

Integrated View of Baseline Protein Expression in Human Tissues

Ananth Prakash,* David García-Seisdedos, Shengbo Wang, Deepti Jaiswal Kundu, Andrew Collins, Nancy George, Pablo Moreno, Irene Papatheodorou, Andrew R. Jones,* and Juan Antonio Vizcaíno*



Cite This: *J. Proteome Res.* 2023, 22, 729–742



Read Online

ACCESS |



Metrics & More



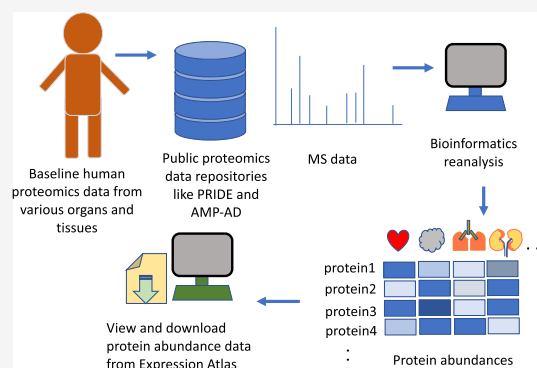
Article Recommendations



Supporting Information

ABSTRACT: The availability of proteomics datasets in the public domain, and in the PRIDE database, in particular, has increased dramatically in recent years. This unprecedented large-scale availability of data provides an opportunity for combined analyses of datasets to get organism-wide protein abundance data in a consistent manner. We have reanalyzed 24 public proteomics datasets from healthy human individuals to assess baseline protein abundance in 31 organs. We defined tissue as a distinct functional or structural region within an organ. Overall, the aggregated dataset contains 67 healthy tissues, corresponding to 3,119 mass spectrometry runs covering 498 samples from 489 individuals. We compared protein abundances between different organs and studied the distribution of proteins across these organs. We also compared the results with data generated in analogous studies. Additionally, we performed gene ontology and pathway-enrichment analyses to identify organ-specific enriched biological processes and pathways. As a key point, we have integrated the protein abundance results into the resource Expression Atlas, where they can be accessed and visualized either individually or together with gene expression data coming from transcriptomics datasets. We believe this is a good mechanism to make proteomics data more accessible for life scientists.

KEYWORDS: mass spectrometry, quantitative proteomics, public data re-use, human proteome



INTRODUCTION

High-throughput mass spectrometry (MS)-based proteomics approaches have matured and generalized significantly, becoming an essential tool in biological research, sometimes together with other “omics” approaches such as genomics and transcriptomics. It is now commonplace to make quantitative measurements of 2,000–3,000 proteins in a single LC–MS run and typically 6,000–7,000 proteins in workflows with fractionation. The most used experimental approach is data-dependent acquisition (DDA) bottom-up proteomics. Among existing DDA quantitative proteomics approaches, label-free is very popular, although labeled approaches such as metabolic labeling (e.g., SILAC) and especially techniques based on the isotopic labeling of peptides (e.g., TMT) are growing in importance. In bottom-up experiments, proteins are first digested into peptides using an enzyme (e.g., trypsin), and typically, several peptides are required per protein to give confidence in the measurement of protein-level quantification across samples. Measured peptide intensity is correlated with absolute protein abundance, but there can be differences depending on individual peptides due to the considerable variation in the ionization efficiency of these peptides. Different peptides can also be detected in different studies, giving rise to variability in protein abundance. One further challenge in quantitative proteomics relates to the “protein inference” problem.¹ In brief, many peptide sequences cannot

be uniquely mapped to a single protein due to common conserved sequences present in different gene families (paralogs). During the last decade, technological advances in MS have led to a large number of studies that have analyzed protein abundances across various human tissues and organs.^{2–5} These efforts are complemented by the comprehensive characterization of the human proteome performed within the Human Proteome Project (HPP),^{6–8} although the HPP has been focused on the identification of proteins without performing any quantitative analysis.

In parallel with the technical developments in chromatography, MS, and bioinformatics, the proteomics community has evolved to largely support open data practices. In brief, this means that datasets are released alongside publications, allowing other groups to check findings or reanalyze data with different approaches to generate new findings. Therefore, in recent years, the amount and variety of shared datasets in the public domain have grown dramatically. This was driven by

Received: July 7, 2022

Published: December 28, 2022



Table 1. List of Proteomics Datasets that were Reanalyzed

Expression Atlas accession number	Proteomics dataset identifier*	Tissues	Organs	Mass spectrometer	Number of MS runs	Number of samples	Fractionation	Number of protein groups†	Number of peptides†	Number of unique peptides†	Number of unique genes (canonical proteins) mapped‡
E-PROT-29	PXD010154 ³⁶	Adrenal gland, Bone marrow, Brain, Colon, Duodenum, Esophagus, Fallopian tube oviduct, Fat adipose tissue, Gallbladder, Heart, Kidney, Liver, Lung, Lymph node, Ovary, Pancreas, Pituitary hypophysis, Placenta, Prostate, Rectum, Salivary gland, Small intestine, Smooth muscle, Spleen, Stomach, Testis, Thyroid, Tonsil, Urinary bladder, Uterine endometrium, Vermiform appendix	Adrenal gland, Bone marrow, Brain, Colon, Duodenum, Esophagus, Fallopian tube oviduct, Fat adipose tissue, Gallbladder, Heart, Kidney, Liver, Lung, Lymph node, Ovary, Pancreas, Placenta, Prostate, Rectum, Salivary gland, Small intestine, Smooth muscle, Spleen, Stomach, Testis, Thyroid, Tonsil, Urinary bladder, Uterine endometrium, Vermiform appendix	Q-Exactive Plus	1,795	37	Yes	19,382	874,825	346,894	12,622
E-PROT-33	PXD005819 ³⁷	Anterior pituitary gland (adenohypophysis)	Brain	LTO-Orbitrap Velos	33	2	Yes	1,936	12,112	10,174	1,207
E-PROT-34	PXD004143 ³⁸	Dorsolateral prefrontal cortex	Brain	Thermo Orbitrap Fusion	80	4	Yes	9,929	203,798	124,355	7,633
E-PROT-35	PXD006233 ³⁹	Middle temporal lobe	Brain	Orbitrap Elite	192	6	Yes	7,030	62,455	46,786	5,264
E-PROT-36	PXD012755 ⁴⁰	Cerebellar hemispheric cortex, Occipital cortex	Brain	OrbitrapVelos Elite	15	15	No	3,033	25,514	20,125	2,264
E-PROT-40	PXD001608 ⁴¹	Colon	Colon	Q-Exactive Plus	30	10	No	5,943	73,053	56,756	4,789
E-PROT-41	PXD002029 ⁴²	Colon	Colon	Q-Exactive Plus	24	8	No	4,608	25,300	22,032	3,211
E-PROT-42	PXD000547, PXD000548 ⁴³	Corpus callosum, Anterior temporal lobe	Brain	LTO-Orbitrap XL	40, 40	2, 2	Yes	1,353, 1,873	8,250, 12,503	6,739, 10,071	868, 1,304
E-PROT-43	PXD010271 ⁴⁴	Liver, Ovary, Pancreas, Substantia nigra	Liver, Ovary, Pancreas, Brain	Velos Orbitrap, Q-Exactive	55	54	No	5,927	75,320	50,652	4,528
E-PROT-44	PXD004332 ⁴⁵	Pineal gland	Brain	LTO-Orbitrap Velos	56	3	Yes	4,953	49,455	38,884	3,692
E-PROT-45	PXD006675 ⁴⁶	Aorta, Aortic valve, Atrial septum, Inferior vena cava, Left atrium, Left ventricle, Mitral valve, Pulmonary artery, Pulmonary valve, Pulmonary vein, Right atrium, Right ventricle, Tricuspid valve, Ventricular septum	Cardiovascular system (Heart)	Q-Exactive HF	347	42	Yes	9,160	161,943	98,692	7,602
E-PROT-46 [§]	PXD008934 ⁴⁷	Left ventricle	Cardiovascular system (Heart)	Q-Exactive	7	7	No	2,977	31,755	25,704	2,294
E-PROT-51 [§]	syn6038852 ⁴⁸	Dorsolateral prefrontal cortex	Brain	Q-Exactive Plus	11	11	No	4,962	56,558	41,312	3,803
E-PROT-52 [§]	syn21444980 ²⁶	Dorsolateral prefrontal cortex	Brain	Q-Exactive Plus	83	83	No	3,846	55,290	39,470	3,127
E-PROT-53 [§]	syn7204174 ⁴⁹	Dorsolateral prefrontal cortex	Brain	Q-Exactive Plus	26	26	No	5,658	101,460	64,060	4,454
E-PROT-54 [§]	syn3606087 ⁵⁰	Dorsolateral prefrontal cortex	Brain	Q-Exactive Plus	11	11	No	4,751	58,302	40,469	3,626
E-PROT-55 [§]	syn4624471 ⁵⁰	Precuneus	Brain	Q-Exactive Plus	13	13	No	4,829	60,129	42,179	3,695
E-PROT-56 [§]	syn7431984 ⁵¹	Temporal cortex	Brain	Q-Exactive Plus	31	31	No	5,997	113,820	68,951	4,675
E-PROT-57 [§]	syn6038797 ⁵²	Frontal pole	Brain	Q-Exactive Plus	53	53	No	5,812	91,285	59,701	4,420
E-PROT-58 [§]	syn21443008 ²⁶	Dorsolateral prefrontal cortex	Brain	Q-Exactive Plus	47	47	No	5,050	86,685	55,386	4,053
E-PROT-61 [§]	PXD012131 ⁵³	Amygdala, Caudate nucleus, Cerebellum, Entorhinal cortex, Inferior parietal lobule, Middle frontal gyrus, Neocortex, Superior temporal gyrus, Thalamus, Visual cortex	Brain	Orbitrap Fusion Lumos	114	15	Yes	10,097	192,397	113,310	7,826
E-PROT-63	PXD020187 ⁵⁴	Umbilical artery	Umbilical artery	nano-ESI Orbitrap-Elite	10	10	No	1,505	10,365	7,702	933
E-PROT-65 [§]	PXD015079 ⁵⁵	Prefrontal cortex, Vermiform appendix	Brain, Vermiform appendix	Q-Exactive HF-X	6	6	No	2,313	24,162	18,892	1,694
TOTAL	24 datasets	67 tissues	31 organs		3,119 MS runs	498 Samples					

^aDataset identifiers starting with “PXD” come from the PRIDE database, and those identifiers starting by “syn” come from the AMP-AD knowledge portal. ^bOnly normal samples within this dataset are reported in this study. However, results from both normal and disease samples are available in Expression Atlas. Unique protein sample batches available in any given dataset are considered as individual samples (e.g., dataset E-PROT-34 (PXD004143) consists of four experiment batches, where materials from two donors are each digested with LysC and trypsin, and therefore these four unique batches are considered as four different samples). ^cNumbers after post-processing. The proteomics results in Expression Atlas can be accessed using the link: <https://www.ebi.ac.uk/gxa/experiments/E-PROT-XX/Results>, where XX should be replaced by the E-PROT accession number shown in the table. The raw proteomics datasets in PRIDE can be accessed using the link: <https://www.ebi.ac.uk/pride/archive/projects/PXDxxxxxx>, where PXDxxxxxx should be replaced by the PRIDE dataset identifier shown in the table.

the establishment and maturation of reliable proteomics data repositories, in tandem with policy recommendations by scientific journals and funding agencies.

The PRIDE database,⁹ which is one of the founding members of the global ProteomeXchange consortium,¹⁰ is currently the largest resource worldwide for public proteomics data deposition. As of October 2022, PRIDE hosts more than 29,500 datasets. Of those, human datasets are by far the majority, representing approximately 40% of all datasets. Public datasets stored in PRIDE (or in other resources) present an opportunity to be systematically reanalyzed and integrated in order to confirm the original results potentially in a more robust manner, obtain new insights, generate new hypotheses, and even be able to answer biologically relevant questions orthogonal to those posed in the original studies. Such integrative meta-analyses have already been successfully employed, especially in genomics and transcriptomics.^{11–13} Therefore, the large availability of public datasets has triggered different types of data re-use activities, including “big data” approaches (e.g.,^{14–16}) and the establishment of new data resources using reanalyzed public datasets as the basis.^{17–19} In this context of data re-use, the main interest of PRIDE is to disseminate and integrate proteomics data into popular added-value bioinformatics resources at the European Bioinformatics Institute (EMBL-EBI), such as Expression Atlas²⁰ (for quantitative proteomics data), Ensembl²¹ (proteogenomics), and UniProt⁷ (protein sequence information including post-translational modifications (PTMs)). The overall aim is to enable life scientists (including those who are non-experts in proteomics) to have improved access to proteomics-derived information. Expression Atlas (<https://www.ebi.ac.uk/gxa/home>) is an added-value resource that enables easy access to integrated information about gene and, recently, protein expression across species, tissues, cells, experimental conditions, and diseases. The Expression Atlas “bulk” Atlas has two sections: baseline and differential atlas. Protein abundance results derived from the reanalysis of DDA public datasets of different sources have started to be incorporated into Expression Atlas. The availability of such results in Expression Atlas makes proteomics abundance data integrated with transcriptomics information in the web interface. We have performed two DDA studies of this type so far. First of all, we reported the reanalysis and integration into the Expression Atlas of 11 public quantitative datasets coming from cell lines and human tumor samples.²² Additionally, we have recently reported the reanalysis and integration of 23 datasets coming from mouse and rat tissues in baseline conditions.²³

There are other public resources providing access to reanalyzed MS-based quantitative proteomics datasets. ProteomicsDB²⁴ provides access to human protein abundance data in addition to other recent (multi-omic) studies carried out on model organisms. Many additional human datasets derived from human tissues have been made publicly available in recent years. Within the HPP, it is important to highlight that ProteomeXchange resources PeptideAtlas²⁵ and MassIVE provide peptide and protein identifications derived from the reanalysis of public human datasets, but their main focus is not quantitative data. Additionally, antibody-based protein abundance information can be accessed via the Human Protein Atlas (HPA).⁴ Here, we report the reanalysis and integration of 24 public human label-free datasets and the incorporation of the results into Expression Atlas as baseline studies.

Experimental Procedures

Datasets. As of September 2020, 3,930 public MS human proteomics datasets were publicly available in PRIDE. We manually filtered these 3,930 human datasets to select suitable datasets for downstream analyses by applying several selection criteria. These selection criteria for the datasets to be reanalyzed were (i) experimental data from healthy tissues in baseline conditions coming from label-free studies where no PTM-enrichment had been performed; (ii) experiments performed on Thermo Fisher Scientific instruments (LTQ Orbitrap, LTQ Orbitrap Elite, LTQ Orbitrap Velos, LTQ Orbitrap XL ETD, LTQ Orbitrap XL ETD, Orbitrap Fusion, and Q-Exactive) because they represent the larger proportion of the relevant public datasets available, and we preferred to avoid the heterogeneity introduced by using data taken from different MS vendors; (iii) availability of detailed sample metadata in the original publication or after contacting the original submitters; and (iv) our previous experience in the team working with some datasets, which were discarded because they were not considered to be usable (data not shown). As a result, 16 human datasets were obtained from PRIDE (Table 1). Additionally, 8 datasets coming from human brain samples (also generated in Thermo Fisher Scientific instruments) were downloaded from a large Alzheimer’s disease (AD) dataset described in,²⁶ which was available via the AMP-AD knowledge portal (<https://adknowledgeportal.synapse.org/>). Due to ethical issues, the AD datasets from the AMP-AD knowledge portal are available under a controlled access agreement (i.e., data made available only to approved users of the data included in the AMP-AD knowledge portal) and were downloaded after obtaining the required authorization.

The sample and experimental metadata were manually curated from their respective publications or by contacting the original authors/submitters. Metadata was annotated using Annotare²⁷ and stored using the investigation description format (IDF) and sample and data relationship format (SDRF) file formats required for their integration in Expression Atlas. The IDF includes an overview of the experimental design, including the experimental factors, protocols, publication information, and contact information. The SDRF file includes sample metadata and describes the relationship between various sample characteristics and the data files included in the dataset.

In addition to the quantification of proteins in healthy tissues representing the baseline conditions described in this study, we also analyzed samples in the same datasets that were from non-healthy/non-normal samples, which were included in the same datasets (which are not discussed in this article, but the results are also available in Expression Atlas). The selected datasets are listed in Table 1, including the original dataset identifiers, tissues and organs included, number of MS runs, and number of samples. The 24 datasets sum up a total of 498 samples from 67 different tissues classified in 31 organs.

Proteomics Raw Data Processing

Datasets were analyzed separately using the same software and search database. Peptide/protein identification and protein quantification were performed using MaxQuant^{28,29} (version 1.6.3.4) on a high-performance Linux computing cluster. The input parameters for each dataset, such as MS1 and MS2 tolerances, digestive enzymes, and fixed and variable modifications, were set as described in their respective

publications, together with two missed cleavage sites. PSM (peptide spectrum match) and protein FDR (false discovery rate) levels were set at 1%. Other MaxQuant parameter settings were left as defaults: maximum number of modifications per peptide: 5, minimum peptide length: 7, maximum peptide mass: 4,600 Da. For a match between runs, the minimum match time window was set to 0.7 s, and the minimum retention time alignment window was set to 20 s. The MaxQuant parameter files are available for download from Expression Atlas. The UniProt human reference proteome release-2019_05 (including isoforms and 95,915 sequences) was used as the target sequence database. The inbuilt MaxQuant contaminant database was used, and the decoy database was generated using MaxQuant at the time of the analysis (on-the-fly) by reversing the input database sequences after the respective enzymatic cleavage. The datasets were run in a multithreading mode with a maximum of 60 threads and 300 GB of RAM per dataset.

Post-Processing

The results coming from MaxQuant for each dataset were further processed downstream to remove potential contaminants, decoys, and protein groups, which had fewer than 2 PSMs. The protein intensities were normalized using the fraction of total (FOT) method, wherein each protein's "iBAQ" intensity value is scaled to the total amount of signal in a given MS run and transformed to parts per billion (ppb).

$$\text{ppb}_{\text{iBAQ}_i} = \left(\text{iBAQ}_i / \sum_{i=1}^n \text{iBAQ}_i \right) \times 1,000,000,000$$

The bioconductor package "mygene"³⁰ was used to assign Ensembl gene identifiers/annotations to the protein groups by mapping the "majority protein identifiers" within each protein group. This step is required for integration into Expression Atlas because, at present, all abundance values have to be in the same reference system to be integrated. The protein groups, whose protein identifiers were mapped to multiple Ensembl gene IDs, were not integrated into Expression Atlas but are available in Supporting Table S1. In the case of a protein group containing isoforms from the same gene, these mapped to a single unique Ensembl gene ID and were not filtered out. In cases where two or more protein groups mapped to the same Ensembl gene ID, their median intensity values were considered. The parent genes, to which different protein groups were mapped, are equivalent to "canonical proteins" in UniProt (https://www.uniprot.org/help/canonical_and_isoforms), and therefore, the term protein abundance is used to describe the protein abundance of the canonical protein throughout the article.

Integration into Expression Atlas

The calculated canonical protein abundances (mapped as genes), validated SDRF files, and summary files detailing the quality of post-processing were integrated into Expression Atlas (release 37, March 2021) as proteomics baseline experiments (E-PROT identifiers are available in Table 1).

Protein Abundance Comparison Across Datasets

Since datasets were analyzed separately, the protein abundances, available in ppb values within each dataset, were converted into ranked bins for comparison of abundances across datasets. The normalized protein abundances per MS run, as described above, were ranked and grouped into five bins, wherein proteins with the lowest protein abundance

values were in bin 1 and those with the highest abundance values were in bin 5. Additionally, distinct tissue regions or organs within a dataset were grouped into batches and binned separately. In this study, "tissue" is defined as a distinct functional or structural region within an "organ". For example, the corpus callosum, anterior temporal lobe, and dorsolateral prefrontal cortex were defined as tissues that are part of the brain (organ), and similarly, the left ventricle, aorta, and tricuspid valve are defined as tissues in the heart (organ).

During the rank-bin transformation, if a protein was not detected in any of the samples within a batch, we did not assign it a bin value but annotated it as an NA (corresponding to not detected) value instead. However, if a protein was not detected in some samples of the batch but had protein abundance values in other samples within the batch, we assigned the lowest bin value 1 to those samples in that batch that were undetected. For example, in a dataset comprising tissue samples from the brain, all samples from tissue regions such as the corpus callosum were grouped into a batch, and the ppb abundances were transformed into bins. If any of the samples within a batch had no abundance values for a protein, they were marked as NA. If some samples within the batch had missing abundance values, the missing abundance values of those samples for that protein were assigned the bin value 1. Binned abundances of those proteins that were detected in at least 50% of the samples in the heart and brain datasets were selected for PCA (principal component analysis). To compare which normalization methods performed better at removing batch effects, the iBAQ protein abundances were also normalized using the ComBat³¹ and Limma³² methods. PCA was performed in R using the Stats package. Pearson's correlation coefficient for all samples was calculated on the basis of pairwise complete observations of bin-transformed iBAQ values in R. Samples were hierarchically clustered on columns and rows using Euclidean distances.

Comparison of the Results with the Protein Abundance Values from the Human Protein Atlas and ProteomicsDB

Results from our analysis were compared with protein abundance data available at the HPA. Abundance profiles of proteins in normal human tissues were downloaded from HPA version 21.0. Protein abundance with reliability scores labeled as "uncertain" were not considered in the comparison. For the purposes of easing the comparison and computing correlation, the categorical protein abundance levels in data downloaded from the HPA were assigned numerical values closely matching the protein abundance bins used in our analysis. Protein abundance levels annotated as "low", "medium", and "high" were assigned values 1, 2, and 3, respectively. The level annotated as "not detected" was assigned NA, and levels annotated as "ascending", "descending", and "not representative" were all assigned a value of 1. For the purpose of this comparison, we re-binned our protein abundance data into just three categories: bins 1, 2, and 3, representing low, medium, and high abundances, respectively. The "randomized edit distance difference" was calculated across all pairs of organs included in this study and HPA. The "randomized edit distance difference" is the difference between the "true edit distance" and the "randomized edit distance" of protein abundance bins. Randomized edit distance difference = $\text{mean}(\text{random edit distance}_{1-n} - \text{true edit distance}_{1-n})$. The "true edit distance" of a protein was computed as the absolute difference between the protein abundance bins of both pairs.

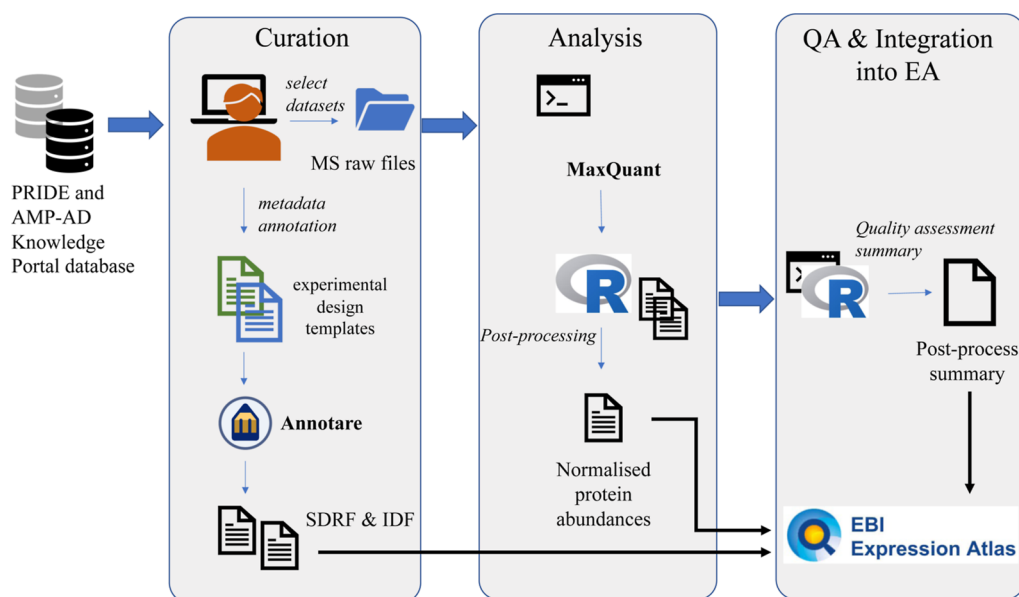


Figure 1. Overview of the study design and reanalysis pipeline. QA: Quality assessment. Reprinted (Adapted or Reprinted in part) with permission from .²⁰ Copyright 2022 EMBL-EBI.

The “randomized edit distance” is calculated as the mean of the absolute difference between the bin value of pair 1 and the randomized bin value of pair 2, after sampling it 10 times, that is, the randomized edit distance = $\text{mean}(\sum_1^{10} [\text{abs}(\text{bin1} - \text{random}(\text{bin2}))])$. This was done using the base *R* package.

Normalized protein intensities from ProteomicsDB³³ were queried for organs that were common in our study (31 organs). Values were obtained using the ProteomicsDB application programming interface. For different tissue samples, we aggregated the normalized intensities using the median of their respective organs. The intensities were log₂ normalized and compared.

Comparison of Label-free Protein Abundances with Protein Abundances Generated Using a TMT Approach

The protein abundances calculated across various baseline human organs/tissues using the TMT-labeling method were obtained from³ (Supporting file “NIHMS1624446-supplement-2”, sheet: “C protein normalized abundance”). Protein abundances of the respective organs measured across different TMT channels and runs were aggregated using the median and log₂ transformed. Different tissue samples from the esophagus, heart, brain, and colon were aggregated into their respective organs. Pearson’s correlation was calculated in *R*.

Organ-Specific Expression Profile Analysis

To investigate the organ-specific protein-based abundance profile, we carried out a modification of the classification scheme done by Uhlén et al.⁴ Briefly, each of the 13,070 canonical proteins that were mapped from the protein groups was classified into one of three categories based on the bin levels in 31 organs: (1) “organ-enriched”: one unique organ with bin values twofold higher than the mean bin value across all organs; (2) “group enriched”: a group of 2–7 organs with bin values twofold higher than the mean bin value across all organs; and (3) “mixed”: the remaining canonical proteins that are not part of the above two categories.

Enriched gene ontology (GO) term analysis was performed by means of the over-representation test, combining the

“organ-enriched” and “group-enriched” mapped gene lists for each organ. The computational analysis was carried out in the *R* environment with the package clusterProfiler³⁴ version 3.16.1 using the function enrichGO() for the GO over-representation test using the parent gene list of all detected canonical proteins as the background set. Setting the *p*-value cut-off to 0.05 and the *q*-value cut-off to 0.05. Additionally, reactome³⁵ pathway analysis was carried out by using mapped gene lists (indicated by the protein groups) and running pathway topology and over-representation analysis. First, the “project to human” option was selected with the combining list of “organ-enriched” and “group-enriched” entities. Afterward, those pathways with a *p*-value >0.05 were filtered out. The hierarchical clustering was done based on the distances calculated on the *p*-values using the gg dendro package in *R*.

RESULTS

Human Baseline Proteomics Datasets

We manually selected 24 label-free publicly available human proteomics datasets coming from PRIDE and from the AMP-AD knowledge portal databases (Table 1). These datasets were selected to represent baseline conditions and therefore included samples annotated as healthy or normal from a wide range of biological tissues. The datasets were restricted to include those label-free datasets generated on Thermo Fisher Scientific Instruments. See more details about dataset selection in the “Methods” section.

In total, the aggregated datasets represent 67 healthy tissues, corresponding to 3,119 MS runs covering 498 samples, coming from 489 individuals. In this study, “tissue” is defined as a distinct functional or structural region within an “organ”. The cumulative CPU time used for the reanalyses was approximately 2,750 h or 114 calendar days. The numbers of protein groups, peptides, and unique peptides identified and protein coverage in each dataset are shown in Table 1.

The resulting protein abundances of all samples have been made available in Expression Atlas. These “proteomics baseline” quantification results can be viewed as abundance

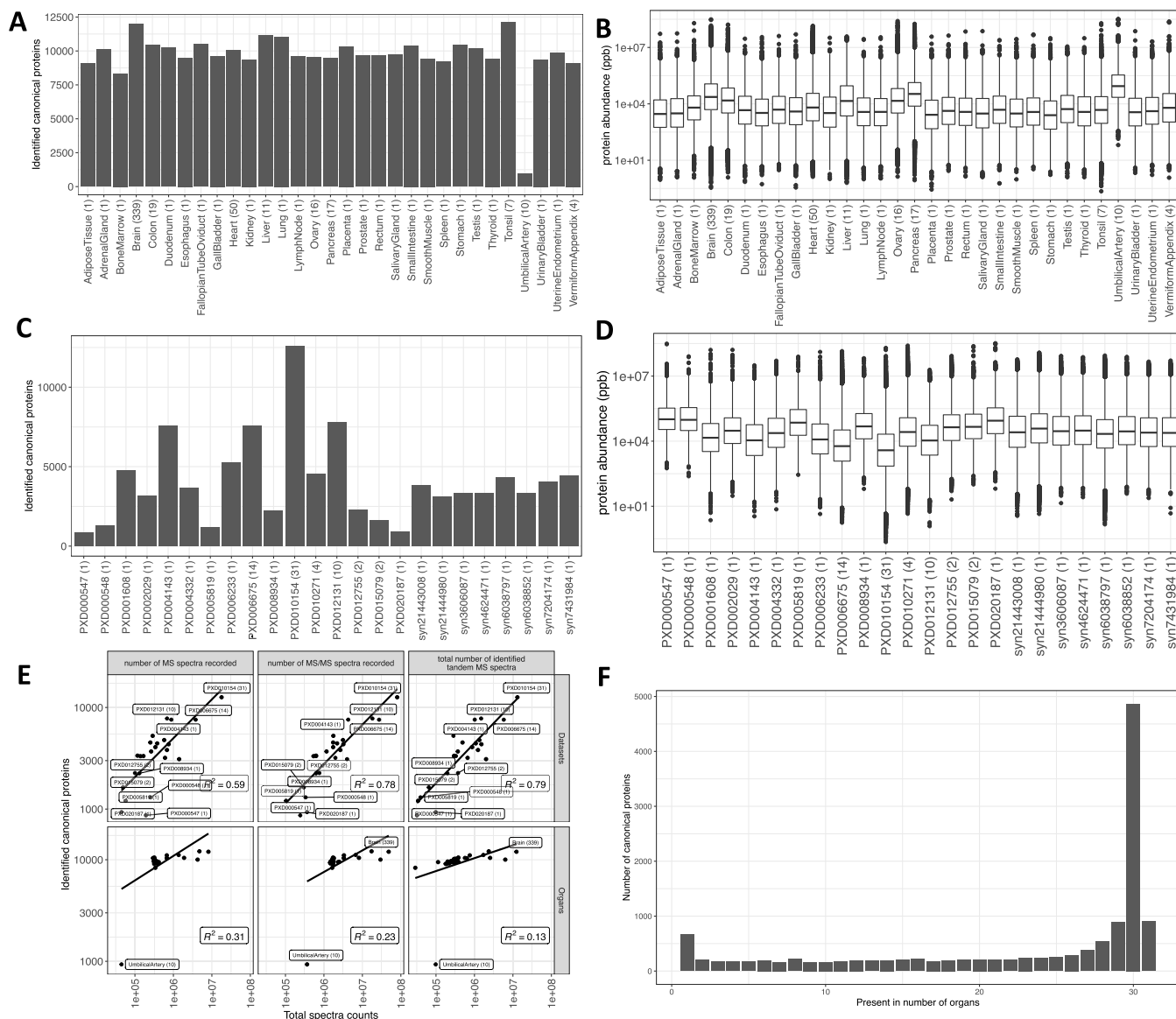


Figure 2. (A) Number of canonical proteins identified across different organs. The number within the parenthesis indicates the number of samples. (B) Range of normalized iBAQ protein abundances across different organs. The number within the parenthesis indicates the number of samples. In panels (A) and (B), the term heart is used in a broader sense to mean the cardiovascular system. (C) Canonical proteins identified across different datasets. The number within the parenthesis indicates the number of unique tissues in the dataset. (D) Range of normalized iBAQ protein abundances across different datasets. The number within the parenthesis indicates the number of unique tissues in the dataset. (E) Comparison of total spectral data with the number of canonical proteins identified in each dataset and organ. (F) Distribution of canonical proteins identified across organs.

heatmaps against the gene symbols, and the quantification matrices can be downloaded as text files together with annotated metadata of donor samples, experimental parameters, and a summary file describing the analysis with representative charts (quality assessment) summarizing the output of the post-processed samples. The protocol for data reanalysis is summarized in Figure 1.

Protein Coverage Across Samples

For simplicity of comparison, we broadly grouped 67 tissues into 31 major types of organs. As explained in “Methods”, we defined “tissue” as a distinct functional or structural region within an “organ”. For example, the corpus callosum, anterior temporal lobe, and dorsolateral prefrontal cortex were all defined as tissues in the brain (which is the “organ”). After

post-processing the output files from MaxQuant, 11,653 protein groups (36.3% of identified protein groups across all datasets) were uniquely present in only one organ and 380 protein groups (1.2%) were ubiquitously observed (Supporting Table S2). This does not imply that these proteins are unique to these organs. Merely, this is the outcome considering the selected datasets.

We mapped the isoforms in the protein groups to their respective parent gene names, which we will use as equivalent to “canonical proteins” in UniProt (see “Methods”) from now on in the article. Overall, 13,070 different genes were mapped from protein identifiers in the protein groups. We denote the term “protein abundance” to mean “canonical protein abundance” from here on. We then estimated the number of proteins identified across organs, which indicated that greater

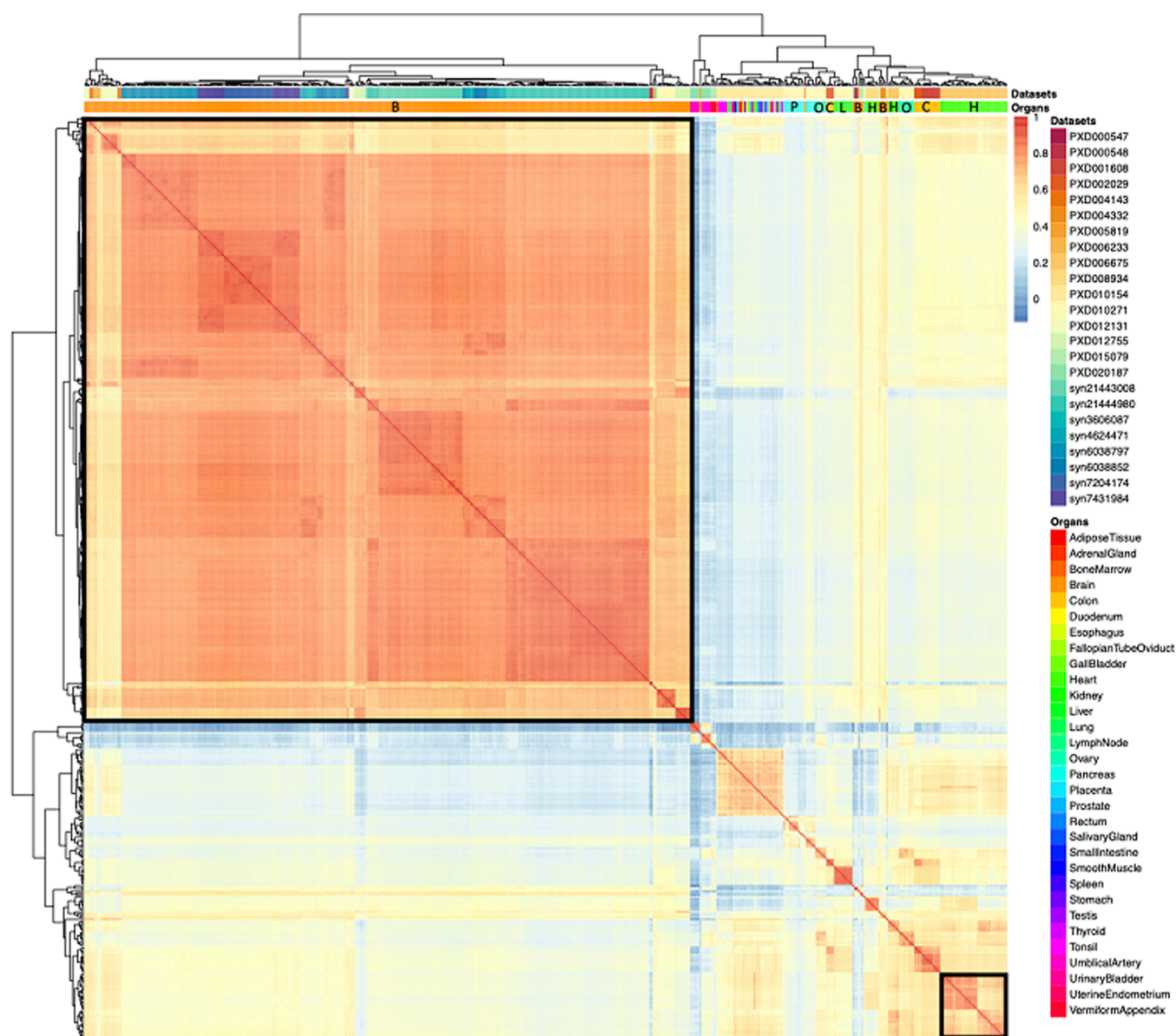


Figure 3. Heatmap of pairwise Pearson correlation coefficients across all samples. The color on the heatmap represents the correlation coefficient, which was calculated using the bin-transformed iBAQ values. The samples are hierarchically clustered on columns and rows using Euclidean distances. The clusters composed of the brain and cardiovascular system (heart) samples are highlighted with black borders. The abbreviations used in the organs' header are B: brain, C: colon, H: heart, L: liver, O: ovary, and P: pancreas.

than 70% of all canonical proteins were present in a majority of organs (Figure 2A,C). We also observed the highest numbers of common proteins in samples from the tonsil (92.2%) and brain (90.9%) and the lowest numbers in samples from the umbilical artery (7.2%).

The higher number of proteins identified in the brain could be attributed to the greater representation of samples (339 samples out of 498, 68.0%). However, tonsils were represented only by seven samples, all of which were derived from one dataset (PXD010154). It is worth noting that the sample preparation protocol for the tonsil samples employed seven different proteases (trypsin, LysC, ArgC, GluC, AspN, LysN, and chymotrypsin) for tissue digestion,³⁶ thus significantly increasing its peptide coverage.³⁶ The sample size of the umbilical artery, which showed significantly lower protein coverage than other organs, was 10 samples.

The largest number of canonical proteins were identified in dataset PXD010154 (Figure 2C), which comprises numerous tissue samples (31 tissues) including samples from tonsils. The dynamic range of protein abundances in all organs is shown in Figure 2B. On the other hand, protein abundances among datasets showed that PXD010154 had the lowest median protein abundances (Figure 2D). We also compared the quantity of spectral data from various organs and datasets with the number of canonical proteins identified in them to detect any organ or dataset that showed enrichment of proteins relative to the amount of data. We observed a linear relation between the number of proteins identified and the amount of spectral data present in the organ samples or datasets (Figure 2E).

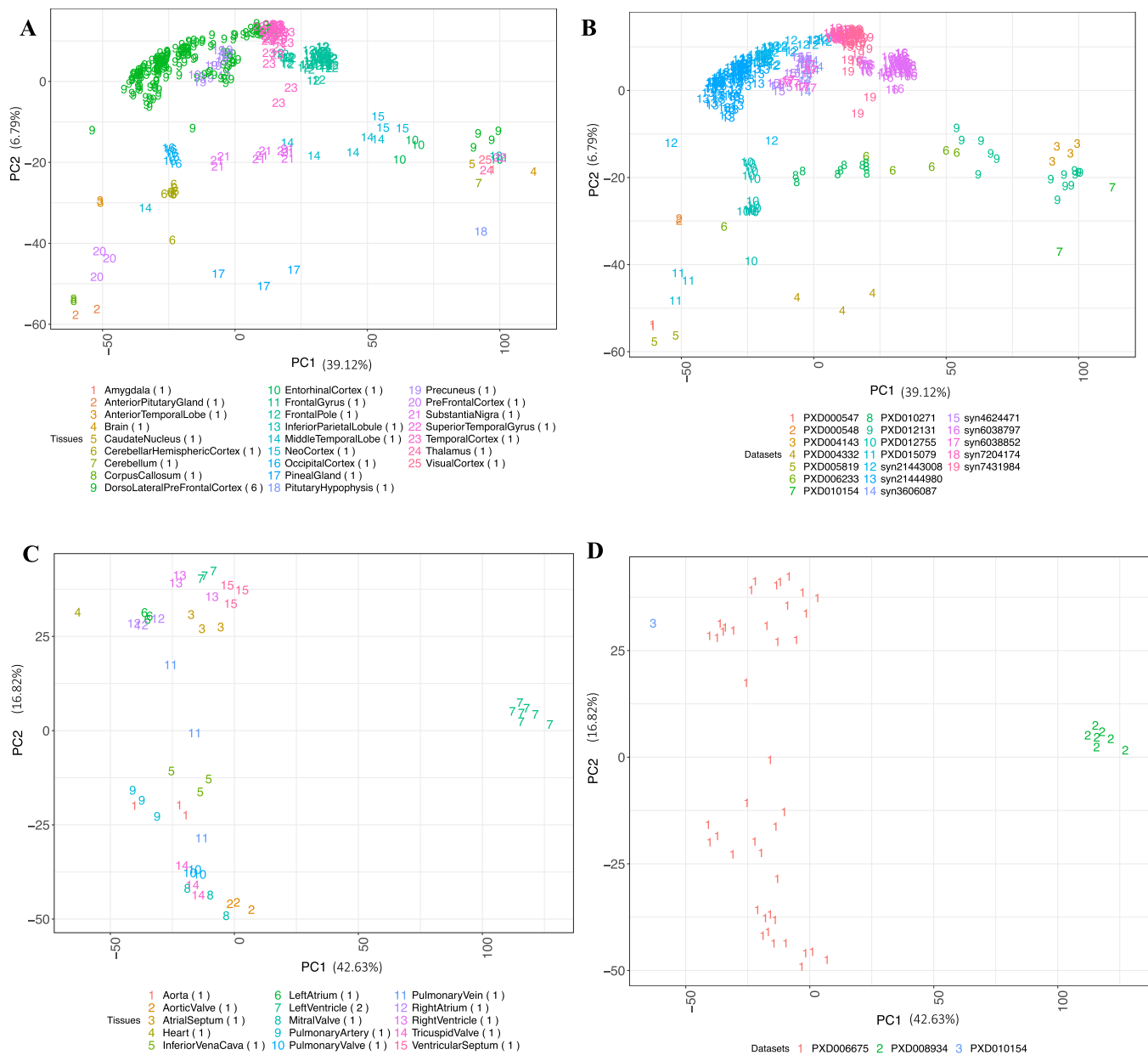


Figure 4. (A) PCA of brain samples colored by the tissue types. (B) PCA of brain samples colored by their respective dataset identifiers. (C) PCA of cardiovascular system (heart) samples colored by the tissue types. (D) PCA of cardiovascular system (heart) samples colored by their respective dataset identifiers. The numbers in parenthesis indicate the number of datasets for each tissue. Binned values of canonical proteins quantified in at least 50% of the samples were used to perform the PCA.

Distribution of Canonical Protein Identifications per Organ

We observed that 37.1% (4,853) of the identified canonical proteins were expressed in 30 different organs (Figure 2F). The low number of proteins identified in umbilical artery (933) samples greatly influenced the protein distribution. As a result, 7.0% (917) of all identified canonical proteins were present in all 31 organs, whereas 4.2% (565) of the identified canonical proteins were uniquely present in one organ. However, it is important to highlight that the list of concrete canonical proteins that were detected in just one organ should be taken with caution since the list is subjected to an inflated FDR due to the accumulation of false positives when analyzing the datasets separately. However, this should not be an issue in the case of proteins detected across five datasets or more since

the number of commonly detected decoy protein hits enabled us to calculate a protein FDR less than 1% (Figure S1 in Supporting Figures).

Protein Abundance Comparison Across Organs

Next, we compared the protein abundances to see how proteins compared across different organs. Inter-dataset batch effects make comparisons challenging. We transformed the normalized iBAQ intensities into ranked bins as explained in "Methods". The bin-transformed protein abundances in all organs are provided in Supporting Table S3.

To compare protein abundance across all organs, a pairwise Pearson correlation coefficient of binned protein abundances was calculated across 498 samples (Figure 3). We observed a good correlation of protein abundance within the brain (median $R^2 = 0.61$) and cardiovascular system (median $R^2 =$

0.41) samples, which represent the two organ groups with the largest number of samples. We tested the effectiveness of various normalization methods in reducing batch effects by performing a PCA on samples coming from the cardiovascular system and brain datasets. The analyzed brain and cardiovascular system samples constituted the largest numbers in the aggregated dataset, including 19 and 3 datasets, respectively. First, we performed PCA on the normalized iBAQ values, wherein the brain samples did not cluster either by tissues or by datasets. However, for cardiovascular system samples, we observed clustering of samples by datasets and not by tissue type (Figure S2 in Supporting Figures). We then tested the ComBat and Limma normalization methods on iBAQ values, which neither showed clustering of samples by tissues nor by datasets for both cardiovascular system and brain samples (Figures S3 and S4 in Supporting Figures).

We then decided to use the bin-transformed protein abundances (see “Methods”). First, we observed that brain samples were clustered together according to their tissue type (Figure 4A). All brain tissue samples, except those coming from the dorsolateral prefrontal cortex (DLPFC), were part of individual datasets. The DLPFC samples were derived from six separate datasets, of which five of them were part of the Consensus Brain Protein Coexpression study.²⁶ The DLPFC samples clustered into two groups: a large group that comprised samples from the Consensus Protein Coexpression study and a smaller cluster with samples from dataset PXD004143 (Figure 4B), indicating that there was still a residual batch effect.

Similarly, we observed cardiovascular system samples clustered according to their tissue types (Figure 4C). All cardiovascular system samples except those coming from the left ventricle were part of an individual dataset. Interestingly, we observed three major clusters: one wherein all valve samples (aortic valve, mitral valve, pulmonary valve, and tricuspid valve) were clustered together. A second cluster was formed where the samples from the ventricles and atriums were clustered in a large group together with other cardiovascular system samples. Finally, left ventricle samples from dataset PXD008934 (Figure 4D) formed a separate cluster, indicating that there were still batch effects, which were not completely removed.

Comparison of Protein Abundance Values with Previous Studies

We first compared the protein abundances resulting from our reanalysis with those reported in the original publications. By comparing the number of protein groups or genes identified in individual datasets, we observed that the differences between our analysis and the original published results ranged from as low as 1.3% (E-PROT-53, dataset syn7204174) to as high as 43.2% (E-PROT-36, dataset PXD012755). Similarly, the difference at the level of identified peptides ranged from a minimum of 0.29% (E-PROT-33, dataset PXD005819) to a maximum of 57.2% (E-PROT-36, dataset PXD012755) (Supporting Table S4). These differences in overall numbers could be due to various factors, including the target protein sequence database, the analysis software, and the version used.

We then compared our results with protein abundance data available in ProteomicsDB 33 and found a good correlation in abundance across various organs. As it can be seen in Figure S5 in Supporting Figures, the highest correlation was found in the salivary gland ($R^2 = 0.75$) and the lowest one in the ovary (R^2

$= 0.52$). However, it should be noted that one of the datasets included in our analysis (dataset PXD010154) is also included in ProteomicsDB. Additionally, we also made a comparison between our protein abundance results and those found in a large study across multiple human organs using TMT-labeling method 3. Figure S6 in Supporting Figures shows the Pearson's correlation of protein abundances between both studies, which was generally lower than in the case of ProteomicsDB data, ranging from 0.22 to 0.48 across various organs.

In addition, we compared our results with protein abundances computed using antibody-based methods, available in the Human Protein Atlas (HPA). First, we performed a qualitative analysis in which we compared the number of proteins identified in matching organs in our analysis with those identified in the HPA. There were 30 organs that were common between both studies (except for the umbilical artery, which was not available in HPA). The comparison results are shown in Figure S7 in Supporting Figures. Our analysis shows that an average of 43.7% of all proteins identified in HPA were also present in our aggregated dataset, with the highest number of commonly identified proteins found in the brain (50.4%) and the lowest number of common proteins was found in adipose tissue (27.2%). On the other hand, an average of 40.4% of proteins were only identified in our analysis and were not present in the results analyzed in HPA. The largest and the lowest number of proteins that were identified only in our analysis were in adipose tissue (61.6%) and in testes (30.2%), respectively. Lastly, an average of 15.8% of the proteins were exclusive to HPA and not identified across any organs in our analysis. Of these proteins, the largest HPA-exclusive group was present in the vermiform appendix (21.6%), and the lowest was found in the adrenal gland (8.9%).

We then compared protein abundances by first transforming the abundances in HPA into numerical bins. Protein abundance data from HPA are annotated in three categorical groups as “low”, “medium”, and “high”, which we converted into three numerical bins 1, 2, and 3, respectively. For the purpose of this comparison, we re-binned our protein abundance data into just three categories: bins 1, 2, and 3, representing low, medium, and high abundances, respectively (see “Methods”). To identify the difference between noise and signal, we calculated the randomized edit distance difference metric across all organs between the two studies (see Methods). The higher “randomized edit distance difference” indicates that there is a difference between signal and random noise. The randomized edit distance difference matrix (Figure S8 in Supporting Figures) shows that the randomized edit distance difference between organs within HPA is low (average randomized edit distance difference = 0.18) compared to that of the organs within our study (average randomized edit distance difference = 0.43). This seems to suggest that the overall protein abundances generated in this study are less noisy than the abundance data available in HPA.

Organ-Elevated Proteome and the Over-Representative Biological Processes

As explained in “Methods”, according to their abundances, canonical proteins were divided in three different groups according to their organ specificity: “organ-enriched”, “group-enriched”, and “mixed” (see Supporting Table S5). We considered elevated canonical proteins, which were classified as an “organ-enriched” or “group-enriched” instead of the “mixed” group. The analysis (Figure 5A) showed that, on

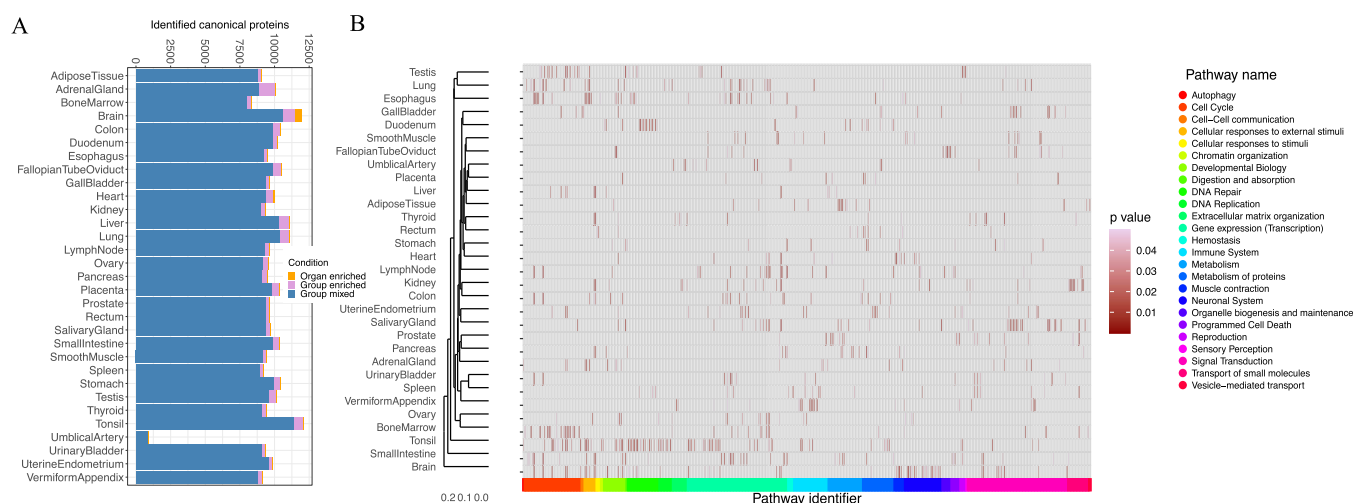


Figure 5. (A) Analysis of organ-specific canonical proteins. The analysis comprises the number of canonical proteins found in 31 organs, classified in three groups: “organ-enriched”, “group-enriched”, and “group mixed”. (B) Pathway analysis of the over-represented canonical proteins, showing the statistically significant representative pathways (p -value < 0.05) in 31 organs. In panels (A) and (B), the term heart is used in a broader sense to mean the cardiovascular system.

average, 3.8% of the total elevated canonical proteins were organ group-specific. The highest ratio was found in the adrenal gland (9.3%), brain (7.5%), and liver (7.1%), and the lowest ratio was found in the gall bladder (2.3%) and umbilical artery (0.1%). In addition, 0.4% of the total canonical proteins were uniquely organ enriched. The highest ratio was found in the brain (3.8%), cardiovascular system (1.4%), and liver (0.5%), and the lowest ratio ($\sim 0.1\%$) was found in the tonsil and uterine endometrium.

Then, we performed a gene ontology (GO) enrichment analysis using the GO terms related to biological processes for those canonical proteins that were “organ-enriched” and “group-enriched”, as shown in Table 2. As a summary, 358 GO terms were found to be statistically significant across all organs (see Supporting Table S6). The terms found were in agreement with the known functions of the respective organs. The brain had the largest number of “organ-enriched” canonical proteins (457), and among the biological processes associated with them stand out the regulatory function on membrane potential (GO:0042391), neurotransmitter transport (GO:0006836), modulation of chemical synaptic transmission (GO:0050804), regulation of *trans*-synaptic signaling (GO:0099177), and potassium ion transport (GO:0006813). The second organ with a greater number of “organ-enriched” canonical proteins was the cardiovascular system (137). The enriched biological processes involved were related to striated muscle cell differentiation (GO:0051146), sarcomere organization (GO:0045214), muscle structure development (GO:0061061), and regulation of myotube differentiation (GO:0010830). As expected, there were common GO terms that were shared between the organs, such as detoxification of inorganic compounds (GO:0061687) in the liver and kidneys, import across the plasma membrane (GO:0098739) in the kidney, brain, and umbilical artery, and processes involved in tissues with high cell division turnover like chromosome segregation (GO:0007059) in the bone marrow and testis.

Next, we performed a pathway-enrichment analysis using reactome³⁵ to analyze canonical proteins that were “organ-enriched” and “group-enriched” (see Supporting Table S7). The heatmap (Figure 5B) shows statistically significant

Table 2. Analysis of the GO Terms for Each Organ Using the Elevated Organ-Specific Canonical Proteins and Group-Specific Ones, as Described in the “Methods” Section

organ	GO ID	description	adjusted p -value
adrenal gland	GO:0031649	heat generation	7.94×10^{-4}
bone marrow	GO:0034080	CENP-A containing nucleosome assembly	1.49×10^{-3}
brain	GO:0042391	regulation of membrane potential	1.71×10^{-11}
fallopian tube oviduct	GO:0044782	cilium organization	2.88×10^{-48}
gallbladder	GO:0017158	regulation of calcium ion-dependent exocytosis	1.17×10^{-2}
cardiovascular system (heart)	GO:0051146	striated muscle cell differentiation	9.58×10^{-6}
kidney	GO:0046942	carboxylic acid transport	4.81×10^{-16}
liver	GO:0097501	stress response to metal ion	1.08×10^{-3}
lung	GO:0003002	regionalization	8.83×10^{-3}
lymph node	GO:0002250	adaptive immune response	1.93×10^{-4}
ovary	GO:0008544	epidermis development	8.42×10^{-7}
placenta	GO:0044706	multi-multicellular organism process	4.84×10^{-3}
testis	GO:0048232	male gamete generation	5.00×10^{-24}
thyroid	GO:0098742	cell–cell adhesion via plasma-membrane adhesion molecules	1.34×10^{-2}
tonsil	GO:0031424	keratinization	3.20×10^{-5}
umbilical artery	GO:0001937	negative regulation of endothelial cell proliferation	1.75×10^{-3}

pathways (p -value < 0.05) across the organs. The total number of pathways found in all the organs was 928, and the largest number of pathways was found in the brain with 67 pathways. The pathways found were consistent with the GO analysis and with the expected function in each organ. We observed a “cell cycle” cluster of over-represented pathways related to the bone

marrow and testis (R-HSA-1640170, R-HSA-69620, R-HSA-73886, R-HSA-2500257, and R-HSA-69618), expected in high cell turnover tissues, the digestion pathway (R-HSA-192456) in the pancreas and stomach, a neuronal system cluster of pathways (R-HSA-112316) in the brain, and pathways related to the transport of small molecules (R-HSA-382551, R-HSA-425407, R-HSA-425393, and R-HSA-425366) in the kidneys.

Integration of Results into Expression Atlas

Protein abundance results from label-free experiments across various tissues were integrated into Expression Atlas. The abundances of each protein are represented in terms of their canonical gene symbols since Expression Atlas is designed as a gene-centric resource. Proteomics results can be accessed using the link www.ebi.ac.uk/gxa/experiments/E-PROT-xx/Downloads (replacing xx with the corresponding identifier for each dataset). For each dataset, the raw, unprocessed MaxQuant output files (proteinGroups.txt) are made available to download together with the input experimental parameters (mqpar.xml) to MaxQuant, as well as the metadata annotation file of each sample. We also provide a summary of the quality assessment of the results. Supporting Figures S9–S12 provide a brief manual on how to access proteomics data in Expression Atlas.

DISCUSSION

We here include a combined analysis of human baseline proteomics datasets representing baseline protein abundance across 67 healthy tissues grouped in 31 organs. This type of study has been enabled by the large amount of data in the public domain, as the proteomics community is now embracing open data policies. The large-scale availability of MS data in public databases such as PRIDE enables integrated metaanalyses of proteomics data covering a wide array of tissues and biological conditions. The main aim of our study was to provide a system-wide baseline protein abundance catalog across various tissues and organs, which could be used as a reference (especially to those non-experts in proteomics) and help to reduce redundant efforts of similar computationally expensive reanalyses.

Unlike what was done in one previous study performed by us,²² and analogously to what we did with a more recent study performed using data generated from baseline rat and mouse tissues,²³ here we analyzed each dataset separately using the same software and the same search protein sequence database. The disadvantage of this approach is that the FDR statistical thresholds are applied at a dataset level and not to all datasets together as a whole, with the potential accumulation of false positives across datasets. However, this does not represent an issue in the case of proteins detected in several datasets (in this particular study, at least five datasets will provide a protein FDR of less than 1%, Figure S1 in Supporting Figures), since the number of commonly detected false positives is reduced in parallel with the increase in the number of common datasets where a given protein is detected. This means that proteins that are only detected in a small number of datasets could potentially be false positives (considering the applied 1% FDR at the protein level), but that does not mean that they are. At that point, researchers should seek confirmation of the existence of the protein (if that is their goal) via alternative sources as well. Different reanalyses of some of the datasets used in this study, with different FDR calculation methods, have been published independently.^{56,57}

In our view, the objective of integrating quantitative proteomics information with other omics data types (in this case, transcriptomics) in resources used by non-proteomics researchers such as Expression Atlas is only feasible in a sustainable manner using a dataset per dataset analysis approach, at least at present. This enables that (i) the computing requirements for the reanalyses are realistic, given the large volume of files included in the potentially very large-combined datasets; (ii) interesting additional datasets could be added at a different time point without having to reanalyze all datasets together again; (iii) future updates in the results are more feasible to perform; and (iv) (semi)-automation of the reanalyses is achievable, making these efforts more sustainable again. As mentioned above, we followed this same overall approach in the recent study that we performed in mouse and rat tissues in baseline conditions.²³ Additionally, we compared our results with previous analogous studies performed in baseline tissues using MS and also the antibody-based data available in the HPA. These comparisons generated quite different results depending on each study.

One of the major bottlenecks was, as reported before, the curation of dataset metadata, consisting of mapping files to samples and biological conditions. Detailed sample and donor metadata is crucial for result reproducibility, and we found detailed metadata available in PRIDE for just a handful of datasets. The required information was either inferred or requested by contacting the respective study's authors. If no responses were obtained, such datasets could not be considered for the reanalysis. Therefore, to aid the reproducibility of results in the future, we need to improve the provision of metadata by data submitters. A format to enable that has been developed (the SDRF-Proteomics format, as part of the new MAGE-TAB-Proteomics format), which can be submitted optionally to PRIDE.⁵⁸ We expect that it will become increasingly used for data submissions to PRIDE once the right tooling is available and submitters have been educated appropriately.

Another one of the major challenges in the reanalysis of a large number of proteomics datasets is the integration of results from different datasets since batch effects are inevitable. We used a rank-binned normalization of abundances, which transformed protein abundances across datasets and samples to bins of 1 to 5. This approach is useful for reducing batch effects, although we acknowledge that there is also a loss of signal through this transformation. We also acknowledge that this method is not ideal in all circumstances, but in our view, it generally works better when compared to popular methods to reduce batch effects, such as ComBat and Limma. Since our method computes protein abundances in terms of their canonical protein and gene identifiers, we acknowledge that using the median of intensities to aggregate abundances over protein groups with isoforms coming from the same canonical protein may not represent the total sum of all proteins and may influence ranking during binning.

Although the combined dataset contains a higher representation of particular tissues (especially brain), we believe it represents the current state of the art with regard to public baseline human proteomics studies carried out in tissues. The analysis search strategy used in this study focused only on detecting known coding protein sequences using the UniProt reference proteome, in the same way as performed in the original studies. Therefore, it was not possible to detect any single amino acid variants or equivalent isobaric combinations

involving PTMs. However, the effect of this limitation on the analysis should, in our view, be relatively small because the type of samples used in this study (healthy tissues) did not involve, for example, tumor samples. The availability of the results through Expression Atlas enables the integration of mRNA and proteomics abundance information, offering an interface for researchers to access this type of information. One possible next step will be the integration of datasets in the differential part of Expression Atlas. The work required there would be more complex at different levels, including the downstream statistical differential analysis. Also, the availability of mapping between the channels (e.g., in TMT and SILAC experiments) and samples is very rare at present. In parallel, work has also started on integrating in Expression Atlas quantitative proteomics data generated using data independent acquisition (DIA) approaches.⁵⁹

The generated baseline protein abundance data can be used for different purposes. For instance, quantitative proteomics data can be used for the generation of co-expression networks and/or the inference of protein complexes. Protein abundance data could also be used to potentially refine the recently developed AlphaFold-based protein complex predictions.⁶⁰ Additionally, it is possible to use artificial intelligence approaches to impute protein abundance values using calculated abundance values as training data.⁶¹ It would also be possible to perform expression correlation studies between the gene and protein expression information. However, this type of study can only be performed optimally if the same samples are analyzed by both techniques, as reported in the original publication for dataset PXD010154.³⁶ It should also be highlighted that a growing number of studies are using non-MS-based proteomics techniques, such as the use of affinity reagents (e.g., the Olink and SomaLogic platforms), due to the increased throughput that they can provide. Initial studies are being performed to compare these with MS approaches.

In conclusion, the results presented here represent a large-scale meta-analysis of public human baseline proteomics datasets. We also show the challenges in this kind of analysis, providing a roadmap for such future studies.

DATA AVAILABILITY

Expression Atlas E-PROT identifiers and PRIDE and AMP-AD original dataset identifiers are included in Table 1.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00406>.

Supporting figures of ComBat and Limma normalization methods; comparison of the results with previous studies; and tutorial on how to browse proteomics abundance data in Expression Atlas (PDF)

Protein groups from all datasets that are mapped to more than one Ensembl Gene ID; median protein abundances (in ppb) for each protein group across various tissue samples in each organ; median binned protein abundances across various organs; comparison of protein and peptide identification numbers across various datasets with the reported ones in their respective original publications; ‘organ-enriched’ and ‘group-enriched’ elevated proteomes in various organs; gene ontology enrichment analysis of ‘organ-enriched’

and ‘group-enriched’ proteins; and reactome pathway-enrichment analysis of “organ-enriched” and “group-enriched” proteins (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Ananth Prakash – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom; Open Targets, Hinxton, Cambridge CB10 1SD, United Kingdom; Email: ananth@ebi.ac.uk

Andrew R. Jones – Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom; orcid.org/0000-0001-6118-9327; Email: jonesar@liverpool.ac.uk

Juan Antonio Vizcaíno – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom; Open Targets, Hinxton, Cambridge CB10 1SD, United Kingdom; orcid.org/0000-0002-3905-4335; Email: juan@ebi.ac.uk

Authors

David García-Seisdedos – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Shengbo Wang – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Deepthi Jaiswal Kundu – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Andrew Collins – Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom

Nancy George – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Pablo Moreno – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

Irene Papatheodorou – European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, United Kingdom; Open Targets, Hinxton, Cambridge CB10 1SD, United Kingdom

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00406>

Author Contributions

A.P., D.G.S., S.W., and D.J.K. selected and curated the datasets. A.P., D.G.S., and S.W. performed analyses. A.C. and A.J. helped in the interpretation of results and designed approach for data normalization. N.G., P.M., and I.P. helped integration of results into Expression Atlas. A.P., D.G.S., A.R.J., and J.A.V. wrote the manuscript. All authors have read and approved the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

First of all, we would like to thank all data submitters who made their datasets available in the public domain (most of the datasets in PRIDE). This work has been funded by Open Targets (project OTAR-043), Wellcome Trust [grant number 208391/Z/17/Z], BBSRC [BB/T019670/1 and BB/T019557/1], and EMBL core funding. We thank Thawfeek Varusai for helping with the pathway analysis using Reactome. We are very grateful to Mathias Walzer and Yasset Perez-Riverol for their useful suggestions and discussions.

ABBREVIATIONS

AD, Alzheimer's disease; DLPFC, dorsolateral prefrontal cortex; FOT, fraction of total; HPA, Human Protein Atlas; GO, gene ontology; iBAQ, intensity-based absolute quantification; IDF, investigation description format; MS, mass spectrometry; PCA, principal component analysis; SDRF, sample and data relationship format

REFERENCES

- (1) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4*, 1419–1440.
- (2) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581.
- (3) Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; Consortium, G. T.; et al. A Quantitative Proteome Map of the Human Body. *Cell* **2020**, *183*, 269–283.e19.
- (4) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419.
- (5) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582–587.
- (6) Adhikari, S.; Nice, E. C.; Deutsch, E. W.; Lane, L.; Omenn, G. S.; Pennington, S. R.; Paik, Y. K.; Overall, C. M.; Corrales, F. J.; Cristea, I. M.; et al. A high-stringency blueprint of the human proteome. *Nat. Commun.* **2020**, *11*, 5301.
- (7) UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.
- (8) Zahn-Zabal, M.; Michel, P. A.; Gateau, A.; Nikitin, F.; Schaeffer, M.; Audot, E.; Gaudet, P.; Duek, P. D.; Teixeira, D.; Rech de Laval, V.; et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* **2020**, *48*, D328–D334.
- (9) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47*, D442–D450.
- (10) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrar, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32*, 223–226.
- (11) Rung, J.; Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **2013**, *14*, 89–99.
- (12) Madrid, L.; Labrador, S. C.; González-Pérez, A.; Sáez, M. E. The Alzheimer's Disease Neuroimaging Initiative, A. Integrated Genomic, Transcriptomic and Proteomic Analysis for Identifying Markers of Alzheimer's Disease. *Diagnostics (Basel)* **2021**, *11* (12), No. 2303.
- (13) Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinf. Biol. Insights* **2020**, *14*, 117793221989905.
- (14) Ochoa, D.; Jarnuczak, A. F.; Viéitez, C.; Gehre, M.; Soucheray, M.; Mateus, A.; Kleefeldt, A. A.; Hill, A.; Garcia-Alonso, L.; Stein, F.; et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **2020**, *38*, 365–373.
- (15) Tiwary, S.; Levy, R.; Gutenbrunner, P.; Salinas Soto, F.; Palaniappan, K. K.; Deming, L.; Berndl, M.; Brant, A.; Cimermanic, P.; Cox, J. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **2019**, *16*, 519–525.
- (16) Gabriels, R.; Martens, L.; Degroove, S. Updated MS(2)PIP web server delivers fast and accurate MS(2) peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.* **2019**, *47*, W295–W299.
- (17) Kustatscher, G.; Grabowski, P.; Schrader, T. A.; Passmore, J. B.; Schrader, M.; Rappsilber, J. Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **2019**, *37*, 1361–1371.
- (18) Ramasamy, P.; Turan, D.; Tichshenko, N.; Hulstaert, N.; Vandermarliere, E.; Vranken, W.; Martens, L. Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context. *J. Proteome Res.* **2020**, *19*, 3478–3486.
- (19) Brunet, M. A.; Lucier, J. F.; Levesque, M.; Leblanc, S.; Jacques, J. F.; Al-Saedi, H. R. H.; Guillo, N.; Grenier, F.; Avino, M.; Fournier, I.; et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.* **2021**, *49*, D380–D388.
- (20) Papatheodorou, I.; Moreno, P.; Manning, J.; Fuentes, A. M.; George, N.; Fexova, S.; Fonseca, N. A.; Füllgrabe, A.; Green, M.; Huang, N.; et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **2020**, *48*, D77–D83.
- (21) Howe, K. L.; Achuthan, P.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M. R.; Armean, I. M.; Azov, A. G.; Bennett, R.; Bhai, J.; et al. Ensembl 2021. *Nucleic Acids Res.* **2021**, *49*, D884–D891.
- (22) Jarnuczak, A. F.; Najgebauer, H.; Barzine, M.; Kundu, D. J.; Ghavidel, F.; Perez-Riverol, Y.; Papatheodorou, I.; Brazma, A.; Vizcaíno, J. A. An integrated landscape of protein expression in human cancer. *Sci. Data* **2021**, *8*, 115.
- (23) Wang, S.; García-Seisdedos, D.; Prakash, A.; Kundu, D. J.; Collins, A.; George, N.; Fexova, S.; Moreno, P.; Papatheodorou, I.; Jones, A. R.; et al. Integrated view and comparative analysis of baseline protein expression in mouse and rat tissues. *PLoS Comput. Biol.* **2022**, *18*, No. e1010174.
- (24) Samarasinghe, P.; Schmidt, T.; Frejino, M.; Gessulat, S.; Reinecke, M.; Jarzab, A.; Zecha, J.; Mergner, J.; Giansanti, P.; Ehrlich, H. C.; et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res.* **2020**, *48*, D1153–D1163.
- (25) Deutsch, E. W.; Lam, H.; Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **2008**, *9*, 429–434.
- (26) Johnson, E. C. B.; Dammer, E. B.; Duong, D. M.; Ping, L.; Zhou, M.; Yin, L.; Higginbotham, L. A.; Guajardo, A.; White, B.; Troncoso, J. C.; et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **2020**, *26*, 769–780.
- (27) Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; Fonseca, N. A.; Petryszak, R.; Papatheodorou, I.; et al. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **2019**, *47*, D711–D715.
- (28) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.
- (29) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319.

- (30) Mark A, T. R.; Afrasiabi, C.; Wu, C. *MyGene: Access MyGene, Info services*. 2014, Version 1.2.3[R/Bioconductor package].
- (31) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118–127.
- (32) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, No. e47.
- (33) Schmidt, T.; Samaras, P.; Frejno, M.; Gessulat, S.; Barnert, M.; Kienegger, H.; Krömer, H.; Schlegel, J.; Ehrlich, H. C.; Aiche, S.; et al. ProteomicsDB. *Nucleic Acids Res.* **2018**, *46*, D1271–D1281.
- (34) Yu, G.; Wang, L. G.; Han, Y.; He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **2012**, *16*, 284–287.
- (35) Wu, G.; Haw, R. Functional Interaction Network Construction and Analysis for Disease Discovery. *Methods Mol. Biol.* **2017**, *1558*, 235–253.
- (36) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L. H.; Meng, C.; et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **2019**, *15*, No. e8503.
- (37) Yelamanchi, S. D.; Tyagi, A.; Mohanty, V.; Dutta, P.; Korbonits, M.; Chavan, S.; Advani, J.; Madugundu, A. K.; Dey, G.; Datta, K. K.; et al. Proteomic Analysis of the Human Anterior Pituitary Gland. *OMICS* **2018**, *22*, 759–769.
- (38) Wingo, T. S.; Duong, D. M.; Zhou, M.; Dammer, E. B.; Wu, H.; Cutler, D. J.; Lah, J. J.; Levey, A. I.; Seyfried, N. T. Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. *J. Proteome Res.* **2017**, *16*, 3336–3347.
- (39) Gallart-Palau, X.; Serra, A.; Lee, B. S. T.; Guo, X.; Sze, S. K. Brain ureido degenerative protein modifications are associated with neuroinflammation and proteinopathy in Alzheimer's disease with cerebrovascular disease. *J. Neuroinflammation* **2017**, *14*, 175.
- (40) Abraham, J. R.; Szoko, N.; Barnard, J.; Rubin, R. A.; Schlatter, D.; Lundberg, K.; Li, X.; Natowicz, M. R. Proteomic Investigations of Autism Brain Identify Known and Novel Pathogenetic Processes. *Sci. Rep.* **2019**, *9*, 13118.
- (41) Bennike, T. B.; Carlsen, T. G.; Ellingsen, T.; Bonderup, O. K.; Glerup, H.; Bøgsted, M.; Christiansen, G.; Birkelund, S.; Stensballe, A.; Andersen, V. Neutrophil Extracellular Traps in Ulcerative Colitis: A Proteome Analysis of Intestinal Biopsies. *Inflammatory Bowel Dis.* **2015**, *21*, 2052–2067.
- (42) Bennike, T. B.; Kastaniegaard, K.; Padurariu, S.; Gaihede, M.; Birkelund, S.; Andersen, V.; Stensballe, A. Proteome stability analysis of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human colon mucosal biopsies. *Data Brief* **2016**, *6*, 942–947.
- (43) Martins-de-Souza, D.; Carvalho, P. C.; Schmitt, A.; Junqueira, M.; Nogueira, F. C.; Turck, C. W.; Domont, G. B. Deciphering the human brain proteome: characterization of the anterior temporal lobe and corpus callosum as part of the Chromosome 15-centric Human Proteome Project. *J. Proteome Res.* **2014**, *13*, 147–157.
- (44) Kushner, I. K.; Clair, G.; Purvine, S. O.; Lee, J. Y.; Adkins, J. N.; Payne, S. H. Individual Variability of Protein Expression in Human Tissues. *J. Proteome Res.* **2018**, *17*, 3914–3922.
- (45) Yelamanchi, S. D.; Kumar, M.; Madugundu, A. K.; Gopalakrishnan, L.; Dey, G.; Chavan, S.; Sathe, G.; Mathur, P. P.; Gowda, H.; Mahadevan, A.; et al. Characterization of human pineal gland proteome. *Mol. Biosyst.* **2016**, *12*, 3622–3632.
- (46) Doll, S.; Dreßen, M.; Geyer, P. E.; Itzhak, D. N.; Braun, C.; Doppler, S. A.; Meier, F.; Deutsch, M. A.; Lahm, H.; Lange, R.; et al. Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* **2017**, *8*, 1469.
- (47) Chen, C. Y.; Caporizzo, M. A.; Bedi, K.; Vite, A.; Bogush, A. I.; Robison, P.; Heffler, J. G.; Salomon, A. K.; Kelly, N. A.; Babu, A.; et al. Suppression of deetyrosinated microtubules improves cardiomyocyte function in human heart failure. *Nat. Med.* **2018**, *24*, 1225–1233.
- (48) Sonnen, J. A.; Larson, E. B.; Haneuse, S.; Woltjer, R.; Li, G.; Crane, P. K.; Craft, S.; Montine, T. J. Neuropathology in the adult changes in thought study: a review. *J. Alzheimers Dis.* **2009**, *18*, 703–711.
- (49) Beach, T. G.; Adler, C. H.; Sue, L. I.; Serrano, G.; Shill, H. A.; Walker, D. G.; Lue, L.; Roher, A. E.; Dugger, B. N.; Maarouf, C.; et al. Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program. *Neuropathology* **2015**, *35*, 354–389.
- (50) O'Brien, R. J.; Resnick, S. M.; Zonderman, A. B.; Ferrucci, L.; Crain, B. J.; Pletnikova, O.; Rudow, G.; Iacono, D.; Riudavets, M. A.; Driscoll, I.; et al. Neuropathologic studies of the Baltimore Longitudinal Study of Aging (BLSA). *J. Alzheimers Dis.* **2009**, *18*, 665–675.
- (51) Allen, M.; Carrasquillo, M. M.; Funk, C.; Heavner, B. D.; Zou, F.; Younkin, C. S.; Burgess, J. D.; Chai, H. S.; Crook, J.; Eddy, J. A.; et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci. Data* **2016**, *3*, 160089.
- (52) Wang, M.; Beckmann, N. D.; Roussos, P.; Wang, E.; Zhou, X.; Wang, Q.; Ming, C.; Neff, R.; Ma, W.; Fullard, J. F.; et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci. Data* **2018**, *5*, 180185.
- (53) McKetney, J.; Runde, R. M.; Hebert, A. S.; Salamat, S.; Roy, S.; Coon, J. J. Proteomic Atlas of the Human Brain in Alzheimer's Disease. *J. Proteome Res.* **2019**, *18*, 1380–1391.
- (54) Mallis, P.; Sokolis, D. P.; Makridakis, M.; Zoidakis, J.; Velentzas, A. D.; Katsimpoulas, M.; Vlahou, A.; Kostakis, A.; Stavropoulos-Giokas, C.; Michalopoulos, E. Insights into Biomechanical and Proteomic Characteristics of Small Diameter Vascular Grafts Utilizing the Human Umbilical Artery. *Biomedicines* **2020**, *8* (8), No. 280.
- (55) Li, P.; Ensink, E.; Lang, S.; Marshall, L.; Schilthuis, M.; Lamp, J.; Vega, I.; Labrie, V. Hemispheric asymmetry in the human brain and in Parkinson's disease is linked to divergent epigenetic patterns in neurons. *Genome Biol.* **2020**, *21*, 61.
- (56) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell. Proteomics* **2015**, *14*, 2394–2404.
- (57) The, M.; Samaras, P.; Kuster, B.; Wilhelm, M. Re-analysis of ProteomicsDB using an accurate, sensitive and scalable false discovery rate estimation approach for protein groups. *Mol. Cell. Proteomics* **2022**, *21*, 100437.
- (58) Dai, C.; Füllgrabe, A.; Pfeuffer, J.; Solovyeva, E.; Deng, J.; Moreno, P.; Kamatchinathan, S.; Kundu, D. J.; George, N.; Fexova, S. A proteomics sample metadata representation for multiomics integration, and big data analysis. *Nat. Commun.* **2021**, *12* (1), No. 5854.
- (59) Walzer, M.; García-Seisdedos, D.; Prakash, A.; Brack, P.; Crowther, P.; Graham, R. L.; George, N.; Mohammed, S.; Moreno, P.; Papatheodorou, I. Implementing the re-use of public DIA proteomics datasets: from the PRIDE database to Expression Atlas. *Sci. Data* **2022**, *9* (1), No. 335.
- (60) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Židek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2022**, 2021, 463034.
- (61) Barzine, M. P.; Freivalds, K.; Wright, J. C.; Opmanis, M.; Rituma, D.; Ghavidel, F. Z.; Jarnuczak, A. F.; Celms, E.; Čerāns, K.; Jonassen, I.; et al. Using Deep Learning to Extrapolate Protein Expression Measurements. *Proteomics* **2020**, *20*, 2000009.