Mol Genet Genomics
DOI 10.1007/s00438-015-1140-5

ORIGINAL ARTICLE

Correlated expression of retrocopies and parental genes in zebrafish

Zaixuan Zhong[1,3] · Liandong Yang[1,3] · Yong E. Zhang[2,3] · Yu Xue[4] · Shunping He[1]

Received: 12 May 2015 / Accepted: 27 October 2015

Abstract Previous studies of the function and evolution of retrocopies in plants, Drosophila and non-mammalian chordates provided new insights into the origin of novel genes. However, little is known about retrocopies and their parental genes in teleosts, and it remains obscure whether there is any correlation between them. The present study aimed to characterize the spatial and temporal expression profiles of retrogenes and their parental genes based on RNA-Seq data from *Danio rerio* embryos and tissues from adult. Using a modified pipeline, 306 retrocopies were identified in the zebrafish genome, most of which exhibited ancient retroposition, and 76 of these showed a Ks < 2.0. Expression of a retrocopy is generally expected to present no correlation with its parental gene, as regulatory regions are not part of the retroposition event. Here, this assumption was tested based on RNA-Seq data from eight stages and thirteen tissue types of zebrafish. However, the result suggested that retrocopies displayed correlated expression with their parental genes. The level of correlation was found to decrease during embryogenesis, but to increase slightly within a tissue using Ks as the proxy for the divergence time. Tissue specificity was also observed: retrocopies were found to be expressed at a more specific level compared with their parental genes. Unlike Drosophila, which has sex chromosomes, zebrafish do not show testis-biased expression. Our study elaborated temporal and spatial patterns of expression of retrocopies in zebrafish, examined the correlation between retrocopies and parental genes and analyzed potential source of regulated elements of retrocopies, which lay a foundation for further functional study of retrocopies.

Keywords Retroposition · Expression pattern · Transcriptome · Specificity

Communicated by J. Cerdá.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-015-1140-5) contains supplementary material, which is available to authorized users.

✉ Shunping He
clad@ihb.ac.cn

Zaixuan Zhong
zhongzaixuan0429@126.com

Liandong Yang
yangliandong1987@163.com

Yong E. Zhang
zhangyong@ioz.ac.cn

Yu Xue
xueyu@hust.edu.cn

[1] The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, Hubei, People's Republic of China

[2] Key Laboratory of the Zoological Systematic and Evolution & State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Beijing 100000, People's Republic of China

[3] University of Chinese Academy of Sciences, Beijing 100039, People's Republic of China

[4] Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, People's Republic of China

Published online: 11 November 2015

Introduction

Retrocopies are intronless gene copies of reverse-transcribed mRNA intermediates, which have been described since the early 1980s. However, these copies have long been dismissed a priori as dead on arrival (Jeffs and Ashburner 1991; Petrov et al. 1996; Zhang et al. 2003, 2004). They have been classified as pseudogenes because of their lack of regulatory elements and the presence of frameshifts and premature stop codons (Mighell et al. 2000). A large number of functional retrogenes and their expression patterns have been systematically studied, mainly in mammals, fruit flies, and rice, (Betran et al. 2002; Makova and Li 2003; Vinckenbosch et al. 2006; Bai et al. 2007; Okamura and Nakai 2008; Kaessmann et al. 2009; Sakai et al. 2011). Other than these reports, there is no systemic study of retrocopies in teleost.

Because a newly created retrocopy usually contains the untranslated and coding regions of the parental gene, but not its promoter, expressed retrocopies generally require regulatory elements for transcription. Four mechanisms underlying the acquisition of such regulatory elements have been discovered in recent years (Kaessmann et al. 2009). The first is co-retroposition, or the inheritance of promoters and enhancers directly from parental genes. This process only occurs when there is leaky transcription from a start site upstream of the parental gene, and it provides the retrogene with a regulatory region similar to that of the source gene (Okamura and Nakai 2008). The second is the use of nearby regulatory elements. For example, regulatory elements might be inserted into introns of host genes and expressed as fusion transcripts (Long and Langley 1993). The third mechanism is the acquisition of de novo regulatory elements (Yamashita et al. 2005), and the final mechanism is integration into chromosomal domains that favor transcription because of open chromatin, which may create an environment that facilitates transcriptional activation (Gilbert et al. 2004). The divergence in the expression of retrocopy pairs might be predicted based on different modes of acquisition of regulatory elements.

Ohno proposed that differences in expression are the first steps in functional divergence between duplicate genes (Ohno 1970). In this study, we examined temporal and spatial patterns of the expression of retrocopies in zebrafish and investigated the divergence in expression between retrocopies and their parental genes. First, we identified retrocopies through genomic surveys using a strict pipeline to ensure that each retrocopy was correctly paired with its parental gene. Then, through systematic analysis of RNA-Seq data from eight stages and 13 tissues, we compared the expression between retrocopies and parental genes. Finally, we found that the expression pattern was surprisingly conserved between retrocopy pairs, especially at different stages. Our study reported first time of the expression profiles, evolutionary age, and selective pressure of retrocopies in zebrafish.

Materials and methods

Retrocopy detection

To identify retrocopies in zebrafish, we adapted an approach that was previously used in humans and modified the discovery procedures (Marques et al. 2005). The genome and 42,157-peptide sequence of zebrafish were obtained from Ensembl (http://www.ensembl.org/). First, a TBLASTN search was performed using whole peptide sequences as queries against genome sequences (Altschul et al. 1997). Second, the TBLASTN results were analyzed using customized Perl scripts, and the DNA sequences that most closely matched the proteins were selected and aligned via Genewise (Birney and Durbin 1997). Using the alignment with the corresponding protein sequences, multi-exon sequences were selected for subsequent analysis. Combining only nearby matches (distance <40 bp) that were unlikely to be separated by introns, adjacent homology matches were merged in a series of parsing steps using customized Perl scripts. We employed the criteria that the query and merged target sequences aligned with one another for more than 50 amino acids and exhibited an amino acid identity of more than 30 %. Next, similarity searches were performed on the merged sequences against multi-exon proteins using FASTA. The best hits were selected for use as candidate proteins (WR 1990). If the merged sequences overlapped and at least three exons of the parental gene and the distance between the two adjacent exons was greater than 70 bp, the sequence was considered to be a retrocopy. For this purpose, 70 bp was set as a threshold based on the following considerations: (1) The majority of introns in the zebrafish genome are larger than 70 bp, which does not include small gaps in parental genes mistakenly annotated as introns (Yeo et al. 2004). (2) At least two introns had been lost by each parental gene, as indicated by the fact that 70 bp is larger than the gap size (40 bp) used in the merging step. To ensure the absence of introns from the retrocopies, the merged sequences with their 10,000 bp flanking regions were compared with those of their candidate parental proteins using Genewise, with a cutoff score of 35. Finally, the absence of introns was checked manually.

Age of retrocopies

The timing of the retroposition events was determined by calculating coding sequence divergence at synonymous sites

(Ks) between each retrocopy and its source gene. Pairwise Ks statistics were estimated using YN00 in PAML4 (Yang 2007). Under current conditions, among the fish included in our analyses, silver carp (*Hypophthalmichthys molitrix*) (data unpublished) was found to exhibit the closest genetic relationship with zebrafish. Blastn searches were performed using retrocopies as queries against the genome sequence of silver carp, with an *E* value of 10^{-5}, to assess the distribution of retrocopies in silver carp and estimate their ages.

Analyses of selective pressure

The Codeml program in PAML 4 was used to compare the selective pressure experienced by retrocopies and their parental genes (Yang 2007). A species tree (((stickleback ((fugu, tetraodon) tilapia)) (platyfish, medaka)) zebrafish), was employed as a guide tree for analysis (Li et al. 2007; Near et al. 2013). The orthologous genes of retrocopies and their parental genes were downloaded from Ensembl (http://asia.ensembl.org/index.html) and aligned using Clustal X, respectively. Orthologous genes may be absent from some species. The tree shows the differences for each gene, which were confirmed manually. A branch model in which zebrafish was set as the foreground branch was used to determine whether positive selection had occurred. We corrected the multiple testing by applying the false discovery rate method (FDR). Genes were considered to have experienced significant purifying selection if the FDR-adjusted *P* value is less than 0.05 and $\omega < 1$.

Library preparation and high-throughput sequencing for testis tissue

Because data were available in the NCBI database for eight stages and nine tissue types, considering the integrity of the data analysis, RNA-Seq of testis tissue from adult zebrafish was performed. Wild-type zebrafish from an AB background were maintained in the zebrafish facilities of the Institute of Hydrobiology, Chinese Academy of Sciences. An adult testis was collected for RNA extraction. Total RNA extraction was performed using TRIZOL reagents from Takara, following the manufacturer's instructions. Total RNA contents were measured using a NanoDrop 2000. Then, 5 μg of total RNA was employed for mRNA isolation with oligo (dT) magnetic beads and fragmented into small pieces in fragmentation buffer at 70 °C for 4 min before cDNA synthesis. First-strand cDNA was synthesized from the cleaved RNA using SuperScript II reverse transcriptase and random hexamer primers, followed by second-strand cDNA synthesis. Finally, the paired-end cDNA library was prepared according to the Illumina's protocols and sequenced for 101 bp at both ends using the Illumina HiSeq™ 2000 platform.

Expression analyses using RNA-Seq

Compared with ESTs and full-length cDNA, RNA-Seq has been found to be more accurate in the quantification of gene expression levels. RNA-Seq data were obtained from eight consecutive developmental stages in zebrafish: the 2–4-cell, 1000-cell (3 hpf), dome (4.5 hpf), shield (6 hpf), bud (10 hpf), 28 hpf (hours post fertilization), 2d, and 5d stages; these data were downloaded from NCBI under accession code GSE32898 (Pauli et al. 2012). RNA-Seq data were also collected from five zebrafish tissues: adult zebrafish ovary, male adult zebrafish head, female adult zebrafish head, whole male adult zebrafish without the head or testis and whole female adult zebrafish without the head or ovary; these data were downloaded from NCBI under accession code ERP000016. Data for heart, brain the kidney and swim bladder were downloaded from NCBI under accession code ERP000447 (Collins et al. 2012). Datasets for the heart, brain, blood, liver and muscle were downloaded from NCBI under accession number (SRA) PRJNA207719 (Kaushik et al. 2013). Since several datasets of the same tissue were available [ERR35545 (Brain) and ERR35546 (Heart) from ERP000447, SRR891511 (Brain) and SRR891495 (Heart) from PRJNA207719], we used the RSeQC-2.5(Wang et al. 2012) to evaluate the quality (Supplementary Fig. 2). The biological replicates were analyzed separately with the same method. Using Tophat 2.0.4, all the obtained RNA-Seq reads and testis data were mapped to the genome using the ultra high-throughput short read aligner Bowtie (Trapnell et al. 2009, 2012; Kim and Salzberg 2011), in which only reads with a length longer than 48 and multihits lower than three were retained. We only allowed 2 bp mismatches in a read. Then, Trinity v2.0.5 (Grabherr et al. 2011) was used to calculate fragments per kilobase of transcript sequence per millions base pairs sequenced (FPKM) for each retrocopy and parental gene. To discriminate the transcription between the parental gene and retrocopy, we used SAM's NH:i:1 flag to ensure the reads mapped to a unique location on the genome. Considering that the paired-end length of RNA-seq read was 152 bp and only 2 bp mismatches were allowed in a read, we thought 5 bp difference in 152 bp was sufficient to differentiate parental gene from retrocopy when mapping reads. We analyzed differences between parental genes and retrocopies using a sliding window of 152 bp with a 1 bp step size implemented via a Perl script. Each pair of parental genes and retrocopies showed at least 5 bp difference within 152 bp.

We investigated the number of RNA-seq reads mapping to intergenic regions to determine the background level of expression. We selected intergenic regions >5 kb in length. To ensure these were not protein-coding regions, we using BlastX search against the non-redundant protein database

(nr) of the National Center for Biotechnology Information, with a cutoff E value of 10^{-8}. Then, we calculated the number of mapping RNA-seq reads for intergenic regions, which was shown in Supplementary Fig. 3. Most (cumulative percent = 95.8 %) of the intergenetic region FPKM values were less than 0.42, which we set as the background level.

To evaluate retrocopy specificity and source genes, we calculated $H(g)$, the Shannon entropy, which is expressed in bits of the expression the vector of gene g. This practice is based on FPKM. The specificity score was defined as $1 - H(g)/\log2(N)$, where N represents the number of points in time or the types of tissue (Pauli et al. 2012). The Pearson product moment correlation coefficient (R) was calculated using FPKM values from different stages and tissue types to compare expression patterns between each retrocopy and its parental gene.

$$H(g) = -\sum_{i=1}^{N} P_i \log_2 P_i \qquad P_i = g_i / g_{\text{sum}}$$

where g is the gene name, g_i is the FPKM for the ith tissue, and g_{sum} is the sum of N tissues.

RT-PCR experiment

Total RNA was extracted from Zebrafish samples (AB strain, Liver) using SV Total RNA Isolation Kit (Promega, Madison, WI, USA). To avoid genomic contamination, we treated Total RNA with DNase I (Promega) to remove genomic DNA. We synthesized single-strand cDNA was using Reverse-Transcription System (Promega). We amplified β-actin to ensure cDNA was successfully synthesized. Unique primers were used to amplify target sequences, and mRNAs without being reverse-transcribed were used as negative controls. The Purified products were sequenced on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, California, United States) and the resulting sequences were deposited in GenBank [GenBank: KP324775 to KP324787].

Shared motif analysis

Regulatory sequences are often located 5′ of proteins, comprising the 5′ UTRs, promoter regions, and other regulatory elements, such as enhancers. A shared motif is defined as a region of high local similarity between two DNA sequences, regardless of their order, orientation, or spacing (Castillo-Davis et al. 2004). Accordingly, d_{SM}, defined here as the relative number of both sequences that did not include a region of significant local similarity, was used to quantify the differences in regulatory elements. An algorithm created by Castillo-Davis et al. (2004) was used to perform the calculations (Waterman and Eggert 1987).

Statistical analysis

SPSS 20.0 software for windows was used for the most statistical analysis. Mann–Whitney U test was performed to analyze the significant difference ($P < 0.05$) of expression level between retrocopies and parental genes. The correlation between retrocopies and parental genes was analyzed by the Pearson correlation test. The One-Sample T Test was performed to compare ω value with 1. R 3.1.2 for Linux was used for Chi squared test to analyze purifying selection of retrocopies and parental genes and two-sample equality test to analyze the distribution of genes belonging to zygotic supercluster in retrocopies.

Results

Number, structure, age distribution, and functional implications of retrocopies

Of the 42,157 protein sequences obtained in this work, 306 were found to be associated with retrocopies in the zebrafish genome using a refined version of a previously published procedure (Chen et al. 2011). Briefly, all the protein sequences were mapped to the genome with TBLASTN (Altschul et al. 1997), and the 11,144 longest regions mapped with the proteins were retained (Fig. 1). The structure of these 11,144 sequences was obtained using Genewise (Birney and Durbin 1997), and only 9911 sequences with introns larger than 40 bp were retained. We mapped these sequences to the merged sequences extracted from the TBLASTN results with Fasta. We retained the 3050 results that exhibited the best scores but no overlap with the position of the protein in the genome, which were regarded as candidate parental genes. Finally, if the sequence of the merged genes overlapped two exons of the parental gene and the distance between the two exons was larger than 70 bp, we confirmed the gene as a candidate retrocopy. In total, 1107 retrocopy candidates were found that satisfied all of our criteria. Considering that these candidates might be duplicates descended from primary retrocopies, we used Genewise to determine whether each lost intron was derived from the parental gene and detected 306 primary retrocopies. We have outlined these analyses in a flow chart format in Fig. 1.

We analyzed the expression patterns of retrocopies showing a Ks value (the number of synonymous substitutions (amino acids) per synonymous site) <2.0, which were relatively young (Supplementary Table 1). None of the retrocopies included in the analysis contained introns, and all

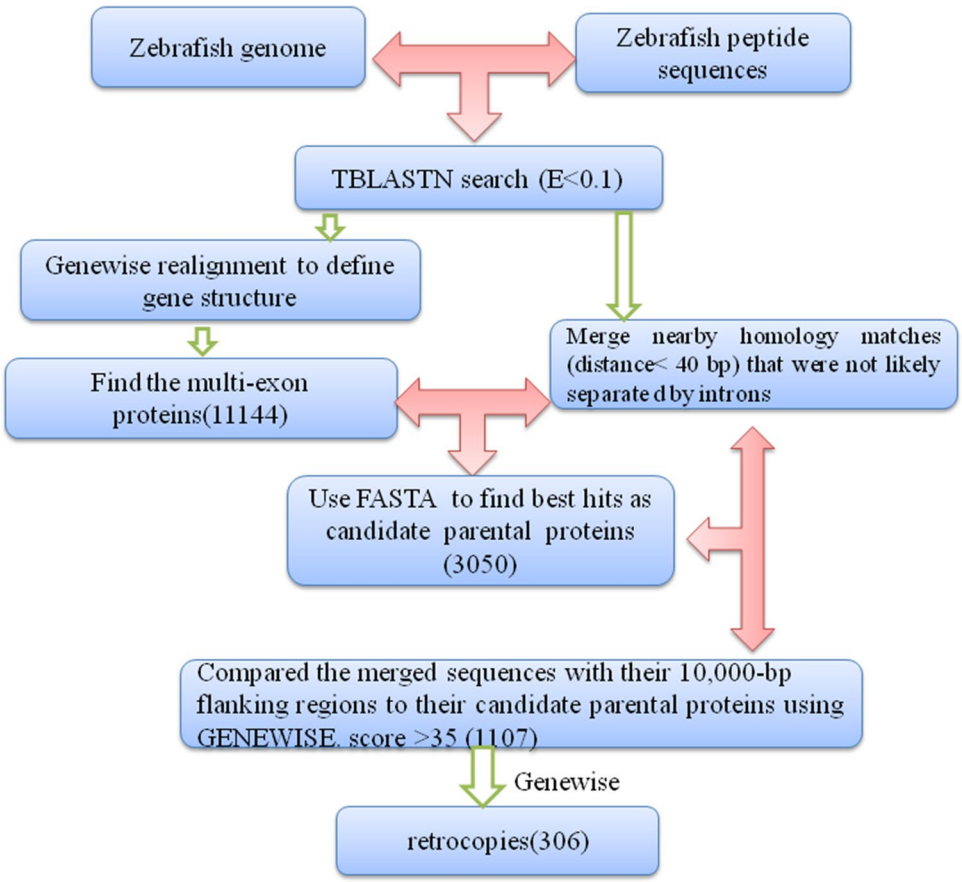

Fig. 1 The pipeline of retrocopies detection. 306 retrocopies were identified in zebrafish genome

had lost at least two introns relative to their putative parental genes. These restrictions ensured that the genes in question had truly emerged via retroposition.

The Ks value between each retrocopy and its parental gene was estimated, and the distribution is shown in Fig. 2. Only a few pairs exhibited Ks values below 1.0, which may indicate that these retrocopies were not as young as had been expected. We verified this result by performing searches of the retrocopies against species with a relatively close genetic relationship with zebrafish. The 76 retrocopies were compared with the genome sequence of *Hypophthalmichthys molitrix* using blastn, and 73 % of the retrocopies showed >50 % nucleotide identity. Another seven retrocopies were not found, and the remainder exhibited little identity.

Despite this result, the 76 retrocopies exhibited structural changes that were not observable in their parental genes. In 11 of these retrocopies, reading frame disruptions were found, such as frameshifts and stop codons, which we defined as retropseudogenes. Strikingly, the mean Ks of these genes was significantly lower than the mean Ks of retrocopies with intact open-reading frames, which were defined as retrogenes (0.23 vs. 1.42, $P = 2.7\text{E-}06$, Mann–Whitney U test), indicating that the retropseudogenes were probably much younger. Among the retrogenes, the

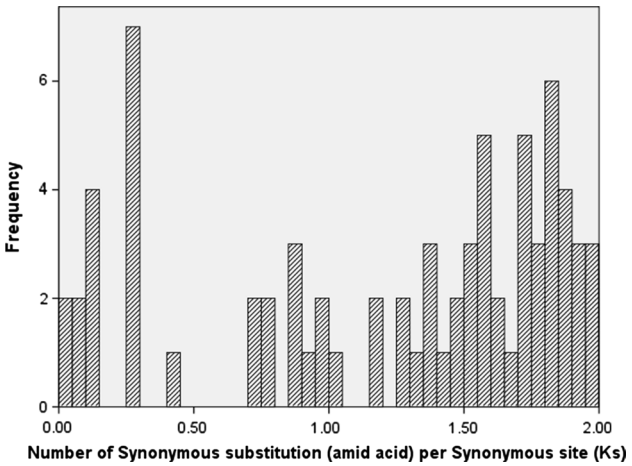

Fig. 2 The distribution of synonymous substitutions per site (k_s) for 76 retrocopies

percentages of retrocopies that formed single-exon genes or one exon of a host gene, or inserted into the introns of host genes or intergenic regions were 44.6, 29.2, 7.7, and 18.5 %, respectively (Supplementary Table 2).

To determine whether the transcription of these retrocopies has any functional implications, the transcription

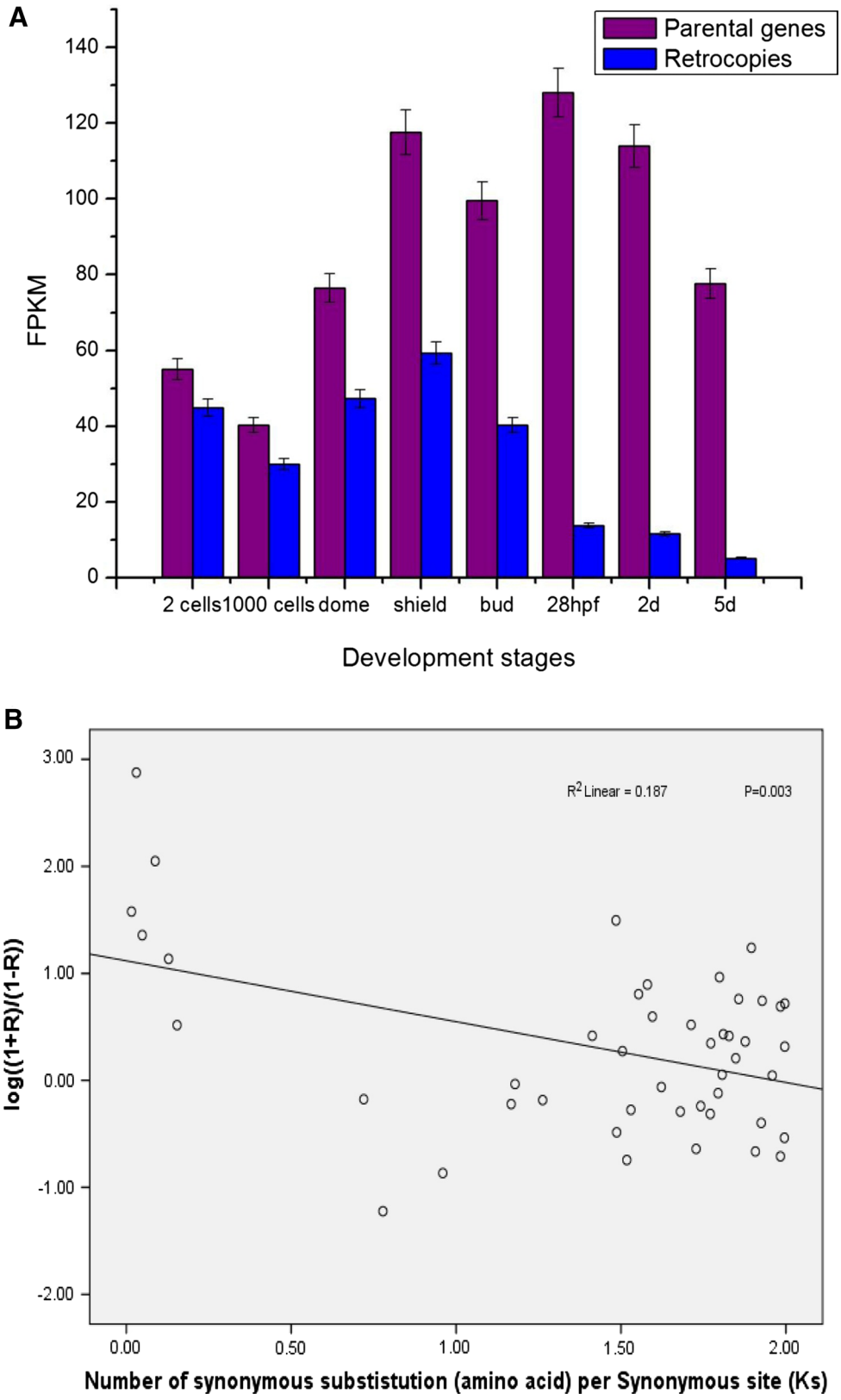

Fig. 3 The FPKM and correlation between retrocopies and their parental genes for stages. **a** Histogram showed FPKM value of retrocopies (*blue bar*) was significant lower than that of parental genes (*brown bar*) ($P < 0.005$, Mann–Whitney Test). **b** Negative correlation between retrocopies and their parental genes was represented by Ks and Y [$\log(1 + R)/(1 - R)$] ($P = 0.003$)

of 65 possibly functional retrocopies with intact ORFs was compared with the transcription of the 11 retropseudogene copies. The mean FPKM of the retrocopies was significantly higher than the mean FPKM of the retropseudogenes in different tissue types and stages (Mann–Whitney U test, Supplementary Table 3). The mean FPKM of

retropseudogene was smaller than the background level in some tissues (Supplementary Table 3). Reverse-transcription polymerase chain reaction (RT-PCR) was used to confirm the expression of these retrocopy genes. Primers were designed for 13 randomly selected retrocopies, and all of these genes were amplified. Information on the primers and the RT-PCR results is presented in Supplementary Table 4 and Supplementary Fig. 1. Given that the most highly transcribed retrocopies in human beings have been identified as functional genes, as indicated by EST analyses, most retrocopies in zebrafish may also be functional (Vinckenbosch et al. 2006).

The ratio of nonsynonymous to synonymous substitutions per site (Ka/Ks) is another indicator of functionality. Here, this ratio was calculated for retrocopy and parental gene pairs. The 11 retropseudogene copies exhibited significantly higher Ka/Ks ($\omega = 0.72$) ($P < 0.001$, Mann–Whitney U test) than the remaining 65 intact retrocopies ($\omega = 0.36$) (Supplementary Table 1). This value was far lower ($P < 0.001$, One-Sample T Test) than the neutral expectation of $\omega = 1.0$, indicating that the intact retrocopies were predominately shaped by purifying selection throughout the evolutionary process.

ω values were estimated separately for retrocopies and their parental genes (48 pairs) among homologous genes using PAML. To determine whether these retrocopies and their parental genes were under purifying, neutral or positive selection, Likelihood ratio testing (LRT) with a χ^2 approximation was also conducted. The results suggested that 24 retrocopies and 18 parental genes had experienced significant purifying selection ($\omega < 1$, $P < 0.05$, Chi squared test, Supplementary Table 6). The mean ω values were 0.21 and 0.17 ($P = 0.75$, Mann–Whitney U test) for the retrocopies and parental genes, respectively.

Retrocopies with detailed annotations were evaluated, and information was retrieved manually from Ensembl (http://asia.ensembl.org/index.html). Only four of these retrocopies (ENSDARG00000089553(Siddiqui et al. 2010), ENSDARG00000037995 (Ye et al. 2010; Peterson et al. 2013), ENSDARG00000079355 (Thisse et al. 2008) and ENSDARG00000011724 (Thisse and Thisse 2004)) have been described in publications.

Temporal expression of newer retrocopies

The dynamics of the temporal expression of retrogenes and their source genes was examined using a subset of the RNA-Seq data, which encompassed eight stages: the 2-cell, 1000-cell, dome, shield, bud, 28 hpf (hours post fertilization), 2d (days), and 5d stages (Fig. 3a). This analysis was limited to 52 gene pairs due to the occurrence of tandem duplication after retroposition events (Hasselmann et al. 2010). The vast majority of the remaining 52 gene pairs (46) showed expression during different stages, indicated by the FPKM higher than the background level 0.42 (Supplementary Table 5). This result provided the basis for a systematic analysis of temporal patterns. Failure to find evidence of expression of these six genes using RNA-Seq data may suggest that they were expressed in other developmental stages and tissues or at very low levels in the analyzed developmental stages. The mean FPKM for the eight stages was calculated, revealing the FPKM of the retrocopies to be significantly ($P =$ 6.40E-4, 6.55E-4, 6.63E-4, 9.71E-4, 1.76E-3, 2.23E-4, 2.22E-4, 3.75E-5, respectively, for 2 cells, 1000cells, dome, shield, bud, 28hpf, 2d, 5d, Mann–Whitney Test) lower than the FPKM of the parental genes. The FPKM of the retrocopies peaked (59.3) at the shield stage, after which it degraded, reaching the lowest level of 5.02 (Fig. 3a). The parental genes did not show any tendency to decrease.

Furthermore, the FPKM values offered a wealth of information for elucidating the correlation of the temporal expression of retrocopies with that of their parental genes. The correlation coefficient (R) was computed between retrocopy pairs (Supplementary Table 5). We considered retrocopies and parental genes showed significantly correlated expression if the FDR-adjusted p value was less than 0.05. Six pairs for which FPKM was zero were eliminated. Among the remaining 46 pairs, 27 (58.7 %) exhibited positive correlation, among which 8 (17.4 %) exhibited a significantly positive correlation ($R > 0.8$, $P < 0.05$, Pearson correlation test), while 2 (4.3 %) showed a significantly negative correlation ($R < -0.7$, $P < 0.05$, Pearson correlation test). To confirm that these retrocopies were expressed in a similar pattern to their source genes, the expression values for two of the most closely correlated pairs were graphed (Fig. 4).

To investigate the relationship between evolutionary time and R, Ks was used as a proxy of the duration of divergence (Makova and Li 2003). The following transformation was used: $Y = \log[(1 + R)/(1 - R)]$ (Gu et al. 2002; Makova and Li 2003; Li et al. 2009; Sakai et al. 2011). Normal linear regression was performed between each pair of Ks and Y values, revealing a moderate significant negative correlation between these parameters (Fig. 3b; $R = -0.43$, $P = 0.003$). Thus, in the temporal profile, newer retrocopies tended to be expressed more similarly to their parental genes than older retrocopies. In conclusion, the temporal expression of retrocopies and parental genes tended to be positively correlated, but this expression decreased with Ks in an approximately linear fashion during evolution.

Tissue specificity of retrocopies and parental genes

To elucidate the dynamics of the divergence of spatial expression, data from 13 types of tissue were downloaded

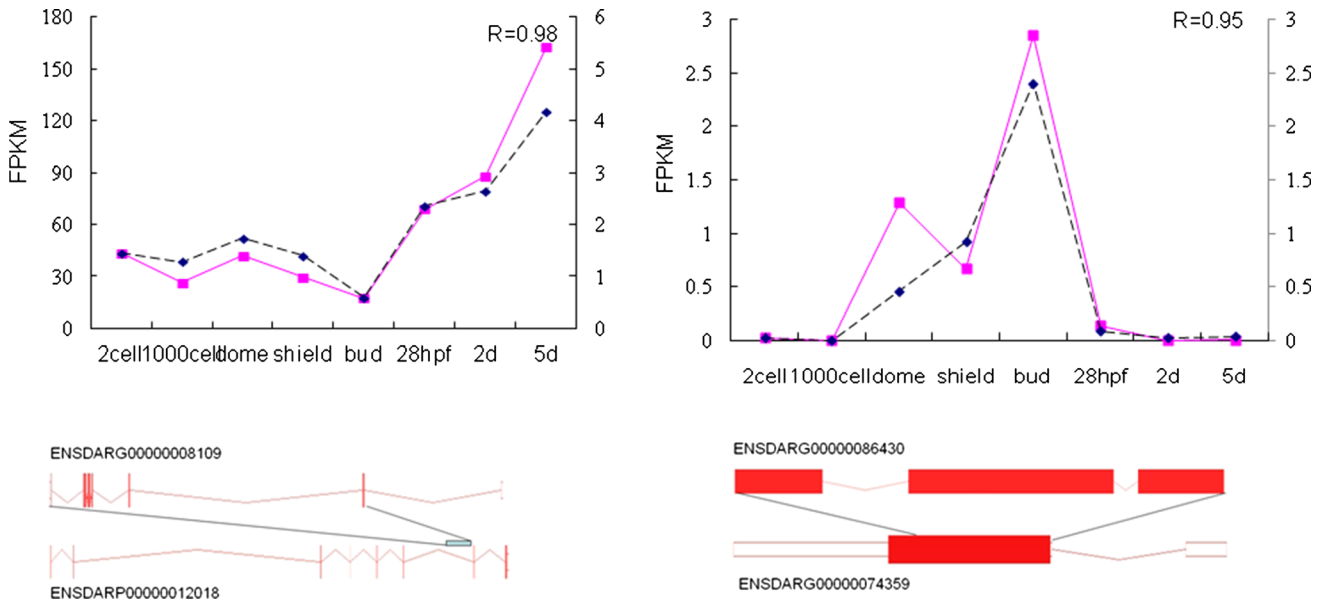

Fig. 4 Graph of expression counts for retrocopies (*dashed lines*) and parental genes (*solid lines*) in eight stages. The values on the *left of axis* represent for parental genes and the *right* for retrocopies. The diagrams under graph represent the structure of host genes (*bottom*) and parental genes (*top*)

and processed. For only two samples with biological replicates freely available, we repeated our analyses and the results of biological replicates were consistent. To streamline our logic, we only used the dataset of SRR891511 (Brain), SRR891495 (Heart) in our manuscript and the result of ERR35545 (Brain), ERR35546 (Heart) is listed in Supplementary File 1.

The Pearson coefficient (R) was calculated for the retrocopies and parental genes (Supplementary Table 5) with controlling a false discovery rate (FDR) of 5 %. 4 exhibited significant positive correlations ($P < 0.05$, Pearson correlation test). Nevertheless, a weak positive correlation ($P = 0.28$) was observed between the transformation $Y = \log [(1 + R)/(1 - R)]$ and Ks (Fig. 5b). The parental genes showed significantly higher mean FPKM values than the retrocopies in the examined tissues (Adjusted P value = 7.15E-4, 3.51E-4, 3.70E-4, 7.01E-4, 3.68E-4, 3.70E-4, 9.84E-5, 3.68E-4, 9.84E-5, 9.84E-5, 3.51E-4, 2.56E-3, 6.07E-4, respectively, Mann–Whitney Test) (Fig. 5a).

However, the retrocopies always tended to exhibit higher tissue specificity than their parental genes. To test this hypothesis, RNA-Seq data were used to as a measurement of spatially specific expression (Yang et al. 2013). A gene was defined as being expressed in a given tissue or organ if its FPKM value was larger than background level 0.42 and at least two reads were mapped to the gene. The mean number of retrocopies that were not expressed in each tissue was significantly greater than the number of parental genes [(20.7 vs. 9.7), $P < 0.001$, Mann–Whitney U test, Table 1). This result may indicate that retrocopies show more specific expression patterns. Another measure of tissue specificity is Shannon entropy (a measure of the breadth of expression, see "Methods" section). The specificity score was calculated (Supplementary Table 7). The mean values of these scores were 0.42 and 0.37 for retrocopies and parental genes, respectively, and the former was significantly higher than the latter ($P = 0.03$, Mann–Whitney U test). Together, all these analyses revealed relatively high spatial specificity to be a trait of the retrocopies in zebrafish.

As testis-specific expression of retrogenes has been demonstrated in humans and fruit flies (Betran and Long 2003; Marques et al. 2005; Vinckenbosch et al. 2006; Kalamegham et al. 2007; Ding et al. 2010), to determine whether there is any significant sex bias in the expression of retrocopies in zebrafish, the mean FPKM values for zebrafish heads, gonads, and bodies without heads or gonads were compared for the two sexes (Fig. 5a). In contrast to what is observed in fruit flies and humans, we did not find any significant difference between the different sexes in zebrafish ($P = 0.82$, $P = 0.92$, and $P = 0.80$, respectively, Mann–Whitney U test, Supplementary Fig. 5).

Two genes are said to exhibit divergent expression in a particular tissue if one gene is expressed in that tissue and the other is not. A retrocopy and its parental gene were considered to exhibit divergent expression if they showed differences in expression in at least one of the tissues studied (Makova and Li 2003). In our data, 35 retrocopies satisfied this criterion, including 10 retrocopies that were expressed

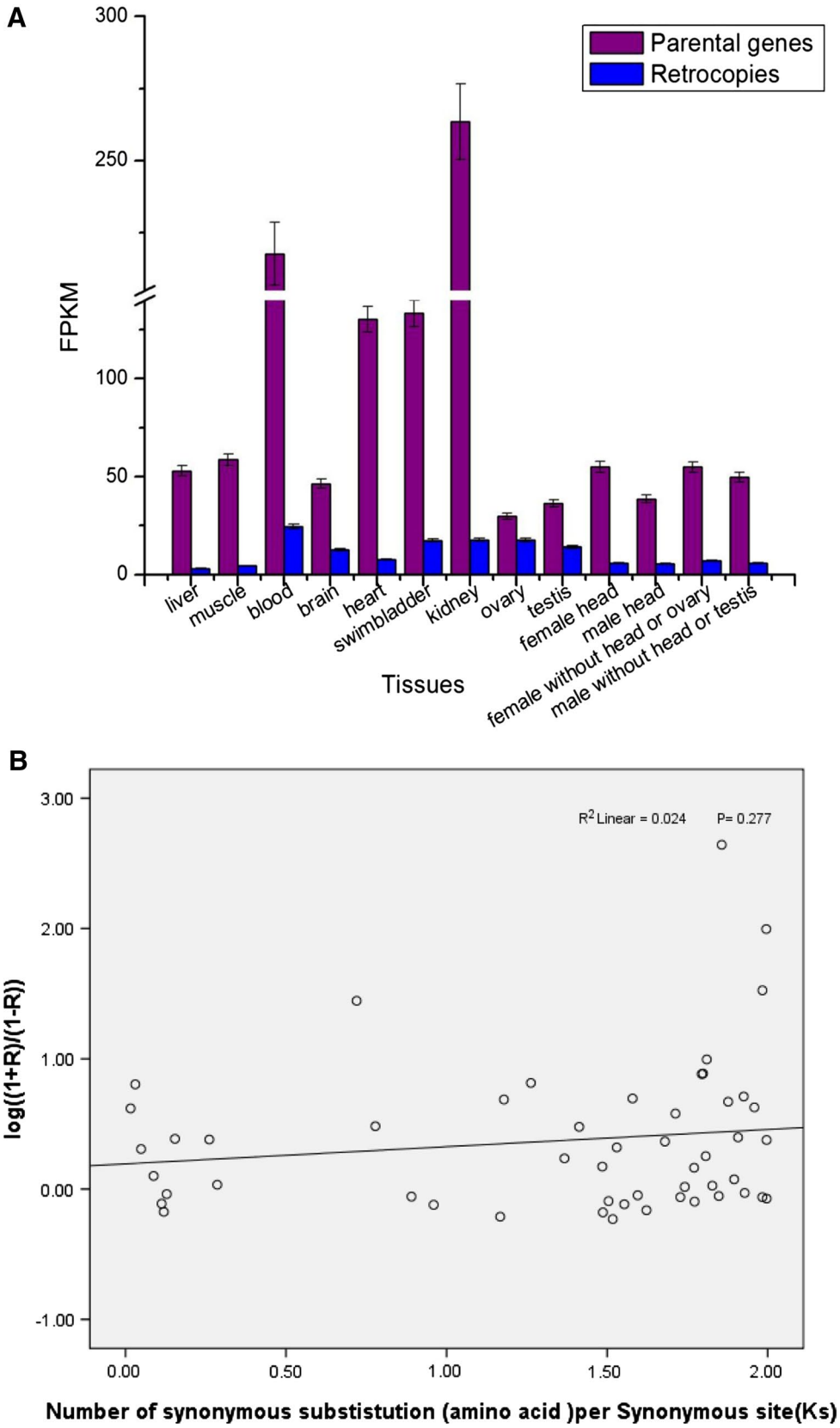

Fig. 5 The FPKM and correlation between retrocopies and their parental genes for tissues. **a** Histogram of FPKM value showed that the expression level retrocopies (*blue bar*) were significantly lower than that of parental genes (*brown bar*) ($P < 0.005$, Mann–Whitney Test). **b** Relationship between retrocopies and their parental genes was represented by Ks and Y [log(1 + R)/(1 − R)]

Table 1 Comparison between retrocopies and their parental genes of the number of expressed genes in different tissues

	Retrogenes		Parental gene	
	No. of unexpressed genes	No. of expressed genes	No. of unexpressed genes	No. of expressed genes
Liver	35	17	18	34
Muscle	26	26	13	39
Blood	22	30	9	43
Brain	18	34	8	44
Heart	23	29	9	43
Swim bladder	13	39	9	43
Kidney	18	34	7	45
Female without head or ovary	21	31	11	41
Male without head or testis	24	28	10	42
Female head	15	37	5	47
Male head	13	39	5	47
Ovary	26	26	14	38
Testis	16	36	8	44
Average	20.7	31.3	9.7	42.3

in a new tissue in which the parental genes showed no expression, whereas 21 retrocopies displayed the opposite condition (Supplementary Table 9). For example, the retrocopy ENSDARP00000056404 was expressed solely in the swim bladder, while its source gene was ubiquitously expressed in 13 tissues. We also calculated the relative divergence between gene pairs as (M1 − M2)/(M1 + M2), where M1 and M2 are the average FPKM for the each of the retrocopies and parental genes, respectively. The relative divergence was negatively correlated with evolutionary age (Ks) (R = −0.34, P = 1.4E-3, Pearson coefficient test).

Promoter region expression and shared motif divergence

To further explore the mechanisms that might underlie the correlations in expression between retrocopies and their parental genes, two possible explanations were proposed.

If promoter regions are co-retroposed with coding sequences, retrocopies and their parental genes may exhibit similar expression patterns (Okamura and Nakai 2008). To confirm such patterns, 35 parental genes that were validated based on flcDNAs (full-length cDNAs) and contained 500-bp upstream sequences that did not overlap with adjacent genes were investigated. Five types of tissue were examined: heart, liver, muscle, brain and blood. Data from these tissues were used to calculate the FPKM for these 500-bp regions. The mean FPKM values of the promoter regions for these five types of tissues were 15.4, 8.3, 10.0, 7.7, and 25.8, respectively, which were significantly lower than the corresponding values for the parental genes [187.8 (P = 0.002), 77.6 (P = 0.03), 84.0 (P = 0.002), 54.5 (P = 0.003) and 313.3 (P = 0.003), Mann–Whitney Test]. Meanwhile, the Pearson coefficient (R) between the FPKM of the promoter region and parental genes was calculated (Supplementary Table 8). Among 32 pairs of promoter regions and parental genes, 21 (65.6 %) showed a significantly positive correlation in expression ($P < 0.05$, Pearson correlation test). The leaky transcription of the promoter regions of the parental genes indicated that the retroposition of promoters was possible (Sakai et al. 2011).

Sequence similarity was detected by examining the upstream sequences between retrocopies and parental genes. Only 11 pairs, in which both the retrocopies and their source genes were validated based on flcDNAs and exhibited 500-bp upstream regions that did not overlap with other genes, were searched. For these gene pairs, the divergence of the 500-bp upstream sequences was investigated using d_{SM}. By definition, d_{SM} is the fraction of both sequences that does not include a region of significant local similarity according to these criteria. For instance, a d_{SM} of 1 indicates an absence of shared motifs, and a d_{SM} of 0 implies a complete sharing of motifs between the sequences. This measure is akin to a distance metric but exhibits a maximum value of 1. Values of $d_{SM} = 1$ are not necessarily equally divergent and should not be compared because they are saturated with sequence differences (Castillo-Davis et al. 2004). In this study, the mean d_{SM} was found to be 0.78, suggesting that the upstream sequences of retrocopy pairs may experience a great deal of divergence. The correlation between d_{SM} and Ks was calculated (Fig. 6). The results showed a moderate correlation in expression between retrocopy pairs (R = 0.610, P = 0.046,

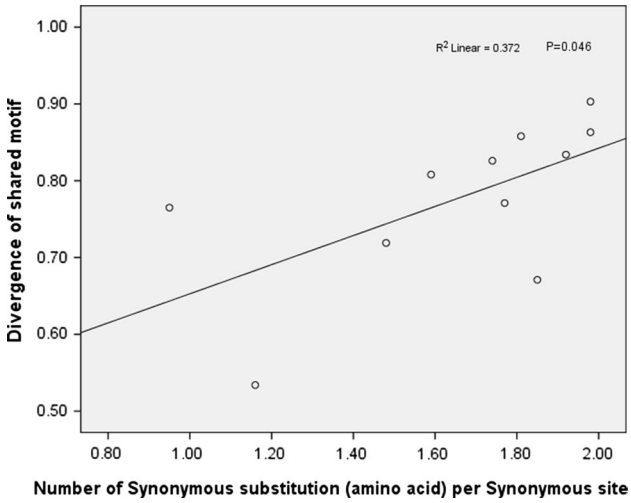

Fig. 6 The relationship between Ks and shared motif divergence. A positive correlation in expression between retrocopy pairs ($R = 0.610$, $P = 0.046$, Pearson correlation test)

Pearson correlation test). Put another way, the older the retrocopy, the more pronounced is the degree of divergence in the shared motif.

Discussion

Retroposition has drawn biologists' attention as a mechanism of gene duplication and a means by which new genes are born (Kaessmann et al. 2009). In particular, some retrogenes identified as new genes have become essential, and lethality occurs when these genes are knocked out (Chen et al. 2010).

Two possible causes of older retrocopies

In the present study, 76 retrocopies were found in the zebrafish genome with a Ks < 2.0 (Fig. 2). In the present work, it was required that the parental genes had lost at least two introns and must have merged sequences with the 10,000-bp flanking regions of the candidate parental proteins. There are far fewer retrocopies in zebrafish than in primates (Marques et al. 2005; Vinckenbosch et al. 2006), which may be due to the different evolutionary fates of retrocopies between mammals and fish.

The Ks values and the differences in blastn results between the retrocopies found in zebrafish and silver carp also confirmed that 73 % of the retrocopies exhibited more than 50 % nucleotide identity. The retrocopies of zebrafish tended to be old, although the conservative approach taken here was biased toward fairly new retrocopies. This result suggested two possibilities, though neither can be proven at this time. First, the retroelements responsible for retroposition are not constant among most of the zebrafish retrocopies produced by LINE1 elements (Chen et al. 2011). It is not clear whether LINE activity decreases, increases, or remains the same in older insertions, though it may be that these insertions are deleted quickly. However, the number of retrocopies exhibits a significant correlation with the number of LINE1 elements (Chen et al. 2011). Second, approximately 320–400Mya, the third genome duplication occurred in the stem lineage of teleost fish (infraclass Teleostei) after divergence from nonteleost ray-finned fish (Amores et al. 1998; Taylor et al. 2003; Van de Peer and Meyer 2003; Hoegg et al. 2004; Jaillon et al. 2004; Meyer and Van de Peer 2005; Nakatani et al. 2007; Amores et al. 2011). In a previous study on retrogenes in zebrafish, Fu concluded that the majority of retrocopies had formed within the past 200 million years under conditions of a synonymous substitution rate of zebrafish genes of 4.13×10^{-9} substitutions per silent site per year (Fu et al. 2010).

As functional retrogenes in mammals, fruit flies, and rice have been systematically discovered and studied, it has become valuable to investigate the function of these retrocopies in fishes. Initially, the evolution of pseudogenes take place under neutral conditions, which means that the associated Ka/Ks values are larger than the ratios obtained for genes subject to functional constraints under purifying selection and smaller than the ratios for genes under positive selection (Torrents et al. 2003). This method has been widely employed in identifying the functionality of genes (Nekrutenko et al. 2002). Suffering from greater selection pressure, intact retrocopies show lower Ka/Ks (ω) values than pseudogenes. To identify the functionality of retrogenes, a stricter criterion should be used: Ka/Ks < 0.5 (Emerson et al. 2004). Using this criterion, 54 retrocopies were found.

Generally, duplicated genes undergo three types of evolutionary fates: subfunctionalization, neofunctionalization, and nonfunctionalization (Kaessmann et al. 2009). A central issue in the evolution of duplicate genes is why so many duplicate genes are retained in a genome, even though the most likely fate for a redundant duplicate is nonfunctionalization (Li et al. 2005).

Temporal and spatial expression patterns of retrocopies

Our present analysis of correlation and divergence between retrocopies and their source genes may shed fresh light on the temporal and spatial expression profiles of these genes. A previous study on the zygotic transition of zebrafish revealed that clusters of genes were characterized by very low expression during pre-MBT stages and showed high abundance from 3.5 hpf or 5.3 hpf, which were referred to as zygotic supercluster, demonstrating the existence of

a mechanism driving development prior to ZGA (Neyfakh 1956; Newport 1982; Korzh 2009). In the present study, we found the level of expression of retrocopies increased from the 1000-cell stage, which is close to the MBT, peaking at the shield stage, close to the post-MBT stage, representing a similar expression pattern to the zygotic supercluster genes before budding (Fig. 3a). However, this situation was not observed among the parental genes. Within 26,206 protein-coding genes (Kettleborough et al. 2013), 1253 genes were included in the zygotic supercluster in zebrafish (Aanes et al. 2011). We found 3 retrocopies belonged to the zygotic supercluster within 76 retrocopies. Two-sample test for equality of proportions with continuity correction was implemented in R 3.1.2 and no significant difference ($P = 0.94$) was found. Our result may indicate that Supercluster actually represented some lineage-specific feature and thus recruited some new retrogenes. During the early stages of development, the expression level of retrocopies tended to decrease, while the expression level of the source genes remained roughly the same. There might be a quantitative rather than a qualitative effect in duplicates (Force et al. 1999; Stoltzfus 1999). Expression may decrease to such a level that both duplicates may be needed to achieve a given function. It was also possible that this result might be a result of the degeneration process. Enzymatic activity was less pronounced for the duplicates compared with the double gene dosage provided by the two gene copies.

We determined the Pearson coefficient (R) to evaluate the correlation between the expression of retrocopies and parental genes and measured the tendencies of Ks and R. Here, Ks was regarded as a more appropriate proxy of the divergence time than Ka, as Ks varies substantially less among genes (Makova and Li 2003), while Ka is strongly affected by selection, which may differ greatly among genes. Ks is less affected by selection, particularly in mammals, in which there is no evidence of strong selection for codons (Urrutia and Hurst 2001). However, Ks is affected by regional variation in the mutation rate within a genome (Lercher et al. 2001; Williams and Hurst 2002). Ks continued to vary among genes, which might partially explain that weak but significant correlation was observed between Ks and Y in our study, as measured based on R. This result suggested that compared with older retrocopies, newer retrocopies showed stronger correlations with parental genes, which has also been validated in rice (Sakai et al. 2011).

Regarding tissue specificity, a classical view is that gene duplication enables duplicates to become specialized in different tissues (Markert 1964; Ohno 1970; Ferris 1979; Force et al. 1999; Gu et al. 2004). One alternative, and not mutually exclusive, hypothesis is that retrocopies might preferentially insert themselves into open, actively transcribed chromatin (Fontanillas et al. 2007). In this study, expression data were systematically screened among different tissues, which collectively revealed a higher specificity among retrocopies than among parental genes (Table 1). The results of this investigation were also evaluated by calculating the Shannon entropy and the number of unexpressed retrocopies. A higher specificity among retrocopies has not been found in rice (Sakai et al. 2011).

Source of regulated elements

Generally, the observation that a significant number of retrocopies have evolved into bona fide genes has raised the issue of how retrocopies can be expressed in their new genomic location. A core promoter and probably other elements, such as enhancers, are necessary for a gene to be expressed at a significant level and in a meaningful way (for example, in tissues in which it can exert a selectively beneficial function) (Kaessmann et al. 2009) (Supplementary Fig. 4). The possible mechanisms for the regulation of retrogenes are as follows: (1) de novo regulatory elements may occur; (2) nearby genes may be used to drive expression (Wang et al. 2000, 2002; Vinckenbosch et al. 2006; Bai et al. 2008); (3) co-retroposition of the 5′ promoter regions of parental genes may occur; and (4) regulatory elements may be embedded within retroposed regions, such as exons and UTRs.

Until recently, although some instances of parental-promoter inheritance had been observed, inheriting parental promoters directly was still considered unlikely (Soares et al. 1985; McCarrey 1987; Shiao et al. 2008). However, a recent study suggested that retrocopies might frequently inherit basic promoters directly from their source genes (Okamura and Nakai 2008). Initially, it was expected that the expression patterns observed in this study would be completely uncorrelated between retrocopies and their source genes. However, both slight and strong correlations were found. This result was attributed to one potential mechanism: the co-retroposition of 5′ promoter regions along with parental coding sequences (Okamura and Nakai 2008). As initially speculated, the FPKM values of upstream sequences were found to be significant lower than the FPKM values of parental genes. By measuring d_{SM}, strong divergence was observed in upstream regions. The older the retrocopies, the more pronounced the promoter divergence, which was consistent with the conclusion that the older retrocopies seemed to exhibit a less pronounced correlation with their parental genes. These results are similar to the results found in *Drosophila* (Bai et al. 2008) and rice (Sakai et al. 2011), in which no clear sequence homology in promoter regions has been detected between retrocopies and their source genes. These observations made it difficult to verify that the cause of the obtained correlation was the retroposition of 5′ regulatory regions from the parental genes.

In the present study, we identified 306 retrocopies. Among these retrocopies, 76 met the Ks threshold of <2.0. We observed correlations between retrocopies and their parental genes, which may be correlated with the promoters. The level of correlation was found to decrease during embryogenesis, but to increase slightly within a tissue using Ks as the proxy for the divergence time. Additionally, we found that the retrocopies formed a ZGA supercluster and showed strong, tissue-specific expression compared with their parental genes. Unlike Drosophila, which has sex chromosomes, zebrafish do not show testis-biased expression. In summary, we systematically elaborated the spatial and temporal expression profiles of retrocopies and their parental genes. The correlation and divergence of expression between retrocopies and parental genes were analyzed, which might facilitate the functional analysis of retrogenes in zebrafish.

Acknowledgments We are thankful to Beide Fu and Ming Zou for their critical comments and suggestions.

Compliance with ethical standards

Funding This study was funded by the grants from Chinese Academy of Sciences (XDB13020100) and National Natural Science Foundation of China (91131014).

Conflict of interest The authors declare that they have no competing interests.

Ethical approval The methods involving animals in this study were conducted in accordance with the Laboratory Animal Management Principles of China. All experimental protocols were approved by the Ethics Committee of the Institute of Hydrobiology, Chinese Academy of Sciences.

Data access The RNA-Seq data have been submitted to the NCBI Sequence Read Archive (SRA) with accession number SRR16957302. The retrocopies sequences have been submitted to Genbank (http://www.ncbi.nlm.nih.gov/genbank/) with accession number from KP324775 to KP324787.

Authors' contributions ZZ developed the algorithm, performed the analyses, and drafted the manuscript. LY participated in algorithm development. YZ participated in the design of the study and data analysis. SH conceived of the study, participated in its design and coordination, and helped to analyze the data. All authors read and approved the final manuscript.

References

Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SG, Lim AY, Hajan HS, Collas P, Bourque G, Gong Z, Korzh V, Alestrom P, Mathavan S (2011) Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res 21(8):1328–1338

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282(5394):1711–1714

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188(4):799–808

Bai Y, Casola C, Feschotte C, Betran E (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol 8(1):R11

Bai Y, Casola C, Betran E (2008) Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genom 9:241

Betran E, Long M (2003) Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics 164(3):977–988

Betran E, Thornton K, Long M (2002) Retroposed new genes out of the X in Drosophila. Genome Res 12(12):1854–1859

Birney E, Durbin R (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. Proc Int Conf Intell Syst Mol Biol 5:56–64

Castillo-Davis CI, Hartl DL, Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14(8):1530–1536

Chen S, Zhang YE, Long M (2010) New genes in Drosophila quickly become essential. Science 330(6011):1682–1685

Chen M, Zou M, Fu B, Li X, Vibranovski MD, Gan X, Wang D, Wang W, Long M, He S (2011) Evolutionary patterns of RNA-based duplication in non-mammalian chordates. PLoS One 6(7):e21466

Collins JE, White S, Searle SM, Stemple DL (2012) Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res 22(10):2067–2078

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, Zhang Y, Zhang G, Dong Y, Yu H, Zhou Q, Wang W (2010) A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. PLoS Genet 6(12):e1001255

Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. Science 303(5657):537–540

Ferris SDaW GS (1979) Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol 12:267–317

Fontanillas P, Hartl DL, Reuter M (2007) Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS Genet 3(11):e210

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4):1531–1545

Fu B, Chen M, Zou M, Long M, He S (2010) The rapid generation of chimerical genes expanding protein diversity in zebrafish. BMC Genom 11:657

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell 118(5):555–566

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652

Gu Z, Nicolae D, Lu HH, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18(12):609–613

Gu Z, Rifkin SA, White KP, Li WH (2004) Duplicate genes increase gene expression diversity within and between species. Nat Genet 36(6):577–579

Hasselmann M, Lechner S, Schulte C, Beye M (2010) Origin of a function by tandem gene duplication limits the evolutionary capability of its sister copy. Proc Natl Acad Sci USA 107(30):13378–13383

Hoegg S, Brinkmann H, Taylor JS, Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59(2):190–203

Jaillon OAJ, Brunet F, (61 co-authors) et al (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957

Jeffs P, Ashburner M (1991) Processed pseudogenes in Drosophila. Proc Biol Sci 244(1310):151–159

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10(1):19–31

Kalamegham R, Sturgill D, Siegfried E, Oliver B (2007) Drosophila mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. Mol Biol Evol 24(3):732–742

Kaushik K, Leonard VE, Kv S, Lalwani MK, Jalali S, Patowary A, Joshi A, Scaria V, Sivasubbu S (2013) Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. PLoS ONE 8(12):e83616

Kettleborough RN, Busch-Nentwich EM, Harvey SA, Dooley CM, de Bruijn E, van Eeden F, Sealy I, White RJ, Herd C, Nijman IJ, Fenyes F, Mehroke S, Scahill C, Gibbons R, Wali N, Carruthers S, Hall A, Yen J, Cuppen E, Stemple DL (2013) A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature 496(7446):494–497

Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12(8):R72

Korzh V (2009) Before maternal-zygotic transition… There was morphogenetic function of nuclei. Zebrafish 6(3):295–302

Lercher MJ, Williams EJ, Hurst LD (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol Biol Evol 18(11):2032–2039

Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends Genet 21(11):602–607

Li C, Orti G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol 7:44

Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J (2009) Expression pattern divergence of duplicated genes in rice. BMC Bioinformatics 10(Suppl 6):S8

Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260(5104):91–95

Makova KD, Li WH (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13(7):1638–1645

Markert CL (1964) Cellular differentiation—an expression of differential gene function. In: Congenital malformations, Internation Medical Congress, p 163–174

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 3(11):e357

McCarrey JR (1987) Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. Gene 61(3):291–298

Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27(9):937–945

Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett 468(2–3):109–114

Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17(9):1254–1265

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrink FT, Friedman M, Wainwright PC (2013) Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc Natl Acad Sci USA 110(31):12738–12743

Nekrutenko A, Makova KD, Li WH (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res 12(1):198–202

Newport JKM (1982) A major developmental transition in early Xenopus embryos: i. Characterization and timing of cellular changes at the midblastula stage. Cell 30:675–686

Neyfakh A (1956) The changes of radiosensitivity in the course of fertilization in the loach Misgurnus fossilis. Dokl Akad Nauk SSSR 109:943–946

Ohno S (ed) (1970) Evolution by Gene Duplication. Springer-Verlag

Okamura K, Nakai K (2008) Retrotransposition as a source of new promoters. Mol Biol Evol 25(6):1231–1238

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22(3):577–591

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Peterson AG, Wang X, Yost HJ (2013) Dvr1 transfers left-right asymmetric signals from Kupffer's vesicle to lateral plate mesoderm in zebrafish. Dev Biol 382(1):198–208

Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in Drosophila. Nature 384(6607):346–349

Sakai H, Mizuno H, Kawahara Y, Wakimoto H, Ikawa H, Kawahigashi H, Kanamori H, Matsumoto T, Itoh T, Gaut BS (2011) Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. Genome Biol Evol 3:1357–1368

Shiao MS, Liao BY, Long M, Yu HT (2008) Adaptive evolution of the insulin two-gene system in mouse. Genetics 178(3):1683–1691

Siddiqui M, Sheikh H, Tran C, Bruce AE (2010) The tight junction component Claudin E is required for zebrafish epiboly. Dev Dyn 239(2):715–722

Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A (1985) RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol 5(8):2090–2103

Stoltzfus A (1999) On the possibility of constructive neutral evolution. J Mol Evol 49(2):169–181

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res 13(3):382–390

Thisse B and Thisse C (2004) Fast release clones: a high throughput expression analysis. ZFIN direct data submission 2

Thisse B, Wright GJ and Thisse C (2008) Embryonic and larval expression patterns from a large scale screening for novel low affinity extracellular protein interactions. ZFIN direct data submission

Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13(12):2559–2567

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

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene

and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578

Urrutia AO, Hurst LD (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159(3):1191–1199

Van de Peer YTJ, Meyer A (2003) Are all fishes ancient polyploids? J Struct Funct Genomics 3:65–73

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103(9):3220–3225

Wang W, Zhang J, Alvarez C, Llopart A, Long M (2000) The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, *Drosophila melanogaster*. Mol Biol Evol 17(9):1294–1301

Wang W, Brunet FG, Nevo E, Long M (2002) Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci USA 99(7):4448–4453

Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28(16):2184–2185

Waterman MS, Eggert M (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. J Mol Biol 197(4):723–728

Williams EJ, Hurst LD (2002) Is the synonymous substitution rate in mammals gene-specific? Mol Biol Evol 19(8):1395–1398

Yamashita R, Suzuki Y, Sugano S, Nakai K (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. Gene 350(2):129–136

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586–1591

Yang L, Zou M, Fu B, He S (2013) Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. BMC Genom 14(1):65

Ye M, Berry-Wynne KM, Asai-Coakwell M, Sundaresan P, Footz T, French CR, Abitbol M, Fleisch VC, Corbett N, Allison WT, Drummond G, Walter MA, Underhill TM, Waskiewicz AJ, Lehmann OJ (2010) Mutation of the bone morphogenetic protein GDF3 causes ocular and skeletal anomalies. Hum Mol Genet 19(2):287–298

Yeo G, Hoon S, Venkatesh B, Burge CB (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci USA 101(44):15700–15705

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13(12):2541–2558

Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20(2):62–67