Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes

Ya-Long Guo\(^1,2,\ast\)

\(^1\)State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China, and
\(^2\)Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Received 22 May 2012; revised 1 November 2012; accepted 23 November 2012; published online 15 January 2013.

\(^\ast\)For correspondence (e-mail yalong.guo@ibcas.ac.cn or ya-long.guo@hotmail.com).

**SUMMARY**

Gene family size variation is an important mechanism that shapes the natural variation for adaptation in various species. Despite its importance, the pattern of gene family size variation in green plants is still not well understood. In particular, the evolutionary pattern of genes and gene families remains unknown in the model plant *Arabidopsis thaliana* in the context of green plants. In this study, eight representative genomes of green plants are sampled to study gene family evolution and characterize the origination of *A. thaliana* genes, respectively. Four important insights gained are that: (i) the rate of gene gains and losses is about 0.001359 per gene every million years, similar to the rate in yeast, *Drosophila*, and mammals; (ii) some gene families evolved rapidly with extreme expansions or contractions, and 2745 gene families present in all the eight species represent the ‘core’ proteome of green plants; (iii) 70% of *A. thaliana* genes could be traced back to 450 million years ago; and (iv) intriguingly, *A. thaliana* genes with early origination are under stronger purifying selection and more conserved. In summary, the present study provides genome-wide insights into evolutionary history and mechanisms of genes and gene families in green plants and especially in *A. thaliana*.

Keywords: *Arabidopsis thaliana*, evolutionary genomics, gene family, gene origination, green plants.

**INTRODUCTION**

Most genes can be grouped into gene families based on sequence similarity. Gene family size is variable across different lineages and may have important functional outcomes related to adaptation or speciation (Lynch and Conery, 2000; Nei and Rooney, 2005).

Various factors can alter the size of gene families (De Bodt et al., 2005; Maere et al., 2005; Hahn et al., 2007; Freeling, 2009; Proost et al., 2011; Tautz and Domazet-Loso, 2011; De Smet and Van De Peer, 2012). First, gene duplication or whole-genome duplication (polyploidy) increases gene family size. Second, gene deletion decreases gene family size. Third, the *de novo* creation of genes creates new gene families. Gene duplication is an ongoing contributor to genome evolution and is of the same order of magnitude as the mutation rate per nucleotide site (Lynch and Conery, 2000, 2003; Gu et al., 2002; Mc lysaght et al., 2002; Simillion et al., 2002; Doyle et al., 2008; Flagel and Wendel, 2009; Van De Peer et al., 2009; Van De Peer, 2011). While duplication and *de novo* creation of genes increase the gene number, gene deletion decreases the gene number either by deletion of single gene or several genes within segmental fragment.

Although variation in gene family has been the subject of many studies, most analyses have focused on only one or a few gene families across a small number of plant genomes (Shiu et al., 2004; Chardon and Damerval, 2005; Kim et al., 2006; Leseberg et al., 2006; Li et al., 2006; Porter et al., 2009; Guo et al., 2011a) due to the limitation of available plant genomes or lack of robust pipelines. Recently, two new methods have been developed to address the evolutionary pattern of genes and gene families at the whole-genome level, including the likelihood approach implemented in CAFE (Computational Analysis of gene Family Evolution) (De Bie et al., 2006) and the gene emergence approach (Domazet-Loso et al., 2007; Cai et al., 2009; Zhang et al., 2010a; Banks et al., 2011). The likelihood approach uses a stochastic birth and death process to model the evolution of gene family size over a phylogeny (Hahn et al., 2005, 2007; De Bie et al., 2006; Demuth et al., 2006). In contrast, the gene emergence approach...
maps genes to different phylogenetic ranks (internodes) according to their emergence by similarity searches in the representative genomes, called as ‘phylostratigraphy’, naming phylogenetic ranks or internodes ‘phylostrata’ (Domazet-Loos et al., 2007; Prachumwat and Li, 2008; Cai et al., 2009; Zhang et al., 2010b, 2011; Banks et al., 2011; David and Alm, 2011). Recently, this approach has been used to study the protein domain evolution and transcription-associated proteins as well (Lang et al., 2010; Kersting et al., 2012; Zhang et al., 2012).

The comparison of gene families among closely related and representative species of green plants (Plantae or Viridiplantae), comprising land plants and green algae, allows a comprehensive study of the genes and gene families evolution in green plants. Eight representative genomes of green plants were compared, including two Arabidopsis genomes (Figure 1). Here I have investigated several fundamental questions regarding genes and gene families evolution in green plants: (i) the distribution of the gene family size; (ii) the rate of the gene gain and loss, the rate of the expansion or contraction per gene family on each lineage; and (iii) for A. thaliana genes in different phylostrata, the variation of genes or gene families fixation rate, the variation of functional annotation, of protein size, of expression level, and the variation of selection strength.

In summary, the analyses of gene families flanking the eight representative genomes of green plants provide in-depth insights into the researches of evolutionary genomics about green plants.

RESULTS

Gene family size distribution

All proteins from the eight genomes are used for the following analyses, since none of them is homologous to transposable elements. Markov Cluster Algorithm (MCL) clustering of all these proteins resulted in 55,989 gene families; where 13.6% of genes are singletons, 4.6% are grouped into two-gene families, and 81.7% are grouped into multi-gene families (Table 1, see Data S1 for the detail).

The clustering of genes in each genome indicates that the number of gene families ranges from 8692

![Figure 1. Phylogeny and divergence time (million years ago) of the eight green plants.](image)

<table>
<thead>
<tr>
<th>Species</th>
<th>Singletons</th>
<th>Two-genes</th>
<th>≥3</th>
<th>Total gene families</th>
<th>Max_gf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>9088</td>
<td>890</td>
<td>697 (4293)</td>
<td>10,675 (15,161)</td>
<td>132</td>
</tr>
<tr>
<td>Physcomitrella patens</td>
<td>11,513</td>
<td>2,081</td>
<td>2,115 (20,241)</td>
<td>15,709 (35,916)</td>
<td>4322</td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>1,717</td>
<td>4,348</td>
<td>2,627 (24,262)</td>
<td>8,692 (34,675)</td>
<td>1,541</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>13,597</td>
<td>2,201</td>
<td>2,589 (37,633)</td>
<td>18,387 (55,632)</td>
<td>2,208</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>7,579</td>
<td>1,850</td>
<td>2,419 (24,308)</td>
<td>11,848 (35,587)</td>
<td>1,258</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>6525</td>
<td>2,006</td>
<td>2,465 (33,235)</td>
<td>12,796 (45,372)</td>
<td>326</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>6551</td>
<td>1,964</td>
<td>2,592 (21,990)</td>
<td>11,107 (32,469)</td>
<td>226</td>
</tr>
<tr>
<td>All eight genomes</td>
<td>36,330</td>
<td>6,548</td>
<td>11,111 (230,225)</td>
<td>55,989 (281,651)</td>
<td>4322</td>
</tr>
</tbody>
</table>

*The number of singletons.

bThe number of gene families (genes) with two genes.

cThe number of gene families (genes) contain at least three genes.

The number of total gene families (genes).

The maximum gene family size.
genes within gene families varies from 15,161 (C. reinhardtii) to 45,372 (P. trichocarpa) (Figure 2a and Table 1). In addition, the proportion of singletons in each genome ranges from 5% (S. moellendorfii) to 59.9% (C. reinhardtii); genes in two-gene families from 7.9% (O. sativa) to 25.1% (S. moellendorfii); and genes in multigene families from 28.3% (C. reinhardtii) to 73.3% (P. trichocarpa). The maximum gene family size varies from 132 (C. reinhardtii) to 4322 (P. patens). Interestingly, for the closely related species within the same genus or family, the ratios of genes within different classes are similar, such as O. sativa and S. bicolor, A. lyrata and A. thaliana. Most genes grouped into multigene families in each genome except C. reinhardtii, in which singletons takes up more than half of total genes (59.9%). However, only 2.4% (217/9088) of singletons in C. reinhardtii can be annotated to Gene Ontology (GO) terms, which are related with cellular process, metabolic process, response to stimulus, developmental process, cellular component organization, and multicellular organisinal process.

Orphan genes and loss of entire gene families in Arabidopsis

There are many orphan genes in each genome, which do not have homologous in other genomes. The proportions of orphan genes in different genomes are different, 61.5% in C. reinhardtii, while other species have a much lower proportion, especially two dicot species A. thaliana (5.3%) and A. lyrata (12.3%) (Figure 2b and Table 2). The maximum orphan gene family size ranges from 17 (A. thaliana) to 4322 (P. patens); the total number of orphan gene families varies from 1328 (A. thaliana) to 10,812 (O. sativa); and the number of genes within orphan gene families varies from 1430 (A. thaliana) to 19,351 (P. patens). Among the eight genomes, 16.8% (S. moellendorfii) to 89.1% (A. thaliana) of orphan gene families are singletons; 4.8% (A. thaliana) to 31.3% (S. moellendorfii) are two-gene families; and 6.2% (A. thaliana) to 51.9% (S. moellendorfii) are multigene families.

To understand the functional characteristics of the orphan genes, the well studied model species A. thaliana was selected as a representative to perform GO functional annotation. 1042 genes (72.9%) of A. thaliana orphan genes (1430) can be annotated to GO terms, though most (93.0%) are classified as unknown biological processes; other terms are other cellular processes, other metabolic processes, protein metabolism, transcription, and response to stress.

Another class of interesting genes represents those with homologues between one Arabidopsis species and at least one of other six species but not in another Arabidopsis species, which could have been deleted from the latter Arabidopsis species. There are 98 gene families (174 genes, the minimum is 1, the maximum is 8, the average is 1.7755) only have A. lyrata genes; and 117 gene families (163 genes, minimum 1, maximum 7, average 1.3932) only have A. thaliana genes. Based on the unique genes in each Arabidopsis species but shared with other species, the gene extinction rate for the entirely lost gene families can be calculated. Assuming the divergence time of A. thaliana and A. lyrata to be 13 million years ago (Beilstein et al., 2010; Guo et al., 2011b) (Figure 1) and normalized to the total gene number, extinction rate in A. thaliana is 0.00050 (174/13 × 27 025) extinctions/million year/gene; in A. lyrata, 0.00038 (163/13 × 32 670) extinctions/million year/gene.

Gene family expansion or contraction

The likelihood approach assumes that there is at least one ancestral gene in the most recent common ancestor (MRCA), thus only those gene families that are shared between C. reinhardtii and at least one of other seven species are included for the analysis. Of the 3296 families present in the MRCA of green plants, 551 gene families have been lost completely in at least one species. It is 2745 gene families that are present across all the eight species most probably represent the ‘core’ proteome of green plants.

Assuming that each species has the same $\lambda$ value, the average rate of gene turnover is estimated to be 0.001379 (gains and losses/gene/million years), the rate of gene family to either expand or contract over time. The estimated rate of gene gains and losses implies that within a single
genome like *A. thaliana*, there are approximately 37,267 new duplicates and 37,267 new losses fixed every million years (0.001379 gains and losses/gene/million years × 27 025 genes).

The average expansions and contractions among different lineages are different (Figure 3). Along the phylogenetic ranks, most lineages expanded, including embryophyte, tracheophyte, angiosperm, and rosid. In contrast, lineages of *C. reinhardtii*, grass, and Arabidopsis contracted, and Arabidopsis genus is the most contracted lineage (average contraction size is \(/C0.547\)).

The likelihood approach can identify individual gene families evolving at the rate of gain or loss significantly higher than the genome-wide average (Hahn et al., 2005). Of the 3296 gene families, 41 are evolved non-randomly at \(P < 0.0001\). Of the 41 rapidly evolved gene families, only 37 contain *A. thaliana* genes, and 34 of which can be annotated to GO terms. Rapidly evolving gene families are associated with various biological processes, including other metabolic processes, other cellular processes, response to stress, protein metabolism, other biological processes, and response to abiotic or biotic stimulus (Tables S1 and S2). Nine of the 41 rapidly evolving gene families have significantly changed \((P < 0.0001)\) on the terminal branch to *A. thaliana*. All of the nine gene families significantly contracted \((P < 0.0001)\), most of which are affiliated with nucleotide binding, protein binding, DNA or RNA binding, hydrolase activity, and structural molecular activity (Table S2).

Among nine significantly changing gene families on the *A. thaliana* terminal, eight of them have *A. thaliana* genes and in total 43 genes, of them 34 have the orthologues in *A. lyrata*. The average of \(dN/dS\) for these genes is 0.1505 (median 0.0466), lower than all proteins of *A. thaliana* within the gene families (average 0.2799, median 0.1786) (Wilcoxon rank sum test, \(P = 4.525 \times 10^{-6}\)); the average of \(dS\) is 0.1480 (minimum 0.0427, maximum 0.3557, median 0.1398), also a little lower than all proteins of *A. thaliana* within the gene families (average 0.2167, median 0.1448) \((P = 0.3325)\). The results indicate there is not any correlation between the gene copy variation and the sequence divergence.

### Table 2 Number of orphan gene families (and genes) in green plants

<table>
<thead>
<tr>
<th>Species</th>
<th>Singletons&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Two-genes&lt;sup&gt;b&lt;/sup&gt;</th>
<th>≥3&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Total gene families&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Max_gfe&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>6619 (6619)</td>
<td>443 (886)</td>
<td>317 (1817)</td>
<td>7379 (9322)</td>
<td>68</td>
</tr>
<tr>
<td><em>Physcomitrella patens</em></td>
<td>9001 (9001)</td>
<td>700 (1400)</td>
<td>368 (8960)</td>
<td>10 069 (19 351)</td>
<td>4322</td>
</tr>
<tr>
<td><em>Selaginella moellendorfii</em></td>
<td>1356 (1356)</td>
<td>1263 (2526)</td>
<td>600 (4184)</td>
<td>3219 (8066)</td>
<td>129</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>9534 (9534)</td>
<td>775 (1550)</td>
<td>503 (4961)</td>
<td>10 912 (16 045)</td>
<td>562</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>3600 (3600)</td>
<td>394 (788)</td>
<td>265 (3663)</td>
<td>4259 (8051)</td>
<td>1258</td>
</tr>
<tr>
<td><em>Populus trichocarpa</em></td>
<td>4667 (4667)</td>
<td>688 (1376)</td>
<td>343 (2055)</td>
<td>5698 (8098)</td>
<td>82</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em></td>
<td>2279 (2279)</td>
<td>272 (544)</td>
<td>180 (1165)</td>
<td>2731 (3988)</td>
<td>99</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>1274 (1274)</td>
<td>34 (68)</td>
<td>20 (88)</td>
<td>1328 (1430)</td>
<td>17</td>
</tr>
</tbody>
</table>

<sup>a</sup>The number of singletons.

<sup>b</sup>The number of gene families (genes) with two genes.

<sup>c</sup>The number of gene families (genes) contain at least three genes.

<sup>d</sup>The number of total gene families (genes).

<sup>e</sup>The maximum gene family size.

![Figure 3](image_url)
Origination of Arabidopsis thaliana genes

The 26,839 A. thaliana genes within gene families are assigned to different phylogenetic ranks (phylostrata) (Figure 4a and Table 3). More than 70% genes are assigned to Viridiplantae (40.6%) and Embryophyte (30.1%) internodes. In addition, fixation rates of gene or gene family for A. thaliana genes, which is the rate of genes (or gene families) appearance on different phylogenetic ranks, are calculated, respectively (Figure 4b and Table S3). A. thaliana internode has the highest gene fixation rate (110,000 genes/MY), while angiosperm internode has the lowest gene fixation rate (6,267 genes/MY).

For A. thaliana genes assigned to different phylostrata, the proportion of A. thaliana genes annotated to GO terms is 86.7% (23,282/26,839) (Figure 4c and Table S4). Among internodes, the proportion ranges from 92.1% (rosid internode) to 57.2% (A. thaliana internode). For the unknown function process, along the phylostrata, Viridiplantae has the lowest proportion, but A. thaliana has the highest proportion. For most other biological processes, Viridiplantae has the highest proportion; A. thaliana has nearly the lowest proportion (Figure S1 and Table S5). Within A. thaliana internal, the most abundant terms are unknown biological processes, followed by other cellular processes, other metabolic processes, and transcription.

Along phylostratigraphy from Viridiplantae to A. thaliana, the medians of protein sizes of the A. thaliana genes in different phylostra reduced from 413 (Viridiplantae) to 77 (A. thaliana) (Figure 4d and Table S6). While the median of protein size for all A. thaliana proteins within the gene families is 351 (minimum 29, maximum 5337, average 407.8). In addition, among phylostratigraphy, the proportion of expressed genes reduced from 90.1% (Viridiplantae internode) to 37.1% (A. thaliana internode) (Figure 5a and Table S7); and the median of expression levels reduced

Table 3 Number of A. thaliana gene families (and genes) assigned to each phylostratum

<table>
<thead>
<tr>
<th>Phylostratum internode</th>
<th>Genes (%)</th>
<th>Gene families (min, max, avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Viridiplantae</td>
<td>10,897 (40.601)</td>
<td>3007 (1, 193, 3.6239)</td>
</tr>
<tr>
<td>2 Embryophyte</td>
<td>8074 (30.083)</td>
<td>2216 (1, 204, 3.6435)</td>
</tr>
<tr>
<td>3 Tracheophyte</td>
<td>741 (2.761)</td>
<td>290 (1, 28, 2.5552)</td>
</tr>
<tr>
<td>4 Angiosperm</td>
<td>1617 (6.025)</td>
<td>732 (1, 25, 2.2090)</td>
</tr>
<tr>
<td>5 Rosid</td>
<td>1076 (4.009)</td>
<td>484 (1, 106, 2.2231)</td>
</tr>
<tr>
<td>6 Arabidopsis</td>
<td>3004 (11.193)</td>
<td>1666 (1, 134, 1.8031)</td>
</tr>
<tr>
<td>7 A. thaliana</td>
<td>1430 (5.328)</td>
<td>1328 (1, 17, 1.0768)</td>
</tr>
</tbody>
</table>

*The total number of genes assigned to each phylostratum in A. thaliana genome, and their percentage of whole genome given in parentheses.

**The total number of gene families assigned to each phylostratum. Within parentheses, min indicates the minimum gene family size; max indicates the maximum gene family size; avg indicates the average of gene family size.

Figure 4. Phylostratigraphic analysis of A. thaliana genes.

(a) Gene appearance.
(b) Gene fixation rate.
(c) Gene ontology annotation. A dashed line indicates the fraction of A. thaliana genes within gene families that can be annotated to GO biological process terms.
(d) Protein sizes. A dashed line indicates the median protein size of A. thaliana genes within gene families. Significances of the deviations from the median protein size of A. thaliana genes within gene families using Wilcoxon rank sum test are shown in the P-value panel.
from 6.985 (Viridiplantae) to 2.841 (A. thaliana) (Figure 5b and Table S7). In total, 79.2% A. thaliana genes are expressed (21 248/26 839); and the median of expression level is 5.988 (min 1.344, max 14.050, average 5.866).

In order to understand the divergence variation along phylostratigraphy from Viridiplantae to A. thaliana, the divergence between A. thaliana genes and their orthologues in A. lyrata were calculated. The medians of dN/dS increased from 0.1389 (Viridiplantae) to 0.4556 (Arabidopsis), and the medians of dS increased from 0.1407 (Viridiplantae) to 0.1553 (Arabidopsis) (Figure 6 and Table S8). The median of dN/dS for all proteins of A. thaliana within the gene families is 0.1786, and the median of dS is 0.1448.

**DISCUSSION**

**Family size distribution**

Markov Cluster (MCL) algorithm, a method for clustering proteins into groups based on similarity, rigorously tested and validated on a number of databases (Van Dongen, 2000; Enright et al., 2002), has been used for gene family analysis in many taxa, including human (Lander et al., 2001), vertebrate and invertebrate (Prachumwat and Li, 2008), and Drosophila (Drosophila 12 Genomes Consortium, 2007). In this study, using the MCL algorithm, most genes in each genome are grouped into multigene families except C. reinhardtii, in which singletons make up more than half of total genes (59.9%). The predominant number of singletons in C. reinhardtii suggests that the C. reinhardtii genome has a lower gene duplicability than other genomes, possibly due to functional constraints or that some genes of C. reinhardtii are too young to have accumulated duplicates like the case of vertebrate-specific genes (Prachumwat and Li, 2008), or there are probably some annotation artifacts in C. reinhardtii. However, only 2.4% of singletons in C. reinhardtii can be annotated to a GO term, thus functional constraint is unlikely the case here. For S. moellendorffii genome, in order to sufficiently

---

**Figure 5.** Expression variation of A. thaliana genes among phylostrata. (a) Expressed genes. A dashed line indicates the expression fraction level of A. thaliana genes within gene families. (b) Expression level. A dashed line indicates the median expression level of A. thaliana genes within gene families using Wilcoxon rank sum test are shown in the P-value panel.

**Figure 6.** Divergence variation of A. thaliana genes among phylostrata, estimated between A. thaliana and A. lyrata. (a) dN/dS. (b) dS. A dashed line indicates the median dN/dS or dS of A. thaliana genes within gene families. Significances of the deviations from the median dN/dS and dS of A. thaliana genes within gene families using Wilcoxon rank sum test are shown in the P-value panel, respectively.

© 2012 The Author

represent the S. moellendorffii genome, all gene models instead of the data set with only one of the two haplotypes are used in this study, thus in the present study for this genome, genes in two-gene families are inflated, and genes in singletons are reduced at the same time accordingly.

In total, 2745 gene families are shared among all the eight species, which must be the ‘core’ proteome of green plants. This number is consistent with another recently published study (Van Bel et al., 2012), in which 2928 core gene families are found via a parsimony method from 25 genomes, although these two studies used different strategy to call ‘core’ proteome. Among the 2928 gene families found in the previous study, 2207 (75%) are found in the present study, when match the representative A. thaliana genes of each core gene family reported in the previous study (Van Bel et al., 2012) to the present study. Interestingly, two recent studies have indicated across green plants, there is a relatively large and well conserved core set of protein domains as well (Kersting et al., 2012; Zhang et al., 2012). In addition, phylogenetic comparative analysis indicated the expansion of transcription-associated proteins is positively correlated with the increase in morphological complexity (Lang et al., 2010).

Orphan genes are protein-coding genes that do not have recognizable homologues in other species. Various factors can lead to rise of the orphan genes (Domazet-Loso and Tautz, 2003; Demuth et al., 2006; Hahn et al., 2007). Theoretically, all of these factors can be classified into six classes: (i) true loss of all functional genes; (ii) the de novo evolution of new genes; (iii) rapid protein evolution in previously existing genes so that they are no longer identified as being part of a pre-existing family; (iv) horizontal gene transfer; (v) annotation artifacts; and (vi) artifacts of the clustering process. Such genes might be involved frequently in specific ecological adaptations or be the raw fuel for micro-evolutionary divergence (Carroll et al., 2001; Domazet-Loso and Tautz, 2003; Yang et al., 2009). Not surprisingly, in C. reinhardtii and P. patens, more than half of the genes within these two genomes are orphan genes, given some genes in C. reinhardtii and P. patens are probably not required for vascular plants. Recently, in Drosophila, it has been demonstrated that orphan genes can quickly become essential and play an essential role in development (Chen et al., 2010).

Similar gene gain/loss rate between green plants and animals

The likelihood method implemented in CAFE version 2.0 (De Bie et al., 2006) is a robust pipeline to study gene family evolution, and has been used widely in genome analyses, including mammalian genomes (Demuth et al., 2006), Drosophila genomes (Hahn et al., 2007), and rhesus macaque genome (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007). Of the 3296 families used for the likelihood analysis, along the phylogenetic ranks, gene families in each lineage expands to different extent except for C. reinhardtii, grass, and Arabidopsis. Consistent with the contraction of A. thaliana at the whole-genome level, all nine rapidly evolving gene families contract on the terminal branch to A. thaliana. In Drosophila, it has been demonstrated that some gene families (defense response, proteolysis, and trypsin activity) are evolving rapidly both in copy number and sequence level (Drosophila 12 Genomes Consortium, 2007; Hahn et al., 2007). However, in Arabidopsis, there is not any correlation between the copy number variation and sequence divergence for the genes of rapidly evolving gene families, which suggested that the relationship between copy number variation and sequence divergence is much more complex.

In the present study, the rate of gene gains and losses is about 0.001379 (gains and losses/gene/million years). This rate is a little lower than some previous estimates. For instance, the average duplication rate of a eukaryotic gene was estimated to be on the order of 0.01/gene/million years and 0.002 for A. thaliana (Lynch and Conery, 2000, 2003). However, this rate is similar to that in Drosophila – 0.0012 (Hahn et al., 2007) and Mammalian – 0.0016 (Demuth et al., 2008). As this result is consistent with the average rate of duplication of a eukaryotic gene, and also eight representative genomes of green plants are used, thus the estimated rate is reliable for A. thaliana relatives and green plants as well.

One important factor that affects the estimation of gene gains or losses rate in certain lineages is polyploidy/whole-genome duplication. Especially for most green plants, including a few angiosperm species studied here, are polyploid, and whole-genome duplication is normally followed by gene loss and diploidization (Simillion et al., 2002; Blanc and Wolfe, 2004; Adams and Wendel, 2005; Doyle et al., 2008; Soltis et al., 2009; Koh et al., 2010; Jiao et al., 2011; Proost et al., 2011). Therefore, at the green plants level, whole-genome duplication likely raises the rate of gene gains and losses. However, with the availability of more genomes from various plant lineages, it would be interesting to address how these polyploid events affects gene gains or losses rate in some specific plant lineages that experienced whole-genome duplication.

Origin of Arabidopsis thaliana genes

The gene emergence approach has been used to study the evolutionary emergence of whole-genome genes in individual species or lineage at different phylogenetic levels (Domazet-Lošo et al., 2007; Domazet-Lošo and Tautz, 2008, 2010a,b; Prachumwat and Li, 2008; Zhang et al., 2010a, 2011; Banks et al., 2011). The significant difference of gene emergence on the phylostratigraphic map supports the assumption that some of these genes have retained a signal of their evolutionary history. In this way, the
phylostratigraphy can imply a picture of evolutionary history using extant species. In this study, more than 70% of *A. thaliana* genes are traced back to 450 million years ago. Interestingly, the *A. thaliana* internode also has the highest gene fixation rate, which might be due to high gene origination rate, or more complete annotation since much more expression data can be used for annotation. Transposition in certain families is a possible cause of the high gene origination rate. A recent study indicates that specific gene families can transpose at specific points in evolutionary time, especially after whole-genome duplication events in the Brassicales (Woodhouse et al., 2011).

The fraction of expressed genes, expression level, and protein size reduced along the phylostratigraphy from Viridiplantae to *A. thaliana*. In yeast, there is an inverse relationship between protein size and protein expression level (Warringer and Blomberg, 2006). Conversely, in *A. thaliana* the oldest genes are much longer and also highly expressed. In addition, the comparison between *A. thaliana* and *A. lyrata* indicates transposable elements and small RNAs contribute to gene expression divergence (Hollister et al., 2011). Most probably, different lineage or species has different regulatory mechanism. For example, even comparing to its sister species *A. lyrata*, *A. thaliana* has much higher intron loss rate (Fawcett et al., 2012) and deletion rate, and also tend to delete longer fragments (Hu et al., 2011).

Interestingly, Viridiplantae internode has the lowest dN/dS, while Arabidopsis internode has the highest. This probably suggests that older genes have been under stronger purifying selection, since older genes are mostly fundamental genes for green plants to survive. Accordingly, the older genes are more conserved with a slow evolution rate (synonymous substitution rate). Evolution rate can be shaped by various factors (Albà and Castresana, 2005; Lar racuente et al., 2008; Singh et al., 2009; Liao et al., 2010; Gaut et al., 2011; Yang and Gaut, 2011). Recent comparisons between two available Arabidopsis genomes provide more robust results. The rates of protein evolution are correlated mainly with expression level and specificity (Slotte et al., 2011); and rates at synonymous sites vary on genomic scale and might be a function of recombination rate and evolution rates at non-synonymous sites, strongly correlated with expression pattern and whether a gene is retained after a whole-genome duplication or not (Yang and Gaut, 2011). In addition, 1939 annotated protein-coding genes with little evidence of expression in *A. thaliana* are shorter and evolved with a two-fold higher rate of non-synonymous substitutions (Yang et al., 2011). This finding is consistent with what is found here, that younger genes are shorter, higher in the non-synonymous substitution rate, and lowly expressed.

Furthermore, it would be interesting to study dN/dS ratio in other species pairs, similar comparisons had been made between *A. thaliana* and popular and between two conifers (Picea sitchensis and Pinus taeda) (Buschiazzo et al., 2012), which indicates low rates of diversification in conifers but much higher dN/dS, possibly due to positive selection. It would also be interesting to do a lineage-specific study about dN/dS ratio, which could be informative for understanding the mechanism of lineage-specific evolutionary events.

Although many other plant species’ genome sequences are available now, the eight genomes here are good representative of green plants with clear phylogeny and divergence times estimated, especially assembled with longer and high quality reads at high coverage, which are believed to be very important to address evolutionary questions at whole-genome level (Demuth and Hahn, 2009; Schneider et al., 2009; Milinkovitch et al., 2010). Therefore, the eight representative green plants’ genomes are sufficient to address those interesting questions at whole-genome level. The comparison of the eight genomes, revealed the evolutionary pattern of green plants’ genes and gene families at an unprecedented resolution. The rate of gene gains and losses is similar to that in animals, some gene families are rapidly evolved and the 2745 gene families are ‘core’ proteome of green plants. Most genes of *A. thaliana* are biased toward ancient genes, and genes with early origination are under stronger purifying selection and more conserved; and gene families of Arabidopsis lineage contract a lot, especially *A. thaliana*, which could be correlated with the shaping of the difference of morphology, physiology, and metabolism among species.

**EXPERIMENTAL PROCEDURES**

**Genome data sets**

Protein sequences from eight genomes of green plants were analysed, including *A. thaliana* (TAIR8) (27 025 proteins) (The Arabidopsis Genome Initiative, 2000), *A. lyrata* (JGI-v1.0, FM6) (32 670 proteins) (Hu et al., 2011), *Populus* (*Populus trichocarpa, JGI-v1.1*) (45 555 proteins) (Tuskan et al., 2006), *rice* (*Oryza sativa, v5.0*) (56 278 proteins) (International Rice Genome Sequencing Project, 2005; Ouyang et al., 2007), *Sorghum* (*Sorghum bicolor, JGI-v1.0*) (35 899 proteins) (Paterson et al., 2009), *Selaginella moellendorfii* (JGI-v1.0) (34 697 proteins) (Banks et al., 2011), *Physcomitrella patens* ssp *patens* (JGI-v1.1) (35 938 proteins) (Rensing et al., 2007), and *Chlamydomonas reinhardtii* (JGI-v3.1) (15 256 proteins) (Merchant et al., 2007). The phylogeny and divergence times of the eight genomes were clarified clearly (Kellogg, 2001; Wikström et al., 2001; Gaut, 2002; Wellman et al., 2003; Magallon and Sanderson, 2005; Zimmer et al., 2007; Beilstein et al., 2010; Lang et al., 2010; Guo et al., 2011b) (Figure 1 and Table S9 for the detail of the divergence time at each node). In addition, the sampled genomes are representative, flanking the most important events in the green plants leading to *A. thaliana*. In order to get rid of possible transposable element sequences within the proteins data sets, all proteins were searched against the repeat library database (repeat masker database version 20090604) using tblastn (*E* value < 1 × 10⁻⁵), and length of hit must be at least half of the query sequence. The putative transposable elements are deleted from the dataset in the following analyses.
Homology searches and gene family construction

Markov Cluster Algorithm (mcl-06-058 package; http://micans.org/mcl/src/) was used to cluster all the eight genomes’ protein sequences into gene families with default parameters (–I 1, –S 6). There are two steps for clustering analysis: (i) blastp was used to match each protein against all other proteins (E value <1 × 10⁻⁵) to produce a raw NCBI BLAST output; and (ii) MCL was used to build Markov Matrix using the parsed similarity data converted from raw NCBI BLAST output, and then generated the final gene families. The clustering results were imported into MySQL database for the following analyses. Genes (gene families) that do not have recognizable homologues in other species are called as orphan genes (gene families).

Gene family size analysis

Genes were classified into one of the following three groups by its family size in the clustering according to Prachumwat and Li (2008): (i) a singleton is a single-copy gene; (ii) a two-gene family has two copies; and (iii) a multigene family has ≥3 copies.

Likelihood analysis of gene gain and loss

CAFE version 2.0 (De Bie et al., 2006), a tool for the statistical analysis of the evolution of the gene family size, was used to study the evolution of gene families. One can estimate gene gain and loss rate among these plant genomes; calculate the average expansion or contraction per gene family on each branch; and identify the gene family that has larger-than-expected contractions or expansions. The below input files and parameters are used. First, the input tree structure file, which is a Newick description of a rooted and bifurcating phylogenetic tree (including branch lengths in units of time) (Figure 1). Second, the data file consists of gene family sizes of each gene family across eight species. Third, λ, the probability of both gene gains and losses per gene per unit time in the phylogeny (CAFE assumes that gene birth and death are equally probable), was calculated by CAFE given the gene families in the data file (the estimated λ is the globally most probable value across all families in the input file or the different lineages’ λ with related options of –f). Here I calculated the λ across all the eight species. Fourth, two other parameters were set as P-value threshold (0.01) and number of random samples used the default value (1000). Fifth, the default Viterbi method was employed to identify the branch with the most unlikely amount of change. For gene families that are evolving at rates of gains and losses significantly higher than the genome-wide average with a P-value below P < 0.0001 (Hahn et al., 2005), Viterbi method was used to identify the branch that is the most likely cause of deviation from the random model significant at P < 0.005 (Hahn et al., 2007).

Phylostratigraphic analyses of Arabidopsis thaliana genes

Phylostratigraphic analysis was performed according to previous study (Domazet-Loso et al., 2007). All genes of A. thaliana were assigned into seven phylogenetic ranks (phylostra) of different ages according to the emergence in the reference genomes on the phylogeny based on the similarity clustering via MCL. The seven gene groups were: (1) A. thaliana, genes that arose along the A. thaliana; (2) Arabidopsis, genes were present in the MRCA of Arabidopsis genus and P. trichocarpa; (3) rosid, genes were present in the MRCA of Arabidopsis genus and S. moellendorfii; (4) Tracheophyte, genes were present in the MRCA of angiosperm and S. bicolor and O. sativa; (5) C20/Phylostratum, genes were present in the MRCA of Tracheophyte and C. reinhardtii. Genes (or gene families) fixation rates were calculated according to the formula, \( r = N/T \); r is the number of genes (or gene families) per million years; N is the number of genes (or gene families) in the phylostratum; T is duration of the interval (million years).

Divergence and polymorphism

There are several steps to estimate the divergence of orthologs: (i) the orthologous gene pairs between A. thaliana and A. lyrata were extracted from a whole-genome alignment analysis (Hu et al., 2011); (ii) a codon-based alignment of orthologous coding sequences was performed with PAL2NAL (Suyama et al., 2006); (iii) dS, the number of synonymous substitutions per synonymous site, and dN/dS ratios, the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site, were calculated for orthologue pairs using the codeml program implemented in phylogenetic analysis by maximum likelihood version 3.15 package (Yang, 1997).

GO annotation and gene expression

Gene ontology annotation was performed using GO implemented at TAIR (October 2009) (http://www.arabidopsis.org/tools/bulk/go/index.jsp) for A. thaliana (Gene Ontology Consortium, 2004); for other species, blast2go was used to do the GO annotation (Conesa et al., 2005). GO Biological Processes were used to estimate the abundance of GO terms. Gene expression data from AtGenExpress Arabidopsis expression atlas (Schmidt et al., 2005), which was generated on the widely available Affymetrix ATH1 array platform, were averaged across all tissues as an estimate of expression level of A. thaliana genes.

ACKNOWLEDGEMENTS

I thank Matthew W. Hahn for suggestions on CAFE analysis; Tomislav Domazet-Loso for suggestions on phylostratigraphy analysis; Manyuan Long, Yong Zhang, Jie Guo, and anonymous reviewers for comments about this work; and Tina Hu for proofreading and improving the manuscript. Especially, I am very grateful to Detlef Weigel and Song Ge for their long-term support and also valuable comments on this work. This work was supported by 100 talents program of Chinese Academy of Sciences and the Max Planck Society.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Data S1. The gene family clustering among all proteins of the eight genomes.

Figure S1. The fraction of A. thaliana genes in each phylostratum.

Table S1. GO annotation of A. thaliana genes in each phylostratum.

Table S2. Functional annotation of 41 rapidly evolved gene families.

Table S3. Fixation rate of A. thaliana genes or gene families in each phylostratum (genes/duration of interval or gene families/duration of interval).

Table S4. The fraction of A. thaliana genes in each phylostratum that can be annotated to GO biological process.

Table S5. GO annotation of A. thaliana genes in each phylostratum in terms of proportion of major biological processes (%).

Table S6. Protein sizes of A. thaliana genes in each phylostratum.

Table S7. Expression of A. thaliana genes in each phylostratum.
Table S8. Divergences between A. thaliana genes of each phyllostratum and their orthologous genes in A. lyrata.
Table S9. Divergence time of each node on the phylogeny of green plants.

REFERENCES

Buschiazzo, E., Ritland, C., Bohlmann, J. and Ritland, K. (2010a) A phylogenetically based transcrip-
tional tool for the study of gene family evolution. Trends Genet. 26, 476–481.


Kersting, A.R., Bornberg-Bauer, E., Moore, A.D. and Grath, S. (2010a) A phylogenetically based transcrip-


Evolution of gene families in green plants


