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

Kai Wei*, Tingting Zhang* and Lei Ma

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

* These authors contributed equally to this work.

ABSTRACT

Housekeeping genes are ubiquitously expressed and maintain basic cellular functions across tissue/cell type conditions. The present study aimed to develop a set of pig housekeeping genes and compare the structure, evolution and function of housekeeping genes in the human–pig lineage. By using RNA sequencing data, we identified 3,136 pig housekeeping genes. Compared with human housekeeping genes, we found that pig housekeeping genes were longer and subjected to slightly weaker purifying selection pressure and faster neutral evolution. Common housekeeping genes, shared by the two species, achieve stronger purifying selection than species-specific genes. However, pig- 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 the classical catalytic serine–histidine–aspartate triad, implying that they have converged for maintaining the basic cellular function, which allows them to adapt to the environment. Human and pig housekeeping genes have varied structures and gene lists, but they have converged to maintain basic cellular functions essential for the existence of a cell, regardless of its specific role in the species. The results of our study shed light on the evolutionary dynamics of housekeeping genes.

Subjects Bioinformatics, Evolutionary Studies, Genetics

Keywords Housekeeping genes, Basal cellular function, Convergent evolution, Gene structure, Human–pig lineage

Submitted 4 December 2017
Accepted 3 May 2018
Published 24 May 2018

Corresponding author
Lei Ma, malei1979@hotmail.com

Academic editor
Hossein Khiabanian

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.4840**

BACKGROUND

Housekeeping genes are typically genes that are consistently expressed across tissues and developmental stages for maintaining basic cellular functions, including basic metabolism, cellular transport and cell cycle (*Butte, Dzau & Glueck, 2001*; *Zhu et al., 2008a*). They have unique genomic features. For example, housekeeping genes have shorter structures (including the intron, coding sequence(CDS) and exon) compared with other genes (*Eisenberg & Levanon, 2003*; *Vinogradov, 2004*), their nucleotide composition is slightly richer in GC than that of tissue-specific genes (*Vinogradov, 2003*), and they have a reduced upstream sequence conservation (*Farré et al., 2007*; *Bellora, Farré & Albà, 2007*). Housekeeping genes are often considered as the minimal gene set needed for normal cellular physiology (*Butte, Dzau & Glueck, 2001*) and are widely used as internal controls

for gene expression experiments and computational biology studies (*Thellin et al., 1999*; *Robinson & Oshlack, 2010*; *Rubie et al., 2005*; *Vandesompele et al., 2002*).

In previous studies, many human housekeeping gene sets have been identified. However, some sets slightly overlap. For example, only 155 genes were shared by three lists of microarray-defined housekeeping genes, including 501, 425 and 567 genes (*Warrington et al., 2000*; *Hsiao et al., 2001*; *Eisenberg & Levanon, 2003*). The low overlap may be explained by several reasons. Firstly, their complex transcriptional organisation may cause diverse definitions of housekeeping genes (*Gingeras, 2007*). Secondly, the expression of some genes may vary depending on experimental conditions (*Greer et al., 2010*). Why these genes vary across conditions needs further investigations. Thirdly, traditional techniques have their own drawbacks. For instance, microarray technology has a 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 (3,804) has increased previous estimates based on microarray data (567) by sixfold (*Eisenberg & Levanon, 2013*). With advances in technology, large-scale RNA-seq has provided new insights into the definition of housekeeping genes. Some studies have suggested that transcripts should be used as housekeeping units, and all transcripts of a gene need to satisfy the criteria (*Gingeras, 2007*; *Gerstein et al., 2007*).

There is no consistent definition of human housekeeping genes. However, studying the genes of animals may be able to provide new information for housekeeping genes. Therefore, a comparative analysis of housekeeping genes between humans and other animals is of great interest. Human housekeeping genes are commonly used as control genes in 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 & Rapacz, 2013*). Therefore, to well define a housekeeping gene set in another animal may be valuable. More importantly, housekeeping genes show very strict conservation in the evolutionary process, so the comparison of evolutionary dynamics will allow a fundamental understanding of evolutionary biology.

As an important meat resource for humans, the pig (*Sus scrofa*) is a well-studied organism. Given the anatomical similarities with humans, pigs are often used as a biomedical model in research (*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 of the basal mechanisms that maintain cell function. In the present study, we identified housekeeping genes in pig using RNA-seq data and then compared their structure and function with human housekeeping genes. In addition, we discussed the impact of selection pressure and convergent evolution on the functional conservation of housekeeping genes. The present study provided detailed information on pig housekeeping genes and their functional features and offered insights into their evolutionary dynamics.

MATERIALS AND METHODS

Data preparation

To define housekeeping gene sets, gene expression datasets were downloaded from the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI, September 2016) (*Kodama, Shumway & Leinonen, 2012*). In addition, pig genomic annotation (*Sus Sscrofa* 10.2) was downloaded from the Ensembl Genome Browser (September 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 (Table S1). The SRA files were downloaded from NCBI and then converted to fastq files by using fastq-dump (*Kodama, Shumway & Leinonen, 2012*). RNA-seq reads were then filtered by IlluQC.pl (*Patel & Jain, 2012*) whilst requiring an average read quality above 20. Then, the reads were aligned to a pig genome sequence (Sus Sscrofa10.2) using TopHat (*Trapnell, Pachter & Salzberg, 2009*; *Külahoglu & Bräutigam, 2014*; *Ghosh & Chan, 2016*). The alignments were then fed to an assembler Cufflinks (*Trapnell, Pachter & Salzberg, 2009*) 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.

Definition of housekeeping genes

Housekeeping genes were defined according to the following criteria: (i) the transcripts could be detected in all 21 tissues (6,072 transcripts); (ii) the transcripts showed low expression variance across tissues: $P > 0.1$ (4,068 transcripts; 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 (3,914 transcripts); and (iv) all transcripts of a housekeeping candidate gene met the above criteria (3,136 genes).

Structure analysis

The structure data of genes were obtained from the Ensembl BioMart (*Kinsella et al., 2011*). Human housekeeping genes were derived from the reference (*Eisenberg & Levanon, 2013*), considering their similar type of data from RNA-seq and stringency of the definition by expression breadth and stability. A total of 3,136 and 3,804 housekeeping genes of pigs and humans were obtained, respectively. The length of various parts of housekeeping genes were compared by Mann–Whitney test (Table 1). In addition, the length of various parts of 3,000 non-housekeeping genes were also compared by random selection in humans and pigs.

Gene Ontology (GO) analysis

The analysis of functional annotations of housekeeping genes was performed using DAVID, ver. 6.7, available on their website (*Huang da, Sherman & Lempicki, 2009a*; *Huang da, Sherman & Lempicki, 2009b*). All expressed genes in the data were used as background. Comparative analysis of housekeeping and non-housekeeping genes between humans and

Table 1 Comparison of housekeeping and non-housekeeping genes between pigs and humans.

Structure	Housekeeping gene			Non-housekeeping gene		
	Pigs	Humans	*P*-value[c]	Pigs	Humans	*P*-value
Total intron length[a]	28,108 ± 173[b]	21,062 ± 297	$1.5e^{-105}$	5,9318 ± 523	47,216 ± 487	$2.7e^{-56}$
5′ UTR length	156 ± 3	125 ± 1.5	$3.7e^{-34}$	207 ± 4.5	234 ± 4.1	$1.6e^{-29}$
3′ UTR length	658 ± 13	549 ± 5	$1.4e^{-73}$	958 ± 7.3	558 ± 4.2	$7.3e^{-65}$
Average exon length per gene	261 ± 3	227 ± 1	$1.8e^{-6}$	249 ± 2.7	265 ± 2.9	$2.4e^{-33}$
CDS length	2,181 ± 10	1,460 ± 5	$8.7e^{-234}$	3,047 ± 11.4	3,124 ± 10.8	$3.1e^{-17}$
Transcript length	3,312 ± 13	2,200 ± 5	$7.7e^{-7}$	4,021 ± 17.1	3,841 ± 14.3	$8.6e^{-94}$
Number of exons	9.2 ± 0.1	8.8 ± 0.2	$1.7e^{-4}$	15.2 ± 0.2	13.6 ± 0.1	$4.2e^{-4}$

Notes.
[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.

pigs 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 a significant level of correlation. Heat map analysis was also conducted through DAVID to visualise a matrix of enriched GO.

Evolutionary features analysis

Evolutionary features of housekeeping and non-housekeeping genes between humans and pigs were compared by calculating the substitution ratio. 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, Stecher & Tamura, 2016*; *Nei & Kumar, 2000*). From the Scope row, select the Overall Average option. For the Gaps/Missing data treatment option, select Pairwise Deletion. The genome sequences of orthologous genes were downloaded from Ensembl BioMart. The dN/dS ratios were calculated to assess the selection pressure (*Hurst, 2002*; *Yang & Nielsen, 2002*; *Dasmeh et al., 2014*). Information of active sites of proteins was obtained from UniProt Knowledgebase (*Boutet et al., 2016*; *Pundir et al., 2015*). Species-specific housekeeping genes that have similar functions were processed to search for their active sites.

RESULTS

Gene expression profile

To identify the housekeeping genes in pigs, we surveyed the expression distribution of 30,585 transcripts across 21 tissues of pigs (see Methods, Fig. 1 and Fig. S1). The detectability of RNA-seq data was high, and only 116 transcripts were undetected in the present study. The 226 transcripts showed tissue-specific expression (expressed in one tissue), whereas 6,072 transcripts were found to be broadly expressed in all 21 tissues (Fig. 1). This finding was consistent with the expression tissue breadth of human genes (*Zhu et al., 2008a*; *Zhu et al., 2008b*; *Eisenberg & Levanon, 2013*).

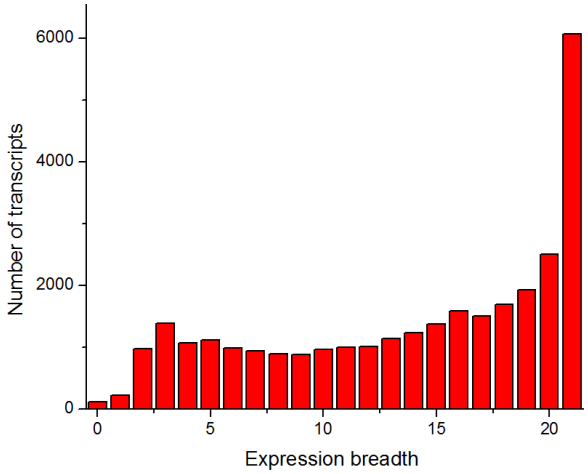

Figure 1 Number of tissues where a given transcript was detected. The expression breadth (horizontal axis) denotes the number of tissues where a given transcript was detected. The zero value of the expression breadth indicates undetected transcripts.

Full-size DOI: 10.7717/peerj.4840/fig-1

Identification of pig housekeeping genes

To obtain the transcripts with ubiquitous expression level across pig tissues, we selected 6,072 transcripts detected in 21 tissues as candidates. The background differences between different sequencing projects resulted in a batch effect between samples, including the difference in sequencing depth and coverage. Therefore, we chose a single sequencing project to assess the uniformity of gene expression. Furthermore, the expression uniformity of candidates in the ERP002055 sequencing project was evaluated using the Kolmogorov–Smirnov test and was accessed using the P-value (*Farajzadeh et al., 2013*). Figure S2 shows the frequencies of candidates with P-value greater than the given cutoff. Approximately 67% of all candidates had P-values greater than 0.1, implying that their expression levels did not significantly vary across tissues and had a high level of 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 4,068 unique transcripts, belonging to 3,754 genes. The housekeeping gene was further restricted into the gene whose transcripts passed the criteria. Altogether, 3,136 genes passed the restriction (File S1), approximately a third of which were unannotated, and 356 genes in pigs possess no orthologues in humans. In addition, housekeeping genes showed a significantly lower number of transcripts (1.22 transcripts on average) compared with whole genes in pig (1.84 transcripts on average) (Mann–Whitney test, $P < 0.05$). Housekeeping genes are always stably expressed in any tissue and environmental condition, but non-housekeeping genes, especially tissue-specific genes, may adjust to different conditions by different transcript isoforms.

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 & Levanon, 2003*; *Eisenberg & Levanon, 2013*). In addition, a lower overlap rate of housekeeping genes between pigs and humans was observed and showed significant difference with any two random sets of genes from pigs and humans (T test, $P < 0.01$).

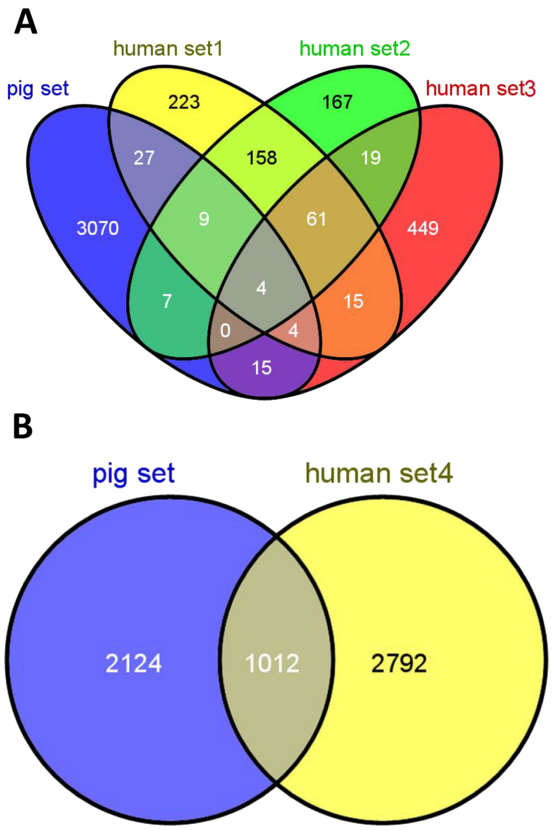

Figure 2 Overlap of housekeeping genes between pigs and humans. Overlap of pig housekeeping gene set identified in the present study (A) with three human gene sets identified by microarray data (*Warrington et al., 2000*; *Hsiao et al., 2001*; *Eisenberg & Levanon, 2003*) and (B) with a human set identified by RNA-seq data (*Eisenberg & Levanon, 2013*).

Full-size DOI: 10.7717/peerj.4840/fig-2

To accurately describe the features, housekeeping genes were grouped into three sets of genes, namely, common housekeeping genes observed in pigs and humans, human-specific housekeeping genes and pig-specific housekeeping genes. We obtained 1,012 common, 2,792 human-specific and 2,124 pig-specific housekeeping genes (Fig. 2B).

Structural comparison of housekeeping genes between pigs and humans

The comparison of length distribution of total intron, 5′ untranslated region (UTR) and CDS in homologous housekeeping genes shows that pig genes have a long length, whereas human genes have a short length (Figs. 3A–3C). Furthermore, Table 1 shows the average lengths of various structures of the housekeeping and non-housekeeping genes that correspond to one another in pigs and humans. All structures of pig housekeeping genes were significantly longer than human housekeeping genes (Table 1), indicating that human housekeeping genes hold a greater impact of gene structure, which were consistent with the previous analyses of pig genomes (*Groenen et al., 2012*). This finding implied that different purifying selection pressures were applied between pigs and humans, showing that selective

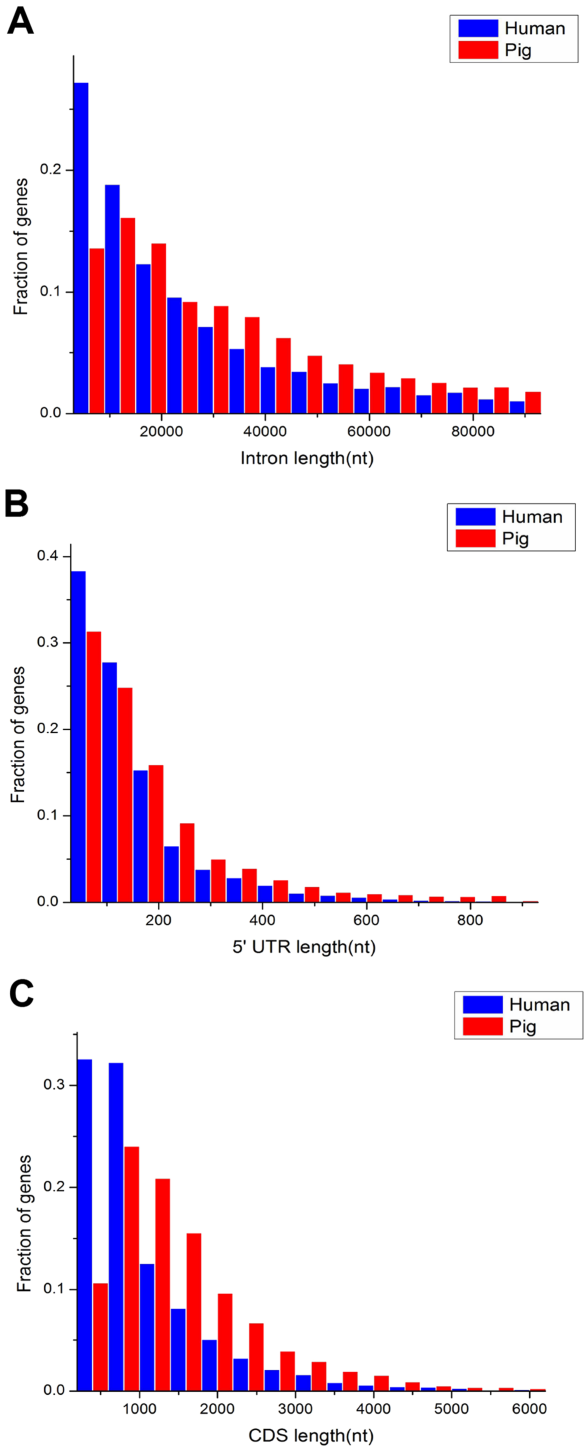

Figure 3 **Comparison of length distribution of housekeeping gene structures between pig and human.** nt, nucleotide(s); 5′ UTR, 5′ untranslated region (UTR); CDS, coding sequence.

Full-size DOI: 10.7717/peerj.4840/fig-3

Table 2 Evolutionary features of housekeeping genes.

Terms	Mouse[a]			Elephant		
	Pigs	Humans	*P*-value[c]	Pigs	Humans	*P*-value
dN	0.084 ± 0.012[b]	0.065 ± 0.01	0.003	0.085 ± 0.010	0.058 ± 0.009	0.001
dS	0.82 ± 5.57	0.70 ± 5.12	0.001	0.76 ± 6.32	0.63 ± 4.67	0.001
dN/dS	0.12 ± 0.011	0.10 ± 0.014	0.004	0.14 ± 0.023	0.11 ± 0.020	0.003

Notes.

[a]Mouse and elephant are outgroups.

[b]The value gives the average and standard error of mean.

[c]The *P*-value was calculated based on the Mann-Whitney test.

pressure may render genes as short as possible for reducing the cost in the transcription process (*Ucker & Yamamoto, 1984*; *Castillo-Davis et al., 2002*). Although the structural length of non-housekeeping genes showed a significant difference, non-housekeeping genes do not show consistent structural features unlike housekeeping genes. For example, the total intron length, 3′ UTR length and transcript length are longer in pigs than in humans, but the 5′ UTR length, average exon length and CDS length are shorter in pigs than in humans (Table 1).

Evolutionary dynamics of housekeeping genes

Evolutionary features of housekeeping genes may provide a deeper understanding of the evolutionary trend of housekeeping genes in different species. For the maintenance of essential function, housekeeping genes are thought to evolve more slowly than other genes (*Zhang & Li, 2004*). To investigate this 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 an outgroup (Files S2 and S3). In addition, the phylogeny of the mouse is close to pigs and may even be closer to humans (*Meredith et al., 2011*). Thus, we also selected elephant (*Loxodonta africana*) as an outgroup to calculate for dN, dS and dN/dS for pig and human housekeeping genes (Files S4 and S5). Generally, synonymous substitutions occur randomly, which may not or slightly suffer from selection pressure and do not appear to change the gene function, but non-synonymous substitutions do not occur randomly, which may be caused by strong selection pressure and change the function of housekeeping genes (*Nei & Kumar, 2000*; *Kimura, 1983*).

In evolutionary analysis, the housekeeping genes between pigs and humans showed significant difference with mouse and elephant as outgroups (Table 2). However, statistical differences were only observed in the dS of non-housekeeping genes between pigs and humans with mouse and elephant as outgroups (Table S2). The selection pressure of non-housekeeping genes between pigs and humans did not show a significant difference. This result may indicate that housekeeping genes show a specific evolutionary feature related to non-housekeeping genes.

The dN followed a power law distribution similar to that of dN/dS with mouse and elephant as outgroups (Fig. 4A, Figs. S3A, S4A and S5A), displaying a relatively large number of genes with a few non-synonymous substitutions and a small fraction of genes

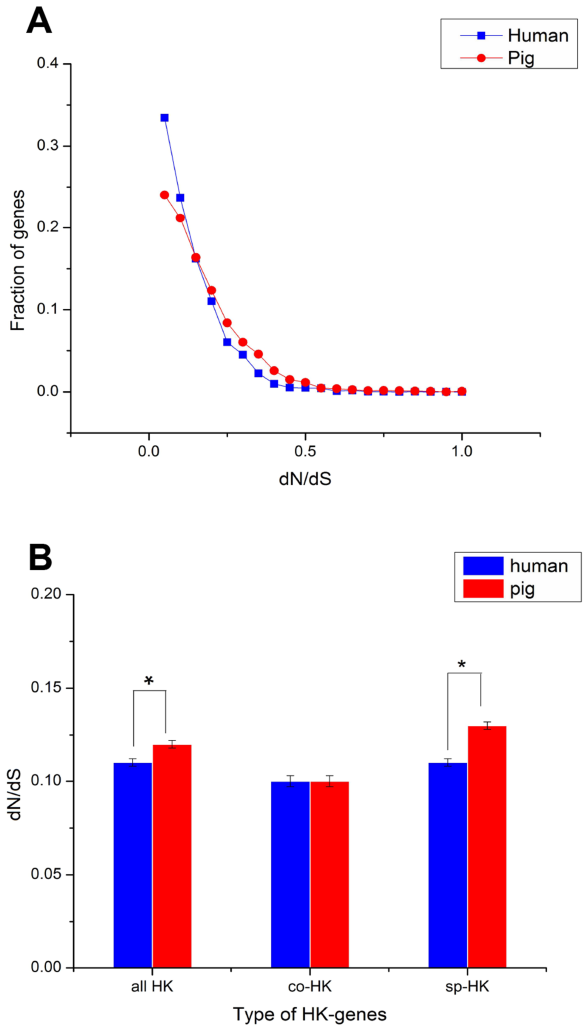

Figure 4 Purifying selection on housekeeping genes with mouse as outgroup. (A) The distribution of the dN/dS ratio. (B) The dN/dS ratios of total (all HK), common (co-HK) and species-specific (sp-HK) housekeeping genes were compared between pig and human (Mann–Whitney test, * denoted $P < 0.05$), respectively.

Full-size DOI: 10.7717/peerj.4840/fig-4

with several substitutions (Fig. 4A and Fig. S4A). In addition, most dN/dS ratios were lower than 1, implying that purifying selection acted on the housekeeping genes to ensure the stability of most genes' functions. The lesser the dN/dS ratio, the stronger the purifying selection. Furthermore, the purifying selection pressure on housekeeping genes was slightly stronger in humans than in pigs (Fig. 4 and Fig. S4).

Although mouse as outgroup showed similar results with elephant as outgroup, but with a lower difference when mouse and elephant is used as the group, respectively (Mann–Whitney test, $P < 0.05$). This result might be caused by the close phylogenetic relationship of mouse and humans (91 Myr ago) compared with pigs (97 Myr ago) and the long phylogenetic time of humans and pigs compared with elephant. Thus, a small difference was obtained when elephant was used as outgroup.

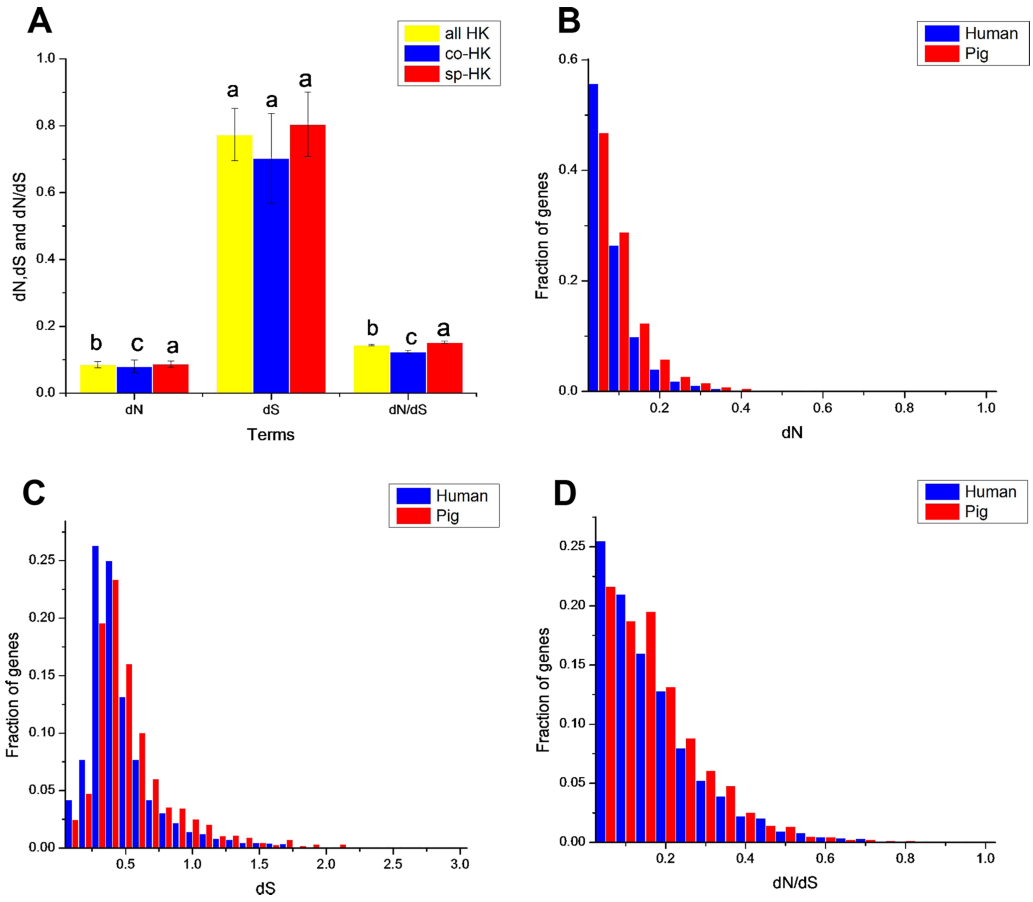

Figure 5 Comparison of evolutionary features of housekeeping genes with mouse as outgroup. (A) dN, dS and dN/dS of all, common and species-specific pig housekeeping genes were compared based on the Mann–Whitney test. In a signal cluster, all such means that share a common English letter are similar; otherwise, they differ significantly at $P < 0.05$. (B)–(D) Distributions of dN, dS and dN/dS of species-specific housekeeping genes in pigs and humans.

Full-size DOI: 10.7717/peerj.4840/fig-5

The dN/dS ratios of common housekeeping genes showed no difference between pigs and humans, but the ratios of species-specific housekeeping genes were significantly lower in humans than in pigs (Mann–Whitney test, $P < 0.05$) (Fig. 4B and Fig. S4B). Furthermore, for both humans and pigs, the dN/dS ratios of common genes were significantly lower than those of species-specific genes (Fig. 5A, Figs. S6 and S7). This result suggested that common housekeeping genes suffered a more stringent purifying selection to remove alleles than species-specific genes.

Moreover, the results of the dN/dS ratios (or dN) also implied that human housekeeping genes have evolved more stable than pig housekeeping genes because the substitution ratio was significantly lower in humans than in pigs (Table 2 and Figs. 5B–5D). This result may indicate that pig housekeeping genes may have wider evolutionary potential than human housekeeping genes. The dS of human species-specific genes had lower values than that of pig genes (Fig. 5C), showing that human housekeeping genes undergo a slower neutral evolution than pig housekeeping genes.

The dS followed an approximately normal distribution (Figs. S3B and S5B), which occurred around a central value (0.77 and 0.63 in pig and human housekeeping genes with mouse as outgroup, respectively). This finding implies the random tendency of synonymous substitutions. No significant difference was noted in the synonymous substitutions between common and species-specific genes within a species (Fig. 5A, Figs. S6 and S7).

Associated function of housekeeping genes

We then characterised the housekeeping genes that enriched the molecular function, biological process, cellular component and disease based on DAVID. The heat map shown in Fig. 6 illustrates the similar enrichment of housekeeping genes between pigs and humans. Briefly, housekeeping genes were predominantly detected as genes associated with GO terms related to basal metabolism that are indispensable for cellular physiology, indicating that housekeeping genes are essential for basic physiological processes (Fig. 6). However, the non-housekeeping genes are mainly associated with the differentiation, development and specific functions of specific tissues or organs (Table S3). This finding shows that humans and pigs have similar basic cellular functions. Although some differences in disease enrichment were noted, many common diseases were found between humans and pigs.

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

Functional convergence

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

The above analysis revealed that the functions of pig and human housekeeping genes were consistent, implying that the 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. The heat map shown in Fig. 7 illustrates the higher 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.

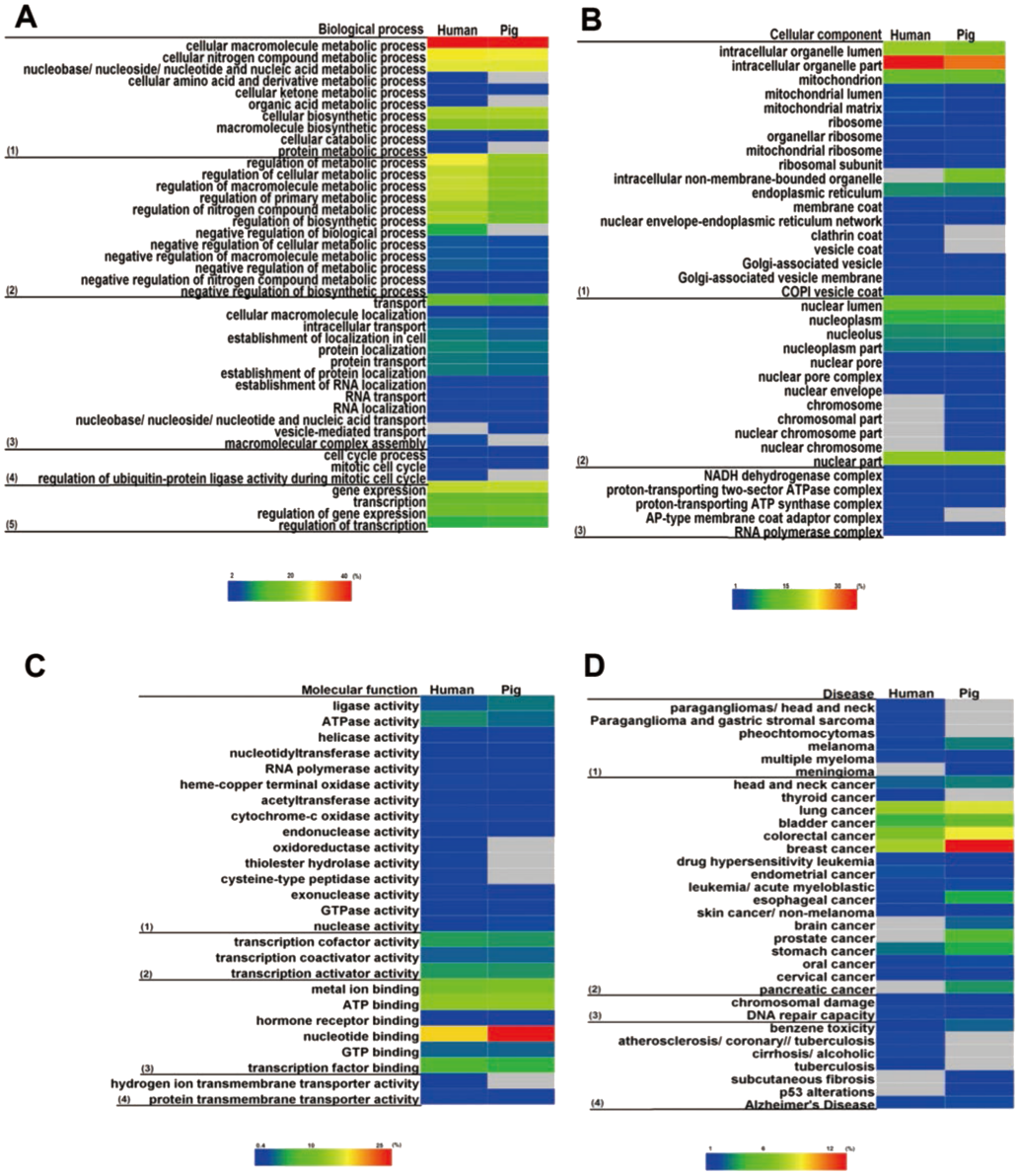

Figure 6 **Functional enrichment analysis for housekeeping genes.** Housekeeping genes were enriched in GO categories of (A) biological process, (B) cellular component, (C) molecular function, and (D) disease. Colour bars show gene frequency from 0. The basal cellular function between pigs and humans showed high consistency. (A): (1) Biological process categories included the basal metabolism, (2) regulation of metabolic processes, (3) cellular transport, (4) cell cycle, and (5) gene expression and regulation. (B): (1) Cellular component categories included organelle, (2) nuclear, and (3) micromolecular complex. (C): (1) Molecular function categories included catalytic activity, (2) transcription factor activity, (3) binding activity, and (4) transporter activity. (D): (1) Disease categories included tumour, (2) cancer, (3) chromosomal damage and repair, and (4) other disease.

Full-size DOI: 10.7717/peerj.4840/fig-6

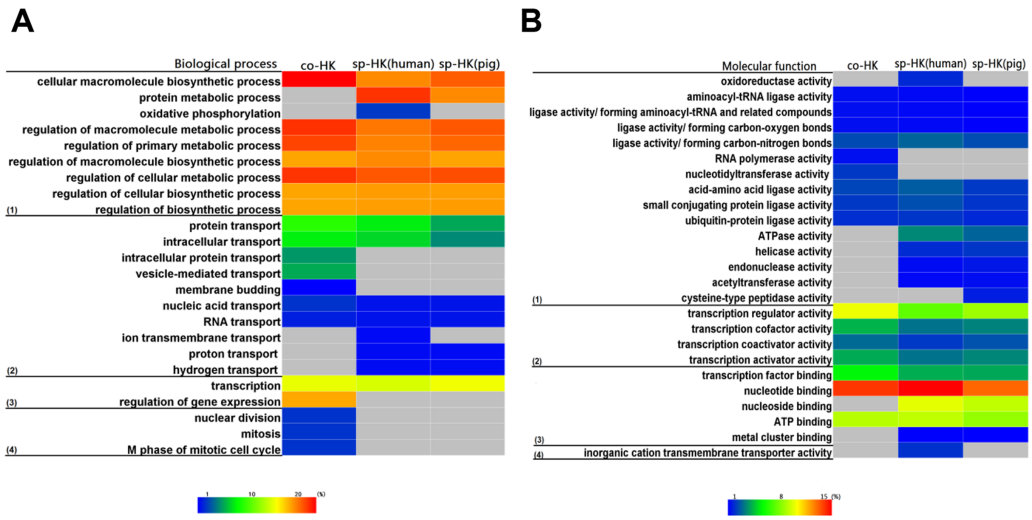

Figure 7 Comparison of functional enrichment analysis. When we compared functional enrichment, common housekeeping genes (co-HK) showed significant difference with species-specific housekeeping genes (sp-HK), but the sp-HK genes between pigs and humans showed very high consistency. Colour bars show gene frequency from 0. (A): (1) Biological process categories included the basal metabolism and regulation, (2) cellular transport, (3) gene expression and regulation, and (4) nuclear division. (B): (1) Molecular function categories included catalytic activity, (2) transcription factor activity, (3) binding activity, and (4) transporter activity.

Full-size DOI: 10.7717/peerj.4840/fig-7

Mechanistic convergence

To understand the mechanistic constraints on the function of housekeeping proteins, we analysed the evolutionary constraints on protein structure, active site feature and chemical reaction centre. We found some similar active site features in housekeeping peptidases (Fig. 8, Table 3), which reflected the intrinsic chemical constraints on enzymes, leading evolution to independently converge on equivalent solutions repeatedly (*Buller & Townsend, 2013*; *Dodson & Wlodawer, 1998*). As housekeeping genes mainly perform basic metabolic pathways of cells and peptidases are the main enzymes that perform these functions, we chose peptidases to study mechanistic convergence. The chemical and physical constraints on enzyme catalysis have caused identical triad arrangements in housekeeping peptidases in the human–pig lineage, such as classical catalytic Ser/His/Asp triad and non-classical variants (Table 3). However, the peptide sequences and their 3D structural profiles totally differed from each other (Figs. 8A and 8B). The 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 aspartate helps orient the histidine residue and neutralise the charge that develops on the histidine during transition states (*Polgar, 2005*; *Ekici, Paetzel & Dalbey, 2008*). Interestingly, almost all proteins in Table 3 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, to our knowledge, these active sites have rarely been mentioned in previous reports.

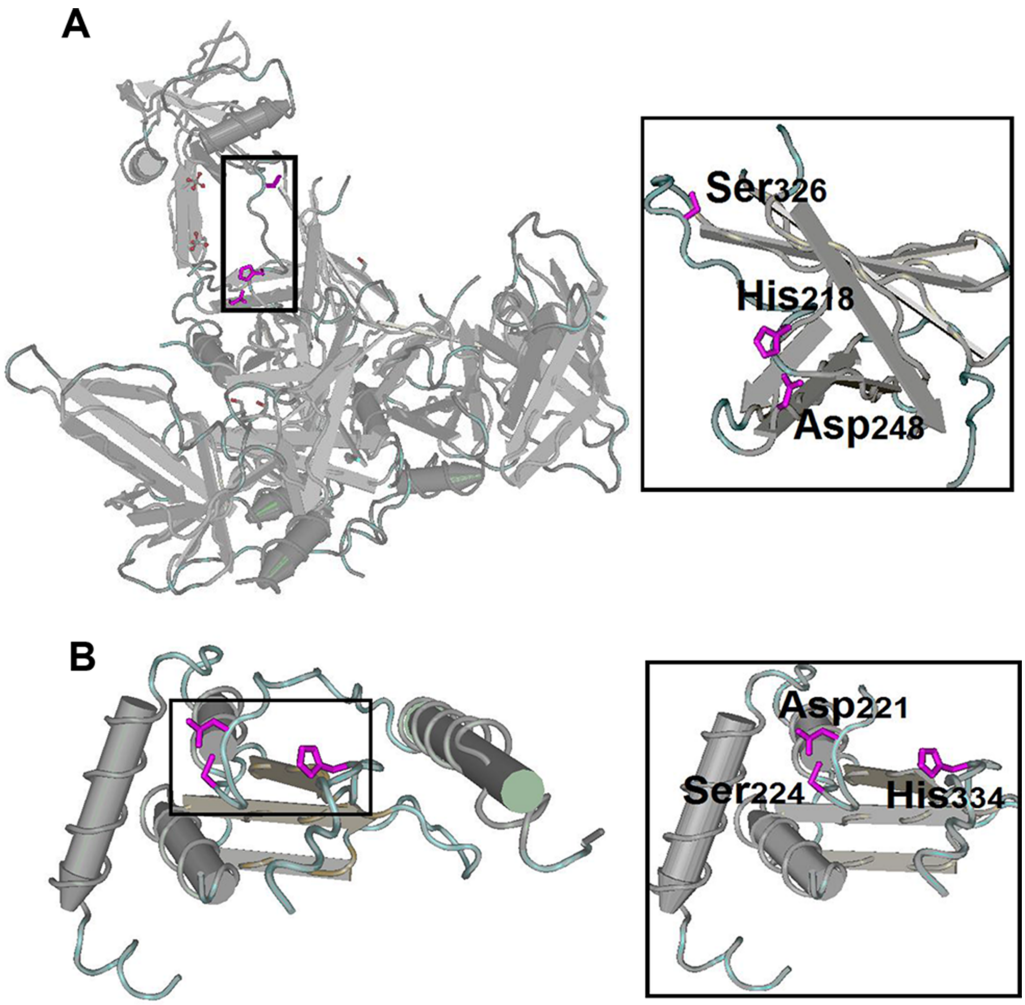

Figure 8 Structures of the 'classical' Ser/His/Asp triad configuration. (A) Serine protease HTRA4 from pigs. (B) OTU domain-containing protein 5 from humans. A zoomed-in view of the catalytic domain is shown to the right of each structure. The side chains of Ser/His/Asp triad are shown in principle.
Full-size DOI: 10.7717/peerj.4840/fig-8

The analysis of housekeeping protein structure and function may reveal several interrelated and previously unrecognised relationships of structure–function constraints. These fundamental constraints have promoted the convergent evolution of housekeeping genes. Although the relationship between mechanistic convergence and functional convergence is unclear in the present study, such finding provides an entry point for our future research.

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 pigs showed a lower overlap relative to the human set as the two sets showed similar physical structure and high homology. Some housekeeping genes, such as *GAPDH*

Table 3 Active site of convergently related peptidases.

Species	Gene	Protein	Nucleophile[a]	General base	Other active site residues
Pigs	*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
Humans	*OTUD5*	OTU domain-containing protein 5	Ser224	His334	Asp221
	SENP6	Sentrin-specific protease 6	Cys1030	His765	Asp917
	USP14	Ubiquitin carboxyl-terminal hydrolase 14	Cys114	His435	
	LONP1	Lon protease homolog, mitochondrial	Ser855	Lys898	

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

and *ACTB*, in humans were not found in our list (*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 verified.

After divergence from a common ancestor, pigs and humans have accumulated differences in the sequence and structure of housekeeping genes. On a molecular level, this phenomenon can occur from random mutation, for example, synonymous substitution. The dS distribution followed an approximately normal distribution, showing a random tendency for synonymous substitutions. Meanwhile, the divergence was also related to adaptive changes. In addition, GC content may affect the distribution of synonymous and non-synonymous substitutions. Hence, we also determined whether dN, dS and dN/dS of housekeeping genes were correlated with the GC content by using mouse as an outgroup. Our results showed that although a strong correlation was found between dS and GC content ($r = 0.48$, $P = 1.94e^{-12}$), dN ($r = -0.087$, $P = 0.013$) and dN/dS ($r = -0.11$, $P = 0.027$) only showed very weak correlations with GC content. Thus, the GC content may not be the main contributing factor to the selection pressure.

Human housekeeping genes were found to be shorter than pig housekeeping genes (Figs. 3A–3C), which facilitates gene expression (*Ucker & Yamamoto, 1984*; *Izban & Luse, 1992*). In addition, the stronger purifying selection in humans than in pigs (Fig. 4A) might result in a lower degree of genetic redundancy. A source of genetic redundancy is convergent evolutionary processes, leading to genes that are close in function but unrelated in sequence, so they may also change the length of the gene structure (*Zhang & Li, 2004*). In other words, human housekeeping genes likely evolved more stable than pig housekeeping genes because of the advantageous and stable living environment. Moreover, humans and pigs have evolved their own species-specific housekeeping genes, which may have led to the formation of the two species, allowing the differentiated fixation of characteristics. In addition, purifying selection was stronger in common than in species-specific housekeeping genes and showed some differences in GO enrichment. This result may indicate that common housekeeping genes are more indispensable than species-specific genes and serve more functions for sustaining life. For example, *GTF2H1* (general transcription factor IIH subunit 1) and *CXXC1* (CXXC finger protein 1) in common housekeeping genes are

crucial for regulating the expression of several genes (*Shiekhattar et al., 1995*; *Butler et al., 2009*), but in species-specific housekeeping genes, they were not enriched.

However, although humans and pigs have diverged for millions of years, both species independently converged towards similar features of housekeeping genes. One of the most unexpected observations was noted in species-specific housekeeping genes. GO enrichment analysis revealed that pig- and human-specific housekeeping genes serve similar functions. In addition, some housekeeping proteins evolved independently to achieve similar active sites, sidechains, catalytic centres or binding sites to complete a similar catalytic reaction or molecular function (*Buller & Townsend, 2013*; *Polgar, 2005*; *Ekici, Paetzel & Dalbey, 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 & Li, 2004*; *Kimura, 1983*), which drives different species-specific genes to perform similar essential functions, regardless of their specific roles in the species.

At present, there is still no large-scale gene expression profile. The current transcriptome sequencing data in pigs may be inadequate to meet the requirement to define housekeeping genes. The accurate definition of housekeeping genes remains an unresolved issue. Therefore, the present set of pig housekeeping genes has limitations, but its characteristics are 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 catalogue 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' should be defined in a hierarchical way related to cell types, growth stages, cell cycles and various physiological conditions and in terms of specific transcript variant (*Zhu et al., 2008a*; *Zhu et al., 2008b*). 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 for analysing the cellular basal function of different species that have some slight differences (*Kumar & Hedges, 1998*; *Meredith et al., 2011*; *Kumar & Subramanian, 2002*).

CONCLUSIONS

The present study offered insight into the general aspects of housekeeping gene structure and evolution. Diverging from the ancestor of humans and pigs, housekeeping genes vary in gene structure and gene list, but they have converged to maintain basic cellular functions essential for the existence of a cell, regardless of their specific role in the species. The results in the present study will shed light on the evolutionary dynamics of housekeeping genes.

ACKNOWLEDGEMENTS

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

ADDITIONAL INFORMATION AND DECLARATIONS

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.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 31272416, 31560310, 31370762.
National High Technology Research and Development Program of China: 2013AA102502.
Scientific Research Foundation of the MHRSS of China.
Scholar Pair-training Program of Shihezi University: SDJDZ201504.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Kai Wei and Tingting Zhang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Lei Ma conceived and designed the experiments, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are provided in the Supplemental Tables.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.4840#supplemental-information.

REFERENCES

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. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

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. *Physiological Genomics* **21(3)**:389–395 DOI 10.1152/physiolgenomics.00025.2005.

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.

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.

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 in Molecular Biology* **1374**:23–54 DOI 10.1007/978-1-4939-3167-5_2.

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.

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.

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 Molecular Biology* **11**:22 DOI 10.1186/1471-2199-11-22.

Brown RS. 2005. Zinc finger proteins: getting a grip on RNA. *Current Opinion in Structural Biology* **15(1)**:94–98 DOI 10.1016/j.sbi.2005.01.006.

Buller AR, Townsend CA. 2013. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proceedings of the National Academy of Sciences of the United States of America* **110(8)**:E653–E661 DOI 10.1073/pnas.1221050110.

Butler JS, Palam LR, Tate CM, Sanford JR, Wek RC, Skalnik DG. 2009. DNA Methyltransferase protein synthesis is reduced in CXXC finger protein 1-deficient embryonic stem cells. *DNA and Cell Biology* **28(5)**:223–231 DOI 10.1089/dna.2009.0854.

Butte AJ, Dzau VJ, Glueck SB. 2001. Further defining housekeeping, or "maintenance," genes focus on "A compendium of gene expression in normal human tissues". *Physiological Genomics* **7(2)**:95–96 DOI 10.1152/physiolgenomics.2001.7.2.95.

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

Chen L, Wang H, Zhang J, Gu L, Huang N, Zhou JM, Chai J. 2008. Structural basis for the catalytic mechanism of phosphothreonine lyase. *Nature Structural & Molecular Biology* **15(1)**:101–102 DOI 10.1038/nsmb1329.

Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. 2014. The influence of selection for protein stability on dN/dS estimations. *Genome Biology and Evolution* **6(10)**:2956–2967 DOI 10.1093/gbe/evu223.

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.

Dodson G, Wlodawer A. 1998. Catalytic triads and their relatives. *Trends in Biochemical Sciences* **23(9)**:347–352 DOI 10.1016/S0968-0004(98)01254-7.

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

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

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends in Genetics* **29(10)**:569–574 DOI 10.1016/j.tig.2013.05.010.

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

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. *Biochemical and Biophysical Research Communications* **438(2)**:346–352 DOI 10.1016/j.bbrc.2013.07.074.

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

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.

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 Research* **17(6)**:669–681 DOI 10.1101/gr.6339607.

Ghosh S, Chan KK. 2016. Analysis of RNA-seq data using TopHat and Cufflinks. *Methods in Molecular Biology* **1374**:339–361 DOI 10.1007/978-1-4939-3167-5_18.

Gingeras TR. 2007. Origin of phenotypes: genes and transcripts. *Genome Research* **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. *Journal of Immunological Methods* **355(1–2)**:76–79 DOI 10.1016/j.jim.2010.02.006.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, Li S, Larkin DM, Kim H, Frantz LA, Caccamo M, Ahn H, Aken BL, Anselmo A, Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, Bosse M, Botti S, Bujie Z, Bystrom M, Capitanu B, Carvalho-Silva D, Chardon P, Chen C, Cheng R, Choi SH, Chow W, Clark RC, Clee C, Crooijmans RP, Dawson HD, Dehais P, De Sapio F, Dibbits B, Drou N, Du ZQ, Eversole K, Fadista J, Fairley S, Faraut T, Faulkner GJ, Fowler KE, Fredholm M, Fritz E, Gilbert JG, Giuffra E, Gorodkin J, Griffin DK, Harrow JL, Hayward A, Howe K, Hu ZL, Humphray SJ, Hunt T, Hornshøj H, Jeon JT, Jern P, Jones M, Jurka J, Kanamori H, Kapetanovic R, Kim J, Kim JH, Kim KW, Kim TH, Larson G, Lee K, Lee KT, Leggett R, Lewin HA, Li Y, Liu W, Loveland JE, Lu Y, Lunney JK, Ma J, Madsen O, Mann K, Matthews L, McLaren S, Morozumi T, Murtaugh MP, Narayan J, Nguyen DT, Ni P, Oh SJ, Onteru S, Panitz F, Park EW, Park HS, Pascal G, Paudel Y, Perez-Enciso M, Ramirez-Gonzalez R, Reecy JM, Rodriguez-Zas S, Rohrer GA, Rund L, Sang Y, Schachtschneider K, Schraiber JG, Schwartz J, Scobie L, Scott C, Searle S, Servin B, Southey BR, Sperber G, Stadler P, Sweedler JV, Tafer H, Thomsen B, Wali R, Wang J, Wang J, White S, Xu X, Yerle M, Zhang G, Zhang J, Zhang J, Zhao S, Rogers J, Churcher C, Schook LB. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491(7424)**:393–398 DOI 10.1038/nature11622.

Hall TMT. 2005. Multiple modes of RNA recognition by zinc finger proteins. *Current Opinion in Structural Biology* **15(3)**:367–373 DOI 10.1016/j.sbi.2005.04.004.

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. *Physiological Genomics* **7(2)**:97–104 DOI 10.1152/physiolgenomics.00040.2001.

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

Huang da W, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4(1)**:44–57 DOI 10.1038/nprot.2008.211.

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

Hurst LD. 2009. Genetics and the understanding of selection. *Nature Reviews Genetics* **10(2)**:83–93 DOI 10.1038/nrg2506.

Izban MG, Luse DS. 1992. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *Journal of Biological Chemistry* **267(19)**:13647–13655.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**:bar030 DOI 10.1093/database/bar030.

Klug A. 1999. Zinc finger peptides for the regulation of gene expression. *Journal of Molecular Biology* **293(2)**:215–218 DOI 10.1006/jmbi.1999.3007.

Klug A. 2010. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Quarterly Review of Biophysics* **43(1)**:1–21 DOI 10.1017/S0033583510000089.

Kodama Y, Shumway M, Leinonen R. 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research* **40(Database issue)**:D54–D56 DOI 10.1093/nar/gkr854.

Kozera B, Rapacz M. 2013. Reference genes in real-time PCR. *Journal of Applied Genetics* **54(4)**:391–406 DOI 10.1007/s13353-013-0173-x.

Külahoglu C, Bräutigam A. 2014. Quantitative transcriptome analysis using RNA-seq. *Methods in Molecular Biology* **1158**:71–91 DOI 10.1007/978-1-4939-0700-7_5.

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

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33(7)**:1870–1874 DOI 10.1093/molbev/msw054.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* **99(2)**:803–808 DOI 10.1073/pnas.022629899.

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.

Lunney JK. 2007. Advances in swine biomedical model genomics. *International Journal of Biological Sciences* **3(3)**:179–184 DOI 10.7150/ijbs.3.179.

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 Research* **18(9)**:1509–1517 DOI 10.1101/gr.079558.108.

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.

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. Oxford: Oxford University Press, 52–72.

Nielsen R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics* **39**:197–218 DOI 10.1146/annurev.genet.39.073003.112420.

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 Molecular Biology* **8**:67 DOI 10.1186/1471-2199-8-67.

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.

Polgar L. 2005. The catalytic triad of serine peptidases. *Cellular and Molecular Life Science* **62(19–20)**:2161–2172 DOI 10.1007/s00018-005-5160-x.

Pundir S, Magrane M, Martin MJ, O'Donovan C. 2015. Searching and navigating UniProt databases. *Current Protocols in Bioinformatics* **50**:1.27.1–1.27.10 DOI 10.1002/0471250953.bi0127s50.

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

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.

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. *Molecular and Cellular Probes* **19(2)**:101–109 DOI 10.1016/j.mcp.2004.10.001.

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.

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. *Nature Methods* **6(5)**:377–382 DOI 10.1038/nmeth.1315.

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. *Journal of Biotechnology* **75(2–3)**:291–295 DOI 10.1016/S0168-1656(99)00163-7.

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

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.

Ucker DS, Yamamoto KR. 1984. Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. *Journal of Biological Chemistry* **259(12)**:7416–7420.

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 Biology* **3**(**7**):RESEARCH0034.1 DOI 10.1186/gb-2002-3-7-research0034.

Vinogradov AE. 2003. Isochores and tissue-specificity. *Nucleic Acids Research* **31**(**17**):5212–5220 DOI 10.1093/nar/gkg699.

Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends in Genetics* **20**(**5**):248–253 DOI 10.1016/j.tig.2004.03.006.

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 b5 relative to that of mitochondrial apocytochrome b5. *Biochemistry* **45**(**46**):13750–13759 DOI 10.1021/bi0615689.

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

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**(**6**):908–917 DOI 10.1093/oxfordjournals.molbev.a004148.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* **21**(**2**):236–239 DOI 10.1093/molbev/msh010.

Zhu J, He F, Hu S, Yu J. 2008a. On the nature of human housekeeping genes. *Trends in Genetics* **24**(**10**):481–484 DOI 10.1016/j.tig.2008.08.004.

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