



## Evolutionary conservation of neighbouring gene pairs in plants

Xiling Liu<sup>a,b</sup>, Bin Han<sup>a,\*</sup>

<sup>a</sup> National Center for Gene Research/Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China

<sup>b</sup> Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

### ARTICLE INFO

#### Article history:

Received 3 November 2008

Received in revised form 12 February 2009

Accepted 16 February 2009

Available online 2 March 2009

Received by J.G. Zhang

#### Keywords:

Convergent pair

Parallel pair

Divergent pair

Intergenic distance

Co-expression

Functional composition

### ABSTRACT

Evolutionary conservation of neighbouring gene pairs has been widely explored in many species, but remains poorly understood in plants. The availability of several plant genome sequences allows for an in-depth investigation of this problem in plants. Here, we analyzed the phylogenetic conservation of physically linked gene pairs in nine plant genomes and compared the conservation in different orientations. We also examined several potential determinants to detect whether they affect the conservation of neighbouring gene pairs. Our results suggested that among the three types of neighbouring gene pairs, closely linked parallel pairs might be the least conserved. Intergenic distance was shown to be a major determinant of linkage conservation, suggesting that the conservation of gene order in plants was determined primarily by chance. The enrichment of housekeeping genes was identified to contribute to the conservation of all three types and the enrichment of genes involved in protein metabolism might contribute to the conservation of parallel pairs. Moreover, a co-expressed signal was detected in conserved divergent pairs, which might be determined by intergenic distance.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

In prokaryotes, conserved genomic context has been widely used to predict functional associations between proteins (Huynen et al., 2000; Rogozin et al., 2002). The prediction is based on gene clusters found in operons where genes are usually functionally associated. In contrast, distribution of genes in eukaryotic genomes is assumed to be largely random because little phenotypic effect is documented from most chromosomal inversions (Hurst et al., 2004). However, recent studies in diverse eukaryotic species have challenged this perception, and shown that genes with similar expression profiles tended to cluster more frequently than expected (Hurst et al., 2004; Niehrs and Pollet, 1999; Zhan et al., 2006). In fact, about 43% and 65% of the analyzed pathways in *Arabidopsis* and human have genes exhibiting significant physical clustering than random distribution across the genome, respectively (Lee and Sonnhammer, 2003). The analysis of rice chromosome 4 tiling array also reveals a positive correlation between transcriptional activity and physical distance of gene on the chromosome (Jiao et al., 2005).

Because genes can be transcribed in either direction, neighbouring gene pairs can be orientated in three distinct types, including parallel

( $\leftarrow\leftarrow/\rightarrow\rightarrow$ ), divergent ( $\leftarrow\rightarrow$ ), and convergent ( $\rightarrow\leftarrow$ ) (Williams and Bowles, 2004). In prokaryotes, parallel pairs are well conserved because of the prevalent operons and divergent pairs are also maintained over several prokaryotic branches mainly for co-regulation, whereas convergent pairs are rapidly lost in evolution (Korbel et al., 2004). However, closely linked divergent pairs are abundant in mammalian genomes, while closely linked parallel pairs are rare (Franck et al., 2008). A study in human genome indicates that promoter segments in some of the divergent pairs can initiate transcription bi-directionally, which also might be the major determinant for maintenance of these pairs (Trinklein et al., 2004). Another study in yeasts reveals that the best predictor for gene order conservation is intergenic distance, followed by the density of essential genes and co-regulation (Poyatos and Hurst, 2007).

The above analyses of gene order conservation mainly depend on the completion of corresponding genome sequencing. However, only three higher plant genomes, including *Arabidopsis*, rice and poplar, have been completely sequenced before 2007. A recent study of the above three plant genomes reveals that the number of closely linked convergent pairs is more than divergent pairs, and conserved gene pairs are significantly more frequent when genes are co-expressed or have the same GO (Gene Ontology) class (Krom and Ramakrishna, 2008). However, parallel pairs are not discussed. Recently, another three plant genomes have been completely sequenced (grape, papaya, and *Physcomitrella patens*) (Jaillon et al., 2007; Ming et al., 2008; Rensing et al., 2008) and the genome sequences of several plants are incompletely assembled, which provide us with a substantially larger

Abbreviations: RMA, robust multi-array average; GO, gene ontology; BP, biological process.

\* Corresponding author. Tel.: +86 21 64845260; fax: +86 21 64825775.

E-mail address: [bhan@ncgr.ac.cn](mailto:bhan@ncgr.ac.cn) (B. Han).

data set to address the important question concerning evolution of neighbouring genes in plants.

In this study we detected the evolutionary conservation of neighbouring gene pairs in nine plant genomes, including four dicots (*Arabidopsis*, grape, papaya, poplar), two monocots (rice and *Sorghum*), a spikemoss (*Selaginella moellendorffii*), a moss species (*Physcomitrella patens*), and a green algae (*Chlamydomonas reinhardtii*). We compared the conservation of physically linked gene pairs in different orientations through plant genome evolution and then examined several potential determinants to detect whether they were responsible for conservation of neighbouring gene pairs in plants, using intergenic distance, gene co-expression, proportion of housekeeping genes, and functional composition as predictors.

## 2. Materials and methods

### 2.1. Protein sequence data

We downloaded protein sequences of nine plant species, including *Arabidopsis* (*Arabidopsis thaliana*; At), grape (*Vitis vinifera*; Vv), papaya (*Carica papaya*; Cp), poplar (*Populus trichocarpa*; Pt), rice (*Oryza sativa*; Os), *Sorghum* (*Sorghum bicolor*; Sb), *Selaginella moellendorffii* (Sm), *Physcomitrella patens* (Pp), *Chlamydomonas reinhardtii* (Cr) and the detailed downloading information is shown in Table S1. The annotation files containing genomic locations of each gene's coding sequence(s) and exon(s) were also downloaded.

The longest splicing variant was chosen when a gene contains more than one splicing variant. After removing splicing variants, 26,536 *Arabidopsis*, 28,377 papaya, 30,434 grape, 45,541 poplar, 55,890 rice, 35,895 *Sorghum*, 22,280 *Selaginella moellendorffii*, 35,935 *Physcomitrella patens*, and 15,144 *Chlamydomonas reinhardtii* protein sequences were retained.

### 2.2. Identification of conserved neighbouring gene pairs

To identify neighbouring gene pairs conserved in *Arabidopsis* and poplar, we first identified orthologs between them, using INPARANOID (version 1.35) program with default settings (Remm et al., 2001; Ren et al., 2007). The INPARANOID program employs BLASTP program (Altschul et al., 1997) to find reciprocal best hits and marks them as orthologs and then adds additional in-paralogs for each orthologous pair (Remm et al., 2001). Only reciprocal best hits as high-quality orthologs in each orthologous cluster were retained for further study (Huminiecki and Wolfe, 2004). The orthologs that are physically adjacent and arrange in the same orientation in both genomes were considered to be conserved neighbouring gene pairs.

The same processes were performed between *Arabidopsis* and each of grape, papaya, rice, *Sorghum*, *Selaginella moellendorffii*, *Physcomitrella patens*, and *Chlamydomonas reinhardtii*. Additionally, 10,861 orthologous clusters between *Arabidopsis* and rice are identified on the InParanoid website (version 6, <http://inparanoid.sbc.su.se/cgi-bin/summary.cgi>), which was almost consistent with what we had identified.

### 2.3. Expected number of conserved neighbouring gene pairs

The expected numbers of conserved neighbouring gene pairs were calculated by two methods. The expected number in the first method was measured by  $(N_c + N_p + N_d) \times f$ , where  $N_c$ ,  $N_p$ , and  $N_d$  were the numbers of conserved convergent, parallel and divergent pairs in each subgroup of a given species,  $f$  equals 25% for convergent or divergent orientation and 50% for parallel orientation. The expected number in the second method was measured by  $(N_c + N_p + N_d) \times \frac{T_i}{T_c + T_p + T_d}$ , where  $N_c$ ,  $N_p$ , and  $N_d$  were the numbers of conserved convergent, parallel and divergent pairs in each subgroup,  $T_c$ ,  $T_p$ , and  $T_d$  were the numbers of all convergent, parallel and divergent pairs in this

subgroup,  $i$  represented  $c$ ,  $p$ , or  $d$  when expected number of convergent, parallel, or divergent pairs was calculated.

### 2.4. Intergenic distance and removal of overlapping genes

The intergenic distance is defined as the length in nucleotides between the annotated coding sequence end of one gene and the annotated coding sequence start of the next gene (Dahary et al., 2005; Ren et al., 2007). Overlapping genes were identified on the basis of the annotated genomic loci of mRNA and were removed from this analysis.

### 2.5. Microarray data

We selected *Arabidopsis* Affymetrix microarray representing 149 conditions of the wild type Columbia (Col-0) from TAIR (<ftp://ftp.arabidopsis.org/home/tair/Microarrays>) (see Table S2). Each chip was filtered to remove the probes which represent more than one gene. If a gene is represented more than once, we removed at first the lower confidence probe(s) of this gene (Affymetrix probes with name “\_x\_at” or “\_s\_at”, Huminiecki and Wolfe, 2004 and Shoja et al., 2007) and then randomly selected one to represent it if it still had more than one probe after removing the lower confidence probe(s). The data were normalized for further calculating Pearson correlation coefficient, using robust multi-array average (RMA) method (Bolstad et al., 2003). The nonparametric test (Wilcoxon signed-rank test) was used to compare the distributions of Pearson correlation coefficients of conserved gene pairs with two control sets: 1) all neighbouring gene pairs with the same orientation; 2) 10,000 randomly selected pairs of genes with no need of neighbourhood between them. If a gene pair with at least one member was not represented in the chips, we would not be able to calculate the Pearson correlation coefficient of it.

We also used this expression data to calculate the number of chips where each gene is expressed, based on the criterion for the detection of expression from the Affymetrix “present” calls (Zhang et al., 2005). Genes expressed in all 149 chips were considered to be housekeeping genes (Lercher et al., 2002).

### 2.6. GO functional categorization

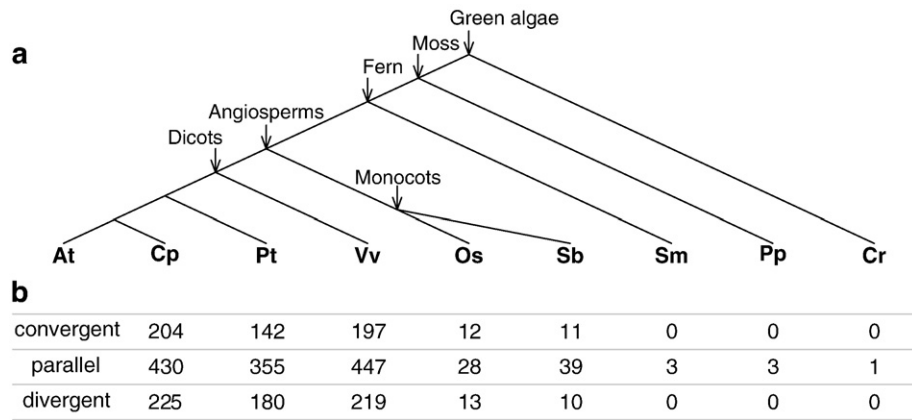
GO annotation of *Arabidopsis* was downloaded from TAIR (<ftp://ftp.arabidopsis.org/home/tair/Ontologies>). The biological process (BP) section of plant-related GO Slim terms was used (Berardini et al., 2004). We calculated the  $p$ -value of each GO Slim term, using hypergeometric distribution, to determine whether this GO Slim term is overrepresented by conserved gene pairs than expected.  $p$ -Values were adjusted with Bonferroni correction for multiple tests.

## 3. Results

### 3.1. Conservation of neighbouring gene pairs

To systematically identify conserved neighbouring gene pairs across plant lineages, we at first identified one-to-one orthologs between *Arabidopsis* and each of three dicots (grape, papaya, poplar), two monocots (rice and *Sorghum*), a spikemoss (*Selaginella moellendorffii*), a moss species (*Physcomitrella patens*), and a green algae (*Chlamydomonas reinhardtii*) (Fig. 1a). In total, 11,731 *Arabidopsis*–papaya, 12,866 *Arabidopsis*–poplar, 10,680 *Arabidopsis*–rice, 11,653 *Arabidopsis*–grape, 10,643 *Arabidopsis*–*Sorghum*, 6916 *Arabidopsis*–*Selaginella moellendorffii*, 6930 *Arabidopsis*–*Physcomitrella patens*, and 3806 *Arabidopsis*–*Chlamydomonas reinhardtii* orthologs were identified.

With the evolutionary distances from these species to *Arabidopsis* increased, the numbers of conserved gene pairs decreased, whereas the number in poplar was out of our expectation (Fig. 1b). Poplar



**Fig. 1.** Number of *Arabidopsis* neighbouring gene pairs conserved across plant evolution. (a) A phylogenetic tree of the nine selected plants. Eight plants, including three dicots (papaya, poplar, grape), two monocots (*Sorghum*, rice), one fern (*Selaginella moellendorffii*), one moss (*Physcomitrella patens*), and one green algae (*Chlamydomonas reinhardtii*) were used to compare with the dicot *Arabidopsis* (Hedges, 2002; Tang et al., 2008). At, Cp, Pt, Vv, Os, Sb, Sm, Pp, and Cr represent *Arabidopsis*, papaya, poplar, grape, rice, *Sorghum*, *Selaginella moellendorffii*, *Physcomitrella patens*, and *Chlamydomonas reinhardtii*, respectively. (b) Numbers under each of the eight plants represent the numbers of *Arabidopsis* neighbouring gene pairs that conserved in the given species.

diverged from *Arabidopsis* about 100–120 million years ago and the divergence between grape and *Arabidopsis* happened earlier (Tuskan et al., 2006; Tang et al., 2008), but the number of conserved gene pairs were larger in grape than in poplar. An additional chromosomal duplication named as salicoid duplication *p* has been reported in poplar (Tang et al., 2008). Once genome undergoes polyploidization, genome rearrangement for diploidization subsequently happens and this process might be used to explain for this exception (Bowers et al., 2003). The conserved neighbouring gene pairs decreased rapidly in monocots and even more rapidly through fern to green algae. In addition, only a few pairs, all of which were with parallel orientation, could be identified in *Selaginella moellendorffii*, *Physcomitrella patens* and *Chlamydomonas reinhardtii*.

If orientation of conserved gene pairs is random, the frequency of convergent, parallel and divergent pairs should be 25%, 50%, and 25% (Franck et al., 2008). However, this assumption would be accurate when the intergenic distance of these three types were similar. So we divided conserved neighbouring gene pairs into subgroups by intergenic distance and supposed the frequency of convergent, parallel and divergent pairs in each subgroup is 25%, 50%, and 25% to calculate the expected number of neighbouring gene pairs (see Materials and methods 2.3). Since most of the conserved neighbouring gene pairs located with intergenic distance  $\leq 1$  kb, we were only concerned with gene pairs with intergenic distance  $\leq 1$  kb when comparing the linkage conservation among convergent, parallel, and divergent pairs. Furthermore, the expected numbers were only calculated in dicots, considering the tiny amount of *Arabidopsis* neighbouring gene pairs conserved in monocots in each subgroup. When the intergenic distance was  $\leq 500$  bp, the number of conserved convergent pairs was higher than expected in all dicots (Table 1). With the intergenic distance increases, there were still two dicots having a higher number of conserved convergent pairs than expected, whereas the numbers of both conserved parallel and divergent pairs were mostly lower than expected. These observations implied that conserved closely linked convergent pairs were enrichment in dicots. However, 32% (3057 of 9616) *Arabidopsis* neighbouring gene pairs with intergenic distance  $\leq 1$  kb were with convergent orientation, which is higher than the supposed frequency (25%) we used in the above method. Hence the enrichment of conserved convergent pairs might be caused by the relative higher frequency of closely linked convergent pairs in *Arabidopsis* genome, but not by well conservation of these pairs. To test this, we applied another method to calculate the expected number based on the number of all convergent, parallel, and divergent pairs in *Arabidopsis* genome within each subgroup (see Materials and methods 2.3). The number of conserved convergent

pairs in ten subgroups was higher than expected when using the first method, whereas in four of the ten subgroups the number was no longer higher than expected when using the second method (Table 1). This observation implied that the enrichment of conserved closely linked convergent pairs derived from the first method was mainly caused by the relative higher frequency of all closely linked convergent pairs in *Arabidopsis* genome. Similarly, three and two subgroups with parallel orientation could be detected to have number higher than expected in the first and second method, respectively. Moreover, the number in five more subgroups with divergent orientation was higher than expected in the second method.

The different numbers of all convergent, parallel, and divergent pairs in other genomes might also affect the comparison of gene order conservation. To avoid this, we selected conserved neighbouring gene pairs between two species with intergenic distance  $\leq 1$  kb in both species and then estimated the percentage of conserved neighbouring gene pairs, in the way of dividing the number of conserved gene pairs by the average total number of neighbouring gene pairs in the two species (Franck et al., 2008; Fig. S1). In all five angiosperms, the percentage of conserved parallel pairs was lower than both of conserved convergent pairs and divergent pairs (Fig. 2), implying that closely linked parallel pairs might be the least conserved. However, there was little difference between the percentages of conserved convergent pairs and divergent pairs.

In prokaryotes, gene pairs conserved in three or more genomes are more likely to be functionally relevant (Rogozin et al., 2002). Using gene pairs conserved in at least three species, instead of two, could partially remove gene pairs that retained by chance. In total, we identified 140 highly conserved convergent pairs, 310 highly conserved parallel pairs, and 153 highly conserved divergent pairs in *Arabidopsis*. These gene pairs were further analyzed for estimating several factors that might be response for the conservation of neighbouring gene pairs.

### 3.2. Intergenic distance

Considering the continuous genomic reshuffling, linked genes with farther distances should have higher chances of rearrangement which will eventually break down the linkage between them. Therefore, conserved neighbouring gene pairs are expected to possess short intergenic distance.

Notably, the conserved pairs with all three orientations have significant shorter intergenic distance, compared with all neighbouring gene pairs with corresponding orientation (Kolmogorov–Smirnov test, all with  $p < 0.01$ ). For instance, about 74% (104 of 140 pairs) of the

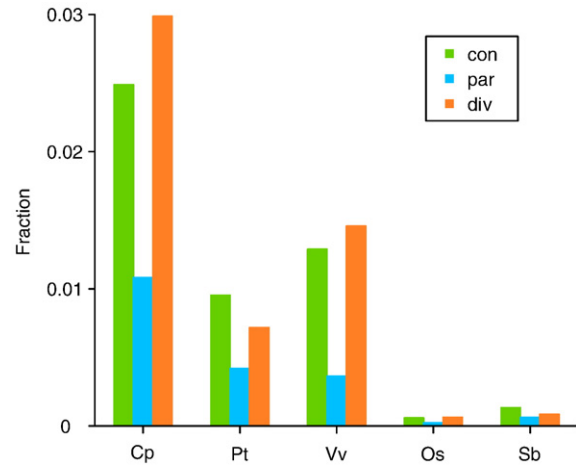
**Table 1**  
Number of *Arabidopsis* neighbouring gene pairs with intergenic distance  $\leq 1$  kb conserved in dicots

	$\leq 250^a$ :1:2:1 <sup>b</sup>		250–500		500–750		750–1k		$\leq 250$		250–500		500–750		750–1k	
	1:2:1 <sup>b</sup>		1:2:1		1:2:1		1:2:1		268:130:101		1167:1250:551		1044:1828:554		578:1660:485	
	Num	Exp	Num	Exp	Num	Exp	Num	Exp	Num	Exp	Num	Exp	Num	Exp	Num	Exp
Cp	Con	<b>21</b> <sup>c</sup>	9	38	<b>51</b>	20	<b>52</b>	29	21	60	51	60	52	45	29	17
	Par	7	18	53	76	73	41	40	7	9	53	64	73	78	40	49
	Div	8	9	<b>48</b>	38	22	20	12	<b>8</b>	28	<b>48</b>	28	22	24	12	14
Pt	Con	<b>10</b>	7	<b>47</b>	33	26	29	15	10	47	47	52	26	36	15	14
	Par	8	13	52	66	<b>68</b>	59	35	<b>8</b>	55	52	<b>68</b>	62	35	35	40
	Div	<b>8</b>	7	32	33	23	16	15	<b>8</b>	24	<b>32</b>	24	<b>23</b>	19	15	12
Vv	Con	<b>15</b>	8	<b>51</b>	40	<b>45</b>	34	<b>26</b>	15	51	51	63	45	41	<b>26</b>	21
	Par	8	17	63	80	<b>70</b>	64	<b>54</b>	8	67	63	70	70	72	54	60
	Div	<b>10</b>	8	<b>45</b>	40	20	34	18	<b>10</b>	30	<b>45</b>	30	20	22	18	17

<sup>a</sup> " $\leq 250$ " represents gene pairs with intergenic distance  $\leq 250$  bp, while "250–500" represents gene pairs with intergenic distance  $> 250$  bp and  $\leq 500$  bp. "500–750 bp" and "750–1k" are similar.

<sup>b</sup> The proportion of convergent, parallel, and divergent pairs. The expected number was calculated based on this proportion (see Materials and methods 2.3).

<sup>c</sup> The number is in bold when it is higher than expected.



**Fig. 2.** Percentage of *Arabidopsis* neighbouring gene pairs with intergenic distance  $\leq 1$  kb conserved in other angiosperms. The conserved neighbouring gene pairs with intergenic distance  $\leq 1$  kb in both species were selected. For each type, the percentage of conserved gene pairs between two species was calculated in the way of dividing the number of conserved gene pairs by the average total number of gene pairs with intergenic distance  $\leq 1$  kb in the two species.

conserved convergent pairs had intergenic distance  $\leq 1$  kb, comparing with 58% (3057 of 5269 pairs) of all convergent pairs with intergenic distance  $\leq 1$  kb (Fig. 3a). For parallel pairs, 46% (142 of 310 pairs) of conserved pairs had intergenic distance  $\leq 1$  kb, while the proportion of all parallel pairs with intergenic distance  $\leq 1$  kb was 35% (4868 of 13,838 pairs) (Fig. 3b). These two proportions of divergent pairs with intergenic distance  $\leq 1$  kb were 51% (78 of 153 pairs) and 27% (1691 of 6284 pairs), respectively (Fig. 3c). Observations above indicated that intergenic distance was a common determinant for conservation of those gene pairs with all three types.

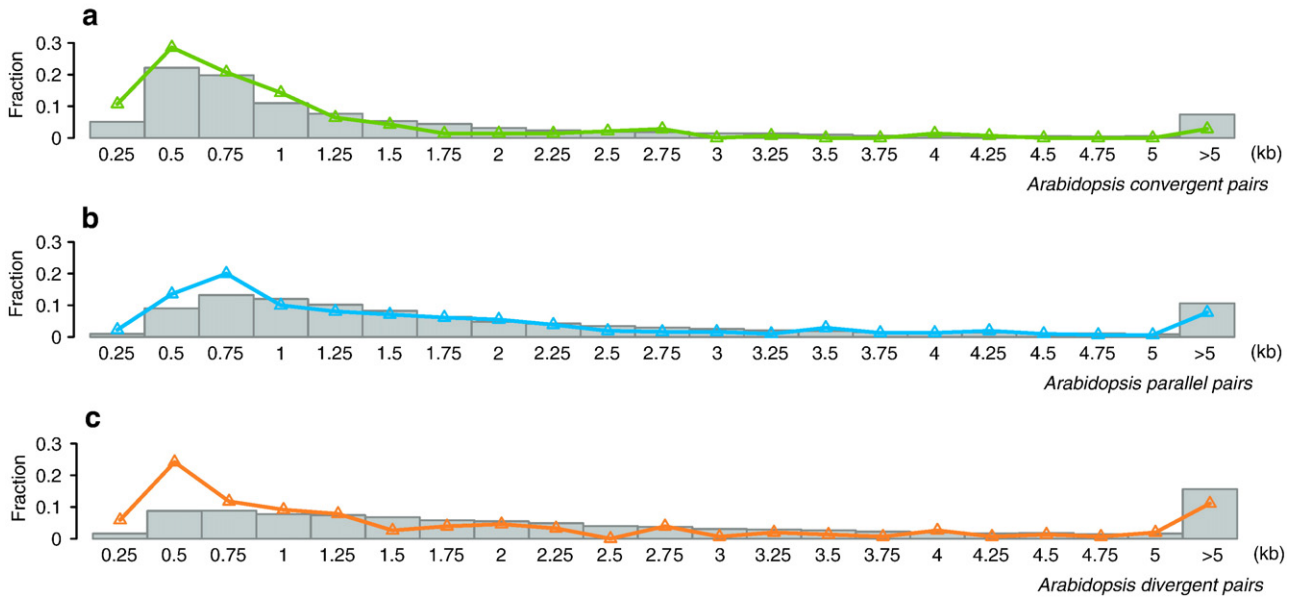
Conserved convergent pairs show significant shorter intergenic distances than the other two types, and conserved divergent pairs have significant shorter intergenic distance than parallel pairs (Kolmogorov–Smirnov test, all with  $p < 0.01$ ). This could explain why the numbers of convergent pairs with intergenic distance  $\leq 1$  kb were mostly higher than expected when we supposed the frequencies of convergent, parallel and divergent pairs are 25%, 50%, and 25% in Result 3.1.

### 3.3. Expression correlation

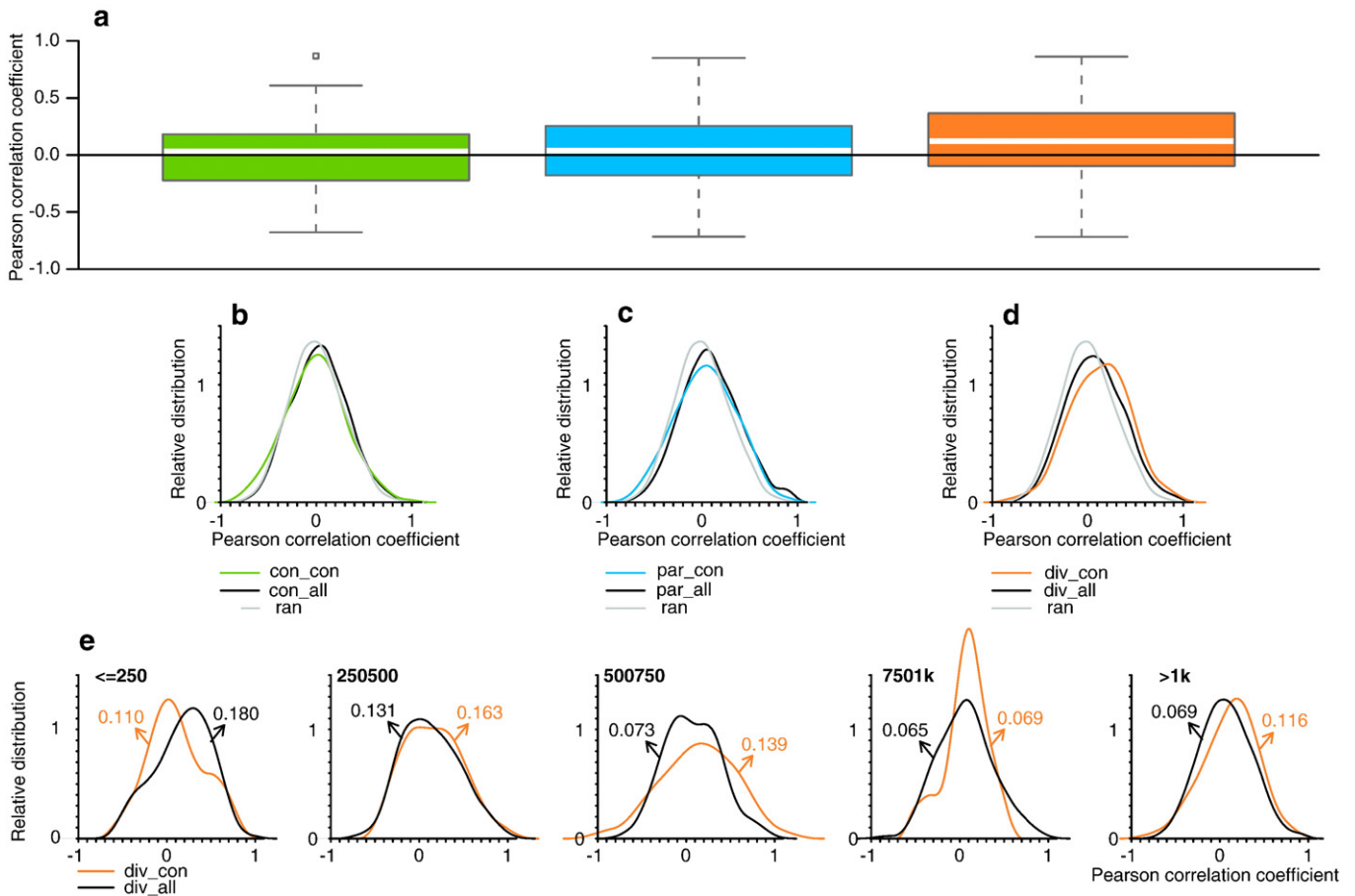
Co-expression has been revealed to be important for divergent pair maintenance in prokaryotes (Korbel et al., 2004) and neighbouring gene pairs in *Arabidopsis* are co-expressed more frequently than expected (Williams and Bowles, 2004; Ren et al., 2005). This led us to address the question of whether co-expression is also important for conservation of neighbouring gene pairs in plants.

We selected *Arabidopsis* Affymetrix microarray representing 149 conditions of the wild type Columbia (Col-0). Of the 26,536 *Arabidopsis* genes, 20,906 had effective probes in these chips (see Materials and methods 2.5). To avoid the co-expressed signals of linked genes independent of conservation, we compared the co-expression of conserved neighbouring gene pairs not only with 10,000 randomly selected pairs of genes but also with all neighbouring gene pairs.

The mean Pearson correlation coefficient of conserved divergent pairs was higher than that of either conserved parallel or conserved convergent pairs (Fig. 4a). It turned out that the conserved divergent pairs were more likely to be co-expressed than all divergent neighbouring pairs as well as 10,000 randomly selected pairs of genes (Wilcoxon signed-rank test,  $p < 0.05$  and  $p < 0.01$ , Fig. 4d), implying that co-expression might contribute to the conservation of divergent pairs. Of 153 conserved divergent pairs, 27 pairs missed the



**Fig. 3.** Distribution of intergenic distances of conserved neighbouring gene pairs in *Arabidopsis*. Histograms of the distribution of intergenic distances of conserved neighbouring gene pairs (dots) and all neighbouring gene pairs with corresponding orientation (bars) are shown. The green (a), blue (b), and orange dots (c) represent conserved gene pairs with convergent, parallel, and divergent orientation, respectively. The numbers under bars (dots) represent intergenic distances. For instance, the bar (dot) located under 0.5 shows the fraction of gene pairs with intergenic distance >0.25 kb and ≤0.5 kb.



**Fig. 4.** Expression correlations of conserved neighbouring gene pairs in *Arabidopsis*. (a) Average Pearson correlation coefficients of conserved neighbouring gene pairs. (b) Density curves are shown for 10,000 pairs of genes randomly selected, all convergent neighbouring gene pairs, and the highly conserved convergent pairs. (c) and (d) represent the expression correlation of gene pairs with parallel and divergent orientation, respectively. The distribution of Pearson correlation coefficients of conserved or all divergent pairs in each subgroup is shown in (e). The mean Pearson correlation coefficient was labelled behind the density curve.

Pearson correlation coefficient and 45 pairs lacked significant Pearson correlation coefficient. Of the 81 remaining pairs with significant Pearson correlation coefficient, 69 pairs were positively correlated and 22 pairs were negatively correlated. However, only 11 pairs with Pearson correlation coefficient  $>0.5$  were highly co-expressed and 2 with Pearson correlation coefficient  $<-0.5$  were highly anti-regulated (Krom and Ramakrishna, 2008). That only a few pairs were highly co-expressed implied that co-expression might play little role in conservation of divergent pairs.

Studies in mammals reveal that adjacent gene pairs with closer distance are more likely to be co-expressed (Semon and Duret, 2006; Purmann et al., 2007). A study in plants observes that closely linked divergent and convergent pairs show strong positive correlation and another study shows that *Arabidopsis* genome contains islands of co-expressed genes (Krom and Ramakrishna, 2008; Zhan et al., 2006). In our study, conserved divergent pairs were revealed to have shorter intergenic distance than that of non-conserved divergent pairs (Kolmogorov–Smirnov test,  $p<0.01$ ). All these reminded us that the contribution to conservation of divergent pairs made by co-expression might overlap with that of short intergenic distance of these pairs. To investigate this, we divided both conserved divergent pairs and all divergent pairs into five subgroups by intergenic distances. Although the mean Pearson correlation coefficient of conserved pairs is higher than that of all pairs in four subgroups (Fig. 4e), no significant co-expressed signal could be detected in conserved divergent pairs in these subgroups. These observations implied that co-expression might be just a by-product of the determinant of intergenic distance. In contrary, the subgroup with closest linkage ( $\leq 250$  bp) had a low Pearson correlation coefficient compared to all divergent pairs with intergenic distance  $\leq 250$  bp, which might be explained by the conclusion that co-expression of linked genes is disadvantageous derived in a recent study in mammalian (Liao and Zhang, 2008).

For the conserved convergent pairs, no significant co-expressed signal was detected when compared with all convergent neighbouring pairs or randomly selected pairs of genes (Fig. 4b). The conserved parallel pairs also lacked a significant co-expressed signal when compared with all parallel neighbouring pairs, although they have a significant higher co-expressed signal comparing with randomly selected pairs of genes (Wilcoxon signed-rank test,  $p<0.05$ , Fig. 4c). Thus, co-expression was not the cause for conservation of parallel pairs and convergent pairs.

### 3.4. Housekeeping genes

Housekeeping genes are constitutively expressed to maintain cellular function (Butte et al., 2001), which have been revealed to evolve more slowly than tissue-specific genes in mammals (Zhang and Li, 2004). An interesting question is whether housekeeping genes are well conserved in plants. If they are, we might expect to detect an enriched signal of housekeeping genes in conserved gene pairs.

Considering genes expressed in all 149 chips to be housekeeping genes (Lercher et al., 2002), 3527 (17%) of the 20906 *Arabidopsis* genes were housekeeping genes. To explore whether there was an enrichment of housekeeping genes in conserved gene pairs, the percentage of housekeeping genes in a given type of conserved gene pairs was compared with that of all neighbouring gene pairs of the same type to reduce the bias of different distribution of different orientation types. For all three types, percentages of housekeeping genes in conserved pairs are significantly higher than that in the entire neighbouring gene pairs (Table 2, parallel pairs with  $p<0.01$  and the others with  $p<0.05$ ).

However, another study suggests that housekeeping genes are not necessarily the genes that expressed in all chips and genes expressed in most chips could also be considered as housekeeping genes (Zhang and Li, 2004). Moreover, housekeeping genes tend to cluster in gene-dense regions in human genome (Lercher et al., 2002). Therefore, the

**Table 2**  
Number of housekeeping genes in conserved neighbouring gene pairs

	Conserved		All		$p^a$
	Num (percent) of hk	Num of all	Num (percent) of hk	Num of all	
Convergent	54(22.3%)	238	1361(16.9%)	8063	2.4e-02*
Parallel	114(21.2%)	538	2578(16.8%)	15,347	9.0e-03**
Divergent	66(23.8%)	277	1791(18.3%)	9811	2.3e-02*

<sup>a</sup>  $p$ -value of Chi-squared test is shown.

\*  $p<0.05$ .

\*\*  $p<0.01$ .

enrichment of housekeeping genes might be caused by the shorter intergenic distances of conserved gene pairs. To avoid the error that might be caused by the strict definition of housekeeping genes and reduce the bias caused by shorter intergenic distance of conserved gene pairs, we divided conserved pairs and all pairs into five subgroups by intergenic distances and observed the distributions of the numbers of chips that gene expressed in conserved pairs and all pairs in each subgroup. For four subgroups with convergent orientation except the subgroup with intergenic distance between 500–750 bp, two peak values were observed when all genes accounted: expressed in few chips and in most chips, while all conserved pairs were almost only enriched for genes that expressed in most chips (Fig. 5a, Kolmogorov–Smirnov test, one with  $p<0.01$  and three with  $p<0.05$ ). Similar results were also observed in four parallel subgroups except the subgroup with intergenic distance  $\leq 250$  bp (Fig. 5b) and two divergent subgroups except subgroups with intergenic distance  $\leq 250$  bp, between 500–750 bp and 750 bp–1 kb. In the five subgroups which we failed to detect significant different distributions between conserved pairs and all pairs, the mean number of chips that gene expressed in conserved gene pairs were all higher than that in corresponding all gene pairs. These observations suggested that the enrichment of housekeeping genes contributed to the conservation of all three types and this factor was affected by the determinant of intergenic distance.

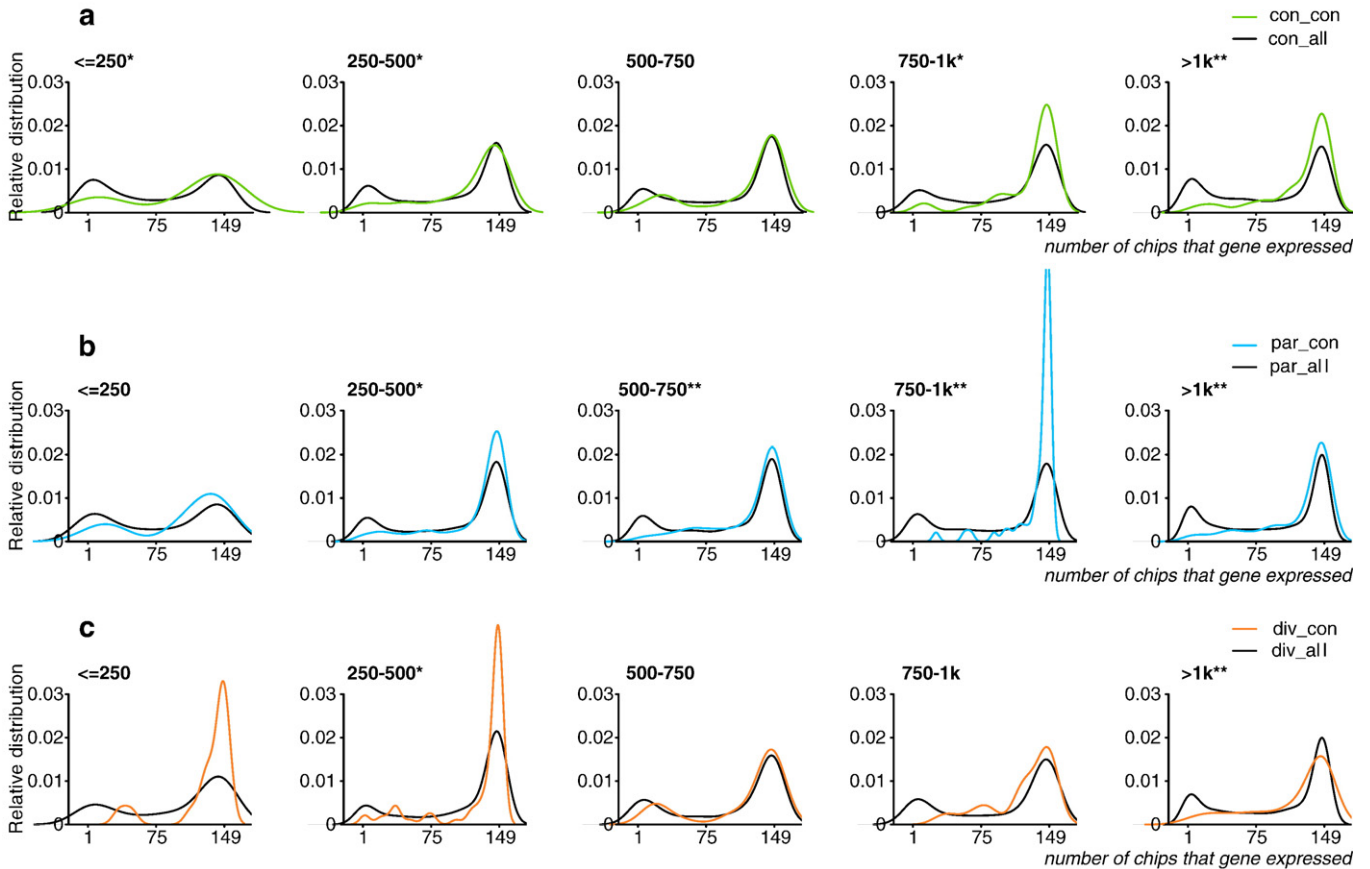
### 3.5. Functional compositions

To find out more specific functional enrichment that might affect the conservation of neighbouring gene pairs in plants, we used the biological process (BP) section of Gene Ontology (GO) terms. In total, 23,340 of 26,536 genes had at least one GO Slim term in the biological process section. When we calculated the statistical significance of functional enrichments in each GO Slim term, the conserved neighbouring gene pairs were compared with the total neighbouring gene pairs with the corresponding orientation to reduce the bias of different distributions of orientations.

After excluding the “other” subcategories, a very weak enrichment of genes involved in protein metabolism is observed in conserved parallel pairs (Table 3,  $p=0.17$ ). Of the 542 genes from conserved parallel pairs, 94 were involved in protein metabolism, implying that participating in protein metabolism might contribute to conservation of parallel pairs. However, the conserved convergent pairs and divergent pairs were found with no significant functional enrichment.

## 4. Discussion

In this study, we conducted a systematic investigation of evolutionary conservation of neighbouring gene pairs in plants. Our study implied that among the three types of neighbouring gene pairs, closely linked parallel pairs might be the least conserved and none of the three types had been obviously supported to be conserved by selection. Evolutionary conservations in prokaryotes or mammals are different from those we observed in plants. In prokaryotes, parallel



**Fig. 5.** Distribution of numbers of chips that gene expressed in conserved neighbouring gene pairs. Genes in all neighbouring gene pairs and in conserved gene pairs were divided into five subgroups by intergenic distances. In each subgroup with a given type, distribution of numbers of chips that gene expressed in conserved neighbouring gene pair was compared with that in all neighbouring gene pairs. The *p*-value was labelled. \*: *p* < 0.05; \*\*: *p* < 0.01.

pairs are both widely conserved and enriched (Korbel et al., 2004). In the mammals, the closely linked divergent pairs are the most abundant and the accumulation of these pairs starts earlier than the others, while the parallel pairs are rare because of latter start of accumulation (Franck et al., 2008).

The strong conservation of parallel pairs in prokaryotes is due to the prevalent operons (Korbel et al., 2004). Genes in prokaryotic operons are transcribed as polycistronic mRNA from a shared promoter. Besides prokaryotes, it has been reported to date that

polycistronic transcriptions have occurred frequently only in trypanosomes (almost all genes) and nematodes (about 25% genes) (Matthews et al., 1994; Blumenthal, 1995; Qian and Zhang, 2008). However, operons are not prevalent in plants (Blumenthal, 1998), which might explain the moderate level of conservation of parallel pairs in plants. Only 60 bicistronic transcripts are identified through systematic scanning of all full-length cDNA in *Arabidopsis* (Thimmapuram et al., 2005). In addition, three highly conserved parallel pairs were on the list of the 60 *Arabidopsis* operons.

**Table 3**  
Functional compositions of conserved neighbouring gene pairs

GO Slim term	Convergent (num)			Parallel (num)			Divergent (num)		
	Con <sup>a</sup>	All	<i>p</i> <sup>b</sup>	Con	All	<i>p</i>	Con	All	<i>p</i>
Transport	17	631	1	39	1233	1	18	824	1
Electron transport or energy pathways	4	252	1	16	593	1	11	323	1
Signal transduction	14	341	1	22	631	1	7	425	1
transcription	15	775	1	35	1266	1	14	777	1
Cell organization and biogenesis	18	563	1	40	1064	1	21	727	1
Developmental processes	17	433	1	34	840	7.5e-01	11	537	1
DNA or RNA metabolism	3	149	1	14	288	7.8e-01	9	225	1
Protein metabolism	43	1284	1	94	2469	1.7e-01	29	1495	1
response to stress	10	342	1	31	721	4.6e-01	16	440	8.6e-01
Response to abiotic or biotic stimulus	14	406	1	37	862	3.1e-01	15	501	1
Other cellular processes	104	3079	6.3e-02	200	5829	3.1e-01	78	3538	1
Other biological processes	156	6192	1	347	11,568	1	167	7297	1
Other metabolic processes	108	3239	7.8e-02	207	6237	9.5e-01	94	3749	1
Total <sup>c</sup>	253	9260	–	542	17,743	–	267	11,080	–

<sup>a</sup> Con represents conserved neighbouring gene pairs.

<sup>b</sup> *p*-value was calculated using hypergeometric distribution and then adjusted with Bonferroni correction.

<sup>c</sup> Only genes that own GO Slim term in biological process section were accounted.

Co-expression is the determinant for conservation of divergent pairs in prokaryotes (Korbel et al., 2004). A higher co-expressed signal was detected in conserved divergent pairs in plants, suggesting implying that co-expression might also be responsible for conservation of these pairs. However, after the intergenic distance was controlled, no significant co-expression signal was detected, suggesting that the higher co-expressed signal might just be a by-product of shorter intergenic distance of these gene pairs, but not be the cause for conservation of them. A recent study in mammalian genomes proposes that co-expression of linked genes is a form of transcriptional interference and is disadvantageous (Liao and Zhang, 2008). Interestingly, the closest linked subgroup of divergent pairs ( $\leq 250$  bp) in our study had a low Pearson correlation coefficient compared to all divergent pairs with intergenic distance  $\leq 250$  bp, which was consistent with the disadvantageous model of co-expression.

Intergenic distance was detected to be a common determinant of conservation of neighbouring gene pairs. Although enrichment of housekeeping genes was contribute to the conservation of all three types, this factor was partially affected by the determinant of short intergenic distance of conserved pairs, implying that the intergenic distance was the major determinant for conservation of neighbouring gene pairs. Genes with shorter intergenic distance have lower chance for their linkage to be broken during evolution. Hence, the major role of intergenic distance in conservation of neighbouring gene pairs implied that gene order evolution in plants was determined primarily by chance.

Recently, three adjacent genes with sequential roles in metabolism have been reported, including two parallel pairs (At5g57590 and At5g57600 involved in biosynthesis of biotin; At1g48850 and At1g48860 involved in aromatic amino acid biosynthesis) and one divergent pair (At3g06850 and At3g06860 involved in the branched-chain amino acids metabolism) (Muralla et al., 2008). The first parallel pair (At5g57590 and At5g57600) was also conserved in poplar (eugene3.01710033 and fgenes4\_pg.C\_scaffold\_171000027), implying that co-functionality might be a mechanism for conservation of some certain neighbouring gene pairs.

Although a recent study identifies and analyzes certain conserved divergent and convergent pairs in *Arabidopsis*, poplar, and rice (Krom and Ramakrishna, 2008), this study could shed new light on conservation of neighbouring gene pair in plants in four aspects. First of all, we detected the evolutionary conservation in nine plant genomes, including four dicots, two monocots, one spikemoss, one moss species, and one green alga. Secondly, the conserved neighbouring gene pairs were identified based on one-to-one orthologs between species and this could eliminate gene pairs resulted from duplication but not conservation. Additionally, we investigated all three types of neighbouring gene pairs. Finally, we did not only focus on finding special conserved gene pairs, but also analyzed the evolutionary conservation of neighbouring gene pairs and determinants of the conservation in three types of gene pairs.

In addition, we removed overlapping genes in this study, because breakage of an overlapping gene pair actually means truncation of a gene and it is very unlikely to happen during evolution. However, we also had done another run of analysis with overlapping genes included in to observe how the results affected by removing of overlapping genes. Although the results regarding the determinants for gene order conservation remained mostly unchanged, convergent pairs seemed to be more conserved when we kept overlapping genes, which might be attributed to the fact that most of the overlapping pairs in *Arabidopsis* had a convergent orientation.

In this study, we identified 140 highly conserved convergent pairs, 310 highly conserved parallel pairs, and 153 highly conserved divergent pairs. Although the conservation of neighbouring gene pairs was mainly caused by their shorter intergenic distance, conserved neighbouring gene pairs were enriched for housekeeping

genes. In addition, conserved parallel pairs might be enriched for genes involved in protein metabolism. Even for the most studied plant—*Arabidopsis*, some genes in the genome still lack predicted function annotation. Detecting the evolutionary conservation of neighbouring gene pairs may facilitate the functional annotation. Therefore, our work will not only extend our understanding of gene order conservation, but also improve the genome annotation in plants.

## Acknowledgments

We would like to acknowledge Yujun Zhang and two anonymous reviewers for helpful suggestions and critical reading of the manuscript. We also acknowledge Wenhao Hu for significant assistance with microarray data analysis. We thank Tao Sang for his help on the manuscript preparation. This research was supported by the grants from the Ministry of Science and Technology of China (The China Rice Functional Genomics Programs, grant numbers 2006AA10A102), the Chinese Academy of Sciences (grant no. KSCW2-YW-N-024), and the National Natural Science Foundation of China (Grant No. 30821004).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.02.012.

## References

- Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Berardini, T.Z., et al., 2004. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* 135, 745–755.
- Blumenthal, T., 1995. Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* 11, 132–136.
- Blumenthal, T., 1998. Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* 20, 480–487.
- Bolstad, B.M., Trizarray, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193.
- Bowers, J.E., Chapman, B.A., Rong, J., Paterson, A.H., 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.
- Butte, A.J., Dzau, V.J., Glueck, S.B., 2001. Further defining housekeeping, or "maintenance," genes focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics* 7, 95–96.
- Dahary, D., Elroy-Stein, O., Sorek, R., 2005. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res.* 15, 364–368.
- Franck, E., Hulsen, T., Huynen, M.A., de Jong, W.W., Lubsen, N.H., Madsen, O., 2008. Evolution of closely linked gene pairs in vertebrate genomes. *Mol. Biol. Evol.* 25, 1909–1921.
- Hedges, S.B., 2002. The origin and evolution of model organisms. *Nat. Rev., Genet.* 3, 838–849.
- Huminiecki, L., Wolfe, K.H., 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* 14, 1870–1879.
- Hurst, L.D., Pal, C., Lercher, M.J., 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev., Genet.* 5, 299–310.
- Huynen, M., Snel, B., Lathe III, W., Bork, P., 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210.
- Jaillon, O., et al., 2007. French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.
- Jiao, Y.L., et al., 2005. A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. *Plant Cell* 17, 1641–1657.
- Krom, N., Ramakrishna, W., 2008. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, *Arabidopsis*, and *Populus*. *Plant Physiol.* 147, 1763–1773.
- Korbel, J.O., Jensen, L.J., von Mering, C., Bork, P., 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* 22, 911–917.
- Lee, J.M., Sonnhammer, E.L.L., 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13, 875–882.
- Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* 31, 180–183.
- Liao, B., Zhang, J., 2008. Coexpression of linked genes in mammalian genomes is generally disadvantageous. *Mol. Biol. Evol.* 25, 1555–1565.

- Matthews, K.R., Tschudi, C., Ullu, E., 1994. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev.* 8, 491–501.
- Ming, R., et al., 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996.
- Muralla, R., et al., 2008. A bifunctional locus (BIO3–BIO1) required for biotin biosynthesis in *Arabidopsis*. *Plant Physiol.* 146, 60–73.
- Niehrs, C., Pollet, N., 1999. Synexpression groups in eukaryotes. *Nature* 402, 483–487.
- Poyatos, J.F., Hurst, L.D., 2007. The determinants of gene order conservation in yeasts. *Genome Biol.* 8, R233.
- Purmann, A., et al., 2007. Genomic organization of transcriptomes in mammals: coregulation and cofunctionality. *Genomics* 89, 580–587.
- Qian, W., Zhang, J., 2008. Evolutionary dynamics of nematode operons: easy come, slow go. *Genome Res.* 18, 412–421.
- Remm, M., Storm, C.E.V., Sonnhammer, E.L.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.
- Ren, X.Y., Fiers, M.W.E.J., Stiekema, W.J., Nap, J.P., 2005. Local coexpression domains of two to four genes in the genome of *Arabidopsis*. *Plant Physiol.* 138, 923–934.
- Ren, X.Y., Stiekema, W.J., Nap, J.P., 2007. Local coexpression domains in the genome of rice show no microsynteny with *Arabidopsis* domains. *Plant Mol. Biol.* 65, 205–217.
- Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69.
- Rogozin, I.B., et al., 2002. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* 30, 2212–2223.
- Semon, M., Duret, L., 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* 23, 1715–1723.
- Shoja, V., Murali, T.M., Zhang, L., 2007. Expression divergence of tandemly arrayed genes in human and mouse. *Comp. Funct. Genom.* 2007, 60964.
- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., Paterson, A.H., 2008. Synteny and collinearity in plant genomes. *Science* 320, 486–488.
- Thimmapuram, J., Duan, H., Liu, L., Schuler, M.A., 2005. Bicistronic and fused monocistronic transcripts are derived from adjacent loci in the *Arabidopsis* genome. *RNA* 11, 128–138.
- Trinklein, N.D., Aldred, S., Hartman, S.J., Schroeder, D.I., Otilar, R.P., Myers, R.M., 2004. An abundance of bidirectional promoters in the human genome. *Genome Res.* 14, 62–66.
- Tuskan, G.A., et al., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604.
- Williams, E.J.B., Bowles, D.J., 2004. Coexpression of neighbouring genes in the genome of *Arabidopsis thaliana*. *Genome Res.* 14, 1060–1067.
- Zhan, S., Horrocks, J., Lukens, L.N., 2006. Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. *Plant J.* 45, 347–357.
- Zhang, L., Li, W., 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* 21, 236–239.
- Zhang, X., Feng, B., Zhang, Q., Zhang, D., Altman, N., Ma, H., 2005. Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. *Plant Mol. Biol.* 58, 401–419.