Open access • Posted Content • DOI:10.1101/391466

Mining unknown porcine protein isoforms by tissue-based map of proteome enhances the pig genome annotation — Source link

Pengju Zhao, Xian Zhong Zheng, Ying Yu, Zhaozheng Hou ...+12 more authors

Institutions: China Agricultural University, Northwestern Polytechnical University, Agricultural Research Service, Shanxi Agricultural University ...+2 more institutions

Published on: 14 Aug 2018 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Proteome, KEGG, Genome project, Reference genome and Genome

Mining Unknown Porcine Protein Isoforms by Tissue-Based Map of Proteome Enhances the Pig Genome Annotation

Journal Pre-proofs

ORIGINAL RESEARCH

Mining Unknown Porcine Protein Isoforms by Tissue-Based Map of Proteome Enhances the Pig Genome Annotation

Pengju Zhao, Xianrui Zheng, Ying Yu, Zhuocheng Hou, Chenguang Diao, Haifei Wang, Huimin Kang, Chao Ning, Junhui Li, Wen Feng, Wen Wang, George E. Liu, Bugao Li, Jacqueline Smith, Yangzom Chamba, Jian-Feng Liu

PII: S1672-0229(21)00035-8
DOI: https://doi.org/10.1016/j.gpb.2021.02.002
Reference: GPB 490

To appear in: *Genomics, Proteomics & Bioinformatics*

Received Date: 14 September 2018
Revised Date: 5 September 2019
Accepted Date: 29 November 2019

Please cite this article as: P. Zhao, X. Zheng, Y. Yu, Z. Hou, C. Diao, H. Wang, H. Kang, C. Ning, J. Li, W. Feng, W. Wang, G.E. Liu, B. Li, J. Smith, Y. Chamba, J-F. Liu, Mining Unknown Porcine Protein Isoforms by Tissue-Based Map of Proteome Enhances the Pig Genome Annotation, *Genomics, Proteomics & Bioinformatics* (2021), doi: https://doi.org/10.1016/j.gpb.2021.02.002

Mining Unknown Porcine Protein Isoforms by Tissue-Based Map of Proteome Enhances the Pig Genome Annotation

Pengju Zhao[1,#], Xianrui Zheng[1,#], Ying Yu[1], Zhuocheng Hou[1], Chenguang Diao[1], Haifei Wang[1], Huimin Kang[1], Chao Ning[1], Junhui Li[1], Wen Feng[1], Wen Wang[2], George E. Liu[3], Bugao Li[4], Jacqueline Smith[5], Yangzom Chamba[6], Jian-Feng Liu[1,*]

[1] *National Engineering Laboratory for Animal Breeding; Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture; College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*

[2] *Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi' an 710072, China*

[3] *Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, U.S. Department of Agriculture, Beltsville 20705, USA*

[4] *Department of Animal Sciences and Veterinary Medicine, Shanxi Agricultural University, Taigu 030801, China*

[5] *The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK*

[6] *Tibet Agriculture and Animal Husbandry College, Linzhi 860000, China*

[#] Equal contribution.

* Corresponding author.

E-mail: liujf@cau.edu.cn (Liu JF).

Running title: *Zhao P et al / Enhancing Pig Genome Annotation by Draft Proteome*

Total words: 9822

Figures: 6

Supplementary Figures: 14

Supplementary tables: 15

Supplemental files: 1

Abstract

A lack of the complete pig proteome has left a gap in our knowledge of the pig genome and has restricted the feasibility of using pigs as a biomedical model. We developed the tissue-based proteome map using 34 major normal pig tissues. A total of 5841 unknown protein isoforms were identified and systematically characterized, including 2225 novel protein isoforms, 669 protein isoforms from 460 genes symbolized beginning with LOC, and 2947 protein isoforms without clear NCBI annotation in current pig reference genome. These newly identified protein isoforms were functionally annotated through profiling the pig transcriptome with high-throughput RNA sequencing of the same pig tissues, further improving the genome annotation of the corresponding protein-coding genes. Combining the well-annotated genes that have parallel expression pattern and subcellular witness, we predicted the tissue-related subcellular components and potential function for these unknown proteins. Finally, we mined 3081 orthologous genes for 52.75% of unknown protein isoforms across multiple species, referring to 68 KEGG pathways as well as 23 disease signaling pathways. These findings provide valuable insights and a rich resource for enhancing studies of pig genomics and biology, as well as biomedical model application to human medicine.

KEYWORDS: Expression pattern; Unknown protein; Pig; Proteome; Subcellular components

Introduction

The domestic pig (*Sus scrofa*) is one of the most popular livestock species predominately raised for human consumption worldwide. Besides its socio-economic importance, pig has been generally recognized as a valuable model species for studying human biology and disease due to its striking resemblances with humans in anatomy, physiology, and genome sequence [1,2]. To date, many porcine biomedical models have been created for exploring etiology, pathogenesis, and treatment of a wide range of human diseases, *e.g.*, Parkinson disease [3], obesity [4], brain disorder [5], cardiovascular, atherosclerotic disease [6], and Huntington's disease [7]. Furthermore, pigs and humans share similarities in the size of their organs, making pig organs potential candidates for porcine-to-human xenotransplantation [8,9]. Recently major efforts have been devoted to the development of tools for further enhancing the value of pigs as a biomedical model for human medicine as well as its role in meat production. Of essential significance is the completion of the assembly of the pig genome sequence (*Sus scrofa*11.1) in recent time. It provides researchers with a vast amount of genomic information, facilitating characterization of individual pig genome as well as genome comparison between pigs and humans.

With the progress of large-scale genome projects, such as Encyclopedia of DNA Elements [10] and Human Proteome Projects [11], many genes have been annotated at RNA and protein levels, and diverse regulatory elements across the human genome were systematically characterized. This creates great opportunities for exploring how genetic variation underlies complex human phenotypes [12]. In particular, a spate of groundbreaking studies was succeeded in building high-resolution maps of the proteome [13−15] in a variety of human tissues and cells. Findings from these studies greatly facilitate the functional annotation of the genome at multiple-omic levels and further improve the understanding of complexity of human phenotypes.

Compared with humans, however, studies of pig proteome are very limited [16,17]. In particular, in-depth identification and characterization of the proteome maps of the pig genome across a broad variety of pig tissues are not yet available. To date, the leading protein database UniProtKB comprised around 1419 reviewed and 34,201 unreviewed pig proteins in Swiss-Prot and TrEMBL respectively. It is far less than the numbers of entries in Swiss-Prot (20,215 proteins) and TrEMBL (159,615 proteins) corresponding to human proteome data. Although the recent update of the pig

PeptideAtlas presented 7139 protein canonical identifications from 25 tissues and three body fluid [18], it is still a limited promotion to whole pig proteome research. In fact, a large number of unreviewed and PeptideAtlas-identified pig proteins were not well annotated in current genome (*Sus scrofa*11.1) due to lack of specific genomic locations and the corresponding assembled RNA transcripts. This suggests that there are still plenty of poorly annotated proteins that were not identified and characterized in previous pig studies. In addition, even if the annotated pig protein-coding genes (PCGs), nearly 20% of which were symbolized beginning with LOC—the orthologs and function of genes have not been determined—that also presented one of the key limitations of pig gene set enrichment analysis. The absence of completive maps for the pig proteome triggers a substantial bottleneck in the progress of refining pig genome annotation and even hinders systematic comparison of omics data between humans and pigs.

Therefore, considering the potential contribution to develop pig proteomic atlases, we conducted in-depth characterization of pig proteome across 34 histologically normal tissues using high-resolution mass spectrometry. Accordingly, we exploited the novel protein firstly identified herein, poorly annotated proteins, and LOC proteins, and defined these as the pig unknown proteins. These unknown proteins were mapped to the latest pig genome (*Sus scrofa*11.1) for confirming their available genomic locations. Jointly profiling the proteome and transcriptome across multiple pig tissues investigated, we found that the majority expression of transcripts was dominated by the expression of a small proportion of protein-coding genes and that most of newly identified protein isoforms herein with relatively higher tissue expressed specificity in contrast to the existing protein-coding genes. We subsequently constructed the tissue-based protein-coding gene spectrum of tissue-enriched, group-enriched, and ubiquitously expressed genes in the pig genome, and determined 452 unknown protein isoforms as the novel candidates of pig housing-keeping genes. Accordingly, we developed pig transcriptomic atlas and subcellular characterization for these unknown protein isoforms to infer their connections with the specific function of tissues. Finally, by systematically comparing the orthologous relationship of these unknown proteins with other multiple species, we further predicted the potential function of these unknown protein isoforms to ensure their availability in future relevant studies. Findings herein will benefit studies and development of pig genome and will allow further investigation of swine gene function and networks of particular interest to the

scientific community.

Results

Tissue-based map of the pig proteome

We explored the pig proteome from 34 tissue (**Figure 1**A) samples using liquid chromatography tandem mass spectrometry (LC-MS/MS). *In silico* analyses (Figure 1B) were then conducted to construct the whole landscape of the pig proteome with a view to furthering pig biological research and human medical studies. The resulting proteome data involved a total number of 21,681,643 MS/MS spectra produced from 680 LC-MS/MS runs (20 runs per tissue).

To exploit convincing peptide evidence for all putative PCGs in the pig genome, we searched the raw MS/MS data by Mascot [19] against multiple protein databases. These included the primary pig database of UniProt [20] for the initial search and two custom-developed databases for sequential searches of unmatched spectra, *i.e.*, (1) RNA sequencing (RNA-seq)-based *de novo* assembly transcriptomic database which included the RNA-seq data generated from the 34 tissues in this study, 1.08 Gb data from an external public expressed sequence tag (EST) database and 953.57Gb from publicly available RNA-seq data (Materials and Methods); and 2) a six-frame-translated pig genome database. Those corresponding matched spectra extracted from each subset of databases were re-searched against the same database by X!Tandem [21] for further filtration, producing the final 5,082,599 peptide spectrum matches (PSMs) at 1% PSM false discovery rate (FDR). Subsequently, Scaffold (version Scaffold_4.4.5, Proteome Software Inc., Portland, OR, USA) was run for MS/MS-based peptide and protein identification, both of them using the local FDR criterion of 0.01.

Totally, we identified 212,154 non-redundant peptides with a median number of 8 unique peptides per gene (Quality assessment of protein identification is shown in Figure S1-S14). Comparison of identified peptides with the largest pig peptide resource PeptideAtlas (http://www.peptideatlas.org/) showed that 49,144 out of 87,909 curated peptides (56%) were confirmed by our identification. The peptides we detected greatly outnumbered those deposited in PeptideAtlas, with a major fraction (77%) found to be novel. A total of 24,431 protein isoforms with median sequence coverage of 30.32% were determined by Scaffold, which corresponded to 19,914 PCGs. To ascertain whether our protein identifications achieved a reasonable false positive error rate, we

additionally validated 31 proteins from different proteogenomic categories. By comparing MS/MS spectra from 71 synthetic peptides with those obtained from our analysis of pig tissues, we obtained 100% of validation (Table S1 and File S1).

Identification and characterization of unknown pig proteins

Classifying all of 24,431 identified protein isoforms (**Figure 2**A) indicated that 16,738 (68.51%) protein isoforms were confirmed by the Uniprot protein evidence, 9204 (37.67%) protein isoforms had evidence from pig PeptideAtlas [18], 17,781 (72.78%) protein isoforms were included in NCBI protein database, and 7910 (32.38%) were supported by all of them. Of all confirmed protein isoforms, 17,781 (85.78%) protein isoforms according to 11,308 PCGs were included in known NCBI annotation, 669 protein isoforms according to 460 PCGs were annotated in the pig genome but classified as uncharacterized LOC genes, and 2947 protein isoforms remained a lack of NCBI annotation support in the pig genome (Figure 2B). The rest of 3703 protein isoforms were identified by MS/MS data for the first time in this study, of which 2225 had higher confidence (PSMs with Mascot ion score > 20 and identification probability > 20%, details in Table S2) can be considered as potential novel proteins.

To further enhance the annotation of PCGs for the current pig genome, we systematically characterized these 5841 feature or/and location unknown protein isoforms detected in current study (*i.e.*, 669 protein isoforms of LOC genes, 2947 protein isoforms without genomic location annotation and 2225 protein novel isoforms firstly identified herein). Considering only 11.5% of protein isoforms had available genomic locations, we mapped the rest of 5172 unknown protein isoforms to the pig reference genome (*Sus scrofa*11.1) by MAKER annotation workflow [22]. First, the low-complexity repeats of pig reference genome were soft-masked by RepeatMasker. Totally, 5172 out of 5841 unknown protein isoforms (non-LOC genes) were aligned to the masked reference genome by BLAST [23]. Sequentially, Exonerate [24] was run to realign and polish the exon–intron boundaries of the unknown genes with the splice-site aware alignment algorithm. A total of 4026 (77.8%) unknown protein isoforms were successfully aligned to the reference genome with > 95% sequence identity and similarity (2073 with the 100% identity and 100% similarity), including 3886 assigned to chromosomes and 140 resided on 23 unplaced scaffolds. More interestingly, we found that the proportion of novel proteins mapped in respective chromosomes was related to the improvements in genomic annotations from *Sus scrofa*10.2 to *Sus*

*scrofa*11.1 for different chromosomes (Figure 2C, R^2 = 0.67, P = 0.0015). This demonstrated that these unknown proteins, especially the novel proteins, were actually ignored in the current pig genome annotation since most of previous studies have been limited to the incomplete annotation of *Sus scrofa*10.2 genome and the small number of tissues investigated.

Comparison of these unknown protein isoforms with the well-annotated proteins revealed that, a major fraction of unknown protein isoforms (39.05%), especially the novel protein isoforms (45.82%), were merely identified in a single tissue that far more than well-annotated protein isoforms. It was most likely due to the tissue-specific and low abundance of these novel proteins. Additionally, further analysis of the reliability for these unknown proteins revealed that a major fraction of them (60.42%) were regarded as the abundant proteins that have more than ten spectral counts [25]. Particularly, although these novel protein isoforms were first identified in this study, almost 60.67% of all were supported by a high spectral count of > 5.

Expression landscape of unknown protein isoforms by profiling pig transcriptome

To further probe potential function of unknown protein isoforms, we characterized the expression landscape of unknown protein isoforms by high-throughput RNA-seq of the 34 tissue samples analyzed in the LC-MS/MS assays.

Approximately 1495 million paired-end reads (376.7G bases per tissue) were obtained through sequencing 116 strand-specific paired-end RNA libraries, of which 1230 million were mapped to the pig genome (*Sus scrofa* 11.1) with an overall pair alignment rate of 88.29% (Table S3). As expected, a total number of 2,486,239 transcripts (the FragmentsPer Kilobase per Millio (FPKM) > 0.1 in at least one tissue) corresponding to 29,270 genes were then assembled and quantified across all tissues, which contained 5250 annotated transcripts corresponding to 3486 known noncoding genes, 7595 potentially novel alternatively spliced transcripts corresponding to 2421 known noncoding genes, 55,328 annotated transcripts corresponding to 20,401 PCGs, 136,537 potentially novel alternatively spliced transcripts corresponding to 15,385 PCGs, and 2,281,529 newly assembled transcripts corresponding to 26,493 genes in the pig genome without annotation information. These findings clearly increased the average number of isoforms per gene (Human-NCBI: 7.27, Pig-NCBI: 2.75, Pig-Identified: 6.60) compared with the existing gene annotation in NCBI (**Figure 3**A).

On the basis of all the currently well-annotated genes (the genes annotated in NCBI),

we constructed a tissue similarity map across the 34 tissues using hierarchical clustering based on the Pearson correlation. As shown in Figure 3B, with the exception of three obvious outliers—adult testis, pancreas and peripheral blood mononuclear cells (PBMC)—the data were clustered into multiple known connected groups: liver and kidney, muscular system (longissimus dorsi and heart), nervous system (retina, brain, and spinal cord), adult immune organs (spleen, salivary gland, and lymph), and bladder tissue (urinary bladder, gall bladder, and oesophagus). These results revealed the expected biology that had a similar expression profile to that of human tissues [13], reflecting the biological similarity between human and pig, as well as the reliability of transcripts we constructed.

Intriguingly, a total of 51.67% (3018) of unknown protein isoforms were successfully confirmed by the transcripts constructed herein, which offered a detailed view of the understanding of unknown proteins. Considering the comparison of unknown protein isoforms with the potential low-expression levels in different tissue, we applied zFPKM normalization method [26] to generate high-confidence estimates of gene expression. The observed zFPKM range of unknown protein isoforms expression ranged from -3.02 to 19.89, showing lower average expression levels (zFPKM = 2.47), especially the novel protein isoforms (zFPKM = 2.21), than the well-annotated PCGs (zFPKM = 3.62). Besides, we also found that these unknown protein isoforms (average 12.1 tissues) tend to be expressed in less tissues than the well-annotated PCGs (average 21.3 tissues), and nearly 39.03% (n = 1178) of unknown protein isoforms were only identified in a single tissue (Figure 3C). The results suggested their specific expression characteristics may be one of the factors that leaded to the incomplete annotation of these unknown protein isoforms.

Screening the protein isoforms expression patterns in each tissue, we observed that the majority expression of transcripts was dominated by the expression of a small proportion of genes in all of the investigated tissues (Table S4). Specifically, the adult pig tissues of prostate, longissimus dorsi, pancreas, gall bladder, *etc.*, had the least complex transcriptome, with 50% expression of the transcripts coming from a few highly expressed genes (3 to 8 transcripts). In contrast, the reproductive tissues (uterus, testis and ovary), expressed more complex transcriptome, with a large number of genes expressed. Similar transcriptomic patterns have also been reported in human tissues [27]. It was surprising that 203 unknown protein isoforms were potentially associated with 148 (13.98%) highly expressed genes, suggesting these unknown protein isoforms

may play an important role in basic function among tissues or organs.

Prediction of unknown proteins function from pig transcriptome

Several approaches for systematic analysis of gene expression across different tissues have indicated that gene expression patterns were usually associated with their biological functions, and genes with the similar functions are more likely to exhibit similar expression patterns [28]. Implementing the similar classification criteria for human genes [13] into the RNA-seq data generated from the multiple pig tissues herein, we classified all 23,887 putative NCBI genes (18,377 PCGs) corresponding to well-annotated 60,578 transcripts and 3018 unknown protein isoforms into three categories for exhibiting their expression features. The numbers of tissue-enriched genes, group-enriched genes, and ubiquitously expressed genes are also displayed as a network plot to show an overview of pig PCGs (**Figure 4**A).

In multicellular organisms, genes expressed in a few tissue types are thought to be tissue-enriched genes which have tissue-specific related functions. We observed 8482 (14%) well-annotated transcripts (5592 genes) and 1178 (39.03%) unknown protein isoforms that have a specific expression in a particular tissue. Furthermore, 16,356 (27%) well-annotated transcripts (9726 genes) and 203 (6.73%) unknown protein isoforms were expressed at least 5-fold higher at the zFPKM level in one tissue compared to the tissue with a second highest expression. Similar to previous studies in humans [13] (Figure 4B), the largest number of tissue-enriched genes were detected in the adult testis, followed by infancy brain, retina, and adult brain. The results reflected that the tissues with complex biological processes usually had more tissue-enriched genes, and these tissue-enriched genes were strongly associated with the function of the corresponding tissues. This can be exemplified by the *RHO* (Rhodopsin) gene that was enriched in retina and was proven to play important roles in retinal pigments [29]. This demonstrates that the tissue specificity can not only confirm the biological characteristics of known genes but also predict basic function of undefined genes in pigs. Accordingly, we successfully updated 1386 tissue-enriched unknown protein isoforms to further explain the functional differences among tissues.

Apart from the genes observed in tissue specificity, some group-enriched genes were over-represented in the group of tissues/organs that together perform closely related functions. Accordingly, we found that a total of 1318 (2.18%) well-annotated

transcripts (948 genes) and 48 (1.59%) unknown protein isoforms were detected and could be grouped into seven types of tissue (Figure 4C). The largest fraction (72.7%) of group-enriched genes belonged to the brain tissue (14.7%), followed by the muscular system (cardiac muscle, longissimus dorsi), adrenal and thymus gland (6.6%), as well as liver and gallbladder (4.5%). Generally, these group-enriched genes have potential role in biological system function, and the expression patterns were usually shown between different species. As exemplified by the group-enriched expression of *MYL3* (myosin light chain 3, a known myosin component) (Figure 4D) and *ENC1* (Ectodermal-Neural Cortex 1, involved in mediating uptake of synaptic material) (Figure 4E), these two genes indicated a similar expression in the muscular system and in brain tissue between humans and pigs. Therefore, 48 of unknown protein isoforms will be the valuable resources for further enriching the functional and comparative genomics between pig and human.

Specifically, we identified 5656 well-annotated transcripts corresponding to 5147 (21.55%) NCBI genes expressed in all pig tissues. Among these genes, a variety of known "housekeeping" genes such as *ACTB*, *GAPDH*, *PGK1*, *RPL19*, (Figure 4F) are usually intracellular and tend to be functionally essential to cell subsistence that involved in metabolism, transcription, and RNA processing or translation [30]. Interestingly, 452 (14.98%) of unknown protein isoforms were detected as the ubiquitously expressed genes, suggesting that the findings of these unknown protein isoforms offered the important supplement to pig genomic annotation. To characterize the set of ubiquitously expression of these unknown genes identified herein, we constructed a co-expression network heatmap that consisted of 24 blocks for assessing ubiquitously expressed gene co-expression interactions across all pig tissues (Figure 4G). Obviously, these unknown protein isoforms have potentially functional connections with the well-annotated genes in the same blocks (Table S5), which can be explained by those genes within modules of a co-expression network involved in similar or related pathways and biological processes [31].

Subcellular characterization of the unknown pig proteome

Proteins with different subcellular locations usually play different roles in physiological and pathological processes. To characterize these unknown pig proteins at the subcellular level, we performed a proteome-wide subcellular classification for all identified pig protein isoforms (n = 24,431) based on the existing prediction methods

[13] (as described in Materials and Methods). A major fraction (72.66%) of pig protein isoforms were predicted to be soluble protein isoforms, followed by 21.55% of membrane protein isoforms and 5.79% of secreted protein isoforms (**Figure 5**A and Table S6). For an in-depth comparative analysis on PCGs, we further clustered all available protein isoforms into four base categories including 14,890 soluble proteins, 3924 membrane proteins, 1053 secreted proteins, and 47 membrane & secreted proteins (Table S7). As shown in Figure 5B, there were only 2.4% of PCGs (*n*=416) with isoforms belonging to two or more categories, which is far less than the 19.3% of PCGs (*n*=3917) with the similar type of isoforms in human [13]. It is worth noting that the novel protein isoforms (84.27%) has a greater proportion of soluble proteins than the known protein isoforms (71.67%). The results hinted that the solubility of soluble proteins in liquids may be one of the reasons that due to some proteins were missed in current pig proteome.

More interestingly, we found that the organ or tissue functions were also related to subcellular of their expressed proteins. Ranking all of identified proteins by their zFPKM value for each tissue, we selected the top 1% to represent their main proteins. As shown in Figure 5C, the higher proportion of membrane proteins were associated with nervous tissues, such as spinal cord, brain, retina. Moreover, muscle tissues have a higher proportion in soluble protein, and the higher proportion of secreted proteins were represented by higher expression especially in some secretory tissues, such as liver, uterus, pancreas, gall bladder, gut.

In addition, similar to human proteome [13], these highly expressed protein especially in secretory tissues usually tended to the secreted components and were representative to the tissue function (Figure 5D). For example, both of LOC100620249 and *PGC* (Progastricsin) genes were highly expressed secreted proteins in the stomach tissue, and the latter is a known secreted protein and constitutes a major component of the gastric mucosa. This demonstrates the LOC100620249 gene more likely has the stomach-related function, which provides a valuable information for enhancing studies of pig genomics and biology.

Inferring orthologous functions of unknown pig proteome across multiple species

To pursue stronger evidence and orthologous functions for these unknown proteins, we further aligned the sequence of each isoform against the top 10 species databases. We adopted two criteria to identify homologous sequences to the newly identified swine

proteins with those of other species: (1) percent identity greater than 80%; and 2) length of homologous sequence longer than 80% of the swine protein sequence. Consequentially, 3081 out of 5841 (52.75%) unknown protein isoforms were inferred to have orthologous in other species. While 90.17% orthologous isoforms (n = 2778) were identified in at least other two species, 36.51% of orthologous isoforms (n = 1125) were the common isoforms for 9 (Chicken) or all 10 species (**Figure 6**A). Interestingly, even the novel protein isoforms still have 43.60% of the orthologous protein isoforms (n = 970) with other species, and almost 73.09% (n = 709) of them were mapped in the pig genome (*Sus scrofa* 11.1) (Figure 6B). The results indicated that the exploited novel proteins herein can be considered as the reliable proteome data that significantly enhance both the pig genome annotation and the current pig protein database.

In addition, compared with the existing orthologues in omabrowser (http://omabrowser.org) and current genome sequences, 3081 of the unknown protein isoforms enriched 12,375 novel pairwise orthologous relationships between pigs and other species (Table S8). These pairwise orthologous relationships of proteins between pigs and other species provided a feasible way to investigate the potential function of corresponding PCGs in the pig genome if these homologous proteins have been well studied in other species. Therefore, considering the most complete set of annotated genes in human proteome, we preformed the functional gene set enrichment for human orthologous proteins of these unknown protein isoforms to speculate their potential function. A functional Gene Ontology (GO) analysis for these unknown protein isoforms showed that most of GO terms describe cell and intracellular part (corrected $P < 0.01$), which provide an important supplement to understand the biological process in pig (Table S9). Meanwhile, by further examining the functional characterization of these unknown protein isoforms, we found 68 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were represented in our unknown proteome (Table S10), mainly involving metabolic pathways (corrected P = 5.2E-20), focal adhesion (corrected P = 2.4E-9), regulation of actin cytoskeleton (corrected P = 5.1E-7). Importantly, we detected 23 disease signaling pathways from the KEGG disease database (corrected $P < 0.05$) that included the metabolism, nervous system, skeletal, muscular, and skin diseases (Table S11). These findings will help us better recognize the potential function of the unknown pig protein isoforms, and provide a valuable resource to support the pig as a biomedical model for human medicine and as donors for porcine-to-human xenotransplantation.

Discussion

Here we presented the landscape of a tissue-based proteome for pigs. Our findings not only offered the verification for 84.84% of the existing pig proteins that have been deposited in the UniprotKB (*n* = 16,738), pig PeptideAtlas (*n* = 9204), and NCBI Protein database (*n* = 17,781), but also identified 2225 novel protein isoforms. Besides, we also detected 669 protein isoforms from uncharacterized LOC genes and 2947 protein isoforms without NCBI annotation in the current reference genome of *Sus scrofa*11.1. Eventually, a total of 5841 unknown protein isoforms were exploited to further optimize the annotation of PCGs for the current pig genome.

We systematically characterized unknown protein proteome for their expression features, subcellular components, and orthologous functions, providing a valuable resource for enhancing studies of pig genomics, as well as offering the opportunities for exploring the potential function of these unknown proteins. Our findings clearly showed that the missing protein annotation in previous studies was due to the three aspects: (1) low-quality assembly in *Sus scrofa*10.2 genome; (2) the specific features that low expression levels, tissue specificity, and greater proportion of soluble components in novel protein isoforms; and 3) the traditional gene prediction and annotation methods are prone to the inevitable errors [32]. The in-depth identification and subcellular characterization of proteome using multiple tissues make it feasible to develop a tissue-based pig proteome map and facilitate studies of functional genomics and relevant research fields. We effectively improved genome annotation for 4026 unknown protein isoforms by mapping their protein sequence to the current pig genome (*Sus scrofa*11.1), of which 3886 were assigned to chromosomes and 140 were resided on 23 unplaced scaffolds.

High-resolution profiling of pig transcriptome allows us to further reveal 1434 unknown protein isoforms that display a tissue- (1386) or group-enriched (48) function expression pattern. In addition, 452 of unknown protein isoforms were ubiquitously expressed in 34 tissues, which raised 7.4% of the potential “housekeeping” gene in the pig genome. These findings provide new insight into understanding the molecular function of the respective tissue or organ. Further inferring the biological function of unknown pig proteome by human orthologous proteome, we found that these unknown protein isoforms were enriched in 68 KEGG pathways and 23 disease signaling

pathways, including the pathways involved in disease of concern for human medicine, such as metabolism, nervous system, skeletal, muscular, and skin diseases. The integrated data of proteome and transcriptome in the 34 pig tissues herein were respectively presented in Table S12 and Table S13, and 5841 unknown protein isoforms with corresponding genomic locations, expression landscapes, subcellular characterizations, orthologous proteins, and predicted functions were also summarized in Table S14. All findings herein will provide valuable insights and resources for enhancing studies of pig genomics and biomedical model application to human medicine in the future.

Materials and methods

Sample acquisitions

The tissue samples and PBMC used for protein identification and mRNA expression analyses were collected from pigs raised in the Ninghe breeding pig farm in Tianjin, China. For purpose of generating profiles of transcriptome and proteome of all major organs and tissues of pigs, we totally collected 34 samples (*i.e.*, 33 pooled tissues and the PBMC) from the nine unrelated Duroc pigs, including three adult male pigs and three female pigs at 200-240 days of age, as well as three male infant pigs at 21-25 days of age. All pig tissues were histologically confirmed to be normal and healthy by an experienced pathologist. An overview of all involved tissues and cell samples is provided in Table S3.

Preparation of pig samples

All samples were snap frozen within 20 min after slaughter and stored in liquid nitrogen until usage. PBMC were isolated using Ficoll-Hypaque PLUS (GE Healthcare, Beijing, China) following the manufacturer's instructions. In brief, the whole blood was first diluted by an equal volume of phosphate buffer solution (PBS). Then, 20 ml of diluted blood was carefully added on top of 10 ml of Ficoll-Hypaque solution in a 50 ml conical tube and centrifuged at 460 g for 20 min at room temperature. After centrifugation, the middle whitish interface containing mononuclear cells was transferred to a new tube, washed by PBS, and centrifuged at 1000 rpm for 10 min twice.

Separation of protein and RNA

Fresh frozen tissue was thawed, cut into small pieces, and extensively washed with precooled PBS. A pool of equal amounts of tissues from three unrelated pigs was homogenized and sonicated in cold lysis buffer. Extraction of 100 μg protein using protein extraction buffer (8 M urea, 0.1% SDS) containing an additional 1 mM phenylmethylsulfonyl fluoride (Beyotime Biotechnology, Shanghai, China) and protease inhibitor cocktail (Roche, California) was kept on ice for 30 min and then centrifuged at 16,000 × g for 15 min at 4 °C. The supernatant was collected and determined with BCA assay (Pierce, Washington) and 10%-20% SDS-PAGE. The cell lysate was stored at −80 °C before LC-MS analysis.

Total RNA was extracted from the pooled tissues via the Trizol method (Invitrogen, Carlsbad, CA, USA) according to standard protocols. RNA degradation and contamination were monitored on 1% agarose gels. The purity and contamination of total RNA was checked using NanoPhotometer (IMPLEN, Los Angeles, CA, USA) and Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Csrlsbad, CA, USA). RNA integrity was measured using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, USA). All pig samples with an RNA Integrity Number (RIN) value greater than 7.0 and at least 5 μg of total RNA were included and batched for RNA sequencing.

Library construction and RNA sequencing

Total RNA of samples meeting quality control (QC) criteria were rRNA depleted, and depleted QC was done using the RiboMinus Eukaryote System v2 and RNA 6000 Pico chip according to the manufacturer's protocol. RNA sequencing libraries were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, England) with 3 μg rRNA depleted RNA according to the manufacturer's recommendation. RNA-seq library preparations were clustered on a cBot Cluster Generation System using HiSeq PE Cluster Kit v4 cBot (Illumina, California) and sequenced using the Illumina Hiseq 2500 platform according to the manufacturer's instructions, to a minimum of 10 G reads per sample (corresponding to 125 bp paired-end reads). The sequenced RNA-Seq raw data for the 34 pig tissues have been deposited in NCBI Sequences Read Archive with the BioProject number PRJNA392949.

Fractionation of peptide mixture using a C18 column

Peptide mixture from each sample was first lyophilized and reconstituted in buffer A (2% ACN, 98% H_2O, pH10). Then, it was loaded onto a Xbridge PST C18 Column, (130 Å, 5 µm, 250 × 4.6 mm, Waters, Massachusetts) on the DIONEX Ultimate 3000 HPLC (Dionex, California, USA) equipped with a UV detector. Mobile phase consists of buffer A and buffer B (90% ACN, 10% H_2O, pH10). The column was equilibrated with 100% buffer A for 25 min before sample injection. The mobile phase gradient was set as follows at a flow rate of 1.0 ml/min: (a) 0–19.9 min: 0% buffer B; (b) 19.9–20 min: 0–4% buffer B; (c) 20–22 min: 4–8% buffer B; (d) 22–42 min: 8–20% buffer B; (e) 42–59 min: 20–35% buffer B; (f) 59–60 min: 35-45% buffer B; (g) 60–61min: 45–95% buffer B; (h) 61–66 min: 95% buffer B; (i) 66–67 min: 95-0% buffer B; and j) 67–91 min: 0% buffer B. A fraction was collected every minute from 24 min to 63 min, and a total of 40 fractions collected were then concentrated to 20 fractions, vacuum dried, and stored at –80°C until further LC-MS/MS analysis.

Liquid chromatography tandem mass spectrometry

Peptide mixture was analyzed on a Q Exactive instrument (Thermo Scientific, Massachusetts) coupled to a reversed phase chromatography on a DIONEX nano-UPLC system using an Acclaim C18 PepMap100 nano-Trap column (75 µm × 2 cm, 2 µm particle size, Thermo Scientific) connected to an Acclaim PepMap RSLC C18 analytical column (75 µm × 25 cm, 2 µm particle size, Thermo Scientific). Before loading, the sample was dissolved in sample buffer, containing 4% acetonitrile and 0.1% formic acid. Samples were washed with 97% mobile phase A (99.9% H_2O, 0.1% formic acid) for concentration and desalting. Subsequently, peptides were eluted over 85 min using a linear gradient of 3%–80% mobile phase B (99.9% acetonitrile, 0.1% formic acid) at a flow rate of 300 nl/min using the following gradient: 3% B for 5 min; 3–5% B for 1 min; 5–18% B for 42 min; 18–25% B for 11 min; 25–30% B for 3 min; 30–80% B for 1 min; hold at 80% B for 5 min; 80–3% B for 0.5 min; and then hold at 3% B for 21.5 min. High mass resolution and higher-energy collisional dissociation (HCD) was employed for peptide identification. The nano-LC was coupled online with the Q Exactive mass spectrometer using a stainless steel emitter coupled to a nanospray ion source. The eluent was sprayed via stainless steel emitters at a spray voltage of 2.3 kV and a heated capillary temperature of 320°C. The Q Exactive instrument was operated in data-dependent mode, automatically switching between MS and MS2. Mass

spectrometry analysis was performed in a data dependent manner with full scans (350-1,600 m/z) acquired using an Orbitrap mass analyzer at a mass resolution of 70,000 at 400 m/z on Q Exactive using an automatic gain control (AGC) target value of 1×10^6 charges. All the tandem mass spectra were produced by HCD. Twenty most intense precursor ions from a survey scan were selected for MS/MS from each duty cycle and detected at a mass resolution of 17,500 at m/z of 400 in Orbitrap analyzer using an AGC target value of 2×10^5 charges. The maximum injection time for MS2 was 100 ms and dynamic exclusion was set to 20s.

Validation of identified Proteins

In total, 71 peptides from 31 proteins (7 known proteins, 11 homologous novel proteins, and 13 non-homologous novel proteins) were randomly selected for peptide synthesis (GL biochem, Shanghai, China) for validation of identified proteins. The synthesized peptide sequences were mixed and processed twice by chromatographic separation using the Thermo EASY-nLC HPLC system and Thermo scientific EASY column. Mass spectral analysis was then performed by Q-Exactive (Thermo Scientific) and processed by Mascot V2.5.1. Finally, all these peptides were compared with those identified from our proteome analysis to verify novel proteins.

QC processing

We conducted a quality control step on raw fastq reads for efficient and accurate RNA-seq alignment and analysis. In this step, raw reads were cleaned up for downstream analyses using the following steps: removal of adapter sequences using BBDuk (http://sourceforge.net/projects/bbtools/) [33]; calculation of the Q20, Q30 and GC content of the clean data for quality control and filtering using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/); and homopolymer trimming to the 3′ end of fragments and removed the N bases from the 3′ end using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

Read mapping and assembly

RNA-seq data were mapped and genome indexed with Hisat 0.1.6-beta 64-bit [34] to the pig genome release version of *Sus scrofa*11.1 (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/003/025/GCF_000003025.6_Sscrofa11.1/). *Sus scrofa*11.1 annotation was used as the transcript model reference for the

alignment, splice junction identification, and for all protein-coding gene and isoform expression-level quantifications. To obtain expression levels for all pig genes and transcripts across all 34 samples, FPKM values were calculated using Stringtie 1.0.4 (Linux_x86_64) [35] with the default parameters. A gene or transcript was defined as expressed if it's FPKM value was measured greater than 0.1 across all tissues. For each tissue, we applied zFPKM normalization method [26] to generate high-confidence estimates of gene expression.

zFPKM level-based classification of genes

Refer to the gene classification in human, we also classified the pig genes into one of three categories based on the zFPKM levels in 34 samples: (1) "Tissue enriched" was only detected in a single tissue as well as at least 5-fold higher at the zFPKM level in one tissue compared to the tissue with a second highest expression; (2) "Group enriched" was detected in all tissues from a groups, and the expression of genes in any tissue from the groups is higher than the tissue that not from the group; and 3) "Expressed in all tissues" was detected in all 34 tissues.

Construction of a reference protein database

To identify novel protein and improve existing proteins annotations in the pig genome, the database for protein searching (MS/MS data searched against protein database) was taken from four different levels using in-house perl scripts, including: (1) UniProt database (*Sus scrofa*); (2) three-frame-translated mRNA de novo sequences from the current study; and 3) six-frame-translated pig genome database. The details are as follows:

Primary database of proteins: resource protein data sets for pig (UniProt version 20150717 containing 34,131 entries, with 1486 Swiss-Prot, 32,643 TrEMBL) were downloaded from the UniProt database (http://www.uniprot.org/).

Secondary database of proteins: it is well known that pig proteins were insufficiently represented by the known detectable proteins, because of the incomplete nature of the pig genome assembly and limited annotation. In our study, three RNA resources were used (Table S15): (1) EST datasets including 34,131 entries from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/susScr3/bigZips/) and 1,676,406 entries from the NCBI database (http://www.ncbi.nlm.nih.gov/nucest). ESTs are normally assembled into longer consensus sequences for three-frame-translated mRNA protein

database using iAssembler version 1.3.2.x64 [36] with default parameter; (2) Paired-End (PE) read libraries including 34 RNA sequencing libraries from our study and 7 previously published article and NCBI database. To construct a complete protein database for three-frame-translated mRNA, we used Trinity (version 2.0.6) [37] for *de novo* transcriptome assembly from RNA-Seq data, and identified potential coding regions within Trinity-reconstructed transcripts by TransDecoder (developed and include with Trinity); and 3) Single-end (SE) reads from 10 previous studies were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/sra/). The method for sequence assembly and coding region prediction were similar to that used for the PE reads. Finally, the translated protein length cutoff was set as 100 amino acids to ensure all database quality and reduce the number of false positive hits.

Tertiary database of proteins: to capture the proteins missed during the laboratory discovery process as far as possible, protein annotation of the pig genome was carried out using the ab initio methods with GeneScan (version 1.0) software [38]. Finally, we compared different protein databases to detect the repetitive protein isoforms (redundancies) among three protein databases using BLASTP software. And then the repetitive protein isoforms were removed and reserved for only one according to the following database priorities: UniProt > *De novo* > *Ab initio*.

Peptide identification based on database searching

All MS/MS data were analyzed using Mascot (version 2.5.1, Matrix Science, London, UK) [39] and X!Tandem (version 2010.12.01.1, The GPM, Rockville, MD, USA) [40]. Mascot was set up to search the pig databases (UniProt, *de novo*, Assembly, *ab initio* database) and the cRAP database (common Repository of Adventitious Proteins, downloaded on 07 Jul 2015, 116 sequences) assuming the digestion enzyme trypsin.

The high resolution peaklist files were converted into Mascot generic format prior to database searching by ProteoWizard (version 3.0.6). X!Tandem was set up to search a subset of the pig databases, also assuming trypsin. The target-decoy option of Mascot and X!Tandem were enabled (decoy database with reversed protein sequences). Mascot and X!Tandem were used to search with a fragment ion mass tolerance of 0.050 Da and a parent ion tolerance of 10.0 PPM. The number of maximums allowed missed cleavage sites was set to 2. All PSMs identified at 1% FDR were set for all samples. Carbamidomethyl of cysteine was specified in Mascot and X!Tandem as a fixed modification. Gln- > pyro-Glu of the n-terminus, oxidation of methionine and acetyl of

the n-terminus were specified in Mascot as variable modifications. Glu- > pyro-Glu of the n-terminus, ammonia-loss of the n-terminus, Gln- > pyro-Glu of the n-terminus, as well as oxidation of methionine and acetylation of the n-terminus were specified in X! Tandem as variable modifications.

Scaffold was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they achieved an FDR < 1% by the Scaffold local FDR algorithm. Protein identifications were accepted if they had an FDR < 1% and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters. In the database searching workflow, unmatched MS/MS spectra generated from the Uniprot database searching were then searched against next level protein database (*De novo*, *Ab initio*).

Mapping the protein isoforms to the pig genome

We attempted to map all unknown protein isoforms against the pig genome using MAKER annotation workflow [22]. First, the low-complexity repeats of pig reference genome were soft-masked by RepeatMasker. Then, the unknown protein isoforms were aligned to the masked reference genome by BLAST [23] for identifying their genomic location roughly. Last, Exonerate [24] was used to realign and polish the exon–intron boundaries of the unknown gene with the splice-site aware alignment algorithm. The house-python script being used to deal with the result: if a successfully aligned protein had 95% identity overall, 95% coverage and the distance from its neighboring exon being less than 50 Kb, it was recorded to be an effectively aligned sequence.

Subcellular prediction and classification of pig proteome

The prediction of pig membrane proteins was carried out similarly to how these proteins were classified in the human proteome. A total of seven methods were used to identify membrane protein topology with different assessment algorithms, for example, topological models, neural networks, support vector machines (SVMs), scale of free energy contributions, and hidden Markov models (HMMs): MEMSAT3 [41], MEMSAT-SVM [42], SPOCTOPUS [43], THUMBUP [44], SCAMPI multi-sequence-version [45], TMHMM [46], and Phobius version 1.01 [47]. In our study, the proteins

were assigned as transmembrane if they were predicted by at least four out of the seven methods.

In accordance with human secretome analysis, the prediction of signal peptides was based on Neural Networks and Hidden Markov models with three software programs: SignalP4.0 [48], SPOCTOPUS and Phobius version 1.01. The proteins, predicted by at least two out of the three methods, to contain a signal peptide were classified as potentially secreted.

Integrating the prediction of pig membrane proteins and of pig secretome proteins, we classified each pig protein into one of three classes: secreted, membrane, or soluble (neither membrane nor secreted protein). In order to compare the proteome between pig and human conveniently, we also constructed four major categories for classifications of the protein-coding genes with multiple protein isoforms: (1) "soluble" just containing soluble category; (2) "secreted" were combined with the soluble/secreted and the secreted categories; (3) "membrane" including soluble/membrane and the membrane groups; and 4) "membrane and secreted isoforms" containing secreted/membrane and soluble/secreted/membrane groups.

Weighted gene co-expression network analysis

In order to reveal the groups of protein coding genes that are functionally related in the whole pig organism, 34 pig tissue data sets were constructed using the WGCNA method. In our study, we mainly used the blockwiseModules function in the WGCNA R package [49] to perform the coexpression network construction, with the following parameters: corType = pearson; maxBlockSize = 30,000; power = 8; minModuleSize = 30; mergeCutHeight = 0.1. The brief function of blockwiseModules automatically constructed a correlation network, created a cluster tree, defined modules as branches, merged close modules, and yielded the module colors and module eigen genes for subsequent analysis (such as visualization by the plotDendroAndColors function).

Functional annotations for pig PCGs

Gene ontology (GO) analysis and KEGG (http://www.genome.jp/kegg/) pathway enrichment analysis were performed and corrected by FDR method with KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/anno_iden.php). GO terms appearing in this study are summarized within three categories: cell component, molecular function, and biological process. In view of the most complete genes annotation in human genome,

we gave priority to those human annotated genes which were homologous to pig genes and utilized them as the background.

Ethical statement

All experimental procedures were performed by a license holder in accordance with the protocol approved by the Institutional Animal Care and Use Committee (IACUC) of China Agricultural University (permit number: DK1023).

Data availability

The RNA-seq raw data for the 34 pig tissues are available in the Sequence Read Archive at NCBI (BioProject: PRJNA392949) and GSA (BioProject: PRJCA004356), and The pig proteomic data have been deposited in PRIDE (PRIDE: PXD006991), which can be accessed at https://www.ebi.ac.uk/pride/archive/projects/PXD006991.

Authors' contributions

J-FL conceived and designed the experiments. PZ performed transcriptome and proteome analyses. CD, HK, and LZ performed pathway analysis and graphic design. XZ, JL, CN, HW, and YC collected samples and prepared for sequencing. XZ and WF assisted the experimental validations. J-FL, PZ, JS, ZH, GL, WW, YY, BL, and JS wrote and revised the paper. All authors read and approved the final manuscript.

CRediT author statement

Pengju Zhao: Software, Data curation, Writing - original draft preparation. **Xianrui Zheng:** Validation, Visualization and Writing - original draft preparation. **Ying Yu:** Writing - review & editing and Supervision. **Zhuocheng Hou:** Formal analysis Investigation. **Chenguang Diao:** Methodology and Data Curation. **Haifei Wang:** Resources and Validation. **Huimin Kang:** Software. **Chao Ning:** Data Curation. **Junhui Li:** Resources. **Wen Feng:** Resources. **Wen Wang:** Writing - review & editing. **George E. Liu:** Writing - review & editing. **Bugao Li:** Writing - review & editing.

Jacqueline Smith: Writing - review & editing. **Yangzom Chamba:** Supervision. **Jian-Feng Liu:** Conceptualization, Writing - original graft, Writing - review & editing, Supervision, Project administration and Funding acquisition.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was financially supported by the National Natural Science Foundations of China (Grant No. 31661143013) and Jinxinnong Animal Science Development Foundation. We thank Ziyao Fan, Yichun Dong, and Kaijie Yang for samples collection. We gratefully acknowledge the assistance from CapitalBio Technology and Beijing Compass Biotechnology Co., Ltd. for high-performance computing platform support.

ORCID

0000-0001-6990-1147 (Pengju Zhao)
0000-0002-1784-7393 (Xianrui Zheng)
0000-0002-4524-0791 (Ying Yu)
0000-0003-4752-2255 (Zhuocheng Hou)
0000-0001-5649-1687 (Chenguang Diao)
0000-0003-2424-9743 (Haifei Wang)
0000-0002-9396-1948 (Huimin Kang)
0000-0001-8247-1700 (Chao Ning)
0000-0002-2810-1765 (Junhui Li)
0000-0001-8485-5345 (Wen Feng)
0000-0002-7801-2066 (Wen Wang)
0000-0003-0192-6705 (George E. Liu)
0000-0002-3844-3726 (Bugao Li)
0000-0002-2813-7872 (Jacqueline Smith)
0000-0003-0982-3364 (Yangzom Chamba)
0000-0002-5766-7864 (Jian-Feng Liu)

References

[1] Cooper DK, Ezzelarab MB, Hara H, Iwase H, Lee W, Wijkstrom M, et al. The pathobiology of pig-to-primate xenotransplantation: a historical review. Xenotransplantation 2016;23:83−105.

[2] Ekser B, Markmann JF, Tector AJ. Current status of pig liver xenotransplantation. Int J Surg 2015;23:240−6.

[3] Bjarkam CR, Nielsen MS, Glud AN, Rosendal F, Mogensen P, Bender D, et al. Neuromodulation in a minipig MPTP model of Parkinson disease. Br J Neurosurg 2008;22:S9−12.

[4] Pedersen R, Ingerslev HC, Sturek M, Alloosh M, Cirera S, Christoffersen BO, et al. Characterisation of gut microbiota in Ossabaw and Gottingen minipigs as models of obesity and metabolic syndrome. PLoS One 2013;8:e56612.

[5] Lind NM, Moustgaard A, Jelsing J, Vajta G, Cumming P, Hansen AK. The use of pigs in neuroscience: modeling brain disorders. Neurosci Biobehav Rev 2007;31:728−51.

[6] Agarwala A, Billheimer J, Rader DJ. Mighty minipig in fight against cardiovascular disease. Sci Transl Med 2013;5:166fs1.

[7] Yan S, Tu Z, Liu Z, Fan N, Yang H, Yang S, et al. A huntingtin knockin pig model recapitulates features of selective neurodegeneration in huntington's disease. Cell 2018;173:989−1002.e13.

[8] Li Y, Fuchimoto D, Sudo M, Haruta H, Lin QF, Takayama T, et al. Development of human-like advanced coronary plaques in low-density lipoprotein receptor knockout pigs and justification for statin treatment before formation of atherosclerotic plaques. J Am Heart Assoc 2016;5:e002779.

[9] Cooper DK. A brief history of cross-species organ transplantation. Proc (Bayl Univ Med Cent) 2012;25:49.

[10] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57−74.

[11] Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, et al. The human proteome project: current state and future direction. Mol Cell Proteomics 2011;10:M111.009993.

[12] Maher B. ENCODE: the human encyclopaedia. Nature 2012;489:46−8.

[13] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347:1260419.

[14] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature 2014;509:575−81.

[15] Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509:582−7.

[16] Fischer D, Laiho A, Gyenesei A, Sironen A. Identification of reproduction-related gene polymorphisms using whole transcriptome sequencing in the large white pig population. G3 (Bethesda) 2015;5:1351−60.

[17] Chen F, Wang T, Feng C, Lin G, Zhu Y, Wu G, et al. Proteome differences in placenta and endometrium between normal and intrauterine growth restricted pig fetuses. PLoS One 2015;10:e0142396.

[18] 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;16:634−44.

[19] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20:3551−67.

[20] UniProt C. UniProt: a hub for protein information. Nucleic Acids Res 2015;43:D204−12.

[21] Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 2003;17:2310−6.

[22] Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 2008;18:188−96.

[23] Mount DW. Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc

2007;2007:pdb top17.

[24] Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 2005;6:31.

[25] Zhou W, Liotta LA, Petricoin EF. The spectra count label-free quantitation in cancer proteomics. Cancer Genomics Proteomics 2012;9:135−42.

[26] Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 2013;14:778.

[27] Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science 2015;348:660−5.

[28] Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol 2010;11:R124.

[29] Yu X, Shi W, Cheng L, Wang Y, Chen D, Hu X, et al. Identification of a rhodopsin gene mutation in a large family with autosomal dominant retinitis pigmentosa. Sci Rep 2016;6:19759.

[30] Ramskold D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 2009;5:e1000598.

[31] Liang YH, Cai B, Chen F, Wang G, Wang M, Zhong Y, et al. Construction and validation of a gene co-expression network in grapevine (*Vitis vinifera. L.*). Hortic Res 2014;1:14040.

[32] Zhang J, Yang MK, Zeng H, Ge F. GAPP: A Proteogenomic Software for Genome Annotation and Global Profiling of Post-translational Modifications in Prokaryotes. Mol Cell Proteomics 2016;15:3529−39.

[33] Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Laboratory. LBNL Report#: LBNL-7065E, 2014. https://escholarship.org/uc/item/1h3515gn

[34] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357−60.

[35] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL.

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33:290−5.

[36] Zheng Y, Zhao L, Gao J, Fei Z. iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences. BMC Bioinformatics 2011;12:453.

[37] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 2011;29:644−52.

[38] Ramakrishna R, Srinivasan R. Gene identification in bacterial and organellar genomes using GeneScan. Comput Chem 1999;23:165−74.

[39] Sadeh NM, Hildum DW, Kjenstad D, Tseng A. Mascot: an agent-based architecture for coordinated mixed-initiative supply chain planning and scheduling. In Workshop on Agent-Based Decision Support in Managing the Internet-Enabled Supply-Chain, at Agents' 99, 1999.

[40] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466−7.

[41] Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 2007;23:538−44.

[42] Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 2009;10:159.

[43] Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 2008;24:2928−9.

[44] Zhou H, Zhou Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 2003;12:1547−55.

[45] Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci U S A 2008;105:7177−81.

[46] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol

1998;6:175−82.

[47] Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027−36.

[48] Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 2011;8:785−6.

[49] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.

Figure Legends

Figure 1 Overview of pig transcriptome-based annotation

A. 34 pig tissues analyzed in this study. 34 representative normal pig tissues were selected as the resource of proteome and transcriptome for exploring convincing evidence of putative PCGs. **B.** The custom pipeline for proteome-based annotation. Four protein databases were used for protein search based on Mascot and X!Tandem software with the same criteria. PCG: protein-coding genes.

Figure 2 Characterization of unknown pig protein isoforms

A. Confirmation of 24,431 identified protein isoforms by other pig protein databases. **B.** Classification of unknown pig protein isoforms. Bar chart and pie chart respectively show the numbers and percentages of three categories in 5841 unknown pig protein isoforms. **C.** Relationship between the improvement of genome quality and novel proteins for each chromosome.

Figure 3 The pig transcriptome in unknown protein isoforms

A. Comparison of number of isoforms expressed per gene between humans and pigs. The box plots compare the number of isoforms expressed per gene within three transcript sets, including known human Ensembl set, known pig Ensembl set, and the newly identified set in the current study. **B.** The heatmap for Pearson correlation of gene expression between 34 tissues. The heatmap was used to reveal the pairwise correlation between all 34 pig tissues. **C.** Bar chart for tissue-based transcriptomic evidence of unknown protein isoforms. The x-axis represents the number of tissues and the y-axis represents the proportion of proteins identified in different number of tissues

.

Figure 4 Expression landscape in pig transcriptome

A. The network plot for the overview of pig PCGs. The red nodes represent the tissue types. Numbers in the yellow, blue, and green nodes respectively represent the number of the genes that are ubiquitously-expressed (expressed in all tissues), tissue-enriched, and group-enriched. G1‑G7 respectively mean immune organs, female reproductive system, male reproductive system, liver and gall bladder, adrenal gland and thymus gland, muscle tissues, and brain tissues. **B.** Numbers of tissue-enriched isoforms for known and unknown protein isoforms. **C.** Numbers of group-enriched isoforms in different tissue groups. **D.** The group-enriched expression of gene *MYL3* in muscular system. The gene levels (FPKM) for *MYL3* gene from different tissue categories (muscular system and non-muscular system) between humans and pigs. **E.** The group-enriched expression of gene *ENC1* in brain tissue. The gene levels (FPKM) for *ENC1* gene from different tissue categories (brain and Non-brain tissues) between humans and pigs. **F.** Expression landscape of ubiquitously expressed genes in 34 tissues. **G.** Hierarchical cluster tree for all ubiquitously expressed genes. 24 modules corresponding to branches are labelled in 24 different colors. FPKM, fragments per kilobase per million.

Figure 5 Classification of subcellular components within pig proteome

A. Pie charts for subcellular location of pig protein isoforms. Pie charts show the percentage of subcellular locations for all pig protein isoforms. **B.** Venn diagram for subcellular location of pig proteins. Venn diagram reveals the number of genes in each of the three main subcellular location categories: membrane, secreted, and soluble. The overlap between the categories gives the number of genes with isoforms belonging to two or all three categories. **C.** The proportion of protein isoforms in 34 tissues to different subcellular components. **D.** The proportion of three subcellular components in 34 tissues. We respectively selected the levels of expression with top 10, top 100, top 1000, and all proteins as protein sets for each tissue.

Figure 6 Orthologous of unknown pig proteome across multiple species

A. The heatmap showing 3081 orthologous isoforms among 10 species. For each isoform, N represents the number of species that pigs shared homology with, whereas

percentage within the color bar means the percentage of genes in all 3081 homologous isoforms for the corresponding N. **B.** The distribution of novel proteins (indicated by the red lines) in each chromosome of the pig genome.

Supplementary materials

File S1 Validation of identified proteins: MS/MS spectra from 71 synthetic peptides with those obtained from analysis of pig tissues

Figure S1 Density distribution for number of unique peptide from 0 to 10

Figure S2 Density distribution for number of unique peptide from 0 to 20

Figure S3 Density distribution for number of unique peptide from 0 to 50

Figure S4 Density distribution for number of unique spectrum from 0 to 10

Figure S5 Density distribution for number of unique spectrum from 0 to 20

Figure S6 Density distribution for number of unique spectrum from 0 to 50

Figure S7 Density distribution for spectrum counts from 0 to 10

Figure S8 Density distribution for spectrum counts from 0 to 20

Figure S9 Density distribution for spectrum counts from 0 to 50

Figure S10 Density distribution for identification probability

Figure S11 The bar plot for number of protein with different peptide bins

Figure S12 The bar plot for number of protein with different coverage bins among 34 tissues

Figure S13 The bar plot for number of protein with different tissues among 10 coverage bins

Figure S14 The bar plot for number of proteins with different coverage bins

Table S1 Validation of 71 peptides from 31 proteins

Table S2 Quality information of novel proteins isoforms

Table S3 Overview of alignment within 34 tissues

Table S4 Gene expression patterns in 34 tissues

Table S5 The co-expression interactions of 6108 ubiquitously expressed gene

Table S6 Subcellular location of the pig proteome (isoform)

Table S7 Subcellular location of the pig proteome (protein)

Table S8 Details of homologous protein with 10 species

Table S9 GO terms for unknown proteome

Note: GO: Gene ontology.

Table S10 KEGG pathways for unknown proteome

Note: KEGG: Kyoto Encyclopedia of Genes and Genomes.

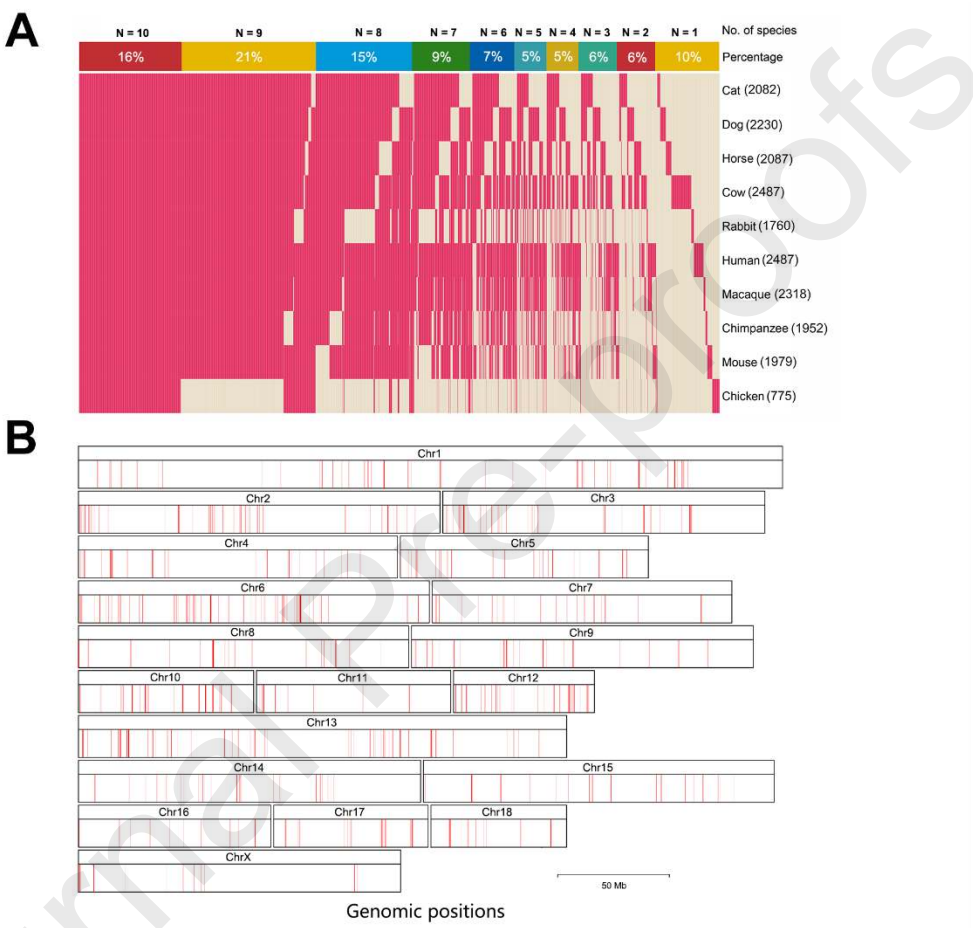
Table S11 KEGG disease pathways for unknown proteome

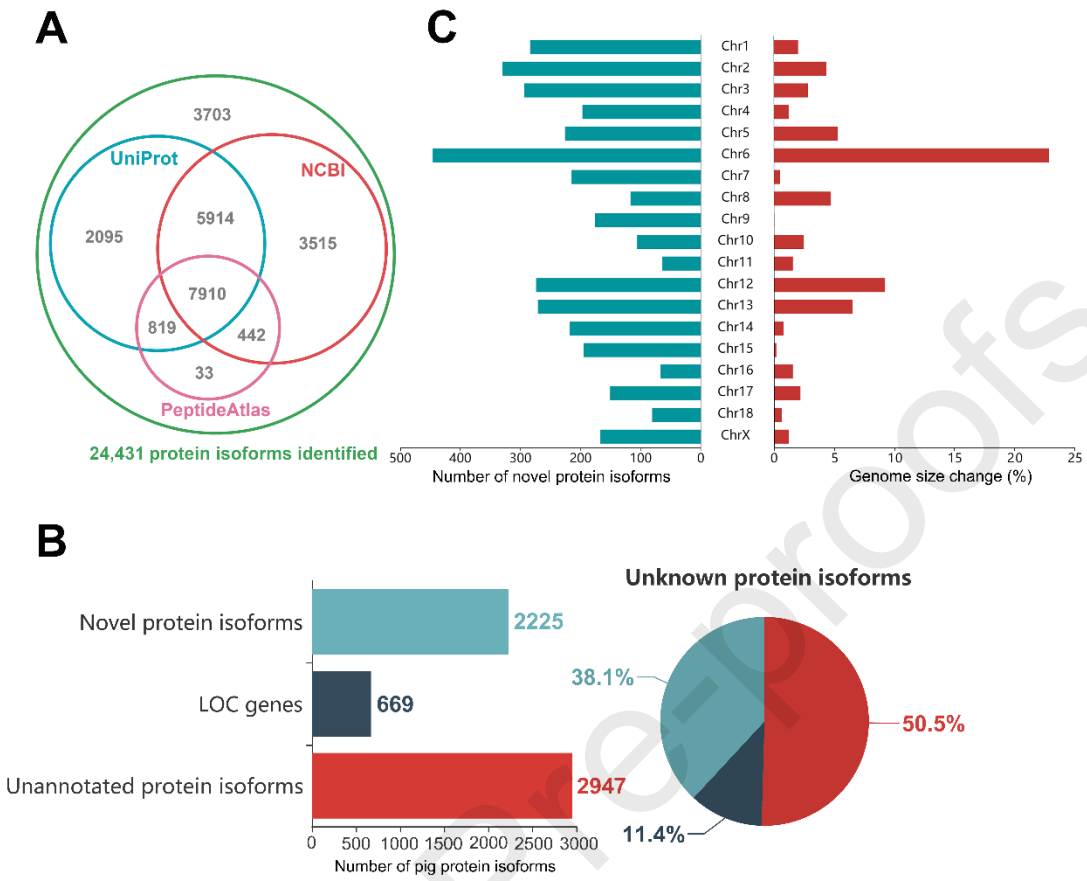
Table S12 Integrated data of proteome

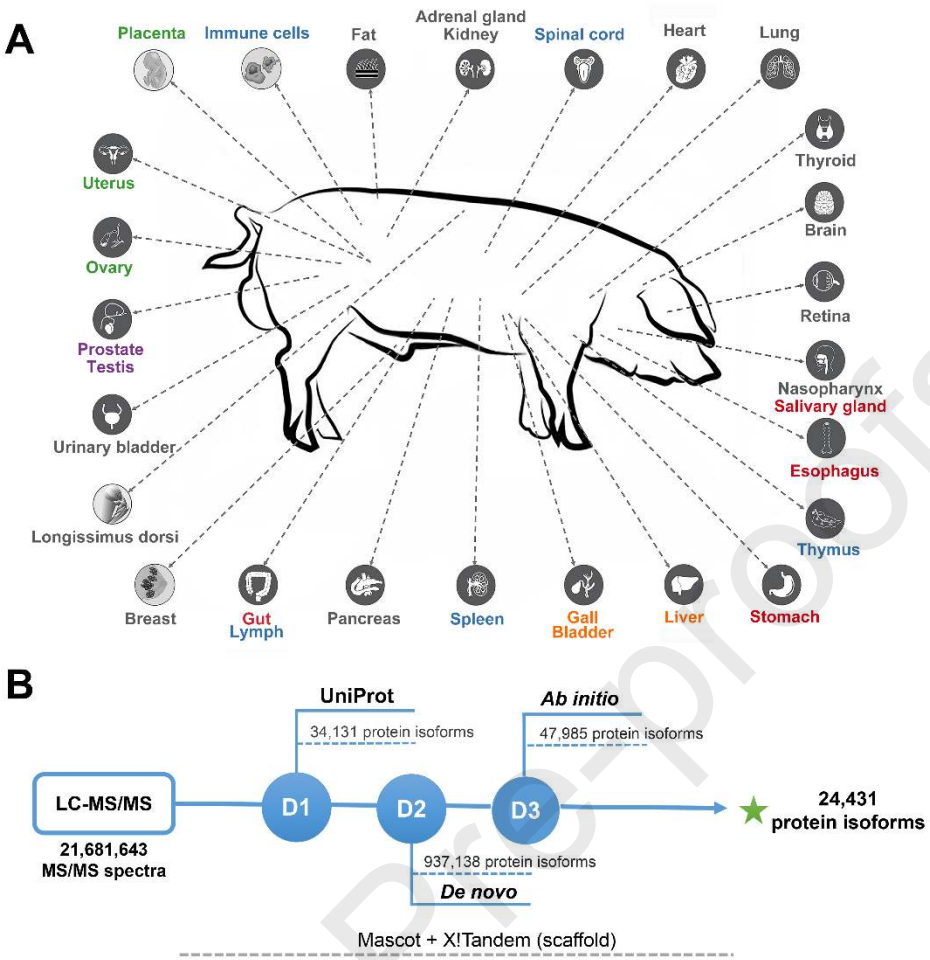
Table S13 Integrated data of transcriptome in the 34 pig tissues

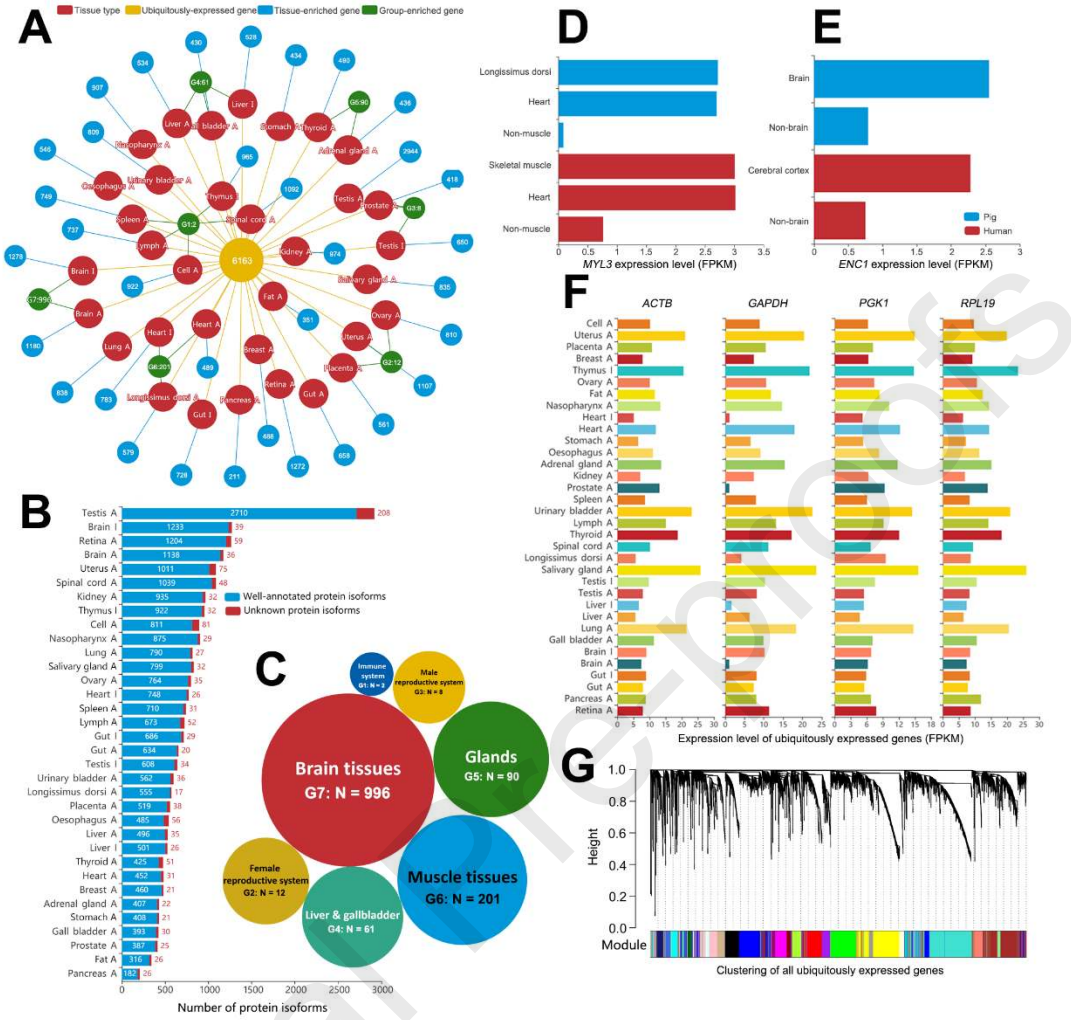
Table S14 Overview of functionally predictive resource for 5841 unknown protein isoforms

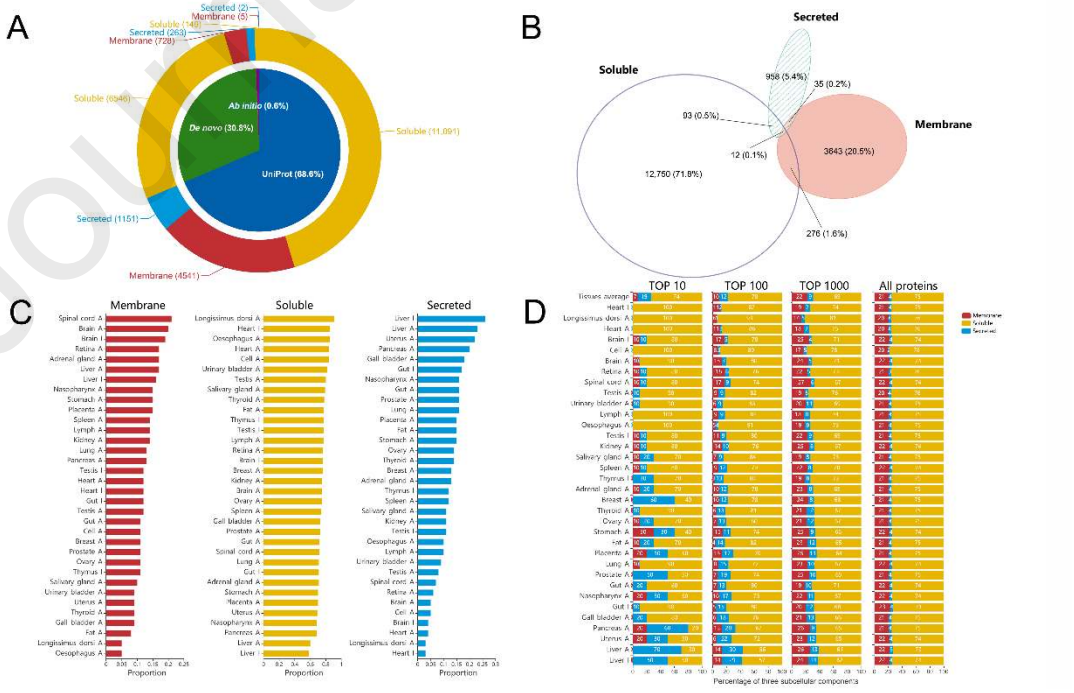
Table S15 RNA-sequencing resource tables








A
Tissue type
Ubiquitously-expressed gene
Tissue-enriched gene
Group-enriched gene
B
Well-annotated protein isoforms
Unknown protein isoforms
Number of protein isoforms
C
Brain tissues
G7: N = 996
Glands
G5: N = 90
Muscle tissues
G6: N = 201
Liver & gallbladder
G4: N = 61
Female reproductive system
G2: N = 12
Male reproductive system
G3: N = 8
Immune system
G1: N = 8
D
Longissimus dorsi
Heart
Non-muscle
Skeletal muscle
MYL3 expression level (FPKM)
E
Brain
Non-brain
Cerebral cortex
Pig
Human
ENC1 expression level (FPKM)
F
ACTB
GAPDH
PGK1
RPL19
Expression level of ubiquitously expressed genes (FPKM)
G
Height
Module
Clustering of all ubiquitously expressed genes


A
Secreted (2)
Membrane (5)
Soluble (140)
Secreted (263)
Membrane (728)
Soluble (6946)
Ab initio (0.6%)
De novo (30.8%)
UniProt (68.6%)
Soluble (11,091)
Secreted (1151)
Membrane (4541)
B
Secreted
Soluble
Membrane
958 (5.4%)
35 (0.2%)
93 (0.5%)
12 (0.1%)
3843 (20.5%)
12,750 (71.8%)
276 (1.6%)
C
Membrane
Soluble
Secreted
Proportion
D
TOP 10
TOP 100
TOP 1000
All proteins
Percentage of three subcellular components