Evolutionary and Molecular Characterization of liver-enriched gene 1

Yanna Dang, Jin-Yang Wang, Chen Liu, Kun Zhang, Peng Jinrong & Jin He

Liver-enriched gene 1 (Leg1) is a newly identified gene with very little available functional information. To evolutionarily and molecularly characterize Leg1 genes, a phylogenetic study was first conducted, which indicated that Leg1 is a conserved gene that exists from bacteria to mammals. During the evolution of mammals, Leg1s underwent tandem duplications, which gave rise to Leg1a, Leg1b, and Leg1c clades. Analysis of the pig genome showed the presence of all three paralogs of pig Leg1 genes (pLeg1s), whereas only Leg1a could be found in the human (hLeg1a) or mouse (mLeg1a) genomes. Purifying force acts on the evolution of Leg1 genes, likely subjecting them to functional constraint. Molecularly, pLeg1a and its coded protein, pig LEG1a (pLEG1a), displayed high similarities to its human and mouse homologs in terms of gene organization, expression patterns, and structures. Hence, pLeg1a, hLeg1a, and mLeg1a might preserve similar functions. Additionally, expression analysis of the three Leg1as suggested that eutherian Leg1as might have different functions from those of zebrafish and platypus due to subfunctionalization. Therefore, pLeg1a might provide essential information about eutherian Leg1a. Moreover, a preliminary functional study using RNA-seq suggested that pLeg1a is involved in the lipid homeostasis. In conclusion, our study provides some basic information on the aspects of evolution and molecular function, which could be applied for further validation of Leg1 using pig models.

Leg1 (liver-enriched gene 1, or C6orf58 homolog) is a newly identified gene with very little available functional information. It is characterized by the presence of Domain of Unknown Function 781 (DUF781 or LEG1 domain) in its encoded protein. Leg1 was first identified in a zebrafish (Danio rerio) microarray study, in which it was named on the basis of its abundance in the liver. Functional experiments later demonstrated that Leg1 is involved in liver development, as knock-down of Leg1 in embryos results in small liver phenotype due to blocked liver expansion. Another functional study of the Leg1 gene was performed in the platypus (Ornithorhynchus anatinus), in which the human C6orf58 paralog MLP encodes monotreme lactation protein (MLP). MLP protein is a secreted protein that is enriched in milk, where it can exert antibacterial activity. Thus, it is presumed that MLP is related to the innate immunity of monotremes during the nipple-less delivery of milk to the hatchlings.

Proteomic studies in eutherian species revealed that the Leg1-encoded N-glycosylated LEG1 protein is mainly present in saliva and seminal plasma. However, no further functional studies were carried out in eutherian animals. Expression profiling analyses of mouse (Mus musculus) and human (Homo sapiens) Leg1s (mLeg1 and hLeg1) reported in the Expression Atlas (www.ebi.ac.uk/gxa/home) showed that the gene is not expressed in the liver or mammary glands, in contrast to studies in monotremes and fishes, implying that eutherian Leg1s might have different biological functions. In addition, the preliminary evolutionary analysis revealed only one copy of Leg1 in humans and mice, whereas the majority of other mammals harboured at least two Leg1 gene copies. Two major clades, each of which consists of C6orf58 orthologs or paralogs, could be identified for mammalian Leg1s in the phylogenetic tree, indicating possible gene duplication event in mammals. Gene duplication resulted from whole genome duplication, unequal crossover, or segmental duplication is an important factor for speciation, adaptation, and gene family expansion. The duplicated paralogs, which are then subjected to evolutionary selection, could conserve their original functions, acquire novel functions (neofunctionalization), maintain a specialized subfunction (subfunctionalization), or lose gene functions (pseudogenization). Since there might have been a duplication event during the evolution of the mammalian Leg1 genes, orthologous Leg1 genes must be cloned and characterized from closely related species to provide information about hLeg1, if there are any.
is the existence of functional constraint on the genes during evolution. Therefore, our study was conducted from two perspectives: 1) evolutionarily, the identification of orthologous genes in model organisms was carried out; 2) molecularly, it was determined whether the chosen orthologous genes show similar characteristics to hLeg1.

To accomplish these two goals, we initially conducted a comprehensive phylogenetic study using all available DUF781 domain harboring proteins (LEG1/LEG1L proteins hereafter), revealing that Leg1 is a conserved gene that exists from bacteria to mammals. Moreover, during mammalian evolution, Leg1 experienced tandem duplications that eventually gave rise to the Leg1a, Leg1b, and Leg1c paralogs. These Leg1 genes are evolutionarily constrained, and in several species, Leg1b and Leg1c copies might have been pseudogenized, leaving Leg1a as the primary form of Leg1 in eutherian genomes, especially in primates.

To study the functional role of hLeg1a, model organisms other than zebrafish (Danio rerio), Caenorhabditis elegans, and Drosophila melanogaster were needed, as no Leg1 copy has been identified in the last two species. Hence, pigs (Sus scrofa) and mice might be better alternatives for elucidating the function of the hLeg1a gene. As the mLeg1a gene has been characterized as presenting only one functional copy, the construction of mLeg1a knockout mice is a straightforward and critical way to elucidate the function of Leg1 in eutherians. The pig is not only an important livestock species but is also highly similar to humans in anatomy, physiology, and metabolism, making it an attractive alternative large animal model for human diseases. Therefore, studying pLeg1 could provide new insights into both agricultural and biomedical applications in addition to its biological mechanism. In contrast to mLeg1a, pigs have three Leg1 gene copies (pLeg1a, pLeg1b, and pLeg1c). Though phylogenetic analysis shows hLeg1a, mLeg1a, and pLeg1a are orthologs, it remains to be determined which of these pig Leg1 copies is molecularly relevant to hLeg1a. In this study, the cloning and characterization of the pig Leg1 genes revealed that pLeg1a was the only one of the genes to be transcriptionally detectable. Additionally, pLeg1a has a similar expression pattern to hLeg1a and mLeg1a. Structural prediction also indicated that pLeg1a, hLeg1a, and mLeg1a are closely related. Finally, RNA-seq was performed to predict the potential function of the Leg1 gene. The results showed that overexpression of pLeg1a affected certain biological processes (e.g., lipid homeostasis) and the level of PPARγ. Therefore, through our study, we provide some basic information regarding the evolution of the Leg1 gene and demonstrate that pLeg1a is evolutionarily and molecularly close to hLeg1a, which could be applied for the further functional annotation of hLeg1a through the use of porcine models.

Materials and Methods

Construction of the phylogenetic tree. To retrieve the sequences for phylogenetic analysis, human (Homo sapiens) LEG1a (NP_001010905.1), mouse (Mus musculus) LEG1a (NP_080612.1), platypus (Ornithorhynchus anatinus) MLP (NP_00131705.1), and zebrafish (Danio rerio) LEG1s (NP_001093552.1, NP_998368.1) were used as queries to search against the non-redundant protein database using the phi-blast algorithm with iterated searches until no further significant hits were found. The obtained hits were initially screened based on an E-value < 0.005, and the redundant sequences, spliced variants, and hits with lengths that were too short were then removed. Then, the NCBI Genome, Ensembl, and UCSC Genome Browsers were used to search for additional annotated or predicted Leg1 gene loci. If there was no Leg1 information available in a species, the surrounding sequences according to synteny were subjected to GENSCAN for the prediction of potential protein-coding genes. Next, the obtained sequences were further screened for the presence of the DUF781 domain using CDD/SPARCLE. Finally, 413 sequences with characteristic DUF781 domains (409 sequences have DUF781 as their sole identifiable domain; Ochotona princeps (XP_004587370.1) and Saccoglossus kowalevskii (XP_006815645.1) LEG1s harbor predicted domains other than DUF781.) were included in subsequent studies (Supplementary spreadsheets 1 and 2). Based on the information provided by the Ensembl Gene Tree and NCBI annotations, vertebrate DUF781 domain containing proteins were named as LEG1 proteins, while those distributed as LEG1-like proteins were designated as LEG1L genes. Correspondingly, the genes were named as Leg1 and Leg1L.

To clearly illustrate Leg1 and Leg1L evolution, the following representative species were chosen: primates (Homo sapiens, Macaca mulatta, Microcebus murinus), rodents (Rattus norvegicus, Mus musculus), Perissodactyla (Equus caballus), Artiodactyla (Bos taurus, Ovis aries, Sus scrofa), carnivores (Canis lupus familiaris, Felis catus), Lagomorpha (Oryctolagus cuniculus), Chiroptera (Pteropus vampyrus, Myotis lucifugus), Echinoderm (Echinoderma variegatus, Loxodonta africana, Sarcophilus harrisii, Ornithorhynchus anatinus, birds (Gallus gallus, Taeniopygia guttata, Apteryx australis mantelli), reptiles (Alligator mississippiensis, Chrysemys picta bellii, Anolis carolinensis, Thamnophis sirtalis), amphibian (Xenopus tropicalis, Xenopus laevis), lobe-finned fish (Latimeria chalumnae), 2 R ray-finned fish (Lepisosteus oculatus), 2 R ray-finned fish (Takifugu rubripes, Danio rerio), 2 R ray-finned fish (Orcochynus mykiss), cartilaginous fishes (Callorhinus milhi, Rhinodon typus), and Hemichordata (Saccoglossus kowalevskii). The coding sequences for Leg1/pseudo-Leg1 and protein sequences were retrieved from the NCBI, Ensembl, or UCSC Genome Browser database (Supplementary spreadsheets 3 and 4).

Multiple sequence alignment was performed using Clustal Omega with default parameters. Maximum likelihood (ML) trees were then constructed using the MEGA7 toolbox with a bootstrap testing for 1,000 times. Protein trees were established using the JTT + G, while the DNA tree was generated via the T(92) + G method. The parameters were chosen based on the BIC and AIC values given by ModelTest-NG and MEGA7. Bayesian trees were established by using MrBayes 3.2.7a.

Gene divergent time was estimated using the ReTime-ML in the MEGA toolbox according to the guideline. The calibration times were retrieved from the TimeTree.

The Gene Structure Display Server was employed to depict the organization of each Leg1 gene organization by comparing the coding sequences against their respective genome sequences.
**Microsynteny analysis.** Microsyntenic analysis was adapted from a previous study\(^{24}\). Briefly, the protein-coding genes adjacent to \(\text{Leg1}/\text{Leg1}^\prime\) were checked based on the available genome annotation data. The analysed species included eutherians, Ornithorhynchus anatinus, Sarcophilus harrisi, Monodelphis domestica, birds (Gallus gallus, Taeniopygia guttata), reptiles (Alligator mississippiensis, Chrysemys picta bellii), amphibians (Xenopus tropicalis, Xenopus laevis), fishes (Danio rerio, Oncorhynchus mykiss, Oncorhynchus tsawatscha, Oncorhynchus kisutch, Salvelinus alpinus, Salmo salar, Labrus bergylta, Sinocyclocheilus anshanensis, Sinocyclocheilus grahami, Sinocyclocheilus rhinoceros, Hippocampus comes, Takifugu rubripes, Oryzias latipes, Callorhinchus millii, Rhinodon typus, Latimeria chalumnae), and invertebrates, diatoms, and bacteria. For those species with incomplete genome information, such as Latimeria chalumnae, in which the contigs -\(\text{leg1}\)-saga3-chdc1 and Ptprk-themis are found in unmapped scaffolds, the synteny information will be partial and speculative based on closely related species. For 3 R and 4 R bony fish, the absence of \(\text{Leg1}\) in some deuterostomes, the \(\text{Leg1}\) found in different chromosomes or linkage groups. To predict the absence of \(\text{Leg1}\) synteny groups were cation due to whole-genome duplication was considered to have occurred when two \(\text{Leg1}\) synteny groups were found in different chromosomes or linkage groups. To predict the absence of \(\text{Leg1}\) in some deuterostomes, the genes adjacent to \(\text{Leg1}\) according to the microsynteny of vertebrates and Saccoglossus kowalevskii were also subjected to BLAST searches against the genomes. When only \(\text{Leg1}\) was absent while other genes were found in the genomes, \(\text{Leg1}\) was considered to have been lost.

**Selection force d\(n/dS\) analysis.** Multiple sequence alignment results generated by Clustal Omega were transferred to a codon alignment analysis using PAL2NAL\(^{27}\). Then, the Z-test of selection in MEGA7 software was used to test overall and pairwise selection force with an alternative hypothesis of \(d_n/dS < 1\), signifying purifying force. The paralogs of \(\text{Leg1}\) from each species were analysed using KaKs Calculator 2.0 to confirm the results from MEGA7 by using the GY-HKY, YN, and \(\gamma\)-YN methods\(^{28}\).

**RNA preparation and gene cloning.** Tissues from the salivary glands (submandibular and parotid), heart, liver, spleen, lung, kidney, brain, small intestine, large intestine, and skeletal muscle of three female Rongchang pigs were collected and kindly provided by Dr. Lei Chen of the Chongqing Academy of Animal Science. These samples were stored in liquid nitrogen and subjected to RNA extraction using Total RNA Kit I (Omegabiotek) according to the manufacturer’s guideline. First-strand cDNA was then synthesized in a 25 μl volume using 1 μg RNA and d(T)\(^{18}\) primer according to the Promega M-MLV reverse transcriptase kit.

To obtain the full coding regions of the \(\text{pLeg1}\) genes, three pairs of primers were designed according to the predicted RNA sequences (\(\text{pleg1a}^\prime\): XM_003121211.1, \(\text{pleg1b}^\prime\): XM_021074892.1, \(\text{pleg1c}^\prime\): XM_021084485.1) spanning the distance from start codons to the stop codons. 3’ rapid amplification of cDNA ends (RACE) was then performed to acquire the 3’ information for \(\text{pLeg1a}^\prime\). Briefly, 1 μg RNA was reverse-transcribed using the 3’RACE oligo dT primer. Then, two rounds of nested PCR were carried out using the primers pairs 3RACE1/\(\text{pleg1a}^\prime\)-3RACEGSP1, and 3RACEL2/\(\text{pleg1a}^\prime\)-3RACEGSP2. 5’RACE was performed using the SMARTer RACE 5’/3’ Kit (Clonetech). Two rounds of nested PCR were carried out via random priming of cDNA with the primer pairs \(\text{pleg1a}^\prime\)-5RACEGSP1/UPM(long) and \(\text{pleg1a}^\prime\)-5RACEGSP2/UPM(short). Two-step PCR reactions were all performed using Phusion High-Fidelity Polymerases (Thermo Fisher) with an annealing temperature of 60 °C. All the amplified fragments were gel-purified (Thermo Fisher) and sent to BGI Genomics for Sanger sequencing.

**Expression profiles of \(\text{pLeg1}\) genes.** RNA samples were prepared from the tissues indicated above. Reverse-transcription PCR (RT-PCR) was first employed to detect the expression patterns of \(\text{pLeg1}\) genes. For each gene, two pairs of primers were designed. The PCR reactions (35 cycles of 94 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s or 1 min) were then performed in 25 μl volume with 2.5 U Taq (Takara), 1 μl cDNA, 400 nM each primer, 200 mM MgCl\(_2\), and 200 μM dNTPs. Subsequently, to obtain a more accurate expression results, quantitative real-time PCR (qRT-PCR) was performed in triplicate for each sample from all the three pigs using a similar protocol to a previous report\(^{27}\). The qRT-PCR results were analysed using the 2\(^-\Delta\Delta C_{\text{T}}\) method\(^{29}\).

**Plasmid construction.** The eukaryotic expression plasmid for \(\text{pLeg1a}^\prime\) was constructed as follows. First, primers including \(\text{BamH}^\prime\) and \(\text{XhoI}\) sites were used to amplify \(\text{pLeg1a}^\prime\) from salivary gland cDNA. Then, the amplified fragment was digested with \(\text{BamH}^\prime\) and \(\text{XhoI}\) and cloned into \(\text{BamH}^\prime\) and \(\text{XhoI}\) sites of pCAG-3 × FLAG to construct pCAG-p\(\text{Leg1a}^\prime\)-3 × FLAG.

**Structural prediction and clustering analysis.** Protein structures were predicted using the Phyre 2 tool\(^{29}\) based on platypus MLP/LEG1c structure (PDB ID: 4V00). The ProCKSI server\(^{30}\) was then employed to compare the structures using the Voronoi algorithm\(^{31}\), and a clustering tree was generated.

**RNA-seq analysis.** HEK293T cells were cultured in DMEM (HyClone) with 10% foetal bovine serum (HyClone) in a 6-well plate until reaching 80% confluence. Then, the cells were transfected with 3 μg of the pCAG-p\(\text{Leg1a}^\prime\)-3 × FLAG vector or 3 μg of the empty pCAG-3 × FLAG vector using Lipofectamine 3000 (Thermo Fisher). After 48 hours, the cells were harvested, and total RNA was prepared. Then, the RNA was sent to Novogene (Beijing) for library construction and sequencing. The libraries were constructed with mRNA and sequenced on the Illumina Hiseq X Ten platform. The obtained reads (Gene Expression Omnibus accession: GSE134920) were assigned directly to hg38 transcripts and analysed by using Salmon (https://combine-lab.github.io/salmon/)\(^{32,33}\). After quantification, differential gene expression was carried out using the DESeq2 package\(^{34}\) with the following parameters (\(P\)-value \(\leq 0.05\) and \(|\log_2 \text{(fold change)}| \geq 1\)). GO and KEGG enrichment of differentially expressed genes was performed by using the Database for Annotation, Visualization, and Integrated Discovery (DAVID)\(^{35,36}\).
Ethical statement. All of the experimental procedures described in the paper followed the guidelines of China Council on Animal Care and were approved by the Animal Welfare Committee of Zhejiang University.

Results

Phylogenetic analysis of Leg1 and Leg1l. First, the human, mouse, platypus, and zebrafish LEG1 protein sequences were used as queries for BLAST searches against the non-redundant protein database using the phi-BLAST algorithm. A total of 413 polypeptides with characteristic DUF781 domains were retained for further analysis. These polypeptides belonged to species from taxa including bacteria (Actinomecete and Proteobacteria), slime mold, diatom, invertebrates (Protozoa, Placozoa, Cnidaria, Echinodermata, Hemichordata), and vertebrates (except for Cyclostomata). Thus, these DUF781 domain encoding genes seem to be conserved across from prokaryotes to primates. However, in plants, fungi, many invertebrates (such as *Caenorhabditis elegans* and *Drosophila melanogaster*), Protochordata, and Cyclostomata, no positive hits for Leg1/Leg1l could be recovered using several gene prediction tools and alignment methods across various databases.

To explore the phylogenetic relationships of these LEG1/LEG1L proteins, all 413 identified proteins, or those from representative vertebrates, were subjected to phylogenetic analysis using the Maximal Likelihood and Bayesian approaches, which resulted in similar topologies (Fig. 1 and S1). According to the analysis, there are three distinct clades of LEG1 in mammals (Fig. 1 and S1). The clade with human *C6orf58* homolog was named as LEG1a, and the clade with *M. murinus* paralog was labeled as LEG1b. The mammalian LEG1 clade grouped with other species is designated as LEG1c. The *Saccoglossus kowalevskii* LEG1L was used as an outgroup. The clustering results are labeled by the square brackets.

Figure 1. Phylogenetic analysis of LEG1 protein sequences in representative vertebrates performed using Bayesian (A) and ML methods (B). Node credibility is shown in A, and bootstrap values are shown in B. The two approaches produced a similar tree topology. Three mammalian LEG1 clades were generated. The one with human *C6orf58* homolog was named as LEG1a, and the clade with *M. murinus* paralog was labeled as LEG1b. The mammalian LEG1 clade grouped with other species is designated as LEG1c. The *Saccoglossus kowalevskii* LEG1L was used as an outgroup. The clustering results are labeled by the square brackets.
other species, the lobe-finned fish (*Latimeria chalumnae*) was grouped first with amphibians and then formed a separate clade with cartilaginous fishes, indicating the evolutionary linkage of amphibians and Crossopterygii37. In invertebrates, *Leg1* was found in Hemichordata and Echinodermata, demonstrating a common evolutionary ancestor of these species with vertebrates as Deuterostomes. Consistent with a previous report, most of the vertebrate *Leg1* genes resided within a microsynteny group between the *Themis* and *SOGA3* genes, except for those of 2 R ray-finned fish, 3R-/4R-teleosts, *Rattus norvegicus* and Bovidae2. To summarize the evolutionary history of the *Leg1* genes in vertebrates, molecular phylogeny in combination with synteny group analysis using the available LEG1 polypeptides was carried out. Figures 1 and 2 show that only one copy of *Leg1* remains in vertebrates after 1 R and 2 R whole-genome duplication events, since *Latimeria chalumnae* and *Lepisosteus oculatus* have only one copy of the *Leg1* gene.

*Latimeria chalumnae* exhibited the same syntenic gene group found in amphibian, reptile, birds, and mammals; however, chromosome rearrangement occurred in *Lepisosteus oculatus*, driving the *SOGA3* gene to a location approximately 10 Mb upstream of *Leg1* in the same linkage group. Additionally, among other bony fish species, only Salmoninae, Sinocyclocheilus, zebrafish (*Danio rerio*), Hippocampus comes, and Labrus bergylta harboured more than one copy of *Leg1*. Further microsyntenic analyses revealed that the *Leg1* copies of the first two taxa resulted from a whole-genome duplication event yielding two copies of *Leg1* genes on different chromosomes, while the other species experienced a tandem duplication event to yield an extra copy of *Leg1*.

**The *Leg1* gene is evolutionarily constrained.** To determine whether *Leg1* genes are evolutionarily constrained, *Leg1* coding sequences (Supplementary spreadsheet 4) from representative vertebrates were analysed for the overall coverage of sequences using the Z-test of selection in the MEGA7 package. The outcome of the analysis showed a *P*-value near 0, indicating that there might be strong purifying selection on the whole vertebrate *Leg1* genes. Furthermore, to determine whether the *Leg1* gene may acquire a new function or preserve its current activity, within species *Leg1* paralogs were subjected to dN/dS calculation. Supplementary table S1 points out that nearly all of the duplicated pairs were subjected to a purifying force; however, there are still some exceptions, including *Rattus norvegicus* Leg1c, zebrafish Leg1, Loxodonta africana Leg1a, Myotis lucifugus Leg1c, and Oryctolagus cuniculus Leg1c. These results indicate that only a small proportion of recently duplicated *Leg1* sequences might have experienced neutral selection or positive selection (e.g., *Oryctolagus cuniculus* dN/dS > 1).

**Molecular cloning and characterization of pLeg1.** Evolutionary analyses indicated that mammalian *Leg1a* orthologs have the highest probability of resembling the function of h*Leg1a*. As mouse *Leg1a* has been cloned and characterized previously, we identified the pig *Leg1a* gene in the current study35. Based on the information provided by the NCBI and Ensembl databases, three putative p*Leg1* genes were identified on chromosome 1 (LOC100511607, LOC100512146, LOC110259407), between the *THEMIS* and *SOGA3* gene loci. Herein, we designated these genes p*Leg1a*, p*Leg1b*, and p*Leg1c* according to their phylogenetic grouping. The lengths of the predicted coding regions of these three genes were 1,014 bp, 1,020 bp, and 684 bp, respectively, spanning a
region of ~80 kb (Fig. 3A). After sequencing the genes using salivary gland RNA, the open reading frame of pLeg1a (GenBank Accession no. MN481509) was found to contain 1,014 bp with a sequence identical to the XM_003121211.1. Information for the other two pLeg1s could not be obtained by molecular cloning. Thus, in the subsequent analysis, the predicted mRNA and protein sequences were used (pLeg1b: XM_021074892.1/XP_020930551.1, pLeg1c: XM_021084485.1/XP_020940144.1). Similar to their human, mouse, and zebrafish counterparts, pLeg1a and pLeg1b have six exons, while pLeg1c only has 5 exons (Fig. 3B).

The pLeg1 genes encode three polypeptides with lengths of 339 aa, 337 aa, and 227 aa. These polypeptides all have a characteristic Domain of Unknown Function 781 (DUF781/LEG1, pfam05612, accession no. cl05272) domain (Fig. 3C). Sequence comparison showed that pLEG1a and pLEG1b are closest to each other with a 66%/79% identity/similarity, while pLEG1c was more distant from the other two pLEG1 proteins, with 19%/32% and 18%/31% identity/similarity. Pairwise comparison also indicated that pLEG1a showed the highest identity/similarity to human LEG1a (Table 1).

Expression profile of pLeg1 genes. To determine the expression pattern of these three pLeg1 copies, RT-PCR was first employed, demonstrating that only pLeg1a was specifically detectable in the salivary glands, while there was no signal for pLeg1b or pLeg1c (Fig. 4A and S3). Then, qRT-PCR was performed to confirm the RT-PCR results. As shown in Fig. 4B, pLeg1a was highly expressed in the salivary glands, and pLeg1b/1c were undetectable in various tissues. BLAST was also employed using the pLeg1 sequences as queries against the EST database, and only hits for pLeg1a were found, mainly from the salivary glands.

Structural prediction implies that LEG1 might retain a conserved function. Currently, only the platypus LEG1c/MLP protein has been structurally resolved. Based on the information provided, other LEG1 protein structures from representative vertebrates were predicted using the Phyre2 tool. The resulting prediction showed similar structures of all LEG1 proteins except for pLEG1c (Fig. 5). To quantify the similarities between different LEG1 proteins, the predicted structural information was submitted to the ProCKSI server, and clustering was finally established. The majority of these LEG1 proteins were clustered in accordance with the phylogenetic tree using sequence information (Fig. S4). The structural tree presented two major branches, in one of which...
Figure 4. Expression patterns of pLeg1 genes in various tissues. (A) RT-PCR analysis of pLeg1a, pLeg1b, and pLeg1c in pig tissues showed that pLeg1a was specifically expressed in the salivary gland (upper panel). pLeg1b and pLeg1c signals could not be obtained in these tissues (the middle two panels). GAPDH was used as internal control (lower panel). (M, marker. He, heart. Li, liver. Sp, spleen. Lu, lung. Ki, kidney. Br, brain. LI, large intestine. SI, small intestine. SM, skeletal muscle. SG, salivary gland. NC, negative control). (B) The expression pattern of pLeg1a was confirmed by qRT-PCR analysis using the expression level of GAPDH as reference. Data are presented as mean ± SEM.

Figure 5. Structural comparison of LEG1 proteins from different species. The platypus MLP protein structure was retrieved from the PDB (4V00), while the others were predicted using Phyre 2. The colors are in rainbow order with red and blue colors indicate the N- and C- termini of LEG1, respectively. All LEG1 proteins exhibit the similar structural prediction result except for pLEG1c, which is slightly different from others due to the lack of the signal peptide.
teleost LEG1s were first grouped with platypus MLP and then clustered together with bird, amphibian, and reptile LEG1s. Eutherian LEG1c, together with rat LEG1a and dog LEG1b, was also grouped within this branch. The other branch mainly contained eutherian LEG1 proteins (LEG1a and LEG1b), although LEG1 from the metathalian *Sarcophilus harrisii* was in this branch as well. These results suggested that the structures of LEG1 proteins are highly similar to each other and were analogous to the phylogenetic results. Additionally, hLEG1a, mLEG1a, and pLEG1a, and pLEG1b are structurally closely related, indicating they may possess similar functions. Therefore, mouse and pig are good models for studying the function of hLEG1 gene.

**Preliminary functional prediction of pLEG1a using RNA-seq.** To functionally predict the role of pLEG1a, HEK293T cells were transfected with pCAG-pLEG1a-3 × FLAG and the empty vector for RNA-seq (GSE134920). A total of 152 genes were differentially expressed (|log2Fold Change| ≥ 1 and P-value ≤ 0.05), among which 85 DEGs were downregulated, and 67 DEGs were upregulated. One of these DEGs was *PPARγ*, which plays a role in the regulation of lipid metabolism and adipocyte differentiation (Fig. 6A). Enrichment analysis showed that some DEGs were enriched in the negative regulation of triglyceride sequestering. In addition, some DEGs were enriched in calcium associated biological processes (Fig. 6B). These results indicate that mammalian *Leg1* genes might be involved in lipid and calcium homeostasis.

**Discussion**

In the current study, we described the evolution of *Leg1/Leg1l* and molecularly characterized *pLeg1* genes, including a phylogenetic study and analysis of their cDNA sequences, expression profiles, exon-intron organization, predicted structures, and potential associated molecular processes. We propose that as genes with unknown function in eutherian species, *pLeg1a, mLeg1a*, and *hLeg1a* might share similar functions indicating that pigs and mice are good models for studying *hLeg1a*.

An initial BLAST analysis showed that *Leg1/Leg1l* exists from prokaryotes to mammals, demonstrating that the *Leg1/Leg1l* gene is conserved. However, in prokaryotes, plants, and Protostomia, copies of *Leg1/Leg1l* were detected in only a few species. One possible reason for this result might have been that *Leg1/Leg1l* could not be found in these species due to poor genome annotations. For example, in the Cyclostomata genome, it was not only *Leg1* but also genes such as *Soga3, Themis*, and *Echdc1* could not be found. However, there was also no signal observed in some well-characterized organisms, such as *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana*, indicating that *Leg1/Leg1l* has been lost and might not be essential in these species. Within Deuterostomia, *Leg1/Leg1l* was identified in the Echinodermata, Hemichordata, and Gnathostomata, but not in Urochordata, Cephalochordata, Cyclostomata, or Xenoturbella. Molecular phylogeny indicates that hemichordates and echinoderms are closely related and form a supragenome referred to as Ambulacarina, in which the *Leg1/Leg1l* gene could be found in this clade (Fig. S1). In Proteochordata, we used the adjacent genes according to the information from vertebrates and hemichordates as queries to search the genomes of its members. The presence of these syntenic genes in combination with the absence of a *Leg1/Leg1l* hit strongly suggested that *Leg1/Leg1l* has been lost in Proteochordata. Therefore, during chordate evolution, *Leg1* might have been lost in protochordates and maintained in jawed vertebrates. This result suggests that *Leg1* may be relevant to specific biological mechanisms or behaviours in jawed vertebrates.

Analysis of the synten of *Leg1* in jawed vertebrates revealed that cartilaginous fish, lobe-finned fish, and tetrapods exhibit a conserved organization in which *Leg1* resides between the *Themis* and SOGA3 genes, while in Actinopterygii, genomic rearrangements have occurred (Fig. 2). Actinopterygii and Sarcopterygii diverged approximately 440 million years ago (Mya), and *Lepisosteus oculatus* is regarded as a bridge connecting teleost and tetrapod species. The syntenic organization has changed in *Lepisosteus oculatus*, suggesting a possible chromosomal rearrangement in the common ancestor of Actinopterygii. This rearrangement was not due to the 2 R genome duplication because evidence suggests that the 2 R genome duplication took place before the divergence of jawed and jawless vertebrates. It is notable that most 3 R ray-finned fish only have one copy of *Leg1*, despite...
the teleost-specific whole-genome duplication\(^{11-16}\). Thus, the Leg1 gene underwent nonfunctionalization after 3 R, which is a common outcome of duplication events\(^7\). There were also some exceptions in 3 R ray-finned fish such as \textit{Labrus bergylta}, zebrafish (\textit{Danio rerio}), and \textit{Hippocampus comes}, in which two tandemly linked \textit{Leg1} copies could be found (Fig. S1). Local tandem duplication may be a better explanation for this phenomenon. Detailed analysis of Salmoninae (\textit{Onchorhyncus mykiss}, \textit{Onchorhyncus tsawytscha}, \textit{Onchorhyncus kisatch}, \textit{Salvelinus alpinus}, \textit{Salmos salar}), \textit{Sinocyclocheilus anhuensis}, and \textit{Sinocyclocheilus grahami} revealed the presence of multiple copies of \textit{Leg1} located on different chromosomes. This phenomenon is most likely due to 4 R genome duplication in Salmonids and Cyprinids\(^{48}\).

Most Sarcopterygii and tetrapod species (birds, reptiles, and amphibians) exhibit preservation of the original syntenic organization of \textit{Leg1} with only one \textit{Leg1} copy; however, in mammals, multiple copies of \textit{Leg1} emerged again (Fig. 2), resulting in the formation of three \textit{Leg1} clades. In one clade, platypus MLP groups with a few other mammalian \textit{Leg1}s in addition to bird and reptile \textit{Leg1}s, suggesting an early duplication event before the divergence of eutherians from other mammals. Then, another duplication drove the formation of \textit{Leg1a} and \textit{Leg1b}. The estimated duplication time for \textit{Leg1c} from other \textit{Leg1}s was about 232.2 Mya, which is earlier than the divergence of eutherians from proto- and metatherians. Then, another duplication splitting \textit{Leg1a} and \textit{Leg1b} occurred 100 Mya, which is around the divergence of eutherian species (Fig. 2)\(^{22}\). As a consequence, proto- and metatherian \textit{Leg1}s are clustered as an outgroup to eutherian \textit{Leg1a} and \textit{Leg1b}. As shown in Fig. S1, \textit{Leg1a} is present in nearly all mammalian species, while some species or the majority of mammalian species lack \textit{Leg1b} or \textit{Leg1c}, respectively. In platypus, MLP has been suggested to exhibit antibacterial activity in the nipple-less delivery of milk to hatchlings\(^6\). However, this mechanism is not needed by eutherian species. Therefore, we speculate that \textit{Leg1b} and especially \textit{Leg1c} are not as essential as \textit{Leg1a} and that they experienced nonfunctionalization and were lost during evolution in most eutherian species. A previous study involving genome information for 49 vertebrates\(^2\), in which platypus, \textit{Sarcophilus harrisii}, \textit{Cavia Porcellus}, \textit{Bos taurus}, and \textit{Oris aries} were grouped together. However, the authors suggested that the grouping did not reflect the evolution of these species. In our study, after adding information from other mammals, we concluded that platypus MLP is a paralog of platypus \textit{C6orf58}, and that MLP/\textit{Leg1c} may present different functions in mammals.

Primate, Scandentia, and Dermoptera form a clade known as Euarchonta, which presents a close relationship to Glires\(^{49-52}\). The analysis of primates, \textit{Tupaia chinensis}, and \textit{Galeopterus variegatus} genomes revealed no \textit{Leg1c}, and only \textit{Microcebus murinus} presented a copy of \textit{Leg1b}, which was significantly different from those found in Glires. Therefore, during the evolution of Euarchonta, \textit{Leg1b} and \textit{Leg1c} were lost. The debris of these nonfunctionalization events can still be observed in the human and \textit{Galeopterus variegatus} genomes, as pseudogenes are clustered in the \textit{Leg1b} clade (Fig. S1D). The Glires genomes are quite diverse; e.g., there are multiple copies of \textit{Leg1} in the \textit{Oryctolagus cuniculus} genome and several rodent genomes, while only \textit{Leg1a} is detected in \textit{Genus Mus}. A possible reason for this situation might be that multiple genome alterations have taken place in this clade, resulting in extreme species diversity, especially in rodents\(^{53}\). Analysis of Laurasiatheria showed that in Carnivora and Cetacea, only \textit{Leg1a} and \textit{Leg1b} are present in the genome, while \textit{Leg1c} has been lost. Suidae and Perissodactyla exhibit all 3 copies of \textit{Leg1}, indicating the preservation of the original syntenic organization after \textit{Leg1} duplication. Among these species, Bovidae species only present \textit{Leg1a} and \textit{Leg1c} copies. Detailed genome analysis indicated that a possible chromosomal inversion with one break site between \textit{Leg1c} and \textit{Leg1a} drove the loss of \textit{Leg1b}, with another break site residing between the SNAP91 and Ripply 2 genes (Fig. S2). Thus, before inversion, the order of the genes should have been SNAP91-Ripply 2-CyB5R4-(...)-SOGA3-Leg1a-Leg1b-Leg1c-Themis, which then became SNAP91-Leg1a-SOGA3-(...)-CYB5R4-Ripply 2-Leg1c-Themis in Bovidae. A similar phenomenon can be found in the \textit{Rattus norvegicus} genome (Fig. S2); however, due to complex genome rearrangement, the precise mechanism resulting in \textit{Rattus norvegicus} synteny needs to be further studied.

To test how evolutionary forces act on the \textit{Leg1} genes, we performed an overall Z-test of selection on \textit{Leg1} sequences from representative vertebrates. Strong purifying selection was suggested by the test, implying that the \textit{Leg1} genes have probably maintained their function during evolution. As paralogs may exhibit different fates after duplication (nonfunctionalization, subfunctionalization, and pseudofunctionalization/nonfunctionalization\(^{54}\)), we evaluated the \(dN/dS\) ratios between paralogous \textit{Leg1} genes within each species. The results shown in Table S1 indicate that most of the duplicates are functionally constrained (\(dN/dS < 1\)), with a few exceptions (e.g., rabbit (\(dN/dS > 1\) and zebrafish (\(dN/dS = 1\)), indicating positive and neutral selection, respectively). Previous \textit{Leg1} functional studies were only carried out in platypus and zebrafish\(^1,2,55,56\). In these studies, different patterns of expression were observed, indicating that subfunctionalization might have taken place in these paralogs, with each \textit{Leg1} copy preserving some aspects of its parental gene functions\(^6,57,58\). In addition, \(hLeg1a\), \(mLeg1a\), and \(pLeg1a\) show significantly distinct expression patterns from their platypus and zebrafish homologs, strongly suggesting subfunctionalization. In summary, our evolutionary analysis indicates that \(mLeg1a\), \(pLeg1a\), and \(hLeg1a\) are evolutionarily closely related and may retain the same functions.

Next, we cloned and characterized pig \textit{Leg1} genes to demonstrate the molecular similarities between \textit{hLeg1a} and \textit{pLeg1a}. Our experiment showed that \textit{pLeg1a} is highly similar to human and mouse homologs in terms of expression, and structure. Three \textit{Leg1} copies were identified on pig chromosome 1 between \textit{Themis} and SOGA3, spanning a region of ~80 kb. \textit{pLeg1a} and \textit{pLeg1b} have six exons, similar to their human and mouse counterparts. However, \textit{pLeg1c} only has 5 exons (Fig. 3). Among these currently identified \textit{Leg1} proteins, \textit{pLeg1a} shows higher similarity/identity with \textit{hLeg1a} than does \textit{mLeg1a}, despite a greater evolutionary distance (Table 1 and Fig. 1). As shown by previous predictions, the characteristic DUFT81 domain follows the signal peptide\(^{1,21}\), which could be detected in all pig \textit{Leg1} proteins except for \textit{pLeg1c}. Additionally, \textit{pLeg1c} only shows 19% and 18% sequence identity to \textit{pLeg1a} and \textit{pLeg1b}, respectively. These results suggest that \textit{pLeg1c} is evolutionarily divergent from its paralogs. Transcriptional analysis showed that \textit{pLeg1a} is specifically expressed in salivary glands, whereas no signal was detected in these tissues for \textit{pLeg1b} or \textit{pLeg1c}. Our results are consistent with those of a microarray analysis demonstrating that \textit{pLeg1a} is highly expressed in the submandibular gland\(^{59}\). Hence,
pig and mouse studies have produced contrary results to those obtained in zebrafish and platypus, in which Leg1 genes are expressed in the liver. Thus, it is unlikely that mammalian Leg1a plays a role in liver development. Interestingly, Leg1c/MLP could be detected in the platypus salivary gland. Therefore, expression analysis suggests that subfunctionalization of Leg1 genes has occurred between mammals and fish. Finally, the structural prediction and clustering analysis using structural information were conducted. The results showed that hLeg1a, mLeg1a, and pLeg1a proteins are highly similar to each other structurally, implying a close functional relationship (Fig. 5 and Fig. S4). Therefore, the above experiments provided some basic evidence that pLeg1a, mLeg1a, and hLeg1a are functionally related.

There are generally two ways of studying gene function: loss of function and gain of function analyses. In this study, we performed an overexpression experiment by transient transfection of HEK293T cells using the pLeg1a expression plasmid to determine which proteins or biological processes would be affected. RNA-seq combined with enrichment analysis showed that several calcium and lipid-related pathways were involved. Among these pathways, we observed that PPARγ displayed significant downregulation (Fig. 6). As previous studies indicate that PPARγ plays a vital role in lipid homeostasis, it is likely that the Leg1a gene also participates in lipid metabolism in mammals.

In conclusion, we cloned and characterized pLeg1a for the first time and demonstrated that it shows high similarity to hLeg1a and mLeg1a from evolutionary and molecular perspectives. Additionally, pLeg1a overexpression would result in the alteration of PPARγ and lipid homeostasis according to functional prediction using RNA-seq. Thus, pLeg1a might be an excellent model for investigating the function of Leg1 genes in mammals in future studies.

Data availability
The materials used in the work are available upon contacting the authors.

Received: 6 August 2019; Accepted: 24 February 2020;
Published online: 06 March 2020

References
1. Chang, C. et al. Liver-enriched gene 1a and 1b encode novel secretory proteins essential for normal liver development in zebrafish. PLoS One 6, e22910 (2011).
2. Enjapoori, A. K. et al. Monot remate lactation protein is highly expressed in monotreme milk and provides antimicrobial protection. Genome Biol. Evol. 6, 2754–2773 (2014).
3. Cheng, W. et al. HNF factors form a network to regulate liver-enriched genes in zebrafish. Dev. Biol. 294, 482–496 (2006).
4. Pini, T. et al. Proteomic Investigation of Ram Spermatooza and the Proteins Conferved by Seminal Plasma. J. Proteome Res. 15, 3700–3711 (2016).
5. Ramachandran, P. et al. Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. J. Proteome Res. 5, 1493–1506 (2006).
6. Reams, A. B. & Roth, J. R. Mechanisms of gene duplication and amplification. Cold Spring Harb. Perspect. Biol. 7, a016592 (2015).
7. Tilley, S. A. & Birshtein, B. K. Unequal sister chromatid exchange. A mechanism affecting Ig gene arrangement and expression. J. Exp. Med. 162, 675–694 (1985).
8. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11, 97–108 (2010).
9. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545 (1999).
10. Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. 5, 28 (2005).
11. Zhu, Z. H., Hu, M. J., Chang, C. Q. & Peng, J. R. Analysis of expression pattern of zebrafish leg1 homologus gene mu-leg1 in mouse. Yi Chuan 34, 1174–1180 (2012).
12. Aigner, B. et al. Transgenic pigs as models for translational biomedical research. J. Mol. Med. (Berl) 88, 653–664 (2010).
13. Zhang, Z. et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 26, 3986–3990 (1998).
14. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomes. J. Mol. Biol. 268, 78–94 (1997).
15. Marchler-Bauer, A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203 (2017).
16. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539 (2011).
17. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33, 1870–1874 (2016).
18. Darriba, D. et al. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol. Biol. Evol. 37, 291–294 (2020).
19. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542 (2012).
20. Tamura, K. et al. Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. USA 109, 19333–19338 (2012).
21. Mello, B. Estimating TimeTrees with MEGA and the TreeTime Resource. Mol. Biol. Evol. 35, 2334–2342 (2018).
22. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819 (2017).
23. Hu, B. et al. GDS2 2.0: an upgraded gene feature visualization server. Bioinformatics 31, 1296–1297 (2015).
24. Lin, J. J., Wang, F. Y., Li, W. H. & Wang, T. Y. The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. Sci. Rep. 7, 15568 (2017).
25. Syuama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–612 (2006).
26. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs, Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8, 77–80 (2010).
27. He, J., Wang, Q., Ye, J., Hu, X. & Li, N. Identification of porcine polycystic kidney disease 1 (PKD1) gene: molecular cloning, expression profile, and implication in disease model. Gene 490, 37–46 (2011).
28. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta CT) Method. Methods 25, 402–408 (2001).
29. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–858 (2015).
30. Barthel, D., Hirst, J. D., Blazewicz, J., Burke, E. K. & Krasnogor, N. ProCCKSI: a decision support system for Protein (structure) Comparison, Knowledge, Similarity and Information. BMC Bioinformatics 8, 416 (2007).
31. Birzele, F., Gewehr, J. E., Csaba, G. & Zimmer, R. Voronoi–fast structural alignment using Voronoi contacts. Bioinformatics 23, e205–211 (2007).
32. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).
33. Sahraei, S. M. E. et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat. Commun. 8, 59 (2017).
34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
35. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13 (2009).
36. Huang, D. W. et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35, W169–175 (2007).
37. Nolan, G. E., Klein, B., McEwen, D. S., Heng, C. M. & Revenko, A. Structural and genetic consequences of the coelacanth Latimeria indicated by haemoglobin sequences. Nature 351, 394–397 (1991).
38. Newman, J. et al. Structural characterization of a novel monotreme-specific protein with antimicrobial activity from the milk of the platypus. Acta Crystallogr. F. Struct. Biol. Commun. 74, 39–45 (2018).
39. Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. & Telford, M. J. Xenoturbella is a deuterostome that eats molluscs. Nature 424, 293–298 (2003).
40. Turbeville, J. M., Schulz, J. R. & Raff, R. A. Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. Mol. Biol. Evol. 11, 648–655 (1994).
41. Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. & Postlethwait, J. H. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188, 799–808 (2011).
42. Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol. Biol. Evol. 26, 47–59 (2009).
43. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. 19, 166 (2018).
44. Amores, A. et al. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714 (1998).
45. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 13, 382–390 (2003).
46. Jaillon, O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946–957 (2004).
47. Ohno, S., Wolf, U. & Atkin, N. B. Evolution from fish to mammals by gene duplication. Genome Biol. 15, 550 (2014).
48. Johnson, K. R., Wright, J. E. Jr. & May, B. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Trends Genet. 15, 169–187 (2001).
49. Lynch, M. & Conery, J. S. The evolution of convergence and consequences of duplicate genes. Science 290, 1151–1155 (2000).
50. Hu, M. et al. Liver-Enriched Gene 1, a Glycosylated Secretory Protein, Binds to FGFR and Mediates an Anti-stress Pathway to Protect Liver Development in Zebrafish. Proc. Natl. Acad. Sci. USA 104, 1159–1164 (2007).
51. Murphy, W. J. et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348–2351 (2001).
52. Waddell, P. J., Kishino, H. & Ota, R. A phylogenetic foundation for comparative mammalian genomics. Genome Inform. 12, 141–154 (2001).
53. Blanga-Kanfi, S. et al. Rodent phology revised: analysis of six nuclear genes from all major rodent clades. BMC Evol. Biol. 9, 71 (2009).
54. Lynch, M. & Conery, J. S. The evolution of convergence and consequences of duplicate genes. Science 290, 1151–1155 (2000).
55. Hu, M. et al. Liver-Enriched Gene 1, a Glycosylated Secretory Protein, Binds to FGFR and Mediates an Anti-stress Pathway to Protect Liver Development in Zebrafish. PLoS Genet. 12, e1005881 (2016).
56. Lin, T. Y. et al. Hypoxia-inducible factor 2 alpha is essential for hepatic outgrowth and functions via the regulation of leg1 transcription in the zebrafish embryo. PLoS One 9, e101980 (2014).
57. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. Proc. Natl. Acad. Sci. USA 104, 1159–1164 (2007).
58. Freeman, T. C. et al. A gene expression atlas of the domestic pig. BMC Biol. 10, 90 (2012).
59. Jones, J. et al. Deletion of PPARgamma in adipose tissues of mice protects against high fat diet-induced obesity and insulin resistance. Proc. Natl. Acad. Sci. USA 102, 6207–6212 (2005).

Acknowledgements
This work was supported by the National Natural Science Foundation of China (Nos. 2018FZ199016, 31830113) and the Fundamental Research Funds for Central Universities (No. 2018FZ199016). We also thank Dr. Ruifei Yang and Dr. Yiqiang Zhao (China Agricultural University) for their suggestions in the computational analyses.

Author contributions
J.H. conceived the study and designed the experiment. Y.D., J.W., and C.L. performed the experiment. Y.D. and J.H. analyzed the data. J.H. wrote the manuscript. Y.D., J.W., C.L., K.Z., P.J. and J.H. edited the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-61208-7.

Correspondence and requests for materials should be addressed to J.H.

Reprints and permissions information is available at www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020