Divergent DNA Methylation Provides Insights into the Evolution of Duplicate Genes in Zebrafish

Zaixuan Zhong,† Kang Du,† Qian Yu,† Yong E. Zhang,†,§ and Shunping He*,†

*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; †Institute of Hydrobiology, University of Chinese Academy of Sciences, Beijing 100039, People’s Republic of China; and §Key Laboratory of the Zoological Systematic and Evolution and §State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Beijing 100000, People’s Republic of China

ABSTRACT The evolutionary mechanism, fate and function of duplicate genes in various taxa have been widely studied; however, the mechanism underlying the maintenance and divergence of duplicate genes in Danio rerio remains largely unexplored. Whether and how the divergence of DNA methylation between duplicate pairs is associated with gene expression and evolutionary time are poorly understood. In this study, by analyzing bisulfite sequencing (BS-seq) and RNA-seq datasets from public data, we demonstrated that DNA methylation played a critical role in duplicate gene evolution in zebrafish. Initially, we found promoter methylation of duplicate genes generally decreased with evolutionary time as measured by synonymous substitution rate between paralogous duplicates (Ks). Importantly, promoter methylation of duplicate genes was negatively correlated with gene expression. Interestingly, for 665 duplicate gene pairs, one gene was consistently promoter methylated, while the other was unmethylated across nine different datasets we studied. Moreover, one motif enriched in promoter methylated duplicate genes tended to be bound by the transcription repression factor FOXD3, whereas a motif enriched in the promoter unmethylated sequences interacted with the transcription activator Sp1, indicating a complex interaction between the genomic environment and epigenome. Besides, body-methylated genes showed longer length than body-unmethylated genes. Overall, our results suggest that DNA methylation is highly important in the differential expression and evolution of duplicate genes in zebrafish.

KEYWORDS divergence duplication methylation transcriptome zebrafish

Gene duplication, which occurs in almost all types of life forms (Kondrashov et al. 2002), is the main source of evolutionary novelty (Ohno 1970) and morphological complexity (Freedling and Thomas 2006). Most teleost, including Danio rerio, experienced genome duplication three times, with the most recent genome duplication dating to 320–400 MYA (Van de Peer and Meyer 2003; Jaillon et al. 2004; Kasahara et al. 2007; Hoegg et al. 2004). The exceptions are common carp and rainbow trout (Xu et al. 2014; Berthelot et al. 2014), both of which have undergone a fourth duplication. Several models of the emergence, maintenance, and evolution of duplicate gene copies have been proposed (Innan and Kondrashov 2010). Duplicate genes can be preserved through subfunctionalization, neofunctionalization, and dosage selection (Conant and Wolfe 2008; Hahn 2009). Nucleotide substitution, cis-regulation, and epigenetic modifications, influence the expression and functional evolution of duplicate genes (Hahn 2009; Betran et al. 2006; Wang et al. 2014; Chang and Liao 2012).

DNA methylation, an epigenetic DNA modification that occurs at cytosine residues, is involved in various important biological processes, such as the regulation of repetitive element expression, the development of early embryogenesis, cell type differentiation, genomic imprinting, and X-inactivation (Bird 2002; Bestor 2000; Smith et al. 2015; Edwards and Ferguson-Smith 2007). Promoter methylation is often associated with transcription repression, whereas intragenic methylation likely controls expression from alternative promoter regions and hinders
transcription elongation (Suzuki and Bird 2008; Maunakea et al. 2010; Bell and Felsenfeld 2000; Kass et al. 1997). Notably, epigenetic silencing of duplicates may aid in functional divergence (Rodin and Riggs 2003), and DNA methylation patterns play an important role in duplicate gene evolution (Widman et al. 2009; Keller and Yi 2014; Feng et al. 2010).

Previous study in humans demonstrated that DNA methylation exhibits striking degrees of evolutionary conservation (Keller and Yi 2014). DNA methylation divergence of duplicate genes is significantly correlated with gene expression divergence (Keller and Yi 2014). Duplicate genes show highly consistent patterns of DNA methylation divergence across multiple tissues due to different frequency of motifs (Keller and Yi 2014). Since zebrafish has been subjected to one more round whole genome duplication (WGD) compared to humans, we wondered whether the aforementioned patterns of DNA methylation of duplicate genes in human were different in zebrafish. In this study, we investigated the relationship between duplicate gene evolution and DNA methylation divergence. For example, we observed how the methylation level changed with evolutionary time (Ks), whether gene expression was coupled with DNA methylation, and how methylation divergence contributed to expression divergence. Since DNA methylation is important for early embryogenesis (Li et al. 1992), we also investigated DNA methylation patterns of duplicate genes during early developmental stages. All these results provided an answer to how DNA methylation influenced evolution of duplicate genes in zebrafish.

MATERIALS AND METHODS

Identification of duplicate genes
All the corresponding nucleotide and protein sequences of zebrafish were retrieved from Ensembl (http://asia.ensembl.org/index.html). To search potential zebrafish duplicate gene pairs, we initially used BLASTP (Altschul et al. 1997) with default parameters. Briefly, each protein sequence was compared against every other protein sequence in the zebrafish genome. Our criteria for whether two genes were considered a gene pair were proposed by Gu et al. (2002): (1) the alignable region between the two protein sequences should be longer than 80% of the longer region; (2) the identity between the two sequences (I) should be $I \geq 30\%$ if the alignable region is longer than 150 amino acids, and $I \geq 0.01$ $n + 4.8 L^{-0.32(1 + \exp(-L/1000))}$ for all other proteins, where $n = 6$, and $L$ is the alignable length between the two proteins (Rost 1999). Based on these initial pairings, gene families were created by performing the Markov Cluster Algorithm (http://micans.org/mcl/) until no additional groups shared a member. For each gene family, we aligned the protein sequences using MUSCLE (Edgar 2004). Using the yn00 module in PAML (Yang 2007), we calculated Ks pairwise, and selected the gene pair with the lowest Ks. We calculated the ratio of nonsynonymous to synonymous substitutions per site (Ka/Ks) of these duplicate genes using PAML4 (Yang 2007) to examine the functional constraints. A Ka/Ks ratio ($\omega > 1$) indicated positive selection, whereas a ratio $< 1$ indicated functional constrain. An LRT was conducted to determine whether Ka/Ks between the duplicate pairs was significantly lower than 0.5 (Zou et al. 2012; I. Wang et al. 2013). The Codeml program or PAML4 was run twice [model = 0 (fixing $\omega = 0.5$), and model = 1] for each pair. Then twice the log likelihood difference of these two runs was compared to a Chi-square distribution with df = 1 (Yang 1998). The false discovery rates (FDR) were controlled using the Benjamini-Hochberg method (Klipper-Aurbach et al. 1995) with an FDR of 5%. $\omega < 0.5$ ($P < 0.01$) may indicate evolutionary constraint.

Comparing the functional domains of duplicate genes
We used Interproscan (Jones et al. 2014) to identify the domains for duplicate genes by scanning the protein domains and important sites to determine any potential functions. The functional divergence of the duplicate copies was detected by comparing the domains. We
attributed two genes to the same functional group if they contained the same domains. The pairs that had distinct domains belonged to different functional groups.

Analysis of DNA methylation data
Methylation data for egg, sperm, testsis, and six stages (16-cell, 32-cell, 64-cell, 128-cell, 1k-cell, and germ ring), were obtained from NCBI with accession number PRJNA188516 (Jiang et al. 2013). RNA-seq data of testis was obtained with SRA number SRR1695730. Each read was separately mapped against Danio rerio Zv9 references (http://asia.ensembl.org/info/data/ftp/index.html) using the software TopHat (Trapnell et al. 2009, 2012). Reads that were longer than 48, and had no more than one multihit, were retained for next procedure. Considering that the high sequence similarity of duplicated genes might lead to the multiple alignment of sequencing reads, read counts in the expression analysis was based on a subset of uniquely aligned reads. For each gene, the normalized expression level was measured by fragments per kilobase of exon per million fragments mapped (FPKM) using Cufflinks (Trapnell et al. 2012).

To evaluate expression specificity of promoter methylated and unmethylated 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 $1 - H(g) / \log_2(N)$, where $N$ represents the number of points in time or the types of tissue (Pauli et al. 2012).

\[
H(g) = - \sum_{i=1}^{N} P_i \log_2 P_i, \quad P_i = \text{FPKM}\text{ of tissue } i
\]

where $g$ is the gene name; $g_i$ is the FPKM for the $i$th tissue; and $\text{FPKM}_\text{sum}$ is the sum of $N$ tissues.

DNA methylation divergence
DNA methylation divergence was calculated as previously described (Keller and Yi 2014; Kim and Yi 2006). We defined promoter methylation divergence (PMD) as $(M_{P1} - M_{P2}) / (M_{P1} + M_{P2})$, where $M_{P1}$ and $M_{P2}$ are the average promoter methylation levels for the first and second gene, respectively. The methylation level was normalized
to the overall methylation level of the pair. Similarly, gene body methylation divergence (GMD) was calculated as \( \frac{M_{G1} - M_{G2}}{M_{G1} + M_{G2}} \).

**Specificity index of DNA methylation**

The stage-specific patterns of DNA promoter methylation were described using the stage specificity index, which was previously used to assess gene expression (Yanai et al. 2005). The specificity index was defined as follows:

\[
SMI = \frac{\sum_{i=1}^{n} (1 - m_i/m_{\text{max}})}{n - 1}
\]

where \( m_i \) is the methylation in stage \( i \), \( m_{\text{max}} \) is the maximum methylation level for a gene across stages, and \( n \) is the number of stage. Thus, a larger SMI indicates a more stage-specific pattern of DNA methylation. We also calculated the divergence of SMI between duplicate gene pairs as \( \frac{(SMI_1 - SMI_2)}{(SMI_1 + SMI_2)} \).

**Motif-enrichment analysis**

We used MEME (Bailey et al. 2006) to identify DNA motifs that were distinguished within the promoter regions of consistently methylated vs. unmethylated duplicate genes. Considering the large number of sequences, we defined the promoter region as the 1000 bases upstream of the transcription start site (TSS), based on a previous study (Keller and Yi 2014). MEME was used to identify the 10 most significantly different motifs in consistently methylated promoters by generating motif position specific priors (PSPs). Then, MAST (Bailey and Gribskov 1998) was used to calculate the frequency of these motifs in the methylated and unmethylated promoter regions. Finally, we used TOMTOM (Gupta et al. 2007) to identify the transcription factor families to which the motifs bound.

**Statistical analysis**

In this study, R3.1.1 for windows was used for most statistical analysis. Pearson’s correlation coefficients were used to measure correlations between methylation and evolutionary time. We used the partial correlation with the “ppcor” package in R to examine the relative correlation between methylation level and gene expression (Y. Wang et al. 2013; Kim and Yi 2006). Multiple testing was corrected by applying the FDR method implemented in R (Storey and Tibshirani 2003).
genes/C24 gene-body methylated. A duplication event that occurred in the stem lineage of teleost is functional. That is, genes compiled in our duplicate gene dataset are largely copies of duplicate gene pairs were under purifying selection.

\[ \text{Ka/Ks} = \text{Likelihood ratio test (LRT)} \]

was calculated for duplicate gene pairs to assess natural selection.

\[ \text{Ka} \text{ to synonymous substitutions per synonymous site (Ks)} \]

ratio of nonsynonymous substitutions per nonsynonymous site (Ka) to synonymous substitutions per synonymous site (Ks) was calculated for duplicate gene pairs to assess natural selection (Yang 2007); the results are shown in Figure 2B. A likelihood ratio test (LRT) of Ka/Ks (\( \omega \)) confirmed that the \( \omega \) of 1851 of 2440 (75.9%) pairs were significantly < 0.5 (Supplemental Material, Table S1), adjusted \( P \)-value < 0.05, suggesting that both copies of duplicate gene pairs were under purifying selection. That is, genes compiled in our duplicate gene dataset are largely functional.

### Data availability

All the corresponding nucleotide and protein sequences of zebrafish were retrieved from Ensembl (http://asia.ensembl.org/index.html). Methylation data for egg, sperm, testis and six stages: 16-cell, 32-cell, 64-cell, 128-cell, 1k-cell and germ ring were obtained from NCBI with accession number PRJNA188516. Available paired end FASTQ sequence files for sperm, egg, 1,000 cells and germ ring were obtained from NCBI with accession number PRJNA188516. RNA-seq data of 16cell, 32-cell and 128-cell were downloaded with accession number PRJNA127881. RNA-seq data of testis was obtained with SRA number SRR1695730.

### RESULTS

#### Generation, quality control, and filter of the duplicate gene dataset

Using a refined version of a previously published procedure, 2440 pairs of duplicate genes with 0.01 < Ks \( \leq \) 2 were obtained in the zebrafish genome (Figure 1). The distribution of Ks is shown in Figure 2A. The number of duplicate genes tended to increase to a peak value, when the Ks value was 1.6. We hypothesized that the third round WGD (whole genome duplication) may cause this peak. Under a substitution rate of 4.13 \( \times \) \( 10^{-9} \) substitutions per silent site per year (Fu et al. 2010), WGD may indicate the birth of duplicate genes \( \sim 387 \) MYA, which was during the time of the third genome duplication event that occurred in the stem lineage of teleost fish (infraclass Teleostei) after the divergence from nonteleost ray-finned fish (Nakatani et al. 2007; Jaillon et al. 2004; Amores et al. 1998, 2011; Taylor et al. 2003; Van de Peer 2003; Meyer and Van de Peer 2005). Such consistency suggests the high quality of our duplicate gene dataset.

The ratio of nonsynonymous substitutions per nonsynonymous site (Ka) to synonymous substitutions per synonymous site (Ks) was calculated for duplicate gene pairs to assess natural selection (Yang 2007); the results are shown in Figure 2B. A likelihood ratio test (LRT) of Ka/Ks (\( \omega \)) confirmed that the \( \omega \) of 1851 of 2440 (75.9%) pairs were significantly < 0.5 (Supplemental Material, Table S1), adjusted \( P \)-value < 0.05, suggesting that both copies of duplicate gene pairs were under purifying selection. That is, genes compiled in our duplicate gene dataset are largely functional.

#### Young duplicates tended to be hypermethylated in promoter regions

The average promoter DNA methylation levels were calculated for nine datasets (egg, sperm, testis, 16-cell, 32-cell, 64-cell, 128-cell, 1k-cell, and germ ring), which exhibited a significant negative correlation with evolutionary time measured by Ks. The relationship for a 16-cell embryo (Pearson’s correlation coefficient, \( R = -0.72, P < 5.46E^{-04} \)) is presented in Figure 3A, and the relationship for the other stages are presented in Table S2. However, the average gene body methylation levels exhibited a positive correlation with evolutionary time (Pearson’s correlation coefficient, \( R = 0.58, P = 7.53E^{-03} \)) (Figure 3B for 16-cell embryos, Table S2 for other stages), indicating that the younger gene tended to have a lower promoter methylation level but higher gene body methylation level.

Interestingly, these trends were obvious when we compared methylation levels of “recent” duplicate pairs, which were single copy in grass carp, but duplicates in zebrafish (\( n = 85 \)), to those of duplicate pairs (Table S3). Duplicates are significantly (\( P < 0.05 \)) less methylated than young duplicates in promoters (Figure 3C), while more methylated than young duplicates in gene bodies (Figure 3D).

To examine this further, we compare the Ks of methylated and unmethylated genes. First, we calculated P_{CG} for the promoter region and gene body of each gene, and used the distribution of P_{CG} as a proxy for the CG methylation level (see Materials and Methods). Only those genes with sufficient CG information (\( n_{cg} \geq 20 \)), and genes for which 60% of cytosine residues were covered by at least two reads were kept (\( n = 2111 \)) (Takuno and Gaut 2012, 2013). The distribution of P_{CG} of promoter and gene body were both bimodal, indicating that CG methylation is not randomly distributed (Figure 4 and Figure S1). The result showed that Ks ratios were significantly higher in body-methylated genes than in unmethylated genes (Table S4, adjusted \( P \)-value < 2.85 \( E^{-12} \). In contrast to that, promoter-methylated genes exhibited lower Ks than unmethylated genes (Table S4, adjusted \( P \)-value < 4.51E-05).

#### Methylation divergence of duplicate genes changed along an evolutionary timescale

To study the dynamics of DNA methylation divergence within duplicate gene pairs, we calculated the relative promoter methylation divergence (PMD) and gene body methylation divergence (GMD) (see Materials and Methods). We observed that the PMD and Ks were positively correlated (Figure 5A for 16-cell, Pearson’s \( R = 0.61 \), adjusted
P value < 4.65E−03 (Figure S2 for other stages). Compared with older duplicate gene pairs, the younger pairs tended to exhibit similar levels of promoter methylation; however, significantly negative correlation was found between the GMD and Ks (Figure 5B, 16-cell, Pearson’s $R = -0.82$, adjusted $P$ value < 4.65E−03, Figure S3 for other stages).

We also calculated the stage specificity index of DNA promoter methylation (SMI, see Materials and Methods), which provided insights into the relative strength of methylation across six early embryo stages (16-cell, 32-cell, 64-cell, 128-cell, 1k-cell, and germ ring). A negative correlation was demonstrated between the relative divergence of SMI and Ks ($R = -0.33$, $P < 2.12E−15$).

**Negative correlation between promoter DNA methylation and gene expression level**

DNA methylation is known to regulate gene expression in mammals and plants, whereby higher levels of promoter methylation silence downstream gene expression (Weber et al. 2007; Zemach et al. 2010a). We hypothesized that a high level of promoter DNA methylation of duplicate promoters is also associated with low expression of duplicate genes in zebrafish. To explore the relative association of promoter methylation and gene body methylation with expression level, we evaluated partial correlation using the “ppcor” package in R (see Materials and Methods). Indeed, expression levels were significantly, negatively correlated with promoter methylation levels (Figure 6, $P < 3.62E−03$), whereas no significant correlation was established between the gene body methylation levels and expression levels.

We are wondering whether differential promoter methylation of duplicate gene pairs resulted in different gene expression. Compared with the promoter methylated genes, unmethylated gene exhibited a significant higher expression level (Table S5, two sample t-test, adjusted $P$ value < 1.43E−02).

Moreover, we used Shannon entropy to measure the breadth of expression. The result indicated that the promoter-unmethylated genes exhibited significantly lower Shannon entropy, suggesting a broader expression than promoter-methylated genes (Figure 7A, 4.98E−03).

**Enrichment of specific DNA motifs in consistent methylated promoters**

Why do promoter-methylated genes show lower gene expression level? One explanation is that some motifs in the promoter region bind to transcription repressors (Keller and Yi 2014). To test this hypothesis, we identified duplicate gene pairs where one copy was consistently

Figure 5 The relationship between methylation divergence and Ks. (A) PMD and Ks are positively correlated (Pearson’s $R = 0.61$, adjusted $P$ value < 4.65E−03). Compared with older duplicate gene pairs, the younger pairs tended to exhibit similar levels of promoter methylation. (B) Significantly negative correlation is found between the GMD and Ks (16-cell, Pearson’s $R = -0.82$, adjusted $P$ value < 4.65E−03).

Figure 6 The relationship between methylation and expression level (Log FPKM). (A) Promoters of duplicate genes are heavily methylated initially, and gradually lose DNA methylation cross evolutionary age. (B) In contrast, gene-body DNA methylation exhibits an increasing pattern.
promoter-methylated while the other copy was promoter-unmethylated across the nine samples we studied; 665 duplicate gene pairs fulfilled our criteria (Table S6).

Here, we examined the mechanism that helps distinguish the two promoters. We used a weight matrix finding algorithm (MEME) (Bailey et al. 2006, 2010), and a motif search tool (MAST) (Bailey and Gribskov 1998), to identify the 10 most significant motifs that discriminated between methylated and unmethylated groups. One motif occurred significantly more often in the unmethylated group than in the methylated groups ($P < 5.86 \times 10^{-8}$, Fisher’s exact test), whereas another motif occurred significantly less often ($P < 3.55 \times 10^{-6}$, Fisher’s exact test) (Figure 7B). Interestingly, the motif enriched in the methylated promoters contained the regions binding to Forkhead box D3 (FOXD3) that were previously identified by TOMTOM (Gupta et al. 2007). FOXD3 has been reported to function as a transcriptional repressor (Guo et al. 2002; Yaklichkin et al. 2007). By contrast, the motif enriched in the unmethylated promoters included Sp1 binding sites, which prevent local DNA methylation (Brandeis et al. 1994). The difference in the methylation levels of promoters may be explained by the presence of these motifs.

**Body-methylated genes showed longer length than body-unmethylated genes**

Previous studies revealed different predictions between body methylation and gene length or exon. For instance, in *Arabidopsis thaliana*, body-methylated genes were significantly longer than unmethylated genes, and have more exons (Takuno and Gaut 2012). Genes with higher CpG were significantly longer than those with lower CpG (Zeng and Yi 2010).

In this study, our data also support these predictions. We found the mean length of the body-methylated genes was significantly longer than the mean length of unmethylated genes (37,582.78 vs. 7655.52 bp for 16 cell, Table S7 for other stages, two sample t-test, adjusted $P$ value $< 2.2 \times 10^{-16}$). Figure 8 shows the distribution of gene length.
expression and functional divergence. In our study, we performed a comparative analysis of the association of epigenetic modification, such as DNA methylation, with the evolutionary divergence of duplicate genes.

**The promoter regions of younger duplicate genes are generally methylated**

It was known that most newly born genes may degrade to pseudogenes. Epigenetic silencing of duplication might play an important role in shifting the loss vs. gain equilibrium (Rodin and Riggs 2003). In our study, we noticed that the promoter regions of younger duplicate genes tended to be methylated, whereas those old duplicates were generally unmethylated (Figure 3, C and D). Remarkably, a similar pattern has also been found in humans (Keller and Yi 2014). Thus, it is possible that newly duplicated genes have the same expression pattern, and are epigenetically silenced in a tissue- or stage-complementary manner, which protects each of the duplicates from “pseudogenization” (Rodin and Riggs 2003).

**Evolutionary conservation of gene body methylation**

According to previous studies, gene bodies consistently exhibit higher levels of methylation compared with promoters (Jjingo et al. 2012). Moreover, our study showed that gene body methylation was significantly correlated with Ks (Pearson’s $R = 0.58$, Figure 3B). In humans, gene-body DNA methylation and Ks are negatively correlated, but this correlation is extremely weak (Pearson’s $R = -0.06$) (Keller and Yi 2014). Gene body methylation is reportedly conserved between plants and animals (Zemach et al. 2010b; Sarda et al. 2012; Zeng and Yi 2010), whereas a study in rice suggests that gene body divergence is associated with Ks (Y. Wang et al. 2013). However, GMD in zebrafish did not exhibit a discernible relationship with Ks. This result suggested that the epigenetic modification of the gene body might be subject to Ks, whereas GMD between duplicate genes was relatively conserved.

Differential gene body DNA methylation covaries with gene length between duplicate genes. It has been hypothesized that body methylation has a functional role, perhaps in transcriptional accuracy or splicing efficiency. Consistently, we demonstrated that methylated duplicate genes have longer gene length.

**Methylation of promoters, rather than that of gene bodies, is associated with transcription levels**

In mammals, DNA methylation of promoter regions is a repressive mark, and depresses gene expression (Zemach et al. 2010b; Elango and Yi 2008; Boyes and Bird 1992; Shen et al. 2007), whereas intragenic methylation is associated with gene expression by controlling the expression from alternative promoter regions (Maunakea et al. 2010). Additionally, in plant genomes, including *A. thaliana* (Arabidopsis), *Oryza sativa* (rice), *Populus trichocarpa* (poplar), and *Chlamydomonas reinhardtii* (green algae), gene-body methylation tends to influence transcription level (Feng et al. 2010; Zemach et al. 2010b; Takuno and Gaut 2013). Even a single CpG within a transcription-factor-binding site potentially influences gene regulation (Ziller et al. 2013). Indeed, we demonstrated that promoter methylation was significantly correlated with gene expression (Figure 6), whereas no significant correlation was observed between gene body methylation and expression level in zebrafish. Meanwhile, promoter-unmethylated genes exhibited significantly lower Shannon entropy, suggesting a broader expression than promoter-methylated genes. The relationship between DNA methylation and transcription level potentially varies between taxa. Consistent with previous studies that raised the notion of “expression

**DISCUSSION**

DNA-mediated duplication, that is, the duplication of chromosomal segments containing genes, has been widely studied (Lynch and Conery 2000; Taylor et al. 2003). DNA duplication is a critical source of genetic innovation that plays a key role in evolution (Assis and Bachtrog 2013). After duplication, genes are subject to a series of processes, including divergence.
reduction model” and “gene dosage balance,” our results indicated that heavy promoter methylation following the duplication event may offset the expression level to avoid detrimental mutations (Rodin and Riggs 2003; Chang and Liao 2012).

Remarkably, the motif-enrichment results revealed that the Forkhead-related transcriptional regulator FOXD3 is present at a significantly higher frequency within the methylated promoters. A previous study of gene expression has suggested that FOXD3 is involved in a negative autoregulatory mechanism (Chiang et al. 2001; Dottori et al. 2001). Moreover, our study demonstrated that promoter methylation divergence of duplicate genes also affects gene expression, indicating that epigenetic divergence potentially influences transcription levels. Comparative genome analysis regarding duplicate genes supports the hypothesis that differential DNA methylation and epigenetic changes play a role in protecting duplicate genes from pseudogenization (Rodin et al. 2005; Cortese et al. 2008).

Genetic influences on DNA methylation variation
Previous study has proved that DNA methylation variation is influenced by genetic and epigenetic changes that are often stably inherited and can influence the expression of nearby genes (Eichten et al. 2013). In this study, we also tried to assess the influence of nucleotide divergence, especially at C nucleotides and CpG di-nucleotides between duplicates, on DNA methylation in zebrafish. First, we tried to identify duplicates that display differential methylation but have little to no sequence variation. These could be more likely candidates of true epigenetic variation. However, since zebrafish went through the third genome duplication 320–400 MYA, duplicate genes diverge greatly in sequence. Then, we carefully assessed the C content of promoter regions between duplicate pairs with a customized Perl script. On the one hand, we narrowed down a set of duplicate genes that have diverged greatly in their C content. We found that 356 duplicate pairs (Table S9) show differential methylation, possibly because the C content was different, which was caused by divergence at the nucleotide level.

Our study gives strong support to the idea that epigenetic divergence of duplicate genes affects gene expression and functional divergence of duplicate genes.

ACKNOWLEDGMENTS
The work was supported by grants from the Chinese Academy of Sciences (XDB13020100) and the National Natural Science Foundation of China (91131014). The authors declare no competing financial interests.

Author contributions: Z.Z. developed the algorithm, performed the analyses, and drafted the manuscript. K.D. participated in algorithm development. Q.Y. participated in data analysis. Y.E.Z. conceived of the study, participated in its design and coordination, and helped analyze the data. All authors read and approved the final manuscript.

LITERATURE CITED
Aanes, H., C. L. Winata, C. H. Lin, J. P. Chen, K. G. Srinivasan et al., 2011 Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. 21(8): 1328–1338.
Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang et al., 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389–3402.
Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya et al., 1998 Zebrafish box clusters and vertebrate genome evolution. Science 282(5394): 1711–1714.
Amores, A., J. Catchen, A. Ferrara, Q. Fontenot, and J. H. Postlethwait, 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.
Assis, R., and D. Bachtrog, 2013 Neofunctionalization of young duplicate genes in Drosophila. Proc. Natl. Acad. Sci. USA 110(43): 17409–17414.
Bailey, T. L., and M. Gribskov, 1998 Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14(1): 48–54.
Bailey, T. L., N. Williams, C. Misle, and W. W. Li, 2006 MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34 (Web Server issue): W369–W373.
Bailey, T. L., M. Boden, T. Whittington, and P. Machanick, 2010 The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics 11: 179.
Bell, A. C., and G. Felsenfeld, 2000 Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405 (6785): 482–485.
Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard et al., 2014 The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5: 3657.
Bestor, T. H., 2000 The DNA methyltransferases of mammals. Hum. Mol. Genet. 9(16): 2395–2402.
Betran, E., Y. Bai, and M. Motiwale, 2006 Fast protein evolution and germ line expression of a Drosophila parental gene and its young retroposed paralog. Mol. Biol. Evol. 23(11): 2191–2202.
Bird, A., 2002 DNA methylation patterns and epigenetic memory. Genes Dev. 16(1): 6–21.
Boyes, J., and A. Bird, 1992 Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J. 11(1): 327–333.
Brandes, M., D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn et al., 1994 Sp1 elements protect a CpG island from de novo methylation. Nature 371(6549): 435–438.
Chang, A. Y., and B. Y. Liao, 2012 DNA methylation rebalances gene dosage after mammalian gene duplications. Mol. Biol. Evol. 29(1): 133–144.
Chiang, E. F., C. I. Pai, M. Wyatt, Y. L. Yan, J. Postlethwait et al., 2001 Two sox9 genes on duplicated zebrafish chromosomes: expression of similar transcription activators in distinct sites. Dev. Biol. 231(1): 149–163.
Conant, G. C., and K. H. Wolfe, 2008 Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9(12): 938–950.
Cortese, R., M. Krispin, G. Weiss, K. Berlin, and F. Eckhardt, 2008 DNA methylation profiling of pseudogene-parental gene pairs and two gene families. Genomics 91(6): 492–502.
Dottori, M., M. K. Gross, P. Labosky, and M. Goulding, 2001 The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. Development 128 (21): 4127–4138.
Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792–1797.
Edwards, C. A., and A. C. Ferguson-Smith, 2007 Mechanisms regulating imprinted genes in clusters. Curr. Opin. Cell Biol. 19(3): 281–289.
Eichten, S. R., R. Briskine, I. Song, Q. Li, R. Swanson-Wagner et al., 2013 Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell 25(8): 2783–2797.
Elango, N., and S. V. Yi, 2008 DNA methylation and structural and functional bimodality of vertebrate promoters. Mol. Biol. Evol. 25(8): 1602–1608.
Feng, S., S. J. Cokus, X. Zhang, P. Y. Chen, M. Bostick et al., 2010 Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. USA 107(19): 8689–8694.
Freeling, M., and B. C. Thomas, 2006 Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16(7): 805–814.
Yang, Z., 2007  PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8): 1586–1591.
Zemach, A., M. Y. Kim, P. Silva, J. A. Rodrigues, B. Dotson et al., 2010a  Local DNA hypomethylation activates genes in rice endosperm. Proc. Natl. Acad. Sci. USA 107(43): 18729–18734.
Zemach, A., I. E. McDaniel, P. Silva, and D. Zilberman, 2010b  Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328 (5980): 916–919.
Zeng, J., and S. V. Yi, 2010  DNA methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of DNA methylation. Genome Biol. Evol. 2: 770–780.
Ziller, M. J., H. Gu, F. Muller, J. Donaghey, L. T. Tsai et al., 2013  Charting a dynamic DNA methylation landscape of the human genome. Nature 500 (7463): 477–481.
Zou, M., G. Wang, and S. He, 2012  Evolutionary patterns of RNA-based gene duplicates in Caenorhabditis nematodes coincide with their genomic features. BMC Res. In 5: 398.

Communicating editor: A. S. McCallion