., Fang, S., Mani, R., Weinfeld, M., Hammel, M., Tainer, J., Schriemer, D., Lees-Miller, S., & Glover, J. (2017). Structural and functional characterization of the PNKP-XRCC4-LigIV DNA repair complex. *Nucleic Acids Res, 45(10), 6238-6251.*

Adamovich, A., Banerjee, T., Wingo, M., Duncan, K., Ning, J., Martins Rodrigues, F., Huang, K., Lee, C., Chen, F., Ding, L., & Parvin, J. (2019). Functional analysis of BARD1 missense variants in homology-directed repair and damage sensitivity. *PLoS Genet, 15(3), e1008049.*

Aparicio, T., Baer, R., & Gautier, J. (2014). DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst), 19, 169-75.*

Belotserkovskaya, R., Raga Gil, E., Lawrence, N., Butler, R., Clifford, G., Wilson, M., & Jackson, S. (2020). PALB2 chromatin recruitment restores homologous recombination in BRCA1-deficient cells depleted of 53BP1. *Nat Commun, 11(1), 819.*

Benitez, A., Liu, W., Palovcak, A., Wang, G., Moon, J., An, K., Kim, A., Zheng, K., Zhang, Y., Bai, F., Mazin, A., Pei, X., & Yuan, F. (2018). FANCA Promotes DNA Double-Strand Break Repair by Catalyzing Single-Strand Annealing and Strand Exchange. *Mol Cell, 71(4), 621-628.e4.*

Bian, L., Meng, Y., Zhang, M., & Li, D. (2019). MRE11-RAD50-NBS1 complex alterations and DNA damage response: implications for cancer treatment. *Mol Cancer, 18(1), 169.*

Blackford, A., & Jackson, S. (2017). ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Mol Cell, 66(6), 801-817.*

Boersma, V., Moatti, N., Segura-Bayona, S., Peuscher, M., Torre, J., Wevers, B., Orthwein, A., Durocher, D., & Jacobs, J. (2015). MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5' end resection. *Nature, 521(7553), 537-540.*

Ceccaldi, R., Rondinelli, B., & D'Andrea, A. (2016). Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol, 26(1), 52-64.*

Chapman, J. R., Barral , P., Vannier , J., Borel, V., Steger, M., Tomas-Loba , A., Sartori, A. A., Adams, I.R., Batista, F.D., Boulton, S.J. (2013). RIF1 Is Essential for 53BP1-Dependent Nonhomologous End Joining and Suppression of DNA Double-Strand Break Resection. *Mol Cell.* 49(5): 858–871.

Clairmont, C., Sarangi, P., Ponnienselvan, K., Galli, L., Csete, I., Moreau, L., Adelmant, G., Chowdhury, D., Marto, J., & D'Andrea, A. (2020). TRIP13 regulates DNA repair pathway choice through REV7 conformational change. *Nat Cell Biol, 22(1), 87-96.*

Costelloe, T., Louge, R., Tomimatsu, N., Mukherjee, B., Martini, E., Khadaroo, B., Dubois, K., Wiegant, W., Thierry, A., Burma, S., Attikum, H., & Llorente, B. (2012). The yeast Fun30 and human SMARCAD1 chromatin remodellers promote DNA end resection. *Nature, 489(7417), 581-4.*

Daley, J., Tomimatsu, N., Hooks, G., Wang, W., Miller, A., Xue, X., Nguyen, K., Kaur, H., Williamson, E., Mukherjee, B., Hromas, R., Burma, S., & Sung, P. (2020). Specificity of end resection pathways for double-strand break regions containing ribonucleotides and base lesions. *Nat Commun, 11(1), 3088.*

de Krijger, I., Föhr B., Pérez S.H., Vincendeau E., Serrat J., Thouin A.M., Susvirkar V., Lescale C., Paniagua I., Hoekman L., Kaur S., Altelaar M., Deriano L., Faesen A.C., Jacobs J.J.L. (2021). MAD2L2 dimerization and TRIP13 control shieldin activity in DNA repair. *Nat Commun, 12(1), 5421.*

El Ghamrasni, S., Cardoso, R., Halaby, M., Zeegers, D., Harding, S., Kumareswaran, R., Yavorska, T., Chami, N., Jurisicova, A., Sanchez, O., Hande, M., Bristow, R., Hakem, R., & Hakem, A. (2015). Cooperation of Blm and Mus81 in development, fertility, genomic integrity and cancer suppression. *Oncogene, 34(14), 1780-9.*

Elbakry, A., Juhász, S., Chan, K., & Löbrich, M. (2021). ATRX and RECQ5 define distinct homologous recombination subpathways. *Proc Natl Acad Sci U S A, 118(3).*

Feng, L., & Chen, J. (2012). The E3 ligase RNF8 regulates KU80 removal and NHEJ repair. *Nat Struct Mol Biol, 19(2), 201-6.*

Feng, Y., Xiang, J., Liu, S., Guo, T., Yan, G., Feng, Y., Kong, N., Li, H., Huang, Y., Lin, H., Cai, X., & Xie, A. (2017). H2AX facilitates classical non-homologous end joining at the expense of limited nucleotide loss at repair junctions. *Nucleic Acids Res, 45(18), 10614-10633.*

Grundy, G., Rulten, S., Zeng, Z., Arribas-Bosacoma, R., Iles, N., Manley, K., Oliver, A., & Caldecott, K. (2013). APLF promotes the assembly and activity of non-homologous end joining protein complexes. *EMBO J, 32(1), 112-25.*

Huang, R., & Zhou, P. (2020). DNA damage response signaling pathways and targets for radiotherapy sensitization in cancer. *Signal Transduct Target Ther, 5(1), 60.*

Iliakis, G., Murmann, T., & Soni, A. (2015). Alternative end-joining repair pathways are the ultimate backup for abrogated classical non-homologous end-joining and homologous recombination repair: Implications for the formation of chromosome translocations. *Mutat Res Genet Toxicol Environ Mutagen, 793, 166-75.*

Juhász, S., Elbakry, A., Mathes, A., & Löbrich, M. (2018). ATRX Promotes DNA Repair Synthesis and Sister Chromatid Exchange during Homologous Recombination. *Mol Cell, 71(1), 11-24.e7.*

Knijnenburg, T., Wang, L., Zimmermann, M., Chambwe, N., Gao, G., Cherniack, A., Fan, H., Shen, H., Way, G., Greene, C., Liu, Y., Akbani, R., Feng, B., Donehower, L., Miller, C., Shen, Y., Karimi, M., Chen, H., Kim, P., Jia, P., Shinbrot, E., Zhang, S., Liu, J., Hu, H., Bailey, M., Yau, C., Wolf, D., Zhao, Z., Weinstein, J., Li, L., Ding, L., Mills, G., Laird, P., Wheeler, D., Shmulevich, I., Monnat, R., Xiao, Y., Wang, C., & Network, C. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep, 23(1), 239-254.e6.*

Lu, G., Duan, J., Shu, S., Wang, X., Gao, L., Guo, J., & Zhang, Y. (2016). Ligase I and ligase III mediate the DNA double-strand break ligation in alternative end-joining. *Proc Natl Acad Sci U S A, 113(5), 1256-60.*

Pond, K., Renty, C., Yagle, M., & Ellis, N. (2019). Rescue of collapsed replication forks is dependent on NSMCE2 to prevent mitotic DNA damage. *PLoS Genet, 15(2), e1007942.*

Pryor, J., Conlin, M., Carvajal-Garcia, J., Luedeman, M., Luthman, A., Small, G., & Ramsden, D. (2018). Ribonucleotide incorporation enables repair of chromosome breaks by nonhomologous end joining. *Science, 361(6407), 1126-1129.*

Rulten, S., Fisher, A., Robert, I., Zuma, M., Rouleau, M., Ju, L., Poirier, G., Reina-San-Martin, B., & Caldecott, K. (2011). PARP-3 and APLF function together to accelerate nonhomologous end-joining. *Mol Cell, 41(1), 33-45.*

Saha, J., & Davis, A. (2016). Unsolved mystery: the role of BRCA1 in DNA end-joining. *J Radiat Res, 57 Suppl 1, i18-i24.*

Sharma, S., Javadekar, S., Pandey, M., Srivastava, M., Kumari, R., & Raghavan, S. (2015). Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell Death Dis, 6, e1697.*

Strande, N., Waters, C., & Ramsden, D. (2012). Resolution of complex ends by Nonhomologous end joining better to be lucky than good? *Genome Integr, 3(1), 10.*

Tang, M., Feng, X., Pei, G., Srivastava, M., Wang, C., Chen, Z., Li, S., Zhang, H., Zhao, Z., Li, X., & Chen, J. (2020). FOXK1 Participates in DNA Damage Response by Controlling 53BP1 Function. *Cell Rep, 32(6), 108018.*

Wang, B., Matsuoka, S., Ballif, B., Zhang, D., Smogorzewska, A., Gygi, S., & Elledge, S. (2007). Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. *Science, 316(5828), 1194-8.*

Wyatt, H., Sarbajna, S., Matos, J., & West, S. (2013). Coordinated actions of SLX1-SLX4 and MUS81-EME1 for Holliday junction resolution in human cells. *Mol Cell, 52(2), 234-47.*

Appendix

Chapter 1 Supplemental Tables and Figures

Table S1.1: Augustus predicted genes and transcripts. Augustus transcript predictions, w/ genomic coordinates of predicted CDS, identified event peptides, identified RefSeq and Ensembl peptides.

Table S1.2: Augustus predicted transcript CDS overlap with Ensembl or RefSeq CDS, 5'UTR, and/or 3'UTR.

Table S1.3: PSK normalized proteogenomics expression matrix of multi-tissue Atlantic salmon proteins.

Table S1.4: Salmon protein clusters in multi-tissue dataset.

Table S1.5: PSK normalized proteogenomics expression matrix of Atlantic salmon liver and primary hepatocyte proteins.

Table S1.6: Salmon protein clusters in liver tissue and primary hepatocyte dataset.

Table S1.7: Chemical defensome gene patterns from (Eide et al., 2021).

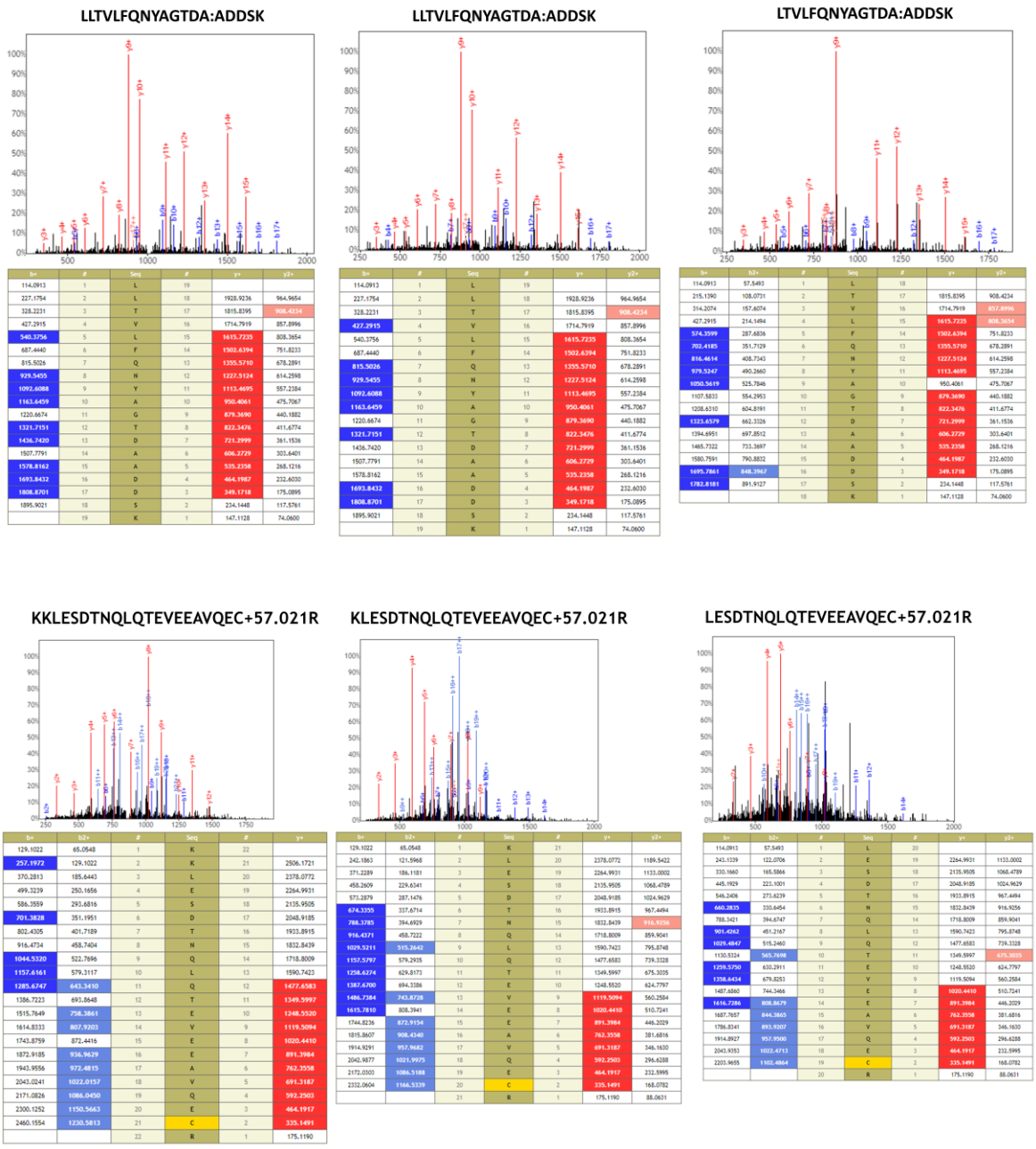
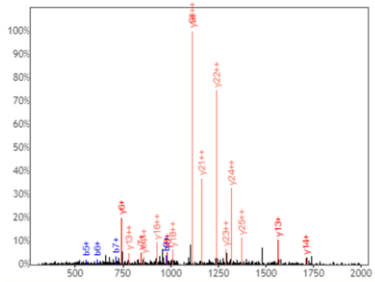


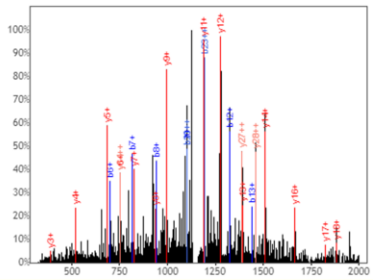
Figure S1.1: Peptide-spectrum matches with b-ion and y-ion matches (+1,+2) for proteogenomic events (L)LTVLFQNYAGTDA:ADDSK and (K)(K)LESDTNQLQTEVEEAVQEC+57.021R.

LDNNTGVYVPTGLAAFPN+0.984ELMHVPQSWAR



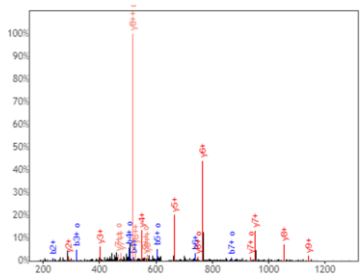
b+	#	Seq	#	y+	y2+
114.0913	1	L	29		
229.1183	2	D	28	3085.4836	1543.2454
343.1612	3	N	27	2970.4567	1485.7320
457.2041	4	N	26	2856.4137	1428.7105
558.2518	5	T	25	2742.3708	1371.6890
615.2733	6	G	24	2641.3231	1321.1652
714.3417	7	V	23	2584.3017	1292.6545
877.4050	8	Y	22	2485.2333	1243.1203
976.4734	9	V	21	2322.1699	1161.5886
1073.5262	10	P	20	2223.1015	1112.0544
1174.5739	11	T	19	2126.0488	1063.5280
1231.5953	12	G	18	2025.0011	1013.0042
1344.6794	13	L	17	1967.9796	984.4934
1415.7165	14	A	16	1854.8955	927.9514
1486.7536	15	A	15	1783.8584	892.4329
1633.8220	16	F	14	1712.8213	856.9143
1730.8748	17	P	13	1565.7529	783.3801
1845.9017	18	N	12	1468.7001	734.8537
1974.9443	19	E	11	1353.6732	677.3402
2088.0284	20	L	10	1224.6306	612.8189
2219.0689	21	M	9	1111.5466	556.2769
2356.1278	22	H	8	980.5061	490.7567
2455.1962	23	V	7	843.4472	422.2272
2552.2490	24	P	6	744.3787	372.6930
2680.3075	25	Q	5	647.3260	324.1666
2767.3396	26	S	4	519.2674	260.1373
2953.4189	27	W	3	432.2354	216.6213
3024.4560	28	A	2	246.1561	123.5817
	29	R	1	175.1190	88.0631

ETGYLHIQGTKPDSAGC+57.021GVNDSPVGLAAYMLEK



b+	b2+	#	Seq	#	y+	y2+
130.0499	65.5286	1	E	33		
231.0975	116.0524	2	T	32	3349.6187	1675.3130
288.1190	144.5631	3	G	31	3248.5710	1624.7891
451.1823	226.0948	4	Y	30	3191.5495	1596.2784
564.2664	282.6368	5	L	29	3028.4862	1514.7467
701.3253	351.1663	6	H	28	2915.4021	1458.2047
814.4094	407.7083	7	I	27	2778.3432	1389.6753
942.4680	471.7376	8	Q	26	2665.2592	1333.1332
999.4894	500.2483	9	G	25	2537.2006	1269.1039
1100.5371	550.7722	10	T	24	2480.1791	1240.5932
1228.6321	614.8197	11	K	23	2379.1314	1190.0694
1325.6848	663.3461	12	P	22	2251.0365	1126.0219
1440.7118	720.8595	13	D	21	2153.9837	1077.4955
1527.7438	764.3755	14	S	20	2038.9568	1019.9820
1598.7809	799.8941	15	A	19	1951.9247	976.4660
1655.8024	828.4048	16	G	18	1880.8876	940.9475
1815.8326	908.4199	17	C	17	1823.8662	912.4367
1872.8540	936.9306	18	G	16	1663.8360	832.4216
1971.9224	986.4649	19	V	15	1606.8145	803.9109
2085.9654	1043.4863	20	N	14	1507.7461	754.3767
2200.9923	1100.9998	21	D	13	1393.7032	697.3552
2288.0243	1144.5158	22	S	12	1278.6762	639.8418
2385.0771	1193.0422	23	P	11	1191.6442	596.3257
2484.1455	1242.5764	24	V	10	1094.5914	547.7994
2541.1670	1271.0871	25	G	9	995.5230	498.2652
2654.2510	1327.6292	26	L	8	938.5016	469.7544
2725.2882	1363.1477	27	A	7	825.4175	413.2124
2796.3253	1398.6663	28	A	6	754.3804	377.6938
2959.3886	1480.1979	29	Y	5	683.3433	342.1753
3090.4291	1545.7182	30	M	4	520.2799	260.6436
3203.5131	1602.2602	31	L	3	389.2395	195.1234
3332.5557	1666.7815	32	E	2	276.1554	138.5813
		33	K	1	147.1128	74.0600

FSTWTDFNNR



b+	#	Seq	#	y+	y2+
148.0757	1	F	10		
235.1077	2	S	9	1140.5069	570.7571
336.1554	3	T	8	1053.4748	527.2411
522.2347	4	W	7	952.4272	476.7172
623.2824	5	T	6	766.3478	383.6776
738.3093	6	D	5	665.3002	333.1537
885.3777	7	F	4	550.2732	275.6402
999.4207	8	N	3	403.2048	202.1060
1113.4636	9	N	2	289.1619	145.0846
	10	R	1	175.1190	88.0631

Figure S1.2: Peptide-spectrum matches with b-ion and y-ion matches (+1,+2) for proteogenomic events LDNNTGVYVPTGLAAFPN+0.984ELMHVPQSWAR, ETGYLHIQGTKPDSAGC+57.021GVNDSPVGLAAYMLEK, and FSTWTDFNNR.

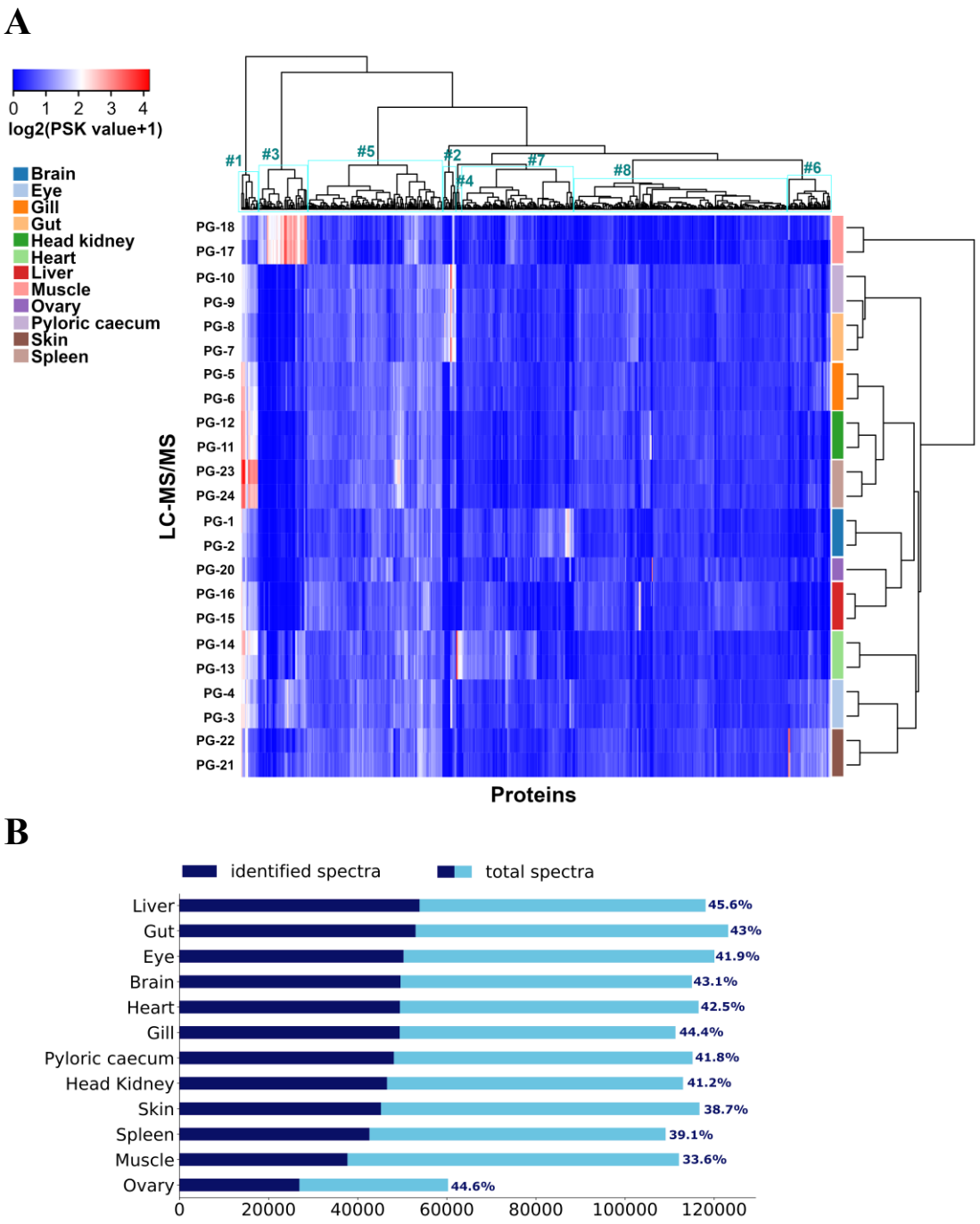


Figure S1.3: (A) Hierarchical clustering of the PSK normalized expression matrix (23 spectrum files, top 600 proteins) in multi-tissue dataset. (**B)** Spectra identified in each of twelve tissues. Identification rates, shown as blue text, are defined as the number of identified spectra divided by the total number of spectra for each tissue type.

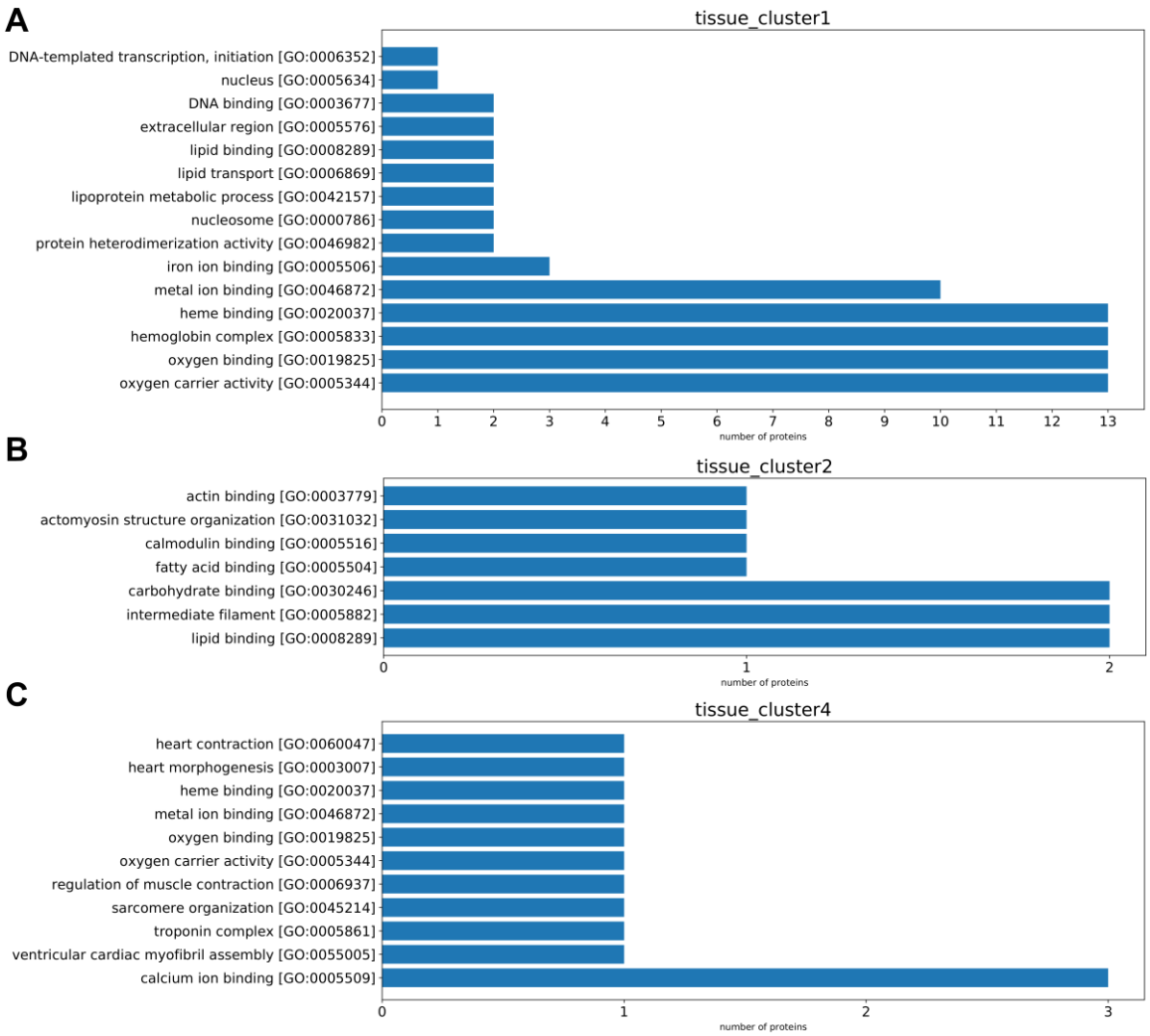


Figure S1.4: G.O. terms for proteins in multi-tissue dataset.

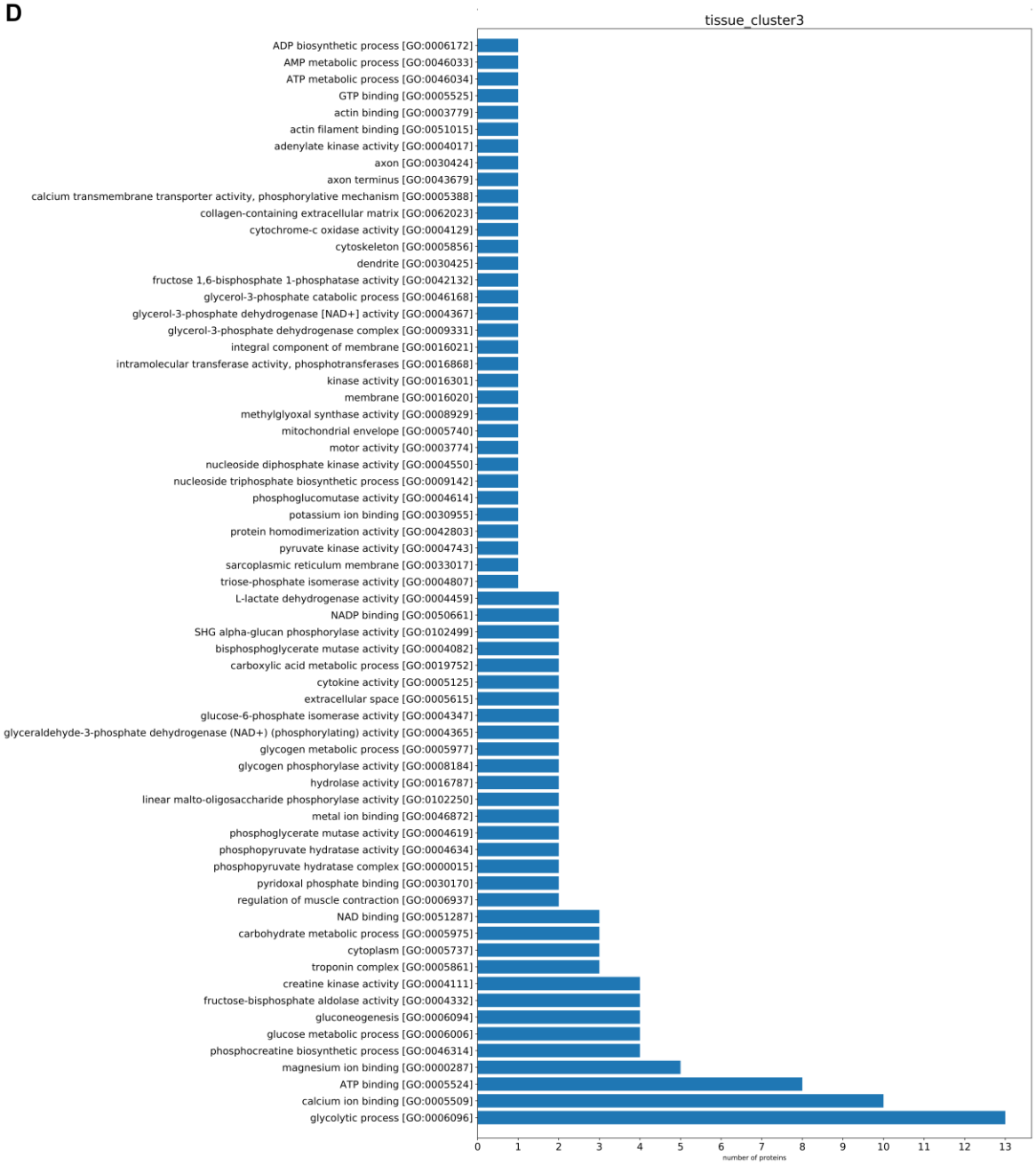


Figure S1.4: G.O. terms for proteins in multi-tissue dataset (Continued).

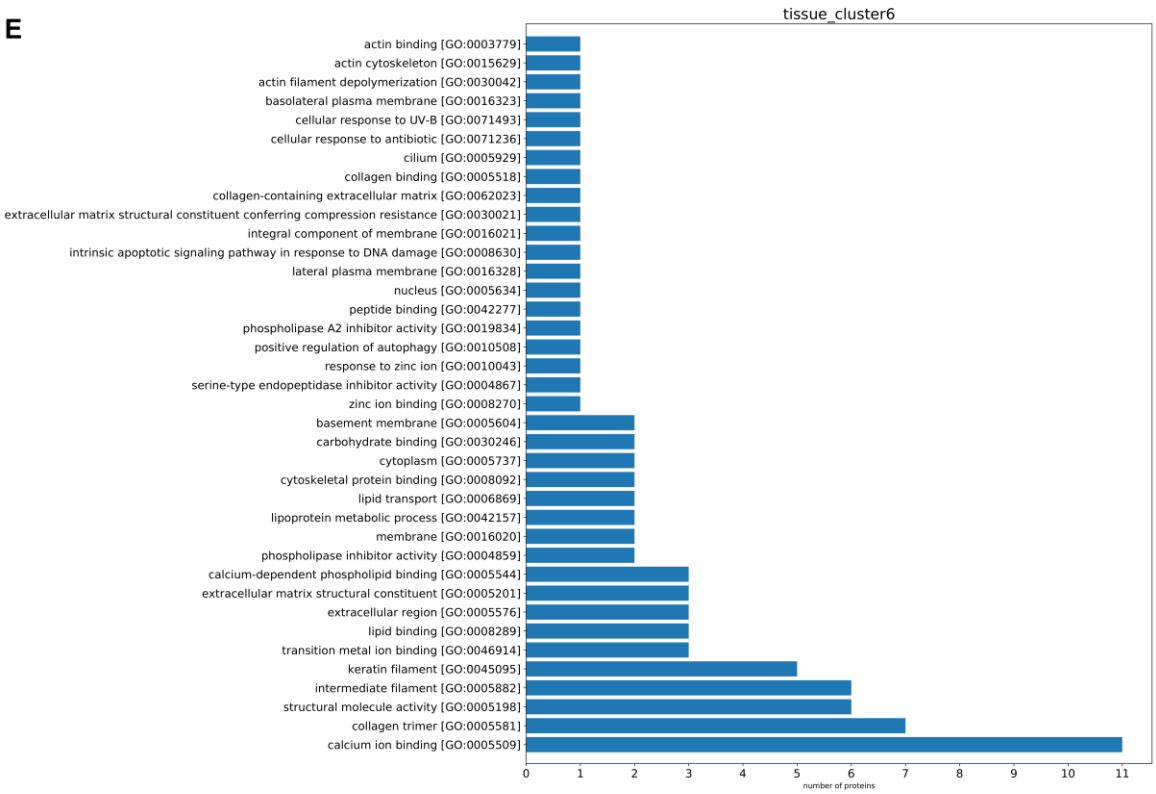


Figure S1.4: G.O. terms for proteins in multi-tissue dataset (Continued).

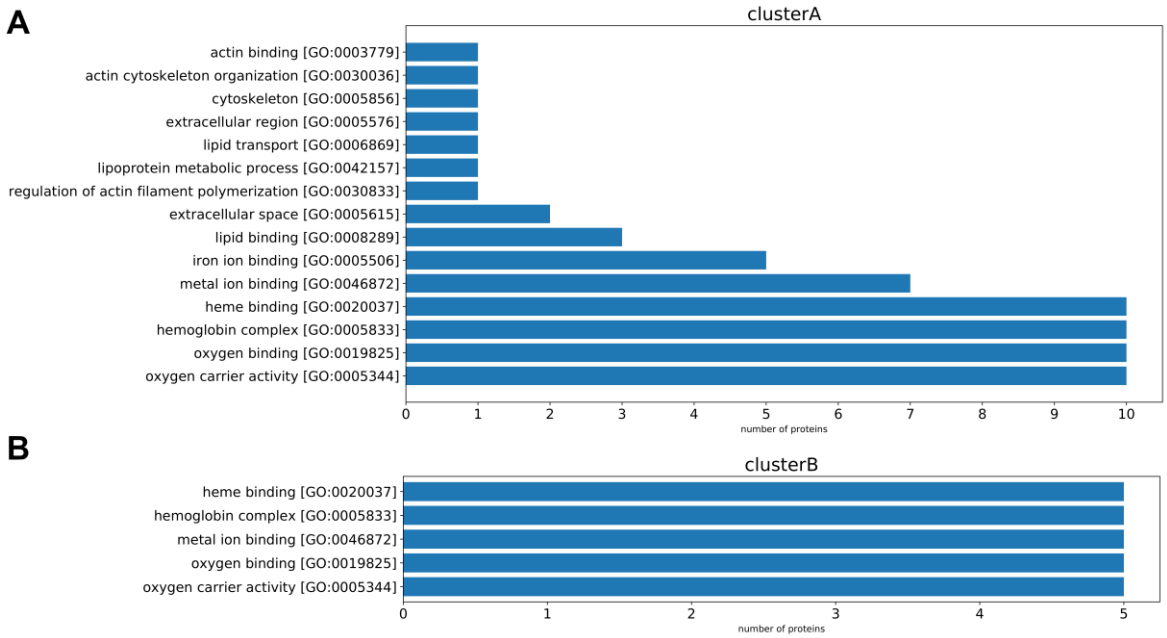


Figure S1.5: G.O. terms for proteins in liver tissue and primary hepatocyte dataset.

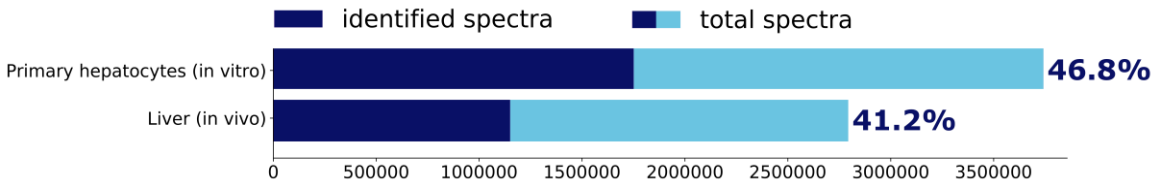


Figure S1.6: Spectra identified in primary hepatocyte and liver tissue. Identification rates are shown as blue text.

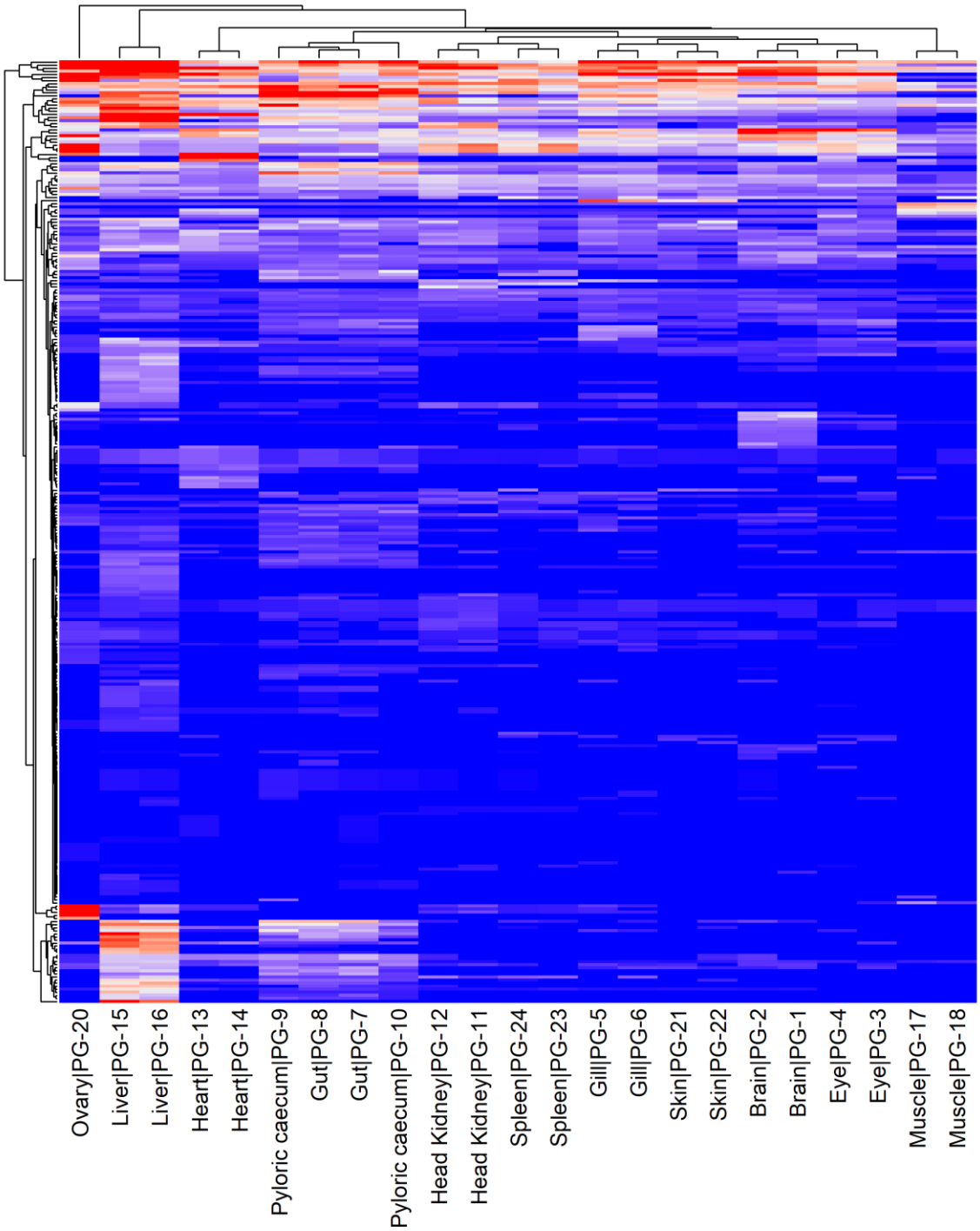


Fig S1.7: Hierarchical clustering of the PSK normalized expression matrix of Atlantic salmon chemical defensome proteins in multi-tissue dataset.

Chapter 2 Supplemental Tables and Figures

Table S2.1: Genera identified by ProteoStorm in 110 individuals suspected of UTI. Related to Figure 2.2. Genera are considered to be identified in an individual if greater than one high-confidence PSM maps uniquely to that genus. High-confidence PSM counts per individual are defined as the average across technical replicates of the number of PSMs passing a 1% FDR (per MS/MS experiment) that belong to a peptide passing a 1% peptide-level FDR (pooled across all MS/MS experiments). Bold text indicates genera reported by a previous study (Yu, et al., 2015). Genera with less than 1 count are not included in the table. Red text indicates genera not identified by the previous study that are identified by ProteoStorm in high abundance.

Table S2.2: Multiscale Bootstrap Resampling Probabilities and Errors. Related to Figure 2.2. Approximately unbiased (au) *p*-values computed by the Pvclust implementation of multiscale bootstrap resampling and reported standard errors (se) for hierarchical clustering of the 64 genera identified by ProteoStorm **(a)** in 110 UTI-suspected individuals and five healthy individuals **(b)** (pvclust; nboot=100,000, method.hclust = "ward.D2", and method.dist="bray).

Table S2.3: Genera identified by ProteoStorm in infant 23. Related to Methods. The previous study (Xiong, et al., 2017) reports species-based taxonomic compositions, and balances spectral counts between shared proteins. In contrast, ProteoStorm reports genera-based taxonomic compositions based on spectral counts from peptides that only map to a single genus. For comparison purposes, we parsed the protein groups and spectral counts file (40168_2017_290_MOESM2_ESM.xlsx) provided by the previous study for genus information, and chose the top 20 genera for each sample (run 1 or 2 for day of life 21, 34, 50). For ProteoStorm, we report the number of PSMs passing a 1% FDR (per MS/MS experiment) that belong to a peptide passing a 1% peptide-level FDR (pooled across all MS/MS experiments). Genera with less than 1 count are not included in the table.

Table S2.4: Variables and Definitions. Related to Methods. A list of variables and definitions used in this manuscript.

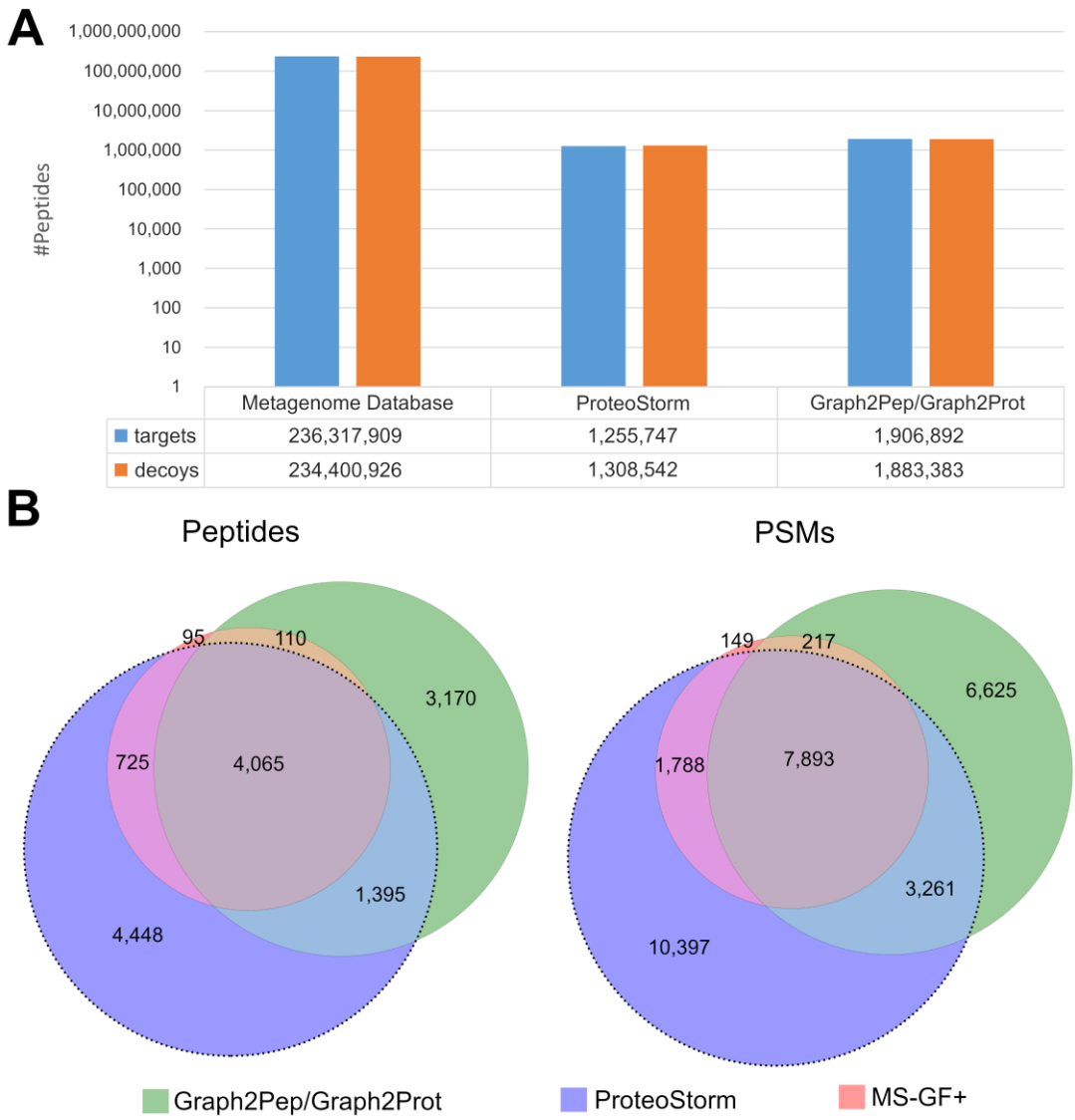


Figure S2.1: Wastewater dataset searched against a metagenome-guided database. Related to Methods. **(a)** When searching a wastewater dataset (150,216 spectra) against a metagenome-guided database (2.47Gb), the initial database of 470,718,835 peptides was reduced to 3,790,275 peptides for Graph2Pep/Graph2Prot (Tang, et al., 2016), which utilizes MS-GF+ for a fully-tryptic first stage search, and 2,564,289 peptides for ProteoStorm. **(b)** Peptides and PSMs identified by ProteoStorm, the Graph2Pep/Graph2Prot approach, and MS-GF+. ProteoStorm identified 10,633 peptides, or 95.9% (4,790) of the MS-GF+ peptides, in 52.6 minutes, while the Graph2Pep/Graph2Prot approach identified 8,740 peptides, or 83.6% (4,175) of the MS-GF+ peptides, in over 3.8 hours. The identification rate is 15.5% for ProteoStorm, 6.7% for MS-GF+, and 12.0% for the Graph2Pep/Graph2Prot approach. Out of 6,295 peptides identified by at least two of the three tools, ProteoStorm identified 98.3% (6,185), while MS-GF+ identified 77.8% (4,900) and the Graph2Pep/Graph2Prot approach identified 88.5% (5,570).

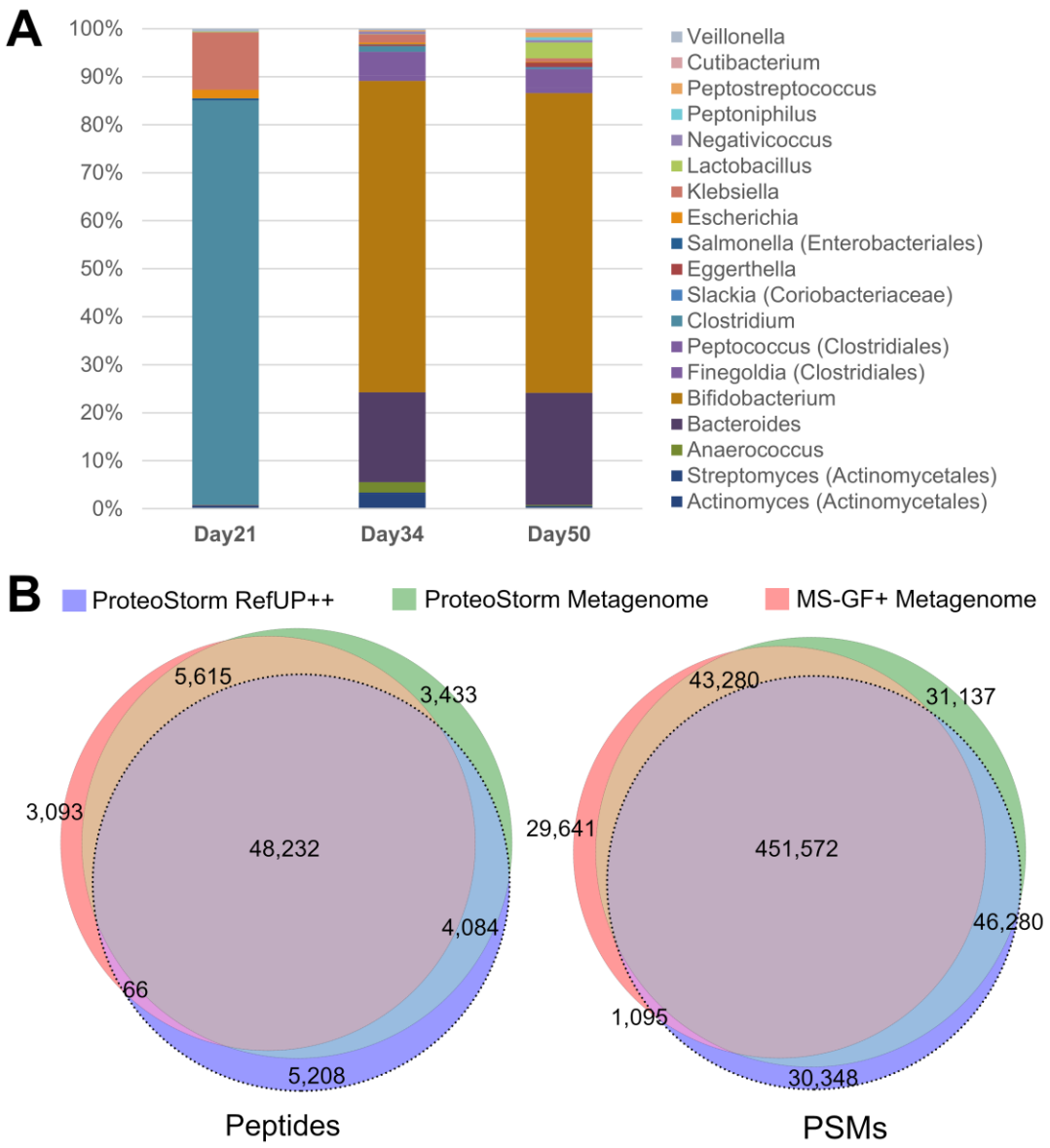


Figure S2.2: Human infant gut dataset. Related to Methods. **(a)** ProteoStorm identified 19 genera when searching a gut dataset (infant 23 at day of life 21, 34, and 50) against RefUp++. Out of 57,590 peptides, 54.9% (31,606) mapped uniquely to a single genus. Day 21 is dominated by *Clostridium* (28,546 PSMs), followed by *Klebsiella* and lesser abundances of *Escherichia*. There is a shift in day 34 to *Bifidobacterium* (33,985 PSMs) and *Bacteroides* (9,756 PSMs), followed by lesser abundances of *Peptococcus*. Day 50 is also dominated by *Bifidobacterium* (36,246 PSMs), with an increase in *Bacteroides* (13,472 PSMs) and *Lactobacillus* (1,894 PSMs). **(b)** Peptides and PSMs identified by ProteoStorm and MS-GF+ when searching against the RefUP++ database or a matched metagenome database. Against the metagenome database, ProteoStorm identified 61,364 peptides, and 94.5% (53,847) of MS-GF+ peptides at an identification rate of 24.2%. Against RefUP++, ProteoStorm identified 57,590 peptides, and 84.7% (48,298) of MS-GF+ peptides at an identification rate of 22.4%.

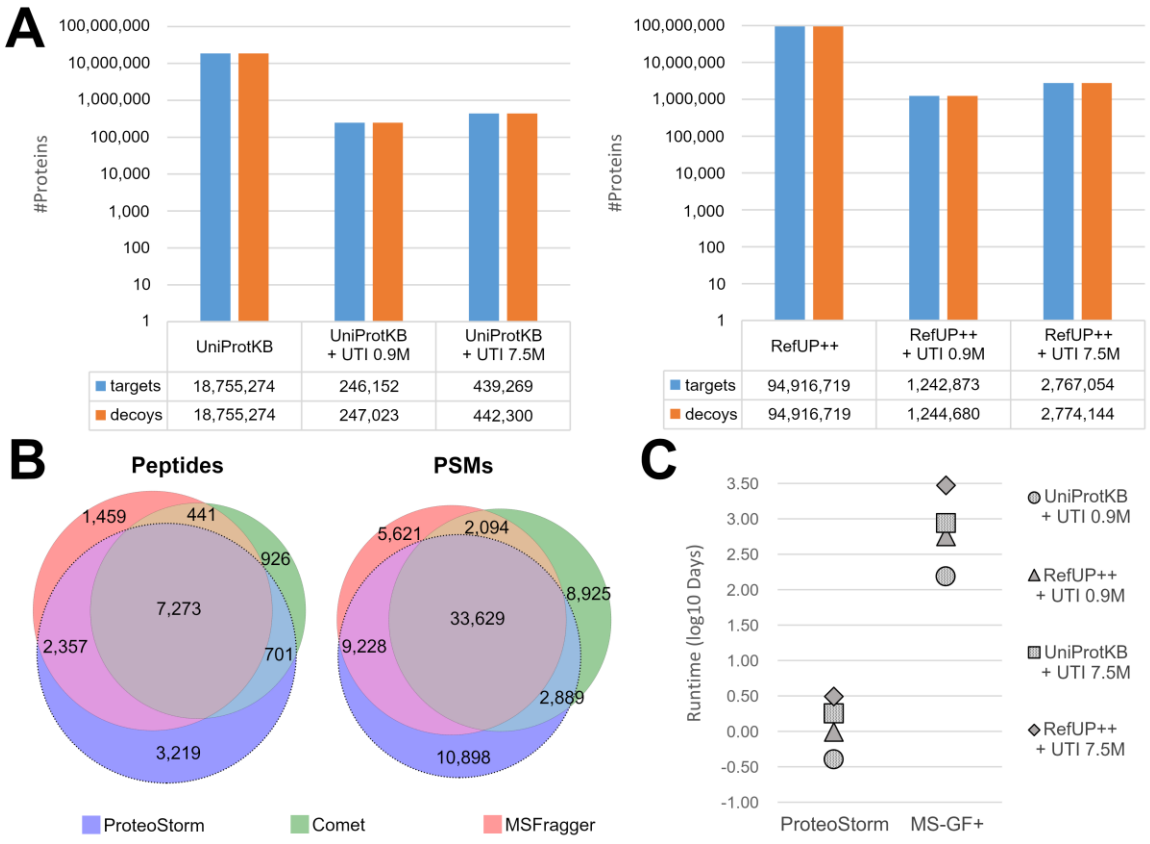


Figure S2.3: UTI dataset: search space reduction, identifications, and runtime. Related to Figure 1. **(a)** When searching the subset UTI dataset (0.9M spectra) and the full UTI dataset (7.5M) against the UniProtKB database, ProteoStorm reduced the initial database of 37,510,548 proteins to 493,175 proteins and 881,569 proteins in the refined protein database, respectively. When searching against the RefUP++ database, ProteoStorm reduced the initial database of 189,833,438 proteins to 2,487,553 proteins and 5,541,198 proteins in the refined protein database, respectively. **(b)** Searching the subset UTI dataset against the UniProtKB database, ProteoStorm identified 56,644 PSMs, Comet identified 47,537 PSMs, and MSFragger identified 50,572 PSMs. Of the 51,338 MS-GF+ PSMs, ProteoStorm identified 97% (49,776), while Comet and MSFragger identified 69.4% (35,640) and 79.9% (41,037), respectively. Of the 47,840 PSMs identified by two of the three tools, ProteoStorm identified 95.6% (45,746), while Comet and MSFragger identified 80.7% (38,612) and 94% (44,951), respectively. The identification rate was 5.98% for ProteoStorm, 5.02% for comet, 5.42% for MS-GF+, and 5.34% for MSFragger. At a 1% peptide-level FDR, ProteoStorm identified 13,550 peptides (p-value 1.56e-12), while Comet and MSFragger identified 9,341 (e-value 7.85e-05) and 11,530 (probability $\geq$ 0.9903) peptides respectively. **(c)** Increasing the number of spectra from 0.9 million to 7.5 million and the initial database size from 7.8Gb (18.8 million sequences) to 34.1Gb (94.9 million sequences), ProteoStorm speedup over MS-GF+ increased from 382x to 953x.

Chapter 3 Supplemental Tables and Figures

Table S3.1: AC049 classified 870 TCGA samples. Includes 14 tumors.

Table S3.2: Highly co-expressed clusters identified using Pvclust.

Table S3.3: Cluster #1509 and #1940 members.

Table S3.4: Cliff's delta matrix of 643 CorEx genes x TCGA and 11 tumor types with at least 10 ecDNA(+) and 10 ecDNA(-) samples.

Table S3.5: Top 643 |LFC| differentially expressed genes identified by DESeq2.

Table S3.6: GO biological process gene sets enriched by up-regulated genes from CorEx set.

Table S3.7: Grouped GO biological process gene sets enriched by up-regulated genes from the CorEx set.

Table S3.8: Cluster #1509 enriched GO biological process terms.

Table S3.9: Genes unique to grouped gene sets enriched by up-regulated CorEx genes.

Table S3.10: Genes found on amplicons in TCGA samples.

Table S3.11: Hand-curated genes for DNA damage repair pathways.

Table S3.12: AC049 classified 1,440 TCGA samples. Includes 24 tumors.

Table S3.13: GO biological process gene sets enriched by down-regulated genes from CorEx set.

Table S3.14: Grouped GO biological process gene sets enriched by down-regulated genes from the CorEx set.

Table S3.15: Up-/down-regulated genes based on 1,440 TCGA samples.

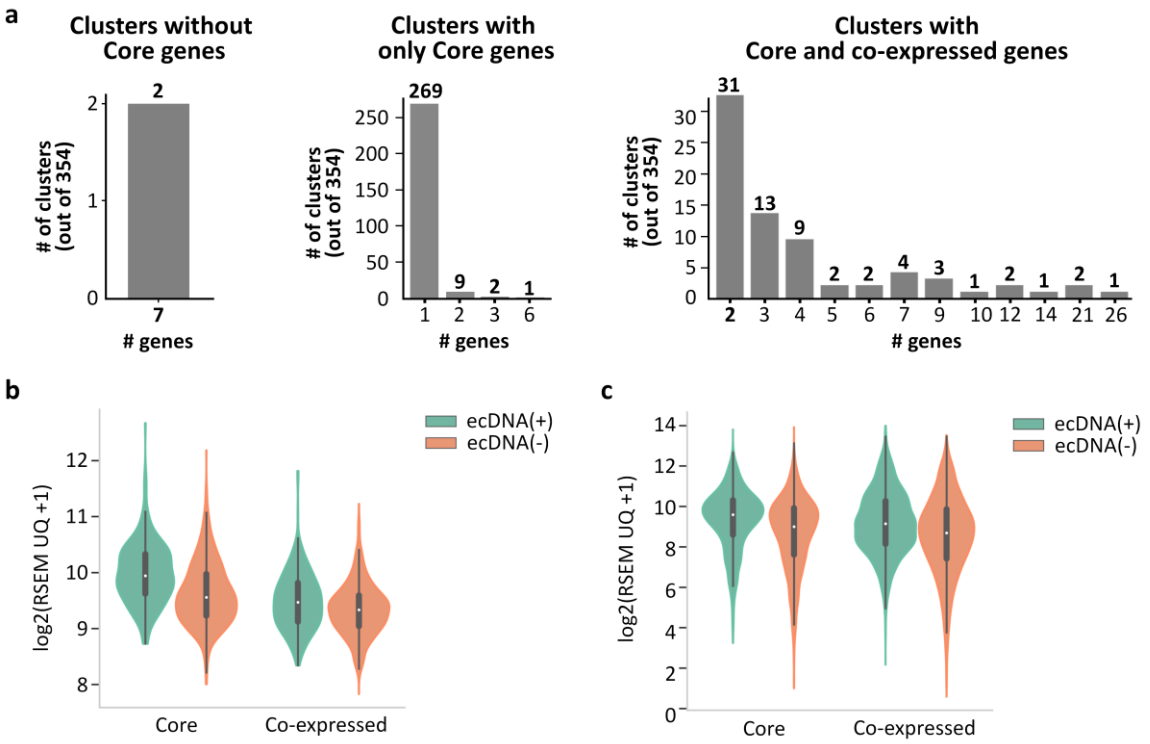


Figure S3.1: (a) Pvclust identified 354 clusters of high significance where members are selected in at least 10 Boruta runs as a cluster. Of the 354 clusters, 2 clusters did not contain any Core genes, 281 clusters only contained Core genes, and 71 clusters contained at least 1 Core and 1 co-expressed gene. **(b)** Cluster #1940 contains a Core gene, RAE1, and a co-expressed gene, CSTF1. **(c)** Cluster #1509 contains 9 Core genes and 12 co-expressed genes.

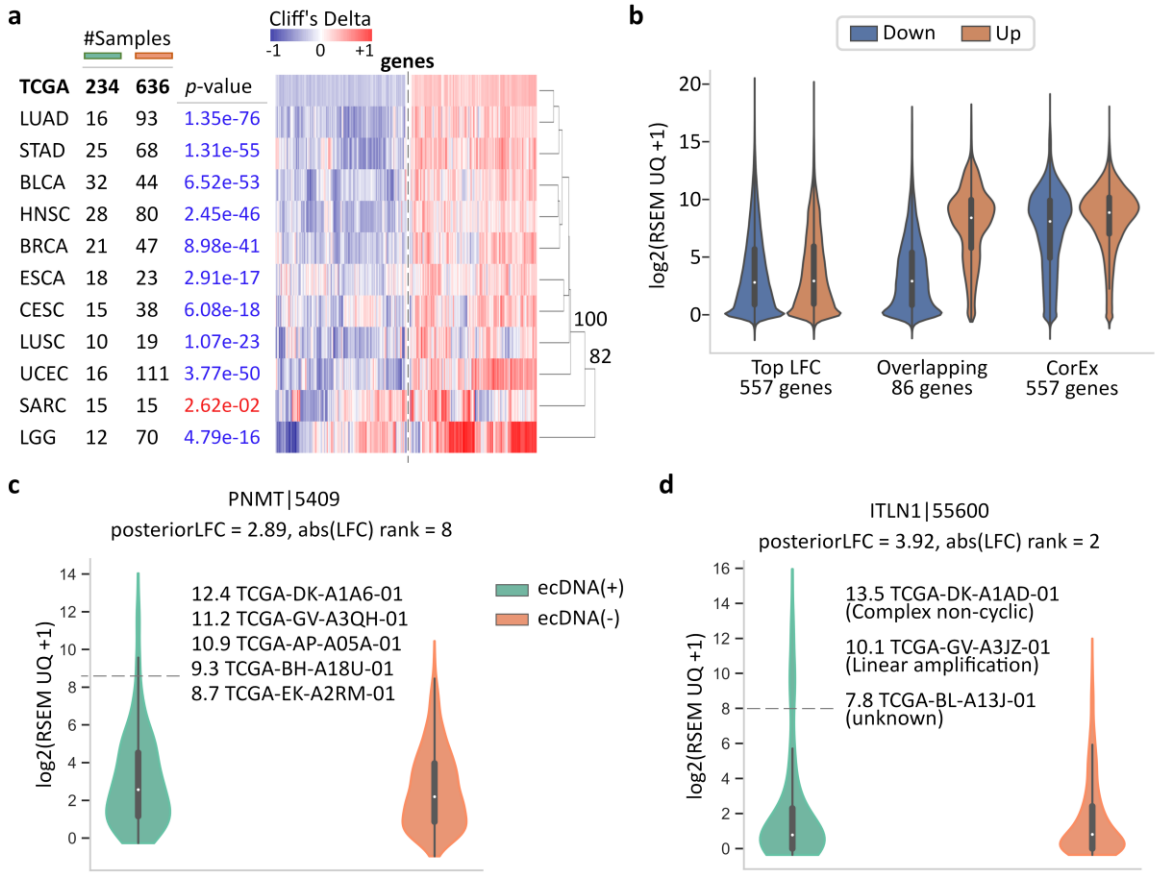


Figure S3.2: (a) Core genes are consistently up- or down-regulated in ecDNA(+) samples across tumor types, with the exception of SARC. **(b)** 86 CorEx genes are part of the Top 643 LFC DE gene set. **(c)** PNMT normalized gene expression in ecDNA(+) vs. ecDNA(-) samples. **(d)** ITLN1 normalized gene expression in ecDNA(+) vs. ecDNA(-) samples.

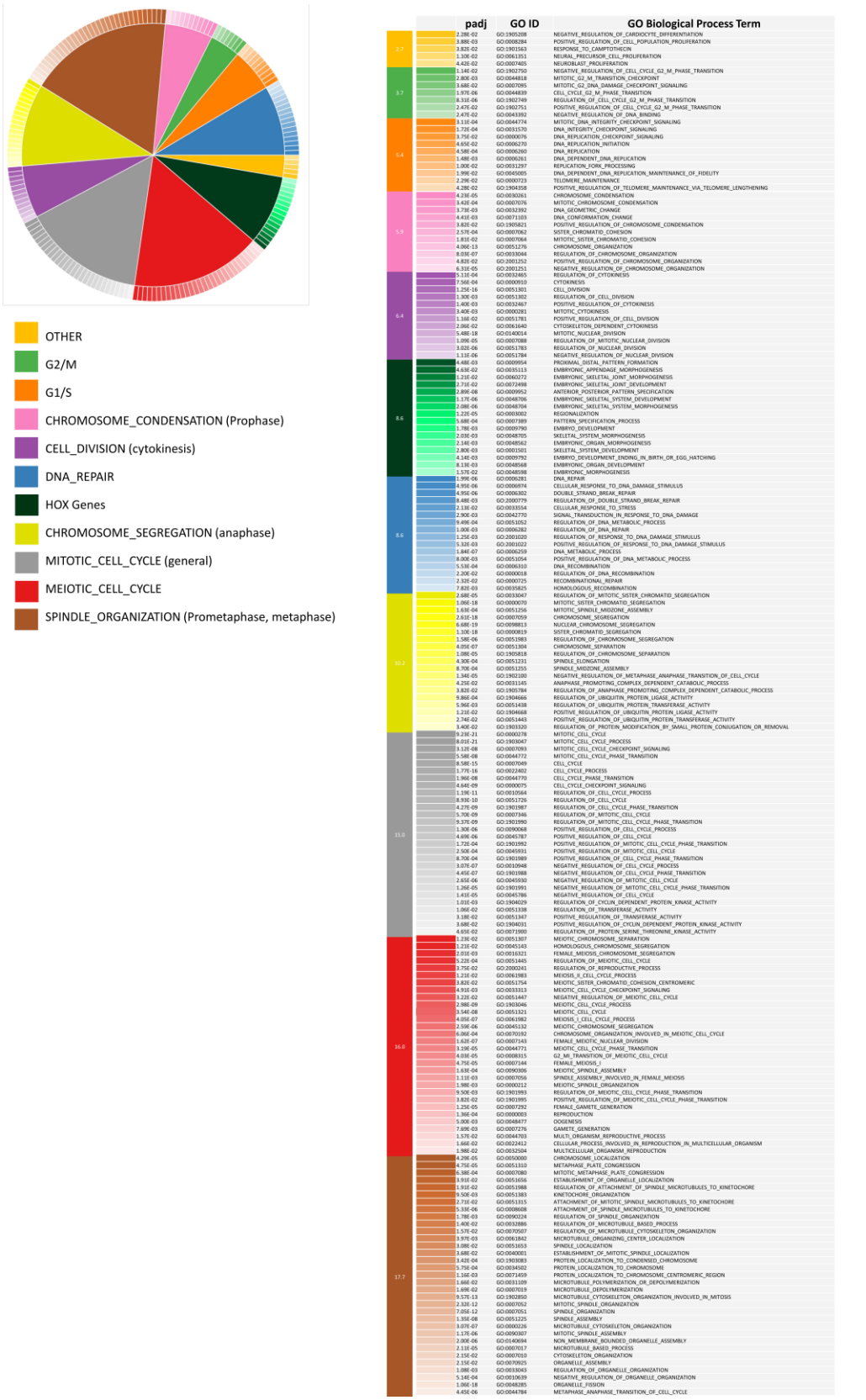


Figure S3.3: 187 enriched GOBP terms represented by 169 up-regulated CorEx genes.

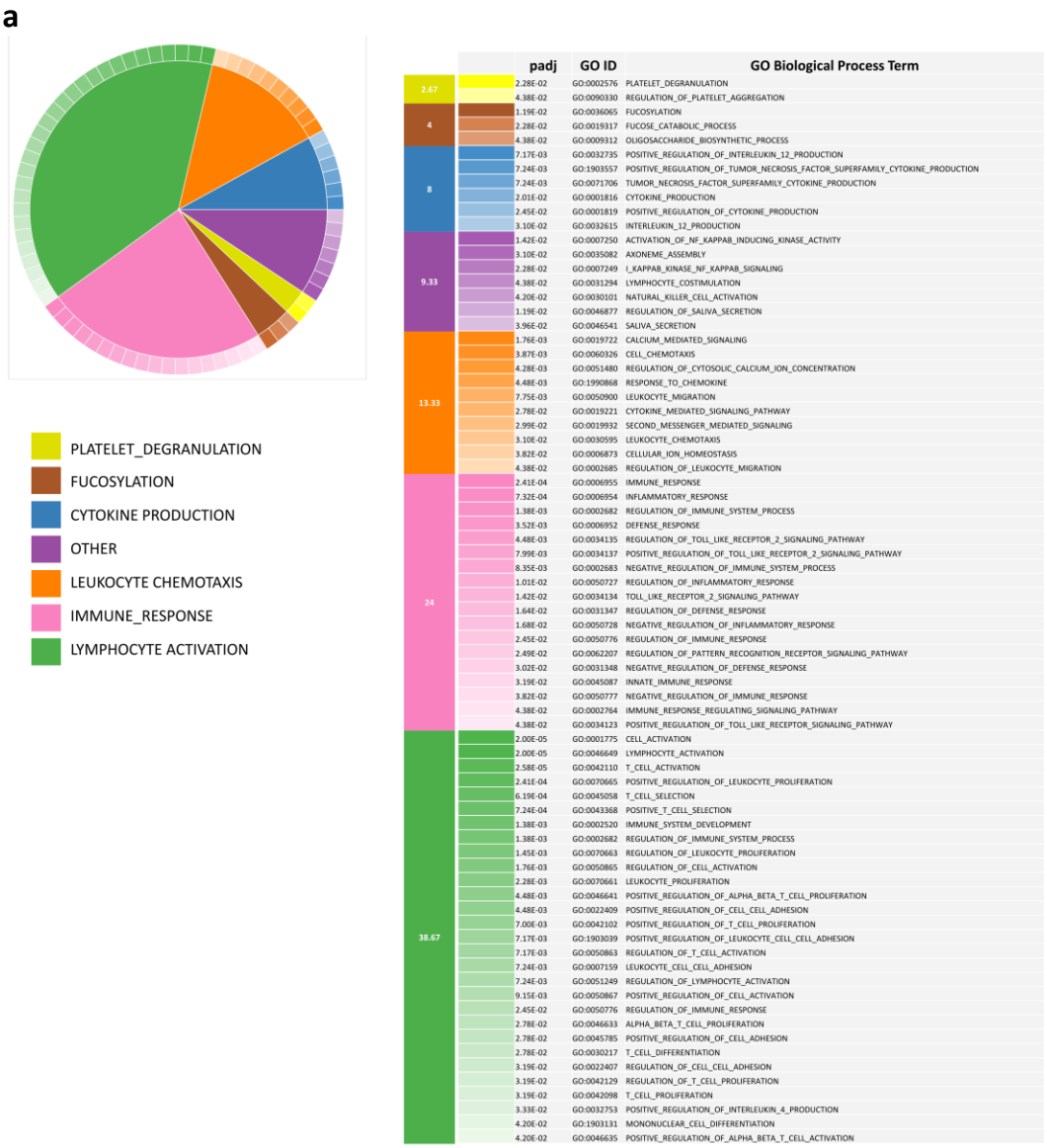


Figure S3.4: (a) 73 enriched GOBP terms represented by 119 down-regulated CorEx genes. **(b)** TME subtyping of TCGA samples classified as ecDNA(+) or ecDNA(-) (AC ver. 049) based on the classifications provided by Thorsson et al., 2018.