Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in
\textit{Arabidopsis} Genomes

Jun Wang$^\dagger$, Feng Tao$^\dagger$, Nicholas C. Marowsky, Chuanzhu Fan$^*$

Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA

$^\dagger$ Equal contributors

Conservation, neofunctionalization and specialization are main evolutionary trajectories for \textit{Arabidopsis} young duplicate genes, and their relative roles change dynamically over evolutionary time.

Funding information:
This work was supported by a start-up fund from Wayne State University to CF.

$^*$Corresponding author email: cfan@wayne.edu
Abstract

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 have been maintained for significantly long periods of evolutionary time. To address the manner in which young duplicate genes were primarily derived 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 *A. 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 with 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.

Key words: *Arabidopsis*, conservation, DNA methylation, histone modification, gene duplication, functionalization, specialization, transcriptome
Introduction

Gene duplication, which generates extra copies from ancestral genes, can provide raw materials for developing new functions and is one major means of contributing to the evolution of genomic novelty (Ohno, 1970; Ohta, 1989; Force et al., 1999; Lynch and Conery, 2000; Conant and Wolfe, 2008; Freeling, 2009; Kaessmann et al., 2009; Chen et al., 2013). Particularly, plant genomes maintain a large abundance of duplicate genes and allow more genomic redundancy, diversity and dynamics than animal systems (Kejnovsky et al., 2009; Kondrashov, 2012; Fischer et al., 2014). For example, approximately 65-85% of Arabidopsis genes are believed to have originated from duplication (Arabidopsis Genome, 2000; Cannon et al., 2004). Both large-scale duplication events, e.g., whole genome duplication and segmental duplication, and small-scale duplication events, e.g., dispersed and tandem duplications, are pervasive in plant genomes (Kaul et al., 2000; Moore and Purugganan, 2005; Ganko et al., 2007; Hanada et al., 2008; Van de Peer et al., 2009; Wang et al., 2013; Vanneste et al., 2014). Since the most recent whole genome duplication in the Arabidopsis lineage occurred before the divergence of Arabidopsis and Brassica genera about 20-43 MYA (Blanc et al., 2003; Beilstein et al., 2010), the duplicate genes that were preserved within genus Arabidopsis after this polyploidization were more likely derived from the small-scale duplication. Previous studies have investigated the extent of small-scale duplication and its potential contributions to the speciation and adaptation of plants (Rizzon et al., 2006; Freeling et al., 2008; Hanada et al., 2008; Freeling, 2009; Carretero-Paulet and Fares, 2012; Rodgers-Melnick et al., 2012; Wang et al., 2013; Zhang et al., 2013; Glover et al., 2015). Although we have gained knowledge about the evolutionary processes of duplicate genes from small-scale duplication events in plants (Duarte et al., 2006; Ganko et al., 2007; Edger and Pires, 2009; Zou et al., 2009; Liu et al., 2011), few study has
systematically addressed the evolutionary trajectories of young plant duplicate genes, which capture the earliest features of duplication and provide detailed information on gene origination (Owens et al., 2013; Wang et al., 2013). Due to the lack of empirical data for ancestral states of paralogs and the lack of available comprehensive genomic and transcriptomic data in plant genomes, how recent duplicates were maintained in plant genomes remained especially inconclusive.

Generally, after gene duplication, if the two copies survive in populations, they undergo four evolutionary trajectories: 1) conservation: the two copies maintain the same function as the ancestral gene; 2) neofunctionalization: one copy develops a novel function whereas the other copy retains the original function; 3) subfunctionalization: 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: 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 recently developed 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 to study the evolutionary processes of preserving duplicate genes in other species.

Here, we systematically address the following two questions: How do young duplicates originate and persist in genomes? What are the underlying mechanisms influencing their different evolutionary trajectories? To answer above questions, we generated and compared the transcriptomic profiles of Arabidopsis thaliana, A. lyrata, and Capsella rubella from high throughput RNA sequencing in five tissues. By taking advantage of the aforementioned phylogenetic approaches, 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 the 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* approximate 5-10 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 origination/retention rate of duplicate genes in these two species. Overall, the origination rate 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 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* (Supporting File 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* (Supporting File 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., 2009; 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 3097 and 3363 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 their *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*, respectively. To set up the level of baseline expression divergence for gene pairs, we also calculated the Euclidean expression distance between single-copy gene (S1) and their *C. rubella* ortholog copy (S2) as $E_{S1,S2}$ for respective *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 $E_{D2,A} > E_{S1,S2}$ and $E_{D1,A} \leq E_{S1,S2}$. For subfunctionalization, both duplicate genes should have different expression profiles from ancestral copy, but the combined duplicate genes 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 genes expression should all differ from the expression of 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 1).

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 (Table S1 and S2). Overall, 192 conservation, 104 neofunctionalization, and 158 specialization cases were classified for the species-specific paralogs. 19 conservation, 23 neofunctionalization, and 36 specialization cases were classified for the genus-specific paralogs (Table 1). 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 those of specialization cases increased (from 34.8% to 46.2%, Fisher test, P=0.03214) over evolutionary time scale (from 5MY to 10 MY).

We observed that sums of Euclidean expression distances of duplicate genes were
positively correlated with the Ks values (Spearman correlation, for *A. thaliana*, rho=0.2663235, P=0.000987; for *A. lyrata*, rho=0.1233875, P=0.01515), suggesting 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 the cases of subfunctionalization were extremely rare (5 and 0 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/the other 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 genes, 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 are 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 2). For species-specific paralogs, we identified 198 conservation cases, 91 neofunctionalization cases, 170 specialization cases, and 0 subfunctionalization case. For genus-specific paralogs, 29 conservation cases, 13 neofunctionalization cases, 35 specialization cases and 1 subfunctionalization case were categorized. The overall results based on expression localization patterns are consistent with the aforementioned results generated by Euclidean distance. This consistence 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 roles 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 The Arabidopsis Information Resources (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, 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%-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.

The 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 then analyzed three alternative metrics, including tissue specificities, relative expression level spectrum across tissues, and selection constraints of duplicate genes (Supporting File S3 and File S4). For the conservation class, duplicate genes have similar tissue specificities and similar relative expression spectra across the five tissues to their ancestral genes (Figure 1A, B). The two copies of the duplicate pairs also have similarly low Ka/Ks (Figure 1C). For the neofunctionalization class, the neofunctionalized duplicate genes have significant higher tissue specificities than their ancestral genes, whereas the tissue specificities of their paralogs are similar to their ancestral genes (Figure 1A, B). The neofunctionalized duplicate genes also have significantly higher Ka/Ks than the other paralogous copies (Figure 1C, Wilcox 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 (Figure 1A, B), and similarly high Ka/Ks, which are significantly higher than the Ka/Ks
of conservation class (Figure 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, FDR q-value < 0.05, see Materials and Methods), suggesting that a large proportion of the 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-value < 0.05), suggesting that they might be driven by positive selection.

**Test the “out of pollen” hypothesis for Arabidopsis young duplicate genes**

Previous studies showed that *Arabidopsis* new 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 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 level in mature pollen compared to in other tissues. Moreover, their expression level is also higher than single-copy genes and the older duplicate genes (i.e., genus-specific duplicate genes) in mature pollen (Figure 2A, B, Supporting File S5). Based on Monte Carlo simulations, we concluded that the younger duplicate genes have significantly higher average pollen expression level 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 level in tri-cellular pollen and mature pollen than in the uninucleate microspores and bi-cellular pollen (Figure 2C, Supporting File S5). 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=0.1990647, P=0.0005238). These results suggest that after new genes originated from pollen, they might acquire more interaction partners over evolutionary time (Figure 2D, Supporting File S5).
The underlying mechanisms contributing to the temporal evolutionary processes of duplicate genes

Firstly, we investigated the duplication modes of young duplicate genes in *A. thaliana*, which could contribute their divergence. Because the *A. lyrata* genome only had scaffold assembly but not chromosome assembly, this analysis only focused 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 3 paralog pairs from segmental duplication. The proportion of dispersed duplication temporally descended over the evolutionary time from 66.31% in younger duplicate genes to 46.55% in older duplicate genes, but the trend is 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* 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, Figure 3) and narrow expression spectrum across tissues (Figure 1B); neofunctionalized plant paralogs have ancestral genes with intermediate tissue specificities (Wilcoxon rank sum test, P < 0.05, Figure 3) and expression spectrum across tissues (Figure 1B); and conserved plant paralogs have
ancestral genes with lower tissue specificities (Wilcoxon rank sum test, P < 0.05, Figure 3) and broader expression spectrum across tissues (Figure 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), membrane part (7.03E-4) and so on (Table S3). The neofunctionalization and specialization classes have the enriched GO related to enzymes participating in either flexible steps or the tips of pathways, e.g. small molecule biosynthetic enzymes participating 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 ions activity (FDR=2.42E-05) and so on (Table S4, S5).

Additionally, we investigated how DNA methylation level and histone modifications abundances 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 1,000bp and downstream 1,000bp flanking regions of transcription start sites TSS), histone modifications abundances 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 modifications abundances and the gene expression level (Figure 4, Spearman’s correlation, see Supporting file S6 for correlation coefficients).

Interestingly, we found that the methylation levels in the promoter regions of younger duplicate genes are significantly higher compared to those of the older duplicate genes (Figure 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 (Figure 5, Wilcoxon rank sum test, P<0.05). Furthermore, younger duplicate genes have lower expression levels compared to the older duplicate genes (Figure 5, Wilcoxon rank sum test, P<0.05). Consistently, we found 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 3), whereas the
active histone modifications abundances in the promoter regions of duplicate genes are positively correlated with their Ks values (Spearman’s correlation, P< 0.01, Table 3).

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; and thus they are very likely to obtain higher expression levels.
Figure 5. The 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, ***P < 0.001.
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 5-10 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 were 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 Arabidopsis and Brassica genera 20-43 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, sub-, neo-, and loss-functionalization) only explained the divergence of duplicates over the long evolutionary time period but not about the scenario 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. Along 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 to survive and thus become pseudogenized later, leading to a decreasing proportion of conservation cases and a relatively increasing proportion of specialization.

Such a plant specific evolutionary pattern for duplicates is different from that in *Drosophila*, where the neofunctionalization of young duplicate genes play a major role in preserving the duplicates, but similar to that in mammals (Assis and Bachtrog, 2013, 2015). This different pattern can be explained 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* and mammals (Assis and Bachtrog, 2013, 2015). This is also supported by previous studies in plant genomes in which subfunctionalization only play a minor role, whereas neofunctionalization and specialization that correspond to the reciprocal neofunctionalization expression in some studies are more important in preserving duplicates and subfunctionalization of duplicated genes were mostly associated with whole genome duplication (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 classified into conservation or specialization categories. One possible solution might be to conduct 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 expressional 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 towards small molecule biosynthetic enzymes. This phenomenon may be accounted by following explanations. Firstly, the change of the functional roles/interaction partners of broad-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 function (neofunctionalization) or develop specialized functions that are distinct from ancestral copies (specialization).

Interestingly, young *Arabidopsis* duplicate genes are predominantly expressed in pollen, especially tri-cellular 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 is corresponded to the “out of testis” phenomenon of new genes in *Drosophila* and mammals (Betran 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”, highly welcome and tolerate genomic and functional
innovations such as new genes (Bernasconi et al., 2004; Kaessmann, 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 age of duplicate genes. These results suggest that after the origin of new
duplicate genes, their expressions are initially repressed by DNA methylation and lack
active histone modifications. This further support that the de-methylation in tri-cellular
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, *A. thaliana*, *A. lyrata*, *C. rubella* and *B. rapa*, for
comparative genomic analysis. The complete genome framework datasets including
assemblies and annotations were downloaded from Phytozome v8.0
(http://www.phytozome.net/) for *A. thaliana* 167 (TAIR release 10 acquired from TAIR),
*A. lyrata* 107 (JGI release v1.0), *C. rubella* 183 (JGI annotation v1.0 on assembly v1),
and *B. rapa* 197 (Annotation v1.2 on assembly v1.1 from brassicadb.org). We generated
RNA sequencing data of five tissues in *A. thaliana*, *A. lyrata* and *C. rubella* (see the
following section). The RNA-seq data of pollen were obtained from Short Read Archive
(SRA) in NCBI with accession SRR022162 (Loraine et al., 2013). The DNA methylation
bisulfite sequencing data of *A. thaliana* and *A. lyrata* genome (from control roots) were
downloaded from the European Nucleotide Archive with the accession number
PRJEB6701 (Seymour et al., 2014). The microarray data of 79 *A. thaliana*
tissues/developmental stages were downloaded from EMBL-EBI under ArrayExpress
with accession number E-AFMX-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 Gene Expression Omnibus of NCBI with series number GSE34190
(Honys and Twell, 2004). The ChIP-seq data of histone modifications from the leaf were
downloaded from Gene Expression Omnibus of NCBI with series number GSE28398
(Luo et al., 2013).

**RNA sequencing**

Plants of the reference lines *A. thaliana* (col-0), *A. lyrata* (MN47) and *C. rubella* (Monte
Gargano) were grown in the greenhouse under normal long-day growth conditions as 16h
(22°C) day/8h (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 6-8 weeks after germination. Floral
buds were collected when plants started bolting. Siliques were collected when young and
fleshy siliques formed. Roots of mature plants (~8 weeks) were dug out from soil and
then thoroughly washed to avoid contamination. Approximate 100 mg of plant tissues
were grounded and treated by TRIzol, and ~200-2000 ng/µl of total RNA were generated.

RNA sequencing libraries were made using the “NEBNext® Ultra™ Directional RNA
Library Prep Kit for Illumina” according to the manufacture’s protocol. Specifically,
mRNA was isolated using the NEBNext Oligo d(T)$_{26}$ and was then fragmented at 94°C
for 15 min. First and second cDNA were sequentially synthesized in succession. 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 (Aglient), 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 oligo sequence locating the downstream side of the adaptor. The loading volume of the pooled libraries was decided using real-time polymerase chain reaction. All libraries were sequenced in one lane of Illumina HiSeq run and then separated by the corresponding tag sequences.

**Identification of young duplicate genes in *Arabidopsis***

We conducted a BLAT search with all peptides against all peptides for *A. thaliana* and *A. lyrata*, respectively. Then, we picked the BLAT matches with identity score \( \geq 40 \) and coverage \( \geq 60\% \) of the peptide length, which are reciprocal best hits of each other. Ks values were calculated for all identified gene pairs using the pipeline previously developed (Zhang et al., 2013). We picked gene pairs as *A. thaliana* species-specific duplicate genes if Ks \( \leq 0.135284 \). We defined gene pairs as *Arabidopsis* genus-specific duplicate genes if \( 0.135284 < \text{Ks} \leq 0.254149 \) and 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 approximately infer the evolutionary divergent time of duplication.

Then we used BLASTP to align the peptide sequence of each duplicate gene against the peptides of *C. rubella* (*A. thaliana* vs. *C. rubella*; and *A. lyrata* vs. *C. rubella*). We only kept the duplicate gene pairs if the two copies of a duplicate gene pair have the same best-hit gene in *C. rubella* genome (BLASTP, \( E \leq 10^{-3} \)). 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 in next section.
We further selected the gene pairs if both copies are 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 expression as
$\text{FPKM} \geq 0.618412$ for $A.\text{thaliana}$, $\text{FPKM} \geq 0.1205161$ for $A.\text{lyrata}$ and $\text{FPKM} \geq$
$0.537131$ for $C.\text{rubella}$, which correspond to the 35% percentile of the exonic FPKM.

**Defining the ortholog relationship**

We determined the ortholog relationship for genes of interest among $A.\text{thaliana}$, $A.\text{lyrata}$, and $C.\text{rubella}$ (Wang et al., 2013). Briefly, we used the developed pipelines from
the UCSC genome browser to construct the whole genome syntenic relationship between
each of $A.\text{thaliana}/A.\text{lyrata}/C.\text{rubella}/B.\text{rapa}$ with the remaining three species. We also
used the BLASTP to search for the reciprocal best hits of peptide sequences among the
above species ($E \leq 10^{-3}$). If a gene is in the syntenic region of its own genome and the
outgroup genome, and has a reciprocal best hit peptide sequence in the outgroup species,
we decided that this gene has the ortholog in the outgroup species, and 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 -G) (Trapnell et al., 2010). Then, the FPKM values of all tested tissues in one
species were quantile-normalized with the “normalize.quantiles” function in R (Assis et
The Affymatrix microarray data were annotated to *A. thaliana* genes with a customized CDF file downloaded from http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/17.1.0/tairg.asp. The data were then processed using the RMA and MAS5 package of R.

The BS-seq data were processed according to the same procedure as previously described (Wang et al., 2014). Briefly, we used trim_galore (v0.3.7, with the options: --length 40 -paired --trim1 --o --retain_unpaired) to remove adaptors and low quality reads and used Bismark (v0.13.0, with the options: --seedlen 40 --seedmms 2 --maqerr 100 --chunkmbs 1024 --minins 100 --maxins 1000 --ambiguous) 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 previously described (Seymour et al., 2014). The methylation level of each region was estimated as the percentage of methylated cytosines over the total mapped cytosines in that region for the three cytosine contexts (CG, CHG, CHH). In our analysis, we only considered methylation in the CG context and the regions in which at least 50% of the cytosines were mapped.

We used Bowtie (v1.1.2) to align the color-space ChIP-seq reads to the *A. thaliana* genome allowing ≤3 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 (v0.6.1p1, with the options: -f sam -r name -s no -t exon -i gene_id -m union) to count the number of the reads in the upstream 1000bp and downstream 1000bp flanking regions of TSS (Anders et al., 2015). The relative histone modification abundance was defined as “n(Histone) x N(Input) / (N(Histone) x n(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 mean 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 $E_{S1,S2}$, because it was
less affected by the extreme values. We also tried various other cutoffs, e.g., mean,
median, standard deviation from mean, median absolute deviation from median with
different constants, semi-interquartile from median and different quantile values. All the
results drew the same conclusion (Table S1 and S2).

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 to 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 $E_{D1,A}$, $E_{D2,A}$, and $E_{D1+D2,A}$ with the cutoff of $E_{S1,S2}$. According to the rules
shown in “Results” and Table 1, 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 duplicate genes” section. 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 $D_{D1,A}$, $D_{D2,A}$, and $D_{D1+D2,A} = 0$ (the same
Based on the D values and the rules shown in “Results” and Table 2, 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 the ANAP (http://gmdd.shgmo.org/Computational-Biology/ANAP/ANAP_V1.1/) (Wang et al., 2012), AtPIN (http://bioinfo.esalq.usp.br/atpin/atpin.pl) (Brandao et al., 2009), Arabidopsis Interactions Viewer (http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi) (Popescu et al., 2007; Popescu et al., 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 non-redundant protein partners for each gene. Because *A. lyrata* does not have such data available, it was excluded in this analysis. For comparison of tissue specificity, a tissue specificity index (τ) was computed for each gene based on the RNA-seq data in the five tissues according to the modified approach by Yanai et al. (Yanai et al., 2005; Assis et al., 2012). τ is in the range from 0 to 1. The higher the τ value, the more tissue specific the gene is expressed in. The mean relative expression levels in heat maps were calculated in each of the tissues for single-copy, 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*, respectively. The gene ontology enrichment analyses were conducted using
Monte Carlo simulation
To estimate whether duplicate genes have a 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 comparison 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 *A. thaliana*
genome, which is the number of younger duplicate genes (187×2) identified in *A. thaliana*. We then computed and compared the average expression level of these
randomly selected genes to 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.
Same approach was applied to calculate the P-value for the older duplicate genes.

Analysis of selection constraints of paralogs
The CDS sequences of each duplicate gene and its ancestral gene were collected and
aligned with MACSE (macse_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 Likelihood Ratio Test (LRT), which tested
whether the likelihood of model 1 was significantly different from that of model 2 by
comparing two times the log likelihood difference. P-values were computed using a
chi-square 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-value ≤ 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. The GenBank accession number of our RNA-seq data is XXXX.

Acknowledgements

Computing & Information Technology of Wayne State University provided grid computing services. The staff at Arizona Genomics Institute, University of Arizona, performed RNA-seq library construction and Illumina HiSeq run.

Authors’ contributions

CF conceived the project. CF and JW designed the experiments. JW, FT and NCM performed the experiments. JW and FT analyzed the data. JW, FT, NCM and CF contributed to the writing of the manuscript.
### Table 1. Classification of evolutionary processes of *Arabidopsis* duplicate genes by measuring the Euclidean expression distance

| Classification         | E\(_{D1,A}\) | E\(_{D2,A}\) | E\(_{D1-D2,A}\) | Species-specific | Genus-specific |
|------------------------|--------------|--------------|-----------------|------------------|----------------|
|                        |              |              |                 | AT   | AL  | Sum | AT   | AL  | Sum |
| Conservation           | ≤\(E_{S1,S2}\) | ≤\(E_{S1,S2}\) | -               | 47   | 145 | 192 | 10   | 9   | 19  |
| Neofunctionalization   | >\(E_{S1,S2}\) | ≤\(E_{S1,S2}\) | -               | 30   | 74  | 104 | 10   | 13  | 23  |
| Subfunctionalization   | >\(E_{S1,S2}\) | >\(E_{S1,S2}\) | ≤\(E_{S1,S2}\) | 1    | 4   | 5   | 0    | 0   | 0   |
| Specialization         | >\(E_{S1,S2}\) | >\(E_{S1,S2}\) | >\(E_{S1,S2}\) | 34   | 124 | 158 | 18   | 18  | 36  |

AT: *A. thaliana*; AL: *A. lyrata*
Table 2. Classification of evolutionary processes of *Arabidopsis* duplicate genes by measuring the binary pattern of expression localization

| Classification          | $D_{D1}$ | $D_{D2}$ | $D_{D1+D2}$ | Species-specific | Genus-specific |
|-------------------------|----------|----------|--------------|-----------------|---------------|
| Conservation            | 0        | 0        | -            | 49              | 17            |
| Neofunctionalization    | 1        | 0        | -            | 30              | 4             |
| Subfunctionalization    | 1        | 1        | 0            | 0               | 1             |
| Specialization          | 1        | 1        | 1            | 33              | 16            |

AT: *A. thaliana*; AL: *A. lyrata*
Table 3. The correlation between the abundances of active histone modifications/DNA methylation and Ks of duplicate genes

| Epigenetic modification | Spearman’s correlation coefficient | P_value     |
|-------------------------|------------------------------------|-------------|
| H3K9Ac                  | 0.1924446                          | 3.52E-05*** |
| H3K4me2                 | 0.2108966                          | 5.563e-06***|
| H3K4me3                 | 0.1471653                          | 0.001626**  |
| H3K36me2                | 0.1759113                          | 0.0001596***|
| H3K36me3                | 0.135351                           | 0.003783**  |
| DNA methylation         | -0.1314893                         | 0.004918**  |

*: p-value<0.05; **: p-value <0.01, ***: p-value <0.0001
Figure Legends

Figure 1. Comparison of functional metrics for paralogs classified in three evolutionary processes

A, Scatterplot of tissue specificities (τ) between duplicates and ancestral genes. B, Heat maps show the mean relative expression levels in five tissues of single-copy (S), duplicates (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 the significance. *P < 0.05, **P<0.01, ***P < 0.001.

Figure 2. “Out of pollen” pattern of Arabidopsis 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 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 red boxes. C, The relative expression level of duplicate genes at the four pollen developmental stages. We computed the relative expression level using the mean expression level of duplicate genes divided by the mean expression level of single-copy genes (Wu et al., 2014). D, The distributions of protein-protein interaction numbers (PPI) of younger (Y) and older (O) duplicate genes. We compared PPI between younger and older duplicates. The significance was computed by using the Wilcoxon rank sum test. *P < 0.05, **P<0.01, ***P < 0.001.
Figure 3. The tissue specificity indexes of ancestral genes

The boxplot shows the distribution of tissue specificity indexes of ancestral genes for three evolutionary classes of duplicates: conservation, neofunctionalization and specialization. We compared the tissue specificity indexes of the three classes with the Wilcoxon rank sum test. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Figure 4. The heatmap of Spearman’s correlation coefficients among DNA methylation level, the abundances of active histone modifications, and gene expression level of duplicate genes

The boxplots show the abundances of promoter (the upstream 1,000bp and downstream 1,000bp flanking regions of TSS) histone modifications, promoter DNA methylation levels, and gene expression levels, for younger and older duplicates, respectively. The abundances of active histone modifications are positively correlated with each other and with gene expression level, whereas DNA methylation level is negatively correlated with the active histone modifications abundances and gene expression level.

Figure 5. The 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”. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. 

36
Supporting information:

Table S1. Classifications based on Euclidean distance with different $E_{S1,S2}$ cutoffs in *A. thaliana*

Table S2. Classifications based on Euclidean distance with different $E_{S1,S2}$ cutoffs in *A. lyrata*

Table S3. The top enriched gene ontologies of duplicate genes in the “Conservation” class

Table S4. The top enriched gene ontologies of duplicate genes in the “Neofunctionalization” class

Table S5. The top enriched gene ontologies of duplicate genes in the “Specialization” class

File S1. The gene expression levels of young duplicate genes in *A. thaliana*

File S2. The gene expression levels of young duplicate genes in *A. lyrata*

File S3. The Ka/Ks ratios and LRT p-value of young duplicate genes in *A. thaliana*

File S4. The Ka/Ks ratios and LRT p-value of young duplicate genes in *A. lyrata*

File S5. The data corresponding to Figure 2.

File S6. The data corresponding to Figure 4.
A nders S, Pyl PT, Huber W (2015) HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166-169

Arabidopsis Genome I (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815

Arunkumar R, Josephs EB, Williamson RJ, Wright SI (2013) Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella grandiflora. Molecular Biology and Evolution 30: 2475-2486

Assis R, Bachtrog D (2013) Neofunctionalization of young duplicate genes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 110: 17409-17414

Assis R, Bachtrog D (2015) Rapid divergence and diversification of mammalian duplicate gene functions. Bmc Evolutionary Biology 15: 138

Assis R, Zhou Q, Bachtrog D (2012) Sex-biased transcriptome evolution in Drosophila. Genome Biol Evol 4: 1189-1200

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America 107: 18724-18728

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J (2006) Positive selection driving diversification in plant secondary metabolism. Proceedings of the National Academy of Sciences of the United States of America 103: 9118-9123

Bernasconi G, As mash TL, Birkhead TR, Bishop JDD, Grossniklaus U, Kubli E, Marshall DL, Schmidt B, Skogsmyr I, Snook RR, Taylor D, Till-Bottraud I, Ward PI, Zeh DW, Hellriegel B (2004) Evolutionary ecology of the prezygotic stage. Science 303: 971-975

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

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res 13: 137-144

Brandao MM, Dantas LL, Silva-Filho MC (2009) APIN: Arabidopsis thaliana protein interaction network. BMC Bioinformatics 10: 454

Braun P, Carvunis AR, Charloteaux B, Dreze M, Ecker JR, Hill DE, Roth FP, Vidal M, Galli M, Balumuri P, Bautista V, Chesnut JD, Kim RC, de los Reyes C, Gilles P, Kim CJ, Matrubutham U, Mirchandani J, Olivares E, Patnaik S, Quan R, Ramaswamy G, Shinn P, Swamlingiah GM, Wu S, Ecker JR, Dreze M, Byrdsong D, Dricot A, Duarte M, Gebreab F, Gutierrez BJ, MacWilliams A, Monachello D, Mukhtar MS, Poulin MM, Reichert P, Romero V, Tam S, Waaliers S, Weiner EM, Vidal M, Hill DE, Braun P, Galli M, Carvunis AR, Cusick ME, Dreze M, Romero V, Roth FP, Tanan M, Yazaki J, Braun P, Ecker JR, Carvunis AR, Ahn YY, Barabasi AL, Charloteaux B, Chen HM, Cusick ME, Dangl JL, Dreze M, Ecker JR, Fan CY, Gai LT, Galli M, Ghoshal G, Hao T, Hill DE, Lurin C, Milenkovic T, Moore J, Mukhtar MS, Pezvzner SJ, Przulj N, Rabello S, Rietman EA, Rolland T, Roth FP, Santhanam B, Schmitz RJ, Spooner W, Stein J, Tanan M, Vandenhaute J, Ware D, Braun P, Vidal M, Braun P, Carvunis AR, Charloteaux B, Dreze M, Galli M, Vidal M, Co
AIM (2011) Evidence for Network Evolution in an Arabidopsis Interactome Map. Science 333: 601-607

Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijo JA, Becker JD, Martienssen RA (2012) Reprogramming of DNA Methylation in Pollen Guides Epigenetic Inheritance via Small RNA. Cell 151: 194-205

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol 4: 10

Carretero-Paulet L, Fares MA (2012) Evolutionary Dynamics and Functional Specialization of Plant Paralogs Formed by Whole and Small-Scale Genome Duplications. Molecular Biology and Evolution 29: 3541-3551

Chen SD, Krinsky BH, Long MY (2013) New genes as drivers of phenotypic evolution (vol 14, pg 645, 2013). Nature Reviews Genetics 14: 745-745

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938-950

Cui X, Lv Y, Chen M, Nikolski Z, Twell D, Zhang D (2015) Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. Mol Plant 8: 935-945

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

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol 23: 469-478

Edger PP, Pires JC (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 17: 699-717

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

Fischer I, Dainat J, Ranwez V, Glenn D, Dufayard JF, Chantret N (2014) Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol 14: 151

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

Freeling M (2009) Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. Annual Review of Plant Biology 60: 433-453
Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D (2008) Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Research 18: 1924-1937

Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in Arabidopsis. Mol Biol Evol 24: 2298-2309

Glover NM, Daron J, Pingault L, Vandepoele K, Paux E, Feuillet C, Choulet F (2015) Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biology 16: 188

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiology 148: 993-1003

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993-1003

He XL, Zhang JZ (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157-1164

Honys D, Twell D (2004) Transcriptome analysis of haploid male gametophyte development in Arabidopsis. Genome Biol 5: R85

Hope ACA (1968) A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society. Series B (Methodological) 30: 582-598

Hughes TE, Langdale JA, Kelly S (2014) The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome Research 24: 1348-1355

Innan H (2009) Population genetic models of duplicated genes. Genetica 137: 19-37

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Research 20: 1313-1326

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

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nieman W, Benito MI, Lin XY, Town CD, Venter JC, Fraser CM, Tabata S, Nakamura Y, Kaneko T, Sato S, Asanizu E, Kato T, Kotani H, Sasamoto S, Ecker JR, Theologis A, Federspiel NA, Palm CJ, Osborne BI, Shinn P, Conway AB, Vysotskaia VS, Dewar K, Conn L, Lenz CA, Kim CJ, Hansen NF, Liu SX, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham PK, Chao Q, Nguyen M, Yu GX, Chen HM, Southwick A, Lee JM, Miranda M, Toriumi MJ, Davis RW, Wambutt R, Murphy G, Dusterhoft A, Stiekema W, Pohl T, Entian KD, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansorge W, Unseld M, Hartmann B, Valge G, Artiguenave F, Weissenhoch J, Quettier F, Wilson RK, de la Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie WR, Mayer K, White O, Bevan M, Lemcke K, Creasy TH, Bielke C, Haas B, Haase D, Maiti R, Rudd S,

CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in Arabidopsis. Mol Biol Evol 24: 2298-2309

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Glover NM, Daron J, Pingault L, Vandepoele K, Paux E, Feuillet C, Choulet F (2015) Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biology 16: 188

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiology 148: 993-1003

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993-1003

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

He XL, Zhang JZ (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157-1164

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Honys D, Twell D (2004) Transcriptome analysis of haploid male gametophyte development in Arabidopsis. Genome Biol 5: R85

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hope ACA (1968) A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society. Series B (Methodological) 30: 582-598

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hughes TE, Langdale JA, Kelly S (2014) The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome Research 24: 1348-1355

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Innan H (2009) Population genetic models of duplicated genes. Genetica 137: 19-37

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Research 20: 1313-1326

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

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

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nieman W, Benito MI, Lin XY, Town CD, Venter JC, Fraser CM, Tabata S, Nakamura Y, Kaneko T, Sato S, Asanizu E, Kato T, Kotani H, Sasamoto S, Ecker JR, Theologis A, Federspiel NA, Palm CJ, Osborne BI, Shinn P, Conway AB, Vysotskaia VS, Dewar K, Conn L, Lenz CA, Kim CJ, Hansen NF, Liu SX, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham PK, Chao Q, Nguyen M, Yu GX, Chen HM, Southwick A, Lee JM, Miranda M, Toriumi MJ, Davis RW, Wambutt R, Murphy G, Dusterhoft A, Stiekema W, Pohl T, Entian KD, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansorge W, Unseld M, Hartmann B, Valge G, Artiguenave F, Weissenhoch J, Quettier F, Wilson RK, de la Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie WR, Mayer K, White O, Bevan M, Lemcke K, Creasy TH, Bielke C, Haas B, Haase D, Maiti R, Rudd S,
development. Science 274: 1537-1540

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Current Opinion in Plant Biology 8: 122-128

Mukhtar MS, Carvunis AR, Dreze M, Epple P, Steinbrenner J, Moore J, Tasan M, Galli M, Hao T, Nishimura MT, Pevzner SJ, Donovan SE, Ghamarsi L, Santhanam B, Romero V, Poulin MM, Gebreab F, Gutierrez BJ, Tam S, Monachello D, Boxem M, Harbort CJ, McDonald N, Gai LT, Chen HM, He YJ, Vandenhaute J, Roth FP, Hill DE, Ecker JR, Vidal M, Beynon J, Braun P, Dangl JL, Conso EUE (2011) Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. Science 333: 596-601

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin, New York,

Ohta T (1989) Role of gene duplication in evolution. Genome 31: 304-310

Owens SM, Harberson NA, Moore RC (2013) Asymmetric Functional Divergence of Young, Dispersed Gene Duplicates in Arabidopsis thaliana. Journal of Molecular Evolution 76: 13-27

Park JI, Semyonov J, Chang CL, Yi W, Warren W, Hsu SYT (2008) Origin of INSL3-mediated testicular descent in therian mammals. Genome Research 18: 974-985

Popescu SC, Popescu GV, Bachan S, Zhang ZM, Gerstein M, Snyder M, Dinesh-Kumar SP (2009) MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. Genes & Development 23: 80-92

Popescu SC, Popescu GV, Bachan S, Zhang ZM, Seay M, Gerstein M, Snyder M, Dinesh-Kumar SP (2007) Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. Proceedings of the National Academy of Sciences of the United States of America 104: 4730-4735

Ranwez V, Harispe S, Desuc L, Douzery EJ (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6: e22594

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. Plos Computational Biology 2: 999-1000

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP (2012) Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. Genome Research 22: 95-105

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nature Genetics 37: 501-506

Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genet 10: e1004785
Sisu C, Pei BK, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang DF, Rutenberg-Schoenberg M, Clark W, Diekhans M, Rozowsky J, Hubbard T, Harrow J, Gerstein MB (2014) Comparative analysis of pseudogenes across three phyla. Proceedings of the National Academy of Sciences of the United States of America 111: 13361-13366

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang XL, Gu HC, Barthes P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS, Kaessmann H (2013) Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Reports 3: 2179-2190

Stoltzfus A (1999) On the possibility of constructive neutral evolution. Journal of Molecular Evolution 49: 169-181

Storey JD (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology 64: 479-498

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society Series B-Statistical Methodology 66: 187-205

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-U174

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends in Plant Science 14: 680-688

Vanneste K, Baele G, Maere S, Van de Peer Y (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Research 24: 1334-1347

Walsh B (2003) Population-genetic models of the fates of duplicate genes. Genetica 118: 279-294

Wang CM, Marshall A, Zhang DB, Wilson ZA (2012) ANAP: An Integrated Knowledge Base for Arabidopsis Protein Interaction Network Analysis. Plant Physiology 158: 1523-1533

Wang J, Marowsky NC, Fan C (2013) Divergent Evolutionary and Expression Patterns between Lineage Specific New Duplicate Genes and Their Parental Paralogs in Arabidopsis thaliana. PLoS One 8: e72362

Wang J, Marowsky NC, Fan C (2014) Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes. PLoS One 9: e110357

Weng JK, Li Y, Mo H, Chapple C (2012) Assembly of an evolutionarily new pathway for alpha-pyrone biosynthesis in Arabidopsis. Science 337: 960-964
Wu DD, Wang X, Li Y, Zeng L, Irwin DM, Zhang YP (2014) "Out of Pollen" Hypothesis for Origin of New Genes in Flowering Plants: Study from Arabidopsis thaliana. Genome Biology and Evolution 6: 2822-2829

Yanai I, Benjamini H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650-659

Yang DL, Zhang GP, Tang K, Li JW, Yang L, Huang H, Zhang H, Zhu JK (2016) Dicer-independent RNA-directed DNA methylation in Arabidopsis. Cell Research 26: 66-82

Yang ZH (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591

Yi X, Du Z, Su Z (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Research 41: W98-W103

Zhang C, Wang J, Marowsky NC, Long M, Wing RA, Fan C (2013) High occurrence of functional new chimeric genes in survey of rice chromosome 3 short arm genome sequences. Genome Biol Evol 5: 1038-1048

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice. Plant Physiology 151: 3-15

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH (2009) Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. PLoS Genet 5: e1000381