Genomics, Proteomics & Bioinformatics, 2024, **22(6)**, qzae083
https://doi.org/10.1093/gpbjnl/qzae083
Advance access publication: 22 November 2024
Method

ProtPipe: A Multifunctional Data Analysis Pipeline for Proteomics and Peptidomics

Ziyi Li [1,2,#], Cory A. Weller [1,2,#], Syed Shah [1,2], Nicholas L. Johnson [1,2], Ying Hao [1], Paige B. Jarreau [1], Jessica Roberts [1], Deyaan Guha [1], Colleen Bereda [1], Sydney Klaisner [1], Pedro Machado [3], Matteo Zanovello [3], Mercedes Prudencio [4,5], Björn Oskarsson [4,5], Nathan P. Staff [6], Dennis W. Dickson [4], Pietro Fratta [3], Leonard Petrucelli [4,5], Priyanka Narayan [7], Mark R. Cookson [1,8], Michael E. Ward [1,9], Andrew B. Singleton [1,8], Mike A. Nalls [1,2,*], Yue A. Qi [1,*]

[1]Center for Alzheimer's and Related Dementias (CARD), National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA
[2]DataTecnica LLC, Washington, DC 20812, USA
[3]UCL Queen Square Motor Neuron Disease Centre, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, University College London, London, WC1N 3BG, UK
[4]Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
[5]Neuroscience Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, FL 32224, USA
[6]Department of Neurology, Mayo Clinic, Rochester, MN 55905, USA
[7]Genetics and Biochemistry Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[8]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA
[9]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA

*Corresponding authors: mike@datatecnica.com (Nalls MA), andy.qi@nih.gov (Qi YA).
#Equal contribution.

Handling Editor: Xiaowen Liu

Abstract

Mass spectrometry (MS) is a technique widely employed for the identification and characterization of proteins, with personalized medicine, systems biology, and biomedical applications. The application of MS-based proteomics advances our understanding of protein function, cellular signaling, and complex biological systems. MS data analysis is a critical process that includes identifying and quantifying proteins and peptides and then exploring their biological functions in downstream analyses. To address the complexities associated with MS data analysis, we developed ProtPipe to streamline and automate the processing and analysis of high-throughput proteomics and peptidomics datasets with DIA-NN preinstalled. The pipeline facilitates data quality control, sample filtering, and normalization, ensuring robust and reliable downstream analyses. ProtPipe provides downstream analyses, including protein and peptide differential abundance identification, pathway enrichment analysis, protein–protein interaction analysis, and major histocompatibility complex (MHC)–peptide binding affinity analysis. ProtPipe generates annotated tables and visualizations by performing statistical post-processing and calculating fold changes between predefined pairwise conditions in an experimental design. It is an open-source, well-documented tool available at https://github.com/NIH-CARD/ProtPipe, with a user-friendly web interface.

Key words: ProtPipe; Data analysis pipeline; Mass spectrometry; Proteomics; Immunopeptidomics.

Introduction

Recent advancements in mass spectrometry (MS) have revolutionized the field of high-throughput proteomics, enabling accurate and rapid analysis of thousands of proteins across many biological samples [1–5]. This enhanced capability allows researchers to characterize the proteome on an unprecedented scale and investigate dynamic changes in protein expression under various experimental conditions [1,6–8]. MS-based proteomics has become an indispensable tool in protein quantification, offering advantages in sample preparation, dynamic range, and absolute quantification at the protein level [9]. Its applications span a wide range of research areas, including proteomics, personalized medicine, and biomarker discovery [9,10]. Beyond total proteome analysis, MS profiling of protein–protein interactions (PPIs) is crucial to understanding cellular processes and disease mechanisms. The integration of affinity purification with MS (AP-MS) has revolutionized the study of PPIs in high-throughput proteomics [11–14]. Additionally, major histocompatibility complex (MHC) immunopeptidomics, a technique that enables the comprehensive characterization of ligands present on the cell surface, has emerged as a powerful approach to enhance our understanding of antigen presentation, immune recognition, and vaccine development. Recent applications of quantitative immunopeptidomics have yielded significant insights into the identification of tumor-specific antigens, characterization of viral epitopes, and exploration of the impact of diseases on the immunopeptidome [15–19].

For proteomics data analysis, numerous software tools focus on MS database search and aid in the qualification of

Received: 5 January 2024; **Revised:** 11 October 2024; **Accepted:** 20 October 2024.
Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China 2024.

peptides and proteins. Available database search engines and integrated proteomics tools include PEAKS [20], DIA-NN [21], Trans-Proteomic Pipeline [22], Quantms [23], OpenMS [24], DIAproteomics [25], and LFQ-Analyst [26]. However, most of these resources and tools are not comprehensive or compatible with modern high-performance computing (HPC) (Table S1). While some database search algorithms offer downstream analysis and visualization capabilities, they primarily focus on specific aspects of MS data analysis with limited visualization options. Tools such as LFQ-Analyst [26], Perseus [27], and Proteome Discoverer do provide downstream analysis; however, users face significant challenges in applying these tools to large-scale datasets on different platforms (*e.g.*, cluster Windows, Linux, and Macs). Graphical user interface (GUI) systems are user-friendly, but high-throughput and large-scale datasets necessitate working on HPC systems. Ensuring compatibility and reproducibility across platforms can be complex, especially when dealing with dependencies and software environments. This lack of comprehensive integration of downstream functionalities highlights the necessity for innovative solutions in MS-based proteomics.

To address these limitations, our proteomics expert group at the Center for Alzheimer's and Related Dementias (CARD) has developed an automated pipeline, ProtPipe, for analyzing post-database search proteomics data. There is a growing need for automating and standardizing data analyses, and ProtPipe addresses these challenges by operating within a Singularity container, where all dependencies are pre-installed. This makes it an open-source software solution that is readily available to the scientific community, ensuring consistency and ease of use across different platforms. Post-analysis functions of other common database search engines are often not versatile enough to accommodate various applications in proteomics, such as AP-MS and immunopeptidomics. ProtPipe provides a unified platform that seamlessly integrates multiple downstream analysis tasks, including differential expression analysis, pathway analysis, PPI analysis, immuno-deconvolution analysis, and robust visualization. All these analyses are available within an easy-to-use environment, which makes it suitable for use on HPC systems as well as standard desktop environments to analyze high-throughput and large-scale datasets.

Method

Ease of use and reproducibility

ProtPipe facilitates reproducibility by operating within a Singularity container, where all necessary dependencies, including downstream analysis R packages, are pre-installed. While ProtPipe does not bundle DIA-NN [21] due to licensing requirements, it provides configuration support for users to easily integrate DIA-NN after installation. ProtPipe is fully capable of integrating and analyzing data outputs from various tools, such as DIA-NN [21], FragPipe [28], and Spectronaut. This approach ensures that while users have access to the complete workflow, they also comply with the software's licensing requirements. This fully open-source project is hosted at our GitHub repository (https://github.com/NIH-CARD/ProtPipe).

To improve accessibility and usability, we offer a web-based platform for downstream analysis. The ProtPipe web application (http://34.42.19.73:8501/) provides a user-friendly, interactive interface for performing differential expression analysis with a single click. It is dedicated to downstream analysis following database searches and does not manage peptide or protein quantification directly. Instead, it visualizes label-free quantitative proteomics datasets preprocessed by tools such as Spectronaut, DIA-NN [21], and FragPipe [28]. Users can upload their preprocessed data to the web application, which then performs the necessary statistical analyses and generates visualizations. Researchers with minimal bioinformatics skills can perform complex analyses and gain insights from their proteomics data. The integration of the web application with ProtPipe ensures a seamless workflow from raw data to meaningful results, enhancing the overall user experience.

Parametrization and command line options

ProtPipe is highly versatile, offering a variety of parameters that can be customized to suit specific instrumental and experimental settings. Users can tailor their analyses according to their unique requirements and experimental conditions. **Table 1** describes the ProtPipe post-analysis steps, while Tables S2 and S3 list the available parameters, along with their descriptions, default values, and examples. These parameters can be specified via the command line, ensuring that the analysis can be precisely controlled and reproducible.

Peptide and protein quantification

ProtPipe does not offer direct protein and peptide quantification due to licensing restrictions. However, users can install DIA-NN, and ProtPipe can run DIA-NN using a configuration file that contains specific arguments and options, enabling researchers to quickly estimate protein abundance. This configuration file can be edited to customize the settings for the database search. Various parameters that can be adjusted to tailor the database search according to the requirements, including specifying input/output files, the reference proteome database (FASTA file), threading, quality thresholds, enzymatic cleavage, modifications, and additional processing options.

Although ProtPipe does not package Spectronaut and FragPipe within the Docker container due to licensing restrictions, it fully supports the integration and analysis of data outputs from these tools. Users can easily import data from Spectronaut and FragPipe in .csv or table format into ProtPipe, allowing them to utilize the advanced data processing features of Spectronaut while benefiting from ProtPipe's extensive downstream analytical capabilities.

For database search parameters performed by Spectronaut, FragPipe [28], and DIA-NN [21], we recommend the library-free module (*i.e.*, direct-DIA), which does not require a preconstructed project-specific spectral library but only uses a proteome sequence file. We selected enzyme specificity (*e.g.*, trypsin) with up to two missed cleavages. Fixed modification was selected with carbamidomethylation of cysteine, while variable modifications were chosen for acetylation of protein N-termini and oxidized methionine. The false discovery rates (FDRs) of peptides and proteins were all set as 1%. We used the Swiss-Prot human proteome, including 20,435 reviewed proteins as the reference database.

Missing values and imputation

ProtPipe employs imputation to handle missing values, a critical technique in proteomics. Missing values in Data-Dependent Acquisition (DDA) commonly result from missing at random (MAR) due to the technical constraints [29]. In contrast, missing values in Data-Independent Acquisition

Table 1 ProtPipe post-analysis steps

Step	Sub-step	Algorithm
Data importation and formatting	Verify the presence of the required documents	Data.table
	Format the data	reshape2
Initial assessment and filtering	Determine the number of detectable protein groups	–
	Check whether the sample intensity distributions are consistent	–
	Evaluate the correlation among all sample pairs	corrplot
	Filter out proteins in low abundance, high variation, or inconsistent data patterns	–
Scaling and normalization	Select a suitable normalization procedure based on the biological background and data properties	–
Data cluster	Perform HC to identify protein groups with similar expression patterns	hccluster
	Conduct PCA to reduce data dimensionality	PCA
	Employ UMAP for low-dimensional data visualization	UMAP
DPE	Utilize the provided experimental design matrix for DPE	–
	Compute T-statistics and perform correction of multiple comparisons	*t*.test
	Create a volcano plot for visualizing DE analysis results	ggplot
	Perform GO pathway analysis	clusterProfile, org.Hs.eg.db
Heatmap	Visualize correlation of protein expression for all pairs of samples	pheatmap
	Optionally, create customized heatmap	–
PPI analysis	Utilize the "stringdb" R package to obtain interaction proteins	stringdb
	Identify candidates that overlap with the string database	ggplot
Immunopeptidome deconvolution	Calculate binding affinities using MHCflurry	MHCflurry
	Generate plots for the binding affinities of HLA allele	ggplot

Note: DPE, differential protein expression; HC, Hierarchical Clustering; PCA, Principal Component Analysis; UMAP, Uniform Manifold Approximation and Projection; DE, differential expression; GO, Gene Ontology; PPI, protein–protein interaction; HLA, human leukocyte antigen.

(DIA) are commonly due to the peptides' low signal intensity, which is below the detection threshold [30]. For DIA data, missing values are treated with zero imputation, assuming non-detectable or absent protein abundance.

Normalization

ProtPipe offers multiple methods for normalizing sample intensities to account for variability across samples and ensure consistent data analysis. Users can choose from three primary normalization techniques: "shift", "scale", or "none". The "shift" method adjusts the sample intensities by adding or subtracting a constant value, aligning the data to a central reference point such as the median. It operates on the premise that most proteins undergo minimal changes in their abundance across samples, and systemic biases impacting protein levels are uniformly distributed. By aligning protein abundance values to a central reference point, median normalization mitigates systematic discrepancies. However, it may not comprehensively capture specific variations in individual protein levels. The "scale" method involves multiplying the sample intensities by a constant factor, normalizing them relative to a reference distribution to adjust for differences in sample concentration. If no normalization is desired, users can select "none", which leaves the data unchanged. Each method provides flexibility in addressing different types of systematic biases and is chosen based on the specific needs of the dataset. ProtPipe utilizes median normalization to adjust protein abundance values at similar scale, ensuring robustness in addressing broad-scale biases.

Quality control and clustering

ProtPipe includes flexible quality control steps, with reasonable default settings, to evaluate data reliability and enhance the signal-to-noise ratio before conducting analyses. Quality control includes steps such as excluding proteins with globally low abundance and samples with low protein group counts. We signify low sample quality by the following criteria. Users can utilize the parameter '--minintensity' to specify a minimum linear intensity value. Samples with intensities below this threshold will be filtered out during the data analysis process. The default value is set to 0. The parameter '--sds' is used to set the number of standard deviations ("N") from the mean to identify samples for filtering based on protein group counts. It allows the user to identify potential outliers or data points that are significantly different from most of the data. The default value is set to 3. The command-line argument '--exclude' is used in ProtPipe to specify samples that should be excluded from the analysis. After removal of outlier proteins or samples, ProtPipe conducts Spearman correlation analysis to assess data reproducibility within samples and across replicates. This analysis generates a figure that demonstrates the consistency and reliability of the dataset by quantifying monotonic relationships through Spearman correlation coefficients (SCCs).

ProtPipe provides three methods for data clustering and visualization: Hierarchical Clustering (HC), Principal Component Analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP). These methods enable users to explore data structure and relationships visually, facilitating comprehensive data interpretation and insight generation.

Differential intensity analysis

ProtPipe conducts differential intensity analysis to identify proteins whose abundances significantly correlate with experimental groups or treatments. Only proteins detected in more than half of all biological replicates are considered. The process begins with data preprocessing using scaled and normalized data. We perform *t*-tests to assess statistical significance between experimental groups or treatments. This helps calculate fold changes (FCs) and *P* values and corrects for multiple comparisons using the Benjamini-Hochberg method. By default, ProtPipe applies $|\log_2 FC| \geq 1$ and adjusted P value ≤ 0.01 to

determine statistically significant differential protein expression. Users can adjust these thresholds using parameters like '--foldchange' and '--fdr'. For visualizing results, ProtPipe offers a volcano plot that highlights proteins with significant changes in intensity. Users can further customize this plot by providing a list of gene symbols for labeling using the '--label' parameter.

Annotation and enrichment analyses utilize the R package clusterProfiler [31], leveraging the Gene Ontology (GO) database [32] to annotate proteins according to their Biological Process, Molecular Function, and Cellular Component. Pathway analysis is performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [33]. Enrichment analyses in Protpipe are ranked on a default adjusted P value threshold of ≤ 0.05, although this threshold can be adjusted using the '--enrich' parameter. The top enriched entries, typically set to a default of 20, are graphically represented.

PPI analysis

ProtPipe initiates PPI analysis by identifying interaction pairs using a matrix-based approach. To account for differing intensity levels between pull-down and negative control samples, ProtPipe employs intergroup normalization and imputes non-detected values in the negative control group, enabling robust differential abundance analysis. Differential proteins are identified by employing a t-test to compare pull-down samples against the negative control, with default $|\log_2 FC| \geq 1$ and adjusted P value ≤ 0.01. Furthermore, ProtPipe integrates data from the STRING database [34], which compiles interactions from experimental data, computational predictions, and text mining. This integration enhances the analysis by validating candidate PPIs and uncovering novel interactions, thereby providing comprehensive insights into functional relationships and interaction networks among the identified proteins.

Immunopeptidome deconvolution

Immunopeptidome deconvolution in ProtPipe facilitates neoantigen prediction, bridging genomics and immunology to enable personalized treatment based on individual genotypes. We utilize the peptide search results obtained from FragPipe [28] for subsequent analysis. By providing the human leukocyte antigen (HLA) typing information of the samples, ProtPipe employs MHCflurry [35], a machine learning tool for predicting potential neoantigens and their binding affinities with the MHC. To ensure high confidence in neoantigen predictions, ProtPipe applies stringent filtering criteria. This includes selecting neoantigens with an MHCflurry affinity score below 200 and an MHCflurry affinity percentile lower than 2%. These criteria are crucial for refining candidate selections and preparing them for further investigation and visualization.

Experimental design and statistical analysis

The MS raw files in case studies 1, 2, and 3 are referred in original publications [36–38]. In case study 4, data were generated from the postmortem frontal cortex tissues collected at Mayo Clinic, including samples from individuals with frontotemporal lobar degeneration with C9 mutation and TDP-43 pathology (C9-FTLD-TDP), as well as healthy controls ($n = 10$ each group). The plasma and cerebrospinal fluid ($n = 4$, each specimen) were subjected to nanoparticle enrichment and fractionation (seer.bio) prior to MS analysis (5 fractions for each sample). The skeletal muscle biopsies were conducted at University College London from patients with inclusion-body myositis ($n = 20$). The neurons, microglia, and astrocytes ($n = 6$) were fully differentiated from KOLF2.1J iPSC line. The statistical analyses and visualization were conducted in our Protpipe as described.

Data acquisition for MS-based proteomics analysis

For the samples we reported in case study 4, the data were generated using MS-based proteomics with our previously published fully automated pipeline [38]. Briefly, protein extraction from biospecimens was conducted using a magnetic beads-based approach coupled with on-beads tryptic digestion (Catalog No. GE45152105050250, Millipore Sigma, Burlington, MA). The resulting peptides were measured and normalized using an automated colorimetric assay on a Bravo robot (Catalog No. G5563A, Agilent Technologies, Santa Clara, CA). The final resulting tryptic peptides were separated on a nano column (75 μm × 500 mm, 2 μm C18 particle) using a 2-h efficient linear gradient [phase B, 2%–35% acetonitrile (ACN)] on an UltiMate 3000 nano-HPLC system (Catalog No. IQLAAAGABHFAPBMBEZ, Thermo Fisher Scientific, Waltham, MA). We used DIA discovery proteomics on a hybrid Orbitrap Eclipse mass spectrometer (Catalog No. FSN04-10000, Thermo Fisher Scientific). Specifically, MS1 resolution was set to 120,000, and MS2 resolution was set to 30,000. For DIA isolation, the precursor range was set to 400–1000 m/z, and the isolation window was 8 m/z, resulting in 75 windows for each scan cycle (3 s). High collision dissociation was used for fragmentation with 30% collision energy. The automatic gain control (AGC) target was set to 800% for MS2 scan (improve MS2 spectra quality). The MS2 scan range was defined as 145–1450 m/z, which covers many fragment ions of typical tryptic peptides.

Results

ProtPipe overview and performance evaluation

ProtPipe is a valuable tool designed for MS-based proteomics, offering an easy-to-use, efficient, and reproducible workflow suitable for users with minimal command line experience. A user-friendly wrapper script runs DIA-NN [21] within a preconfigured Singularity image, enabling users to efficiently extract protein abundance data from raw MS outputs. Notably, ProtPipe script accommodates protein abundance estimates from database searching engines, DIA-NN [21], Spectronaut, and FragPipe [28]. Moreover, ProtPipe includes a preconfigured R environment, which enables seamless post-processing of the protein abundance estimates. Researchers can utilize this environment to generate comprehensive quality control reports, perform diverse analyses, and create visualizations, all within an accessible framework (**Figure 1**). The post-analysis workflow of ProtPipe is designed to handle various stages of data processing, ensuring comprehensive analysis and accurate interpretation of proteomics datasets. The steps involved in the post-analysis workflow are detailed below and summarized in Table 1.

In our performance evaluation, we focused on key metrics essential for efficient data analysis. We analyzed CPU usage, memory consumption, network activity, and disk utilization across varying sample sizes and computational resources (**Table 2**). Interestingly, despite the database searching stage being expected to primarily consume CPU resources, our analysis revealed significant CPU-idle time during this stage. The insights

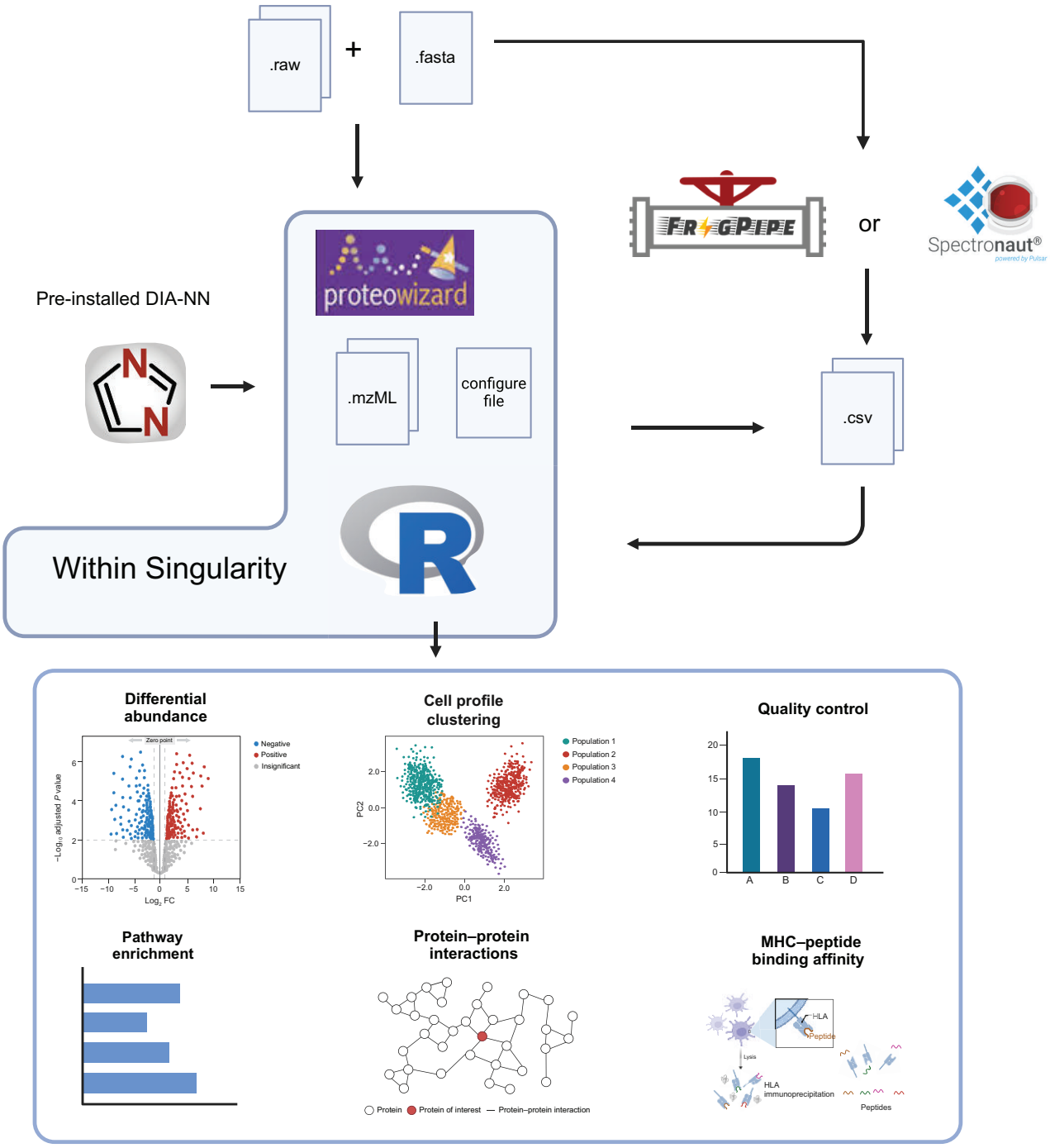

Figure 1 The comprehensive workflow of ProtPipe, a multifunctional data analysis pipeline for proteomics and peptidomics
The database search is performed using pre-installed DIA-NN for DIA data, offline Spectronaut for DIA data, or offline FragPipe for DDA data and immunopeptidomics data. The resulting .csv files (*e.g.*, protein groups or peptides) can then be analyzed by ProtPipe. ProtPipe generates figures and datasets for quality control, clustering analysis, differential abundance analysis, pathway enrichment analysis, PPI network analysis, and MHC–peptide binding affinity predictions. DIA, Data-Independent Acquisition; PPI, protein–protein interaction; FC, fold change; PC, principal component; MHC, major histocompatibility complex; HLA, human leukocyte antigen.

gained from this evaluation were instrumental in optimizing ProtPipe for robust and efficient proteomics data analysis.

Case study 1: analysis of a large-scale proteomics dataset

For our first case study, we generated a dataset by applying the proteomics methodology to characterize human induced pluripotent stem cells (iPSCs) and derived neurons [38,39]. ProtPipe is designed to encompass a comprehensive suite of data quality control measures, including Number of Protein Groups, Distribution of Protein Intensity, and Correlation Among Replicates. This first case study analysis detected more than 7500 proteins across the diverse samples (Figure 2A). The substantial assembly of identified proteins attests to the efficacy of our experimental protocols.

Table 2 Resource requirements for various stages of data processing with ProtPipe

Step	Sample number	Sample size	CPU	Threads	Running time	Memory used
Convert raw to .mzML	42	44.1 Gb	2	1	109 min	2 GB
DIA-NN	42	44.1 Gb	10	10	1138 min	12 GB
DIA-NN	42	44.1 Gb	10	5	2016 min	12 GB
DIA-NN	10	10.7 Gb	10	10	355 min	12 GB
DIA-NN	10	10.7 Gb	10	5	813 min	12 GB
Post analysis	42	4.1 Mb	2	1	25 s	20 MB

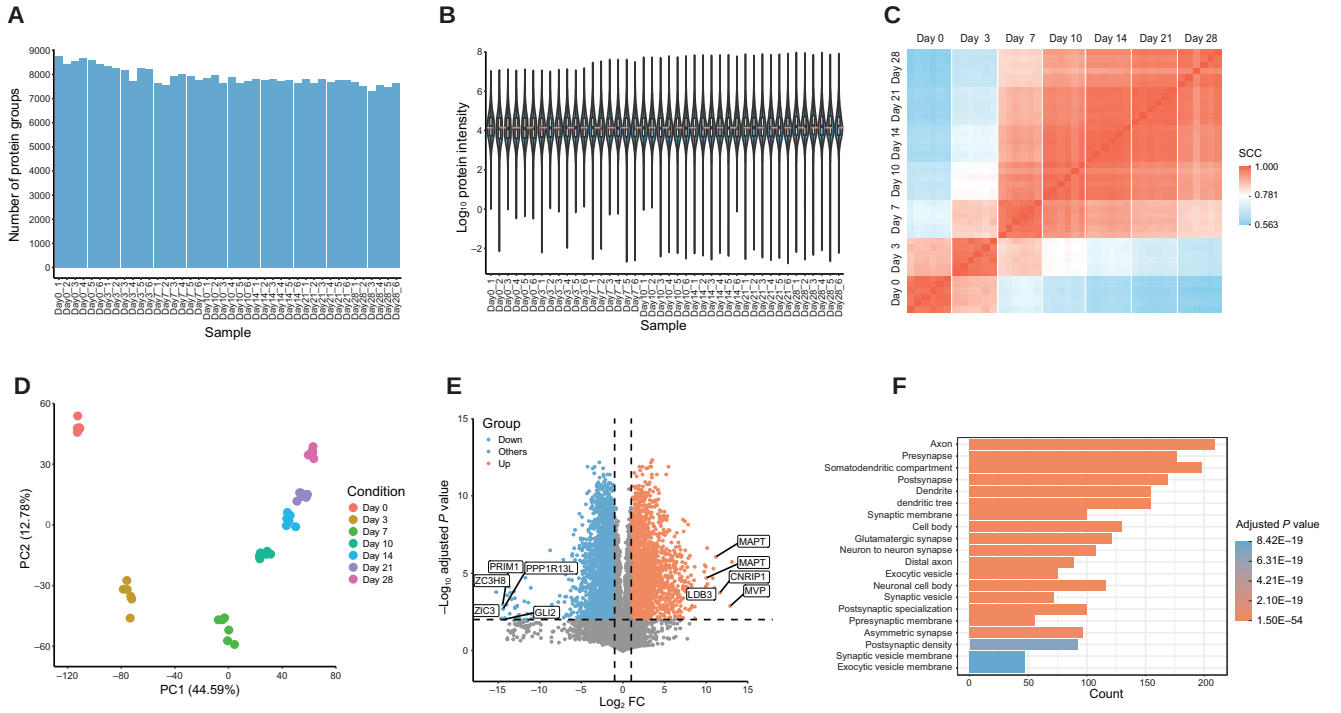

Figure 2 Analysis of a large-scale proteomics dataset

A. Distribution of identified protein groups. **B**. Distribution of protein intensity. **C**. Correlation among biological replicates. **D**. PCA plot. **E**. Volcano plot showing differential abundance analysis results between two groups. Orange dots represent up-regulated proteins, blue dots indicate down-regulated proteins, and gray dots signify proteins with no significant alterations. **F**. Set of bar charts present GO terms related to Cellular Component. PCA, Principal Component Analysis; SCC, Spearman correlation coefficient; GO, Gene Ontology.

We evaluated the distribution of protein intensity values (Figure 2B) to ascertain uniformity and coherence. A protein intensity profile that is evenly distributed across the samples serves as a compelling indicator of data quality, with its consistency underscored. We found a uniform distribution signifying the absence of skewness or predisposed biases within the data. For the correlation analysis, we provide both correlation analysis and visualization tools to assess data quality (Figure 2C). The strong correlation observed among replicates underscores the reliability and consistency of the proteomics data obtained using ProtPipe. If any samples exhibit lower correlation, they can be excluded from the analysis. ProtPipe offers a selection of cluster methods, allowing for comprehensive analysis. For this case study dataset, ProtPipe facilitated the generation of three distinct cluster analyses: PCA (Figure 2D), UMAP (Figure S1A), and HC (Figure S1B). These methods provide users with diverse perspectives on the data, enhancing the interpretability and depth of the analysis.

Differential abundance analysis in proteomics involves comparing the abundance of proteins in different experimental conditions. We used the Student's *t*-test to test for differences in protein abundance while accounting for inherent data variability in proteomics. Based on the provided design matrix, ProtPipe generates specific comparisons as intended. We illustrate this with an example of the comparison between fully differentiated neurons (on day 28) and iPSCs. During our analysis, the volcano plot served as a dynamic tool that allowed for flexibility in parameter settings, ensuring visualization aligned with the nuances of our data. By default, significant differences in abundance require $|\log_2 FC| \geq 1$ and adjusted P value ≤ 0.01 (Figure 2E). These specified threshold values can be adjusted based on the specific context of the investigation. ProtPipe is designed to highlight key genes that exhibit significant changes. The top 5 genes that show the most pronounced up- or down-regulation within the comparison are labeled by default. Specific user-defined genes of interest can also be labeled.

After identifying differentially abundant proteins, we wanted to elucidate the functional context and implications of these alterations to ultimately understand how they may impact cellular functions. The pathway analysis for this dataset encompassed a comprehensive examination of GO terms, which include Cellular Component (Figure 2F), Biological

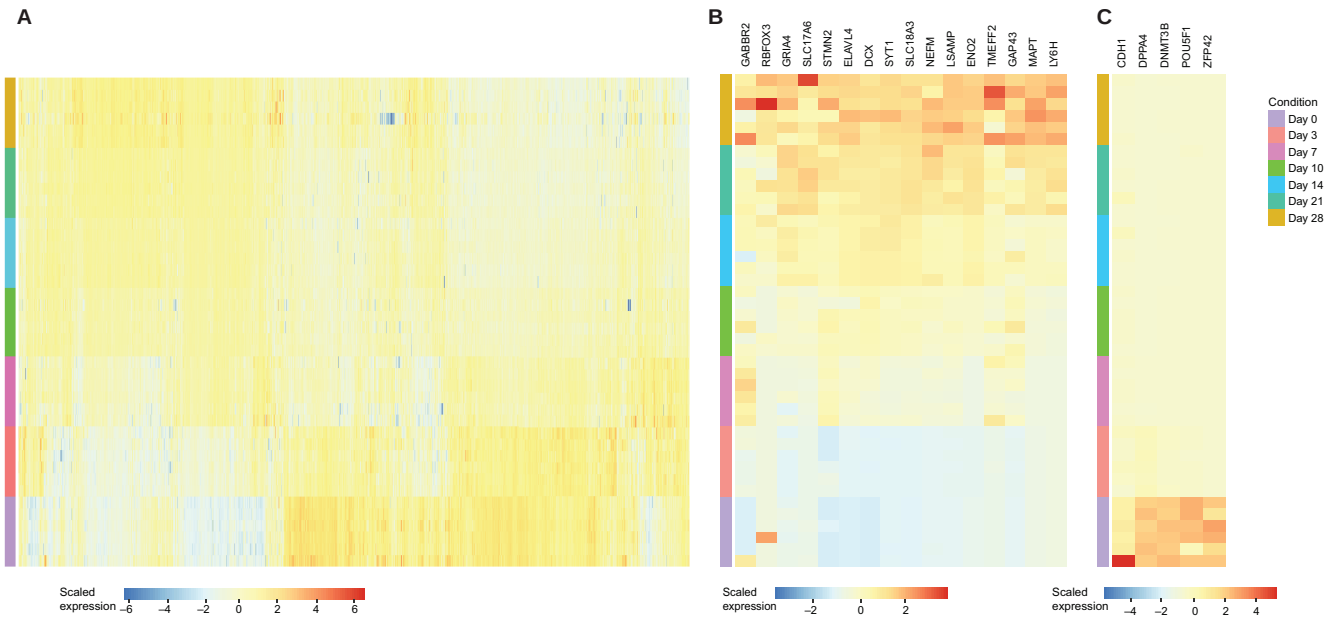

Figure 3 Heatmap analysis of protein intensity patterns
A. Heatmap showing protein intensity patterns across all detected proteins in the samples. **B**. Customized heatmap generated using a select gene list focusing on marker genes closely associated with neuron cells. **C**. Customized heatmap generated using a provided gene list highlighting marker genes unique to iPSCs. iPSC, induced pluripotent stem cell.

Process (Figure S1C), and Molecular Function (Figure S1D) categories. These three categories provide a holistic understanding of the functional context of genes and proteins, shedding light on their cellular localizations, roles in biological processes, and molecular activities. Next, heatmap is employed to visualize patterns of protein intensity across the samples. By default, the ProtPipe generated heatmap including all detected proteins (**Figure 3**A) or a customized dataset with select gene lists. Our heatmaps encompassed marker proteins closely associated with neurons (Figure 3B) and pluripotency (Figure 3C), respectively. ProtPipe's heatmap analysis empowers researchers to conduct more targeted investigations aligned with their research objectives and accommodates a wide range of analytical needs.

Case study 2: protein–protein interactome study

AP-MS enables PPI study. As a showcase, we used ProtPipe to analyze a dataset originating from our previous publication focused on UNC13A, a genetic risk factor associated with amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) [37]. To identify proteins binding to UNC13A, we conducted an AP-MS experiment. During the quality control phase, we assessed the number of detected protein groups and the distribution of protein abundance within both the negative control group and the UNC13A pull-down group (**Figure 4A and B**). Notably, the control group exhibited a limited number of detected protein groups and lower protein intensity levels. We calculated sample correlations to demonstrate the reproducibility of replicate experiments (Figure 4C). The volcano plot showed a statistically significant enrichment in the presence of UNC13A RNA when contrasted with a negative control RNA (Figure 4D). Furthermore, we conducted GO enrichment analysis focusing on UNC13A RNA-binding proteins (RBPs), revealing that the most prominently enriched biological processes encompassed RNA metabolism, mRNA processing, and RNA splicing (Figure S2A). Additionally, our molecular function analysis demonstrated significant enrichment in terms pertaining to RNA binding (Figure S2B). In terms of cellular component analysis, our findings exhibited notable enrichments within the nuclear body and spliceosome categories (Figure S2C).

In PPI analysis, we utilized the STRING database which offers extensive insights into these complex molecular associations. The Venn diagram in Figure 4E illustrates the intersection of interacting proteins identified through the STRING database and those obtained from experimental data. This visualization effectively highlights the overlap between computational predictions and experimental findings, providing a comprehensive view of potential PPIs. It reveals how many proteins are consistently identified by both methods, thereby validating the experimental results against known interaction databases. In addition, the rank plot (Figure 4F) presents proteins ranked by their FCs in abundance. Yellow dots on this plot indicate proteins that are previously known to interact with the target protein, according to the STRING database. This plot not only ranks proteins by their differential expression but also visually emphasizes those with established interactions, reinforcing the biological relevance of the findings. Together, these analyses provide a robust framework for validating and interpreting proteomics data within the context of known PPI networks.

Case study 3: identification of allele-specific peptides with immunopeptidomics

ProtPipe contains a module designed to facilitate the deconvolution of immunopeptides to specific MHC alleles in cells and biological samples. As a case study, we deployed the pipeline within the context of MS-based proteogenomic profiling [36], aiming to ascertain potential immunogenic peptides presented by each of MHC class I alleles in melanoma and EGFR-mutant lung adenocarcinoma. The total count of

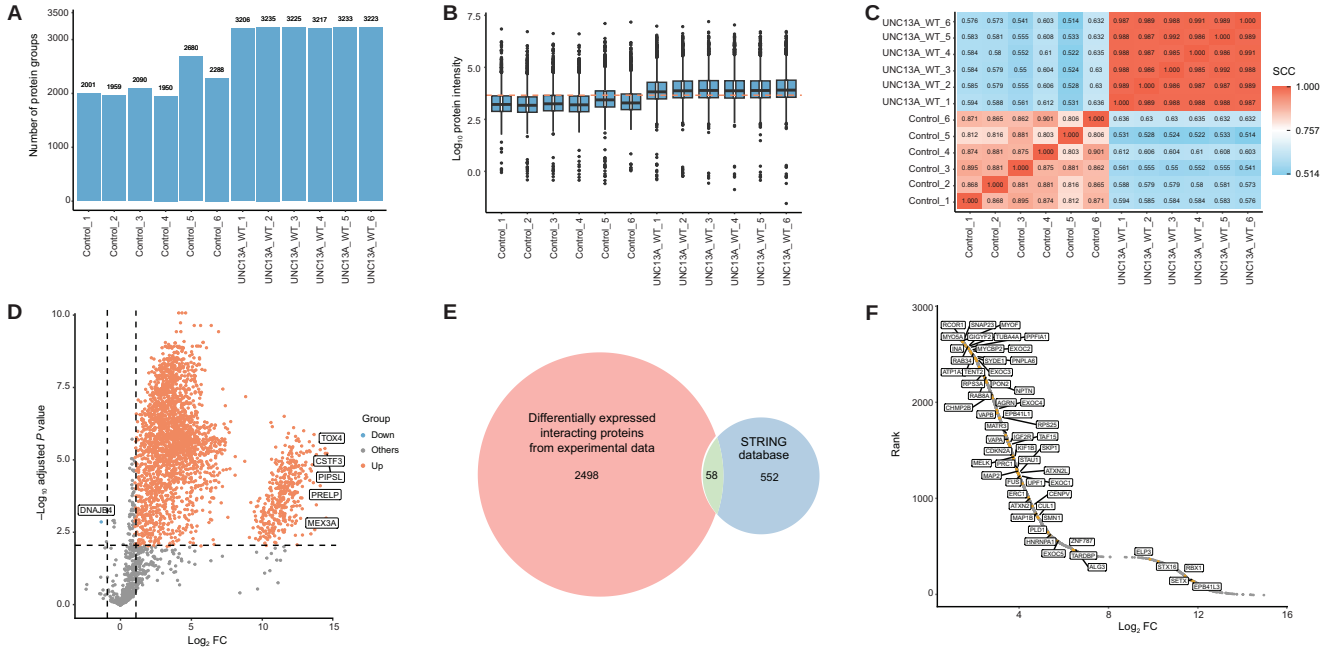

Figure 4 Comprehensive analysis of UNC13A PPIs

A. and **B**. Bar chart and box plot depicting the number of detected protein groups (A) and the distribution of protein intensity (B) in both the control group and the UNC13A pull-down group. **C**. Reproducibility assessment by replication correlation. **D**. Identification of enriched proteins. Volcano plot illustrates proteins significantly enriched in the presence of UNC13A RNA compared to a negative control. **E**. Venn diagram showing the overlap of the interacting proteins obtained by leveraging the STRING database and the potential interacting proteins obtained from experimental data. **F**. Rank plot of the proteins based on their FCs in abundance. Yellow dots on the plot represent proteins that are previously known to interact with the target protein, as obtained from the STRING database.

peptides identified within class I immunopeptidome from cancer cell lines and tumors range from 1520 to 3509 (**Figure 5A**). With MHC genotyping, we deconvoluted the total pool of MHC-bound peptides to their corresponding MHC alleles in PC9 cells (Figure 5B). It is noteworthy that HLA-C*07:02 exhibits the highest count of binding peptides. Subsequently, we examined peptides demonstrating strong binding affinities with specific HLA alleles, notably observing numerous peptides with robust binding affinities to HLA-A*02:06 (Figure 5C). To emphasize the peptides exhibiting the most potent binding affinities, a ranking plot of peptide–HLA binding affinities with each peptide specific binding to a particular HLA is denoted by distinct colors (Figure 5D). Notably, the top 5 peptides demonstrating the strongest binding affinities are all associated with HLA-A*02:06 (Figure 5D). The plot of peptide binding patterns across various HLA alleles highlights the top 3 peptides with the most robust binding affinities for each respective allele (Figure 5E). Furthermore, all MHC–peptide binding affinity analyses were conducted in the H1975 cells, as demonstrated in Figure S3.

Case study 4: a proteomics database related to central nervous system

To showcase the reliably and repeatedly identified proteins in common human specimen and cell line models in neurodegenerative disease research using MS-proteomics, we compiled multiple in-house generated proteomics datasets of central nervous system (CNS)-related cell lines and tissues. These datasets were sourced from a wide range of brain-related tissues and cellular contexts, encompassing human frontal cortex tissue, plasma, cerebrospinal fluid (CSF), and skeletal muscle biopsies, as well as human iPSC-derived neuronal, microglial, and astrocytic cells. Despite plasma and muscle not being directly related to CNS, they are highly relevant to CNS biomarker discovery because of their easy accessibility. Using ProtPipe, we assessed the protein group count within each of these samples (**Figure 6A**). The samples derived from human tissues exhibited a comparatively lower number of detected protein groups. Particularly, the plasma and CSF samples, processed by nanoparticle-based fractionation using Seer bio, displayed the most modest protein group counts, typically hovering around 5000 protein groups. We performed UMAP on this dataset with the aim of examining data clustering patterns (Figure 6B). The CSF and plasma samples clustered tightly together. The iPSC-derived microglial, astrocytic, and neuronal cells demonstrated distinct clustering patterns. Among these, the iPSC-derived neuronal cells exhibited a clustering pattern more closely aligned with brain samples compared to other tissues and iPSC-derived microglial and astrocytic cells. Subsequently, our objective was to elucidate the protein expression patterns across the entirety of the samples.

Using ProtPipe we generated a heatmap focusing exclusively on the subset of proteins that were consistently detectable across all samples, totaling 399 proteins (Figure 6C). Notably, our observations revealed a distinct subset of proteins exhibiting particularly pronounced expression within specific cellular contexts. Remarkably, this subset of 356 proteins exhibited a discernible capacity to effectively distinguish between different cell types based on the clustering patterns observed in the heatmap. We also conducted the pathway analysis of these proteins that were consistently detectable

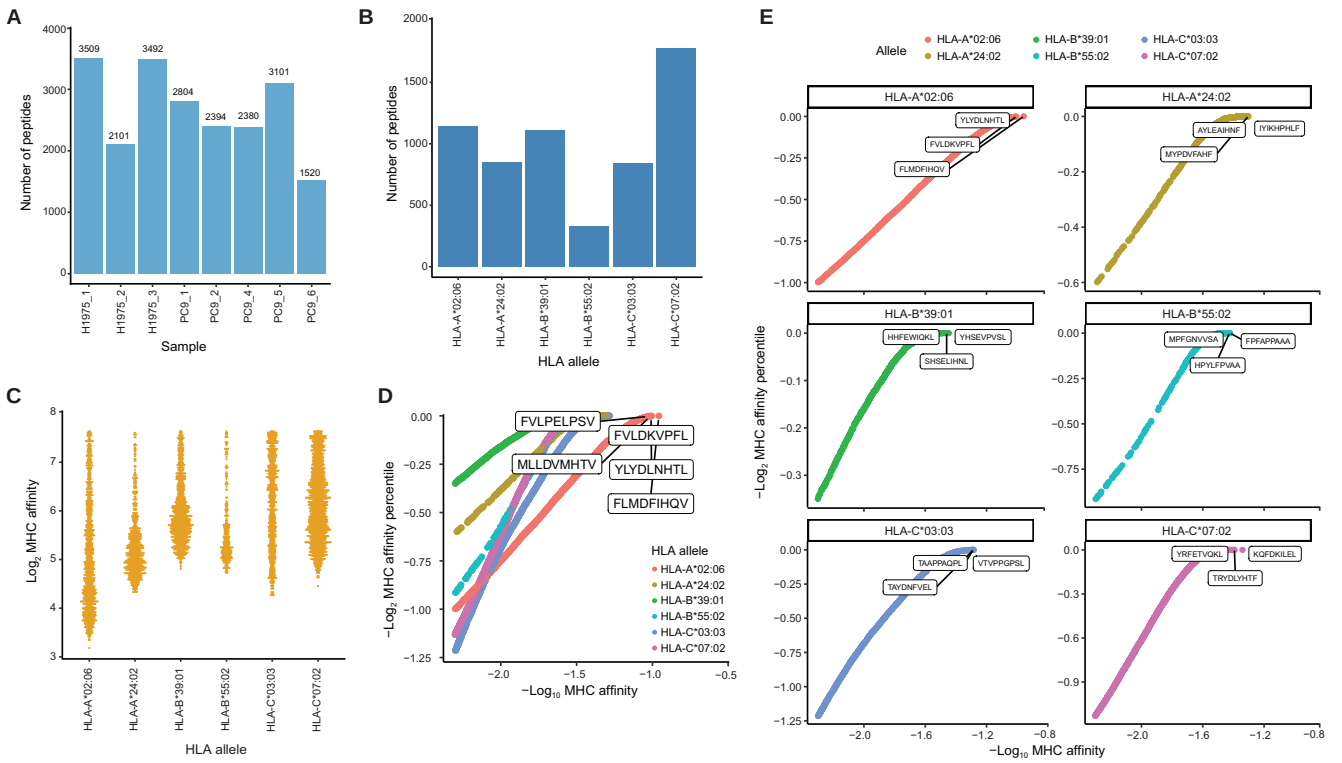

Figure 5 MHC-bound peptide deconvolution

A. The count of peptides identified in class I immunopeptidome from various cancer cell lines and tumors. **B**. The count of peptides with HLA alleles has strong affinities. The cutoff of binding affinities < 200. **C**. The peptides' binding affinities for each HLA allele with specific cutoff criteria for MHCflurry affinity (< 200) and percentile (< 2). **D**. Ranking plot of peptide–HLA binding affinities. The ranking plot visualizes the top 5 peptides exhibiting the strongest binding affinities. **E**. Peptide binding patterns by HLA alleles. Top 3 peptides with robust binding affinities for each respective HLA allele are shown.

across all samples (Figure 6D), where pathways related to actin binding and protein folding are significantly enriched. Since actin assembles are highly conserved, this CNS-related 356 protein panel can serve as a surrogate marker to differentiate different CNS cell and tissue types. Thus, this serves as a valuable resource for researchers seeking to identify benchmark protein characteristic of distinct cell types and tissues.

Discussion

The integration of MS with various proteomics methodologies, such as AP-MS, immunopeptidomics, and total protein proteomics, helps researchers to investigate the intricate dynamics of biological systems. In this context, ProtPipe was designed to handle the complexities of modern MS data, offering a comprehensive toolkit for investigating these dynamics. By supporting a range of proteomics methodologies, ProtPipe showcases its ability to enhance MS-based proteomics data analysis. It serves as a multifunctional pipeline meticulously crafted to streamline and automate the processing and analysis of high-throughput proteomics and peptidomics datasets.

The pipeline encompasses a suite of features including data quality control, imputation, sample filtering, and normalization, which collectively contribute to the reliability and robustness of downstream analysis. ProtPipe offers researchers a spectrum of downstream analysis capabilities. Moreover, it empowers researchers with pathway enrichment analysis, enabling the exploration of functional implications of identified proteins. The PPI analysis feature allows for the examination of intricate networks underlying biological systems. Additionally, it provides the valuable capability of MHC–peptide binding affinity prediction. Immunopeptides sequenced by MS can be deconvoluted to specific MHC alleles using our pipeline, providing a comprehensive rank list of all peptides with corresponding MHC alleles. Through deconvolution, ProtPipe helps researchers to identify and prioritize peptides that have highest binding affinities. This process enables researchers to focus their efforts on characterizing peptides that are most relevant for immune recognition and subsequent therapeutic interventions, thereby streamlining the immunopeptidome analysis workflow and enhancing the efficiency of immunotherapy and cancer vaccine development.

Importantly, ProtPipe is not only a robust tool but also an accessible one. It is an open-source software solution, readily available to the scientific community with a parallel webapp is. A key feature of ProtPipe is its deployment in a containerized environment, ensuring stable and easily accessible locked-in versions that perform reliably. This approach minimizes dependency issues and provides a consistent user experience. To ensure the sustainability and usability, we will update ProtPipe along with the advancements or upgrades in dependent software. ProtPipe's modular design facilitates seamless integration of new versions of its dependent tools and libraries. Regular updates and community contributions will help keep ProtPipe compatible with the latest developments in the field. By maintaining an active repository and

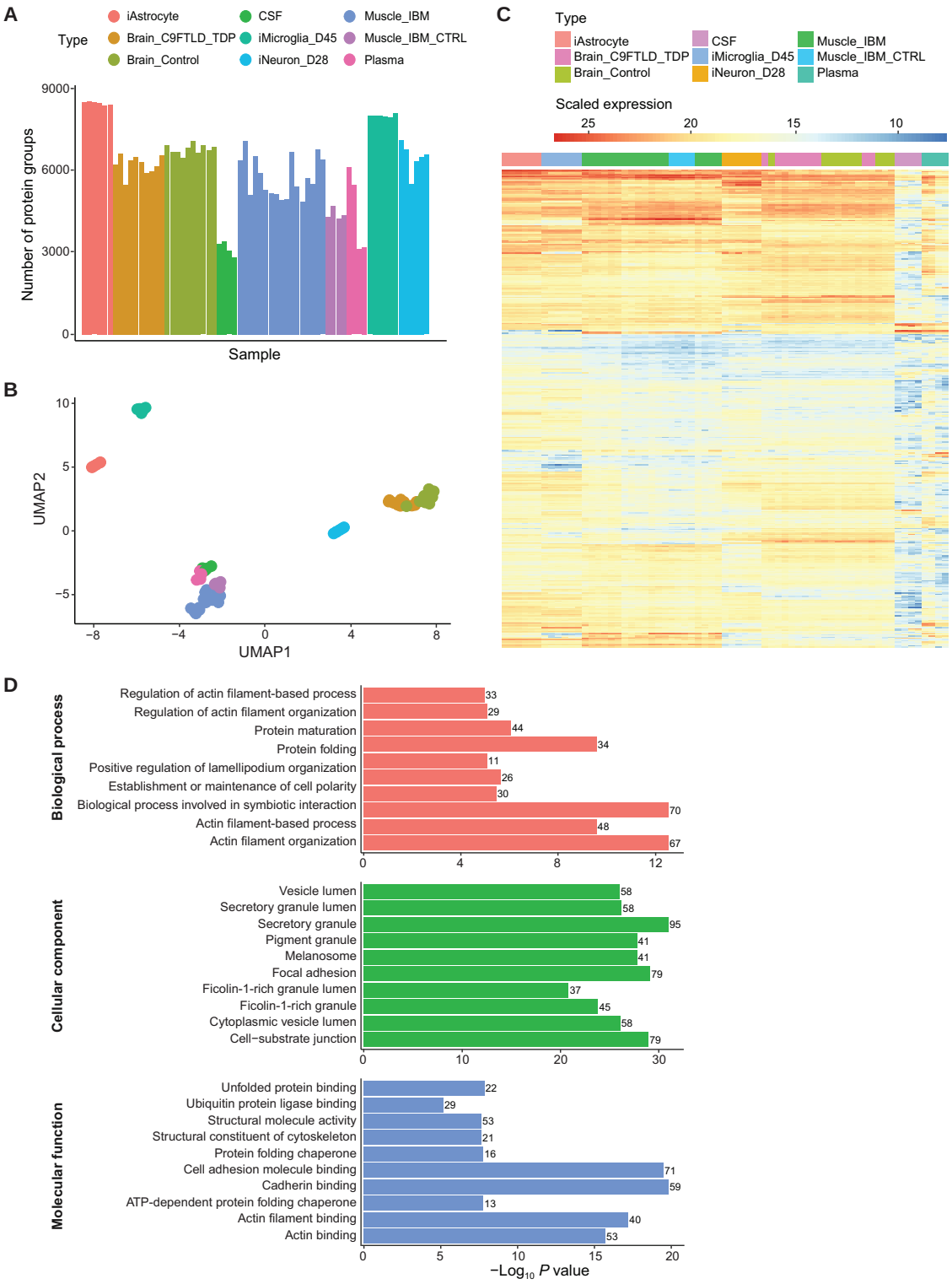

Figure 6 A proteomics database related to CNS

A. The overview of the protein group counts of various tissue and cellular sources, including human brain samples, plasma, CSF, muscle cells, and human iPSC-derived neuronal, microglial, and astrocytic cells. **B**. The UMAP plot. **C**. Comprehensive heatmap of protein expression patterns across all detected proteins in the dataset. **D**. GO enrichment analysis of a subset of 356 proteins consistently detectable across all samples. CNS, central nervous system; CSF, cerebrospinal fluid; UMAP, Uniform Manifold Approximation and Projection.

clear documentation, users can easily implement updates and improvements, ensuring that ProtPipe remains a cutting-edge tool for proteomics research. We intend to continue to update ProtPipe for other analysis functions. ProtPipe's GitHub repository features an issue page where users can seek help and report any problems, fostering a collaborative environment for continuous improvement.

In conclusion, ProtPipe represents a significant advancement in MS-based proteomics data analysis. Its multifunctional capabilities enable researchers to extract meaningful

insights from high-throughput proteomics and peptidomics datasets efficiently and reliably, advancing our understanding of complex biological phenomena.

Ethical statement

All human specimens used in this work were conducted under the NeuroBioBank (NBB) Brain and Tissue Repositories Material Transfer Agreement ID 2771 with the National Institute on Aging (NIA) of the National Institutes of Health (NIH) as the recipient.

Code availability

ProtPipe is publicly available at https://github.com/NIH-CARD/ProtPipe.

Data availability

The MS raw files of case study 4 have been deposited to the ProteomeXchange Consortium via the PRIDE [40] partner repository (ProteomeXchange: PXD047657). The processed protein or peptide abundance data, design matrix, heatmap gene list, and HLA typing .csv files for case studies 1, 2, and 3 can be found at https://github.com/NIH-CARD/ProtPipe/tree/main.

CRediT author statement

Ziyi Li: Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Cory A. Weller:** Software, Writing – review & editing. **Syed Shah:** Software. **Nicholas L. Johnson:** Software. **Ying Hao:** Data curation, Investigation. **Paige B. Jarreau:** Writing – review & editing. **Jessica Roberts:** Investigation. **Deyaan Guha:** Validation. **Colleen Bereda:** Validation. **Sydney Klaisner:** Validation. **Pedro Machado:** Resources. **Matteo Zanovello:** Resources. **Mercedes Prudencio:** Resources. **Björn Oskarsson:** Resources. **Nathan P. Staff:** Resources. **Dennis W. Dickson:** Resources. **Pietro Fratta:** Resources. **Leonard Petrucelli:** Resources. **Priyanka Narayan:** Resources. **Mark R. Cookson:** Writing – review & editing. **Michael E. Ward:** Resources. **Andrew B. Singleton:** Supervision, Project administration. **Mike A. Nalls:** Conceptualization, Supervision, Project administration, Writing – review & editing. **Yue A. Qi:** Conceptualization, Data curation, Investigation, Supervision, Project administration, Writing – review & editing. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzae083).

Competing interests

Mike A. Nalls, Cory A. Weller, Nicholas L. Johnson, Syed Shah, and Ziyi Li's participation in this project was part of a competitive contract awarded to DataTecnica LLC by the National Institutes of Health (NIH) to support open science research. Mike A. Nalls also currently serves on the scientific advisory board for Character Bio Inc., and is a scientific founder at Neuron23 Inc. Björn Oskarsson serves as a consultant for Columbia University/Tsumura Inc., MediciNova, Biogen, uniQure, Amylyx, and Mitsubishi, and has research grants from Columbia University/Tsumura Inc., Biogen, MediciNova, Cytokinetics, Mitsubishi, Calico, Novartis, Sanofi, Ashvattha, and TARGET ALS. Other authors have declared no competing interests.

Acknowledgments

This work was supported by grants from the Intramural Research Program of the NIH, NIA, NIH, Department of Health and Human Services (Grant No. ZIAAG000534), as well as the National Institute of Neurological Disorders and Stroke, the NIH, USA (Grant Nos. RF1 NS120992 and U54 NS123743) to Pedro Machado. We thank the NIH HPC system (*Biowulf*, http://hpc.nih.gov) for making this work possible.

ORCID

0000-0002-5165-8772 (Ziyi Li)
0000-0001-6965-5599 (Cory A. Weller)
0000-0002-2094-3386 (Syed Shah)
0009-0004-1464-5050 (Nicholas L. Johnson)
0000-0001-5053-4850 (Ying Hao)
0000-0002-3963-2641 (Paige B. Jarreau)
0009-0001-4103-9222 (Jessica Roberts)
0009-0004-1955-7818 (Deyaan Guha)
0000-0002-9435-6141 (Colleen Bereda)
0009-0001-5033-0388 (Sydney Klaisner)
0000-0002-8411-7972 (Pedro Machado)
0000-0003-3343-1547 (Matteo Zanovello)
0000-0002-4894-4858 (Mercedes Prudencio)
0000-0002-1725-9866 (Björn Oskarsson)
0000-0001-6760-3859 (Nathan P. Staff)
0000-0001-7189-7917 (Dennis W. Dickson)
0000-0002-7554-1632 (Pietro Fratta)
0000-0002-7554-1632 (Leonard Petrucelli)
0000-0002-9328-4959 (Priyanka Narayan)
0000-0002-1058-3831 (Mark R. Cookson)
0000-0002-5296-8051(Michael E. Ward)
0000-0001-5606-700X (Andrew B. Singleton)
0000-0003-0319-4325 (Mike A. Nalls)
0000-0003-1914-8710 (Yue A. Qi)

References

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
[2] Wu CC, Yates JR 3rd. The application of mass spectrometry to membrane proteomics. Nat Biotechnol 2003;21:262–7.
[3] Wepf A, Glatter T, Schmidt A, Aebersold R, Gstaiger M. Quantitative interaction proteomics using mass spectrometry. Nat Methods 2009;6:203–5.
[4] Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem 2012;404:939–65.
[5] Rozanova S, Barkovits K, Nikolov M, Schmidt C, Urlaub H, Marcus K. Quantitative mass spectrometry-based proteomics: an overview. Methods Mol Biol 2021;2228:85–116.
[6] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature 2003;425:737–41.

[7] Yen HC, Xu Q, Chou DM, Zhao Z, Elledge SJ. Global protein stability profiling in mammalian cells. Science 2008;322:918–23.
[8] Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 2009;11:49–79.
[9] Zhu W, Smith JW, Huang CM. Mass spectrometry-based label-free quantitative proteomics. J Biomed Biotechnol 2010; 2010:840518.
[10] Levin Y, Schwarz E, Wang L, Leweke FM, Bahn S. Label-free LC-MS/MS quantitative proteomics for large-scale biomarker discovery in complex samples. J Sep Sci 2007;30:2198–203.
[11] Gavin AC, Maeda K, Kuhner S. Recent advances in charting protein–protein interaction: mass spectrometry-based approaches. Curr Opin Biotechnol 2011;22:42–9.
[12] Chen GI, Gingras AC. Affinity-purification mass spectrometry (AP-MS) of serine/threonine phosphatases. Methods 2007; 42:298–305.
[13] Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol Syst Biol 2007;3:89.
[14] Liu X, Salokas K, Tamene F, Jiu Y, Weldatsadik RG, Ohman T, et al. An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. Nat Commun 2018;9:1188.
[15] Schuster H, Peper JK, Bosmuller HC, Rohle K, Backert L, Bilich T, et al. The immunopeptidomic landscape of ovarian carcinomas. Proc Natl Acad Sci U S A 2017;114:E9942–51.
[16] Liu C, Song X, Nisbet R, Gotz J. Co-immunoprecipitation with Tau isoform-specific antibodies reveals distinct protein interactions and highlights a putative role for 2N Tau in disease. J Biol Chem 2016;291:8173–88.
[17] Wu T, Guan J, Handel A, Tscharke DC, Sidney J, Sette A, et al. Quantification of epitope abundance reveals the effect of direct and cross-presentation on influenza CTL responses. Nat Commun 2019;10:2846.
[18] Laumont CM, Daouda T, Laverdure JP, Bonneil E, Caron-Lizotte O, Hardy MP, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun 2016;7:10238.
[19] Chong C, Muller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun 2020;11:1293.
[20] Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 2003;17:2337–42.
[21] Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods 2020; 17:41–4.
[22] Deutsch EW, Mendoza L, Shteynberg DD, Hoopmann MR, Sun Z, Eng JK, et al. Trans-Proteomic Pipeline: robust mass spectrometry-based proteomics data analysis suite. J Proteome Res 2023;22:615–24.
[23] Dai C, Pfeuffer J, Wang H, Zheng P, Kall L, Sachsenberg T, et al. quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. Nat Methods 2024; 21:1603–7.
[24] Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 2016;13:741–8.
[25] Bichmann L, Gupta S, Rosenberger G, Kuchenbecker L, Sachsenberg T, Ewels P, et al. DIAproteomics: a multifunctional data analysis pipeline for data-independent acquisition proteomics and peptidomics. J Proteome Res 2021;20:3758–66.
[26] Shah AD, Goode RJA, Huang C, Powell DR, Schittenhelm RB. LFQ-Analyst: an easy-to-use interactive web platform to analyze and visualize label-free proteomics data preprocessed with MaxQuant. J Proteome Res 2020;19:204–11.
[27] Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 2016;13:731–40.
[28] Yu F, Teo GC, Kong AT, Frohlich K, Li GX, Demichev V, et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat Commun 2023;14:4154.
[29] Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci Rep 2018;8:663.
[30] Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J Proteome Res 2016;15:1116–25.
[31] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[32] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[33] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34.
[34] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021; 49:D605–12.
[35] O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell Syst 2020;11: 412–8.e7.
[36] Qi YA, Maity TK, Cultraro CM, Misra V, Zhang X, Ade C, et al. Proteogenomic analysis unveils the HLA class I-presented immunopeptidome in melanoma and EGFR-mutant lung adenocarcinoma. Mol Cell Proteomics 2021;20:100136.
[37] Koike Y, Pickles S, Estades Ayuso V, Jansen-West K, Qi YA, Li Z, et al. TDP-43 and other hnRNPs regulate cryptic exon inclusion of a key ALS/FTD risk gene, UNC13A. PLoS Biol 2023; 21:e3002028.
[38] Reilly L, Lara E, Ramos D, Li Z, Pantazis CB, Stadler J, et al. A fully automated FAIMS-DIA mass spectrometry-based proteomic pipeline. Cell Rep Methods 2023;3:100593.
[39] Pantazis CB, Yang A, Lara E, McDonough JA, Blauwendraat C, Peng L, et al. A reference human induced pluripotent stem cell line for large-scale collaborative studies. Cell Stem Cell 2022;29: 1685–702.e22.
[40] Bhargava S, Jankowski J. The PRIDE database resources in 2023. Nephrol Dial Transplant 2023;39:4–6.

Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China 2024.

Genomics, Proteomics & Bioinformatics, 2024, 22, 1–12
https://doi.org/10.1093/gpbjnl/qzae083
Method