An integrated view of baseline protein expression in human tissues

Ananth Prakash[1,2,*], David García-Seisdedos[1], Shengbo Wang[1], Deepti Jaiswal Kundu[1], Andrew Collins[3], Nancy George[1], Pablo Moreno[1], Irene Papathedourou[1,2], Andrew R. Jones[3], Juan Antonio Vizcaíno[1,2 *]

[1] European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD. United Kingdom.

[2] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD. United Kingdom.

[3] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom.

*Corresponding authors.

Dr. Ananth Prakash. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Phone: + 44 (0) 1223 492610. Email: ananth@ebi.ac.uk

Dr. Juan Antonio Vizcaíno. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Phone: + 44 (0) 1223 492686. Email: juan@ebi.ac.uk.

Keywords

Mass spectrometry, quantitative proteomics, public data re-use, human proteome

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 expression data in a consistent manner. We have reanalysed 25 public proteomics datasets from healthy human individuals, to assess baseline protein abundance in 32 organs. We defined tissue as a distinct functional or structural region within an organ. Overall, the aggregated dataset contains 68 healthy tissues, corresponding to 3,167 mass spectrometry runs covering 501 samples, coming from 492 individuals.

We compared protein expression between the different organs, studied the distribution of proteins across organs, and identified proteins, as well as their isoforms, that are uniquely expressed in certain organs. We also 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 expression results into the resource Expression Atlas, where it can be accessed and visualised either individually or together with gene expression data coming from transcriptomics datasets.

Introduction

High-throughput mass spectrometry (MS)-based proteomics approaches have matured and generalised 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 labelled-approaches such as metabolic-labelling (e.g., SILAC) and especially techniques based on the isotopic labelling 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 only somewhat correlated with absolute protein abundance due to considerable variation in the ionisation efficiency of different 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).

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 re-analyse data with different approaches to generate new findings. Therefore, in recent years, the amount and variety of shared datasets in the public domain has grown dramatically. This was driven by the establishment and maturation of reliable proteomics data repositories, in tandem with policy recommendations by scientific journals and funding agencies.

The PRIDE database [2], which is one of the founding members of the global ProteomeXchange consortium [3], is currently the largest resource worldwide for public proteomics data deposition. As of August 2021, PRIDE hosts more than 23,000 datasets. Of those, human datasets are by far the majority, representing nearly 40% of all datasets. Public datasets stored in PRIDE (or in other resources) present an opportunity to be systematically reanalysed and integrated, in order to confirm the original results potentially in a more robust manner, obtain new insights 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. Therefore, the large availability of public datasets has triggered different types of data reuse activities, including "big data" approaches (e.g. [4-6]) and the establishment of new data resources using re-analysed public datasets as the basis [7-9]. In this context of data reuse, 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 Atlas [10] (for quantitative proteomics expression data), Ensembl [10] (proteogenomics) and UniProt [11] (protein sequences information including post-translational modifications (PTMs)). The overall aim is to enable life scientists (including those who are non-experts in proteomics) to have 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 expression 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 expression data integrated with transcriptomics information in the web interface. Recently, we reported the re-analysis and integration into Expression Atlas of 11 public quantitative datasets coming from cell lines and human tumour samples [11].

There are other public resources providing access to reanalysed MS-based quantitative proteomics datasets. Proteomics DB [12] provides access to, among others to data coming from one of the first drafts of the human proteome [13] in addition to other recent (multi-omic) studies carried out on model organisms, for instance *Arabidopsis thaliana* [14]. Many additional human datasets coming from human tissues have been made publicly available in recent years. Additionally, antibody-based protein expression information is provided *via* the Human Protein Atlas [15] , which nowadays is probably the main reference resource for the understanding the global distribution of human proteins. Here, we report the reanalysis and integration of 25 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. The selection criteria for the datasets to be reanalysed were: i) experimental data from tissues in label-free studies where no PTM-enrichment had been performed; ii) experiments performed on Thermo Fischer 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 coming 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, 17 human datasets were selected from PRIDE (Table 1). Additionally, 8 datasets coming from human brain samples (also generated in Thermo Fischer Scientific instruments) were downloaded from a large Alzheimer's Disease (AD) dataset described in [16], which was available via the AMP-AD Knowledge Portal (https://adknowledgeportal.synapse.org/). The AD datasets from the AMP-AD Knowledge Portal are available under a controlled access agreement and were downloaded after obtaining the required authorisation.

The sample and experimental metadata was manually curated from their respective publications or by contacting the original authors/submitters. Metadata was annotated using Annotare [17] 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 baseline conditions described in this study, we also analysed 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 manuscript, 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 25 datasets sum up a total of 501 samples from 68 different tissues classified in 32 organs.

Proteomics raw data processing

Datasets were analysed separately, using the same software and search database. Peptide/protein identification and protein quantification were performed using MaxQuant [18, 19] (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, 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 default: maximum number of modifications per peptide: 5, minimum peptide length: 7, maximum peptide mass: 4,600 Da. For match between runs, the minimum match time window was set to 0.7 seconds and the minimum retention time alignment window was set to 20 seconds. The MaxQuant parameter files are available for download from Expression Atlas. The UniProt human reference proteome release-2019_05 (including isoforms, 95,915 sequences) was used as the target sequence database. The inbuilt MaxQuant contaminant database was used and the decoy database were generated by 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 normalised using the Fraction of Total (FOT) method, wherein each protein "iBAQ" intensity value is scaled to the total amount of signal in a given MS run and transformed to parts per billion (ppb).

$$ppb_iBAQ_i = \left({iBAQ_i}/{\sum_{i=1}^{n} iBAQ_i} \right) x\ 1{,}000{,}000{,}000$$

The bioconductor package 'mygene' [20] 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 expression 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 Supplementary file 1. 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 the different protein groups were mapped to 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 manuscript.

Integration into Expression Atlas

The calculated canonical protein abundances (mapped as genes), the 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 analysed separately, the protein abundances, available in ppb values within each dataset were converted into ranked bins for comparison of abundances across datasets. The normalised protein abundances per MS run, as described above, were ranked and grouped into 5 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 were binned separately. In

this study, 'tissue' is defined as a distinct functional or structural region within an 'organ'. For example, corpus callosum, anterior temporal lobe, dorsolateral prefrontal cortex were defined as tissues that are part of the brain (organ) and similarly left ventricle, aorta and tricuspid valve are defined as tissues in 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 (not available) 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 brain, all samples from tissue regions such as 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. Proteins that were detected in at least 50% of the samples in heart and brain datasets were selected for PCA (Principal Component Analysis). PCA was performed in R (The R Stats package) using binned abundance values. A Pearson correlation coefficient (r_p) for all samples was calculated on pairwise complete observations of bin transformed iBAQ values in R. Samples were hierarchically clustered on columns and rows using Euclidean distances.

Organ-specific expression profile analysis

To investigate the organ-specific protein-based expression profile, we carried out a modification of the classification scheme done by Uhlén *et al.* [15]. Briefly, each of the 13,271 canonical proteins that were mapped from the protein groups, was classified into one of three categories based on the bin levels in 32 organs: (1) "Organ-enriched": one unique organ with bin values 2-fold higher than the mean bin value across all organs; (2) "Group enriched": group of 2-7 organs with bin values 2-fold 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) terms 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 clusterProfiler [21] 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 pvalueCutoff to 0.05 and the qvalueCutoff to 0.05. Additionally, Reactome [22] pathway analysis was carried out by using mapped gene lists (indicated by the protein groups) and running pathway-topology and over-representation analysis. First, "Project to human" option was selected with the combining list of "Organ-enriched" and "Group enriched" entities. Afterwards, those pathways with p-value > 0.05 were filtered out.

The hierarchical clustering was done based on the distances calculated on the p-values using the ggdendro package in R.

Expression Atlas accession number	Proteomics dataset identifier*	Tissues	Organs	Number of MS runs	Number of samples	Number of protein groups†	Number of peptides†	Number of unique peptides†	Number of unique genes (canonical proteins) mapped†
E-PROT-29	PXD010154 [23]	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	1,795	37	19,382	874,825	346,894	12,622
E-PROT-33	PXD005819 [24]	Anterior pituitary gland (adenohypophysis)	Brain	33	2	1,936	12,112	10,174	1,207
E-PROT-34	PXD004143 [25]	Dorsolateral prefrontal cortex	Brain	80	4	9,929	203,798	124,355	7,633
E-PROT-35	PXD006233 [26]	Middle temporal lobe	Brain	192	6	7,030	62,455	46,786	5,264
E-PROT-36	PXD012755 [27]	Cerebellar hemispheric cortex, Occipital cortex	Brain	15	15	3,033	25,514	20,125	2,264
E-PROT-37	PXD001325 [28]	Breast (tumour distant)	Breast	48	3	1,816	12,100	10,424	1,165
E-PROT-40	PXD001608 [29]	Colon	Colon	30	10	5,943	73,053	56,756	4,789
E-PROT-41	PXD002029 [30]	Colon	Colon	24	8	4,608	25,300	22,032	3,211
E-PROT-42	PXD000547, PXD000548 [31]	Corpus callosum, Anterior temporal lobe	Brain	40, 40	2, 2	1,353, 1,873	8,250, 12,503	6,739, 10,071	868, 1,304
E-PROT-43	PXD010271 [32]	Liver, Ovary, Pancreas, Substantia nigra	Liver, Ovary, Pancreas, Brain	55	54	5,927	75,320	50,652	4,528

E-PROT-44	PXD004332 [33]	Pineal gland	Brain	56	3	4,953	49,455	38,884	3,692
E-PROT-45	PXD006675 [34]	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	Heart	347	42	9,160	161,943	98,692	7,602
E-PROT-46 §	PXD008934 [35]	Left ventricle	Heart	7	7	2,977	31,755	25,704	2,294
E-PROT-51 §	syn6038852 [36]	Dorsolateral prefrontal cortex	Brain	11	11	4,962	56,558	41,312	3,803
E-PROT-52 §	syn21444980 [16]	Dorsolateral prefrontal cortex	Brain	83	83	3,846	55,290	39,470	3,127
E-PROT-53 §	syn7204174 [37]	Dorsolateral prefrontal cortex	Brain	26	26	5,658	101,460	64,060	4,454
E-PROT-54 §	syn3606087 [38]	Dorsolateral prefrontal cortex	Brain	11	11	4,751	58,302	40,469	3,626
E-PROT-55 §	syn4624471 [38]	Precuneus	Brain	13	13	4,829	60,129	42,179	3,695
E-PROT-56 §	syn7431984 [39]	Temporal cortex	Brain	31	31	5,997	113,820	68,951	4,675
E-PROT-57 §	syn6038797 [40]	Frontal pole	Brain	53	53	5,812	91,285	59,701	4,420
E-PROT-58 §	syn21443008 [16]	Dorsolateral prefrontal cortex	Brain	47	47	5,050	86,685	55,386	4,053
E-PROT-61 §	PXD012131 [41]	Amygdala, Caudate nucleus, Cerebellum, Entorhinal cortex, Inferior parietal lobule, Middle frontal gyrus, Neocortex, Superior temporal gyrus, Thalamus, Visual cortex	Brain	114	15	10,097	192,397	113,310	7,826
E-PROT-63	PXD020187 [42]	Umbilical artery	Umbilical artery	10	10	1,505	10,365	7,702	933
E-PROT-65 §	PXD015079 [43]	Prefrontal cortex, Vermiform appendix	Brain, Vermiform appendix	6	6	2,313	24,162	18,892	1,694
TOTAL	**25 datasets**	**68 tissues**	**32 organs**	**3,167 MS runs**	**501 Samples**				

Table 1. List of proteomics datasets that were reanalysed. The breast tissue sample analysed here is a normal/healthy one and was categorised as 'tumour distant' in the original publication. *Dataset identifiers starting with 'PXD' come from the PRIDE database and those identifiers starting by 'syn' come from the AMP-AD Knowledge Portal. §Only normal

samples within this dataset are reported in this study. However, results from both normal and disease samples are available in Expression Atlas. † Numbers after post-processing.

Results

Human baseline proteomics datasets

We manually selected 25 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 Fischer Scientific Instruments. See more details about dataset selection in the 'Methods' section. (Figure 1).

In total the aggregated datasets represent 68 healthy tissues, corresponding to 3,167 MS runs covering 501 samples, coming from 492 individuals. The cumulative CPU time used for the reanalyses was approximately 2,752 hours or 114 calendar days. The numbers of protein groups, peptides, unique peptides identified and protein coverage in each dataset are shown in Table 1.

The resulting protein abundances of all samples were made available in Expression Atlas. These 'proteomics baseline' quantification results can be viewed as expression 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) summarising the output of the post-processed samples.

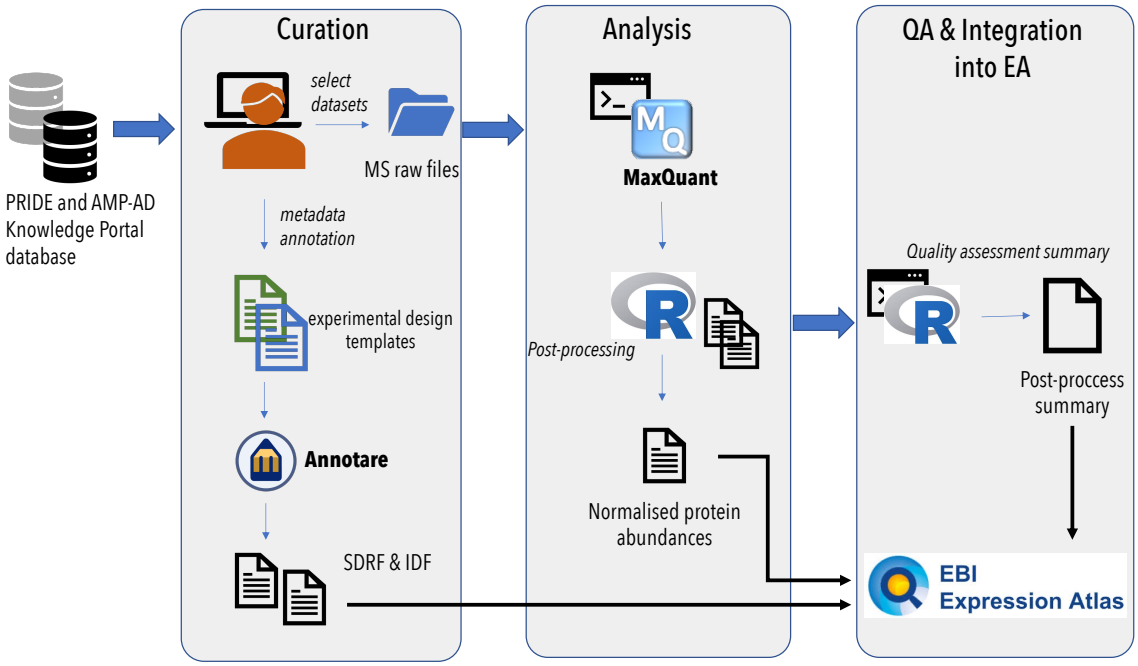

Figure 1. An overview of the study design and reanalysis pipeline. QA: Quality assessment.

Protein coverage across samples

For simplicity of comparison, we broadly grouped 68 tissues into 32 major types of organs. As explained in 'Methods', we defined 'tissue' as a distinct functional or structural region within an 'organ'. For example, corpus callosum, anterior temporal lobe, dorsolateral prefrontal cortex were all defined as tissues in brain (which is the 'organ'). After post-processing the output files from MaxQuant, 11,585 protein groups (36.1%) were uniquely present in only one organ and 183 protein groups (0.57%) were ubiquitously observed (Supplementary file 2).

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 manuscript. Overall, 13,271 different genes were mapped coming from the protein groups. We denote the term 'protein abundance' to mean 'canonical protein abundance' from here on, except in cases where we specifically discuss protein isoforms. We then estimated the number of proteins identified across organs, which indicated that greater than 70% of all canonical proteins were present in a majority of organs (Figure 2A). We also observed the highest numbers in samples from tonsil (91.0%) and brain (90.1%) and the lowest numbers in samples from umbilical artery (7.0%) and breast (8.7%) tissue samples.

The higher number of proteins identified in brain could be attributed to the greater representation of samples (339 samples out of 501, 67.7%). Tonsil samples had the fewest sample size in comparison (7 samples) and were all derived from one dataset (PXD010154). The sample preparation protocol for the tonsil samples employed seven different proteases

(Trypsin, LysC, ArgC, GluC, AspN, LysN and Chymotrypsin) for tissue digestion [23], thus significantly increasing its peptide coverage [23]. The sample sizes of breast and umbilical artery, both of which showed significantly lower protein coverage than other organs, were 3 and 10 samples, respectively.

The largest number of canonical proteins were identified in dataset PXD010154 (Figure 2B), which comprises numerous tissue samples (31 tissues) including samples from tonsil. The dynamic range of protein abundances in all organs is shown in Figure 2C. 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).

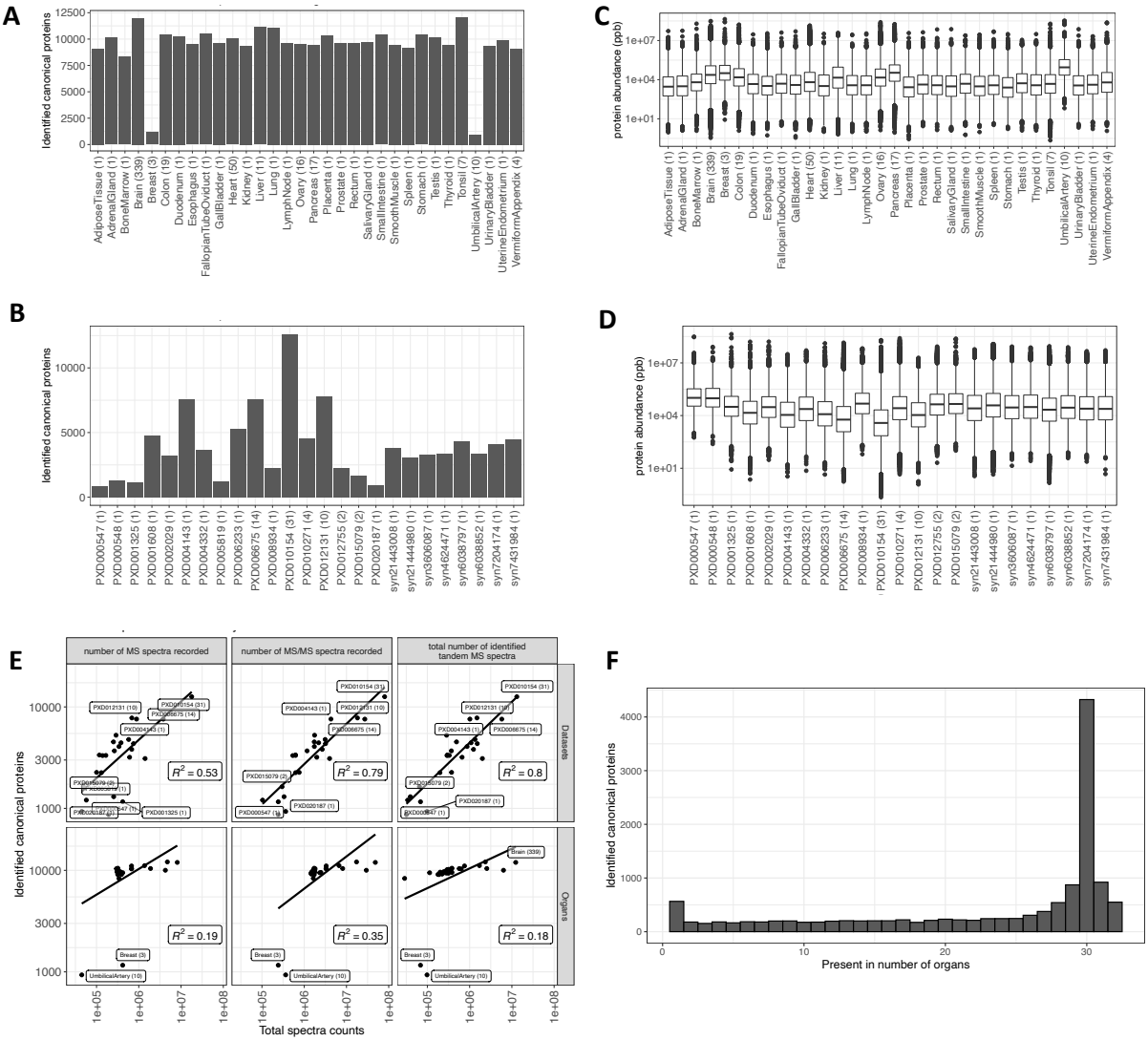

Figure 2. (A) Number of canonical proteins identified across different organs. The number within the parenthesis indicates the number of samples. (B) Canonical proteins identified across different datasets. The number within the parenthesis indicate the number of unique tissues in the dataset. (C) Range of normalised iBAQ protein abundances across different organs. The number within the parenthesis indicates the number of samples. (D) Range of normalised iBAQ protein abundances across different datasets. The number within parenthesis indicate 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.

Distribution of canonical protein identifications per organ

We observed that 32.6% (4,324) of the identified canonical proteins were expressed in 30 different organs (Figure 2F). The low number of proteins identified in breast (1,162) and umbilical artery (933) samples greatly influenced the protein distribution. As a result, nearly 4.1% (550) of all identified canonical proteins were present in all 32 organs, whereas 4.2% (565) of the identified canonical proteins were present uniquely in one organ, which included 422 proteins uniquely expressed in brain, 109 uniquely expressed in heart, 20 proteins uniquely expressed in colon, 7 proteins uniquely expressed in adrenal gland, 2 proteins uniquely expressed in pancreas, and 1 protein uniquely expressed each in breast, liver, ovary, umbilical artery and vermiform appendix (Supplementary file 2).

Identification of unique protein isoforms in organs

From the list of protein groups identified, we next studied which protein isoforms were uniquely expressed in any organ. Among the 11,585 protein groups that were identified in only one unique organ, we selected those instances where the protein group comprised of only one protein isoform, to which a peptide was mapped, and that was also not part of any other protein groups of the same parent gene.

We found 25 protein isoforms that were identified in only one organ (Supplementary file 3). Among these organ-specific proteins 7 isoforms were identified in brain, 4 in tonsil and 3 in heart. The organ-specific isoforms identified in tonsil are probably due to the deeper peptide coverage resulting from the use of multiple enzymes. From the 25 identified organ-specific isoforms, 1 isoform (protein accession number Q9NZ56-2) was supported by tissue-specific expression annotation in UniProt, while 2 isoforms were annotated as expressed in the identified organ but also in other organs, 8 isoforms were annotated as expressed in different tissues whereas 14 isoforms had no tissue specific expression annotation in UniProt.

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

normalised iBAQ intensities into ranked bins as explained in 'Methods'. The bin transformed protein abundances in all organs are provided in Supplementary file 2.

To compare protein expression across all organs, a pairwise Pearson correlation coefficients (r_p) of binned protein abundances was calculated across 501 samples (Figure 3). We observed a good correlation of protein expression within the brain (median $r_p = 0.78$) and heart (median $r_p = 0.63$) samples, which represent the two organ groups with the largest number of samples. We tested the effectiveness of the bin transformation method in reducing batch effects, by performing a PCA on samples coming from heart and brain datasets. The brain and heart tissue samples analysed constituted the largest numbers, including 19 and 3 datasets, respectively. 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 [16]. 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 heart samples clustered according to their tissue types (Figure 4C). All heart samples except those coming from left ventricle were part of an individual dataset. Interestingly, we observed 3 major clusters: one wherein all valve samples (aortic valve, mitral valve, pulmonary valve and tricuspid valve) were clustered together. A second cluster where the samples from ventricles and atriums were clustered in a large group together with other heart samples. Finally, left ventricle samples from dataset PXD010154 (Figure 4D) formed a separate cluster indicating that there were still batch effects which were not completely removed.

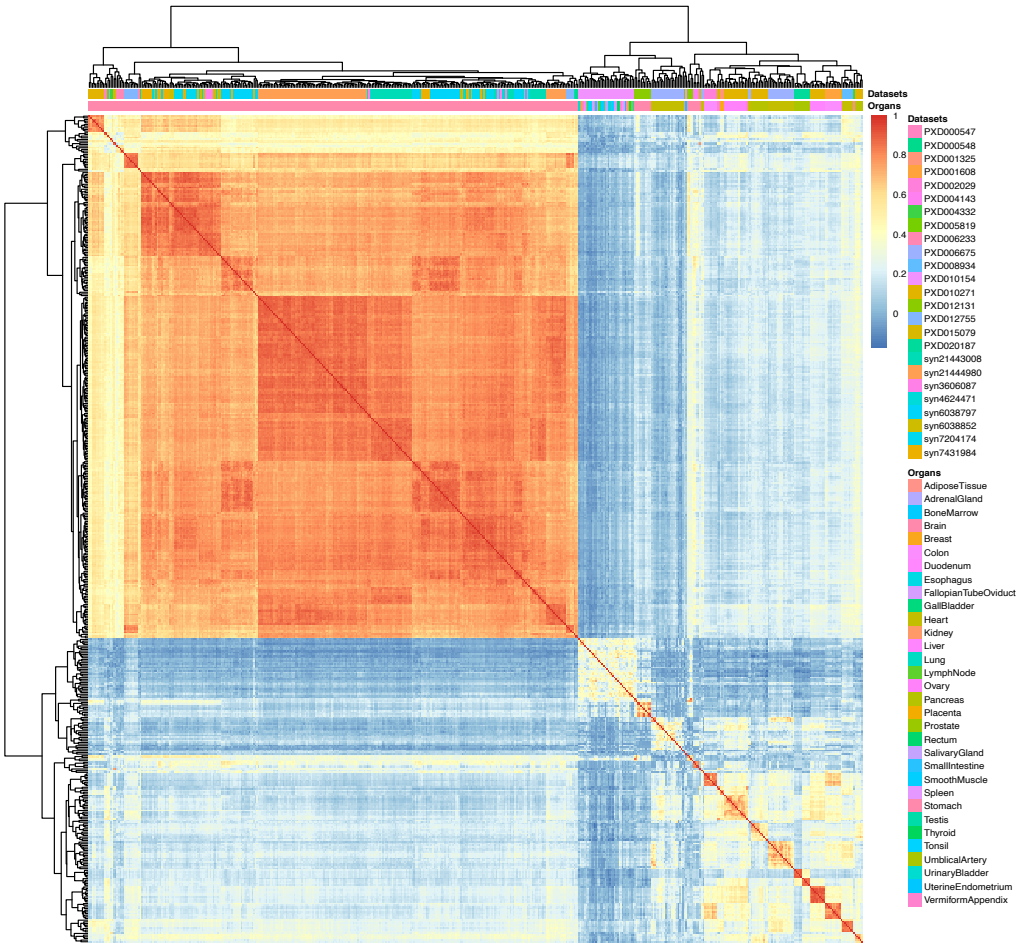

Figure 3. Heatmap of pairwise Pearson correlation coefficients across all samples. Colours on the heatmap represents the correlation coefficient and was calculated using the bin transformed iBAQ values. The samples are hierarchically clustered on columns and rows using Euclidean distances.

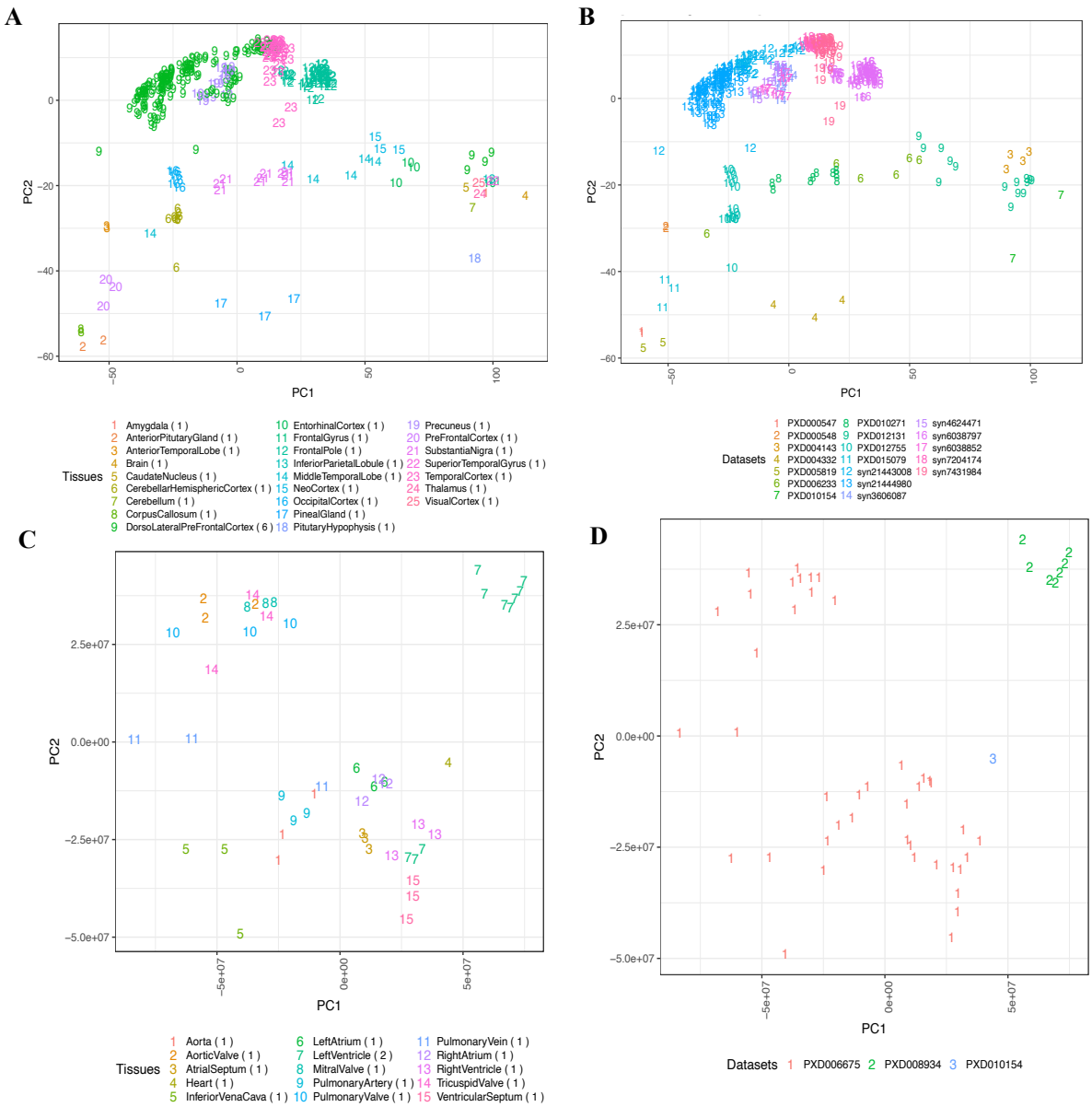

Figure 4. (A) PCA of brain samples coloured by the tissue types. (B) PCA of brain samples coloured by their respective dataset identifiers. (C) PCA of heart samples coloured by the tissue types. (D) PCA of heart samples coloured by their respective dataset identifiers. The numbers in parenthesis indicate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

The organ elevated proteome and the over-representative biological processes

As explained in 'Methods', according to their expression, canonical proteins were divided in three different groups according to their organ-specificity: "Organ-enriched", "Group enriched" and "Mixed" (see Supplementary file 4). We considered elevated canonical proteins those which were classified as an "Organ-enriched" or "Group enriched" instead of

the “Mixed” group. The analysis (Figure 5A) showed that on 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 (7.2%), and the lowest ratio in gall bladder (2.3%), breast (0.2%) and umbilical artery (0.1%). In addition, 0.4% of the total canonical proteins were unique organ-enriched. The highest ratio was found in brain (3.8%), heart (1.7%) and bone marrow (0.6%) and the lowest ratio (~0.1%) was found in tonsil, breast 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”, is shown in Table 2. As a summary, 326 GO terms were found statistically significative across all organs (see Supplementary file 5). 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 (455), among the biological processes associated stand out the regulatory function on membrane potential (GO:0042391), ion transport (GO:0043269), synaptic transmission (GO:0051932), secretory pathway (GO:1903305) and neurotransmitter transport (GO:0006836). The second organ with a greater number of “organ-enriched” canonical proteins was heart (137). The enriched biological processes involved were related with muscle contraction (GO:0006936) and structure (GO:0061061), heart contraction (GO:0060047) and regulation of heart rate (GO:0002027). As expected, there were common GO terms that were shared between the organs, such as: detoxification of inorganic compound (GO:0061687) and organic acid transport (GO:0015849) in liver and kidney, anion transmembrane transport (GO:0098656) in kidney, brain and umbilical artery, processes involved in tissues with high cell division turnover like chromosome segregation (GO:0007059) in bone marrow and testis.

Organ	GO ID	Description	p-value
Adrenal gland	GO:0031649	Heat generation	$2.82*10^{-7}$
Bone marrow	GO:0007059	Chromosome segregation	$3.54*10^{-6}$
Brain	GO:0042391	Regulation of membrane potential	$3.67*10^{-17}$
Fallopian tube oviduct	GO:0044782	Cilium organization	$1.52*10^{-48}$
Gallbladder	GO:0017158	Regulation of calcium ion-dependent exocytosis	$1.34*10^{-6}$
Heart	GO:0051146	Striated muscle cell differentiation	$1.63*10^{-9}$
Kidney	GO:0046942	Carboxylic acid transport	$4.54*10^{-15}$
Liver	GO:0010273	Detoxification of copper ion	$8.66*10^{-7}$
Lung	GO:0045165	Cell fate commitment	$1.81*10^{-6}$
Lymph node	GO:0002250	Adaptive immune response	$4.99*10^{-10}$
Ovary	GO:0008544	Epidermis development	$1.70*10^{-10}$
Placenta	GO:0044706	Multi-multicellular organism process	$2.78*10^{-5}$
Testis	GO:0048232	Male gamete generation	$1.16*10^{-30}$

Thyroid	GO:0006590	Thyroid hormone generation	$8.96*10^{-6}$
Tonsil	GO:0018149	Peptide cross-linking	$5.95*10^{-5}$
Umbilical artery	GO:0001937	Negative regulation of endothelial cell proliferation	$6.01*10^{-6}$

Table 2. Analysis of the GO terms for each organ using the elevated organ-specific canonical proteins and group-specific as described in the 'Methods' section.

Next, we performed a pathway-enrichment analysis using Reactome [22] to analyse canonical proteins that were "organ-enriched" and "group-enriched" (see Supplementary file 6). The heatmap (Figure 5B) shows the statistically significant pathways, (p-value < 0.05) across the organs. The total number of pathways found in all the organs were 945, 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 bone marrow and testis (R-HSA-69620, R-HSA-73886, R-HSA-774815 and R-HSA-1640170), expected in high cell turnover tissues, the digestion pathway (R-HSA-192456) in 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-425393, R-HSA-425366 and R-HSA-382551) in kidney.

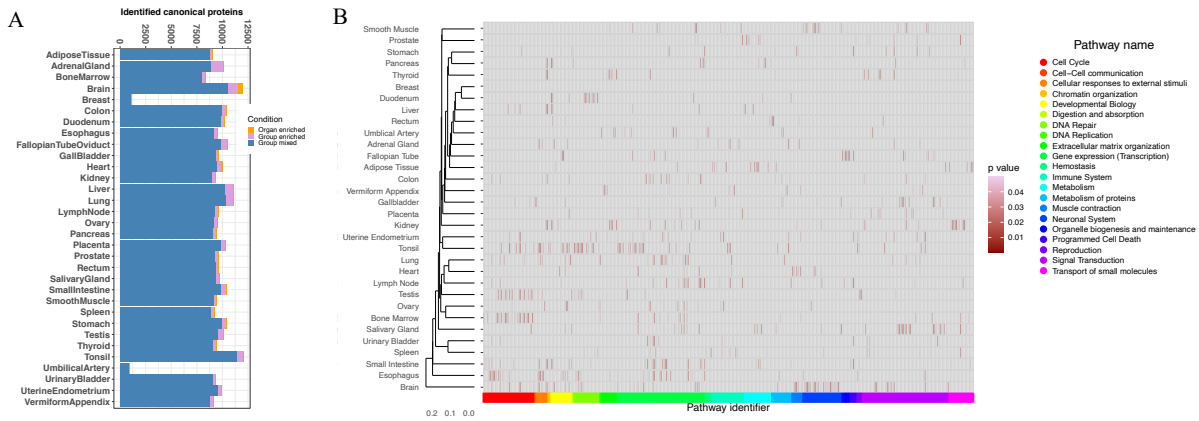

Figure 5. (A) Analysis of organ-specific canonical proteins. The analysis comprises the number of canonical proteins found in 32 organs, classified in three groups: "organ-enriched", "group enriched" and "group mixed". (B) Pathway analysis of the over-represented canonical proteins, showing the statistically significant representative pathways (p-value < 0.05) in 32 organs.

Discussion

We here include a combined analysis of human baseline proteomics datasets representing baseline protein expression across 68 healthy tissues grouped in 32 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 meta-analyses of proteomics data covering a wide array of tissues and biological conditions. The main aim of our study was to provide a systems-wide baseline protein expression catalogue across various tissues and organs, which could be used as a reference and help to reduce redundant efforts of similar computationally expensive reanalyses.

Unlike what was done in one previous study performed by us [11] here we analysed each dataset separately using the same software and the same search protein sequence database. The disadvantage of this approach is that the FDR statistical threshold 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, 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. This enables that: (i) 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 without having to reanalyse all datasets together again; (iii) future updates in the results are feasible to perform; and (iv) (semi)-automation of the reanalyses is achievable, making again these efforts more sustainable.

One of the major bottlenecks was, as reported before, the curation of dataset metadata, consisting in 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 either was inferred or were 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 reproducibility of results in the future, we need to improve the provision of metadata by data submitters. A format to do that has been developed recently (an adapted SDRF-Proteomics format, as part of MAGE-TAB-Proteomics), which can be submitted optionally at present [44]. 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 reanalyses of a large number of proteomics datasets is the integration of results from data since batch effects are inevitable. We used a rank-binned normalisation of abundances, which transformed protein abundances across datasets and samples to bins of 1 to 5. This approach is useful to reduce batch effects, although we acknowledge there is also loss of signal through this transformation. 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 availability of the results through Expression Atlas enables the integration of mRNA and proteomics expression information, offering an interface for researchers to access this type of information. The next step will be the integration of datasets in the differential part of Expression Atlas. The work required is more complex there at different levels, including the downstream statistical differential analysis. Also, availability of the mapping between the channels (e.g. in TMT, SILAC experiments) and the samples is very rare at present. In parallel, work has also started in integrating in Expression Atlas proteomics data generated using Data Independent Acquisition (DIA) approaches [45].

The generated baseline protein expression data can be used with different purposes. For instance, quantitative proteomics data can be used for the generation of co-expression networks and/or the inference of protein complexes. Additionally, it is possible to use artificial intelligence approaches to impute protein expression values using calculated expression values as training data [46]. It would also be possible to perform expression correlation studies between gene and protein expression information. However, this type of studies can be performed more optimally if the same samples are analysed by both techniques, as reported in the original publication for dataset PXD010154 [23].

In conclusion the results presented here represent a pilot study of a large-scale meta-analyses of public human baseline proteomics datasets. We demonstrate its feasibility and show the challenges in this kind of analyses, providing a roadmap for such future studies.

Authors' contributions

AP, DGS, SW, DJK selected and curated the datasets. AP, DGS and SW performed analyses. AC and AJ helped in the interpretation of results and designed approach for data normalisation. AP, DGS, JAV wrote the manuscript. All authors have read and approved the manuscript.

Data availability

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

Acknowledgements

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 suggestions and discussions.

Abbreviations

AD: Alzheimer's Disease
DLPFC: Dorsolateral PreFrontal Cortex
FOT: Fraction Of Total
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

Supplementary Material

Supplementary files can be found here:
https://drive.google.com/drive/u/1/folders/1BkZKAQlxgwwQp8Ug1_fOVfXwycaZfknW

References

1. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.
2. Perez-Riverol, Y., et al., *The PRIDE database and related tools and resources in 2019: improving support for quantification data.* Nucleic Acids Res, 2019. **47**(D1): p. D442-D450.
3. Vizcaino, J.A., et al., *ProteomeXchange provides globally coordinated proteomics data submission and dissemination.* Nat Biotechnol, 2014. **32**(3): p. 223-6.
4. Ochoa, D., et al., *The functional landscape of the human phosphoproteome.* Nat Biotechnol, 2020. **38**(3): p. 365-373.
5. Tiwary, S., et al., *High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis.* Nat Methods, 2019. **16**(6): p. 519-525.
6. Gabriels, R., L. Martens, and S. Degroeve, *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**(W1): p. W295-W299.
7. Kustatscher, G., et al., *Co-regulation map of the human proteome enables identification of protein functions.* Nat Biotechnol, 2019. **37**(11): p. 1361-1371.
8. Ramasamy, P., et al., *Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context.* J Proteome Res, 2020. **19**(8): p. 3478-3486.
9. Brunet, M.A., et al., *OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes.* Nucleic Acids Res, 2021. **49**(D1): p. D380-D388.
10. Papatheodorou, I., et al., *Expression Atlas update: from tissues to single cells.* Nucleic Acids Res, 2020. **48**(D1): p. D77-D83.
11. Jarnuczak, A.F., et al., *An integrated landscape of protein expression in human cancer.* Sci Data, 2021. **8**(1): p. 115.
12. Samaras, P., et al., *ProteomicsDB: a multi-omics and multi-organism resource for life science research.* Nucleic Acids Res, 2020. **48**(D1): p. D1153-D1163.
13. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.
14. Mergner, J., et al., *Mass-spectrometry-based draft of the Arabidopsis proteome.* Nature, 2020. **579**(7799): p. 409-414.
15. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
16. Johnson, E.C.B., 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**(5): p. 769-780.
17. Athar, A., et al., *ArrayExpress update - from bulk to single-cell expression data.* Nucleic Acids Res, 2019. **47**(D1): p. D711-D715.
18. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
19. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nat Protoc, 2016. **11**(12): p. 2301-2319.
20. Mark A, T.R., Afrasiabi C, Wu C, *mygene: Access MyGene.Info services. Version 1.2.3.* R/Bioconductor package, 2014.

21. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.
22. Wu, G. and R. Haw, *Functional Interaction Network Construction and Analysis for Disease Discovery.* Methods Mol Biol, 2017. **1558**: p. 235-253.
23. Wang, D., et al., *A deep proteome and transcriptome abundance atlas of 29 healthy human tissues.* Mol Syst Biol, 2019. **15**(2): p. e8503.
24. Yelamanchi, S.D., et al., *Proteomic Analysis of the Human Anterior Pituitary Gland.* OMICS, 2018. **22**(12): p. 759-769.
25. Wingo, T.S., et al., *Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain.* J Proteome Res, 2017. **16**(9): p. 3336-3347.
26. Gallart-Palau, X., et al., *Brain ureido degenerative protein modifications are associated with neuroinflammation and proteinopathy in Alzheimer's disease with cerebrovascular disease.* J Neuroinflammation, 2017. **14**(1): p. 175.
27. Abraham, J.R., et al., *Proteomic Investigations of Autism Brain Identify Known and Novel Pathogenetic Processes.* Sci Rep, 2019. **9**(1): p. 13118.
28. Groessl, M., et al., *Proteome profiling of breast cancer biopsies reveals a wound healing signature of cancer-associated fibroblasts.* J Proteome Res, 2014. **13**(11): p. 4773-82.
29. Bennike, T.B., et al., *Neutrophil Extracellular Traps in Ulcerative Colitis: A Proteome Analysis of Intestinal Biopsies.* Inflamm Bowel Dis, 2015. **21**(9): p. 2052-67.
30. Bennike, T.B., et al., *Proteome stability analysis of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human colon mucosal biopsies.* Data Brief, 2016. **6**: p. 942-7.
31. Martins-de-Souza, D., et al., *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**(1): p. 147-57.
32. Kushner, I.K., et al., *Individual Variability of Protein Expression in Human Tissues.* J Proteome Res, 2018. **17**(11): p. 3914-3922.
33. Yelamanchi, S.D., et al., *Characterization of human pineal gland proteome.* Mol Biosyst, 2016. **12**(12): p. 3622-3632.
34. Doll, S., et al., *Region and cell-type resolved quantitative proteomic map of the human heart.* Nat Commun, 2017. **8**(1): p. 1469.
35. Chen, C.Y., et al., *Suppression of detyrosinated microtubules improves cardiomyocyte function in human heart failure.* Nat Med, 2018. **24**(8): p. 1225-1233.
36. Sonnen, J.A., et al., *Neuropathology in the adult changes in thought study: a review.* J Alzheimers Dis, 2009. **18**(3): p. 703-11.
37. Beach, T.G., et al., *Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program.* Neuropathology, 2015. **35**(4): p. 354-89.
38. O'Brien, R.J., et al., *Neuropathologic studies of the Baltimore Longitudinal Study of Aging (BLSA).* J Alzheimers Dis, 2009. **18**(3): p. 665-75.
39. Allen, M., et al., *Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases.* Sci Data, 2016. **3**: p. 160089.
40. Wang, M., et al., *The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease.* Sci Data, 2018. **5**: p. 180185.

41. McKetney, J., et al., *Proteomic Atlas of the Human Brain in Alzheimer's Disease.* J Proteome Res, 2019. **18**(3): p. 1380-1391.
42. Mallis, P., et al., *Insights into Biomechanical and Proteomic Characteristics of Small Diameter Vascular Grafts Utilizing the Human Umbilical Artery.* Biomedicines, 2020. **8**(8).
43. Li, P., et al., *Hemispheric asymmetry in the human brain and in Parkinson's disease is linked to divergent epigenetic patterns in neurons.* Genome Biol, 2020. **21**(1): p. 61.
44. Dai, C., et al., *A proteomics sample metadata representation for multiomics integration, and big data analysis.* bioRxiv, 2021: p. 2021.05.21.445143.
45. Walzer, M., et al., *Implementing the re-use of public DIA proteomics datasets: from the PRIDE database to Expression Atlas.* bioRxiv, 2021: p. 2021.06.08.447493.
46. Barzine, M.P., et al., *Using Deep Learning to Extrapolate Protein Expression Measurements.* Proteomics, 2020. **20**(21-22): p. e2000009.