Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes

Jun Wang, Feng Tao, Nicholas C. Marowsky, and Chuanzhu Fan*

Department of Biological Sciences, Wayne State University, Detroit, Michigan 48202

Gene duplication is a primary means to generate genomic novelties, playing an essential role in speciation and adaptation. Particularly in plants, a high abundance of duplicate genes has been maintained for significantly long periods of evolutionary time. To address the manner in which young duplicate genes were derived primarily from small-scale gene duplication and preserved in plant genomes and to determine the underlying driving mechanisms, we generated transcriptomes to produce the expression profiles of five tissues in Arabidopsis thaliana and the closely related species Arabidopsis lyrata and Capsella rubella. Based on the quantitative analysis metrics, we investigated the evolutionary processes of young duplicate genes in Arabidopsis. We determined that conservation, neofunctionalization, and specialization are three main evolutionary processes for Arabidopsis young duplicate genes. We explicitly demonstrated the dynamic functionalization of duplicate genes along the evolutionary time scale. Upon origination, duplicates tend to maintain their ancestral functions; but as they survive longer, they might be likely to develop distinct and novel functions. The temporal evolutionary processes and functionalization of plant duplicate genes are associated with their ancestral functions, dynamic DNA methylation levels, and histone modification abundances. Furthermore, duplicate genes tend to be initially expressed in pollen and then to gain more interaction partners over time. Altogether, our study provides novel insights into the dynamic retention processes of young duplicate genes in plant genomes.

Generally, after gene duplication, if the two copies survive in populations, they undergo four evolutionary trajectories: (1) conservation, in which the two copies maintain the same function as the ancestral gene; (2) neofunctionalization, in which one copy develops a
novel function whereas the other copy retains the original function; (3) subfunctionalization, in which the two copies develop different functions from each other but work together to compensate for the entire function of the ancestral gene; and (4) specialization, in which the two copies evolve different functions from each other, and their overall function is also different from the ancestral gene, which encompasses the processes of both neofunctionalization and subfunctionalization (Ohno, 1970; Force et al., 1999; Stoltzfus, 1999; He and Zhang, 2005; Assis and Bachtrog, 2013). The four evolutionary trajectories of retained duplicate genes have been supported by both theoretical models and substantial empirical evidence (Mena et al., 1996; Force et al., 1999; Lynch and Force, 2000; Walsh, 2003; Loppin et al., 2005; Benderoth et al., 2006; Kleinjan et al., 2008; Park et al., 2008; Innan, 2009; Ding et al., 2010; Weng et al., 2012). Furthermore, a set of analysis metrics was developed recently to quantitatively distinguish the four evolutionary trajectories of gene duplication by applying a phylogenetic comparison of the transcriptomic data of closely related *Drosophila* and mammal species (Assis and Bachtrog, 2013, 2015). Using expression profiles as proxies for function, the expression distances of two duplicate genes in *Drosophila* and mammal species to their ancestral gene in outgroup species were compared with that of single-copy genes to their outgroup orthologous genes (Assis and Bachtrog, 2013, 2015). These analysis metrics provide a valuable resource with which we can study the evolutionary processes of preserving duplicate genes in other species.

Here, we systematically address two questions. How do young duplicates originate and persist in genomes? What are the underlying mechanisms influencing their different evolutionary trajectories? To answer these questions, we generated and compared the transcriptomic profiles of *A. thaliana*, *A. lyrata*, and *Capsella rubella* from high-throughput RNA sequencing (RNA-seq) in five tissues. By taking advantage of the aforementioned phylogenetic approaches and combining expression profiles with gene functions, selection constraints, duplication mechanisms, and epigenetic modifications, we unraveled how young duplicate genes were maintained in Arabidopsis genomes.

RESULTS

Identification of Lineage-Specific Young Duplicate Genes in Arabidopsis Genomes

Based on the *A. thaliana*, *A. lyrata*, *C. rubella*, and *Brassica rapa* genomes, we used a homolog search based on sequence similarity and the rate of synonymous substitutions (Ks) to search for newly duplicated genes in Arabidopsis genomes (see “Materials and Methods”). We identified 187 *A. thaliana* species-specific duplicate gene pairs that originated fewer than 5 MYA and 58 Arabidopsis genus-specific duplicate gene pairs that originated after Arabidopsis split from *C. rubella* and before the divergence of *A. lyrata* and *A. thaliana* approximately 5 to 10 MYA (see “Materials and Methods”). We identified 450 *A. lyrata* species-specific paralog pairs that originated fewer than 5 MYA and 58 Arabidopsis genus-specific paralog pairs that originated between 5 and 10 MYA. The different numbers of species-specific duplicates between *A. thaliana* and *A. lyrata* might reflect the differential originiation/retention rate of duplicate genes in these two species. Overall, the origination rates of duplicate genes in *A. thaliana* and *A. lyrata* are estimated as 37 genes and 90 genes per genome per million years, respectively, which is much faster than in animals (Long et al., 2013).

Of these duplicate gene pairs identified from the *A. thaliana* genome, 112 (59.89%) *A. thaliana* species-specific paralog pairs and 38 (65.52%) Arabidopsis genus-specific paralog pairs are expressed in at least one tested tissue of *A. thaliana* (Supplemental Table S1). Of these duplicate gene pairs identified from the *A. lyrata* genome, 347 (77.11%) *A. lyrata* species-specific paralog pairs and 40 (68.97%) Arabidopsis genus-specific paralog pairs are transcribed in at least one tested tissue of *A. lyrata* (Supplemental Table S2). Pseudogenes generally have no/low expression to result in large expression distances from the ancestral states, but they may not reveal the retention processes of functional duplicate genes (Zou et al., 2009a; Sisu et al., 2014). To avoid the confounding effect from pseudogenes, therefore, we focused on the expressed paralog pairs in the following analyses. Additionally, we identified 3,097 and 3,363 single-copy genes in *A. thaliana* and *A. lyrata*, respectively, which are expressed in at least one tissue of Arabidopsis and their orthologs in *C. rubella*.

Classifying the Evolutionary Trajectories of Duplicate Genes Using the Euclidean Distance of Expression

By applying the phylogenetic comparison and previously developed quantitative analysis metrics (Assis and Bachtrog, 2013, 2015), we investigated the evolutionary processes of the identified expressed paralogs in *A. thaliana* and *A. lyrata*. We first computed the Euclidean distance of relative expression profiles between one duplicate copy (D1) and the *C. rubella* ancestral copy (A) as $E_{D1,A}$ between the other duplicate gene (D2) and the ancestral copy as $E_{D2,A}$, and between the combined two duplicate genes and the ancestral copy as $E_{D1+D2,A}$ for each of duplicate gene pairs in *A. thaliana* and *A. lyrata*. To set up the level of baseline expression divergence for gene pairs, we also calculated the Euclidean expression distance between a single-copy gene (S1) and the *C. rubella* ortholog copy (S2) as $E_{S1,S2}$ for *A. thaliana* and *A. lyrata*.

Next, we compared $E_{D1,A}$, $E_{D2,A}$, and $E_{D1+D2,A}$ with $E_{S1,S2}$ for each of the paralog pairs to infer their evolutionary trajectories individually. Presumably, $E_{S1,S2}$ represents the expected Euclidean expression distance between the genes in sister species; thus, we can compare $E_{D1,A}$, $E_{D2,A}$, and $E_{D1+D2,A}$ with $E_{S1,S2}$ to define a set of rules to classify cases into conservation, neofunctionalization,
subfunctionalization, and specialization. In detail, for conservation, the expression profiles of both duplicate genes should be similar to the ancestral copy. Thus, the expectation is $E_{D1,A} \leq E_{S1,S2}$ and $E_{D2,A} \leq E_{S1,S2}$. For neofunctionalization, the expression profile of one duplicate copy should be different from the ancestral copy and the other paralogous copy should be similar to the ancestral copy. Thus, the expectation is $E_{D1,A} > E_{S1,S2}$ and $E_{D2,A} \leq E_{S1,S2}$, or $E_{D1,A} > E_{S1,S2}$ and $E_{D2,A} \leq E_{S1,S2}$. For subfunctionalization, both duplicate genes should have different expression profiles from the ancestral copy, but the combined duplicate gene expression should be similar to the expression of the ancestral copy. Hence, the expectation is $E_{D1,A} > E_{S1,S2}$, $E_{D2,A} > E_{S1,S2}$, and $E_{D1+D2,A} \leq E_{S1,S2}$. For specialization, the expression profiles of the two duplicate genes and combined duplicate gene expression should all differ from the expression of the ancestral copy. Therefore, the expectation is $E_{D1,A} > E_{S1,S2}$, $E_{D2,A} > E_{S1,S2}$, and $E_{D1+D2,A} > E_{S1,S2}$ (Table I).

We chose the median absolute deviation from the median as the cutoff of $E_{S1,S2}$ because this cutoff was more resilient to extreme values in the distribution and could robustly represent the skewed distribution. To validate the robustness of our analysis, we also tested a number of alternative cutoffs, and all of them generated similar results (Supplemental Tables S3 and S4). Overall, 192 conservation, 104 neofunctionalization, and 158 specialization cases were classified for the species-specific paralogs, and 19 conservation, 23 neofunctionalization, and 36 specialization cases were classified for the genus-specific paralogs (Table I). These results also suggest that conservation, neofunctionalization, and specialization are the three main evolutionary processes of young duplicate genes in Arabidopsis. In addition, the relative proportion of conservation cases decreased (from 42.3% to 24.4%; Fisher test, $P = 0.002121$) and that of specialization cases increased (from 34.8% to 46.2%; Fisher test, $P = 0.000987$; for $A. lyrata$, $P = 0.1233875$, $P = 0.03214$) over the evolutionary time scale (from 5 to 10 million years).

We observed that sums of Euclidean expression distances of duplicate genes were positively correlated with Ks values (Spearman correlation; for $A. thaliana$, $\rho = 0.2663235$, $P = 0.009987$; for $A. lyrata$, $\rho = 0.1233875$, $P = 0.01515$), suggesting that the expression of duplicate genes diverged over time. This is consistent with the above pattern of the changing proportions of conservation and specialization cases. Even if we categorized the paralogs according to species, conservation, neofunctionalization, and specialization were still the three primary evolutionary trajectories of young duplicate genes, and changes in their proportions over evolutionary time still occurred. We have observed that cases of subfunctionalization were extremely rare (five and zero in $A. thaliana$ and $A. lyrata$, respectively), so we excluded this category for the later analyses.

### Classification of Evolutionary Trajectories Using Expression Localization Patterns

To further verify the classification of paralog evolutionary trajectories, we performed an alternative analysis based on expression localization patterns (Assis and Bachtrog, 2013). A binary index, $D$, was used to measure the expression localization change. Presumably, if two genes are expressed in the same tissues, $D = 0$; and if they are expressed in different tissues, $D = 1$. We calculated $D$ between one duplicate/another duplicate/combined two duplicates and ancestral genes as $D_{D1,A}$, $D_{D2,A}$, and $D_{D1+D2,A}$ respectively. For conservation, both duplicate genes are expressed in the same tissues as the ancestral gene, so the expectation is $D_{D1,A} = 0$ and $D_{D2,A} = 0$. For neofunctionalization, the copy with neofunction is expressed in the different tissues from the ancestral gene, but the other paralogous copy is expressed in the same tissue as the ancestral gene. Thus, the expectation is $D_{D1,A} = 0$ and $D_{D2,A} = 1$ or $D_{D1,A} = 1$ and $D_{D2,A} = 0$. For subfunctionalization, both duplicates are expressed in different tissues between themselves, but these tissues overlap with the tissues of the ancestral gene. Thus, the expectation is $D_{D1,A} = 1$, $D_{D2,A} = 1$, and $D_{D1+D2,A} = 0$. Finally, for specialization, the expression of two duplicate genes and combined duplicate copies is in the different tissues from the ancestral gene; thus, the expectation is $D_{D1,A} = 1$, $D_{D2,A} = 1$, and $D_{D1+D2,A} = 1$.

By applying the above rules, we classified the evolutionary processes of paralogs with the binary index $D$ (Table II). For species-specific paralogs, we identified 198 conservation cases, 91 neofunctionalization cases, 170 specialization cases, and zero subfunctionalization case. For genus-specific paralogs, 29 conservation cases, 13 neofunctionalization cases, 35 specialization cases, and one subfunctionalization case were categorized. The overall results based on expression localization

### Table I. Classification of evolutionary processes of Arabidopsis duplicate genes by measuring the Euclidean expression distance

<table>
<thead>
<tr>
<th>Classification</th>
<th>$E_{D1,A}$</th>
<th>$E_{D2,A}$</th>
<th>$E_{D1+D2,A}$</th>
<th>Species Specific</th>
<th>Genus Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation</td>
<td>$\leq E_{S1,S2}$</td>
<td>$\leq E_{S1,S2}$</td>
<td>$-$</td>
<td>AT</td>
<td>47</td>
</tr>
<tr>
<td>Neofunctionalization</td>
<td>$&gt; E_{S1,S2}$</td>
<td>$\leq E_{S1,S2}$</td>
<td>$-$</td>
<td>30</td>
<td>74</td>
</tr>
<tr>
<td>Subfunctionalization</td>
<td>$&gt; E_{S1,S2}$</td>
<td>$&gt; E_{S1,S2}$</td>
<td>$\leq E_{S1,S2}$</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Specialization</td>
<td>$&gt; E_{S1,S2}$</td>
<td>$&gt; E_{S1,S2}$</td>
<td>$&gt; E_{S1,S2}$</td>
<td>34</td>
<td>124</td>
</tr>
<tr>
<td>Total</td>
<td>$112$</td>
<td>$347$</td>
<td>$459$</td>
<td>38</td>
<td>40</td>
</tr>
</tbody>
</table>
patterns are consistent with the aforementioned results generated by Euclidean distance. This consistency suggests the robustness of our classification results and supports our conclusion that neofunctionalization, conservation, and specialization are the three major evolutionary processes and the dynamic process of conservation and specialization in preserving young duplicates in Arabidopsis over evolutionary time.

Expression Data from TAIR Supporting the Classification of the Evolutionary Trajectories of Paralogs

To search the independent evidence for paralog classification based on RNA-seq profiling, we retrieved the plant ontology annotations (e.g. guard cells, pollen, etc.) from TAIR for each of the A. thaliana duplicate genes. Because the expression of the ancestral genes is not available at TAIR, we only classified the duplicates into conserved and divergent (including subfunctionalization, neofunctionalization, and specialization) cases. To define conserved or divergent cases based on the TAIR plant ontology annotations, we first counted the total number of structures in which duplicate genes are expressed (defined as A, including the expressed structures of both genes or either copy) and then counted the number of structures in which the two gene copies are differentially expressed (defined as B, namely, the structures where one gene is expressed but the other gene is not). We defined the two duplicate genes as divergently expressed if the ratio of B/A is larger than 10%; otherwise, they were defined as conserved. We also validated this classification using different B/A ratios as cutoffs (from 10% to 50%), which yielded similar results. Overall, based on TAIR expression structure profiling, 52 and 98 paralog pairs were classified as conserved and divergent cases, respectively. This is consistent with the aforementioned analysis based on RNA-seq profiling, in which 57 and 93 paralog pairs were categorized as conserved and divergent cases.

Functional Measurement Supporting the Classification of the Evolutionary Fates of Paralogs

To further confirm the biological relevance of our classification results and assess the biological function of the classified paralogs, we analyzed three alternative metrics: tissue specificities, relative expression level spectrum across tissues, and selection constraints of duplicate genes. For the conservation class, duplicate genes have similar tissue specificities and similar relative expression spectra across the five tissues to their ancestral genes (Fig. 1, A and B). The two copies of the duplicate pairs also have similarly low Ka/Ks (Fig. 1C). For the neofunctionalization class, the neofunctionalized duplicate genes have significantly higher tissue specificities than their ancestral genes, whereas the tissue specificities of their paralogs are similar to their ancestral genes (Fig. 1, A and B). The neofunctionalized duplicate genes also have significantly higher Ka/Ks than the other paralogous copies (Fig. 1C; Wilcoxon rank-sum test, $P = 0.02233$). For the specialization class, both duplicate genes have significantly higher tissue specificities than their ancestral genes, differential expression spectra between themselves and also from the ancestral genes (Fig. 1, A and B), and similarly high Ka/Ks, which are significantly higher than the Ka/Ks of the conservation class (Fig. 1C; Wilcoxon rank-sum test, $P = 0.03868$).

Additionally, we found that 145 duplicate genes out of 150 paralog pairs in A. thaliana, and 189 duplicate genes out of 387 paralog pairs in A. lyrata, have Ka/Ks significantly smaller than 1 using the likelihood ratio test (LRT; false discovery rate [FDR] $q < 0.05$; see “Materials and Methods”), suggesting that many duplicate genes are under functional constraints. The Ka/Ks of one duplicate gene in A. thaliana and seven duplicate genes in A. lyrata are significantly larger than 1 (LRT; FDR $q < 0.05$), suggesting that they might be driven by positive selection (Supplemental Tables S5 and S6).

Test of the Out-of-Pollen Hypothesis for Arabidopsis Young Duplicate Genes

Previous studies showed that new Arabidopsis genes tend to express in mature pollen (Liu et al., 2011; Wu et al., 2014; Cui et al., 2015). Based on this observation, we asked whether our identified new duplicate genes also have a biased tissue expression pattern. Using both available RNA-seq and microarray data in A. thaliana, we showed that the younger duplicate genes (i.e. species-specific duplicate genes) have higher expression levels in mature pollen compared with most of the other tissues. Moreover, their expression level inferred from RNA-seq also is higher than that of single-copy genes and older
Figure 1. Comparison of functional metrics for paralogs classified in three evolutionary processes. A, Scatterplot of tissue specificities (tao) between duplicates and ancestral genes. B, Heat maps show the mean relative expression levels in five tissues of single-copy (S), duplicate (D1 and D2), and ancestral (A) genes, with darker colors depicting higher values. C, Distributions of Ka/Ks for duplicates (D1 and D2). Using the Wilcoxon rank-sum test, we compared the distributions of Ka/Ks between two duplicate copies to estimate significance: *, P < 0.05.
Figure 2. Out-of-pollen pattern of *A. thaliana* young duplicate genes. A, Heat map shows the mean expression levels measured with RNA-seq in six tissues (pollen, silique, root, stem, floral bud, and leaf) of single-copy (S), younger duplicate (Y), and older duplicate (O) genes, with darker colors depicting higher values. B, Heat map shows the mean expression levels measured using microarray data in 47 tissues/developmental stages, with bright red depicting lower values and bright green depicting higher values. The expression levels in pollen are indicated with the red box. C, Relative expression levels of duplicate genes at the four pollen developmental stages. We computed relative expression levels using the mean expression level of duplicate genes divided by the mean expression level of...
duplicate genes (i.e. genus-specific duplicate genes) in mature pollen (Fig. 2, A and B; Supplemental Table S7). Based on Monte Carlo simulations, we concluded that the younger duplicate genes have significantly higher average pollen expression levels than the background ($P = 0.009170$), but the older duplicate genes do not ($P = 0.208645$). Among the four pollen developmental stages, the duplicate genes have higher expression levels in tricellular pollen and mature pollen than in the uninucleate microspores and bicellular pollen (Fig. 2C; Supplemental Table S7). Additionally, by comparing the numbers of protein-protein interactions, we found that younger duplicate genes have significantly fewer protein-protein interactions than the older duplicate genes (Wilcoxon rank-sum test, $P = 0.0008756$). The numbers of protein-protein interactions of duplicate genes are significantly positively correlated to their Ks (Spearman correlation; $\rho = 0.181855$, $P = 5.142e-05$). These results suggest that, after new genes originated from pollen, they might have acquired more interaction partners over evolutionary time (Fig. 2D; Supplemental Table S7).

The Underlying Evolutionary Mechanisms Contributing to the Temporal Evolutionary Processes of Duplicate Genes

First, we investigated the duplication modes of young duplicate genes in *A. thaliana*, which might contribute to their divergence. Because the *A. lyrata* genome only had scaffold assembly but not chromosome assembly, this analysis focused only on the *A. thaliana* genome. Overall, the great majority of duplicate genes were derived from small-scale gene duplication, including 91 tandem duplications and 151 dispersed duplications. We only identified three paralog pairs from segmental duplication. The proportion of dispersed duplication temporally descended over evolutionary time, from 66.31% in younger duplicate genes to 46.55% in older duplicate genes, but the trend was opposite for tandem duplication, from 32.62% in younger duplicate genes to 51.72% in older duplicate genes. This suggested that dispersed duplicates might be less favorable to be retained than tandem duplicates over time (Fisher test, $P = 0.006224$).

Previous studies in *Drosophila* spp. showed that the functional diversities of ancestral genes might impact the evolutionary processes of young duplicate genes (Assis and Bachtrog, 2013). Specifically, the conserved paralogs have the narrow ancestral function, specialized paralogs have the moderate ancestral function, and neofunctionalized duplicates have the broad ancestral function (Assis and Bachtrog, 2013). However, we observed a completely different pattern in plant genomes, as shown in our Arabidopsis genome analyses.

Namely, specialized plant paralogs have ancestral genes with high tissue specificities (Wilcoxon rank-sum test, $P < 0.001$; Fig. 3) and narrow expression spectrum across tissues (Fig. 1B); neofunctionalized plant paralogs have ancestral genes with intermediate tissue specificities (Wilcoxon rank-sum test, $P < 0.05$; Fig. 3) and expression spectrum across tissues (Fig. 1B); and conserved plant paralogs have ancestral genes with lower tissue specificities (Wilcoxon rank-sum test, $P < 0.05$; Fig. 3) and broader expression spectrum across tissues (Fig. 1B).

Consistently, Gene Ontology (GO) term enrichment analysis of duplicates in the three classes indicates that the enriched GO terms of the conservation class are associated with highly connected functions such as macromolecular complex (FDR = 9.52E-07), protein complex (FDR = 1.45E-04), and membrane part (FDR = 7.03E-4; Supplemental Table S8). The neofunctionalization and specialization classes have enriched GO terms related to enzymes participating in either flexible steps or the tips of pathways, including small molecule biosynthetic enzymes participating in plant secondary metabolism, such as amidine-lyase activity (FDR = 0.0182), N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity (FDR = 0.0182), ferric-chelate reductase activity (FDR = 2.42E-05), oxidoreductase activity, and oxidizing metal ion activity (FDR = 2.42E-05; Supplemental Tables S9 and S10).

Additionally, we investigated how DNA methylation level and histone modification abundance of duplicate...
genes change over time and whether they would impact the evolutionary processes of duplicate genes. For histone modifications, we explicitly interrogated five types of histone marks associated with the activation of genes: H3K9Ac, H3K4me2, H3K4me3, H3K36me2, and H3K36me3, in the aerial tissue of A. thaliana (Luo et al., 2013). We computed the Spearman’s correlation coefficients among DNA methylation level in the promoter regions (including the upstream 1,000-bp and downstream 1,000-bp flanking regions of the transcription start site [TSS]), histone modification abundance in the promoter regions, and expression level in the leaf for 27,081 genes with data available in the genome. As expected, the abundances of active histone modifications are positively correlated with each other and the gene expression level, whereas the DNA methylation level is negatively correlated with the active histone modification abundances and the gene expression level (Fig. 4; Spearman’s correlation; for correlation coefficients, see Supplemental Table S11).

Interestingly, we found that the methylation levels in the promoter regions of younger duplicate genes are significantly higher compared with those of the older duplicate genes (Fig. 5; Wilcoxon rank-sum test, \( P < 0.05 \)). In contrast, the abundances of active histone modifications in the promoter regions of younger duplicate genes are significantly lower than those of the older duplicate genes (Fig. 5; Wilcoxon rank-sum test, \( P < 0.05 \)). Furthermore, younger duplicate genes have lower expression levels compared with older duplicate genes (Fig. 5; Wilcoxon rank-sum test, \( P < 0.05 \)). Consistently, we found that DNA methylation levels in the promoter regions of duplicate genes are significantly negatively correlated with their Ks values (Spearman’s correlation, \( P < 0.01 \); Table III), whereas the active histone modification abundances in the promoter regions of duplicate genes are significantly positively correlated with their Ks values (Spearman’s correlation, \( P < 0.01 \); Table III). Overall, these results suggest that duplicate genes at the early origination stage have higher levels of DNA methylation and lower abundances of active histone modifications in the promoter regions. However, as the duplicate genes become old and are preserved in the genome, they could acquire higher abundances of active histone modifications and their DNA methylation levels decrease; thus, they are very likely to obtain higher expression levels.

**DISCUSSION**

Gene duplication plays a central role in plant diversification as a key process that forms the raw material necessary for adaptive evolution. Our analyses focused on the evolutionary processes of young duplicate genes that originated fewer than 5 to 10 MYA. This identification was based on the stringent homology and Ks cutoffs, which distinguish our analyses from previous studies focusing on all duplicates or old duplicates from whole-genome duplication. Indeed, most of the identified young duplicates originated from tandem and dispersed duplications, which is consistent with the observation that the most recent whole-genome duplication in the Arabidopsis lineage occurred before the split of the Arabidopsis and Brassica genera 20 to 43 MYA (Blanc et al., 2003; Beilstein et al., 2010).

Among the four distinct evolutionary processes to maintain the paralogs after gene duplication, our analysis explicitly demonstrates that conservation, neofunctionalization, and specialization are the three primary evolutionary trajectories of young duplicate genes in Arabidopsis. Interestingly, the proportion of
the conservation cases decreased and that of the specialization cases increased over time. The traditional models of duplicate gene divergence (e.g., duplication-degeneration-complementation, subfunctionalization, neofunctionalization, and lost function) only explained the divergence of duplicates over the long evolutionary time period but not about the situation immediately after duplication. In contrast, our study captured the initial and dynamic patterns of evolutionary trajectories of duplicate genes. Our observations might be explained by two alternative but not exclusive hypotheses. First, after gene duplication, the two gene copies initially maintained similar ancestral functions. After a longer evolutionary time, however, both copies underwent genetic and epigenetic divergence and finally developed distinct novel functions. Therefore, some paralogs that were initially categorized as conserved would become specialized or neofunctionalized later. Since the conserved duplicates more likely tend to be highly connected genes than the specialized or neofunctionalized duplicates (as shown above), the second explanation is that the duplicates of former genes might have more difficulties surviving and thus become pseudogenized later, leading to a decreasing proportion of conservation cases and a relatively increasing proportion of specialization.

The conservation cases decreased and that of the specialization cases increased over time. The traditional models of duplicate gene divergence (e.g., duplication-degeneration-complementation, subfunctionalization, neofunctionalization, and lost function) only explained the divergence of duplicates over the long evolutionary time period but not about the situation immediately after duplication. In contrast, our study captured the initial and dynamic patterns of evolutionary trajectories of duplicate genes. Our observations might be explained by two alternative but not exclusive hypotheses. First, after gene duplication, the two gene copies initially maintained similar ancestral functions. After a longer evolutionary time, however, both copies underwent genetic and epigenetic divergence and finally developed distinct novel functions. Therefore, some paralogs that were initially categorized as conserved would become specialized or neofunctionalized later. Since the conserved duplicates more likely tend to be highly connected genes than the specialized or neofunctionalized duplicates (as shown above), the second explanation is that the duplicates of former genes might have more difficulties surviving and thus become pseudogenized later, leading to a decreasing proportion of conservation cases and a relatively increasing proportion of specialization.

**Figure 5.** Dynamic methylation levels, histone modification abundances, and gene expression levels of duplicate genes. This figure shows the comparison of the abundances of five active histone modifications, DNA methylation levels, and gene expression levels of the younger duplicates (Y) and the older duplicates (O) using the Wilcoxon rank-sum test. The abundances of the five active histone modifications of Y are significantly lower than those of O. The DNA methylation levels of Y are significantly higher than those of O. The gene expression levels of Y are significantly lower than those of O. *, P < 0.05; **, P < 0.01; and ***, P < 0.001.

### Table III. Correlations between the abundances of active histone modifications/DNA methylation and Ks of duplicate genes

***, P < 0.001; and ***, P < 0.0001.

<table>
<thead>
<tr>
<th>Epigenetic Modification</th>
<th>Spearman’s Correlation Coefficient</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K9Ac</td>
<td>0.1924446</td>
<td>3.52E-05***</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>0.2108966</td>
<td>5.563e-06***</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.1471653</td>
<td>0.001626**</td>
</tr>
<tr>
<td>H3K36me2</td>
<td>0.1759113</td>
<td>0.0001596***</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>0.135351</td>
<td>0.003783**</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>−0.1314893</td>
<td>0.004918**</td>
</tr>
</tbody>
</table>
Such a plant-specific evolutionary pattern for duplicates is different from that in Drosophila spp., where the neofunctionalization of young duplicate genes plays a major role in preserving the duplicates, but similar to that in mammals (Assis and Bachtrog, 2013, 2015). This different pattern can be explained in that plant genomes are more accommodating for genomic and functional redundancy and diversity to adapt to various environments. Moreover, we found rare cases of subfunctionalization in Arabidopsis, which is the same pattern as in Drosophila spp. and mammals (Assis and Bachtrog, 2013, 2015). This is also supported by previous studies in plant genomes in which subfunctionalization only plays a minor role, whereas neofunctionalization and specialization, which correspond to the reciprocal neofunctionalization expression in some studies, are more important in preserving duplicates (Ganko et al., 2007; Liu et al., 2011; Hughes et al., 2014). Alternatively, the rare cases of subfunctionalization of plant duplicate genes may be due to the stringent criteria of our methodology for identification, which cause some subfunctionalization cases to be classified into conservation and specialization categories. One possible solution might be to conduct the same analysis using expression data from more tissues (Assis and Bachtrog, 2015).

In plants, the conserved, neofunctionalized, and specialized duplicates display a sequential manner of expression pattern, ranging from low tissue specificities/broad expression spectra in conservation to high tissue specificities/narrow expression spectra in specialization. Furthermore, the functions of neofunctionalized and specialized duplicates are biased toward small-molecule biosynthetic enzymes. This phenomenon may be accounted for by the following explanations. First, the change of the functional roles/interaction partners of broadly expressed duplicate genes might be detrimental to the organisms under purifying selection. Alternatively, an increasing dose of their ancestral function might be beneficial to species adaptation, for example, by buffering the ancestral genes (conservation). Lastly, the duplicate genes with high tissue specificities that have products at either flexible steps or at the tips of pathways are favorable to evolve novel functions (neofunctionalization) or to develop specialized functions that are distinct from ancestral copies (specialization).

Interestingly, young Arabidopsis duplicate genes are predominantly expressed in pollen, especially tricellular and mature pollen. We demonstrate that the initial out-of-pollen duplicates can gain more protein-protein interactions over time. The out-of-pollen pattern has been found in previous studies (Liu et al., 2011; Wu et al., 2014) and corresponds to the out-of-testis phenomenon of new genes in Drosophila spp. and mammals (Betrán et al., 2002; Emerson et al., 2004). Two underlying reasons might account for this. First, these tissues may provide a permissive environment for gene expression, such as the relaxed epigenetic context in plant pollen, where the tissue undergoes developmental chromatin remodeling (Calarco et al., 2012; Soumillon et al., 2013; Wu et al., 2014). Second, these tissues may experience fast evolutionary rates due to their haploid nature and the strong selection forces they undergo and are associated with sexual selection and sperm/pollen competition. Thus, these tissues may appear as evolutionarily accommodating, welcoming and tolerating genomic and functional innovations such as new genes (Bernasconi et al., 2004; Kaesmann, 2010; Arunkumar et al., 2013). More importantly, both methylation levels and the abundances of active histone modifications in promoter regions are significantly and temporally associated with the ages of duplicate genes. These results suggest that, after the origin of new duplicate genes, their expression is initially repressed by DNA methylation and deficiency of active histone modifications, but they will lose DNA methylation and gain active histone modifications over evolutionary time. This is further support that the demethylation in tricellular and mature pollen can provide a unique outlet for newly originated duplicate genes to be expressed. After the new genes are expressed in pollen, they might have chances to become established in the genome by gaining specific/broad expression spectra and more interaction partners.

MATERIALS AND METHODS

Plant Species and Genomic Sequence Data Sets

We chose four closely related species, Arabidopsis thaliana, Arabidopsis lyrata, Capsella rubella, and Brassica rapa, for comparative genomic analysis. The complete genomes framework data sets including assemblies and annotations were downloaded from Physotzome version 8.0 (http://www.phytozome.net/) for A. thaliana 167 (TAIR release 10), A. lyrata 107 (IGI release version 1.0), C. rubella 183 (IGI annotation version 1.0 on assembly version 1), and B. rapa 197 (annotation version 1.2 on assembly version 1.1 from brassicadb.org). We generated RNA-seq data of five tissues in A. thaliana, A. lyrata, and C. rubella (see below). The RNA-seq data of pollen were obtained from the Short Read Archive in NCBI with accession number SRP022162 (Loraine et al., 2013). The DNA methylation bisulfite sequencing data of A. thaliana and A. lyrata (from control roots) were downloaded from the European Nucleotide Archive with accession number PRJEB6701 (Seymour et al., 2014). The microarray data of 79 A. thaliana tissues/developmental stages were downloaded from EMBL-EBI (European Bioinformatics Institute) under ArrayExpress with accession number E-APMX-9 (Schmid et al., 2005). We removed 32 tissues/developmental stages with redundancy or mutants and kept 47 tissues in our analyses. The microarray data of four primary pollen developmental stages in A. thaliana were obtained from the Gene Expression Omnibus of NCBI with series number GSE34190 (Horvys and Twell, 2004). The chromatin immunoprecipitation (ChIP) sequencing data of histone modifications from the leaf were downloaded from the Gene Expression Omnibus of NCBI with series number GSE28398 (Luo et al., 2013).

RNA-Seq

Plants of the reference lines A. thaliana (Columbia-0), A. lyrata (MN47), and C. rubella (Monte Gargano) were grown in the greenhouse under normal long-day growth conditions as 16-h (22°C) day / 8-h (18°C) night. Total RNAs of leaves, stems, siliques, floral buds, and roots were isolated using TRIzol reagent (Life Technologies). Leaves were collected from rosettes. Stems were cut from mature plants around 6 to 8 weeks after germination. Floral buds were collected when plants started bolting. Siliques were collected when young and fleshy siliques formed. Roots of mature plants (approximately 8 weeks) were dug out from soil and then thoroughly washed to avoid contamination. Approximately 100 mg of plant tissues was ground and treated by TRIzol, and approximately 200 to 2,000 ng µL⁻¹ total RNA was generated. RNA-seq libraries were made.
using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina according to the manufacturer’s protocol. Specifically, mRNA was isolated using the NEBNext Oligo d(T)~18 and then fragmented at 94°C for 15 min. First and second cDNAs were simultaneously synthesized in success. After end repairing and dA tailing of the cDNA library, the multiplex adaptors (index primers set 1 and set 2) were ligated to the two ends of the cDNAs. Finally, each library was enriched by 14 cycles of amplification. The quality of cDNA libraries with the multiplex adaptors was checked using Bioanalyzer (Agilent), and the quantity of cDNA was measured using Qubit (Invitrogen). High-quality libraries were then pooled based on the balance of the concentration of each library. Fifteen libraries were mixed into one pool, and each library had a different oligonucleotide sequence locating the downstream side of the adaptor. The loading volume of the pooled libraries was decided using real-time PCR. All libraries were sequenced in one lane of the Illumina HiSeq run and then separated by the corresponding tag sequences.

Identification of Young Duplicate Genes in Arabidopsis

We conducted a BLAST search with all peptides against all peptides for A. thaliana and A. lyrata. Then, we picked the BLAST matches with identity scores of 40 or greater and coverage of 60% or greater of the peptide length, which are reciprocal best hits of each other. Ks values were calculated for all identified gene pairs using the pipeline developed previously (Zhang et al., 2013). We picked gene pairs as Arabidopsis species-specific duplicate genes if Ks \( \leq 0.135284 \). We defined gene pairs as Arabidopsis genus-specific duplicate genes if Ks \( \leq 0.135284 \) or Ks \( \leq 0.237983 \) for genomes of A. thaliana and A. lyrata, respectively. Ks values of 0.135284, 0.254149, and 0.237983 were calculated from the mean Ks values of single-copy orthologs between A. thaliana and A. lyrata, between A. thaliana and C. rubella, and between A. lyrata and C. rubella to approximate in fer the evolutionary divergence time of duplication.

Then, we used BLASTP to align the peptide sequence of each duplicate gene against the peptides of C. rubella (A. thaliana versus C. rubella and A. lyrata versus C. rubella). We only kept the duplicate gene pairs if the two copies of a duplicate gene pair had the same best-hit gene in the C. rubella genome (BLASTP, E \( \leq 10^{-7} \)). We required the best-hit C. rubella gene as the ortholog gene of one copy of the duplicate gene pair. The C. rubella orthologs of the duplicate genes could reveal their ancestral states. The detailed procedure to determine the ortholog of a gene in the outgroup species is described below. We further selected the gene pairs if both copies were expressed in at least one of the tested tissues regardless of two copies expressed in the same or different tissues. Based on the quantile-normalized fragments per kilobase of exon per million fragments (FPKM) of the RNA-seq data in the intergenic and exonic regions, we took the cutoff of exonic FPKM \( \geq 0.537131 \) for A. thaliana, FPKM \( \geq 0.237983 \) for A. lyrata, and FPKM \( \geq 0.237983 \) for C. rubella.

Defining the Ortholog Relationship

We determined the ortholog relationship for genes of interest among A. thaliana, A. lyrata, and C. rubella (Wang et al., 2013). Briefly, we used the developed pipelines from the UCSC genome browser to construct the whole-genome syntetic relationship between each of A. thaliana / A. lyrata / C. rubella / B. rapa with the remaining three species. We also used BLASTP to search for the reciprocal best hits of peptide sequences among these species (E \( \leq 10^{-10} \)). If a gene is in the syntenic region of its own genome and the outgroup genome, and it has a reciprocal best-hit peptide sequence in the outgroup species, we decided that this gene has the ortholog in the outgroup species and that the best reciprocal hit gene in the outgroup species is its ortholog counterpart.

Processing RNA-seq, Microarray, BS-Seq, and Histone Modification Data

The RNA-seq data were mapped to the corresponding genomes using Tophat (tophat-2.0.10; Kim et al., 2013) with default parameter settings. The expression intensities were measured in FPKM with Cufflinks (cufflinks-1.3.0, with the options–p 10 –b –u –C –Trapnell et al., 2010). Then, the FPKM values of all tested tissues in one species were quantile normalized with the normalized quantiles function in R (Assis et al., 2012). The Affymetrix microarray data were annotated to A. thaliana genes with a customized CDF file downloaded from http://brainarray.mbrl.med.umich.edu/Brainarray/Database/CustomCDF/

17.1.0/tairg.asp. The data were then processed using the RMA and MASS package of R.

The BS-seq data were processed according to the same procedure as described previously (Wang et al., 2014). Briefly, we used trim_galore (version 0.3.7, with the options–length 40 –paired –trim1 –o –retain_unpaired) to remove adapters and low-quality reads and used Bismark (version 0.13.0, with the options–non_directional –prefix –seedlen 40 –seedsms 2 –maqger 100 –chankmbs 1024 –minbins 100 –maxbins 1000–ambiguqous) to map, process the reads, and generate the methylation report. We then removed the differentially methylated sites between two sample replicates using the Fisher test and conducted binomial tests to define methylated sites, following the procedures described previously (Seymour et al., 2014). The methylation level of each region was estimated as the percentage of methylated Cs over the total mapped Cs in that region for the three C contexts (CG, CHG, and CHH). In our analysis, we only considered methylation in the CG context and the regions in which at least 50% of the Cs were mapped.

We used Bowtie (version 1.1.2) to align the color-space ChiP sequencing reads to the A. thaliana genome, allowing three or fewer mismatches (with the options–v 3 –a –m 1 –best –strata –C –p 10 –S; Langmead et al., 2009). We only considered the reads that are uniquely mapped to the genome. We used HTSeq (version 0.6.1p1), with the options–s sam–r name–m no–a exon–i gene_id –m union to count the number of reads in the upstream 1,000-bp and downstream 1,000-bp flanking regions of the TSS (Anders et al., 2015). The relative histone modification abundance was defined as \( n(\text{Histone}) \times N(\text{Input})/(N(\text{Histone}) \times N(\text{Input})) \), where \( n \) is the sum of the reads in the TSS-surrounding regions and \( N \) is the number of all the mapped reads. Histone and Input represent the specific histone modification and ChiP input, respectively (Yang et al., 2016).

Identification of Evolutionary Processes Using Expression Data

For Euclidean expression distance comparison, we first computed the Euclidean expression distance between duplicate genes and their ancestral genes and between single-copy orthologous genes. We used the median absolute deviation (with constant = 0.6 for A. thaliana and A. lyrata) from the median as the cutoff for E\( _{\text{D1,SL}} \) because it was less affected by the extreme values. We also tried various other cutoffs (e.g., mean, median, \( \text{sd} \), median absolute deviation from the median with different constants, semi-interquartile from the median, and different quantiles) values. All the results drew the same conclusion (Supplemental Tables S3 and S4).

Smaller cutoffs lead to fewer conservation cases and more specialization cases, whereas larger cutoffs lead to more conservation cases and fewer specialization cases. The expression localization patterns and other measurements are most consistent with the classification with the median absolute deviation from the median as the cutoff. Therefore, the median absolute deviation from the median is a robust cutoff to measure a skewed distribution, and we chose this cutoff to present the remainder of our results. We then compared E\( _{\text{D1,LA}} \), E\( _{\text{D1,SL}} \), and E\( _{\text{D1,D2,LA}} \) with the cutoff of E\( _{\text{D1,SL}} \). According to the rules shown in “Results” and Table I, we classified the evolutionary processes of duplicates using the Euclidean expression distance. For expression localization pattern comparison, we first determined whether a gene was expressed in certain tissues with the criteria that we mentioned in "Identification of Young Duplicate Genes in Arabidopsis." Then, depending on whether the duplicate genes were expressed in the same set of tissues as the ancestral genes, we set expression localization comparison indexes D\( _{\text{D1,LA}} \), D\( _{\text{D1,D2,LA}} \), and D\( _{\text{D1,D2,LA}} \) = 0 (the same tissues) or 1 (different tissues). Based on the D values and the rules shown in “Results” and Table II, we classified the evolutionary processes with expression localization patterns.

Analyses of Functional Metrics of Paralogs

The protein-protein interaction data of A. thaliana were collected from ANAP (http://gmds.shgmo.org/Computational-Biology/ANAP/ANAP_V1.1/; Wang et al., 2012), AtPIN (http://bioinfo.esalq.usp.br/atpin/atpin.pl; Brandão et al., 2009), Arabidopsis Interactions Viewer (http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi; Popescu et al., 2007, 2009; Braun et al., 2011), Arabidopsis Predicted Interactome (https://www.arabidopsis.org/portals/proteome/proteinInteract.jsp), and Plant Interactome Database (http://interactome.dfci.harvard.edu/A_thaliana/; Braun et al., 2011; Mukhtar et al., 2011). All of the protein-protein interaction data were pooled together to generate the nonredundant protein partners for each gene. Because A. lyrata does not have such data available, it was
excluded from this analysis. For the comparison of tissue specificity, a tissue specificity index (\(\tau\)) was computed for each gene based on the RNA-seq data in the five tissues according to the modified approach by Yanai et al. (2005) and Assis et al. (2012). \(\tau\) is in the range from 0 to 1. The higher the \(\tau\) value, the more tissue specific the gene expression. The mean relative expression levels in heat maps were calculated in each of the tissues for single-copy genes, duplicate genes, and ancestral genes. To allow the comparison between duplicates and their ancestral genes, which are in the outgroup species \(C.\ rubella\), the mean expression level of \(C.\ rubella\) ancestral genes in each tissue was normalized with the mean expression level of single-copy genes in the corresponding tissue for \(A.\ thaliana\) and \(A.\ lyrata\). The GO enrichment analyses were conducted using http://structuralbiology.cau.edu.cn/PlantGSEA (Yi et al., 2013).

Monte Carlo Simulation

To estimate whether duplicate genes have higher average pollen expression level than the background, we conducted 1,000,000 Monte Carlo sampling testing. Monte Carlo significance test procedures consist of the observation of the observed data with random samples generated in accordance with the hypothesis being tested (Hope, 1968; Emerson et al., 2004). For each simulation, we randomly selected 374 genes from the \(A.\ thaliana\) genome, which is the number of younger duplicate genes (187 \(\times\) 2) identified in \(A.\ thaliana\). We then computed and compared the average expression level of these randomly selected genes with the average expression level of the younger duplicate genes in pollen. We counted the number of simulations with the average pollen expression levels of the randomly selected genes higher than that of younger duplicate genes, and then divided it by 1,000,000. This mathematical derivative is treated as the \(P\) value for the younger duplicate genes with higher average pollen expression level than the background. The same approach was applied to calculate the \(P\) value for the older duplicate genes.

Analysis of Selection Constraints of Paralogs

The coding sequences of each duplicate gene and its ancestral gene were collected and aligned with MACSE (mace_v0.9b1; Ranwez et al., 2011). The branch model of PAML was used to compute the Ka/Ks ratio for duplicate genes (Yang, 2007). We tested two branch models: model 1, with both the branch-specific Ka/Ks and background Ka/Ks varying freely; and model 2, with the branch-specific Ka/Ks fixed at 1 and background Ka/Ks varying freely. We then conducted the LRT, which tested whether the likelihood of model 1 was significantly different from that of model 2 by comparing \(2\times\) the log likelihood difference. \(P\) values were computed using a \(y^2\) distribution with 1 degree of freedom (Yang, 1998). Then, for each \(P\) value, we computed the corresponding FDR \(q\) value using the \(q\)-value package of R. We took \(q \leq 0.05\) as the cutoff of significance (Storey, 2002; Storey et al., 2004).

All the intermediate steps were carried out with custom PERL and R scripts.

Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries from NCBI under SRA accession number SRP080799.

Supplemental Data

The following supplemental materials are available.

Supplemental Table S1. The gene expression levels of young duplicate genes in \(A.\ thaliana\).

Supplemental Table S2. The gene expression levels of young duplicate genes in \(A.\ lyrata\).

Supplemental Table S3. Classifications based on Euclidean distance with different \(E_{0.5,0.2}\) cutoffs in \(A.\ thaliana\).

Supplemental Table S4. Classifications based on Euclidean distance with different \(E_{0.5,0.2}\) cutoffs in \(A.\ lyrata\).

Supplemental Table S5. The Ka/Ks ratios and LRT \(P\) value of young duplicate genes in \(A.\ thaliana\).

Supplemental Table S6. The Ka/Ks ratios and LRT \(P\) value of young duplicate genes in \(A.\ lyrata\).

Supplemental Table S7. The data corresponding to Figure 2.

Supplemental Table S8. The top enriched gene ontologies of young duplicate genes in the “Conservation” class.

Supplemental Table S9. The top enriched gene ontologies of young duplicate genes in the “Neofunctionalization” class.

Supplemental Table S10. The top enriched gene ontologies of young duplicate genes in the “Specialization” class.

Supplemental Table S11. The data corresponding to Figure 4.

ACKNOWLEDGMENTS

We thank Computing and Information Technology of Wayne State University for providing grid computing services and the staff at the Arizona Genomics Institute, University of Arizona, for performing RNA-seq library construction and the Illumina HiSeq run.

Received July 27, 2016, accepted August 1, 2016; published August 2, 2016.

LITERATURE CITED


Downloaded from www.plantphysiol.org on August 31, 2016 - Published by www.plantphysiol.org Copyright © 2016 American Society of Plant Biologists. All rights reserved.


