Divergent and convergent evolution of housekeeping genes in human-pig lineage

Kai Wei [1] , **Tingting Zhang** [1] , **Lei Ma** Corresp. [1]

[1] College of Life Science, Shihezi University, Shihezi, Xinjiang, China

Corresponding Author: Lei Ma
Email address: malei1979@hotmail.com

Housekeeping genes are ubiquitously expressed and maintain basic cellular function across tissue/cell types conditions. The present study aimed to develop a set of pig housekeeping genes and compare characteristics of structure, evolution and function of housekeeping genes in the human-pig lineage. Using RNA sequencing data, we identified a list of 3,136 pig housekeeping genes. Comparing to human homologous counterparts, we found pig housekeeping genes were longer and subjected to slight weaker purifying selection pressure and faster neutral evolution. Common housekeeping genes, shared by the two species, have stronger purifying selection than species-specific genes. But pig-specific and human-specific housekeeping genes have similar functions. Some species-specific housekeeping genes have evolved independently to form similar protein-active sites or structure, such as classical catalytic serine-histidine-aspartate triad and zinc finger features, implying that they have converged for maintaining the basic cellular function, which led to equivalent solutions for adapting to the environment. Human and pig housekeeping genes have varied in their structure and gene list, but they have converged on the maintenance of basic cellular functions essential for the existence of a cell, regardless of its specific role in the species. The results shed light on the evolutionary dynamics of housekeeping genes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3450v1 | CC BY 4.0 Open Access | rec: 6 Dec 2017, publ: 6 Dec 2017

Divergent and convergent evolution of housekeeping genes in human-pig lineage

Kai Wei †, Tingting Zhang †, Lei Ma *

College of Life Science, Shihezi University, Shihezi City, Xinjiang Province, China

†These authors contributed equally to this work

*Corresponding author

Lei Ma: malei1979@hotmail.com

Abstract

Housekeeping genes are ubiquitously expressed and maintain basic cellular function across tissue/cell types conditions. The present study aimed to develop a set of pig housekeeping genes and compare characteristics of structure, evolution and function of housekeeping genes in the human-pig lineage. Using RNA sequencing data, we identified a list of 3,136 pig housekeeping genes. Comparing to human homologous counterparts, we found pig housekeeping genes were longer and subjected to slight weaker purifying selection pressure and faster neutral evolution. Common housekeeping genes, shared by the two species, have stronger purifying selection than species-specific genes. But pig-specific and human-specific housekeeping genes have similar functions. Some species-specific housekeeping genes have evolved independently to form similar protein-active sites or structure, such as classical catalytic serine-histidine-aspartate triad and zinc finger features, implying that they have converged for maintaining the basic cellular function, which led to equivalent solutions for adapting to the environment. Human and pig housekeeping genes have varied in their structure and gene list, but they have converged on the maintenance of basic cellular functions essential for the existence of a cell, regardless of its specific role in the species. The results shed light on the evolutionary dynamics of housekeeping genes.

Keywords: Housekeeping genes; Gene structure; Basal cellular function; Convergent evolution; Pig

Background

Housekeeping genes are typically genes consistently expressed across tissues and developmental stages for the maintenance of basic cellular functions (Butte et al.2001; Zhu et al.2003). They have unique genomic features, including gene structure (Eisenberg and Levanon 2003;

Vinogradov 2004), nucleotide composition (Vinogradov 2003), and upstream sequence conservation (Farré et al.2007; Belloraet al.2007). They are often considered as the minimally essential gene set for normal cellular physiology (Butte et al.2001) and are widely used as internal controls for gene expression experiments as well as computational biology studies (Thellin et al.1999; Robinson and Oshlack 2010;Rubie et al.2005; Vandesompele et al.2002).

In previous studies, many human housekeeping gene sets have been identified. However, some sets have little overlap. For example, only 155 genes were shared by three lists of microarray-defined housekeeping genes, including 501, 425 and 567 genes, respectively (Warrington et al.2000; Hsiao et al.2001; Eisenberg and Levanon 2003). The low overlap may be explained by several reasons. First, their complex transcriptional organization may cause diverse definitions of housekeeping genes (Gingeras 2007). Second, the expression of some housekeeping genes may vary depending on experimental conditions (Greer et al.2010). The question of why these genes vary across conditions awaits further investigations. Third, traditional techniques have their own drawbacks. For instance, the microarray technology has limited dynamic range and sensitivity, and also suffers from poor detectability and reproducibility for low-copy and transiently-expressed genes (Marioni et al.2008; Fu et al.2009; Bradford et al.2010; Draghici et al.2006).

RNA sequencing (RNA-seq) data greatly improve the detectability of housekeeping genes. For example, the amount of human housekeeping genes revisited by the RNA-seq data has increased ten-fold the previous estimates based on microarray data (Eisenberg and Levanon 2013). With advances in technology, large-scale RNA sequencing has provided new insights into the definition of housekeeping genes. Some studies have suggested that transcripts should be used as housekeeping units (Gingeras 2007; Gerstein et al.2007).

The comparative analysis of housekeeping genes between human and other animals is of great interest. Human housekeeping genes are commonly used as control genes in the real-time quantitative polymerase chain reaction (qRT-PCR) for other animals. However, whether human genes can be used as references for other animals remains unclear. For instance, the most commonly used human reference genes (e.g. *ACTB* and *GAPDH*) do not always apply to all tissues of different organisms (Brattelid et al.2010; Kozera et al.2013). Therefore, to well define a housekeeping genes set in another animal may be valuable.

As an important meat resources for humans, the pig (*Sus Scrofa*) is a well-studied organism. And because of anatomical similarities with humans, the pig is often used as a biomedical model in research as well (Lunney 2007; Rolandsson et al.2002; Lee et al.2009; Becker et al.2010). Surveying pig housekeeping genes may help pave the way for a greater understanding basal mechanisms that maintain cell function. In the present study, we identified housekeeping genes in pig using the RNA-seq data, and then compared their structure and function with human orthologs. In addition, we discussed the impact of selection pressure and convergent evolution on functional conservation of housekeeping genes. The present study provided detailed information of pig housekeeping genes and their functional features, and offered insights into evolutionary dynamics on them.

Materials and Methods

Data preparation

In order to define housekeeping gene sets, the gene expression datasets were downloaded from Sequencing Read Achieve (SRA) database of National Center for Biotechnology Information (NCBI, Sep, 2016) (Kodama et al.2012). In addition, pig genomic annotation (*Sus Sscrofa*10.2) was downloaded from the Ensembl Genome Browser (Sep, 2016) (Kinsellaet al.2011). The RNA-seq dataset of 14 experiments were used to identify housekeeping genes, which were derived from 21 tissues (heart, spleen, liver, kidney, lung, musculus longissimus dorsi, occipital cortex, hypothalamus, frontal cortex, cerebellum, endometrium, mesenterium, greater omentum, backfat, gonad, ovary, placenta, testis, blood, uterine and lymph nodes), containing a total of 131 samples(Supplementary material1: Table S1). The SRA files were downloaded from the NCBI and then converted to fastq files using fastq-dump (Kodama et al.2012). RNA-seq reads were then filtered by IlluQC.pl (Patel and Jain 2012) while requiring an average read quality above 20, and then were aligned to pig genome sequence (Sus Sscrofa10.2) using Tophat (Trapnell et al.2009; Külahoglu et al.2014; Ghosh S, Chan et al.2016). The alignments were then fed to an assembler Cufflinks (Trapnel et al.2010) to assemble aligned RNA-seq reads into transcripts and estimate their abundances, which were measured in Fragments Per Kilobase of exon per Million fragments mapped (FPKM).

To define housekeeping genes

Housekeeping genes were defined according to the following criteria: (i) the transcripts could be detected in all 21 tissues; (ii) the transcripts showed low expression variance across tissues: *P* > 0.1 (Kolmogorov-Smirnov test); (iii) no exceptional expression in any single tissue; that is, the expression values were restricted within the fourfold range of the average across tissues; and (iv) all transcripts of a housekeeping candidate gene met the above criteria.

Structure analysis of housekeeping genes

The structure data of genes were taken from the Ensembl BioMart (Kinsella et al.2011). Human housekeeping genes were derived from the reference (Eisenberg and Levanon 2013), considering its similar type of data and stringency of the definition. We obtained 3,136 and 3,804 housekeeping genes of pig and human, respectively. Length of various parts of housekeeping genes between them were compared by Mann-Whitney test (Table 1).

Gene ontology analysis of housekeeping genes

The analysis of functional annotations of housekeeping genes was performed using DAVID, ver. 6.7, available on their website (Huang da et al.2009; Huang da et al.2009). All expressed genes in the data were used as background. Comparative analysis of housekeeping genes between human and pig was performed. The false discovery rates (FDR) were calculated to estimate the extent to which genes were enriched in GO categories (Ashburner et al.2000). Probabilities less than 0.01 were used as the cut-off value and considered to show significant level of the correlation. Heat map analysis was also conducted through DAVID outcomes to visualize a matrix of enriched GO.

Evolutionary feature analysis of housekeeping genes

The number of non-synonymous substitutions per non-synonymous site (dN) and the number of synonymous substitutions per synonymous site (dS) were estimated using the Nei-Gojobori method embedded in MEGA 7.0 (Z-test, P<0.05)(Kumar et al.2016; Nei and Kumar 2000). From the Scope row, select the Overall Average option. For the Gaps/Missing data treatment option, select Pairwise Deletion. The genome sequence of orthologous genes were downloaded from Ensembl BioMart. The dN/dS ratios were calculated to assess selection pressure (Hurst 2002; Yang and Nielsen 2002; Dasmeh et al.2014). The information of active sites and zinc fingers of proteins were obtained from UniProt Knowledgebase (UniProtKB) (Boutet et al.2016; Pundir et al.2015). Species-specific housekeeping genes that have similar function were processed to search their active sites or zinc fingers.

Results

Gene expression profile

To identify the housekeeping genes in pig, we surveyed the expression distribution of 30,585 transcripts across 21 tissues of pig (see Methods, Figure 1, Supplementary material 1: Figure S1). The detectability of RNA-seq data was high, and only 116 transcripts undetected in the present study. The 226 transcripts showed tissue-specific expression(expressed in one tissue), whereas 6072 transcripts was found broadly expressed in all tissues (Figure 1). This finding was consistent with the expression tissue-breadth of human genes (Zhu et al.2008; Eisenberg and Levanon 2013).

Identification of pig housekeeping genes

To obtain the transcripts with the ubiquitous expression level across pig tissues, we selected the transcripts detected in all tissues and then obtained 6072 candidates. The background differences between different sequencing projects result in batch effect between samples, including difference of sequencing depth and coverage. Therefore, we chose a single sequencing project to assess the uniformity of gene expression, which contains a larger sample size. Furthermore, the expression uniformity of those candidates in ERP002055 sequencing project was tested by the Kolmogorov-Smirnov (K-S) test and then was accessed by the *P*-value of the test(Farajzadeh et al.2013). Figure S2 of Supplementary material 1 represents the frequencies of the candidates with the *P*-value being greater than the given cutoff. For about 67% of all candidates, the *P*-values were above 0.1, implying their expression levels were not significantly varied across tissues and had a high level of the expression uniformity. Therefore, we defined the cutoff of the uniform level as $P > 0.1$ for the following analyses, which resulted in a list of 4068 unique transcripts, belonging to 3754 genes. The housekeeping gene was further restricted into the gene whose all transcripts passed the criteria. Altogether, the 3,136 genes passed the restriction (Supplementary material 2), about a third of which were unannotated.

Figure 2 shows the overlap of pig housekeeping genes identified in the present study with previously reported human housekeeping genes (Warrington et al.2000; Hsiao et al.2001; Eisenberg and Levanon 2003; Eisenberg and Levanon 2013). In order to more accurately describe the features, housekeeping genes were grouped into three sets of genes, namely, common housekeeping genes observed both in pig and human, human-specific and pig-specific housekeeping genes. We obtained 1,012 common, 2,792 human-specific and 2,124 pig-specific housekeeping genes, respectively.

Structure comparison of housekeeping genes between pig and human

The comparison of length distribution of total intron, 5’ untranslated region (UTR) and coding sequence (CDS) in homologous housekeeping genes shows that pig genes dominates the fraction of long length whereas human genes are prone to short length (Figure 3A - C). Furthermore, Table 1 compares the average lengths of various structures of the housekeeping genes that correspond to one another in pig and human. All structures of pig housekeeping genes were significantly longer than human's (Table 1), which were consistent with the previous analyses of pig genomes (Groenen et al.2012), implying that different purifying selection pressures were applied between pig and human. Selective pressure may make gene as short as possible for reducing the cost in the transcription process (Ucker and Yamamoto 1984; Castillo-Davis et al.2002).

Evolutionary dynamics of housekeeping genes

Evolutionary features of housekeeping genes may provide a deeper understanding for the evolutionary trend of housekeeping gene in different species. For the maintenance of essential function, housekeeping genes are thought to evolve more slowly than other genes (Zhang and Li 2004). To survey that feature, the number of non-synonymous substitutions per non-synonymous site (dN), the number of synonymous substitutions per synonymous site (dS) and dN/dS ratio were calculated for pig and human housekeeping genes using mouse(*Mus musculus*) as outgroup (Supplementary material 3 and 4), respectively. Generally, synonymous substitutions occurred randomly and do not appear to change the gene function, but the non-synonymous substitutions occurred nonrandomly, which may change the function of housekeeping genes and suffer strong selection pressure (Nei and Kumar 2000, Kimura 1983).

The dN followed a power law distribution similar to that of the dN/dS (Figure 4A, Supplementary material 1: Figure S3A), displaying a relatively large number of genes with a few non-synonymous substitutions and a small fraction of genes with much more substitutions

(Figure 4A). In addition, most of the dN/dS ratios were lower than one, implying that purifying selection have acted on housekeeping genes to ensure the stability of most of genes' function. The less the dN/dS ratio is, the stronger purifying selection is. Furthermore, purifying selection pressure on housekeeping genes were slightly stronger in human than in pig (Figure 4A, B).

The dN/dS ratios of common housekeeping genes showed no difference between pig and human, but the ratios of species-specific housekeeping genes were significantly lower in human than in pig (Mann-Whitney test, $P < 0.05$) (Figure 4B, Figure 5D). Furthermore, for both human and pig, the dN/dS ratios of common genes were significantly lower than species-specific genes (Figure 5A for pig and Supplementary material 1: Figure S4 for human). This result suggested that common housekeeping genes suffered more stringent purifying selection to remove alleles than species-specific genes.

On the other side, these results of the dN/dS (or dN) also implied that human housekeeping genes have evolved more stably than pig genes (Figure 5B-D). The dS of human species-specific genes were prone towards lower values than pig genes (Figure 5C), showing that human housekeeping genes have slower neutral evolution than pig housekeeping genes.

The dS followed an approximately normal distribution (Supplementary material 1：Figure S3B), occurring to be around a central value (0.77 and 0.63 in pig and human housekeeping genes, respectively). This finding implies the random tendency of synonymous substitutions. There was no statistic difference in the synonymous substitutions between common and species-specific genes within a species (Figure 5A for pig and Supplementary material 1: Figure S4 for human).

In addition, considering the mouse is close to human and pig in phylogeny, and may be more close to human(Meredith et al. 2011). So, we also selected elephant (*Loxodonta africana*) as outgroup to calculate dN,dS, and dN/dS for pig and human housekeeping genes, respectively(Additional 5 and 6). Furthermore, all analyses of evolutionary dynamics were performed to verify foregoing results using elephant as outgroup, and the results is similar to the previous analysis of mouse as outgroup (Supplementary material 7).

Associated function of housekeeping genes

We then characterized the housekeeping genes that enriched molecular function, biological process, cellular component, and disease, respectively, based on the Database for Annotation, Visualization, and Integrated Discovery (DAVID) program. The heat map shown in Figure 6 illustrates the similar enrichment of housekeeping genes between pig and human. Briefly, housekeeping genes were predominantly detected as the genes associated with Gene Ontology (GO) terms related to basal metabolism that are indispensable for cellular physiology, indicating housekeeping genes are essential for basic physiological processes (Figure 6).

It was worth noting that many pig housekeeping genes were enriched in human diseases, especially in several cancers with high mortality rates: breast cancer, lung cancer and colorectal cancer (Figure 6D). This finding may be beneficial for studies of human disease (Tu et al.2006), given that pig may not have some human risk genes. For instance, alcohol-induced cirrhosis was enriched in human housekeeping genes, but not in pig.

Functional convergence

Interestingly, the functional enrichment analyses showed a coherent trend in pig and human housekeeping genes although the low overlap of gene lists and the difference in gene structure between the two species were found. For example, for biological process, pig and human showed a slight difference in the GO term enrichment (Figure 6A). In addition, similar trends were also observed in the active molecules that related to basic metabolism and gene expression (Figure 6B and C).

The above analysis revealed that functions of housekeeping genes between pig and human were consistent, implying that selection pressure may preclude the species-differentiation of housekeeping genes for the maintenance of basal cellular functions, especially for species-specific housekeeping genes. To confirm this conjecture, we performed functional enrichment analysis for common and species-specific housekeeping genes, respectively. The heat map shown in Figure 7 illustrates the more similarity between two species-specific terms than between common and species-specific terms. These results indicated housekeeping genes suffered strong selection pressure for maintaining normal life activities, and human and pig species-specific housekeeping genes converged on the basal cellular function.

Mechanistic convergence

To understand the mechanistic constraints on the function of housekeeping proteins, we analyzed the evolutionary constraints on protein structure, active site feature and chemical reaction center. We found some similar active site features in housekeeping peptidases (Figure 8, Table 2), which reflected the intrinsic chemical constraints on enzymes, leading evolution to independently converge on equivalent solutions repeatedly (Buller and Townsend 2013; Dodson and Wlodawer

1998). The chemical and physical constraints on enzyme catalysis have caused identical triad arrangements in housekeeping peptidases in human-pig lineage, such as classical catalytic Ser/His/Asp triad and non-classical variants (Table 2). However, the peptide sequences and three-dimensional structure profiles of them were totally different (Figure 8A and B). Classical Ser/His/Asp catalytic triad is a universal phenomenon in the serine protease class (E.C. 3.4.21), where serine is the nucleophile, histidine is the general base or acid, and the aspartate helps orient the histidine residue and neutralize the charge that develops on the histidine during the transition states (Polgar 2005; Ekici et al.2008). Interestingly, almost all proteins in Table 2 contained histidine as an active site to provide a proton receptor (Wang et al.2006). In addition, Cys/His and Glu/His/Asp in peptidases also evolved convergent; however, these active sites have rarely been mentioned in previous reports to our knowledge.

Structural convergence

Moreover, many housekeeping proteins tended to form common zinc finger features involved in the regulation of gene expression (Figure 9, Supplementary material 1: Table S2 and S3). For example, C_2H_2 type is one of major zinc fingers in transcription factors (Wolfe et al.2000; Li et al.2004). This analysis of housekeeping protein structure and function revealed several interrelated and previously unrecognized relationships of structure–function constraints. These fundamental constraints have promoted the convergent evolution of housekeeping genes, especially for species-specific housekeeping genes and low homology genes.

Discussion

In the present study, we defined a set of pig housekeeping genes with a wide range of expression and low expression variation across tissues. The present set of housekeeping genes in pig showed lower overlap with a human set. Some housekeeping genes of human were not in our list, such as *GAPDH* and *ACTB* (Barber et al.2005;de Jonge et al.2007; Nygard et al.2007), thus whether human housekeeping genes can be used as reference controls for other species remains to be further verified.

After divergence from common ancestor, pig and human have accumulated difference in the sequence and structure of housekeeping genes. On a molecular level, that can happen from random mutation, for example, the synonymous substitution. The dS distribution followed an approximately normal distribution, showing a random tend of synonymous substitutions. On the other side, the divergence was also related to adaptive changes. Human housekeeping genes were found to be shorter than pig genes (Figure 3A - C). The possible reason is food intake and stored energy is less in.human than pig, so the shorter structure is good for human to consume less time and cost in the process of gene expression (Ucker and Yamamoto 1984; Izban and Luse 1992). In addition, the stronger purifying selection in human comparing to pig (Figure 4A) might result in a lower degree of genetic redundancy as well (Zhang and Li 2004). In other words, human housekeeping genes would have evolved more stably than pig, because advantageous and stable living environment. Moreover, human and pig have evolved their own species-specific housekeeping genes, which might lead to the formation of the two species, allowing differentiated fixation of characteristics. In addition, purifying selection is stronger in common than in species-specific housekeeping genes and show some differences in GO enrichment. This may indicate common housekeeping genes were more indispensable than species-specific and involve more functions for sustain life. Such as *GTF2H1* (general transcription factor IIH subunit 1) and *CXXC1* (CXXC finger protein 1) in common are crucial for regulation of many of gene expression(Shiekhattar et al.1995; Andersen et al.2001), but in species-specific housekeeping genes were not enrichment.

However, although human and pig have been divergent for millions of years, both species independently converged towards similar features of housekeeping genes. One of the most unexpected observations stemmed from species-specific housekeeping genes. The GO enrichment analysis revealed that pig-specific and human-specific housekeeping genes have similar functions. In addition, some housekeeping proteins evolved independently to have similar active sites, sidechains, catalytic centers or binding sites to complete similar catalytic reaction or molecular function (Buller and Townsend 2013; Polgar 2005; Ekici et al.2008; Brannigan et al.1995; Chen et al. 2008; Klug 2010; Klug 1999; Hall 2005; Brown 2005), although these proteins showed very low homology with each other. They have "converged" on the maintenance of basic cellular functions, which led to equivalent solutions for adapting to the environment (Nielsen 2005; Hurst 2009). Functional similarity across species may be caused by adaptive evolution (Zhang and Li 2004; Kimura 1983), which drive different species-specific genes to perform similar essential functions, regardless of its specific role in species.

As known, it is still under investigation to attain large-scale gene expression profile. The current transcriptome sequencing data in pig may be inadequate to meet the requirement to define the housekeeping genes. The accurate definition of housekeeping genes is still an unresolved issue. Therefore, the present set of pig housekeeping genes had limitations, but it successfully offered some instances, the characteristics of which were similar to those reported in previous studies. As new technologies emerge, high-quality deep-sequencing transcriptome profiling data may open up opportunities to improve the stringency in defining housekeeping genes and narrowing the catalog of housekeeping genes that are expressed in a single cell (Tang et al.2009). Furthermore, the advancement of statistical methods will greatly improve housekeeping gene detection. More specifically, the concept of "housekeeping" or "maintenance" should be defined in a hierarchical way related to cell types, growth stages, cell cycles as well as various physiological conditions, and in terms of specific transcript variant (Zhu et al.2008). Thus, we will be able to observe several sets of housekeeping genes in a single species. In addition, more stringent sets of housekeeping genes will also provide powerful support for structural and functional genomics, especially to analyze the cellular basal function of different species (Kumar and Hedges 1998; Meredith et al.2011; Kumar et al.2002).

Conclusions

The present study offered insight into the general aspects of housekeeping gene structure and evolution. Diverging from the ancestor of human and pig, housekeeping genes have varied in gene structure and gene list, but they have converged on the maintenance of basic cellular function that are essential for the existence of a cell, regardless of their specific role in species. The results in the present study will shed light on the evolutionary dynamics of the housekeeping genes.

Declarations

Ethics approval and consent to participate

We reused public data from the NCBI database and did not report on or involve the use of any another animal data.

Availability of data and material

All data generated or analysed during this study are included in this published article and its supplementary information files.

Funding

The research was supported by the National Natural Science Foundation of China (31272416, 31560310 and 31370762), the National High Technology Research and Development Program of China (863 program, 2013AA102502), the Scientific Research Foundation of the MHRSS of China for the Returned Overseas Chinese Scholars and the Scholar Pair-training Program of Shihezi University (SDJDZ201504). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Kai Wei and Lei Ma designed the study. Kai Wei and Tingting Zhang performed the data analyses and drafted the manuscript. Lei Mai revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank all of the contributors of the RNA-seq data sets and the anonymous reviewers for helpful suggestions on the manuscript. We thank Dave Baab for copyediting the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author detail

[1] College of Life Science, Shihezi University, Shihezi City, Xinjiang Province, China

References

Butte AJ, Dzau VJ, Glueck SB. 2001. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". Physiol. Genomics, 7(2):95-96.

Zhu J, He F, Song S, Wang J, Yu J. 2008. How many human genes can be defined as housekeeping with current expression data? BMC Genomics, 9:172. doi: 10.1186/1471-2164-9-172.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19(7):362-365. doi:10.1016/S0168-9525(03)00140-9.

Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 20(5):248-253. doi:10.1016/j.tig.2004.03.006.

Vinogradov AE. 2003. Isochores and tissue-specificity. Nucleic Acids Res. 31(17):5212-5220. doi:10.1093/nar/gkg699.

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. 8(7):R140. doi: 10.1186/gb-2007-8-7-r140.

Bellora N, Farré D, Albà MM. 2007. Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Genomics, 8:459. doi: 10.1186/1471-2164-8-459.

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. 1999. Housekeeping genes as internal standards: use and limits. J. Biotechnol. 75(2-3):291-295. doi:10.1016/S0168-1656(99)00163-7.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11(3):R25. doi: 10.1186/gb-2010-11-3-r25.

Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, Brittner B, Ludwig B, Schilling M. 2005. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Mol. Cell Probes. 19(2):101-109. doi:10.1016/j.mcp.2004.10.001.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3(7):RESEARCH0034.1. doi:10.1186/gb-2002-3-7-research0034.

Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol. Genomics, 2(3):143-147.

Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, MacDonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR. 2001. A compendium of gene expression in normal human tissues. Physiol. Genomics, 7(2):97-104.

Gingeras TR. 2007. Origin of phenotypes: genes and transcripts. Genome Res. 17(6):682-690. doi:10.1101/gr.6525007.

Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L. 2010. Housekeeping genes; expression levels may change with density of cultured cells. J. Immunol. Methods, 355(1-2):76-79. doi:10.1016/j.jim.2010.02.006.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18(9):1509-1517. doi:10.1101/gr.079558.108.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics, 10:161. doi: 10.1186/1471-2164-10-161.

Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ. 2010. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics, 11:282. doi: 10.1186/1471-2164-11-282.

Draghici S, Khatri P, Eklund AC, Szallasi Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22(2):101-109. doi:10.1016/j.tig.2005.12.005

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. Trends Genet. 29(10):569-574. doi:10.1016/j.tig.2013.05.010.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17(6):669-681. doi:10.1101/gr.6339607.

Brattelid T, Winer LH, Levy FO, Liestol K, Sejersted OM, Andersson KB. 2010. Reference gene alternatives to Gapdh in rodent and human heart failure gene expression studies. BMC Mol. Biol. 11:22. doi: 10.1186/1471-2199-11-22.

Kozera B, Rapacz M. 2013. Reference genes in real-time PCR. J. Appl. Genet. 54(4):391-406. doi:10.1007/s13353-013-0173-x.

Lunney JK. 2007. Advances in swine biomedical model genomics. Int J Biol Sci. 3(3):179-184.

Rolandsson O, Haney MF, Hagg E, Biber B, Lernmark A. 2002. Streptozotocin induced diabetes in minipig: a case report of a possible model for type 1 diabetes? Autoimmunity, 35(4):261-264.

Lee L, Alloosh M, Saxena R, Van Alstine W, Watkins BA, Klaunig JE, Sturek M, Chalasani N. 2009. Nutritional model of steatohepatitis and metabolic syndrome in the Ossabaw miniature swine. Hepatology, 50(1):56-67. doi:10.1002/hep.22904.

Becker ST, Rennekampff HO, Alkatout I, Wiltfang J, Terheyden H. 2010. Comparison of vacuum and conventional wound dressings for full thickness skin grafts in the minipig model.

International journal of oral and maxillofacial surgery. 39(7):699-704. doi:10.1016/j.ijom.2010.03.016.

Kodama Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40(Database issue):D54-56. doi: 10.1093/nar/gkr854.

Farajzadeh, L, Hornshoj H, Momeni J, Thomsen B, Larsen K, Hedegaard J, Bendixen C, Madsen LB. 2013. Pairwise comparisons of ten porcine tissues identify differential transcriptional regulation at the gene, isoform, promoter and transcription start site level. Biochem. Biophys. Res. Commun. 438(2):346-352. doi:10.1016/j.bbrc.2013.07.074.

Patel RK, Jain M. 2012. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PloS one., 7(2):: e30619. doi:10.1371/journal.pone.0030619.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat:discovering splice junctions with RNA-Seq. Bioinformatics. 25(9): 1105-1111. doi:10.1093/bioinformatics/btp120.

Külahoglu C, Bräutigam A. 2014. Quantitative Transcriptome Analysis Using RNA-seq. Methods Mol. Biol. 1158:71-91. doi: 10.1007/978-1-4939-0700-7_5.

Ghosh S, Chan KK. 2016. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol. Biol.1374:339-361. doi: 10.1007/978-1-4939-3167-5_18.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren J, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28(5):511-515. doi: 10.1038/nbt.1621.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33(7):1870-1874. doi:10.1093/molbev/msw054.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics, Oxford University Press,Oxford, pp.52-72.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18(9):486. doi:10.1016/S0168-9525(02)02722-1.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19(6):908-917. doi:10.1093/oxfordjournals.molbev.a004148.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. protoc. 4(1):44-57. doi: 10.1038/nprot.2008.211.

Huang da W, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1):1-13. doi: 10.1093/nar/gkn923.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25(1):25-29. DOI: 10.1038/75556.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Biol. Evol. 21(2):236-239. doi: 10.1093/molbev/msh010.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge, U.K.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature, 392(6679):917-920. doi:10.1038/31927.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik, E, Simao TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science, 334(6055):521-524. doi: 10.1126/science.1211028.

Barber RD, Harmer DW, Coleman RA, Clark BJ. 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol. Genomics, 21(3):389-395. doi:10.1152/physiolgenomics.00025.2005

de Jonge HJ, Fehrman RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A. 2007. Evidence based selection of housekeeping genes. PloS one, 2(9):e898. doi:10.1371/journal.pone.0000898.

Freilich S, Massingham T, Bhattacharyya S, Ponsting H, Lyons PA, Freeman TC, Thornton JM. 2005. Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. Genome Biol. 6(7):R56. doi:10.1186/gb-2005-6-7-r56.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. Trends Genet. 24(10):481-484. doi:10.1016/j.tig.2008.08.004.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature, 491(7424):393-398. doi: 10.1038/nature11622.

Ucker DS, Yamamoto KR. 1984. Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. J. Biol. Chem. 259(12):7416-7420.

Izban MG, Luse DS. 1992. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. J. Biol. Chem. 267(19):13647-13655.

Nielsen R. 2005. Molecular Signatures of Natural Selection. Annu. Rev. Genet. 39:197-218. doi: 10.1146/annurev.genet.39.073003.112420.

Hurst LD. 2009. Genetics and the understanding of selection. Nat. Rev. Genet. Doi:10(2):83-93. 10.1038/nrg2506.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. methods, 6(5):377-382. doi:10.1038/nmeth.1315.

Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. 2014. The influence of selection for protein stability on dN/dS estimations. Genome Biol. Evol. 6(10):2956-67. doi: 10.1093/gbe/evu223.

Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. 2006. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics, 7:31. doi: 10.1186/1471-2164-7-31.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA. 99(2):803-808. doi:10.1073/pnas.022629899

Nygard AB, Jorgensen CB, Cirera S, Fredholm M. 2007. Selection of reference genes for gene expression studies in pig tissues using SYBR green qPCR. BMC Mol. Biol. 8:67. doi: 10.1186/1471-2199-8-67.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat. Genet. 31(4):415-418. doi:10.1038/ng940.

Buller AR, Townsend CA. 2013. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. Proc Natl Acad Sci USA. 110(8):E653-661. doi: 10.1073/pnas.1221050110.

Polgar L. 2005. The catalytic triad of serine peptidases. Cell Mol. Life Sci. 62(19-20):2161-2172. doi: 10.1007/s00018-005-5160-x

Ekici OD, Paetzel M, Dalbey RE. 2008. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. Protein Sci. 17(12):2023-2037. doi: 10.1110/ps.035436.108.

Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, Tomchick DR, Murzin AG. 1995. A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. Nature, 378(6555):416-419. doi:10.1038/378416a0.

Chen L, Wang H, Zhang J, Gu L, Huang N, Zhou JM, Chai J. 2008. Structural basis for the catalytic mechanism of phosphothreonine lyase. Nat. Struct. Mol. Biol. 15(1):101-102. doi:10.1038/nsmb1329.

Wang LJ, Sun N, Terzyan S, Zhang XJ, Benson DR. 2006. A Histidine/Tryptophan π-Stacking Interaction Stabilizes the Heme-Independent Folding Core of Microsomal Apocytochrome b5Relative to that of Mitochondrial Apocytochrome b5. Biochemistry 45 (46): 13750 -13759. doi: 10.1021/bi0615689.

Wolfe SA, Nekludova L, Pabo CO. 2000. DNA recognition by Cys2His2 zinc finger proteins. Annu. Rev. Biophys. Biomol. Struct. 29:183-212. doi: 10.1146/annurev.biophys.29.1.183.

Li L, He S, Sun JM, Davie JR. 2004. Gene regulation by Sp1 and Sp3. Biochemistry and cell biology , 82(4):460-471. doi: 10.1139/o04-045.

Klug A. 2010. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Q. Rev. Biophys. 43(1):1-21. doi:10.1017/S0033583510000089.

Klug A. 1999. Zinc finger peptides for the regulation of gene expression. J. Mol. Biol. 293(2):215-218. doi: 10.1006/jmbi.1999.3007.

Hall TM. 2005. Multiple modes of RNA recognition by zinc finger proteins. Curr. Opin. Struct. Biol. 15(3):367-373. doi:10.1016/j.sbi.2005.04.004.

Brown RS. 2005. Zinc finger proteins: getting a grip on RNA. Curr. Opin. Struct. Biol. 15(1):94-98. doi:10.1016/j.sbi.2005.01.006.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios, I. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol. Biol. 1374:23-54. doi:10.1007/978-1-4939-3167-5_2.

Pundir S, Magrane M, Martin MJ, O'Donovan C. 2015. Searching and Navigating UniProt Databases. Curr. Protoc. Bioinformatics, 50:1.27.1-10. doi: 10.1002/0471250953.bi0127s50.

Dodson G, Wlodawer A. 1998. Catalytic triads and their relatives. Trends Biochem. Sci. 23(9):347-352. doi:10.1016/S0968-0004(98)01254-7

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science, 334(6055):521-524. doi: 10.1126/science.1211028

Shiekhattar R, Mermelstein F, Fisher RP, Drapkin R, Dynlacht B, Wessling HC, Morgan DO, Reinberg D. 1995. Cdk-activating kinase complex is a component of human transcription factor TFIIH. Nature, 374(6519):283-287. doi:10.1038/374283a0.

Lee JH, Voo KS, Skalnik DG. 2001. Identification and characterization of the DNA binding domain of CpG-binding protein. J. Biol. Chem. 276(48):44669-44676. doi: 10.1074/jbc.M107179200.

Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Perez-Enciso M. 2011. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics, 12:552. doi:10.1186/1471-2164-12-552.

Martinez-Montes AM, Fernández A, Pérez-Montarelo D, Alves E, Benitez RM, Nuñez Y, Óvilo C, Ibañez-Escriche N, Folch, JM, Fernández AI. 2016. Using RNA-Seq SNP data to reveal potential causal mutations related to pig production traits and RNA editing. Anim. Genet. 48(2):151-165. doi:10.1111/age.12507.

Wang T, Jiang A, Guo Y, Tan Y, Tang G, Mai M, Liu H, Xiao J, Li M, Li X. 2013. Deep sequencing of the transcriptome reveals inflammatory features of porcine visceral adipose tissue. Int. J. Biol. Sci. 9(6):550-556. doi:10.7150/ijbs.6257.

Pérez-Montarelo D, Madsen O, Alves E, Rodriguez MC, Folch JM, Noguera JL, Groenen MA, Fernández AI. 2014. Identification of genes regulating growth and fatness traits in pig through hypothalamic transcriptome analysis. Physiol. Genomics, 2014, 46(6):195-206. doi:10.1152/physiolgenomics.00151.2013.

Jiang S, Wei H, Song T, Yang Y, Peng J, Jiang S. 2013. Transcriptome comparison between porcine subcutaneous and intramuscular stromal vascular cells during adipogenic differentiation. PloS one. 8(10):e77094. doi:10.1371/journal.pone.0077094.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CKL, Chen L, Ma J. et al. 2013. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat. Genet. 45(12):1431-1438. doi:10.1038/ng.2811.

Samborski A, Graf A, Krebs S, Kessler B, Bauersachs S. 2013. Deep sequencing of the porcine endometrial transcriptome on day 14 of pregnancy. Biol. Reprod. 88(4):84. doi:10.1095/biolreprod.113.107870.

Zhang X, Huang L, Wu T, Feng Y, Ding Y, Ye P, Yin Z. 2015. Transcriptomic Analysis of Ovaries from Pigs with High And Low Litter Size. PloS one. 10(10):e0139514. doi:10.1371/journal.pone.0139514.

Endale Ahanda ML, Fritz ER, Estelle J, Hu ZL, Madsen O, Groenen MA, Beraldi D, Kapetanovic R, Hume DA, Rowland RR, Lunney JK, Rogel-Gaillard C, Reecy JM, Giuffra E. 2012. Prediction of altered 3'- UTR miRNA-binding sites from RNA-Seq data: the swine leukocyte antigen complex (SLA) as a model region. PloS one. 7(11):e48607. doi:10.1371/journal.pone.0048607.

Liu H, Nguyen YT, Nettleton D, Dekkers JC, Tuggle CK. 2016. Post-weaning blood transcriptomic differences between Yorkshire pigs divergently selected for residual feed intake. BMC Genomics, 17:73. doi: 10.1186/s12864-016-2395-x.

Rahman KM, Camp ME, Prasad N, McNeel AK, Levy SE, Bartol FF, Bagnell CA. 2016. Age and Nursing Affect the Neonatal Porcine Uterine Transcriptome. Biol. Reprod. 2016, 94(2):46. doi:10.1095/biolreprod.115.136150.

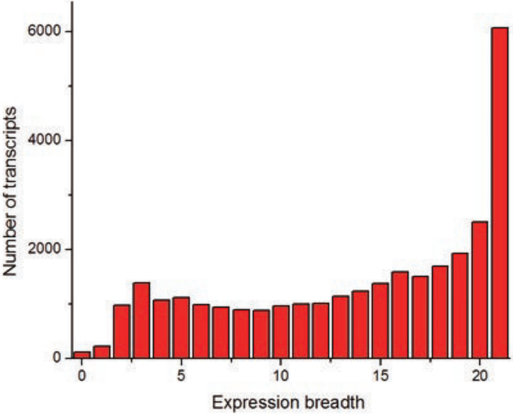
Miller LC, Bayles DO, Zanella EL, Lager KM. 2016. Effects of Pseudorabies Virus Infection on the Tracheobronchial Lymph Node Transcriptome. Bioinform. Biol. Insights. 9(Suppl 2):25-36. doi: 10.4137/BBI.S30522.

Samborski, A, Graf A, Krebs S, Kessler B, Reichenbach M, Reichenbach HD, Ulbrich SE, Bauersachs S. 2013. Transcriptome changes in the porcine endometrium during the preattachment phase. Biol. Reprod. 89(6):134. doi: 10.1095/biolreprod.113.112177.

Figure 1(on next page)

The number of tissues where a given transcript was detected.

The expression breadth (horizontal axis) denotes the number of tissues where a given transcript was detected. The zero value of the expression breadth indicates undetected transcripts.

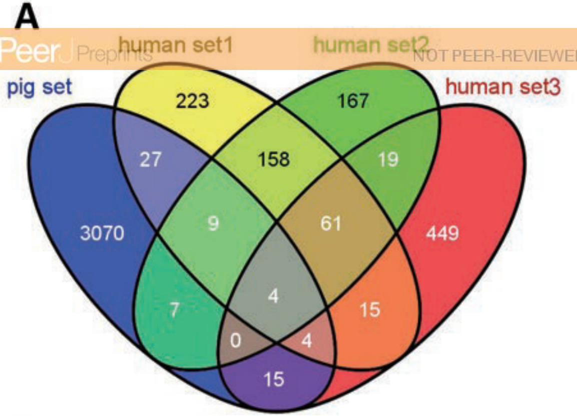


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3450v1 | CC BY 4.0 Open Access | rec: 6 Dec 2017, publ: 6 Dec 2017

Figure 2(on next page)

Overlap of housekeeping genes between pig and human.

Overlap of pig housekeeping gene set identified in the present study(A) with three human gene sets identified by microarray data (Warrington et al.2000; Hsiao et al.2001; Eisenberg and Levanon 2003) and (B)with a human set identified by RNA-seq data (Eisenberg and Levanon 2013).


A
human set1
human set2
pig set
human set3
223
167
27
158
19
3070
9
61
449
4
7
15
0
4
15

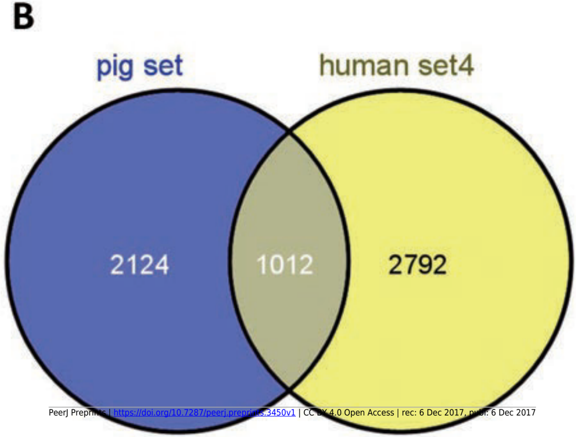

B
pig set
human set4
2124
1012
2792

Figure 3(on next page)

Comparison of length distribution of homologous housekeeping gene structures between pig and human.

nt, nucleotide(s); 5’UTR, 5’untranslated region (UTR); CDS, coding sequence.

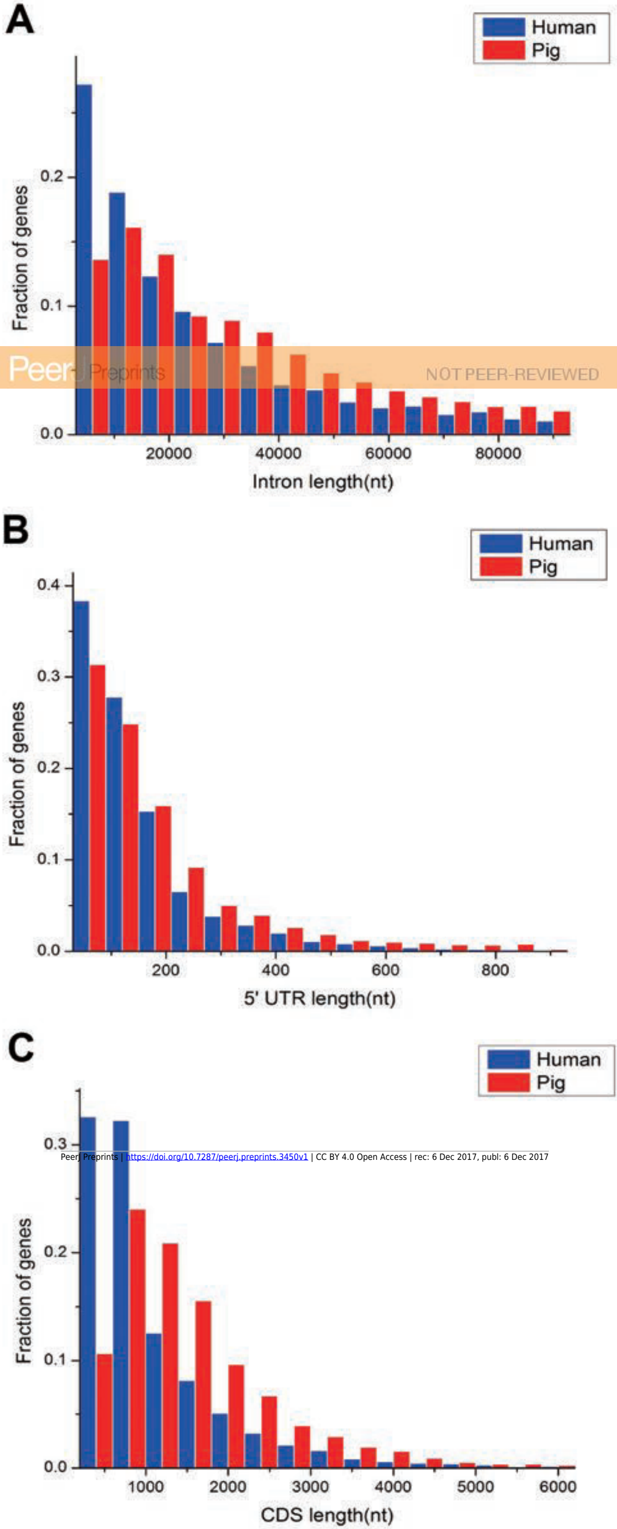

A
Human
Pig
Fraction of genes
0.0
0.1
0.2
20000
40000
60000
80000
Intron length(nt)
B
Human
Pig
Fraction of genes
0.0
0.1
0.2
0.3
0.4
200
400
600
800
5' UTR length(nt)
C
Human
Pig
Fraction of genes
0.0
0.1
0.2
0.3
1000
2000
3000
4000
5000
6000
CDS length(nt)

Figure 4(on next page)

Purifying selection on housekeeping genes.

(A) The distribution of the dN/dS ratio. (B) The dN/dS ratios of total (all HK), common (co-HK) and species-specific (sp-HK) housekeeping genes were compared between pig and human (Mann-Whitney test, * denoted *P* < 0.05), respectively.

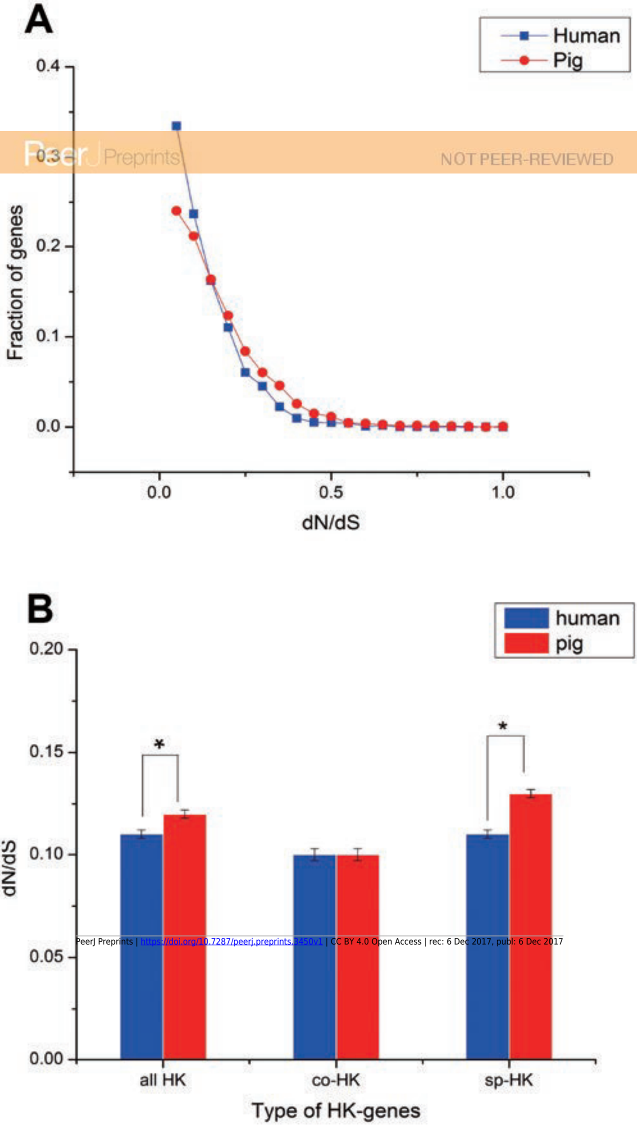

A
Human
Pig
0.4
0.3
0.2
0.1
0.0
Fraction of genes
0.0
0.5
1.0
dN/dS
B
human
pig
0.20
0.15
0.10
0.05
0.00
dN/dS
*
*
all HK
co-HK
sp-HK
Type of HK-genes

Figure 5(on next page)

Comparison of evolutionary features of housekeeping genes.

(A) The dN, dS and dN/dS of all, common and species-specific of pig housekeeping genes were compared based on the Mann-Whitney test, respectively. All such means which share a common English letter are similar; otherwise, they differ significantly at $p < 0.05$. (B) - (D) Distributions of dN, dS and dN/dS of species-specific housekeeping genes in pig and human.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3450v1 | CC BY 4.0 Open Access | rec: 6 Dec 2017, publ: 6 Dec 2017

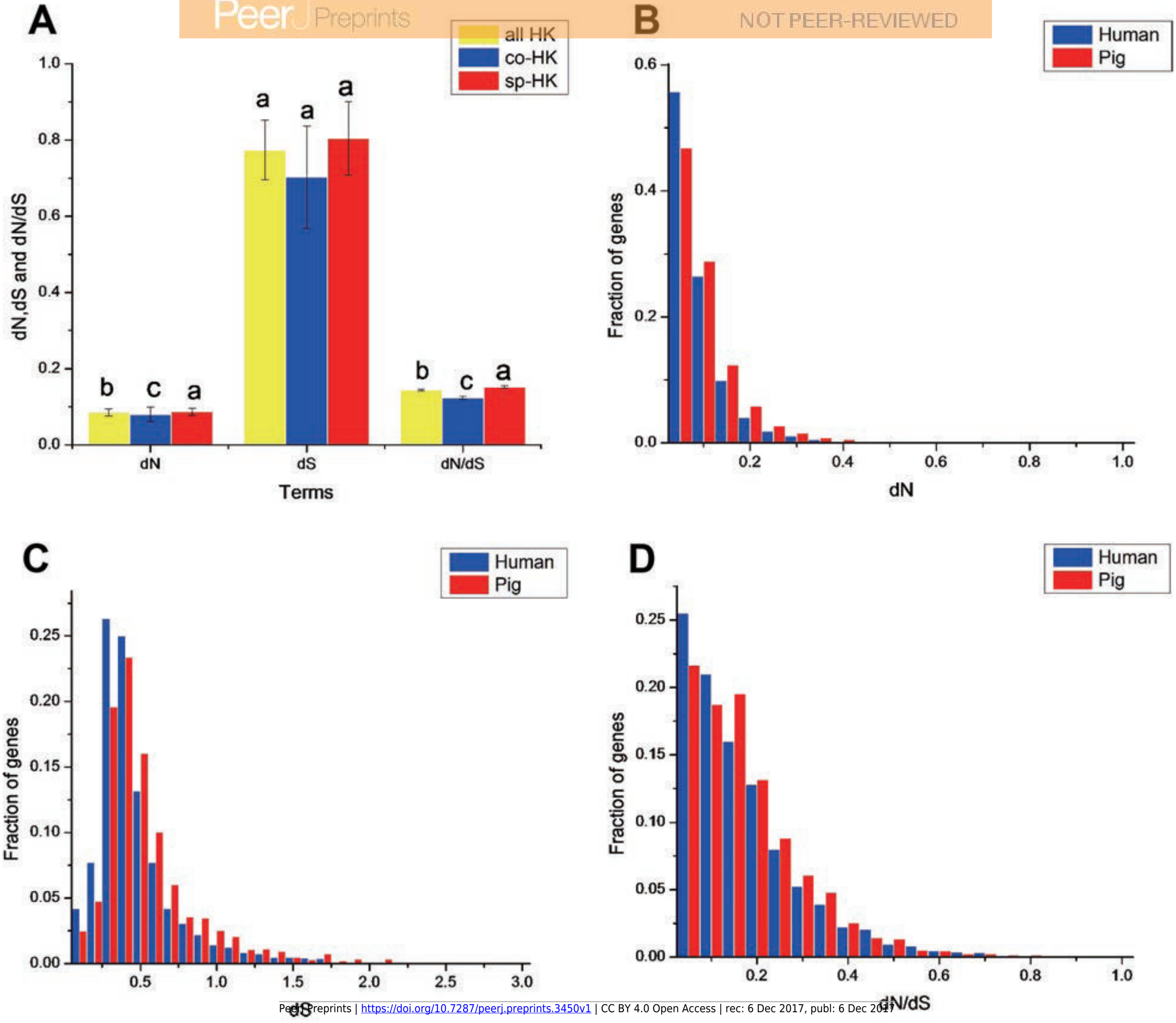

A
all HK
co-HK
sp-HK
dN,dS and dN/dS
1.0
0.8
0.6
0.4
0.2
0.0
b
c
a
a
a
a
b
c
a
dN
dS
dN/dS
Terms
B
Human
Pig
Fraction of genes
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
dN
C
Human
Pig
Fraction of genes
0.25
0.20
0.15
0.10
0.05
0.00
0.5
1.0
1.5
2.0
2.5
3.0
dS
D
Human
Pig
Fraction of genes
0.25
0.20
0.15
0.10
0.05
0.00
0.2
0.4
0.6
0.8
1.0
dN/dS

Figure 6(on next page)

Functional enrichment analysis for housekeeping genes.

Housekeeping genes were enriched in GO categories of (A)biological process, (B) cellular component, (C) molecular function, (D) molecular functions . The basal cellular function between pig and human showed high consistency. (A) (1) Biological process categories included the basal metabolism, (2) regulation of metabolic processes, (3) cellular transport, (4) cell cycle, (5)gene expression and regulation. (B) (1) Cellular component categories included organelle, (2) nuclear, (3) micromolecular complex. (C) (1) Molecular function categories included catalytic activity, (2) transcription factor activity, (3)binding activity, (4) transporter activity. (D) (1) Disease categories included tumour, (2) cancer, (3) chromosomal damage and repair, (4) other disease.

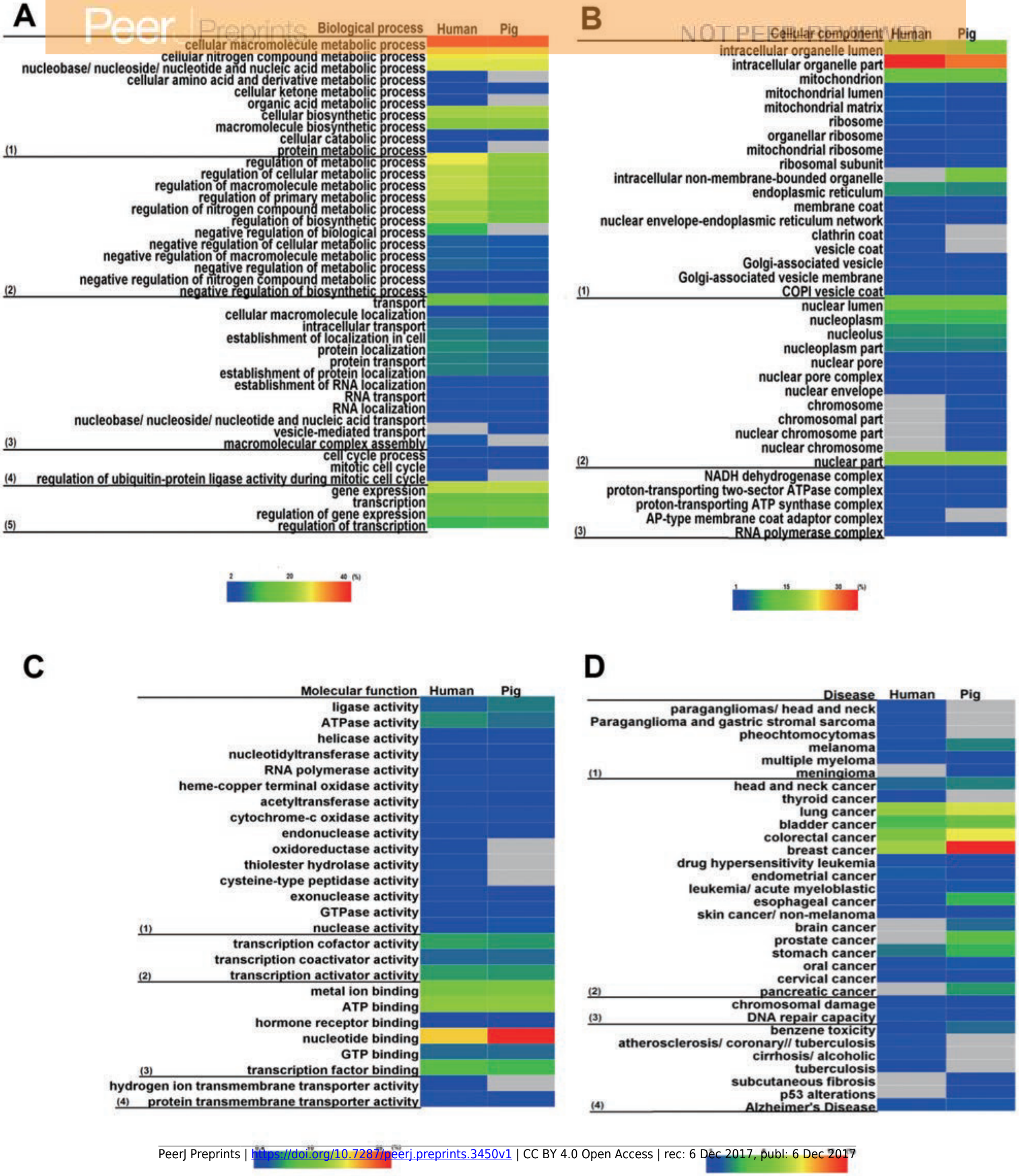

A
Biological process
Human
Pig
cellular macromolecule metabolic process
cellular nitrogen compound metabolic process
nucleobase/ nucleoside/ nucleotide and nucleic acid metabolic process
cellular amino acid and derivative metabolic process
cellular ketone metabolic process
organic acid metabolic process
cellular biosynthetic process
macromolecule biosynthetic process
cellular catabolic process
(1)
protein metabolic process
regulation of metabolic process
regulation of cellular metabolic process
regulation of macromolecule metabolic process
regulation of primary metabolic process
regulation of nitrogen compound metabolic process
regulation of biosynthetic process
negative regulation of biological process
negative regulation of cellular metabolic process
negative regulation of macromolecule metabolic process
negative regulation of metabolic process
negative regulation of nitrogen compound metabolic process
(2)
negative regulation of biosynthetic process
transport
cellular macromolecule localization
intracellular transport
establishment of localization in cell
protein localization
protein transport
establishment of protein localization
establishment of RNA localization
RNA transport
RNA localization
nucleobase/ nucleoside/ nucleotide and nucleic acid transport
vesicle-mediated transport
(3)
macromolecular complex assembly
cell cycle process
mitotic cell cycle
(4)
regulation of ubiquitin-protein ligase activity during mitotic cell cycle
gene expression
transcription
regulation of gene expression
(5)
regulation of transcription
2
20
40
(%)
B
Cellular component
Human
Pig
intracellular organelle lumen
intracellular organelle part
mitochondrion
mitochondrial lumen
mitochondrial matrix
ribosome
organellar ribosome
mitochondrial ribosome
ribosomal subunit
intracellular non-membrane-bounded organelle
endoplasmic reticulum
membrane coat
nuclear envelope-endoplasmic reticulum network
clathrin coat
vesicle coat
Golgi-associated vesicle
Golgi-associated vesicle membrane
(1)
COPI vesicle coat
nuclear lumen
nucleoplasm
nucleolus
nucleoplasm part
nuclear pore
nuclear pore complex
nuclear envelope
chromosome
chromosomal part
nuclear chromosome part
nuclear chromosome
(2)
nuclear part
NADH dehydrogenase complex
proton-transporting two-sector ATPase complex
proton-transporting ATP synthase complex
AP-type membrane coat adaptor complex
(3)
RNA polymerase complex
1
15
30
(%)
C
Molecular function
Human
Pig
ligase activity
ATPase activity
helicase activity
nucleotidyltransferase activity
RNA polymerase activity
heme-copper terminal oxidase activity
acetyltransferase activity
cytochrome-c oxidase activity
endonuclease activity
oxidoreductase activity
thiolester hydrolase activity
cysteine-type peptidase activity
exonuclease activity
GTPase activity
(1)
nuclease activity
transcription cofactor activity
transcription coactivator activity
(2)
transcription activator activity
metal ion binding
ATP binding
hormone receptor binding
nucleotide binding
GTP binding
(3)
transcription factor binding
hydrogen ion transmembrane transporter activity
(4)
protein transmembrane transporter activity
0.4
10
(%)
D
Disease
Human
Pig
paragangliomas/ head and neck
Paraganglioma and gastric stromal sarcoma
pheochtomocytomas
melanoma
multiple myeloma
(1)
meningioma
head and neck cancer
thyroid cancer
lung cancer
bladder cancer
colorectal cancer
breast cancer
drug hypersensitivity leukemia
endometrial cancer
leukemia/ acute myeloblastic
esophageal cancer
skin cancer/ non-melanoma
brain cancer
prostate cancer
stomach cancer
oral cancer
cervical cancer
(2)
pancreatic cancer
chromosomal damage
(3)
DNA repair capacity
benzene toxicity
atherosclerosis/ coronary// tuberculosis
cirrhosis/ alcoholic
tuberculosis
subcutaneous fibrosis
p53 alterations
(4)
Alzheimer's Disease

Figure 7(on next page)

Comparison of functional enrichment analysis.

When we compared functional enrichment, common housekeeping genes (co-HK) showed significant difference with species-specific housekeeping genes (sp-HK), but the sp-HKgenes between pig and human showed very high consistency. (A) (1) Biological process categories included the basal metabolism and regulation, (2) cellular transport, (3)gene expression and regulation, (4) nuclear division. (B) (1) Molecular function categories included catalytic activity, (2)transcription factor activity, (3) binding activity, (4) transporter activity.

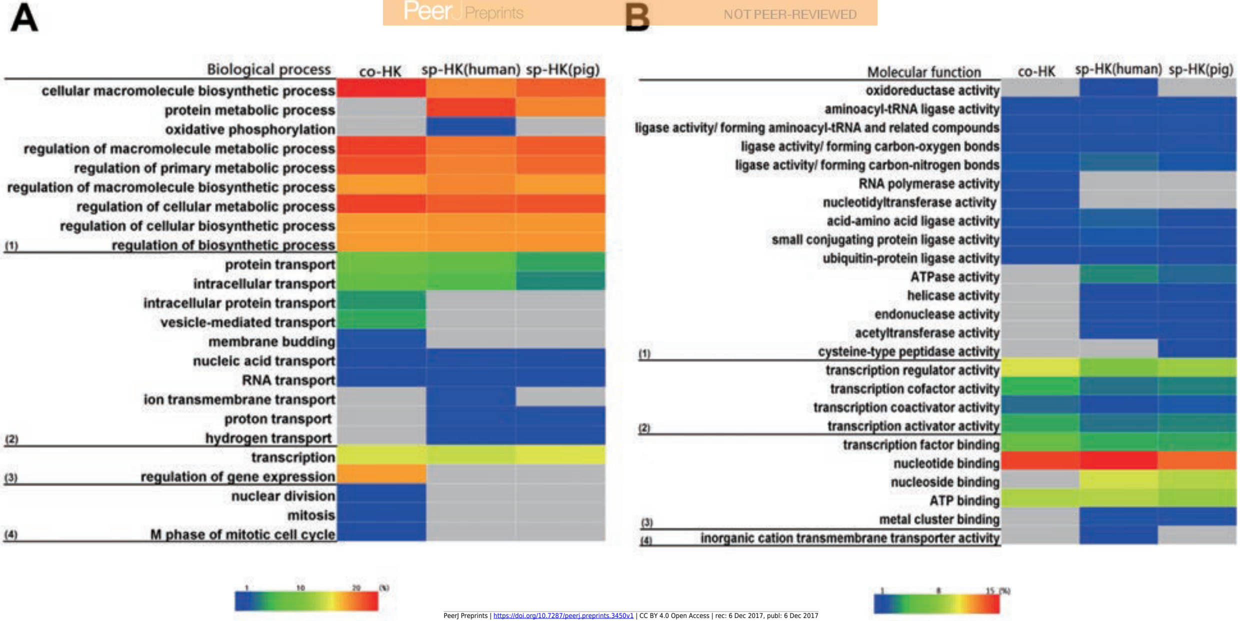

A
Biological process
co-HK
sp-HK(human)
sp-HK(pig)
cellular macromolecule biosynthetic process
protein metabolic process
oxidative phosphorylation
regulation of macromolecule metabolic process
regulation of primary metabolic process
regulation of macromolecule biosynthetic process
regulation of cellular metabolic process
regulation of cellular biosynthetic process
(1) regulation of biosynthetic process
protein transport
intracellular transport
intracellular protein transport
vesicle-mediated transport
membrane budding
nucleic acid transport
RNA transport
ion transmembrane transport
proton transport
(2) hydrogen transport
transcription
(3) regulation of gene expression
nuclear division
mitosis
(4) M phase of mitotic cell cycle
1
10
20
(%)
B
Molecular function
co-HK
sp-HK(human)
sp-HK(pig)
oxidoreductase activity
aminoacyl-tRNA ligase activity
ligase activity/ forming aminoacyl-tRNA and related compounds
ligase activity/ forming carbon-oxygen bonds
ligase activity/ forming carbon-nitrogen bonds
RNA polymerase activity
nucleotidyltransferase activity
acid-amino acid ligase activity
small conjugating protein ligase activity
ubiquitin-protein ligase activity
ATPase activity
helicase activity
endonuclease activity
acetyltransferase activity
(1) cysteine-type peptidase activity
transcription regulator activity
transcription cofactor activity
transcription coactivator activity
(2) transcription activator activity
transcription factor binding
nucleotide binding
nucleoside binding
ATP binding
(3) metal cluster binding
(4) inorganic cation transmembrane transporter activity
1
8
15
(%)

Figure 8(on next page)

Structures of the “classical” Ser/His/Asp triad configuration.

(A) Serine protease HTRA4 from pig. (B) OTU domain-containing protein 5 from human. A zoomed-in view of the catalytic domain is shown to the right of each structure. The side chains of Ser/His/Asp triad are shown in principle.

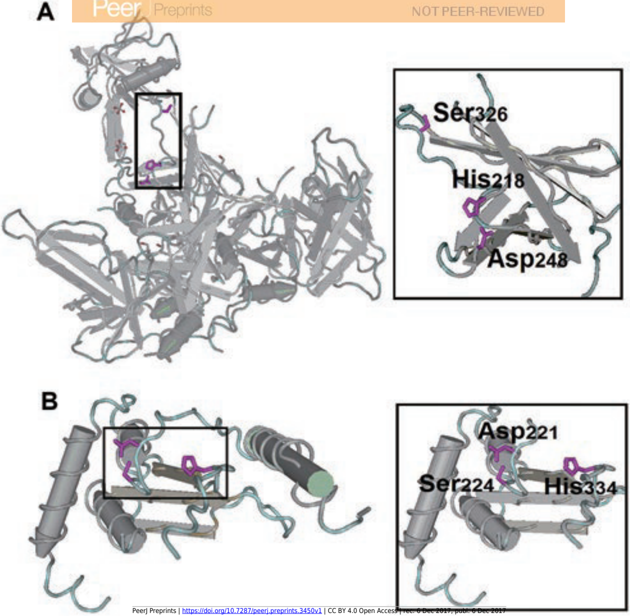

A
Ser326
His218
Asp248
B
Asp221
Ser224
His334

Figure 9(on next page)

Convergent evolution of regulatory proteins towards forming common zinc finger.

The number of zinc fingers per gene was standardized through dividing the number of each type of zinc finger by the number of proteins containing the zinc finger.

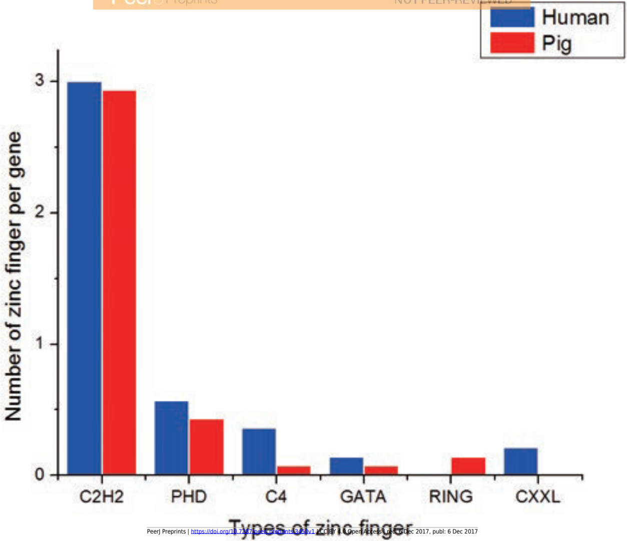

Human
Pig
Number of zinc finger per gene
0
1
2
3
C2H2
PHD
C4
GATA
RING
CXXL
Types of zinc finger

Table 1(on next page)

Comparison of housekeeping genes between pig and human

[a] The length is measured in nucleotides. [b] The value gives the average and standard error of mean. [c] The *p*-value was calculated based on the Mann-Whitney test. UTR, untranslated region; CDS, coding sequence.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3450v1 | CC BY 4.0 Open Access | rec: 6 Dec 2017, publ: 6 Dec 2017

Table 1 Comparison of housekeeping genes between pig and human

Structure	Pig	Human	*P*-value [c]
Total intron length [a]	28,108±173 [b]	21,062±297	$1.5e^{-105}$
5' UTR length	156±3	125±1.5	$3.7e^{-34}$
3' UTR length	658±13	549±5	$1.4e^{-73}$
Average exon length per gene	261±3	227±1	$1.8e^{-6}$
CDS length	2,181±10	1,460±5	$8.7e^{-234}$
Transcript length	3,312±13	2,200±5	$7.7e^{-7}$
Number of exons	9.2±0.1	8.8±0.2	$1.7e^{-4}$

[a] The length is measured in nucleotides. [b] The value gives the average and standard error of mean. [c] The *p*-value was calculated based on the Mann-Whitney test. UTR, untranslated region; CDS, coding sequence.

Table 2(on next page)

Active site of convergently related peptidases.

[a] the number following amino acid represents the position of the amino acid in protein.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3450v1 | CC BY 4.0 Open Access | rec: 6 Dec 2017, publ: 6 Dec 2017

Table 2 Active site of convergently related peptidases

Species	Gene	Protein	Nucleophile [a]	General base	Other active site residues
Pig	BLMH	Bleomycin hydrolase	Cys73	His372	Asn396
	AFG3L2	AFG3-like protein 2	Glu575	His574	Asp649
	HTRA4	Serine protease HTRA4	Ser326	His218,	Asp248
	CAPN7	Calpain-7	Cys290	His458	Asn478
Human	OTUD5	OTU domain-containing protein 5	Ser224	His334	Asp221
	SENP6	Sentrin-specific protease 6	Cys1030	His765	Asp917
	USP14	Ubiquitin carboxyl-terminal hudrolase 14	Cys114	His435	
	LONP1	Lon protease homolog, mitochondrial	Ser855	Lys898	

[a] the number following amino acid represents the position of the amino acid in protein.