Article

Chinese Science Bulletin

SPECIAL TOPIC
Bioinformatics

August 2010 Vol.55 No.22: 2359–2365
doi: 10.1007/s11434-010-3027-5

Cis-regulatory change and expression divergence between duplicate genes formed by genome duplication of *Arabidopsis thaliana*

CHEN KeNian[1,2], ZHANG YanBin[3], TANG Tian[1] & SHI SuHua[1*]

[1] *State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen (Zhongshan) University, Guangzhou 510275, China;*
[2] *Department of Biotechnology, Guangzhou Medical College, Guangzhou 510182, China;*
[3] *College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China*

Received April 14, 2009; accepted August 13, 2009

As an index of functional divergence, expression divergence between duplicate gene copies has been observed and correlated with protein coding sequence divergence and bias in gene functional classes. However, the changes in the *cis*-regulatory region of the duplicate genes which is thought to have important role in expression divergence, has not been explored on the genome-wide scale. We analyzed functional genomics data for a large number of duplicated gene pairs formed by ancient polyploidy events in *Arabidopsis thaliana*. The divergence in *cis*-regulatory regions between two copies is positively correlated with the magnitude difference of expression. Moreover, we find that highly expressed duplicate gene pairs have a more diverged *cis*-regulatory region than weakly expressed gene pairs. We also show that the correlation between expression functional constraint and protein functional constraint is different in old and young duplicate pairs. Our results suggest that *cis*-regulatory sequence divergence contributes to the expression divergence of duplicate genes formed by genome-wide duplication. *Cis*-regulatory region diverges faster in highly expressed duplicate pairs. The diversify selection strengths that act on *cis*-regulatory region and protein coding region are negatively correlated in young duplicate pairs under expression constraint.

genome duplication, duplicate gene *cis*-regulatory divergence, expression divergence, *Arabidopsis*, evolution

Citation: Chen K N, Zhang Y B, Tang T, et al. *Cis*-regulatory change and expression divergence between duplicate genes formed by genome duplication of *Arabidopsis thaliana*. Chinese Sci Bull, 2010, 55: 2359–2365, doi: 10.1007/s11434-010-3027-5

Since Ohno [1], "evolution by gene duplication" has been recognized as a general principle of biological evolution [2]. The most spectacular gene duplication is the whole-genome duplication via polyploidization which is most prominent in plants [3–5]. The model plant *Arabidopsis* is believed to have experienced at least three ancient polyploidy events [6–8]. The remnants of these polyploidy events comprise ~23% of the genes in current *Arabidopsis* genome and form a large set of duplicated chromosomal segments, which have been identified and ordered in different age classes by several groups using slightly different approaches [6,9]. Previous studies indicated that functional divergence is the most likely fate of duplicate genes retained in genome [4,10–12]. However, the driving force of divergence is still poorly understood [13].

Although there are good reasons to consider the *cis*-regulatory region of the duplicate genes as well, to date, almost all studies have focused on protein coding sequences [2,4,12,14–18], mainly due to the lack of a biologically relevant measure of *cis*-regulatory evolution that relates directly to gene expression [3,4,15,19]. Accordingly, Castillo-Davis et al. [20] described a method called the shared motif method (SMM) to quantify functional regulatory changes in *cis*-regulatory regions.

In this paper, we adopted the SMM to investigate the questions of how *cis*-regulatory region changes contribute to paralogs expression divergence in *Arabidopsis*, the effect of genome duplication on *cis*-regulatory region evolution,

*Corresponding author (email: lssssh@mail.sysu.edu.cn)

and the relationship between protein and regulatory sequence evolution in duplicates.

To address these questions, duplicate genes derived from polyploidy events that are still lying in chromosomal segments are excellent materials for several reasons. Firstly, genes duplicated via retro-transposition would lose regulatory sequences and include additional sequences at flanking regions. Tandem duplication by unequal crossing over might even not include the entire coding and/or regulatory sequences [2,15,21]. Thus, duplicate genes created via these mechanisms differed immediately after duplication, making the analysis of *cis*-regulatory region divergence complicated. Secondly, genome duplication derived duplicates were created simultaneously, divergence time between pairs belonging to the same age class were the same [7,9,10]. Thirdly, segmental duplicate genes were extensively identified by several groups and yielded similar results, the reliability of data was high, and the effect of genome rearrangement was less than other types of duplicate genes [4,8].

1 Materials and methods

(i) Sequence data. All gene and genomic sequence information, including intergenic distances, protein coding sequences (CDS), upstream sequences from transcription start sites, were obtained from TAIR database release 8. For genes that has more than one transcript (annotated splice variants), only the first instance (id with ".1" suffix) was used for study. It is usually the longest transcript of the gene.

(ii) Expression data. Gene expression information was obtained from Nottingham *Arabidopsis* Stock Centre's microarray database (NASCArrays). The dataset contained 62 ATH1 Affymetrix *Arabidopsis* microarray expression intensities under various experimental conditions and tissues were utilized by Blanc et al. [4]. Microarray probe intensities were normalized using MAS5.0 algorithm, i.e. the top 2% and bottom 2% of signal intensities were excluded, then the mean was calculated. The original signal values were scaled such that the mean was made equal to 100. Expression values were averaged among replicates. 128 Genes with the potential for cross-hybridization (marked with the "x" suffix on their probe ID) were discarded. We also excluded any probes that matched multiple genes. Thus, the potential of cross-hybridization was largely reduced. The genes without one expression value >150 were classified as 'no expressed' genes, which meant that the genes were weakly expressed. The genes left were classified as 'expressed' genes. The magnitude expression difference was defined as the absolute difference between maximum expression values of two copies.

(iii) Protein sequence analysis. To identify dispersed pairs, the FASTA method described by Gu et al. [22] was used. Briefly, after excluded mitochondrial and cytochondrial proteins an all against all FASTA search was conducted with $E<10$. Two protein sequences were identified as pair if (1) the aligned region was >80% of the longer protein, and (2) the identity between two proteins was $\geqslant$ 30% for alignments longer than 150 amino acids or $\geqslant (0.01n + 4.8L^{-0.32[1+\exp(-L/1000)]})$ otherwise, where L is the alignable length between two proteins and n=6. Tandem duplicates were identified as duplicate genes located within 100 kb each other, separated by less than 2 non-homologous genes. Dispersed pairs were selected from families containing only two members and excluded segmental and tandem duplicates. Duplicate gene pairs derived by genome duplication were retrieved from Blanc et al. [4]. The dataset contained 2584 'young' duplicate pairs from the most recent polyploidy events, and 1372 'old' duplicate pairs formed in two older polyploidy events.

For each duplicate pair, coding sequences were aligned by CLUSTALW [23] using the amino-acid translation of each sequence followed by back-translation into DNA sequence alignment. The maximum likelihood estimation of K_a, K_s, K_a/K_s ratio values were obtained using CODEML [24] program in the PAML package[25].

(iv) Regulatory sequence analysis. The *cis*-regulatory sequence analysis was achieved by using SMM (shared motifs method) which was described in details by Castillo-Davis et al. [20]. Briefly, a shared motif is defined as a region of high local similarity between two given DNA sequences without considering their order, orientation, or spacing.

The SMM value was defined as the fraction of both sequences containing shared motifs. The SMM software (sharmot) was obtained from Castillo-Davis et al., and was used to calculate the divergence of upstream sequence of 100, 500, 1000, 1500 bp from transcription start sites (TSS) between each duplicate pair.

To obtain the SMM distribution of upstream sequences of 3000 bp from TSS, we modified the SMM method, and created a sliding window version of SMM. The window size was 300 bp, and the step size was 100 bp, sliding over upstream sequences of 3000 bp from TSS, the average SMM value and standard error were calculated for every window. For each pairs group (i.e. Dispersed-E, Dispered-N, Old-E, Young-E, Old-N, Young-N pairs respectively), the average SMM value and standard error of each window were calculated and used to draw the distribution line.

Correlation and linear multiple regression analysis was performed using R statistical package.

2 Results and discussion

2.1 Identification and classification of duplicate genes

A FASTA method described by Gu et al. [22] was used to identify duplicate pairs. As mentioned earlier, we mainly focused on duplicate genes formed by genome duplication when analyzing the correlation between *cis*-regulatory divergence and expression divergence. We also utilized the gene families containing only two members when studying

the interplay between the expression constraint and *cis*-regulatory divergence, because there was less influence of other family members.

A total of 1148 dispersed duplicate pairs (i.e. not tandem duplicates, not segmental duplicates) were identified.

Genome duplication in *Arabidopsis* has been extensively studied, and polyploidy-derived duplicated gene pairs that still lying on segments have been identified by several groups using slightly different approaches [6–9], because most of their data are overlapped, the slight difference should not significantly affect the investigation. Therefore, we only performed analysis using dataset from Blanc et al. [4], which had been utilized in their study. In this dataset, 2584 'young' duplicate pairs came from the most recent polyploidy events, and 1372 'old' duplicate pairs formed in two older polyploidy events.

We first classified all genes as 'not expressed' or 'expressed' according to their expression values across 62 microarray experiments (see materials and methods for details). In brief, the genes have expression value >150 in at least one experiment were classified as 'expressed' else 'not expressed'. Genes subject to cross-hybridization were excluded, and only those genes for which a unique probe set (probe ID with '_at' extension, without suffix) was available on the ATH1 microarray were retained, thus the effect of potential cross-hybridization was reduced to the minimum. Among the 1148 dispersed duplicate pairs, both copies of 595 pairs were 'expressed' (Dispersed-E), either or both copies of 553 pairs were 'not expressed' (Dispersed-N).

Among the 2584 young duplicate pairs, both copies of 1125 pairs were 'expressed' (Young-E pairs), either or both copies of 1459 pairs were 'not expressed' (Young-N pairs). For the 420 old duplicate pairs, both copies were 'expressed' (Old-E pairs), and for the 728 old pairs, either or both copies were 'not expressed' (Old-N pairs) (Table 1).

The correlation analysis of *cis*-regulatory sequences divergence and expression difference were carried out using Young-E and Old-E pairs, both copies of which were 'expressed'.

2.2 *Cis*-regulatory sequences divergence and expression magnitude changes

Expression profile correlation is often used as an amenable indicator of functional divergence between duplicate genes. However, gene expression functionality has a lot of aspects. For example, gene expression can change in spatial, temporal, and environmental dimensions, as well as in the expression levels (i.e. the transcript abundance). The former changes were referred to as changes in relative expression and the latter as changes in expression magnitude in [20]. It is impossible to summarize all of them with just a single measure. In analyzing expression patterns and *cis*-regulatory region divergence of C.elegans, Castillo-Davis et al. observed a correlation between *cis*-regulatory sequence divergence and the differences in the magnitude of expression between duplicate genes in C.elegans, by utilizing a 'shared motif method' (SMM, see Materias and methods for details) [20] .

We adopted the SMM method to measure divergence between upstream sequences of each duplicate pair, both copies of which were 'expressed', and related it to the difference of maximum expression value. Keep in mind that our purpose is not to identify high confidence *cis*-regulatory motifs but rather to detect global quantitative trends. Because the average size of regulatory regions is not known, we first calculated that the average intergenic sequences length of *Arabidopsis* genome was about 1600 bp, and then we looked for shared motifs 100, 500, 1000 and 1500 bp upstream from annotated transcription start sites.

For 420 Old-E pairs, we observed a highly significant correlation between dsm (i.e. 1- proportion of shared motifs in total sequence) and difference in gene expression magnitude (Spearman correlation $rs = 0.169$, $P<10^{-3}$) for upstream sequences of 1500 bp (Figure 1). Shorter or longer upstream sequences were less correlated with expression difference (data not shown). For the 1125 Young-E pairs, we observed a significant correlation between dsm and difference in gene expression magnitude (Spearman correlation $rs = 0.068$, $P<0.05$) for upstream sequences of 500 bp. Note that the upstream sequence length that gives the significant correlation are not necessarily the same between 'old' and 'young' duplicate pairs, because SMM is not a motif discovery method, it only detects the local similarities between given sequences, and if the divergence time of the two sequences is not long enough, the similarity in the non-motif regions of the upstream sequence will probably obscure the detection of shared motifs. On the other hand, Lynch et al. [26]

Table 1 Statistics of duplicate gene pair groups[a)]

Symbol	Meaning	Number of pairs
Dispersed-E	Both copies are 'expressed' according to our criteria	595
Dispersed-N	Either or both copies are 'not expressed'	553
Young-E	Both copies are 'expressed' according to our criteria	1125
Young-N	Either or both copies are 'not expressed'	1459
Old-E	Both copies are 'expressed' according to our criteria	420
Old-N	Either or both copies are 'not expressed'	728

a) The number of pairs in each group. '-E' means both copies of the duplicate gene pair are classified as 'expressed'. '-N' means either or both copies of the duplicate gene pair are classified as 'not expressed'.

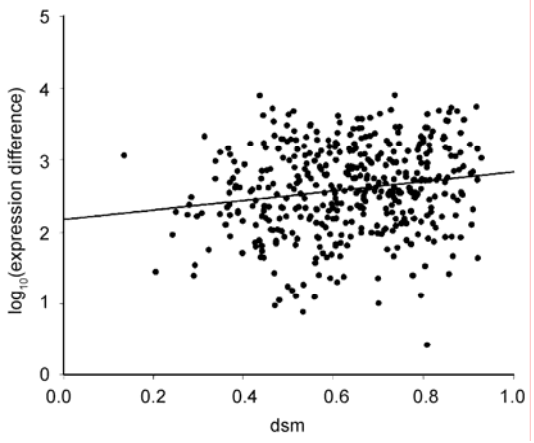

Figure 1 Correlation between dsm and gene expression magnitude difference. A highly significant positive correlation between expression difference and dsm (1- proportion of shared motifs in total sequences) of upstream sequence 1500 bp from transcription start site of Old-E pairs.

argued that the mutational events and population-genetic mechanisms that led to the short-term preservation of duplicate genes were not necessarily the same as those exhibited by well-established paralogs. Under this view, the old duplicate pairs are well-established paralogs, and the young duplicate pairs, at least part of them, are probably still experiencing a birth and death process. So it is not surprising to see a higher significant correlation between expression divergence and *cis*-regulatory divergence in old duplicate pairs than in young duplicate pairs.

Another question is whether the correlation between dsm and magnitude difference in expression is caused only by their correlation with K_s (synonymous substitutions per site), i.e. the substitution rate, respectively. To test whether dsm really contributes to expression difference, we applied a multiple regression analysis (multiple regression formula: expression difference ~ dsm + K_a + K_s) to old and young duplicate pairs respectively, and found that, for old duplicate pairs, dsm did contribute to expression difference, whereas K_s did not, and K_a had a negative correlation with the expression difference, which was consistent with the observation that strength of purifying selection acting on protein sequence and expression was correlated in *Arabidopsis* [27]. For young duplicate pairs, the correlation between dsm and the expression differences were largely due to their correlation with K_s respectively. These results taken together, imply that dsm correlates with a functional difference in the gene expression magnitude is not a simple consequence of substitution rate (K_s), at least in old duplicate pairs, and for young pairs, this method might be less powerful due to the fact that the divergence between the upstream sequence is not as large enough as old duplicate pairs to discriminate the 'real motifs' from non-coding DNA that is under little or no selective constraint.

2.3 Highly expressed duplicate pairs have more diverged *cis*-regulatory sequences

Papp et al. [28] observed that the evolutionary rates of protein coding sequences were slower for highly expressed duplicate genes in yeast, which suggests a higher functional constraint on highly expressed genes' protein coding sequences.

We first compared the K_a/K_s ratio of 'expressed' and 'not expressed' duplicate pairs in dispersed duplicate pairs (i.e. Dispersed-E pairs and Dispersed-N pairs) previously identified. And then we performed the same analysis on young and old age groups respectively (i.e. Young-E, Young-N, Old-E, Old-N group of pairs as mentioned earlier). The K_a/K_s ratio is a traditional index of functional constraints on protein sequence. The smaller the K_a/K_s ratio is, the stronger the functional constraints are. Consistent with the finding in yeast, the K_a/K_s ratio of 'expressed' pairs is smaller than that of 'not expressed' pairs (Dispersed-E vs. Dispersed-N $P< 2.2\times10^{-16}$, Young-E vs. Young-N $P < 2.2\times10^{-16}$, Old-E vs. Old-N $P < 1.914\times10^{-12}$ Wilcoxon U-test), which indicates a larger functional constraint on protein coding sequences of duplicate genes that are both 'expressed'.

It is interesting to know whether the expression level will influence the evolutionary rates (i.e. the divergence) of duplicate genes' *cis*-regulatory sequences as well and how.

To answer this question, we calculated the overall 'shared motif proportion' (smm value) for 1500 bp upstream sequences of 'expressed' and 'not expressed' pairs for 'dispersed', 'old' and 'young' duplicate gene, respectively. The lower the smm value, the larger the divergence is. We found that, for 'dispersed' pairs the smm value was significantly higher in 'not expressed' pairs than in 'expressed' pairs ($P = 2.501\times10^{-17}$, Wilcoxon U-test). This also holds for old duplicates ($P = 0.037$, Wilcoxon U-test). For young duplicates, although the trend is obvious, the smm value was not significantly higher in 'not expressed' pairs ($P = 0.35$, Wilcoxon U-test).

Because the distribution of *cis*-regulatory motifs on upstream sequence is not even, only depending on the overall smm value for upstream sequences of 1500 bp might make it look too sloppy, and looking into the shared motif distribution may give us more insight. We then apply a sliding window version SMM method (see materials and methods) to upstream sequence of 3000 bp for each duplicate pair, and obtain an average smm distribution for 'expressed' and 'not expressed' duplicate pairs (Figure 2).

Figure 2 clearly shows that the 'expressed' pairs have more diverged *cis*-regulatory sequences (lower smm value) than the 'not expressed' pairs. A straight forward explanation is that for 'expressed' gene pairs, the redundancy of the gene product in a specific tissue should be reduced to maintain the viability of the organism, An economic way to achieve this goal is through the changes in the *cis*-regulatory region which drives tissue-specific expression [29].

Evolution models of duplicate gene retention state that sub-function of a gene might involve the expression of a gene in a specific tissue, cell lineage, or developmental stage, or individual functional domains within the protein coding region of the gene [29]. Thus, functional divergence might be achieved either through the complementary loss of regulatory elements and/or protein domain change [3,15,29]. Consider that the diversify selection constraints act on *cis*- and protein region as a whole, diverge of one region may relax the constraints on the other region, and thus yield the above observation.

2.4 Expression functional constraint and protein functional constraint

Recent studies have shown that the fate of a duplicated gene largely depends on its function [30,31]. According to this finding and combined with the above results, we can infer that the expression functional constraints (ED/K_s ratio, using dsm/K_s to give similar result) should positively correlate with protein function constraints (K_a/K_s ratio), because these two indexes both indicate a gene's functional importance.

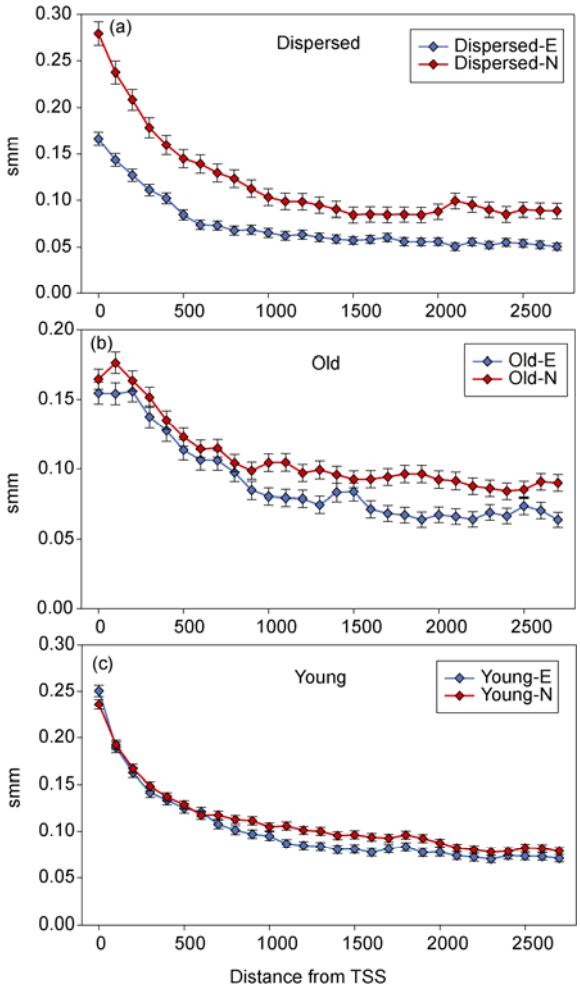

Figure 2 Average smm distribution of upstream sequence 3000 bp from TSS. The average smm value (proportion of 'shared motifs' in total sequence) distribution for 'expressed' (blue) and 'not expressed' (red) pair groups. (a) Dispersed-E pairs vs. Dispersed-N pairs; (b) Old-E pairs vs. Old-N pairs; (c) Young-E pairs vs. Young-N pairs.

For 595 Dispersed-E pairs and 420 Old-E pairs, there is a highly significant positive correlation between the ED/K_s ratio and the K_a/K_s ratio. (Dispersed-E pairs : $rs = 0.663$, $P < 2.2\times10^{-16}$, Spearman correlation Figure 3(a); Old-E pairs: $rs = 0.423$, $P < 2.2\times10^{-16}$; Figure 3(b)). For 1125 Young-E pairs, interestingly, we found a highly significant negative correlation between the ED/K_s ratio and the K_a/K_s ratio ($rs = -0.103$, $P < 0.001$ Spearman correlation Figure 3(c)).

One explanation of the discrepancy between the Old-E pairs and Young-E pairs is that, unlike Old-E pairs, Young-E pairs, at least partially, are still under a strong diversify selection constraint compare with the well-established Old-E pairs. Given diversify selection acting on genes' expression function and protein function as a whole, the divergence of expression function may relax the selection constraint on protein function, or vice versus.

If this is true, then for the Young-N pairs, the correlation of expression function constraint and protein function constraint would be similar to those of the Old-E pairs, because both copies of Young-N pairs are 'not expressed' and hence under a less or no expression functional constrain.

Because Young-N pairs are weakly expressed, whose expression value might be subjected to large noise, we use dsm/K_s ratio instead of using ED/K_s ratio as index of the expression functional constraint. Consistent with this explanation, there is highly significant positive correlation between dsm/K_s ratio and K_a/K_s ratio in Young-N pairs ($rs = 0.081$, $P < 0.01$, Spearman correlation Figure 3(d).

3 Conclusions

We studied *cis*-regulatory sequence divergences and evaluated their relative contribution of differences in expression between copies of duplicate genes derived by polyploidy events. Although the nature of the data analyzed associated with high levels of noise and other factors may also contribute to expression divergence (i.e. epigenetic difference), we still found a significant positive correlation between *cis*-regulatory divergence and expression difference in duplicate genes formed by polyploidy events especially in old pairs on the whole-genome scale. This implies that the divergence of *cis*-regulatory region leads to expression difference in duplicate genes at least in that formed by polyploidy events.

By classifying the duplicate pairs as 'expressed' and 'not expressed', we studied the influence of expression constraint on the *cis*-regulatory sequence divergence, and found that the duplicate pair under high expression pressure tended to have more diverged *cis*-regulatory region. The negative correlation between expression constraint and protein

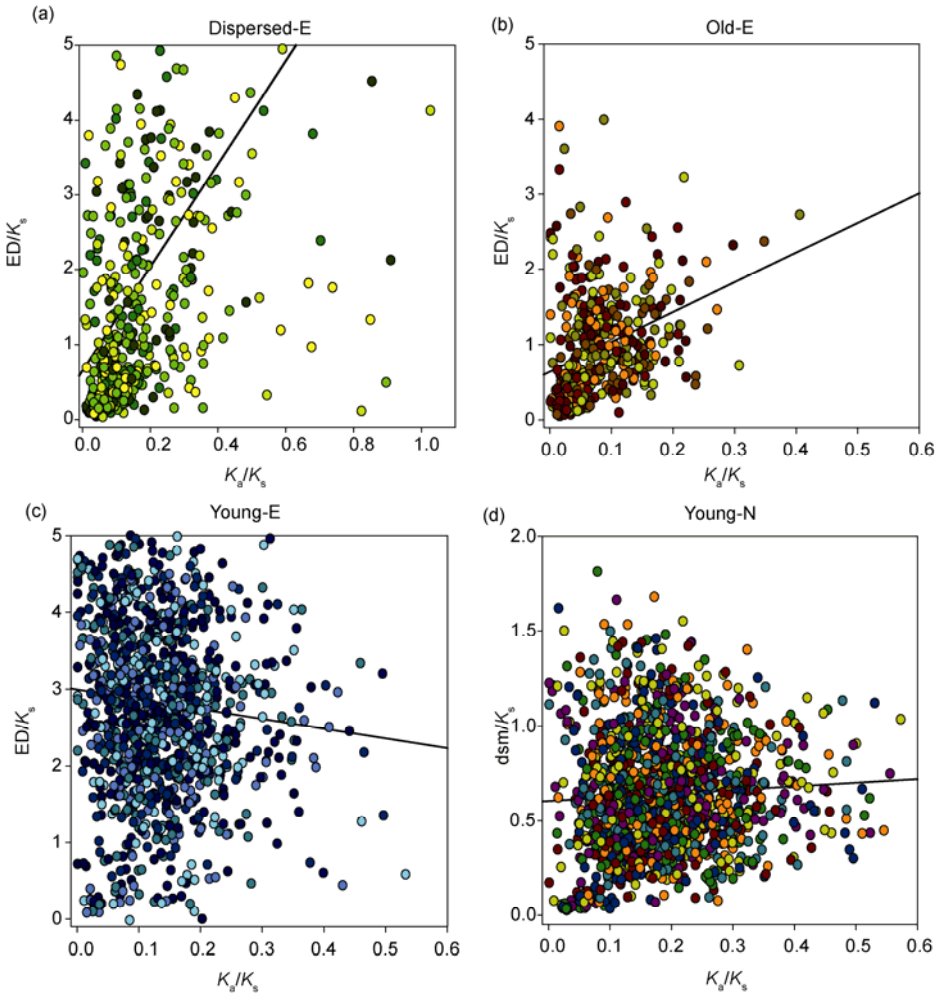

Figure 3 Correlation between the ED/K_s ratio and the K_a/K_s ratio. For Dispersed-E, Old-E pair groups, the log10 (expression difference)/K_s ratio is positively correlated with K_a/K_s ratio. For Young-E pairs, a negative correlation is found. For Young-N pairs, instead of ED/K_s ratio, a dsm/K_s ratio was used, and a positive correlation is found.

functional constraint in Young-E pairs indicates that diversified selection constrain act on *cis*-regulatory region and protein coding region as a whole. The changes of one region may relax the diversified selection pressure on another region.

The correlation between expression functional constraint and protein functional constraint also imply that duplicate genes' fate are largely determined by their nature [31,32], i.e. the functional importance of themselves at least on the genome-wide scale.

This work was supported by the National Basic Research Program of China (2007CB815701, 2009ZX08010-0178), National Natural Science Foundation of China (30730008, 40976081, 30970208). We thank Xionglei He, Renchao Zhou, Hua Bao for comments and valuable suggestions, and all members of the Shi lab for their support. We are grateful to David E. Boufford and anonymous reviewers for helping making revision of the manuscript.

1 Ohno S. Evolution by Gene Duplication. Berlin, New York: Springer-Verlag, 1970. 160
2 Zhang J Z. Evolution by gene duplication: An update. Trends Ecol Evol, 2003, 18: 292–298
3 Lockton S, Gaut B S. Plant conserved non-coding sequences and paralogue evolution. Trends Genet, 2005, 21: 60–65
4 Blanc G, Wolfe K H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell, 2004, 16: 1679–1691
5 Reinisch A J, Dong J M, Brubaker C L, et al. A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: Chromosome organization and evolution in a disomic polyploid genome. Genetics, 1994, 138: 829–847
6 Vision T J, Brown D G, Tanksley S D. The origins of genomic duplications in *Arabidopsis*. Science, 2000, 290: 2114–2117
7 Simillion C, Vandepoele K, Van Montagu M C, et al. The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci USA, 2002, 99: 13627–13632
8 Bowers J E, Chapman B A, Rong J, et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature, 2003, 422: 433–438
9 Blanc G, Hokamp K, Wolfe K H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res, 2003, 13: 137–144
10 Van De Peer Y, Taylor J S, Braasch I, et al. The ghost of selection past: Rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol, 2001, 53: 436–446

11 Wagner A. Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol, 2002, 19: 1760–1768
12 Haberer G, Hindemitt T, Meyers B C, et al. Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of Arabidopsis. Plant Physiol, 2004, 136: 3009–3022
13 Park C, Makova K D. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. Genome Biol, 2009, 10: R10
14 Lynch M, Conery J S. The evolutionary fate and consequences of duplicate genes. Science, 2000, 290: 1151–1155
15 Lynch M, Katju V. The altered evolutionary trajectories of gene duplicates. Trends Genet, 2004, 20: 544–549.
16 Casneuf T, De Bodt S, Raes J, et al. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol, 2006, 7: R13
17 Ganko E W, Meyers B C, Vision T J. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Bio Evol, 2007, 24: 2298–2309
18 Schmid M, Davison T S, Henz S R, et al. A gene expression map of *Arabidopsis thaliana* development. Nat Genet, 2005, 37: 501–506
19 Simpson P, Ayyar S. Evolution of *cis*-regulatory sequences in Drosophila. Adv Genet, 2008, 61: 67–106
20 Castillo-Davis C I, Hartl D L, Achaz G. *Cis*-regulatory and protein evolution in orthologous and duplicate genes. Genome Res, 2004, 14: 1530–1536
21 Hurles M. Gene duplication: The genomic trade in spare parts. PLoS Biol, 2004, 2: E206
22 Gu Z, Cavalcanti A, Chen F C, et al. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol Biol Evol, 2002, 19: 256–262
23 Thompson J D, Higgins D G, Gibson T J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, 22: 4673–4680
24 Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol, 1994, 11: 725–736
25 Yang Z. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci, 1997, 13: 555–556
26 Lynch M, Katju V. The altered evolutionary trajectories of gene duplicates. Trends Genet, 2004, 20: 544–549
27 Ganko E W, Meyers B C, Vision T J. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol, 2007, 24: 2298–2309
28 Pal C, Papp B, Hurst L D. Highly expressed genes in yeast evolve slowly. Genetics, 2001, 158: 927–931
29 Force A, Lynch M, Pickett F B, et al. Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 1999, 151: 1531–1545
30 Blanc G, Wolfe K H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell, 2004, 16: 1679–1691
31 Casneuf T, De Bodt S, Raes J, et al. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol, 2006, 7: R13
32 Maere S, De Bodt S, Raes J, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA, 2005, 102: 5454 –5459