Selection for Higher Gene Copy Number after Different Types of Plant Gene Duplications

Corey M. Hudson1,*, Emily E. Puckett2, Michâel Bekaert3, J. Chris Pires1,2, and Gavin C. Conant1,3

1Informatics Institute, University of Missouri
2Division of Biological Sciences, University of Missouri
3Division of Animal Sciences, University of Missouri

*Corresponding author: E-mail: cmhkbd@mail.missouri.edu.

Abstract

The evolutionary origins of the multitude of duplicate genes in the plant genomes are still incompletely understood. To gain an appreciation of the potential selective forces acting on these duplicates, we phylogenetically inferred the set of metabolic gene families from 10 flowering plant (angiosperm) genomes. We then compared the metabolic fluxes for these families, predicted using the Arabidopsis thaliana and Sorghum bicolor metabolic networks, with the families’ duplication propensities. For duplications produced by both small scale (small-scale duplications) and genome duplication (whole-genome duplications), there is a significant association between the flux and the tendency to duplicate. Following this global analysis, we made a more fine-scale study of the selective constraints observed on plant sodium and phosphate transporters. We find that the different duplication mechanisms give rise to differing selective constraints. However, the exact nature of this pattern varies between the gene families, and we argue that the duplication mechanism alone does not define a duplicated gene’s subsequent evolutionary trajectory. Collectively, our results argue for the interplay of history, function, and selection in shaping the duplicate gene evolution in plants.

Key words: dosage selection, genome duplication, gene duplication.

Introduction

The contribution of gene duplication to evolution has long been a topic of interest (Taylor and Raes 2004), but in the last 10 years there has been a resurgence of interest in the varied fates of such duplications (Zhang et al. 2002; Kondrashov and Koonin 2004; Adams and Wendel 2005; Aury et al. 2006; Rodriguez et al. 2007; Barker et al. 2008; Liang et al. 2008; Ha et al. 2009; Innan and Kondrashov 2010; Ramsey 2011). Among those fates, the important roles played by genetic drift and simple changes in the gene “dosage” are increasingly appreciated. In several contributions, Lynch et al. have argued that the relatively small population sizes of multicellular eukaryotes could result in the fixation of many gene duplications through nonadaptive processes (Force et al. 1999; Lynch and Conery 2003; Lynch 2007). These processes, of course, still occur under the overall umbrella of natural selection. For instance, selection may act on gene dosage in one of two ways. First and most obviously, duplication of a gene may increase the rate of transcription and hence the translation of the encoded protein, increasing its abundance. We have previously referred to this possibility as a selection on “absolute” dosage (Bekaert et al. 2011). If a higher protein expression is selectively beneficial, we expect copy number polymorphisms will be fixed (Blanc and Wolfe 2004a; Kondrashov and Kondrashov 2006). The second possibility is that of a selection on the “relative dosage,” where an event affecting one of several genes that have coevolved together (i.e., a single gene duplication or differential paralog loss after polyploidy) introduces selective costs. This concept is known as the “dosage balance hypothesis” (Freeling 2009) and has been explored by a number of authors (Papp et al. 2003; Freeling and Thomas 2006; Birchler and Veitia 2007; Edger and Pires 2009). Here, we focus on the role of absolute dosage selection in determining duplicate fates.

As the first complete genome sequences became available, their patterns of gene duplication were explored to understand, among other questions, the role of natural selection in duplicate gene fixation (Lynch and Conery 2000; Gu et al. 2002; Wagner 2002; Gu et al. 2003). Those duplications...
had multiple origins, including whole-genome duplications (WGDs or polyploidy), as well as segmental, tandem, and retro-duplications (referred to here collectively as small-scale duplications or SSDs; Cannon et al. 2004; Thomas et al. 2006; Freeling 2009). The preponderance of polyploids among angiosperms (Wendel 2000) has led plant biologists to focus on understanding the patterns of the duplicate gene loss and the retention following WGD events (Bowers et al. 2003; Blanc and Wolfe 2004a; Blanc and Wolfe 2004b; De Bodt et al. 2005; Maere et al. 2005; Pfeil et al. 2005; Sterck et al. 2005; Cui et al. 2006; Freeling and Thomas 2006; Paterson et al. 2006; Schranz and Mitchell-Olds 2006; Town et al. 2006; Tuskan et al. 2006; Tang, Wang, et al. 2008; Barker et al. 2009; Edgar and Pires 2009; Soltis et al. 2009; Wood et al. 2009; Duarte et al. 2010; Coate et al. 2011; Jiao et al. 2011; Schnable et al. 2011). In this work, we consider gene families with members derived from both SSD and WGD. These families are inferred from 10 angiosperm genomes: seven dicots (Arabidopsis, papaya, soybean, Medicago truncata, poplar, peach, and grape) and three monocots (Brachypodium distachyon, rice, and sorghum).

Of course, the taxa examined have a long history of polyploidy. Within the eudicots, the oldest genome duplication event, γ, was an ancient hexaploidy that characterizes the Rosidae (sensu Soltis et al. 2011), if not the core eudicots (Gunneridae sensu Jaillon et al. 2007; Lyons, Pedersen, Kane, Alam, et al. 2008; Lyons, Pedersen, Kane, Freeling, et al. 2008; Ming et al. 2008; Freeing 2009; Argout et al. 2011; Jiao et al. 2011; Shulaev et al. 2011; Soltis et al. 2011). Comparative genomics suggest that the lineage leading to poplar (Populus trichocarpa) underwent an additional WGD event, whereas that of the thale cress (Arabidopsis thaliana) had two: β and α. That these two duplications are independent is suggested by their absence in both grape (Vitis vinifera) and papaya (Carica papaya; fig. 1; Jaillon et al. 2007; Ming et al. 2008; Tang, Wang, et al. 2008; Freeling 2009). Analysis of the nonsynonymous substitution rates in the soybean (Glycine max) genome has revealed two WGD events post-WGD-γ: one shared with peanut (Arachis hypogaea), a basal legume, and a more recent soybean-specific duplication (Bertoli et al. 2009; Schmutz et al. 2010). The 3:1 ratio of grape to rice (Oryza sativa) genomic segments suggests that the γ paleohexaploid is dicot-specific (Jaillon et al. 2007). However, cereal monocots also have a WGD event, ρ, basal to their radiation (Paterson et al. 2004); rice, sorghum (Sorghum bicolor) and purple false brome (B. distachyon) show no evidence of further WGD events (Throude et al. 2009; Vogel et al. 2010).

There is mounting evidence that the gene duplications created by WGD and by SSD differ in their ultimate fates (Seoighe and Wolfe 1999; Papp et al. 2003; Blanc and Wolfe 2004a, 2004b; Cannon et al. 2004; Aury et al. 2006; Thomas et al. 2006; Hakes et al. 2007; Conant and Wolfe 2008; Freeling 2008; Edgar and Pires 2009; Freeling 2009; Coate et al. 2011). To cite just one example (relevant to this work), Maere et al. (2005) found that the ion transporters were overretained after WGD but underretained following SSD. The study of Arabidopsis WGDs by Blanc and Wolfe (2004a) reached similar conclusions but also found that genes involved in phosphate metabolism were significantly overretained following the recent WGD—α.

We are interested in whether the dosage effects are a strong predictor of duplicate retention, and here, we have taken both a “high level” phylogenomic approach and a “low-level” single-gene approach to look for evidence of such selection. Our first analysis extends our previous work in Arabidopsis, where we found an association between metabolic flux and some, but not all, of the Arabidopsis WGDs (Bekaert et al. 2011). Specifically, we hypothesize that genes in families with high flux will be, on average, over duplicated. Given that we have previously found significant differences in duplication propensity between cellular compartments (Bekaert et al. 2011; Hudson and Conant 2011), we also test for a relationship of duplicability and compartment. Additionally, we hypothesized that the WGD-produced and SSD-produced gene duplications will differ in their postduplication selective constraints. We evaluate this by narrowing our focus to a group of ion transporters. Such transporters have been found to have an outsized influence on the metabolic flux (Kacser and Burns 1981 notwithstanding; Brown et al. 1998; Pritchard and Kell 2002). Furthermore, their evolutionary behavior is distinct from other metabolic genes following both SSD and WGD (Lin and Li 2010; Bekaert and Conant 2011). Given the complexity of plant genome evolution, limiting our analysis to single gene families also has the advantage of allowing us to carefully distinguish WGD from SSD.

Materials and Methods

Estimation of Metabolic Flux

As previously described (Bekaert et al. 2011), we used the Systems Biology Research Tool v2.0.0 (Wright and Wagner 2008) to perform flux-balance analysis on the A. thaliana and S. bicolor metabolic networks (de Oliveira Dal’Molin et al. 2010a, 2010b). We estimated the maximal biomass production possible under photosynthetic conditions (a fixed level of photon import allowed, sugar imports forbidden) for both A. thaliana and S. bicolor networks (Pearson’s correlation of flux r = 0.638, P < 10−15) and nonphotosynthetic conditions (photon import forbidden, fixed sugar imports allowed) for the A. thaliana network. Flux-balance analysis was also run by limiting the biomass and maximizing either photon import or sugar import (Bekaert et al. 2011), the results were similar and qualitatively the same. Because the distinctions between the two networks are in the photosynthetic reactions and because the sorghum metabolic network is derived from the Arabidopsis one, inclusion of the sorghum root data would be less informative and is
hence omitted. In each case, we also made every possible reaction knockout whereby a given reaction’s flux is constrained to zero and the remainder of the network is reoptimized. After knockout, all fluxes were normalized by the value of the biomass flux. Then, for each reaction, we selected the observed maximum flux, across all conditions.
By doing so, we find what is essentially an upper bound on the flux of each reaction. It would obviously be desirable to also estimate the sensitivity of the network to changes in flux through each reaction. However, we do not have kinetic data for the entire network and these values cannot be estimated with flux-balance analysis. Instead, we compared this maximal flux to the duplication status of each reaction node. In cases where there was more than one flux value associated with a gene family, all possible flux values for that family were used in our association analyses, meaning that large gene families will not tend to be biased toward high flux because they encompass more reactions.

Gene Family Identification

We used the list of *A. thaliana* enzymes from the de Oliveira Dal’Molin et al. (2010a) metabolic network to identify enzyme gene families in the genomes of 10 flowering plants (fig. 1; *A. thaliana*, *B. distachyon*, *C. papaya*, *G. max*, *M. trunculata*, *O. sativa*, *P. trichocarpa*, *Prunus persica*, *S. bicolor*, and *V. vinifera*). The Arabidopsis Genome Initiative 2000; Young et al. 2005; Ouyang et al. 2006; Tuskan et al. 2006; Jaillon et al. 2007; Ming et al. 2008; Paterson et al. 2009; Schmutz et al. 2010; The International Brachypodium Initiative 2010; Jung et al. 2009; Vogel et al. 2010). Homologous relationships were inferred using GenomeHistory (Conant and Wagner 2002), which calculated the nonsynonymous substitution rate ($K_a$) for all gene pairs with BLAST scores lower than 0.0001. Gene families were identified by single-linkage clustering with a cutoff in nonsynonymous divergence of $K_a \leq 0.20$ for *A. thaliana*/A. thaliana comparisons and $K_a \leq 0.30$ for all other comparisons (Powell et al. 2008). Gene pairs with $K_a$ values below these thresholds were treated as nodes connected by an edge in the provisional gene family networks. These $K_a$ parameters were selected after analyzing the results of using different $K_a$ thresholds. For each threshold, we iteratively removed single edges from the provisional gene families. The chosen $K_a$ thresholds were the largest values that did not cause a noticeable change in the constituency of the provisional gene families when any single edge was removed (data not shown). Families with fewer than four member genes were excluded. We used these gene families to associate a gene tree with each gene in the *S. bicolor* metabolic network.

Of the 138 pathways involved in Arabidopsis central metabolism, six contain no enzymes in the gene families we analyzed. Two of them (1,4-dichlorobenzene degradation and C21-steroid hormone metabolism) contain enzymes only present in Arabidopsis. The transport of α-d-glucose from the cytoplasm to the external cellular component contains a gene (ATSG14570) that our pipeline split into gene families that were too small to analyze phylogenomically.

Phylogenomics of Gene Families

Multiple sequence alignments of the protein sequences for each gene family were computed with MUSCLE v3.6 (Edgar 2004) using default parameters. Codon alignments were deduced from those alignments having 50 or more amino acids. We then inferred maximum likelihood gene trees using RAxML v7.0.4 (Stamatakis et al. 2008) with a general time-reversible model and discrete approximation of the gamma distribution (GTR + Γ). Confidence values were assigned to the gene trees from 100 bootstrap replicates. A relatively limited number of replicates were computed because we only wished to use these bootstrap statistics to identify nodes in the phylogeny with low support (<65%) prior to gene tree/species tree reconciliation. We thus recounted all inferred gene trees with the species tree in figure 1 (Moore et al. 2007; Wang et al. 2009). To do so, we used NOTUNG v2.6 (Chen et al. 2000) to infer the most parsimonious pattern of the gene duplication and loss. Gene tree nodes with less than 65% bootstrap support were treated as polytomies and allowed to rearrange in order to minimize the number of duplications and/or losses (in practice, choosing support value thresholds between 50% and 80% produced similar results; data not shown). Using these parsimony reconstructions, we calculated the number of duplications (and number of duplications per species) for each gene tree.

Manual Annotation of Transporter Gene Trees

Coding sequences for the nine annotated PHT1s, one PHT2, three PHT3s, six PHT4s, and eight NHXs of *A. thaliana* were downloaded from TAIR (Swarbreck et al. 2008). A BLASTP search of the A. thaliana genome with these 19 and 8 sequences identified no further phosphate or sodium transporters in the genome. We then used BLASTP to search for ion transporter homologs in the genomes of papaya and poplar. We retained genes with BLAST E-values less than $10^{-20}$ as putative members of a given transporter family. Our homology estimation procedure always placed genes from *C. papaya* and *P. trichocarpa* into only a single *A. thaliana* transporter family. Gene trees were constructed as detailed above. In the case of the NHXs, one *A. thaliana* gene (AT2g01980) aligned poorly with the other NHXs and was excluded from the alignment and gene tree.

We manually assigned nodes in these phylogenies as either speciation or duplication events (fig. 2). Nodes connecting genes from the same species were labeled as duplication events where nodes connecting genes from different species were labeled speciation events. Because we were working with only a handful of genes, it was possible to make a more accurate distinction between SSD and WGD genes.
for these transporters than was possible for the genome-scale analyses. Thus, whole-genome duplicates were inferred in cases where the paralogs fit into distinct paralogous synten blocks from the Plant Whole Genome Duplication Database (PGDD; Tang, Bowers, et al. 2008). Nodes connecting gene paralogs that could not be assigned using the PGDD were inferred to be SSDs (fig. 2). These manual duplication or speciation designations agreed with the automatic assessments of NOTUNG.

Selective Constraint Following Speciation and Duplication Events in Five Families of Ion Transporters

The selective constraint (ratio of nonsynonymous substitutions to synonymous substitutions, i.e., $K_a/K_s$), for each gene tree was estimated by maximum likelihood under the MG/GY94 codon model (Goldman and Yang 1994; Muse and Gaut 1994); for details, see Conant et al. (2007). We tested three nested models of evolution: requiring all branches to have the same value of $K_a/K_s$ (R_Null), allowing different values of $K_a/K_s$ for branches following a speciation node from those following a duplication node (R_Dupl), and a model with differing values of $K_a/K_s$ for branches following speciation, whole genome, and small-scale duplications (R_WGD). We compared these three models with nested likelihood ratio tests and evaluated statistical significance using the $\chi^2$ distribution, knowing that R_WGD has one more free parameter than R_Dupl, which in turn has one more parameter than R_Null.

Analysis of Constraints by Gene Ontology Slim Annotation

Gene ontology slim (GO Slim) annotations were obtained for each A. thaliana gene from TAIR. GO Slim categories were further condensed (supplementary table 1, Supplementary Material online) and transferred to our gene families. Spearman’s rank correlations between the flux and

---

**Fig. 2.**—Ion transporter gene trees used in this study. Branches demarcating speciation events are colored orange, whole-genome duplication events green, and non-WGDs purple. (a) High-affinity phosphate transporters AtPHT1; 1-AtPHT1; 9 with 16 P. trichocarpa and 7 C. papaya homologs. (b) Low-affinity phosphate transporters AtPHT2; 1 with 2 poplar and 1 papaya homologs. (c) Mitochondrial phosphate transporters AtPHT3; 1-ATPHT3; 3 with 6 poplar and 1 papaya homologs. (d) Chloroplast phosphate transporters AtPHT4; 1-AtPHT4; 6 with 10 poplar and 6 papaya homologs. (e) Sodium ion transporters AtNHX1-AtNHX6, AtNHX8 with 6 poplar and 4 papaya homologs.
number of duplications in each gene family were calculated in SAS (v9.2.1, Cary, NC) for all the cellular compartments and functions. Note that gene families could appear in more than one compartment or functional group. We applied a Bonferroni multiple-test correction equal to the number of either compartments or functional groups analyzed, resulting in the respective values of α, 0.0055 and 0.0042.

We also used the Wilcoxon rank test (SAS v9.2.1) to ask if the number of duplications per gene family differed for each cellular compartment or function as compared with the reminder of the genome. We used the same Bonferroni multiple-test corrections as previously.

Results

Computing Gene Families and Flux Values

We estimated the flux through each biochemical reaction in the Arabidopsis and sorghum metabolic networks using flux-balance analysis (Orth et al. 2010), maximizing the production of new cell mass for a fixed input of either light energy in both Arabidopsis and sorghum (in photosynthetic tissues) or carbohydrates for Arabidopsis (in nonphotosynthetic tissues, see Materials and Methods). We included the sorghum network to be sure that the differences in C3 and C4 photosynthesis were not greatly biasing our results.

Maximal flux values ranged from 0 to 3865120 (arbitrary flux-balance units) in the Arabidopsis leaf, from 0 to 6156740 in the Arabidopsis root, and from 0 to 2560860 in sorghum, when the biomass production is maximized and scaled to 1000 units. We then coupled those data to a set of cross-genome gene families identified from the 10 plant genomes (Materials and Methods). We included the sorghum network to be sure that the differences in C3 and C4 photosynthesis were not greatly biasing our results.

Table 1

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>All Flux Values</th>
<th>Excluding Null-Flux</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρ</td>
<td>p</td>
</tr>
<tr>
<td>All conditions</td>
<td>0.245</td>
<td>&lt;10⁻¹⁵</td>
</tr>
<tr>
<td>C3 leaves</td>
<td>0.218</td>
<td>&lt;10⁻¹⁵</td>
</tr>
<tr>
<td>C4 leaves</td>
<td>0.176</td>
<td>&lt;10⁻⁸</td>
</tr>
<tr>
<td>Roots</td>
<td>0.223</td>
<td>&lt;10⁻¹⁵</td>
</tr>
<tr>
<td></td>
<td>0.227</td>
<td>&lt;10⁻¹⁵</td>
</tr>
<tr>
<td>C3 leaves</td>
<td>0.203</td>
<td>&lt;10⁻¹⁵</td>
</tr>
<tr>
<td>C4 leaves</td>
<td>0.163</td>
<td>&lt;10⁻⁷</td>
</tr>
<tr>
<td>Roots</td>
<td>0.211</td>
<td>&lt;10⁻¹⁵</td>
</tr>
</tbody>
</table>

Flux values equaling 0 can have confounding biological and computational meanings.

Association of Flux and Duplication is neither Taxa nor Duplication-Mechanism Specific

As described, these species share a history of WGD (fig. 1). We summed the number of duplications on each branch in figure 1, separating those with lineage-specific WGDs from those without. Duplications in both groups are significantly and positively correlated with maximum flux (WGD: r = 0.111, P < 0.05; SSD: r = 0.094, P < 0.05). Of course, the branches containing WGDs will also have some background level of SSD, meaning that the duplications on these branches will not be exclusively due to WGD. However, the similarity in correlations seen between the two types of branch suggests that a more careful accounting of duplicates is unlikely to yield different results. Similarly, we found significant positive associations of duplication and flux for the monocot subtree as well as the eudicot tree with A. thaliana removed (P < 0.05). The similarity of the results for these subtrees implies that our results are not specific to Arabidopsis, even though one of the primary metabolic networks used is from this organism. Among the terminal nodes with rice and soybean show significant associations of flux and duplication after a Bonferroni multiple-testing correction (P < 0.00256). Unfortunately, for the remainder of the tip taxa, it is difficult to distinguish between the lack of an association and the lack of sufficient numbers of duplicates to discern if that association might exist. Similarly, the flux values inferred from the sorghum C4 leaves show a mixed pattern of associations and lack thereof depending on the precise data set used (0.1689 ≤ P ≤ 0.9653).
Table 2

<table>
<thead>
<tr>
<th>Cellular Compartment</th>
<th>n</th>
<th>r'</th>
<th>p</th>
<th>Z^d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>56</td>
<td>0.320</td>
<td>0.016</td>
<td>3.481</td>
<td>0.0005</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>74</td>
<td>0.133</td>
<td>0.258</td>
<td>4.910</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Chloroplast and plastids</td>
<td>273</td>
<td>0.275</td>
<td>&lt;0.0001</td>
<td>−0.646</td>
<td>0.518</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>134</td>
<td>0.434</td>
<td>&lt;0.0001</td>
<td>1.012</td>
<td>0.311</td>
</tr>
<tr>
<td>Plasma membrane</td>
<td>97</td>
<td>0.135</td>
<td>0.187</td>
<td>7.371</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>44</td>
<td>0.304</td>
<td>0.045</td>
<td>−1.161</td>
<td>0.246</td>
</tr>
<tr>
<td>Golgi apparatus</td>
<td>12</td>
<td>0.401</td>
<td>0.196</td>
<td>2.280</td>
<td>0.023</td>
</tr>
<tr>
<td>Cell wall</td>
<td>52</td>
<td>0.343</td>
<td>0.013</td>
<td>3.210</td>
<td>0.001</td>
</tr>
<tr>
<td>Extracellular</td>
<td>51</td>
<td>0.089</td>
<td>0.532</td>
<td>5.115</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Bold values are significant at a Bonferroni corrected α = 0.0055.

a Wilcoxon rank test of difference across compartments (positive values: overduplication; negative values: underduplication).

Association of Flux and Duplication Extends Across Compartments and Functional Annotations

Gene families were associated with GO Slim annotations (supplementary table 1, Supplementary Material online) for both cellular compartment and function. We found significant Spearman's correlations between the flux and duplication rate for the metabolic gene families from the chloroplast and mitochondria (table 2). Likewise, gene families that have a role in DNA or RNA binding or metabolism, hydrolase activity, and responses to stimuli or stress had significant correlations between the number of duplications and flux (table 3).

Table 3

<table>
<thead>
<tr>
<th>Function</th>
<th>n</th>
<th>r'</th>
<th>p</th>
<th>Z^d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell organization and biogenesis</td>
<td>29</td>
<td>0.209</td>
<td>0.274</td>
<td>1.010</td>
<td>0.312</td>
</tr>
<tr>
<td>Developmental processes</td>
<td>20</td>
<td>−0.132</td>
<td>0.578</td>
<td>1.693</td>
<td>0.090</td>
</tr>
<tr>
<td>DNA or RNA binding or metabolism</td>
<td>26</td>
<td>0.760</td>
<td>&lt;0.0001</td>
<td>−2.284</td>
<td>0.022</td>
</tr>
<tr>
<td>Electron transport</td>
<td>7</td>
<td>0.860</td>
<td>0.013</td>
<td>0.460</td>
<td>0.645</td>
</tr>
<tr>
<td>Hydrolyase</td>
<td>114</td>
<td>0.310</td>
<td>&lt;0.001</td>
<td>−1.661</td>
<td>0.097</td>
</tr>
<tr>
<td>Kinase activity</td>
<td>62</td>
<td>−0.031</td>
<td>0.817</td>
<td>1.167</td>
<td>0.243</td>
</tr>
<tr>
<td>Nucleic acid or Nucleotide binding</td>
<td>94</td>
<td>0.062</td>
<td>0.554</td>
<td>0.011</td>
<td>0.991</td>
</tr>
<tr>
<td>Protein binding or metabolism</td>
<td>121</td>
<td>0.139</td>
<td>0.128</td>
<td>1.463</td>
<td>0.143</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>13</td>
<td>0.104</td>
<td>0.735</td>
<td>2.450</td>
<td>0.014</td>
</tr>
<tr>
<td>Stimulus or stress response</td>
<td>199</td>
<td>0.302</td>
<td>&lt;0.0001</td>
<td>2.650</td>
<td>0.008</td>
</tr>
<tr>
<td>Transferase activity</td>
<td>166</td>
<td>0.211</td>
<td>0.006</td>
<td>−0.833</td>
<td>0.405</td>
</tr>
<tr>
<td>Transports or transporters</td>
<td>56</td>
<td>−0.034</td>
<td>0.801</td>
<td>1.755</td>
<td>0.079</td>
</tr>
</tbody>
</table>

Bold values are significant at a Bonferroni corrected α = 0.0042.

To determine whether duplication rates differed among compartments or classes, we used Wilcoxon rank-sum test (Z-scores in tables 2 and 3). Although gene families could appear in more than one annotation group, families located in the nucleus, cytosol, plasma membrane, cell wall, and extracellular space were significantly overduplicated compared with all other gene families (table 2). No functional categories were significantly overduplicated (table 3).

Selection on Sequence Evolution of Ion Transporters

We chose to analyze the ion transporters because of their interesting role as potential chokepoints. In the metabolic networks used in this analysis, the gene families representing transporters have significantly higher flux than nontransporter gene families (Mann–Whitney one-tailed P < 10⁻¹⁵). However, we found no significant correlation between the flux and duplicability among transporter gene families (P = 0.599). Therefore, we chose to look at the fine-scale differences in selection in two classes of ion transporters, phosphate and sodium. These elements have distinct roles in the growth and development of plants and hence potentially differing duplication dynamics. Phosphate transporters import an essential macronutrient, while sodium transporters primarily limit the import of potentially toxic sodium (Rausch and Bucher 2002; Kronzucker and Britto 2011). By narrowing our focus to just these 5 gene families and limiting ourselves to the three species (A. thaliana, P. trichocarpa, and C. papaya), it is possible to manually isolate SSD and WGD events. This inference in turn allows us to assess if the strength of selection differs following WGD, SSD, and speciation.

Phosphate Transporters

Phosphate transporters in A. thaliana are divided into four gene families. These families include the high-affinity transporters (PHT1; Mudge et al. 2002; Poirier and Bucher 2002), which import ions across the plasma membrane, and the mitochondrial (PHT3; Hamel et al. 2004) and chloroplast (PHT4; Guo et al. 2008) transporters, which act in their respective organelles. Finally, low-affinity (PHT2) phosphate transporters are also localized to the chloroplast (Versaw and Harrison 2002). We inferred gene phylogenies for the four phosphate transporter families and for one sodium transporter family (see Materials and Methods). Although the topology of phosphate transporter gene families is easily reconciled to the species tree, none of the clades contained the 4:2:1 ratio of A. thaliana to P. trichocarpa to C. papaya genes that would be expected if all transporters had been retained following the α, β, and P. trichocarpa–WGDs and no SSDs had been retained (fig. 2a and b). The average selective constraint (K_s/K_a) for PHT gene families varies considerably from 0.076 in high-affinity transporters to 0.207 in low-affinity transporters (table 4). The lowest K_s/K_a corresponds to the family with the largest observed number of
seem to be under considerably higher selective constraint than gene orthologs; however, SSDs with the fewest duplications (low-affinity transporters: 1 duplication; Table 4) have significantly higher $K_a/K_s$ values than small-scale duplicates, indicating a weaker selective constraint following WGD. In all cases, the branches following duplications into those due to WGD show significantly higher $K_a/K_s$ values for whole-genome duplicates are not significantly different than those for SSDs. Among the mitochondrial transporters, whole-genome duplicates have significantly higher $K_a/K_s$ than small-scale duplicates, indicating a weaker selective constraint following WGD. The counts of WGDs versus SSDs per gene family are statistically uninformative (Fisher’s Exact test: $P = 0.75$).

**Discussion**

**Selection on Plant Gene Duplications**

Although it has been hypothesized that a substantial fraction of the surviving duplicate genes in the genomes of multicellular eukaryotes might be due to the neutral fixation of duplicates (Lynch and Conery 2003), other potential forces could also be involved (Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010). Here, we have taken both a low-level and a high-level approach to look for evidence of selection in the process of gene and genome duplications in the plants.

**Selection, Sequence Evolution, and Ion Transporters**

Part of our analysis focused on sequence evolution in two families of ion transporters. Transporters sometimes appear to be the limiting step in metabolic pathways (Brown et al. 1998; Pritchard and Kell 2002), a fact that may partly explain why their evolution after both SSD and WGD is distinct from other metabolic genes (Lin and Li 2010; Bekaert and Conant 2011). Limiting our analysis to single gene families also allows us to carefully distinguish WGD and SSD events and to model the selective constraints acting on these genes.

| Table 4 |
| Selective Constraint Estimated with Three Models of Gene Evolution for Ion Transporters of A. thaliana, C. papaya, and P. trichocarpa |

<table>
<thead>
<tr>
<th>Model</th>
<th>Branches</th>
<th>$K_a/K_s$</th>
<th>$K_a/K_s$</th>
<th>$K_a/K_s$</th>
<th>$K_a/K_s$</th>
<th>$K_a/K_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R_Null</td>
<td>All</td>
<td>0.076</td>
<td>0.207</td>
<td>0.114</td>
<td>0.148</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15379.5</td>
<td>3577.0</td>
<td>5893.8</td>
<td>24869.3</td>
<td>11210.1</td>
</tr>
<tr>
<td>R_Dupl</td>
<td>Speciation</td>
<td>0.063a</td>
<td>0.156a</td>
<td>0.080a</td>
<td>0.123a</td>
<td>0.062a</td>
</tr>
<tr>
<td></td>
<td>Duplication</td>
<td>0.082a</td>
<td>0.415a</td>
<td>0.133a</td>
<td>0.249a</td>
<td>0.031a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15376.7</td>
<td>3570.2</td>
<td>5894.1</td>
<td>24847.0</td>
<td>11188.5</td>
</tr>
<tr>
<td>R_WGD</td>
<td>Speciation</td>
<td>0.063</td>
<td>—</td>
<td>0.080a</td>
<td>0.123</td>
<td>0.061a</td>
</tr>
<tr>
<td></td>
<td>WGD</td>
<td>0.081</td>
<td>—</td>
<td>0.190a</td>
<td>0.233</td>
<td>0.067a</td>
</tr>
<tr>
<td></td>
<td>SSD</td>
<td>0.085</td>
<td>—</td>
<td>0.112a</td>
<td>0.265</td>
<td>0.018a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15376.6</td>
<td>—</td>
<td>5891.2</td>
<td>24846.7</td>
<td>11175.3</td>
</tr>
</tbody>
</table>

a Bold values indicate a significant improvement over the model immediately above at $P < 0.05$; nested likelihood ratio test (distributed $\chi^2$, $P < 0.05$, degrees of freedom = 1).

b No small scale duplications in PHT2, so model R_Dupl is equivalent to model R_WGD.

c WGD: determined by syntenic paralogy using the Plant Genome Duplication Database (Tang, Bowers, et al. 2008).

d SSD: determined either by a lack of syntenic paralogy and/or by tandem duplication status.

discussions (high-affinity transporters: 19 duplications), whereas the highest $K_a/K_s$ values correspond to the family with the fewest duplications (low-affinity transporters: 1 duplication). This observation is, however, without statistical significance. In all cases, the branches following gene duplications show significantly higher $K_a/K_s$ than do those following speciation (Table 4; but note that the small size of the low-affinity family limits the strength of our conclusion for that family). We also investigated selective constraints associated with duplication mechanism by dividing the branches following duplications into those due to WGD and to SSD. Here, the difference in selective constraint is less clear: for the high-affinity and chloroplast phosphate transporters, the $K_a/K_s$ values for whole-genome duplicates are not significantly different than those for SSDs. Among the mitochondrial transporters, whole-genome duplicates have significantly higher $K_a/K_s$ than small-scale duplicates, indicating a weaker selective constraint following WGD. The counts of WGDs versus SSDs per gene family are statistically uninformative (Fisher’s Exact test: $P = 0.75$).

**Sodium Transporters**

The angiosperm sodium ion transporters (NHX) are a single gene family responsible for keeping Na⁺ concentrations at nontoxic levels (Rodríguez-Rosales et al. 2008). The sodium ion transporters have a lower average $K_a/K_s$ than do any of the phosphate transporter families (0.049 versus 0.076–0.207). Curiously, among these transporters, paralogs have significantly lower $K_a/K_s$ values than do orthologs, indicating no release in a selective constraint after duplication (Table 4). Genes duplicated by WGD seem to be under slightly less selective constraint than gene orthologs; however, SSDs seem to be under considerably higher selective constraint than either.
Hughes (1993) found no evidence for the relaxation of selective constraint among 17 genes in the tetraploid frog Xenopus laevis. Kondrashov et al. (2002) found that recent paralogs were under significantly lower selective constraints than orthologs, whereas others (Lynch and Conery 2000; Kondrashov et al. 2002; Zhang et al. 2003; Jordan et al. 2004) have found evidence for a decrease in selective constraint immediately following duplication. This relaxation appears to be temporary; Jordan et al. (2004) found that the average strength of purifying selection acting on old duplicates was higher than for nonduplicated genes. This observation presumably reflects the situation after the fate determining mutation, which breaks the selective symmetry of two duplicates and sends them down differing paths (Innan and Kondrashov 2010). Among PHTs, our results parallel those of Jordan et al. (2004) in finding a general relaxation of selective constraint after ion transporter duplication. This result is not supported among the NHX transporters. This difference may be due to the limited evolutionary paths opened by a duplication of sodium transporters compared with that of phosphate transporters (Kronzucker and Britto 2011).

We also extended our analysis to differences in constraint between SSD- and WGD-produced duplicates. We had no a priori hypothesis on which mechanism would impart higher selective constraint, and, in fact, we found both possible outcomes.

Associations Between Duplication Propensity and Metabolic Flux

We also made a large-scale analysis of the patterns of evolution in the metabolic network. To our knowledge, this analysis represents the first high-level phylogenomic-scale study of gene duplication and metabolism in angiosperms (for studies of metabolism following WGD in other organisms, see Gout et al. 2009; van Hoek and Hogeweg 2009). By focusing on metabolism, we can ask whether duplications are randomly distributed across the network (as might be expected if drift were the only force at work) or show biases in the patterns of fixation. Notably, we find that there is a statistically significant relationship between duplication propensity and each enzyme's predicted flux. This analysis follows our work on absolute and relative dosage among the Arabidopsis WGD duplicates (Bekaert et al. 2011), where we found that reactions with high flux were enriched for enzymes coded by duplicate genes produced by the ancient β event (but not the more recent α event). Here, we have shown that the relationship between the flux and duplication is not specific to Arabidopsis but a more general pattern in plants. Although it is certainly not the case that all gene duplications are associated with high-flux reactions (the association magnitudes found are small), selection for increased gene dosage (Kondrashov and Kondrashov 2006; Conant and Wolfe 2008) is an attractive explanation for the fixation of some of these duplicates. In fact, examples of plant duplications apparently fixed by such selection are well known (van Hoof et al. 2001; Widholm et al. 2001).

Because the association of flux with duplication holds for both SSD and WGD events, we propose that different types of selective environment favor dosage-based duplicates produced by the two mechanisms. Thus, SSD may be useful in situations where the increased dosage would be beneficial at the tips of a pathway or in secondary metabolism: this is likely the case for the copper tolerance duplication in bladder campion (Silene vulgaris; van Hoof et al. 2001). However, as Kacser and Burns (1981) pointed out, for most metabolic pathways, it is unlikely that a single reaction is flux limiting, meaning that a single gene duplication is unlikely to alter the flux in such a pathway. WGD is a potential route to increased flux in such situations, and it appears that such selection may have occurred after a WGD in the ancestor of bakers’ yeast (Saccharomyces cerevisiae; Conant and Wolfe 2007; Merico et al. 2007; van Hoek and Hogeweg 2009).

Taking these analyses to the subcellular level, we find strong correlations between flux and duplication in the mitochondria and chloroplast, but not in the cytosol. This result suggests that the general association between flux and duplication is primarily driven by reactions in these compartments, an unsurprising conclusion given the roles of the chloroplast and the mitochondria as the plant cell’s anabolic and energy-yielding centers. These patterns also accord well with our prior analyses of the compartmental evolution in the Arabidopsis and human metabolic networks (Bekaert et al. 2011; Hudson and Conant 2011).

Gene and Genome Duplication, Selection, and Contingency

Although a WGD that occurs in a particular individual is much less likely to be selectively neutral than an SSD (Vieta 2005), it does not follow that there should be strong selection at every locus duplicated in such an event. Although it might therefore appear that WGD produces a large class of duplicate genes that evolve more or less neutrally after WGD, this hypothesis is difficult to reconcile with observations such as the dosage balance hypothesis. To distinguish between these two hypotheses, one might consider what the sources of variation in the selective constraint are for the set of WGD-produced duplicate genes in a genome. In fact, the number of sources of variation in constraint among duplicates at large (Duret and Mouchiroud 2000; Pál et al. 2003; Drummond et al. 2006; Vitkup et al. 2006) suggests the importance of “contingency” in duplicate evolution. In other words, a duplicate’s fate will depend on both its intrinsic properties (including factors studied here, such as function, cellular compartment, and duplication mechanism) as well as the environment in which it finds itself at birth.
Supplementary Material

Supplementary table 1 is available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Patrick Edger and our molecular evolution course for helpful discussions. This work was supported by the National Library of Medicine Biomedical and Health Informatics Training Fellowship (LM007089-19) to C.M.H., the University of Missouri Life Sciences Fellowship to E.E.P., the US National Science Foundation (DBI 0501712 and DBI 0638536) to J.C.P., and the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri to G.C.C.

Literature Cited


Selection and Gene Duplication

Hakes L, Pinney J, Lovell S, Oliver S, Robertson D. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 8:R209.


Thomas B, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.


