A peer-reviewed version of this preprint was published in PeerJ on 24 May 2018.

View the peer-reviewed version (peerj.com/articles/4840), which is the preferred citable publication unless you specifically need to cite this preprint.

https://doi.org/10.7717/peerj.4840
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.
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; Bellora et 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. \textit{ACTB} and \textit{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 resource for humans, the pig (\textit{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 Scrofa 10.2) was downloaded from the Ensembl Genome Browser (Sep, 2016) (Kinsella et 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 material 1: 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 Scrofa 10.2) 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 (Trapnell 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.
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.
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.
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.

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)
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, C₂H₂ 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 \textit{GAPDH} and \textit{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 trend 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 \textit{GTF2H1} (general transcription factor IIH subunit 1) and \textit{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


**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.
**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).
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.
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.
**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.
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.
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-HK genes 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.
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.
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.
Table 1 (on next page)

Comparison of housekeeping genes between pig and human

- The length is measured in nucleotides. 
- The value gives the average and standard error of mean. 
- The p-value was calculated based on the Mann-Whitney test. UTR, untranslated region; CDS, coding sequence.
## Table 1 Comparison of housekeeping genes between pig and human

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

$^a$ 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.

* the number following amino acid represents the position of the amino acid in protein.
### Table 2 Active site of convergently related peptidases

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

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