SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Signaling and regulatory biology
• RESEARCH PAPER •

February 2017 Vol.60 No.2: 178–188
doi: 10.1007/s11427-015-0348-5

Significant variations in alternative splicing patterns and expression profiles between human-mouse orthologs in early embryos

Geng Chen[1,2], Jiwei Chen[1], Jianmin Yang[1], Long Chen[1], Xiongfei Qu[1], Caiping Shi[1], Baitang Ning[3], Leming Shi[2,3], Weida Tong[3], Yongxiang Zhao[4*], Meixia Zhang[5**] & Tieliu Shi[1***]

[1]*The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China;*
[2]*Center for Pharmacogenomics, School of Pharmacy, Fudan University, Shanghai 201203, China;*
[3]*National Center for Toxicological Research, US Food and Drug Administration, Jefferson AR 72079, USA;*
[4]*Biological Targeting Diagnosis and Therapy Research Center, Guangxi Medical University, Nanning 530021, China;*
[5]*Department of Ophthalmology, West China Hospital, Sichuan University, Chengdu 610041, China*

Received November 18, 2015; accepted February 11, 2016; published online July 4, 2016

Human and mouse orthologs are expected to have similar biological functions; however, many discrepancies have also been reported. We systematically compared human and mouse orthologs in terms of alternative splicing patterns and expression profiles. Human-mouse orthologs are divergent in alternative splicing, as human orthologs could generally encode more isoforms than their mouse orthologs. In early embryos, exon skipping is far more common with human orthologs, whereas constitutive exons are more prevalent with mouse orthologs. This may correlate with divergence in expression of splicing regulators. Orthologous expression similarities are different in distinct embryonic stages, with the highest in morula. Expression differences for orthologous transcription factor genes could play an important role in orthologous expression discordance. We further detected largely orthologous divergence in differential expression between distinct embryonic stages. Collectively, our study uncovers significant orthologous divergence from multiple aspects, which may result in functional differences and dynamics between human-mouse orthologs during embryonic development.

ortholog, alternative splicing, RNA-seq, early embryo, gene expression

Citation: Chen, G., Chen, J., Yang, J., Chen, L., Qu, X., Shi, C., Ning, B., Shi, L., Tong, W., Zhao, Y., Zhang, M., and Shi, T. (2017). Significant variations in alternative splicing patterns and expression profiles between human-mouse orthologs in early embryos. Sci China Life Sci 60, 178−188. doi: 10.1007/s11427-015-0348-5

INTRODUCTION

Orthologs are homologous genes in distinct species derived from a speciation event, whereas paralogs often belong to the same species and result from a duplication event. The mouse has been broadly used as a model organism to study human biology as they are inexpensive, easy to raise, propagate rapidly, and have short life cycles (Elso et al., 2008; Gharib and Robinson-Rechavi, 2011). Moreover, many genes are conserved between humans and mice, especially for human-mouse orthologs (Mouse Genome Sequencing et al., 2002). Orthologs are widely used to infer species phylogenies and the function of uncharacterized genes in other organisms, based on known functions (Blair and Hedges, 2005; Ciccarelli et al., 2006; Koonin, 2005). A general assumption is that orthologous genes usually have equivalent biological functions in different organisms (Dolinski and Botstein, 2007). However, this assumption has been challenged by evidence of divergence between human and

*Corresponding author (email: yongxiangzhao@126.com)
**Corresponding author (email: zhangmeixia@medmail.com.cn)
***Corresponding author (email: tlshi@bio.ecnu.edu.cn)

mouse orthologs (Ginis et al., 2004; Liu et al., 2010, 2011; Nehrt et al., 2011; Yashiro et al., 2000). To correctly use mouse genes to understand the normal and pathological functions of their human orthologs, more research is required to assess genome-wide similarities and discrepancies between human-mouse orthologs (Gabaldon and Koonin, 2013; Studer and Robinson-Rechavi, 2009).

Several previous studies have compared orthologous human and mouse genes from different perspectives; however, further investigations are needed owing to prior technology/methodology limitations. Low conservation was observed in alternative splicing patterns for human-mouse orthologs (Nurtdinov et al., 2003; Takeda et al., 2008). In addition, microarray data suggested that human and mouse genes may be conserved in their gene expression patterns (Liao and Zhang, 2006; Xing et al., 2007), but the direct use of Pearson's correlation coefficient to measure gene expression similarity has been questioned for its inappropriateness and bias (Pereira et al., 2009; Piasecka et al., 2012; Qian et al., 2010). Currently, the higher resolution and larger dynamic range of RNA-Seq has made it the preferred approach for gene expression profiling (Marioni et al., 2008; Wang et al., 2009). Employing the *Z*-score as a measure of expression similarity using RNA-Seq data, a prior study observed greater expression similarities between human-mouse orthologs than those of within-species paralogs (Chen and Zhang, 2012). However, the fact that only one member of an orthologous pair was expressed was usually ignored by previous studies; this in fact also contributes to discordant expression between human-mouse orthologs. In addition, a systematic comparison of human-mouse orthologs relative to sequential developmental stages has not been performed. The advances in single-cell RNA-Seq provide a unique chance to further explore these relationships, enabling investigations on detailed gene activities in cells (Ramskold et al., 2012; van der Vegt et al., 2009).

Here, we dissected the similarities and divergences of alternative splicing patterns and expression profiles for 15,600 human-mouse orthologous pairs. In this comparison, we observed large differences in isoform number, alternative splicing modes, and GC content. We then compared expression patterns and profile changes for human-mouse orthologs by using single-cell RNA-Seq data from paired human and mouse early embryos (sequential stages from oocyte to morula except the zygote owing to the absence of mouse zygote data). We found that orthologous expression similarities varied across different embryonic stages, as many human-mouse orthologs have discrepancies in their expression profiles. Furthermore, we also detected a notable fraction of human-mouse orthologs only expressed in either human or mouse embryos. Differential expression analysis between distinct embryonic stages revealed great disparities between human-mouse orthologs in terms of expression changes. Moreover, the orthologous expression divergences increased as more combined embryonic stages were investigated.

RESULTS

Large differences in alternative splicing patterns and sequence content between human-mouse orthologs

To conduct this study, we first obtained 15,600 one-to-one human and mouse orthologous gene pairs using Ensembl BioMart (Kinsella et al., 2011). Based on Ensembl annotation (version 72) (Flicek et al., 2013), we found that orthologous human genes could generally encode more isoforms than their mouse counterparts could ($P<10^{-15}$, Wilcoxon test). On average, each orthologous human gene possessed 7.77 isoforms, which is much larger than the 3.88 transcripts per mouse ortholog. 68.6% of human orthologs were predicted to encode more isoforms than their mouse orthologs, while the reverse was true for 15.8% of genes (Figure 1A, Figure S1 in Supporting Information). Specifically, we observed 408 extreme examples wherein the orthologous human genes were predicted to encode a large number (⩾20) of isoforms, compared to their mouse orthologs possessing disproportionately fewer (⩽5) of isoforms (Table S1 in Supporting Information). About 81% and 53% of transcripts for orthologous human and mouse genes, respectively are generated by exon skipping (ES, cassette exon), and the total number of isoforms produced by ES for these human orthologs was three-fold greater than that of mouse orthologs. Furthermore, human-mouse orthologs were also divergent in other annotated alternative splicing modes, including constitutive exon (CNE), alternative 3′ splice sites (A3SS), alternative 5′ splice sites (A5SS), and intron retention (IR) (Figure 1B). Ensembl protein annotation indicated that each orthologous human gene could encode 4.36 unique proteins on average, almost twice that of (2.25) the corresponding mouse orthologs. Several distinct transcripts (different untranslated regions but identical coding sequence) can be produced from the same gene through alternative splicing and transcription (Pal et al., 2011) to encode an identical protein. This process was found to be more common for human genes (6,001 cases) than for their mouse orthologs (2,290 cases).

To study the sequence differences between human and mouse orthologs, we investigated their GC content and sequence identity. We found that GC content between human and mouse orthologs was significantly correlated (Pearson's correlation: 0.88, $P<10^{-15}$). Moreover, 45% GC content could be regarded as a demarcation point for these human-mouse orthologs (Figure 1C). The GC content of orthologous mouse genes was generally higher than that of their human counterparts when the GC content of orthologous human genes was below 45% ($P<10^{-15}$, Wilcoxon test). However, the trend was reversed when the GC percent of human orthologs was higher than 45% ($P<10^{-15}$, Wilcoxon test). Although the sequence identity for most hu

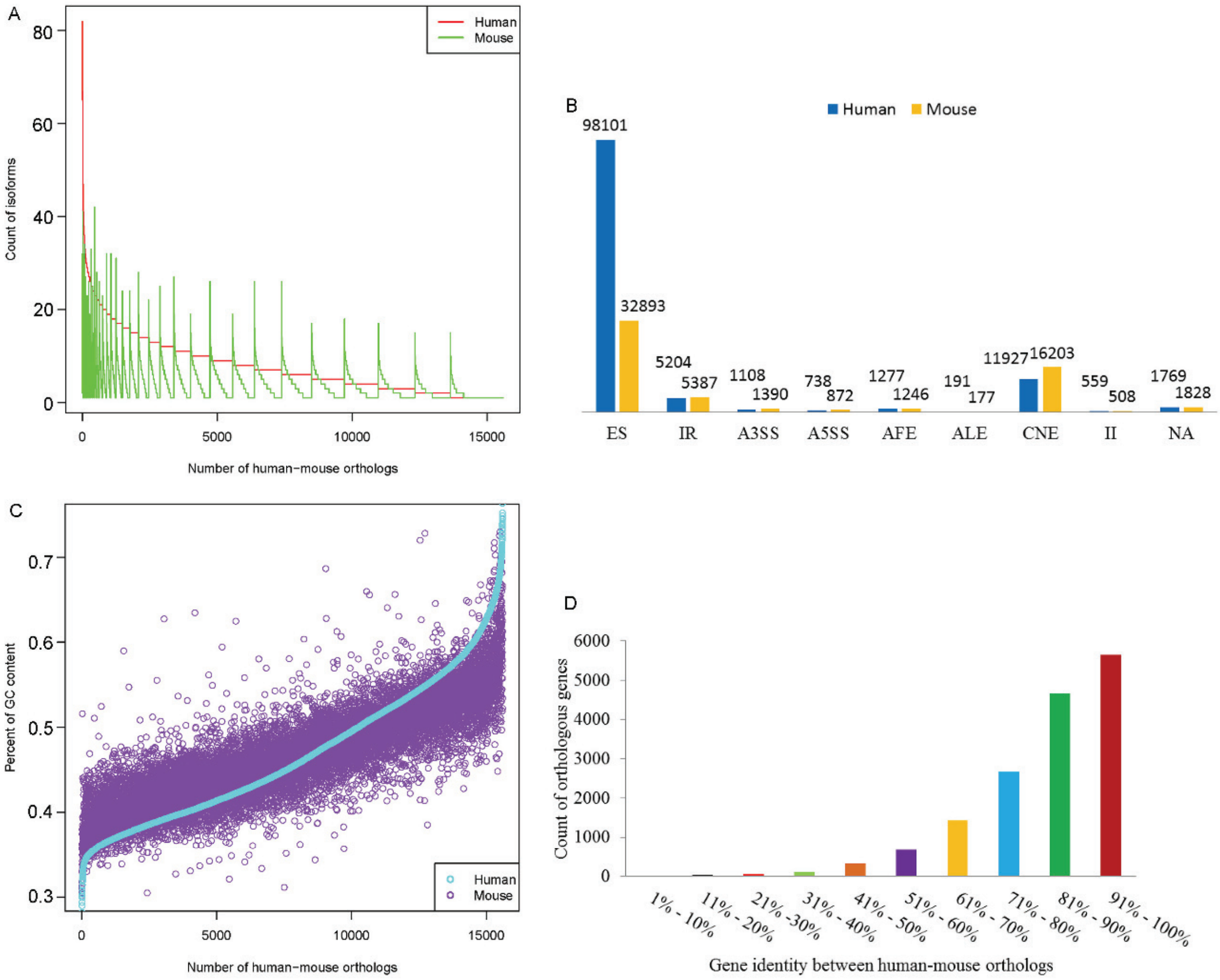

Figure 1 Alternative splicing modes and sequence content comparison between human-mouse orthologs. A, Isoform comparison for human-mouse orthologs based on Ensembl annotation. The isoform counts for orthologous human genes were drawn in descending order accompanied with their paired orthologous mouse isoform counts. B, Alternative splicing mode comparison between human-mouse orthologs. ES, cassette exon or exon skipping. IR, intron retention. A3SS, alternative 3′ sites. A5SS, alternative 5′ sites. AFE, alternative first exon. ALE, alternative last exon. CNE, constitutive exon. II, intron isoform. NA, no annotated Ensembl splicing mode. C, GC content comparison of human-mouse orthologs. The GC percentages of orthologous human genes were shown in ascending order along with that of their mouse orthologs. D, Distribution of the gene sequence identities for human-mouse orthologs.

man-mouse orthologous pairs was determined to be greater than 50%, 461 orthologous pairs showed lower sequence identity (<50%), suggesting that these orthologs may be evolving fast (Figure 1D). These results indicate that many human-mouse orthologs vary in their alternative splicing patterns and sequence content. However, we did not observe significant correlations between GC content and sequence identity or gene isoform number.

Expression comparison between human-mouse orthologs in early embryos

To explore the expression similarities and differences in human-mouse orthologs in pre-implantation embryos, we compared their expression profiles using single-cell RNA sequencing data from a previous study (Xue et al., 2013). The RNA-Seq data we used included embryonic stages from oocyte to morula except zygote of both human and mouse, and each stage contained at least two replicates (total 44 samples, Table S2 in Supporting Information). We separately mapped RNA-Seq datasets to the human (GRCh37/hg19) and mouse (GRCm38) reference genomes with TopHat2 (Kim et al., 2013), and then quantified gene and transcript expression using Cufflinks (Trapnell et al., 2010). To compare expression similarity of orthologous genes in an appropriate manner, we first transformed each gene expression value into a *Z*-score by using the logarithm of fragments per kilobase of exon model per million mapped fragments (FPKM) (Chen and Zhang, 2012). Next, we individually calculated the Spearman's correlation coefficient, based on *Z*-score values, to evaluate the expression similarity between human-mouse orthologs in different gene sequence identity intervals (see Methods).

The expression similarities between human-mouse orthologs in different embryonic stages have large variances. We found that orthologous expression similarity in the morula was significantly higher than in 8-cell embryos (P=0.0034, Wilcoxon test), which was also higher than that in other embryonic stages (Figure 2). Orthologous expression similarities in oocytes, 2-cell embryos, and 4-cell embryos were all lower than pronucleus expression similarity (P<0.015, Wilcoxon test). In addition, we observed that expression similarities between orthologous genes in different embryonic stages generally declined with decreasing sequence identity in the range of 100% to 80%; however, expression similarity became irregular when orthologous gene sequence identity was less than 80%. Overall, expression similarities between human-mouse orthologs in both the pronucleus and morula stages significantly diminished with decreasing orthologous gene sequence identity (Spearman's correlation coefficient, r=0.71, P=0.012; r=0.84, P=0.001, respectively), while those of other embryonic stages were not as obvious (r<0.5). These variances could result from orthologous expression discordance in distinct embryonic stages.

Although in general, orthologous human genes exhibited similar expression patterns with their mouse orthologs to some extent, there were many minor differences (Figure 3). We found that detectable genes and transcripts in human and mouse embryos varied greatly with 0.1 FPKM as the threshold. Fewer orthologous genes were detected in the oocyte, pronucleus, 2-cell, and 4-cell human embryos than in corresponding mouse embryos; however, this was not the case in 8-cell and morula embryos (Figure 4A). 6,920 and 6,907 orthologous genes (5,294 of them being orthologous pairs) were detected across these six human and mouse embryonic stages, respectively. Although the majority (>65%) of detectable orthologous genes in the same stage of human and mouse embryos were paired, a notable number of human-mouse orthologs had only one detectable member. However, after considering the expression across six different embryonic stages, only 38.76% of 13,993 detectable human-mouse orthologs (wherein, one or both members of

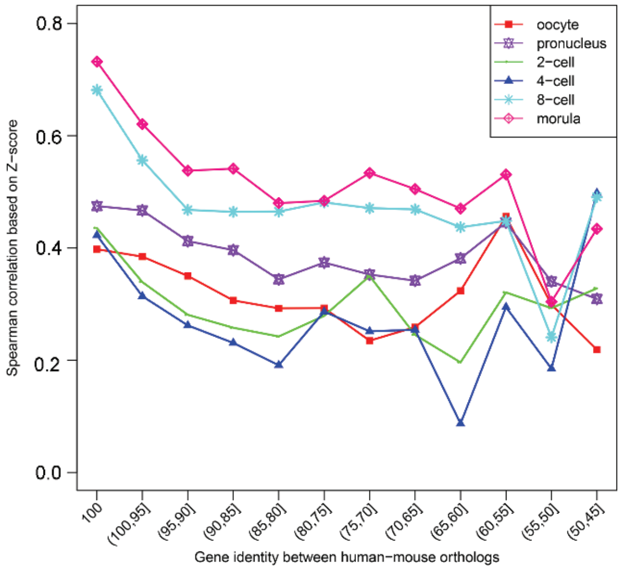

Figure 2 Expression similarities with gene sequence identities for human-mouse orthologs in different embryonic stages. The expression similarities of human-mouse orthologs were calculated based on the Z-scores, which were transformed from the FPKM values of genes. Z-score-based expression similarity was only performed on human-mouse orthologs for which both members of the pair were detectable (>0.1 FPKM).

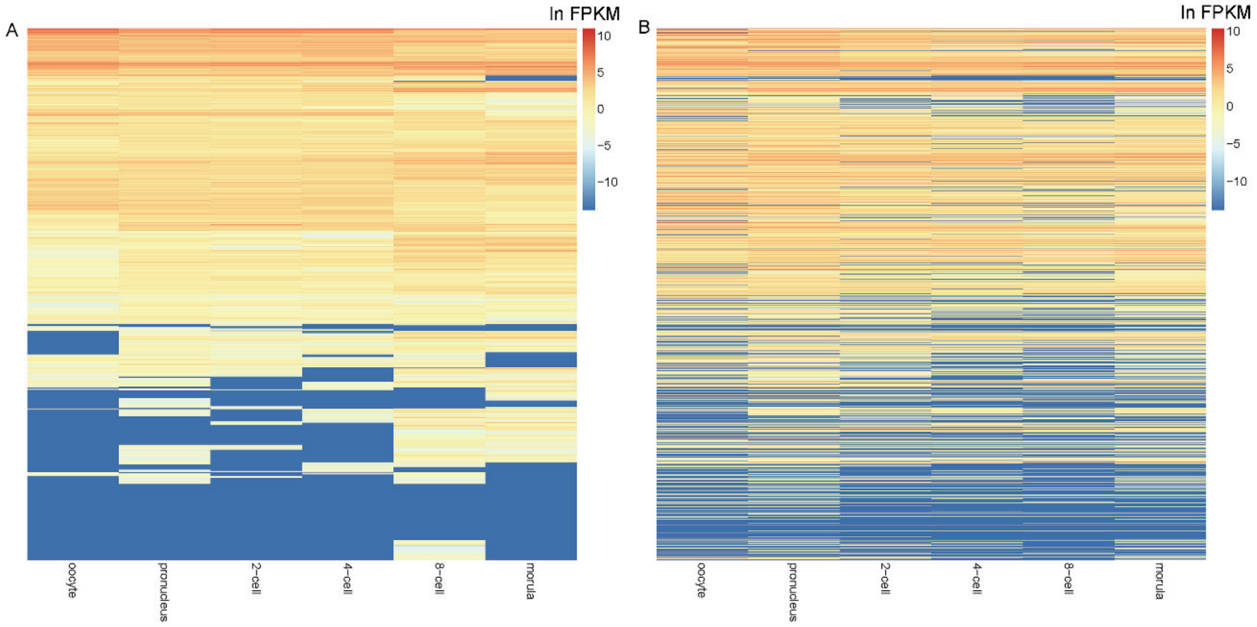

Figure 3 Expression patterns of orthologous human and mouse genes. A, Expression clustering for orthologous human genes. B, Expression pattern of orthologous mouse genes. The order of mouse genes showed in the graph is paired with the orthologous human genes clustered in Figure 3A.

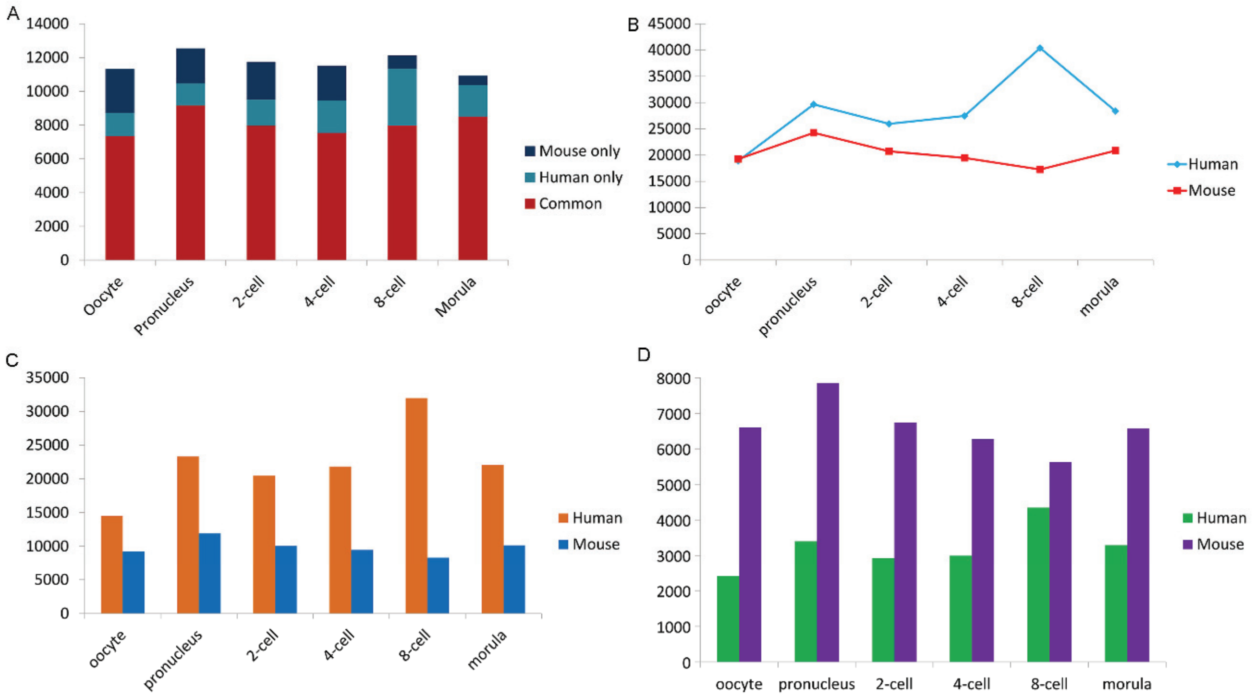

Figure 4 Comparison of expressed genes and isoforms between human-mouse orthologs. A, Distribution of expressed orthologs (human and mouse) in each embryonic stage. Common, both members of an orthologous human-mouse pair were detectable (>0.1 FPKM). Human only, only the human gene member was detectable for an orthologous pair. Mouse only, only the mouse gene member was detectable for an orthologous pair. B, Total detectable transcripts for human-mouse orthologs in each embryonic stage. C, Detected ES (exon skipping) transcript comparison for human-mouse orthologs. D, Detected CNE (constitutive exon) transcript comparison between human-mouse orthologs.

an orthologous pair were detected in at least one stage) had both members detected or not detected. Thus, for many human-mouse orthologs, both members do not turn gene expression on (or off) at the same stage. This discordance would become greater with the consideration of covering distinct stages. Moreover, we observed that hundreds of orthologous human genes were specifically expressed at >1 FPKM in a certain embryonic stage, whereas their corresponding mouse orthologs were not detectable in embryos (<0.1 FPKM) (Figure S2 in Supporting Information). Functional annotation using database for annotation, visualization and integrated discovery (DAVID) (Huang da et al., 2009) showed that these human-mouse orthologs were mainly enriched for terms such as polymorphism, alternative splicing, membrane, and glycoprotein (Table S3 in Supporting Information). At the isoform level, except for the oocyte, the total detectable transcripts encoded by these orthologous genes were notably greater in human embryonic stages than in the relevant mouse embryos (Figure 4B). However, the average detectable isoform number for orthologous human genes was higher than that of mouse orthologs in every embryonic stage. The largest discrepancy was observed at the 8-cell embryonic stage, wherein more than 23,000 human gene transcripts were detected compared to their orthologs. Furthermore, the alternative splicing event of ES was much more prevalent for orthologous human genes in each embryonic stage than that of mouse orthologs (Figure 4C). This could be a major reason for the larger number of isoforms generated by human genes. However, greater number of CNE was detected in corresponding mouse embryos (Figure 4D). Accordingly, different usage of alternative splicing modes between human-mouse orthologs could be a factor accounting for their distinct expression patterns. Taken together, these results suggest that human-mouse orthologs have certain discrepancies in their expression patterns, including alternative splicing, and this divergence is associated with embryonic stages.

Expression profile of orthologous human and mouse splicing regulators

We further checked the expression profile of 171 pairs of orthologous human and mouse splicing regulators including SR (serine-arginine-rich RNA-binding) and hnRNPs (heterogeneous nuclear ribonucleoproteins) proteins that have been previously reported (Grosso et al., 2008) (Table S4 in Supporting Information). This factors are crucial for regulating alternative splicing in gene expression (Yeo, 2005). At the gene level, majority (>89%) of these orthologous human and mouse splicing regulators were expressed (>0.1 FPKM) and many of them had relatively higher expression in each embryonic stage (Figure S3 in Supporting Information). However, at the isoform level, human orthologous splicing regulators generated more isoforms than corresponding mouse orthologs, especially at the 8-cell embryo stage (1,011 isoforms in total for human and 395 for mouse). ES and CNE are the two most abundant types of alternative splicing for both orthologous human and mouse splicing regulators; however, ES is more prevalent for human splicing regulators, and CNE is more prevalent for mouse splicing regulators. Moreover, the expression similarities between human and mouse splicing regulators in these six embryonic stages ranged from 0.35 to 0.67, with the highest being in the morula and the lowest being in the 4-cell stage. Accordingly, the results suggested that orthologous splicing regulators were widely expressed in both human and mouse early embryos, but differed in alternative splicing modes and isoform quantity. The divergence of orthologous splicing regulators may contribute to differences in splicing patterns between orthologous human and mouse genes.

Expression similarities and differences of orthologous transcription factors

Transcription factors (TFs) play crucial roles in the initiation and regulation of gene expression. We compared 699 pairs of human-mouse orthologous TFs that were separately obtained from TFClass (Wingender et al., 2013) and TFdb (Kanamori et al., 2004) databases. Most (82.83%) of these human-mouse orthologous TF pairs were highly conserved in terms of gene sequence identity ⩾80%. Ensembl annotation showed that, in general, orthologous human TFs encode a greater number of isoforms than their mouse counterparts ($P<10^{-15}$, Wilcoxon test; average 6.77 per orthologous human TF and 3.98 for each orthologous mouse TF). We observed divergent expression patterns between many orthologous human and mouse TFs (Figure 5). The expression similarities (Z-score based on Spearman's correlation coefficient) for orthologous human-mouse TFs, in different embryonic stages, ranged from 0.15 (4-cell) to 0.49 (morula). Except for 8-cell and morula embryos, the detectable orthologous mouse TFs in embryonic stages were greater than the detectable orthologous human TFs. However, a greater number of expressed isoforms could be detected from human orthologs in all embryonic stages except the oocyte, showing the distinct alternative splicing patterns between human-mouse TF orthologs. Furthermore, only 24.45% of 638 detectable human-mouse orthologous TF pairs exhibited a similar trend across different stages, wherein both members of a pair were detected or not detected, in a given stage. We also observed more than a dozen human-mouse orthologous TFs with only one member of a pair detectable in corresponding embryos. Specifically, nine orthologous human TFs were expressed in all embryonic stages, but their mouse orthologs were not detectable in any embryonic stage. Moreover, five orthologous mouse TFs were detected in every embryonic stage, but their human orthologs were not detected in any stages. Interestingly, these 14 human-mouse orthologs (nine human specific and five mouse specific) are involved in several important biological processes, including the control and regulation of differentiation and development (according to GeneCards annotation) (Rebhan et al., 1998). Hence, findings suggest that a portion of human-mouse orthologous TFs have large differences in their alternative splicing patterns and expression profiles, which could further affect the expression patterns of their target genes.

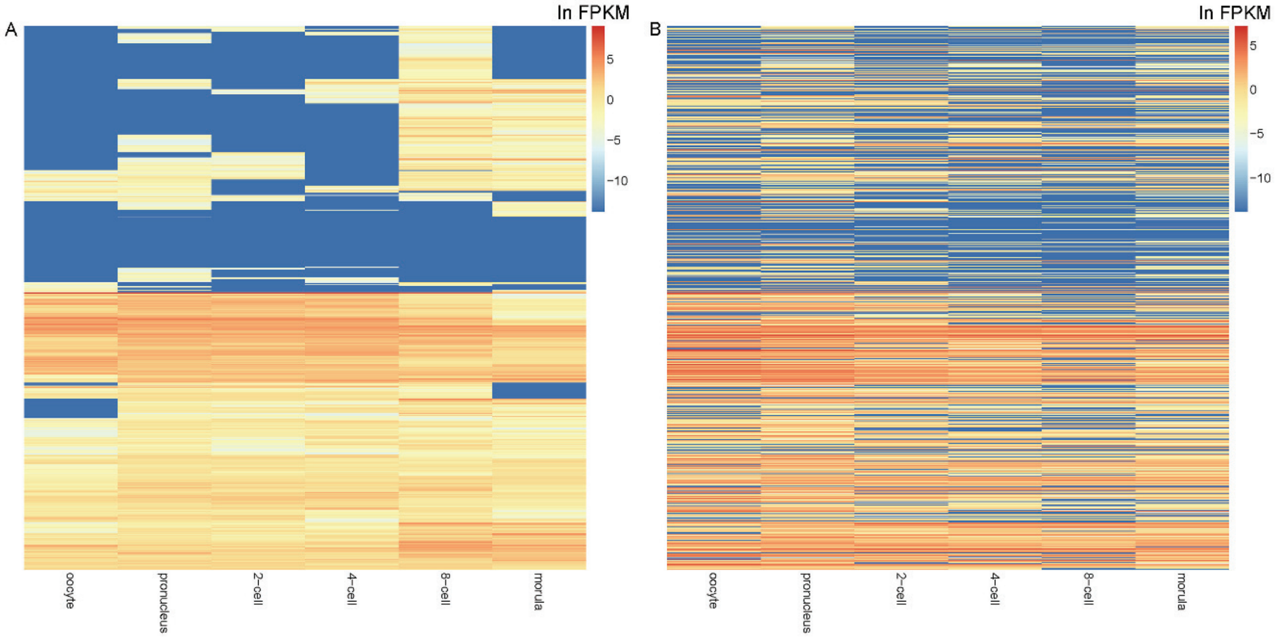

Figure 5 Expression patterns of orthologous human and mouse TFs. A, Expression clustering for orthologous human TFs. B, Expression pattern of orthologous mouse TFs. The order of mouse TFs showed in the graph is paired with the orthologous human TFs clustered in Figure 5A.

Divergence in expression changes between human-mouse orthologs

To gain insight into the discordance of expression changes of human-mouse orthologs between distinct embryonic stages, we conducted differential expression calling using Cuffdiff (Trapnell et al., 2013). Four groups of oocyte versus pronucleus, 2-cell versus 4-cell, 4-cell versus 8-cell, and 8-cell versus morula were compared. Surprisingly, we observed a large discrepancy in human-mouse orthologs in these comparisons (Figure 6). For each group comparison, the great majority of differentially expressed (P<0.05) orthologous human genes do not correspond with their mouse counterparts. Only a small fraction of human-mouse orthologs had both members consistently up-regulated or down-regulated in a specific group comparison. Moreover, some orthologous human genes were up-regulated (or down-regulated) in a certain embryonic stage, whereas the expression of their mouse orthologs was opposite. For instance, 628 human and 235 mouse orthologous genes were differentially expressed when comparing 4-cell and 8-cell embryos. However, in most cases, only one member of an orthologous pair was differentially expressed. Only two human-mouse orthologous pairs had both members concordantly differentially expressed. 14 human-mouse orthologous pairs were expressed in an opposite manner (with one member up-regulated while the other was down-regulated).

The discordance between human-mouse orthologous pairs with only one member differentially expressed, or two members expressed in a reverse trend, may account for the distinctions in embryonic development between humans and mice, to some extent. Functional enrichment analyses for these involved human-mouse orthologs using DAVID (Huang da et al., 2009) suggested that they were mainly enriched for terms including phosphoprotein, alternative splicing, and nucleus (Table S5 in Supporting Information). Intriguingly, we also found six human-mouse orthologs with both members differently expressed in a reverse trend, wherein group comparison identified them to be functionally related to embryonic development (Table S6 in Supporting Information). Moreover, the discordance in differential expression between human-mouse orthologs increased when considering the constant expression changes across distinct developmental stage comparisons. For example, although a total of 576 human-mouse orthologous pairs were concordantly differentially expressed or exhibited in-

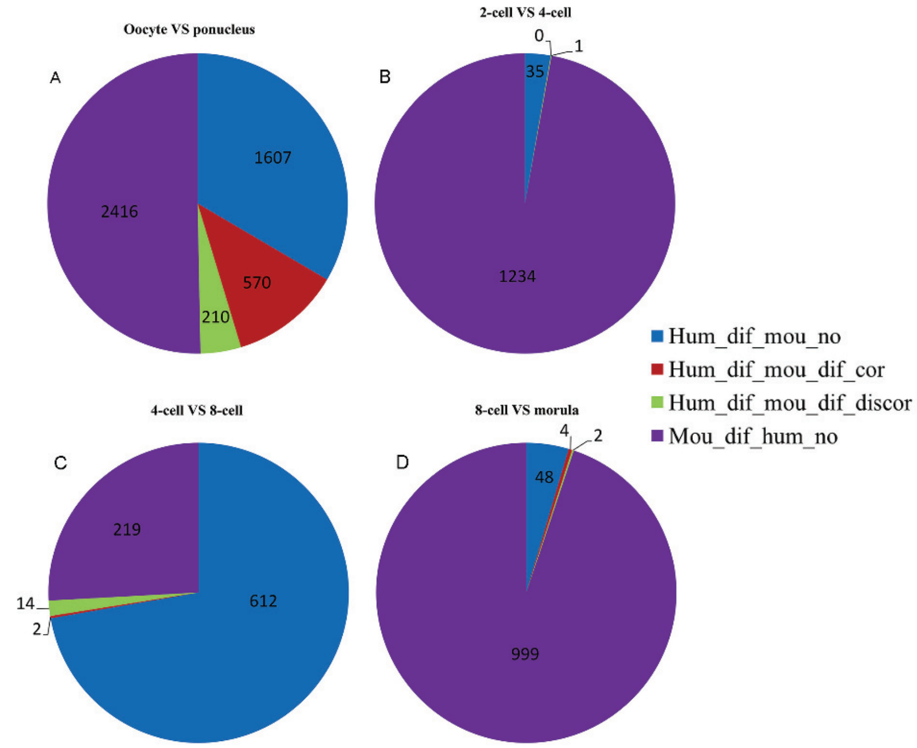

Figure 6 Differential expression profiling of human-mouse orthologs. A, Detected differentially expressed orthologous human and mouse genes in oocyte versus pronucleus. Hum_dif_mou_no, only the human gene member was differentially expressed for an orthologous pair. Hum_dif_mou_dif_cor, both members of a human-mouse orthologous pair were differentially expressed and their expression changes were in the same trend (both were up-regulation or down-regulation). Hum_dif_mou_dif_discor, both members of a human-mouse orthologous pair were differentially expressed but their expression changes were in the reverse trend (one was up-regulated while the other was down-regulated). Mou_dif_hum_no, only the mouse gene member was differentially expressed for an orthologous pair. B, Detected differentially expressed human and mouse orthologs in the comparison of 2-cell versus 4-cell. C, Detected differentially expressed human and mouse orthologs in the comparison between 4-cell and 8-cell. D, Detected differentially expressed human and mouse orthologs in the comparison of 8-cell against morula.

significant expression changes in each group comparison (but differentially expressed in at least one group), 197 orthologous pairs of did not show consistent expression changes across these four groups. Accordingly, many human-mouse orthologs were greatly divergent in expression changes between different embryonic stages and might closely associate with the discordant embryonic development between humans and mice.

DISCUSSION

Previous studies mainly compared human-mouse orthologs based on one or two aspects of alternative splicing modes and expression patterns; few studies have investigated expression similarities and differences of human-mouse orthologs in sequential stages. We compared the alternative splicing patterns, expression patterns, and expression changes of one-to-one human-mouse orthologs in paired early embryos. Orthologous human genes could generally produce more transcripts and proteins than their mouse counterparts could. This was supported by both Ensembl annotation and our expression analyses of single-cell RNA-Seq data from human and mouse pre-implantation embryos. These human-mouse orthologs were divergent in their alternative splicing patterns and the two notably distinct splicing modes were ES and CNE. Although ES is the main splicing event for both human and mouse orthologs, ES is far more prevalent for orthologous human genes than their mouse counterparts are. This is also responsible for orthologous human genes having a larger number of isoforms than mouse orthologs. In contrast, a larger number of CNEs were found for orthologous mouse genes in different embryonic stages. The expression divergence of orthologous splicing regulators may play an important role in these disparities in alternative splicing patterns between human-mouse orthologs. These observations largely benefit from the high-resolution of single-cell RNA-Seq technology. Human-mouse orthologs are also different in their GC content. An intriguing phenomenon was that orthologous human genes with high GC content (>45%) generally have mouse orthologs with lower GC content and vice versa. Therefore, large variances were generated, in terms of gene structure, between human-mouse orthologs during evolution.

Alternative splicing determines how genes generate specific isoforms, which exert particular functions in a certain condition. Different isoforms could play distinct roles, while the sum of their expression comprises their gene expression level. Accordingly, alternative splicing and expression levels determine the qualitative and quantitative traits of genes, respectively. We found that human-mouse orthologs have certain similarities in overall expression but substantial discrepancies exist in early embryos. The orthologous expression similarities in distinct embryonic stages were different with the highest being in the morula stage. After four cycles of cell division, the zygote divides into the morula containing totipotent cells, and then the morula differentiates into the blastocyst. Different expression similarities in disparate embryonic stages could account for cell division divergences between humans and mice. However, the limitations of single-cell RNA-Seq technology may account for the expression similarity in the morula being higher than that of the former stages, as the loss of mRNA in sample preparation and other sources of technical noise for single-cell RNA-Seq could significantly affect the expression profiling of different embryonic stages (Shapiro et al., 2013). The expression similarity calculation did not take into account those human-mouse orthologs with only one member expressed and this set of orthologs was often ignored by prior studies, despite having a significant impact on expression divergence. Ignoring this would lead to observed higher expression similarities between human-mouse orthologs. We uncovered a remarkable number of human-mouse orthologs of this type in each embryonic stage, suggesting that these orthologs have expression specificity in corresponding human and mouse embryos. Many human and mouse orthologous TFs also exhibited expression variances and some were associated with the regulation of embryonic development. Discordant TF expression could influence the expression of target genes, and further contribute to the divergent expression profiles between human-mouse orthologs in different embryonic stages.

We revealed great variation in differential expression between distinct embryonic stages; this has rarely been previously studied. In the majority of cases, one member of a pair had significant expression changes while its counterpart did not. In addition, some human-mouse orthologs are involved in biological processes related to embryonic development, but their two members were differentially expressed in an opposite trend. Only a minority of human-mouse orthologs had concordant expression changes. Notably, the expression discrepancies of human-mouse orthologs increased with embryonic stage. Consequently, those differences of alternative splicing patterns and expression profiles between human-mouse orthologs could cause distinct orthologous functional dynamics (Keren et al., 2010; Roux and Robinson-Rechavi, 2011), which might associate with embryonic developmental discrepancies between humans and mice.

We uncovered notable differences in human-mouse orthologs in terms of method of alternative splicing, GC content, and expression profiles during embryonic development. To correctly use mouse genes to infer the function of human genes, more investigations with advanced technologies and methodologies are needed to further explore the similarities and discrepancies between human-mouse orthologs. RNA-Seq technologies including single-cell RNA sequencing are evolving fast, and sequencing accuracy, output, and read length will be greatly improved with innovative technology (McGettigan, 2013; Ozsolak and

Milos, 2011). These advances will be of further benefit for orthologous expression comparison at the transcriptomic level. In addition, the continuous improvement of tandem mass spectrometry (MS/MS) for the protein identification, quantification, and modification analyses (Angel et al., 2012; Domon and Aebersold, 2006; Nagaraj et al., 2011) could identify orthologous conservation and divergence at the protein level. Accordingly, it is anticipated that more developmental stages will be compared at both the transcriptomic and proteomic levels, for more comprehensive comparisons of orthologous expression profiles.

METHODS

Public data used in the study

We obtained the human and mouse orthologous genes from Ensembl BioMart (Kinsella et al., 2011) and only used the one-to-one human-mouse orthologous pairs for further analysis. We also downloaded the Ensembl human and mouse gene annotation files (GTF format of Ensembl version 72) from Ensembl (Flicek et al., 2013). For comparing orthologous human-mouse genes, we also obtained the GC content, alternative splicing modes and identities of human-mouse orthologous pairs using Ensembl BioMart. To investigate the orthologous human-mouse TF (transcription factor) genes, we also downloaded the human and mouse TFs from TFClass (Wingender et al., 2013) and TFdb (Kanamori et al., 2004) databases, respectively. Those TFs of human and mouse that are not orthologous pairs were excluded.

Single-cell RNA-Seq data of human and mouse embryos

The single-cell RNA-Seq datasets of human and mouse embryos were downloaded from NCBI Gene Expression Omnibus (GEO) with the accession number GSE44183 (Xue et al., 2013). Because of the absence of mouse zygote RNA-Seq data, the RNA sequencing data of human zygotes were not used in this study. Six groups of RNA-Seq data from different stages of human and mouse embryos were analyzed, including oocyte, pronucleus, 2-cell, 4-cell, and 8-cell and morula. These RNA-Seq data were sequenced using the Illumina HiSeq2000 platform and the related information can be found in Table S2 in Supporting Information.

Gene and isoform expression quantification

To quantify the expression of human and mouse genes, we first aligned human and mouse RNA-Seq data to the human reference genome (GRCh37/hg19) and mouse reference genome (GRCm38), respectively, using TopHat2 (version 2.0.9) (Kim et al., 2013). Parameters of $-r$=0 were employed for TopHat2 according to the cDNA fragment length selection of human and mouse embryos; other parameters were in default. Next, we estimated the gene and isoform expression of human and mouse genes in each embryonic stage using Cufflinks (version 2.1.1) (Trapnell et al., 2010), based on the mapping results from TopHat2. Human and mouse gene annotation files, in GTF format, and the parameters of $-b$ for bias correction and $-u$ for multiple mapped reads correction, were employed by Cufflinks in expression quantification. We only enabled Cufflinks to estimate the expression of genes and isoforms in the human and mouse reference GTF annotation files and did not set the argument of assemble novel transcripts for Cufflinks.

Calculation of *Z*-score and Spearman's correlation coefficient

We used methods similar to that of a previous study (Chen and Zhang, 2012) to calculate the *Z*-score of human-mouse orthologs to normalize gene expression levels between human and mouse. Specifically, we first transformed the FPKM values estimated by Cufflinks into $\log_e$(FPKM) and then calculated the *Z*-score ($Z=\frac{x-\mu}{\sigma}$) for each orthologous gene. Only those human-mouse orthologs with both members having FPKM values >0.1 in a certain stage were included. To study the relationship between expression similarity and gene sequence identity, we divided the human-mouse orthologs into different groups according to the gene sequence identity interval and separately computed the Spearman's correlation coefficient. The Spearman's correlation coefficient of human-mouse orthologs in each embryonic stage was calculated based on the *Z*-score.

Differential expression calling

To analyze expression changes for human-mouse orthologs between different stages of embryos, we separately called the differentially expressed human and mouse genes using Cuffdiff (version 2.1.1) (Trapnell et al., 2013), with parameters of $-b$ and $-u$ for mapping corrections enabled. We divided the human and mouse embryos into four groups for comparison: oocyte against pronucleus, 2-cell against 4-cell, 4-cell against 8-cell and 8-cell against morula. After differential expression calling, we compared the detected differentially expressed genes from human-mouse orthologous pairs group by group. We also examined the continuous expression changes for these orthologous human and mouse genes across these four groups.

Functional enrichment and gene ontology analyses

We performed functional annotation for orthologous human and mouse genes with DAVID (Huang da et al., 2009) and also referenced the GO (gene ontology) terms extracted from Ensembl BioMart. We uploaded those selected genes to the DAVID bioinformatics resources 6.7 and then employed the functional annotation tools to conduct gene function enrichment analysis. For the selection of GO terms and functional annotation clustering, a Fisher Exact *P*-value of

0.05 was chosen as the cutoff.

Author contributions *Geng Chen, Yongxiang Zhao, Meixia Zhang and Tieliu Shi conceived and designed the study; Geng Chen, Jiwei Chen, Jianmin Yang, Long Chen, Xiongfei Qu, and Caiping Shi performed analyses. Geng Chen, Leming Shi, Weida Tong, and Tieliu Shi wrote the manuscript. Geng Chen, Meixia Zhang and Tieliu Shi revised the manuscript. All the authors read and approved the final manuscript.*

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We would like to thank Li Zhang and Rong Cheng from East China Normal University for their helpful discussions. We are also grateful to the Supercomputer Center of East China Normal University. This work was supported by the China Human Proteomics Project (2014DFB30010), the National High Technology Research and Development Program of China (2015AA020104,), the National Natural Science Foundation of China (31071162), and the Graduate School of East China Normal University.*

Angel, T.E., Aryal, U.K., Hengel, S.M., Baker, E.S., Kelly, R.T., Robinson, E.W., and Smith, R.D. (2012). Mass spectrometry-based proteomics: existing capabilities and future directions. Chem Soc Rev 41, 3912–3928.

Blair, J.E., and Hedges, S.B. (2005). Molecular phylogeny and divergence times of deuterostome animals. Mol Biol Evol 22, 2275–2284.

Chen, X., and Zhang, J. (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8, e1002784.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Dolinski, K., and Botstein, D. (2007). Orthology and functional conservation in eukaryotes. Annu Rev Genet 41, 465–507.

Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. Science 312, 212–217.

Elso, C., Lu, X., Morrison, S., Tarver, A., Thompson, H., Thurkow, H., Yamada, N.A., and Stubbs, L. (2008). Germline translocations in mice: unique tools for analyzing gene function and long-distance regulatory mechanisms. J Natl Cancer Inst Monogr, 91–95.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S.M. (2013). Ensembl 2013. Nucleic Acids Res 41, D48–55.

Gabaldon, T., and Koonin, E.V. (2013). Functional and evolutionary implications of gene orthology. Nat Rev Genet 14, 360–366.

Gharib, W.H., and Robinson-Rechavi, M. (2011). When orthologs diverge between human and mouse. Brief Bioinform 12, 436–441.

Ginis, I., Luo, Y.Q., Miura, T., Thies, S., Brandenberger, R., Gerecht-Nir, S., Amit, M., Hoke, A., Carpenter, M.K., Itskovitz-Eldor, J., and Rao, M.S. (2004). Differences between human and mouse embryonic stem cells. Dev Biol 269, 360–380.

Grosso, A.R., Gomes, A.Q., Barbosa-Morais, N.L., Caldeira, S., Thorne, N.P., Grech, G., von Lindern, M., and Carmo-Fonseca, M. (2008). Tissue-specific splicing factor gene expression signatures. Nucleic Acids Res 36, 4823–4832.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57.

Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., and Suzuki, H. (2004). A genome-wide and nonredundant mouse transcription factor database. Biochem Biophys Res Commun 322, 787–793.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11, 345–355.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011, bar030.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309–338.

Liao, B.Y., and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol 23, 530–540.

Liu, H., Chen, C.H., Espinoza-Lewis, R.A., Jiao, Z., Sheu, I., Hu, X., Lin, M., Zhang, Y., and Chen, Y. (2011). Functional redundancy between human *SHOX* and mouse *Shox2* genes in the regulation of sinoatrial node formation and pacemaking function. J Biol Chem 286, 17029–17038.

Liu, Z., Miner, J.J., Yago, T., Yao, L., Lupu, F., Xia, L., and McEver, R.P. (2010). Differential regulation of human and murine P-selectin expression and function *in vivo*. J Exp Med 207, 2975–2987.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18, 1509–1517.

McGettigan, P.A. (2013). Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17, 4–11.

Mouse Genome Sequencing, C., Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., and Lander, E.S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 7, 548.

Nehrt, N.L., Clark, W.T., Radivojac, P., and Hahn, M.W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7, e1002073.

Nurtdinov, R.N., Artamonova, II, Mironov, A.A., and Gelfand, M.S. (2003). Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12, 1313–1320.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature Rev Genet 12, 87–98.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L.C., Dahmane, N., and Davuluri, R.V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. Genome Res 21, 1260–1272.

Pereira, V., Waxman, D., and Eyre-Walker, A. (2009). A problem with the correlation coefficient as a measure of gene expression divergence. Genetics 183, 1597–1600.

Piasecka, B., Robinson-Rechavi, M., and Bergmann, S. (2012). Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. Bioinformatics 28, 1865–1872.

Qian, W., Liao, B.Y., Chang, A.Y., and Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet 26, 425–430.

Ramskold, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P., and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30, 777–782.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 14, 656–664.

Roux, J., and Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res 21, 357–363.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14, 618–630.

Studer, R.A., and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? Trends Genet 25, 210–216.

Takeda, J., Suzuki, Y., Sakate, R., Sato, Y., Seki, M., Irie, T., Takeuchi, N., Ueda, T., Nakao, M., Sugano, S., Gojobori, T., and Imanishi, T. (2008). Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. Nucleic Acids Res 36, 6386–6395.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515.

van der Vegt, B., de Bock, G.H., Hollema, H., and Wesseling, J. (2009). Microarray methods to identify factors determining breast cancer progression: potentials, limitations, and challenges. Crit Rev Oncol Hematol 70, 1–11.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63.

Wingender, E., Schoeps, T., and Donitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res 41, D165–170.

Xing, Y., Ouyang, Z., Kapur, K., Scott, M.P., and Wong, W.H. (2007). Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol Biol Evol 24, 1283–1285.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J.Y., Horvath, S., and Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597.

Yashiro, K., Saijoh, Y., Sakuma, R., Tada, M., Tomita, N., Amano, K., Matsuda, Y., Monden, M., Okada, S., and Hamada, H. (2000). Distinct transcriptional regulation and phylogenetic divergence of human *LEFTY* genes. Genes Cells 5, 343–357.

Yeo, G.W. (2005). Splicing regulators: targets and drugs. Genome Biol 6, 240.

SUPPORTING INFORMATION

Figure S1	Isoform comparison for human-mouse ortholgos based on the Ensembl annotation. The isoform counts for orthologous mouse genes were drawn in descending order accompanied with their paired orthologous human isoform counts.
Table S1	Extreme examples where human genes hold large number of isoforms (>=20) compare to mouse (<=5)
Table S2	The single-cell RNA-Seq data used in this study
Table S3	Functional annotation of those human and mouse specific orthologs
Table S4	Human and mouse orthologous splicing regulators
Table S5	Functional annotation of differentially expressed orthologous genes
Table S6	Embryonic development associated human-mouse orthologs that their two members with reverse expression changes.

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.