Integrated analysis of baseline protein expression in pig tissues

Shengbo Wang[1†], Andrew Collins[3†], Ananth Prakash[1,2†], Silvie Fexova[1], Irene Papatheodorou[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.

†All three authors have contributed equally, and they wish to be considered as joint first authors.

Prof. Andrew R. Jones. Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom. Email: Andrew.Jones@liverpool.ac.uk.

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

Abstract

The availability of an increasingly large amount of public proteomics datasets presents an opportunity for performing combined analyses to generate comprehensive organism-wide protein expression maps across different organisms and biological conditions. *Sus scrofa*, the domestic pig, is a model organism relevant for food production and for human biomedical research. Here we reanalyzed 14 public proteomics datasets from the PRIDE database coming from pig tissues to assess baseline (without any biological perturbation) protein abundance in 14 organs, encompassing a total of 20 healthy tissues from 128 samples. The analysis involved the quantification of protein abundance in 599 mass spectrometry runs.

We compared protein expression patterns among different pig organs and examined the distribution of proteins across these organs. The number of canonical proteins per dataset ranged from 789 (dataset PXD002918, biceps femoris) to 6,062 (dataset PXD012636, heart). Then, we studied how protein abundances compared across different datasets. Of particular interest, we conducted a comparative analysis of protein expression between pig and human tissues (from a previous study performed following the same methodology), revealing a high degree of correlation in protein expression among orthologs, particularly in brain, kidney, heart, and liver samples. Additionally, we analysed the differences in quantitative expression of proteins across organs between human and pig organs.

We have integrated the protein expression results into the Expression Atlas resource for easy access and visualisation of the protein expression data individually or alongside gene expression data.

Introduction

In recent years, high-throughput mass spectrometry (MS)-based proteomics methods have made significant advances and have become essential tools in biological research [1]. These improvements are the result of significant developments in MS instrumentation, chromatographic methods, sample preparation automation and computational analysis [2]. The dominant experimental technique for MS-based proteomics has historically been data-dependent acquisition (DDA) bottom-up proteomics [3]. Among the quantitative techniques, label-free DDA approaches are well accepted. However, Data Independent Acquisition (DIA) approaches are currently becoming increasingly popular.

In parallel to the technical developments, in recent years, the proteomics community has embraced open data practices, leading to a substantial increase in the availability of shared datasets in the public domain. The field has mirrored the progress witnessed in genomics and transcriptomics. The PRIDE database [4], as part of the ProteomeXchange consortium [5], is the most used proteomics data repository worldwide. The availability of extensive public proteomics datasets has paved the way for various applications, including meta-analysis studies involving the reanalysis and integration of quantitative proteomics datasets [6-8]. By systematically reanalysing these datasets, original findings can be updated, confirmed and/or strengthened. Moreover, novel insights beyond the scope of the original studies can be obtained through alternative reanalysis strategies to those used in the original studies.

To enable the access to proteomics data by the wider scientific community, PRIDE is developing data dissemination and integration pipelines with existing popular resources at the European Bioinformatics Institute (EBI). Expression Atlas [9] (https://www.ebi.ac.uk/gxa/home) is a well-established database for gene expression data,

and has more recently incorporated protein expression information derived from reanalysed datasets into its 'bulk' section. As a result, the integration of proteomics expression/protein abundance data with transcriptomics information, primarily from RNA-Seq experiments, enhances the comprehensive understanding of molecular expression across various biological contexts. This approach ensures the long-term accessibility and integration of proteomics data, benefiting researchers, including those without expertise in proteomics, in their exploration of multi-omics information.

We have already performed combined analyses of baseline (without any perturbation) protein expression for human [7], mouse and rat tissues [6]. Here we are reporting an analogous study of baseline protein expression in the model organism *Sus scrofa,* the domestic pig. The study of pig proteomics datasets is crucial for advancing food production, animal welfare, and human biomedical research, as it offers insights into genetic and environmental factors affecting farm animal production and leverages the close genetic and proteomic similarities between pigs and humans [10, 11]. The resource PeptideAtlas provided a few years ago a build for pig, including extensive peptide and protein identification data, but is not providing protein abundance information [8]. Additionally, PaxDB recently released a new version 5.0 [12], providing expression data coming from different vertebrates including *Sus scrofa,* but quantitative data is based on spectral counting, a semi-quantitative technique. Also, no tissue specific information is provided, apart from liver. To the best of our knowledge, we are providing the first combined quantitative analysis of label-free DDA datasets.

Here, we report the reanalysis and integration of 14 public label-free pig baseline tissue datasets including 14 organs and a total of 20 healthy tissues from 128 samples. The results were incorporated into Expression Atlas as baseline studies. Additionally, we report a comparative analysis of protein expression across pig and human tissues.

Methods

1. Datasets

The PRIDE database hosted 165 publicly available MS proteomics datasets of *Sus scrofa* as of October 2022. For this study, we manually selected datasets based on several predefined criteria, which included: (i) label-free DDA studies from tissues without enrichment for post-translational modifications; (ii) datasets generated using Thermo Fisher Scientific instruments to avoid the heterogeneity introduced by data generated by other platforms; and (iii) datasets with sufficient sample metadata, manually curated from the original publication. This resulted in the identification of 14 pig datasets for further analysis.

Sample and experimental metadata were manually curated using Annotare [13], and adhering to the Investigation Description Format (IDF) and Sample-Data Relationship Format (SDRF) files [14], which are needed for integration of the data into Expression Atlas. The IDF file contains an overview of the experimental design, including details on experimental factors, protocols, publication information, and contact information. The SDRF file contains complementary information - the sample metadata that describes the relationships between various sample characteristics and the associated data files within the dataset.

2. Proteomics raw data processing

All datasets underwent analysis using MaxQuant version 2.0.3.0 [15] in multi-threaded mode on a Linux high-performance computing cluster for peptide/protein identification and protein quantification. Input parameters for each dataset, including MS1 and MS2 tolerances, digesting enzymes, fixed and variable modifications, were set according to the specifications provided in their respective publications and accounting for two missed cleavage sites. The false discovery rate (FDR) at both the peptide spectrum match (PSM) and protein levels was

set to 1%. The remaining parameters of MaxQuant were set to the default values: a maximum of 5 modifications per peptide, a minimum peptide length of 7 amino acids, and a maximum peptide mass of 4,600 Da. For the “match between runs” option, a minimum match time window of 0.7 seconds and a minimum retention time alignment window of 20 seconds were applied. MaxQuant parameter files can be downloaded from Expression Atlas. The *Sus scrofa* UniProt [16] Reference proteome release-2021_04 (including isoforms, 49,865 sequences) was used as the target sequence database for the pig datasets. MaxQuant uses a built-in database of contaminants, and a decoy database was generated by reversing the input database sequences following the respective enzymatic digestion.

3. Post-processing

The post-processing of MaxQuant results followed the methodology detailed in previous publications [6]. In short, after removing the protein groups labelled as potential contaminants, decoys, and those with less than 2 PSMs, the protein intensities in each sample were normalised by scaling the iBAQ intensity values with the total signal in each MS run and converting to parts per billion (ppb).

$$ppb_iBAQ_i = (iBAQ_i / \sum n_{i=1} iBAQ_i) \times 1{,}000{,}000{,}000$$

UniProt protein accessions, from the MaxQuant output- proteinGroups.txt file, were mapped to their Ensembl gene identifiers (ENSSSCG) using the ID mapping dataset (Release 2022/03) at UniProt website (https://www.uniprot.org/id-mapping) [17]. The resulting id mapping data file (idmapping_selected.tab), by default maps UniProt protein accessions to Ensembl gene identifiers of all pig breeds (i.e., Landrace, Pietrain, etc.) rather than to the reference breed, thus leading to multiple Ensembl gene identifiers being returned per UniProt protein identifier. To resolve this, we downloaded the *Sus scrofa* model Ensembl fasta peptide dump (Release 11.1 gene set, https://ftp.ensembl.org/pub/release-110/fasta/sus_scrofa/pep/Sus_scrofa.Sscrofa11.1.pep.all.fa.gz) and used this as a filter to

keep only gene-mappings specific to the reference pig breed. We used the reference pig Ensembl gene identifiers for further downstream analysis.

During downstream post-processing we removed protein groups which mapped to more than one gene identifier and for cases where two or more protein groups mapped to the same gene identifier, protein intensities were aggregated using the median value. 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.

4. Integration into Expression Atlas

The normalised protein abundances along with the validated SDRF files, summary files detailing the post-processing quality assessment and MaxQuant parameter files (mqpar.xml) are available to download from Expression Atlas. Table 1 describes the datasets and their corresponding E-PROT identifiers.

5. Protein abundance comparison across datasets

The normalised protein abundances (in ppb values) within each dataset were transformed into ranked bins as described in [7]. Briefly, normalised protein abundance (ppb) of each MS run was sorted from lowest to highest and binned into 5 equal length bins. Proteins ranked in the lowest bin (bin 1) represent the lowest abundance, and correspondingly proteins ranked in bin 5 have the highest abundance. To analyse and compare the data effectively, protein abundances from 'tissues' were grouped into 'organs'. For example, the Ileum and jejunum ‘tissues’ were grouped as ‘small intestine’. Similarly, triceps, biceps femoris, diaphragm and skeletal muscle were grouped as ‘muscle’. When combining tissues into organs, median bin values were used.

Proteins detected in at least 50% of the samples were selected for UMAP (Uniform Manifold Approximation and Projection) [18] representation and analysed for binned abundance values using the R programming language (https://www.R-project.org/). Pearson correlation coefficients (rp) were calculated for all samples based on paired complete observations and used to generate a heatmap. Missing values were marked as NA (not available). For each organ, the median rp was calculated from all paired rp values of its respective sample. Columns and rows of the samples were clustered hierarchically using Euclidean distances.

6. Comparison of protein expression values between pig and human tissues

The orthologous genes for pig and human were obtained following the procedure described in Ensembl BioMart [19]. Briefly, we first selected the “Ensembl genes 110”, then chose “Human genes (GRCh38.p14)”, clicked on “Filters” in the left menu, then unfold the “MULTI SPECIES COMPARISONS” box, ticked the “Homolog filters” option and chose “Orthologous Pig Genes” from the drop-down menu. Then, we clicked on “Attributes” in the left menu and unfold the “Pig ORTHOLOGS” box and selected the pig gene ID and pig gene name. Finally, we clicked on the “Results” button (top left) to download the orthologous genes between human and pig. The orthologous gene list was filtered to include only parent gene identifiers from pig samples in this study and the parent genes of human samples described in our previous study using human baseline tissue samples [7].

We used the calculation of “edit distance” [7] of a protein, which was computed as the difference between two pairs of protein abundance bins in pigs and humans. The following categories were used to classify their groups of protein expression samples: (1) "Group A": protein abundance is similar between human and pig tissues; (2) "Group B": protein abundance is higher in human tissues when compared to pig tissues; (3) "Group C ": Protein abundance is higher in pig tissues when compared to human tissues.

Enriched gene ontology (GO) terms analysis was performed using the mapped gene lists for each organ in each group ("Group A", "Group B", or "Group C") as foreground. The computational analysis was carried out in the R programming language with the package clusterProfiler [20] version 3.16.1, by using the function enrichGO() for the GO over-representation test, setting the gene list of all three groups as the background. The pvalueCutoff was set to 0.05 and the qvalueCutoff to 0.05.

The one-to-one mapped ortholog identifiers were used to compare pig and human protein intensities. Additionally, their normalised protein abundance (using ppb values) in 10 organs (adipose tissue, brain, colon, heart, kidney, liver, lung, pancreas, small intestine, and spleen) was used to assess pairwise correlations. Linear regression was calculated using the linear fit 'lm' method in the R programming language.

7. Correlation between gene (RNA-seq) and protein expression

One pig RNA-seq experimental baseline dataset (the only one available) was obtained from Expression Atlas (dataset E-MTAB-5895). The dataset was composed of pig samples from a Duroc breed [21]. Transcriptomics data had been previously collated in Expression Atlas, and FPKMS (Fragments Per Kilobase of transcript per Million mapped reads) data were computed by iRAP (https://github.com/nunofonseca/irap) based on the raw data, which were first averaged based on technical replicates, and then quantile normalised within each set of biological replicates using limma [22] and finally averaged again over all biological replicates. Biological metadata were collected in the SDRF format, consistent with the proteomics data.

8. Comparison of protein abundance with spectral counting values from PaxDB

We compared the protein abundances generated in our study with the protein abundance data from PaxDB version 5.0 (https://www.pax-db.org/) [23] available for *Sus scrofa*. Normalised iBAQ abundances were compared with the spectral counting abundances for liver, the only matching organ available. This comparison was not possible for other pig

organs as other data in PaxDB are labelled as ‘whole organism’. Ensembl gene ids (ENSSSCG) were mapped to protein ids (ENSSSCP) in PaxDB using the Ensembl BioMart, as described in this tutorial [19].

Results

1. Pig proteomics datasets

In summary, we obtained protein expression data from 20 healthy tissues in 14 organs, coming from 14 public datasets. The analyses covered a total of 599 MS runs from 128 samples that were annotated as healthy/control/non-treated samples, thus representing baseline protein expression. Non-control/disease samples associated with these datasets were also analysed but are not discussed here. Normalised protein abundances values (as ppb) from both control/healthy/non-treated and disease/treated tissue samples are available to view as heatmaps in Expression Atlas. The protein abundances along with sample annotations, the sample quality assessment summary and experimental parameter inputs for MaxQuant can be downloaded from Expression Atlas as text files. The total number of proteins and peptides identified in these datasets are shown in Table 1.

Expression Atlas accession number	Proteomics dataset identifier*	Breed	Tissues	Organs	Mass spectrometer	Number of MS runs	Number of samples	Fraction	Number of protein groups†	Number of peptides †	Number of unique peptides †	Number of unique genes (canonical proteins) mapped†
E-PROT-111	PXD001800 [24]	Landrace	retina	eye	Q Exactive Plus	48	16	yes	4,579	36,992	28,055	3,643
E-PROT-113	PXD002918 [25]	NA	biceps femoris	muscle	LTQ Orbitrap XL	6	6	no	1,272	9,148	6,163	841
E-PROT-114	PXD003204 [26]	Landrace	bile duct, heart, spleen, lung, brain, liver, diaphragm, pancreas, kidney	biliary system, heart, spleen, lung, brain, liver, diaphragm, pancreas, kidney	LTQ Orbitrap Elite	108	9	Yes	5,969	85,655	57,338	5,065
E-PROT-115	PXD009577 [27]	NA	heart	heart	Q Exactive HF	30	6	yes	4,258	47,975	32,956	3,582

E-PROT-117	PXD011360 [28]	Landrace × Piétrain	cecum, colon, Ileum, Jejunum	intestine	Q Exactive Plus	16	16	no	4,716	55,392	41184	3,981
E-PROT-118	PXD011536 [29]	NA	liver	liver	Q Exactive HF-X	5	5	no	2,356	18,816	15,078	1,773
E-PROT-119	PXD011755 [30]	House swine	eye	eye	LTQ Orbitrap XL	204	12	yes	1,337	7,039	5,660	888
E-PROT-120	PXD012636 [31]	NA	heart	heart	Q Exactive HF	120	4	yes	7,747	151,177	94,341	6,982
E-PROT-122	PXD014893 [32]	NA	biceps femoris, triceps muscles	muscle	Q Exactive HF-X	8	4	yes	1,287	13,065	8,527	1,004
E-PROT-124	PXD016003 [33]	NA	biceps femoris, triceps muscles	muscle	Q Exactive HF-X	8	4	yes	2,379	22,409	15,650	1,885
E-PROT-125	PXD017671 [34]	NA	liver	liver	Q Exactive HF-X	8	8	no	3,128	32,215	24,585	2,576
E-PROT-126	PXD019852 [35]	NA	heart	heart	Q Exactive HF-X	8	8	no	1,527	13,511	10,085	1,143
E-PROT-130	PXD026910 [36]	NA	subcutaneous adipose tissue, mesenteric adipose tissue	adipose tissue	Q Exactive HF	10	10	no	2,586	23,718	17,676	2,107
E-PROT-131	PXD027772 [37]	Landrace ×Swabian-Hall	skeletal muscle, heart	skeletal muscle, heart	Q Exactive HF	20	20	no	3,242	39,140	25,993	2,797
TOTAL	14 datasets		20 tissues	14 organs		599 MS runs	128 samples					

Table 1. List of pig proteomics datasets and their main characteristics.

2. Protein coverage across organs and datasets

A total of 7,767 protein groups were identified from the reanalysis of the 14 pig datasets, among which 2,164 protein groups (27.9%) were uniquely present in only one organ and 523 protein groups (6.7%) were ubiquitously observed (Table S1 in Supporting File 2). However, it should be emphasised that a specific list of typical proteins detected in only one organ should be treated with caution, as the FDR of this list will be amplified due to the accumulation of false positives when the datasets were analysed individually. For proteins detected in four or more datasets, this should not be a problem, as from the common

number of decoy protein hits across datasets, a protein FDR of less than 1% could be inferred for those proteins (Figure S1 in Supporting File 1).

Protein groups were mapped to 7,780 genes (which are equivalent to canonical proteins, the term that we will be using from now on in the manuscript). The largest number of canonical proteins was detected in samples from heart (6,264, 80.5% of the total) and the lowest number in samples from adipose tissue (1,913, 24.6%) and from the biliary system (1,983, 25.5%) (Figure 1A). The lower number of proteins identified in the biliary system could be attributed to the smallest sample size (only one sample out of 128, 0.08%). Dataset PXD012636, a dataset containing pig heart samples, which was fractionated, provided the highest number of detected canonical proteins (6,062, 77.9%), whereas the smallest number of proteins were detected in dataset PXD002918 (biceps femoris, 789, 10.1%, non-fractionated dataset) (Figure 1B).

We studied the normalised protein abundance distribution in organs (Figure 1C) and found that all organs had similar median abundances. However, one cannot attribute biological meaning to these observations, since the method of normalisation by definition fixes each sample to have the same “total abundance”, which then gets shared out amongst all proteins. The normalised protein abundance distribution in datasets indicated a lower than median abundance detected in the dataset PXD012636 (heart), as a direct result of more proteins being detected overall in this dataset (Figure 1D). In terms of distribution of proteins detected per organ, most proteins were found in just one organ (Figure 1E).

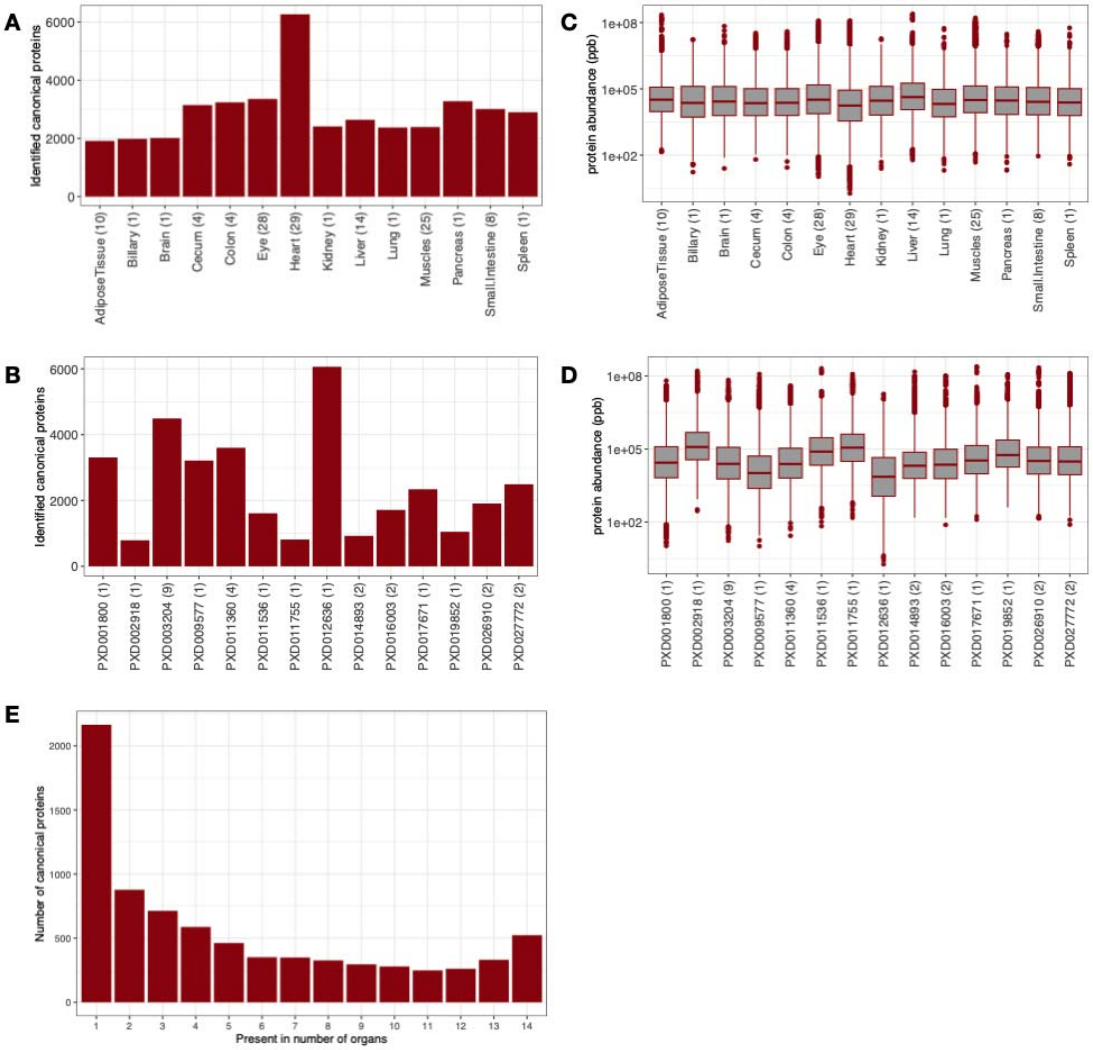
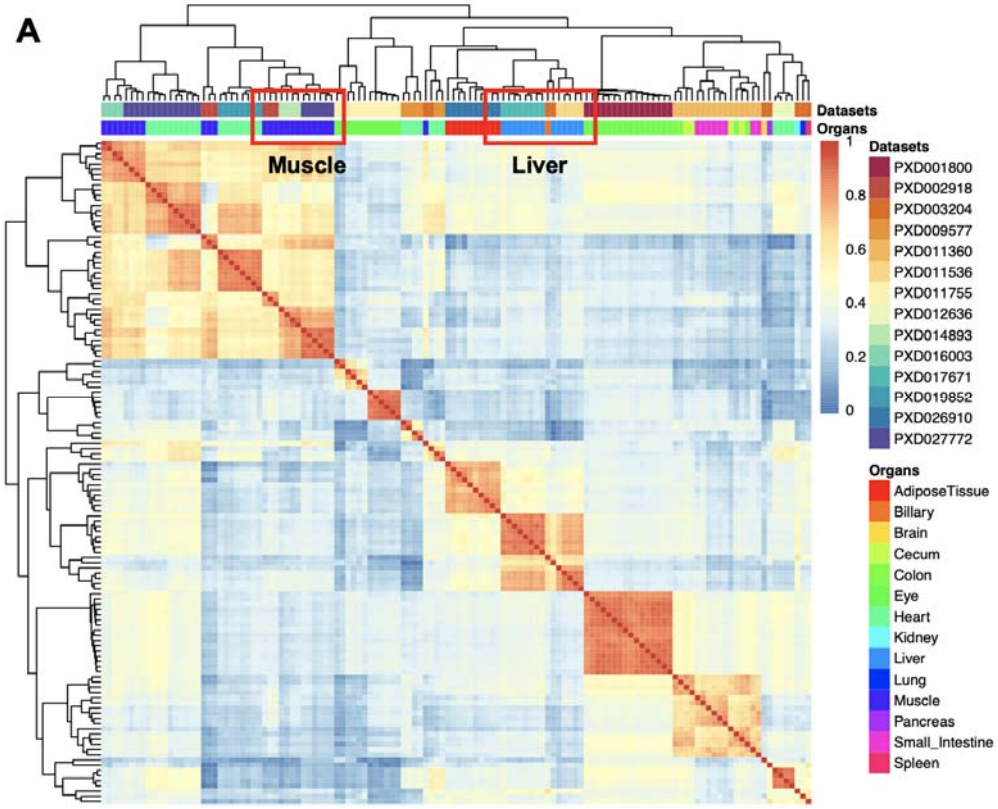


Figure 1. Distribution of canonical proteins detected per organ and dataset. (**A**) Number of canonical proteins identified across different pig organs. The number in brackets denotes the number of samples. (**B**) Number of canonical proteins identified in different datasets. The number in the brackets indicates the number of unique tissues in the dataset. (**C**) Range of normalised iBAQ protein abundances across different organs. The number in brackets indicates the number of samples. (**D**) Range of normalised iBAQ protein abundances across different datasets. The number in brackets indicates the number of unique tissues in the dataset. (**E**) Distribution of canonical proteins identified across different organs.

3. Protein abundance comparison across organs and datasets

Next, we studied how protein abundances compared across different datasets and organs. To make protein abundance values more comparable between datasets, we transformed the normalised iBAQ intensities into ranked bins as explained in 'Methods': i.e., proteins included in bin 5 are highly abundant whereas proteins included in bin 1 are expressed in the lowest abundances (among the detected proteins). We found that 494 (6.3%) proteins were expressed in at least 3 organs, with a median bin value greater than 4 (not including 4). At the other end of the scale, 337 (4.3%) canonical proteins were expressed in at least 3 organs at a median bin value less than 2 (not including 2). The bin transformed abundances in all organs and in all datasets are provided in the Table S2 and S3 in Supporting File 2.

To compare protein expression across all pig organs, we calculated pairwise Pearson correlation coefficients (rp) for the 128 samples (Figure 2A). We observed a good correlation of protein expression within the liver (median rp = 0.77) and muscle (median rp = 0.65) samples. We then performed a cluster analysis using UMAP [18] on all samples to test the effectiveness of the bin-transformed method in reducing batch effects (Figure 2B). We observed that samples from different datasets belonging to the same organ were generally clustered together. For example, liver samples from datasets PXD003204, PXD011536 and PXD017671 clustered together (colour blue in Figure 2B). Additionally, muscle samples from PXD002918 (tissue biceps femoris), PXD003204 (tissue diaphragm), PXD014893 (tissue biceps femoris and triceps muscle) and PXD016003 (tissue biceps femoris, triceps muscle) clustered together too (colour purple). Similarly, heart samples from datasets PXD003204, PXD009577, PXD012636 and PXD019852 clustered together as well (colour dark green).


A
Muscle
Liver
Datasets
Organs
1
0.8
0.6
0.4
0.2
0
Datasets
PXD001800
PXD002918
PXD003204
PXD009577
PXD011360
PXD011536
PXD011755
PXD012636
PXD014893
PXD016003
PXD017671
PXD019852
PXD026910
PXD027772
Organs
AdiposeTissue
Billary
Brain
Cecum
Colon
Eye
Heart
Kidney
Liver
Lung
Muscle
Pancreas
Small_Intestine
Spleen

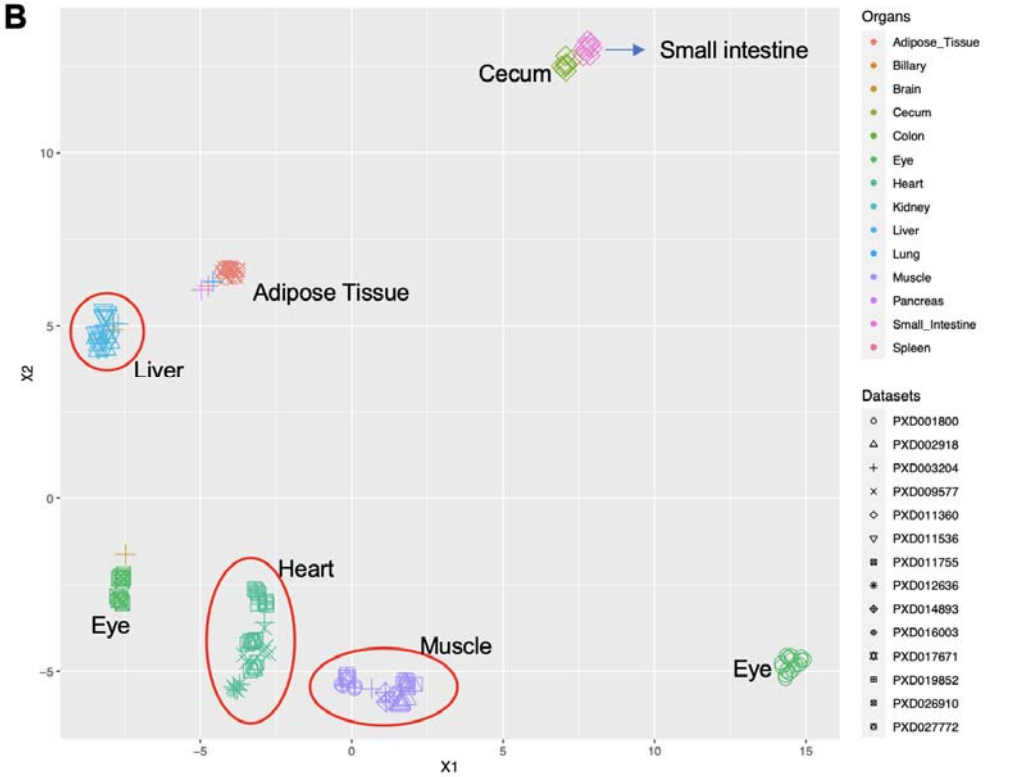

B
Small intestine
Cecum
Adipose Tissue
Liver
Heart
Eye
Muscle
Eye
X1
X2
10
5
0
-5
-5
0
5
10
15
Organs
Adipose_Tissue
Billary
Brain
Cecum
Colon
Eye
Heart
Kidney
Liver
Lung
Muscle
Pancreas
Small_Intestine
Spleen
Datasets
PXD001800
PXD002918
PXD003204
PXD009577
PXD011360
PXD011536
PXD011755
PXD012636
PXD014893
PXD016003
PXD017671
PXD019852
PXD026910
PXD027772

Figure 2. A. Heatmap of pairwise Pearson correlation coefficients across all pig samples among datasets and organs. The colours on the heatmap represent the correlation coefficients calculated using the bin transformed values. Hierarchical clustering of the columns and rows of the samples was performed using Euclidean distances. **B.** UMAP representation among different datasets and organs. Groups of datasets coming from the same organ are highlighted.

4. Comparative analysis of pig and human protein expression

We also performed a comparative analysis of protein abundances (in bins) between the pig baseline tissue datasets with the results obtained in our previous analogous study involving human datasets coming from baseline tissues, which was performed using the same overall methodology [7] (Table S4 in Supporting File 2). Pig is often used as a model organism for human biomedical research and then it is interesting to compare protein expression in both organisms. We calculated metrics to study the differences between the protein abundances for all organs found in the two studies (see 'Material and Methods' for full details), as shown in Figure 3 (also see Table S5 in Supporting File 2). Three groups of proteins were found according to their protein expression levels: (i) "Group A": protein expression is similar between human and pig tissues; (ii) "Group B": protein expression is higher in human tissues; and (iii) "Group C ": protein expression is higher in pig tissues.

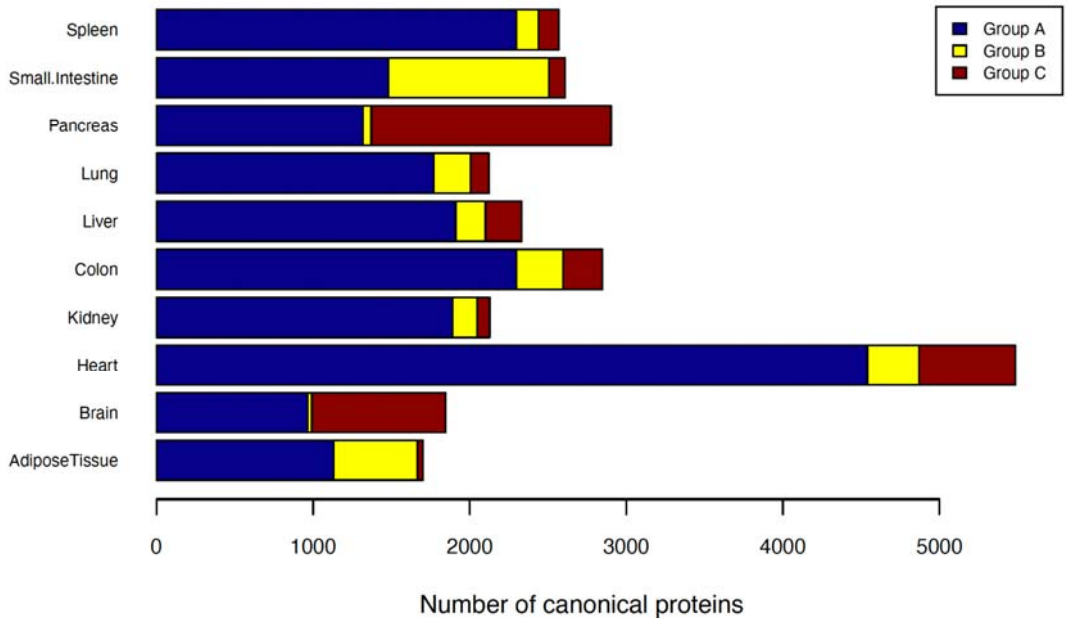


Figure 3. Organ specificity of canonical proteins based on edit distances between pig and human. The canonical proteins detected both in pig and human samples, were classified into three groups: “Group A” (similar protein expression levels between human and pig), “Group B” (higher protein expression in human tissues) and “Group C” (higher protein expression in pig tissues).

We found that for pig, protein expression levels were higher in pancreas, brain, and heart than in the corresponding human tissues, whereas protein expression levels were higher in the human's small intestine and adipose tissue, when compared to the corresponding pig tissues. For other organs, the number of proteins in “Group B” and “Group C” were quite similar. Since different sizes and counts of datasets have been used for the different organs in both species, the organ-level trends reflect the results found in our studies. In our view, they are not necessarily meaningful for understanding species level differences. Instead, for individual proteins, or groups of related proteins, the comparison gives a potentially useful guide to relative protein abundance between orthologous pairs in individual organs.

We then performed a GO enrichment analysis [38] per organ of proteins included in the three groups with using GO terms related to biological processes. We found 75 GO terms to be statistically significant in all organs overall (see all enriched GO terms in Table S6 in Supporting File 2), and in particular in brain, colon, small intestine, heart and spleen. For instance, in brain we found processes enriched for 'head development' (GO:0060322), 'central nervous system development' (GO:0007417) and 'vesicle-mediated transport in synapse' (GO:0099003), among others. We also observed 'mitochondrial gene expression' (GO:0140053), 'mitochondrial translation' (GO:0032543), and 'mitochondrial respiratory chain complex assembly' (GO:0033108) enriched in the small intestine. For the whole list, also including the GO terms enriched for other organs, see Table S6 in Supporting File 2.

5. Comparison of protein abundances across orthologs between pig and human datasets

In a previous study we compared the expression of canonical proteins found in three different species: human, mouse and rat [6]. Here, we used the same approach to compare canonical protein expression between humans [7] and pig organs. Overall, 13,248 detected human canonical proteins were compared with 7,800 detected pig canonical proteins (Table S7 in Supporting File 2). When comparing protein abundance (in ppb), we only considered the corresponding orthologous genes with unambiguous (one-to-one) mappings, which resulted in 6,811 common protein orthologs.

When comparing the protein expression of orthologs in human and pig, we observed a relatively overall high correlation in protein abundance in heart (R^2 = 0.62) and liver (R^2 = 0.53), medium correlation for brain (R^2 = 0.44), colon (R^2 = 0.42), kidney (R^2 = 0.40) and

spleen (R^2 = 0.39), and low correlation in small intestine (R^2 = 0.18), lung (R^2 = 0.21), pancreas (R^2 = 0.26), and adipose tissue (R^2 = 0.27) (Figure 4).

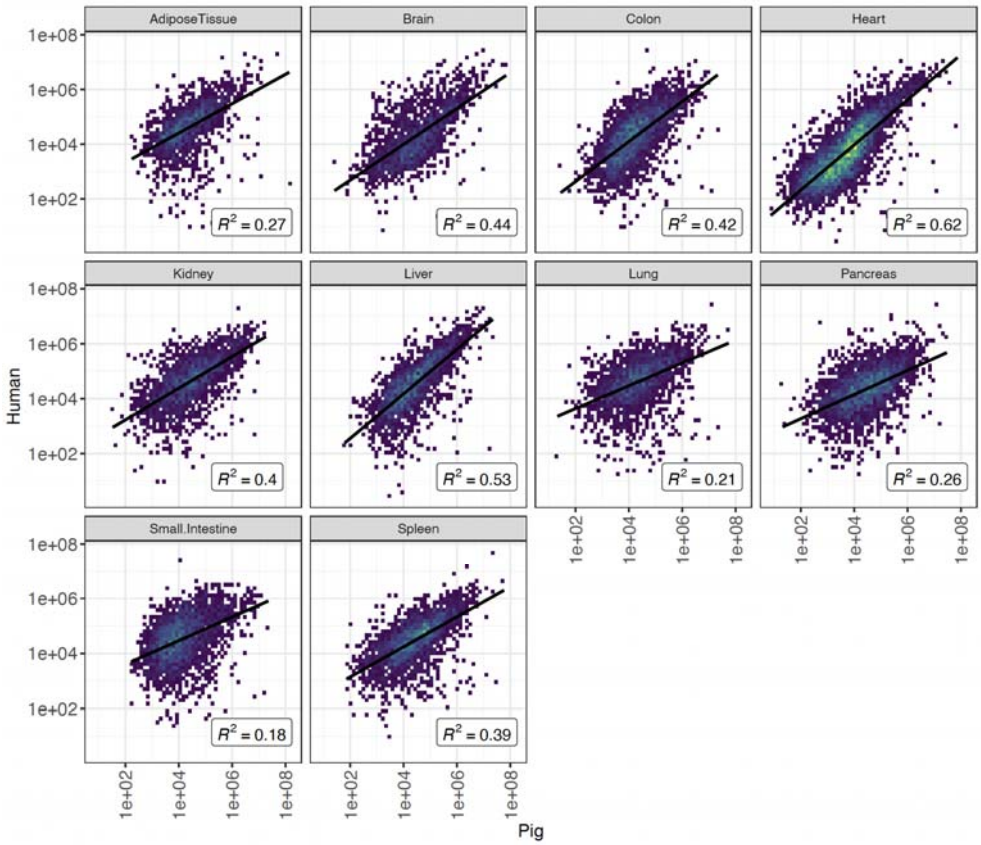


Figure 4. Comparison of protein abundance between orthologs of pig and human in various organs.

We also investigated the correlation of protein expression across different organs in pigs across the datasets used in this study (Figure 5). The data sometimes follow expected distributions whereby tissues predicted to be more similar to each other in terms of biological function have higher pairwise correlations e.g., small intestine and colon (R^2 = 0.87) and spleen and pancreas (R^2 = 0.78). However, other high correlations were found between, between lung and spleen (R^2 = 0.83) and lung and pancreas (R^2 = 0.71) - likely due to these data originating from a single study. The correlation between the brain and the remaining

organs was generally low, as it would be expected. A plotted representation of the abundance of all sorted proteins for both species is provided in Supporting File 3.

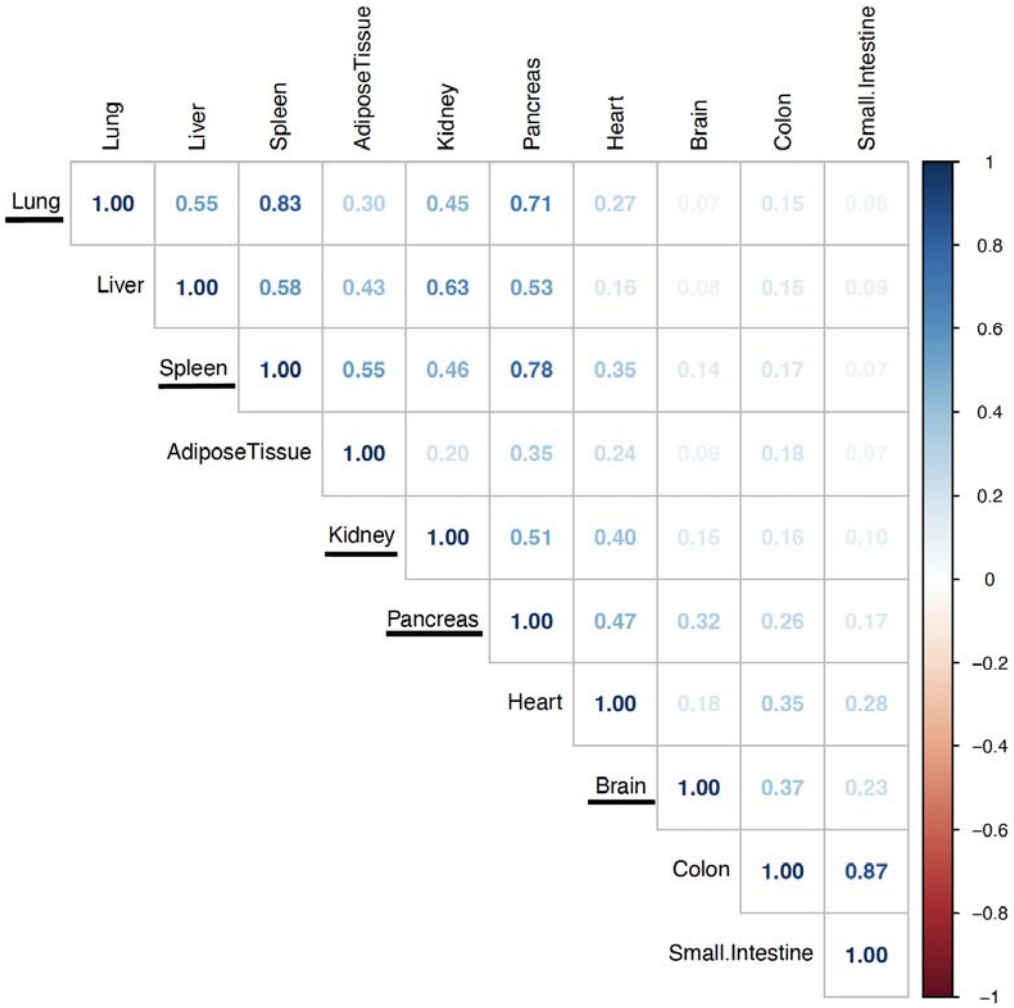


Figure 5. Protein expression correlation among all pig organs across the datasets included in this study. The organs underlined all come from a single dataset (PXD003204).

6. Comparison between gene (RNA-seq) and protein expression, and comparison with protein expression data in PaxDB

We also investigated the correlation between gene (RNA-seq based) and protein expression in baseline tissue pig datasets. For that we used the only suitable dataset available in Expression Atlas (dataset E-MTAB-5895) [21]. We compared the normalised iBAQ protein abundances (ppb) with the baseline RNA-seq expression (FPKM). To compare expression across different organs we grouped the RNA-seq expression from various tissues into their respective organs by using their median values. We did not observe a strong correlation between protein and RNA expression across various organs (Figure S2 in Supporting File 1). The lowest correlation between protein and RNA expression was observed in lung (R^2 = 0.06) and the highest was observed in heart (R^2 = 0.20). In our view, these low correlations could be due to the inherent limitations in this comparison i.e. (i) samples are not paired; and (ii) the different pig breeds used (Duroc for the RNA-Seq study and several others for the protein expression studies, see details in Table 1).

In addition, we compared the protein abundance of organ liver generated in this study with data from the PaxDB resource generated using the spectral counting method. We observed that protein abundance values (fraction of total (FOT) normalised ppb values) calculated using iBAQ in this study correlated till a limited extent (R^2 = 0.47) with the semi-quantitative values available in PaxDB for (Figure S3 in Supporting File 1). Unfortunately, it was not possible to perform this correlation for other organs (see details in 'Material and Methods').

Discussion

We have previously performed three meta-analysis studies involving the reanalysis and integration in Expression Atlas of public quantitative proteomics datasets coming from cell lines and human tumour samples [8], from human baseline tissues [7], and from mouse and rat baseline tissues [6]. Here we have reanalysed 14 public proteomics datasets coming from pig tissues in baseline conditions. Our overall aim was to provide a system-wide baseline protein expression catalogue across various pig organs. We have used the same

methodology as in the study involving baseline human tissues (and in the mouse/rat study), which enabled a comparison of protein expression levels across human and pig organs. To the best of our knowledge this is the first metanalysis study for pig at a protein expression level, in this case using label-free DDA data. The resource PaxDB version 5.0 [12] includes pig data generated using spectral counting, but their granularity is limited, providing only organ-specific expression for liver.

As done before, we reanalysed each dataset separately using MaxQuant and the same 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. However, as also explained before [6, 7], using a dataset per dataset analysis approach is in our view the only sustainable manner to reanalyse and integrate quantitative proteomics datasets in resources such as Expression Atlas, where gene expression datasets are stored following the same dataset per dataset approach. It is also important to highlight that the number of commonly detected protein false positives is reduced in parallel with the increase in the number of common datasets where a given protein is detected. In this case, for proteins detected in four or more datasets, a protein FDR of less than 1% can be inferred (Figure S1 in Supporting File 1).

This overall study of protein expression is relevant in different contexts. First of all, systems biology research of the domestic pig is of immediate relevance for food production and animal welfare. Additionally, the domestic pig is a model organism for human biomedical research. Furthermore, the diversity of available pig models is rapidly expanding. Some possible applications of these models are research in nutrition, inflammation, and host–microbial crosstalk. In this context, pig models present great opportunities, because the pig, like humans, is an omnivore, with very similar nutritional requirements, digestive and immune systems, and gut microbial components [10]. It is also important to highlight that

mini-pig's models are being increasingly used in drug development. Animals are still requested in the safety testing of new drug candidates, and minipigs are a potential non-rodent alternative to the use of non-human-primates (NHP), due to ethical considerations [39]. The use of alternative *in vitro* models is still challenging due to complex biological responses in various organ systems following drug-treatment.

Future directions in analogous studies will involve: (i) the inclusion of additional species, e.g. other model organisms or other species of economic importance; (ii) studies focused on particular diseases or physiological states; (iii) the inclusion of differential proteomics datasets in addition to baseline studies; and (iv) reanalysis of DIA datasets (e.g. [40]). In conclusion we here present a meta-analysis study of public pig baseline proteomics datasets from the PRIDE database. We also performed a comparative analysis across human and pig protein abundance. The resulting protein expression data has been made available *via* Expression Atlas.

Data availability

Expression Atlas E-PROT identifiers and the PRIDE original dataset identifiers used for the reanalysis are included in Table 1.

Acknowledgements

We would like to thank all data submitters who made their datasets available via PRIDE and ProteomeXchange. This work has been funded by Open Targets, BBSRC [BB/T019670/1 and BB/T019557/1] and EMBL core funding. We would also like to thank Andrew Leach and the rest of the team involved in the Open Targets "Target Safety" project, for helpful discussions.

Supporting information

- Supporting File 1:
  - Figure S1: Distribution of common reverse decoy hits across the number of datasets.
  - Figure S2: Correlation between gene (RNA-seq based) and protein expression in baseline tissue pig datasets.
  - Figure S3: Correlation of protein abundances across organ liver compared between PaxDB and the results found in this study.
- Supporting File 2:
  - Table S1: Median protein abundances (in ppb) for each protein group across various tissue samples included in each organ.
  - Table S2: Median binned protein abundances across various tissue samples in each pig organ.
  - Table S3: Median binned protein abundances across various pig datasets.
  - Table S4: Median binned protein abundances of human and pig orthologs across all organs.
  - Table S5: Elevated proteomes of three different groups in various organs after applying edit distance between human and pig homologous genes.
  - Table S6: Gene Ontology enrichment analysis of the three different groups of proteins, considering their level of expression in pig and human tissues.
  - Table S7: Protein abundances (ppb) considering only one to one mapping between human and pig orthologs across all organs.
- Supporting File 3: Figure illustrating the binned protein abundances of all one-to-one mapped orthologs across ten common organs in human and pig.

Abbreviations

DDA: Data Dependent Acquisition

DIA: Data Independent Acquisition

FDR: False Discovery Rate

GO: Gene Ontology

iBAQ: intensity-based absolute quantification

IDF: Investigation Description Format

MS: Mass Spectrometry

NHP: Non-Human Primate

ppb: parts per billion

PCA: Principal Component Analysis

SDRF: Sample and Data Relationship Format

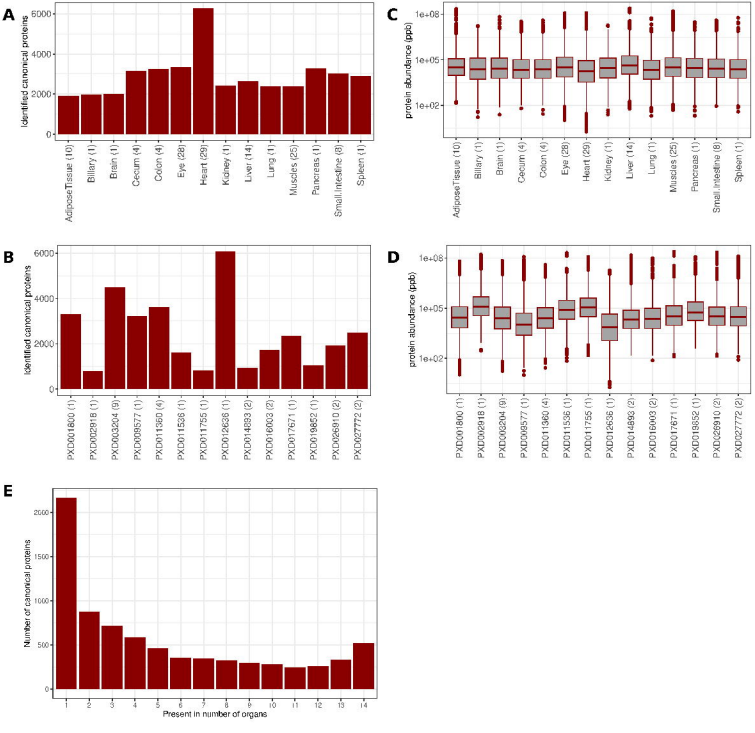
UMAP: Uniform Manifold Approximation and Projection

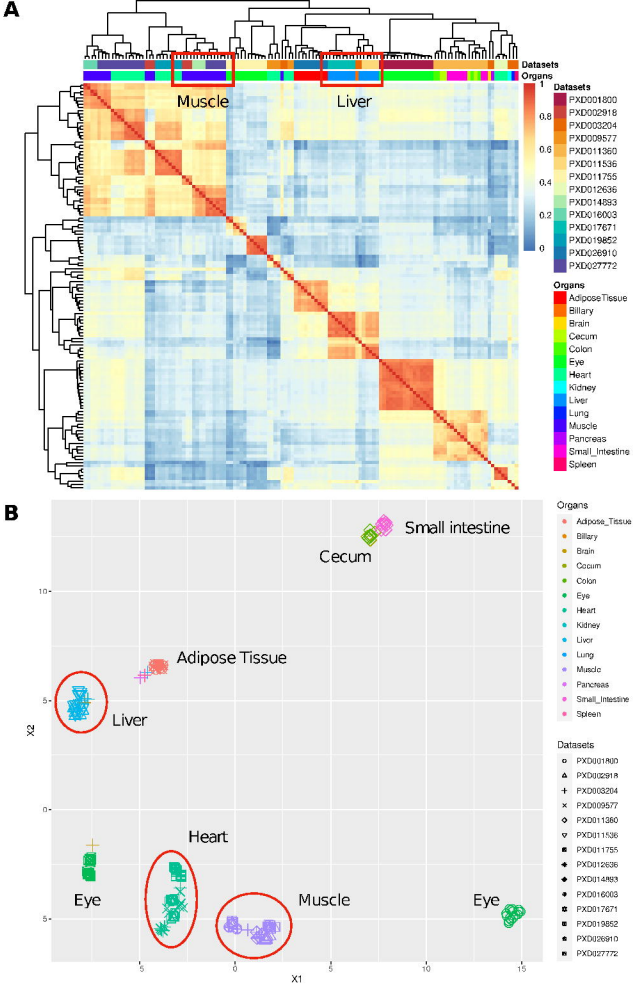
References

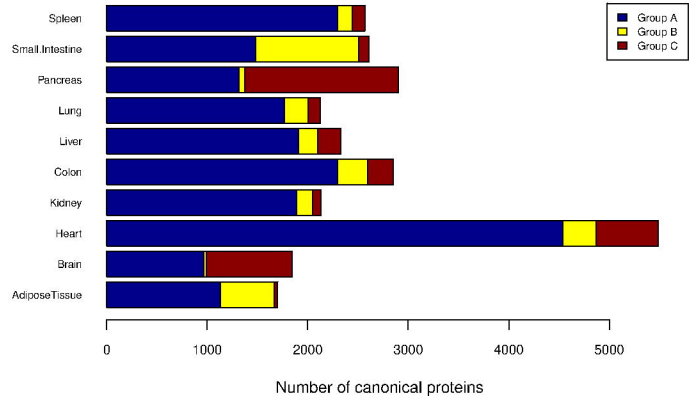
1. Messner CB, Demichev V, Wang Z, Hartl J, Kustatscher G, Mulleder M, et al. Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology. Proteomics. 2023;23 7-8:e2200013. doi:10.1002/pmic.202200013.
2. Eliuk S and Makarov A. Evolution of Orbitrap Mass Spectrometry Instrumentation. Annu Rev Anal Chem (Palo Alto Calif). 2015;8:61-80. doi:10.1146/annurev-anchem-071114-040325.
3. Guan S, Taylor PP, Han Z, Moran MF and Ma B. Data Dependent-Independent Acquisition (DDIA) Proteomics. J Proteome Res. 2020;19 8:3230-7. doi:10.1021/acs.jproteome.0c00186.
4. Perez-Riverol Y, Bai J, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 2022;50 D1:D543-D52. doi:10.1093/nar/gkab1038.
5. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014;32 3:223-6. doi:10.1038/nbt.2839.
6. Wang S, Garcia-Seisdedos D, Prakash A, Kundu DJ, Collins A, George N, et al. Integrated view and comparative analysis of baseline protein expression in mouse and rat tissues. PLoS Comput Biol. 2022;18 6:e1010174. doi:10.1371/journal.pcbi.1010174.
7. Prakash A, Garcia-Seisdedos D, Wang S, Kundu DJ, Collins A, George N, et al. Integrated View of Baseline Protein Expression in Human Tissues. J Proteome Res. 2023;22 3:729-42. doi:10.1021/acs.jproteome.2c00406.
8. Jarnuczak AF, Najgebauer H, Barzine M, Kundu DJ, Ghavidel F, Perez-Riverol Y, et al. An integrated landscape of protein expression in human cancer. Sci Data. 2021;8 1:115. doi:10.1038/s41597-021-00890-2.
9. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Res. 2022;50 D1:D129-D40. doi:10.1093/nar/gkab1030.
10. Hesselager MO, Codrea MC, Sun Z, Deutsch EW, Bennike TB, Stensballe A, et al. The Pig PeptideAtlas: A resource for systems biology in animal production and biomedicine. Proteomics. 2016;16 4:634-44. doi:10.1002/pmic.201500195.
11. Bassols A, Costa C, Eckersall PD, Osada J, Sabria J and Tibau J. The pig as an animal model for human pathologies: A proteomics perspective. Proteomics Clin Appl. 2014;8 9-10:715-31. doi:10.1002/prca.201300099.
12. Huang Q, Szklarczyk D, Wang M, Simonovic M and von Mering C. PaxDb 5.0: Curated Protein Quantification Data Suggests Adaptive Proteome Changes in Yeasts. Mol Cell Proteomics. 2023;22 10:100640. doi:10.1016/j.mcpro.2023.100640.
13. Athar A, Fullgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. 2019;47 D1:D711-D5. doi:10.1093/nar/gky964.
14. Dai C, Fullgrabe A, Pfeuffer J, Solovyeva EM, Deng J, Moreno P, et al. A proteomics sample metadata representation for multiomics integration and big data analysis. Nat Commun. 2021;12 1:5854. doi:10.1038/s41467-021-26111-3.
15. Cox J and Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26 12:1367-72. doi:10.1038/nbt.1511.
16. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51 D1:D523-D31. doi:10.1093/nar/gkac1052.

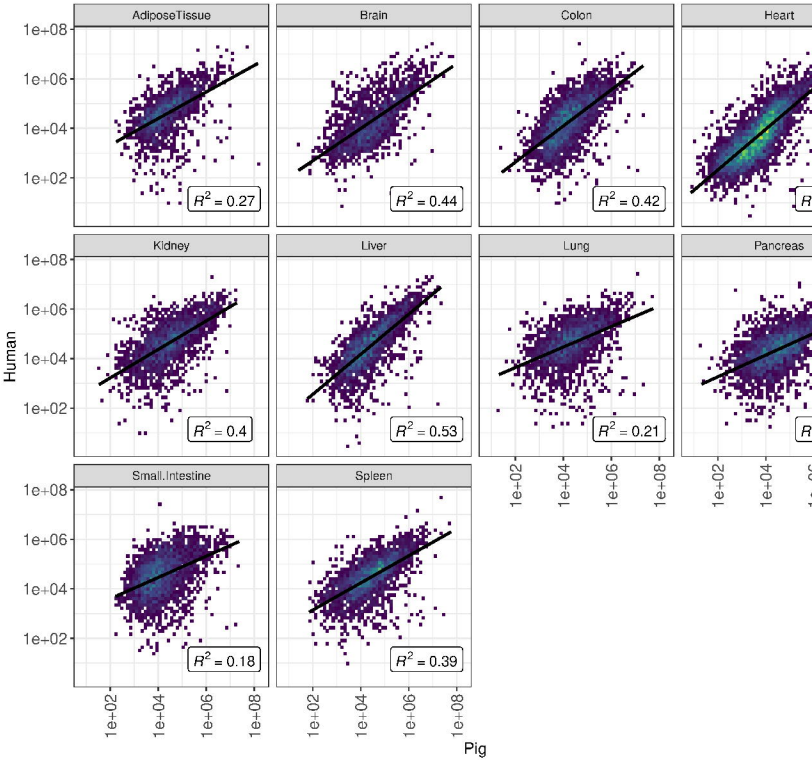
17. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50 D1:D988-D95. doi:10.1093/nar/gkab1049.
18. Leland McInnes JH, James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018; doi:https://doi.org/10.48550/arXiv.1802.03426.
19. Herrero J: How to get all the orthologous genes between two species. https://www.ensembl.info/2009/01/21/how-to-get-all-the-orthologous-genes-between-two-species/ (2009).
20. Yu G, Wang LG, Han Y and He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16 5:284-7. doi:10.1089/omi.2011.0118.
21. Summers KM, Bush SJ, Wu C, Su AI, Muriuki C, Clark EL, et al. Functional Annotation of the Transcriptome of the Pig, Sus scrofa, Based Upon Network Analysis of an RNAseq Transcriptional Atlas. Front Genet. 2019;10:1355. doi:10.3389/fgene.2019.01355.
22. Smyth G, Hu Y, M. R and J. S: limma powers differential expression analyses for RNA-sequencing and microarray studies. https://www.bioconductor.org/packages/release/bioc/html/limma.html (2015).
23. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D and von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015;15 18:3163-8. doi:10.1002/pmic.201400441.
24. Cehofski LJ, Kruse A, Kjaergaard B, Stensballe A, Honore B and Vorum H. Proteins involved in focal adhesion signaling pathways are differentially regulated in experimental branch retinal vein occlusion. Exp Eye Res. 2015;138:87-95. doi:10.1016/j.exer.2015.06.011.
25. Frohlich T, Kemter E, Flenkenthaler F, Klymiuk N, Otte KA, Blutke A, et al. Progressive muscle proteome changes in a clinically relevant pig model of Duchenne muscular dystrophy. Sci Rep. 2016;6:33362. doi:10.1038/srep33362.
26. Marx H, Hahne H, Ulbrich SE, Schnieke A, Rottmann O, Frishman D, et al. Annotation of the Domestic Pig Genome by Quantitative Proteogenomics. J Proteome Res. 2017;16 8:2887-98. doi:10.1021/acs.jproteome.7b00184.
27. Federspiel JD, Tandon P, Wilczewski CM, Wasson L, Herring LE, Venkatesh SS, et al. Conservation and divergence of protein pathways in the vertebrate heart. PLoS Biol. 2019;17 9:e3000437. doi:10.1371/journal.pbio.3000437.
28. Troscher-Mussotter J, Tilocca B, Stefanski V and Seifert J. Analysis of the Bacterial and Host Proteins along and across the Porcine Gastrointestinal Tract. Proteomes. 2019;7 1 doi:10.3390/proteomes7010004.
29. Backman M, Flenkenthaler F, Blutke A, Dahlhoff M, Landstrom E, Renner S, et al. Multi-omics insights into functional alterations of the liver in insulin-deficient diabetes mellitus. Mol Metab. 2019;26:30-44. doi:10.1016/j.molmet.2019.05.011.
30. Schmelter C, Funke S, Treml J, Beschnitt A, Perumal N, Manicam C, et al. Comparison of Two Solid-Phase Extraction (SPE) Methods for the Identification and Quantification of Porcine Retinal Protein Markers by LC-MS/MS. Int J Mol Sci. 2018;19 12 doi:10.3390/ijms19123847.
31. Linscheid N, Santos A, Poulsen PC, Mills RW, Calloe K, Leurs U, et al. Quantitative proteome comparison of human hearts with those of model organisms. PLoS Biol. 2021;19 4:e3001144. doi:10.1371/journal.pbio.3001144.
32. Moretti A, Fonteyne L, Giesert F, Hoppmann P, Meier AB, Bozoglu T, et al. Somatic gene editing ameliorates skeletal and cardiac muscle failure in pig and human models of Duchenne muscular dystrophy. Nat Med. 2020;26 2:207-14. doi:10.1038/s41591-019-0738-2.
33. Regensburger AP, Fonteyne LM, Jungert J, Wagner AL, Gerhalter T, Nagel AM, et al. Detection of collagens by multispectral optoacoustic tomography as an imaging

biomarker for Duchenne muscular dystrophy. Nat Med. 2019;25 12:1905-15. doi:10.1038/s41591-019-0669-y.
34. Riedel EO, Hinrichs A, Kemter E, Dahlhoff M, Backman M, Rathkolb B, et al. Functional changes of the liver in the absence of growth hormone (GH) action - Proteomic and metabolomic insights from a GH receptor deficient pig model. Mol Metab. 2020;36:100978. doi:10.1016/j.molmet.2020.100978.
35. Hinrichs A, Riedel EO, Klymiuk N, Blutke A, Kemter E, Langin M, et al. Growth hormone receptor knockout to reduce the size of donor pigs for preclinical xenotransplantation studies. Xenotransplantation. 2021;28 2:e12664. doi:10.1111/xen.12664.
36. Flenkenthaler F, Landstrom E, Shashikadze B, Backman M, Blutke A, Philippou-Massier J, et al. Differential Effects of Insulin-Deficient Diabetes Mellitus on Visceral vs. Subcutaneous Adipose Tissue-Multi-omics Insights From the Munich MIDY Pig Model. Front Med (Lausanne). 2021;8:751277. doi:10.3389/fmed.2021.751277.
37. Stirm M, Fonteyne LM, Shashikadze B, Lindner M, Chirivi M, Lange A, et al. A scalable, clinically severe pig model for Duchenne muscular dystrophy. Dis Model Mech. 2021;14 12 doi:10.1242/dmm.049285.
38. Mi H, Muruganujan A, Casagrande JT and Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8 8:1551-66. doi:10.1038/nprot.2013.092.
39. Jensen VF, Jensen NK, Schefe LH, Sigh J, Akintomide A, Kaaber K, et al. The Non-Human Primate in Safety Assessment of a Bifunctional Long-Acting Insulin Analogue. Int J Toxicol. 2023;42 3:254-68. doi:10.1177/10915818231156898.
40. Walzer M, Garcia-Seisdedos D, Prakash A, Brack P, Crowther P, Graham RL, et al. Implementing the reuse of public DIA proteomics datasets: from the PRIDE database to Expression Atlas. Sci Data. 2022;9 1:335. doi:10.1038/s41597-022-01380-9.


A
Identified canonical proteins
6000
4000
2000
0
AdiposeTissue (10)
Billiary (1)
Brain (1)
Cecum (4)
Colon (4)
Eye (28)
Heart (29)
Kidney (1)
Liver (14)
Lung (1)
Muscles (25)
Pancreas (1)
Small intestine (8)
Spleen (1)
C
protein abundance (ppb)
1e+08
1e+05
1e+02
AdiposeTissue (10)
Billiary (1)
Brain (1)
Cecum (4)
Colon (4)
Eye (28)
Heart (29)
Kidney (1)
Liver (14)
Lung (1)
Muscles (25)
Pancreas (1)
Small intestine (8)
Spleen (1)
B
Identified canonical proteins
6000
4000
2000
0
PXD001800 (1)
PXD002918 (1)
PXD003204 (9)
PXD009577 (1)
PXD011360 (4)
PXD011538 (1)
PXD011755 (1)
PXD012638 (1)
PXD014893 (2)
PXD016003 (2)
PXD017671 (1)
PXD019852 (1)
PXD026910 (2)
PXD027772 (2)
D
protein abundance (ppb)
1e+08
1e+05
1e+02
PXD001800 (1)
PXD002918 (1)
PXD003204 (9)
PXD009577 (1)
PXD011360 (4)
PXD011538 (1)
PXD011755 (1)
PXD012638 (1)
PXD014893 (2)
PXD016003 (2)
PXD017671 (1)
PXD019852 (1)
PXD026910 (2)
PXD027772 (2)
E
Number of canonical proteins
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Present in number of organs


A
Datasets
Organs
Muscle
Liver
1
0.8
0.6
0.4
0.2
0
Datasets
PXD001800
PXD002918
PXD003204
PXD009577
PXD011360
PXD011536
PXD011755
PXD012636
PXD014893
PXD016003
PXD017671
PXD019852
PXD026910
PXD027772
Organs
AdiposeTissue
Billary
Brain
Cecum
Colon
Eye
Heart
Kidney
Liver
Lung
Muscle
Pancreas
Small_Intestine
Spleen
B
Small intestine
Cecum
Adipose Tissue
Liver
Heart
Eye
Muscle
Eye
X1
X2
10
5
0
5
5
0
5
10
15
Organs
Adipose_Tissue
Billary
Brain
Cecum
Colon
Eye
Heart
Kidney
Liver
Lung
Muscle
Pancreas
Small_Intestine
Spleen
Datasets
PXD001800
PXD002918
PXD003204
PXD009577
PXD011360
PXD011536
PXD011755
PXD012636
PXD014893
PXD016003
PXD017671
PXD019852
PXD026910
PXD027772


Spleen
Small.Intestine
Pancreas
Lung
Liver
Colon
Kidney
Heart
Brain
AdiposeTissue
Group A
Group B
Group C
0
1000
2000
3000
4000
5000
Number of canonical proteins


AdiposeTissue
Brain
Colon
Heart
Kidney
Liver
Lung
Pancreas
Small.Intestine
Spleen
R^2 = 0.27
R^2 = 0.44
R^2 = 0.42
R^2 = 0.4
R^2 = 0.53
R^2 = 0.21
R^2 = 0.18
R^2 = 0.39
1e+08
1e+06
1e+04
1e+02
Human
Pig

	Lung	Liver	Spleen	AdiposeTissue	Kidney	Pancreas	Heart	Brain	Colon	Small.Intestine
Lung	1.00	0.55	0.83	0.30	0.45	0.71	0.27	0.07	0.15	0.08
Liver		1.00	0.58	0.43	0.63	0.53	0.16	0.08	0.15	0.09
Spleen			1.00	0.55	0.46	0.78	0.35	0.14	0.17	0.07
AdiposeTissue				1.00	0.20	0.35	0.24	0.09	0.18	0.07
Kidney					1.00	0.51	0.40	0.15	0.16	0.10
Pancreas						1.00	0.47	0.32	0.26	0.17
Heart							1.00	0.18	0.35	0.28
Brain								1.00	0.37	0.23
Colon									1.00	0.87
Small.Intestine										1.00