Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution

Anna R. Kersting, Erich Bornberg-Bauer, Andrew D. Moore*, and Sonja Grath*

Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Muenster (WWU), Germany

*Corresponding author: E-mail: radmoore@uni-muenster.de; s.grath@uni-muenster.de.

Accepted: 9 January 2012

Abstract

Plant genomes are generally very large, mostly paleopolyploid, and have numerous gene duplicates and complex genomic features such as repeats and transposable elements. Many of these features have been hypothesized to enable plants, which cannot easily escape environmental challenges, to rapidly adapt. Another mechanism, which has recently been well described as a major facilitator of rapid adaptation in bacteria, animals, and fungi but not yet for plants, is modular rearrangement of protein-coding genes. Due to the high precision of profile-based methods, rearrangements can be well captured at the protein level by characterizing the emergence, loss, and rearrangements of protein domains, their structural, functional, and evolutionary building blocks. Here, we study the dynamics of domain rearrangements and explore their adaptive benefit in 27 plant and 3 algal genomes. We use a phylogenomic approach by which we can explain the formation of 88% of all arrangements by single-step events, such as fusion, fission, and terminal loss of domains. We find many domains are lost along every lineage, but at least 500 domains are novel, that is, they are unique to green plants and emerged more or less recently. These novel domains duplicate and rearrange more readily within their genomes than ancient domains and are overproportionally involved in stress response and developmental innovations. Novel domains more often affect regulatory proteins and show a higher degree of structural disorder than ancient domains. Whereas a relatively large and well-conserved core set of single-domain proteins exists, long multi-domain arrangements tend to be species-specific. We find that duplicated genes are more often involved in rearrangements. Although fission events typically impact metabolic proteins, fusion events often create new signaling proteins essential for environmental sensing. Taken together, the high volatility of single domains and complex arrangements in plant genomes demonstrate the importance of modularity for environmental adaptability of plants.

Key words: plant genome evolution, modular evolution, whole-genome duplication, evolution of stress response.

Introduction

The wealth of genomic data has governed a number of insightful studies on genome evolution. To date, most studies have concentrated on gene duplications, gene family expansion or reduction, selective sweeps or signals of selection using site-based statistics. An alternative approach to studying genome evolution utilizes the modular nature of proteins. Most proteins are composed of one or many protein domains, which are the units of protein structure, function, and evolution (Söding and Lupas 2003; Moore et al. 2008). The majority of proteins can be described using a small set of domains, which, despite the ever-increasing amount of available sequence data, grows at only moderate speed. In contrast, the number of domain arrangements, that is, the combination of these domains in proteins, continues to rapidly grow (Levitt 2009; Yang et al. 2009). The study of domain rearrangements across large phyla has provided a detailed understanding of modular protein evolution (Björklund et al. 2005; Ekman et al. 2007; Fong et al. 2007; Wang and Caetano-Anolles 2009; Yang et al. 2009) and has demonstrated that domain rearrangements, paired with the occasional formation of novel domains (Moore and Bornberg-Bauer 2012), create an enormous degree of protein diversity (Apic et al. 2001; Levitt 2009; Yang et al. 2009). The majority of eukaryotic proteins have more than one domain (Apic et al. 2001; Ekman et al. 2005; Yang et al.

2009), and while many domains are found in few arrangements, only few domains are versatile and form a wide array of different arrangements (Weiner et al. 2008; Cohen-Gihon et al. 2011). Rearrangement events at the protein level are easy to detect, and the key mechanisms are thought to be fusion, fission, and terminal deletion (Björklund et al. 2005; Weiner et al. 2006). These events are likely fueled by a series of underlying genetic events such as nonallelic homologous recombination, exon-shuffling, nonhomologous end joining or transposition (Babushok et al. 2007; Buljan et al. 2010). However, with few exceptions (e.g., Oshima et al. 2010), traces of the genetic mechanisms of rearrangement swiftly decay. Buljan et al. (2010) explored the genetic events that facilitate domain gain events to existing arrangements. Their results provide support to the notion that domains are typically added at either terminus. The key mechanism for such domain gain events involves the joining of exons between genes or terminal exon extension. The study of domain content evolution in eukaryotes has illustrated that domain loss and gain are frequent events (Moore and Bornberg-Bauer 2011; Zmasek and Godzik 2011). Whereas lost domains tend to be of catalytic nature, gained domains tend to be regulatory. Despite the diverse studies that have explored modular evolution across many species as well as in restricted clades, to date no study has quantitatively addressed the topic of modularity in a set of plant species. However, modular evolution may be of particular importance for plants, as they face a challenge that many other species do not—they cannot easily evade environmental changes because of their sessile nature. In particular the fusion of genes, and consequently of domain arrangements, allows for "jumps" in protein evolution and may govern truly novel genetic phenotypes. Hence such fusion proteins may exhibit great adaptive potential. Indeed, recent findings suggest that chimeric genes formed by gene fusion can be found in regions of selective sweeps (Rogers and Hartl 2012).

Fusion events have been shown to be associated with regulatory proteins such as the metazoan bHLH transcription factors (Amoutzias et al. 2005) or the MIKC-type MADS-box transcription factor proteins in plants (Veron et al. 2007; Shan et al. 2009). Innovation of transcription factor families is often the result of duplication events, which may occur in chromosomal regions with high recombination rates. Furthermore, it has been illustrated that duplication events in combination with high recombination rates are strong forces in genome evolution (Lang et al. 2010).

Duplications have been more frequently described for plants than elsewhere and plant genome evolution is special in several aspects. First, plant genomes are repeat-rich and transposable elements have a particularly prominent role in creating retrocopies of genes, for example, in monocots (Bennetzen 2005; Baucom et al. 2009; Baucom, Estill et al. 2009). Second, several whole-genome duplication (WGD) events have created many large genomes with various degrees of ploidy within a relatively short period of time. 35% of all vascular plants are recent polyploids (Wood et al. 2009). Moreover, angiosperms have undergone up to four rounds of WGD in roughly 320 Myr, with one WGD common to all seed plants 319 Ma and one WGD common to all angiosperms 192 Ma (van de Peer et al. 2009; Jiao et al. 2011). Although polyploidy events pose a genomic challenge to their host and most polyploidy events are considered a "dead end" for evolution (Mayrose et al. 2011), it has been suggested that polyploidy, be it the result of autopolyploidy or allopolyploidy, may occasionally provide a starting point for evolutionary innovation (Freeling et al. 2006; van de Peer et al. 2009). The benefit of an increased amount of genetic material might be to allow for swift adaptation to extreme environments (van de Peer et al. 2009). For example, the increased heterozygosity resulting from polyploidy impacts the wiring of signaling cascades and can facilitate strong variation in gene expression (Osborn et al. 2003). Numerous studies have also explicitly explored the impact of WGD in plants at the genomic level, for example, by exploring duplicate retention rates (Hanada et al. 2008; Tang et al. 2008; Zheng et al. 2009), gene dosage effects (Freeling et al. 2006; Misook et al. 2007; Bekaert et al. 2011), or recombination rates (Akhunov et al. 2003). WGDs may enhance the potential for diversification and speciation (van de Peer et al. 2009), yet the details remain poorly understood.

As genomic stability is largely influenced by genome size and repeat content (Bennetzen 2005), one might speculate that plants have high rates of recombination and hence exhibit a high number of domain rearrangements. Indeed, comparative studies have illustrated that angiosperms exhibit higher recombination rates than vertebrates (Kejnovsky et al. 2009). However, to date, no study has explored the extent of modular protein evolution in plants.

Given their large genome size, higher recombination rates, and the inability to flee upon environmental challenges, it seems likely that plants may utilize their abundant genomic material to facilitate rapid evolutionary innovation. Consequently, the benefits of modular domain rearrangements might be particularly pronounced, since the ability of modular evolution to swiftly implement changes to the protein repertoire may be a key process in both exploiting existing and creating functionalities. So far, all studies on the evolutionary dynamics and the adaptive potential of domain rearrangements have been reported for bacteria (Enright and Ouzounis 2001), metazoa (Ekman et al. 2007), or fungi (Cohen-Gihon et al. 2011), but none for plants.

In this report, we explore the nature of modular evolution in 29 green plant species (Viridiplantae) with taxa ranging from green algae to liliopsida and eudicotyledons. Our aim is to understand the evolutionary dynamics by studying the frequency of individual modular events such as fusion, fission, or terminal loss. We apply a maximum parsimony-based approach to reconstruct events placing this study into a phylogenomic framework and quantitatively address the role of domain emergence and domain rearrangements. Furthermore, we explore the speed with which new domains, and their arrangements, are gained and lost; how many of these events are clade or species-specific and whether event "hotspots" can be found amongst the phylogenies of the considered species. Finally, we employ several functional analyses based on the Gene Ontology (GO) classification (Ashburner and Lewis 2002) to shed light on the potential adaptive benefits of domain emergence and rearrangements during plant genome evolution.

Materials and Methods

Proteomes and Domain Annotation

Comparative analyses of protein domains and their arrangements were performed on the following 29 plant genomes: *Arabidopsis thaliana* v9.0 (The Arabidopsis Initiative 2000); *Arabidopsis lyrata* v1.0 (Hu et al. 2011); *Carica papaya* v1.0 (Ming et al. 2008); *Citrus sinensis* v1.0 (Sweet Orange Genome Project 2010); *Citrus clementine* v0.9 (Haploid Clementine Genome International Citrus Genome Consortium 2011); *Eucalyptus grandis* v1.0 (Eucalyptus grandis Genome Project 2010); *Mimulus guttatus* v1.1; *Aquilegia coerulea*; *Theobroma cacao* v1.0 (Argout et al. 2011); *Glycine max* v1.0 (Schmutz et al. 2010); *Medicago truncatula* v3.0 (Young et al. 2005); *Lotus japonica* v1.0 (Young et al. 2005); *Populus trichocarpa* v2.0 (Tuskan et al. 2006); *Ricinus communis* v1.0 (Chan et al. 2010); *Manihot esculenta* v1.1; *Malus domestica* (Velasco et al. 2010); *Prunus persica* v1.0 (International Peach Genome Initiative 2010); *Cucumus sativa* v1.0 (Huang et al. 2009); *Vitis vinifera* v1.0 (Jaillon et al. 2007); *Setaria italica* v2.0 (Setaria italica Genome Sequencing Project 2011); *Zea mays* v4a.53 (Schnable et al. 2009); *Sorghum bicolor* v1.4 (Dubchak et al. 2009); *Oryza sativa* v6.1 (Go et al. 2002); *Brachypodium distachyon* v1.0 (Vogel et al. 2010); *Phoenix dactylifera* v2.0 (Al-Dous et al. 2011); *Selaginella moellendorffii* v1.0 (Banks et al. 2011); *Physcomitrella patens* v1.5 (Rensing et al. 2008); *Chlamydomonas reinhardtii* v4.0 (Merchant et al. 2007); *Ostreococcus lucimarinus* v2.0 (Palenik et al. 2007); and *Micromonas pusilla* v3.0 (Worden et al. 2009).

We rooted the tree ~1.700 Ma by including *Trichoplax adhaerens* v1.0 (Srivastava et al. 2008), *Rhizopus oryzae* (Ma et al. 2009) and *Drosophila melanogaster* v5.11 (Adams et al. 2000). Phylogenetic relationships for all 32 species (29 plants and 3 outgroups) used for this study are given in supplementary figure 1 (Supplementary Material online). If several splice variants were present for one protein, we excluded all but the longest transcript. All proteomes were scanned for domains with the pfam_scan utility and HMMER3.0 (Eddy 2011) against the Pfam-A and Pfam-B models obtained from Pfam (v.24) (Finn et al. 2008). For the annotation of Pfam-A domains, we used the model-defined gathering threshold and query sequences were required to match at least 30% of the defining model (Buljan et al. 2010). Pfam-B domains were annotated using an *E* value cutoff of 0.001 (Ekman et al. 2007). Pfam-A domains with clan membership were mapped to their clans and domains of type "repeat" or "motif" were collapsed into one large domain instance (Ekman et al. 2005; Forslund et al. 2007).

Reconstruction of the Ancestral Domain State; Domain Gain, Loss, and Emergence

We reconstructed ancestral domain contents using a maximum parsimony approach as follows: the tree (see fig. 1*B*) was traversed twice, first from leaves to root then from root to leaves. Domain presence or absence is determined by majority rule. During first traversal (leaves → root), the state of domain *d* is set to present at a node *n*, if *d* is present in the majority of leaves of the subtree rooted in *n* (leaves of *n*). Similarly, *d* is set to absent at *n*, if *d* is absent in the majority of leaves of *n*. If there is no state majority for *d* in the child nodes of *n* (i.e., there is an identical proportion of presence and absence states for *d* in the leaves), the state of *d* at *n* is set to unknown. As traversal continues toward the root, *d* is set to present (absent) at *n* as soon as the majority of leaves of *n* exhibit the present (absent) state. Ergo, present and unknown are resolved to present, while unknown and absent are resolved to absent. The first traversal terminates at the root node. All unknown states at the root node are set to present (note that this root includes the outgroups). During the second traversal (root → leaves), unknown states are resolved by setting them to the state of their ancestor. We used a combination of custom-made python scripts and the ETE2 package (Huerta-Cepas et al. 2010) for tree traversal. Branch lengths of the tree (Soltis et al. 2002; Choi et al. 2004; Magallón and Sanderson 2005; Hedges et al. 2006; Cartwright and Collins 2007; Anderson and Janßen 2009; Bhattacharya et al. 2009; Bremer et al. 2009; Forest and Chase 2009; Herron et al. 2009; Wang and Caetano-Anolles 2009; Lang et al. 2010; Reineke et al. 2011) and whole-genome duplication events (Blanc and Wolfe 2004; Schnable et al. 2009; van de Peer et al. 2009; Jiao et al. 2011) were extracted from the literature.

We performed a Blast (Altschul et al. 1997) search to identify recently duplicated proteins. Proteins with a similarity of 75% or more and an *E* value $\leq 10^{-20}$ were considered to be paralogs. We employed a synteny analysis to distinguish between tandem and segmental duplications. Two genes were considered to be tandem duplicates if they were five or less genes apart. Paralogs with more than five genes between them were considered to be a result of a segmental duplication event (Hanada et al. 2008).

Domain gain and loss events along branches were calculated by comparison of domain content at a given node with

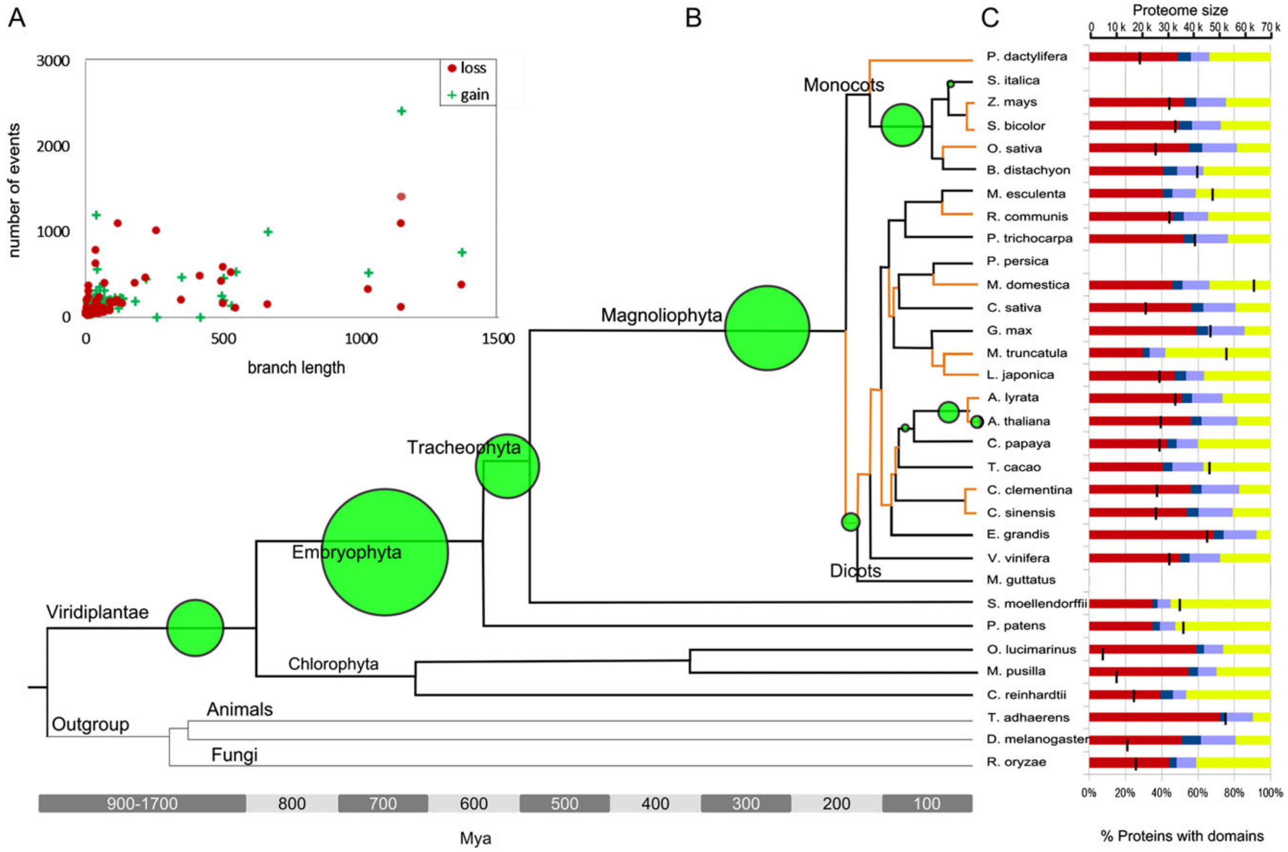

Fig. 1.—Domain gain, loss, emergence and proteome coverage of 26 plant genomes. (*A*) Correlation of domain gain and loss with branch length. Both gain and loss correlate significantly with branch length (gain: $\rho = 0.6$, $P < 0.001$; loss: $\rho = 0.63$, $P < 0.001$). (*B*) Phylogenetic relationship of all species used in this study. For each branch, the size of the green circle corresponds to the number of domain emergence events along the branch. Branches colored in red indicate that the gain and/or loss at this branch is higher than the average gain and/or loss rates. Exact values for domain gain, loss, and emergence are given in supplementary table 2 (Supplementary Material online). (*C*) Domain coverage for proteins. The lower axis (percentage of proteins with domains) displays the proportion of proteins with only Pfam-A domains (red), only Pfam-B domains (dark blue), both Pfam-A and Pfam-B domains (light blue), and without any protein domain annotation (yellow). The upper axis displays proteome size indicated as vertical black line for each species. Statistics for three species (*Setaria italica*, *Prunus persica*, and *Mimulus guttatus*) that are still under Fort Lauderdale restriction are not provided.

the domain content of its ancestor. We distinguish between "gained" domains, which are all domains found present at a node while absent in its ancestor, and "emerged" domains, which are gained domains which can only be found within Viridiplantae. Ergo, emerged domains are a subset of the gained domains. Emerged domains were determined by scanning gained domains with HMMER3.0 against NCBI NR and Integr8 (Kersey et al. 2005). Gained domains, which are not present in the outgroups were also scanned against NCBI NR to determine the kingdoms where these domains are present (supplementary table 6, Supplementary Material online). Domain event rates (gain and loss) were calculated by dividing the number of events predicted to occur along a given branch by the branch length (in million years).

Given the evidence that novel domains are frequently enriched in structural disorder (Buljan et al. 2010; Moore and Bornberg-Bauer 2012), we predicted disorder in domains classified as emerging. VSL2.0 (Obradovic et al. 2005) was used to detect structural disorder in domain sequences. Emerged domains were divided into four bins (Viridiplantae, Embryophyta, Tracheophyta, and Magnoliophyta), corresponding to their emergence nodes. Domains that emerged after the Magnoliophyta node were pooled into one "RECENT" bin. To compare disorder of emerged domains with old domains (i.e., domains that exist at the root), a bin "OLD" was constructed consisting of 500 randomly picked domains occurring in the root. In addition, we constructed a "RANDOM" bin consisting of 100 randomly selected domains, which exist at the root. To account for sampling bias, we repeated the random selection 100 times. Statistical inference was conducted with the kruskalmc test of the R package pgirmess (Siegel and Castellan 1988; R Development Core Team 2008).

We quantified domain emergence and explored a set of attributes (Moore and Bornberg-Bauer 2012). Domain frequency, $d(f)$, is defined as the absolute frequency of a domain across all plant genomes used for the analysis. The domain rate $x(d)$ of domain d is defined as the domain frequency divided by the number of plants in which d occurs. The domain success rate corresponds to the domain rate divided by the node age (in million years) at which the domain first emerged. The prevalence $P(d)$ of a domain d is the number of plants with d divided by the number of plants with the emergence node of d as an ancestor.

Functional Analysis of Domains

Where available, GO (Ashburner and Lewis 2002) annotation of proteomes was obtained from PLAZA 2.0 (Proost et al. 2009); Blast2GO (Götz et al. 2008) with default settings was used to functionally annotate the remaining proteomes. Comparative functional analyses were performed by assessing GO-term overrepresentation (overrepresentation analysis, ORA) in two separate steps. First, for emerging domains, we performed the functional analysis indirectly by using the GO annotation of arrangements that harbor at least one emerging domain, similar to a previous approach (Moore and Bornberg-Bauer 2012). Statistical inference was conducted using the R package TopGo (Alexa et al. 2006). As universe, we used the GO annotation of all proteins in our data set; the sample consisted of arrangements with emerging domains. Second, for assessing functional overrepresentation of arrangements in events (such as fusion or fission), we again conducted an ORA using the proteins GO annotation, however, our sample here was the arrangement set that results from a specific event (e.g., all gained arrangements explainable by a fusion event). *P* value transformed TermClouds were created by logarithmic transformation of the False Discovery Rate (FDR)-corrected (Benjamini and Hochberg 1995) *P* value obtained from the ORA, such that term size represents the significance of the GO term. Visualization was created using Wordle (http://www.wordle.net/) with the transformed *P* value as a custom scaling factor.

Reconstruction of the Ancestral Domain Arrangements State, Arrangement Gain, and Loss

We defined domain arrangements as ordered sets of domains for each protein. For the analysis of arrangements in this study, only Pfam-A domains were used. Ancestral states for arrangements were reconstructed as previously described. Similarly, arrangement gain and loss was determined by comparing current and ancestral states.

Determination of Arrangement Rates

For each gained arrangement, we applied a search algorithm to determine the possible mechanism that led to its formation. We considered the four most important mechanisms of modular rearrangements—fusion, fission, terminal deletion, and domain addition (Björklund et al. 2005; Pasek et al. 2006; Weiner et al. 2006; Buljan et al. 2010). The algorithm assigns a fusion event when two ancestral arrangements can be fused to form the gained arrangement. A gained arrangement is considered to be the result of fission if an ancestral arrangement can be split to give rise to the new arrangement; both products of the split are required to be present in the current node. In contrast, for terminal deletion, only one product of the split (the gained arrangement) may be present in the current node (the other product is considered to be lost). The algorithm counts a domain addition event when the newly gained arrangement contains a domain that is absent in the ancestral node.

Note that in general, any new arrangement can be explained by a sufficiently large "chain" of events. However, since the likelihood of events is not available, we make no assumptions about the relative costs of each mechanism and therefore are not able to determine the most likely chain. Instead, we focus on single-step solutions, that is, on cases where a newly gained arrangement can be explained by a single event. Using this strategy, we can differentiate between arrangements with exact solution (i.e., the formation can be explained by exactly one mechanism), arrangements with nonambiguous solution (i.e., only one mechanism explains the arrangement but there are several events possible) and arrangements with ambiguous solution (i.e., conflicting solutions of different types). All arrangements with solution are referred to as "simple gains," whereas all other arrangements are considered to be "complex gains."

Results

Domain Coverage

In plants, on average, 50% of the proteome residues were found to be covered by domain annotation; the residue coverage ranges from 30% to 70% (supplementary table 1 and fig. 2, Supplementary Material online). For an average of 35% of the residues, for each plant, a Pfam-A domain can be detected, whereas Pfam-B domains affect 15% of all residues. Residue coverage levels for all species are given in supplementary table 1 (Supplementary Material online).

At the protein level, the coverage distribution is more diverse (supplementary table 1 and fig. 2, Supplementary Material online). On average, 70% of the proteins for one plant species have at least one Pfam-A or Pfam-B domain. Fifty percent of the proteins contain only Pfam-A domains, 14% contain only Pfam-B domains, and 6% contain both Pfam-A and Pfam-B domains (fig. 1*C*). All protein coverage values are given in supplementary table 1 (Supplementary Material online). The total number of proteins containing Pfam-A and Pfam-B domains is highly variable between the different proteomes (fig. 1*C*, supplementary table 1, Supplementary Material online).

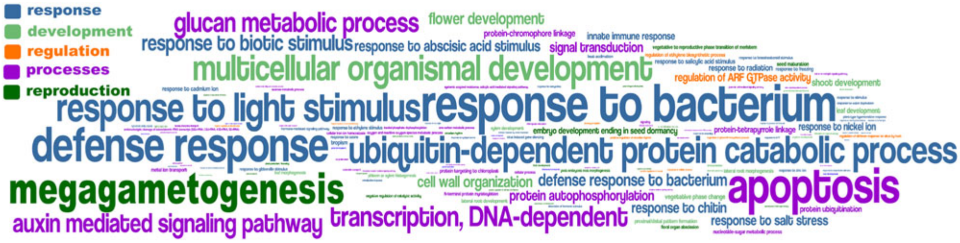

Fig. 2.—Gene Ontology (GO) terms associated with emerging domains. GO terms affected by emergence were tested for overrepresentation using the TopGO package and all terms present in plants as universe (for details, see Materials and Methods). The font size corresponds to the value of significance obtained for this term. Significance was determined after correction for multiple testing using FDR (Benjamini and Hochberg 1995) correction at $P < 0.01$. The vast majority of GO terms is related to stimulus response, development, reproduction, regulation, and plant-specific metabolic processes.

Domain Emergence

To investigate domain gain, loss, and emergence across the considered plants, we reconstructed the ancestral domain content at each internal node of the tree (see also Materials and Methods; supplementary fig. 1, Supplementary Material online). In total, 545 domains emerged in the plant kingdom, that is, these domains are exclusively found in Viridiplantae. The largest amount of domain emergence within plants occurs along the branch leading to Embryophyta, which sees the emergence of 262 domains (fig. 1*B*). A total of 114 and 66 domains emerge along the branches to Magnoliophyta and Tracheophyta, respectively. Fifty-one domains emerged prior to the split of Embryophyta and the green algae and 52 domains are the result of recent emergence events and can only be found within Magnoliophyta (see also Discussion below) (fig. 1*B*).

Radiation and Functional Impact of Emerging Domains

Next, we assessed whether emerged domains confer specific functionalities and whether these might provide adaptive benefit. We assessed functional overrepresentation using GO categories and TopGO (Alexa et al. 2006) (see Materials and Methods for details). We find that GO terms prefixed by response_to are overrepresented along with functionalities related to reproduction, developmental mechanisms, and metabolic processes (fig. 2).

We binned emerging domains according to their point of emergence (for details, see Materials and Methods) and ranked them by their frequency *d*(*f*). The 5% highest ranked domains from each age bin (supplementary table 3, Supplementary Material online) were subject to further investigation as these can be considered to be particularly "successful" emerging domains. Among these, we find domains with plant-specific functions such as flowering control, auxin regulation, fruit development, cell wall development, and plant organelle recognition. Furthermore, we detected domains related to the F-box protein family, to transcription factors and to DNA binding. For the majority of emerging domains, direct functional annotation is difficult—the largest proportion (85%) of all emerging domains in plants are domains of unknown function (DUFs) or belong to the set of poorly annotated Pfam-B domains. We assessed functional overrepresentation using the function of proteins that obtain emerging domains—we are hence not exploring which functional modules emerge but rather which protein functionalities undergo innovation (by the addition of an emerging domain).

There is increasing evidence that young domains can exhibit higher levels of structural disorder than established domains (Buljan et al. 2010; Moore and Bornberg-Bauer 2012). We examined the degree of structural disorder in emerging domains. The results indicate that emerging domains are significantly enriched in intrinsic disorder, more than in randomly chosen domains (see Materials and Methods; supplementary fig. 3, Supplementary Material online). Furthermore, the younger a domain, the higher the degree of disorder.

Domain Gain and Loss

Domain gain and loss are frequent events in plant evolution, and there is a strong variation between different branches (fig. 1*A*). Nevertheless, both gain and loss rates correlate significantly with branch length (Spearman rank correlation, gain: $\rho = 0.6$, $P < 0.001$; loss: $\rho = 0.63$, $P < 0.001$). On average, plants have a domain gain rate of 6.64/Myr and a domain loss rate of 6.11/Myr (fig. 1*A*, supplementary table 2 and fig. 9, Supplementary Material online). In monocots, the average domain gain rate (6.7/Myr) is lower than the domain loss rate (7.4/Myr), whereas in eudicots the situation is reversed; eudicots show a loss rate of 7.4/Myr and a gain rate of 8.3/Myr (supplementary table 2 and fig. 9, Supplementary Material online). Some branches exhibit very high loss rates, such as the branch leading to *P. dactylifera*, the branches to the two Fabaceae *M. truncatula* and *L. japonica*, and the branches to the two Andropogoneae *Z. mays* and *S. bicolor* (fig. 1*B*).

Gain, Loss, and Distribution of Arrangements

We next explored the dynamics of arrangement gain and loss. After determining the presence/absence of arrangements at ancestral nodes (for details, see Materials and

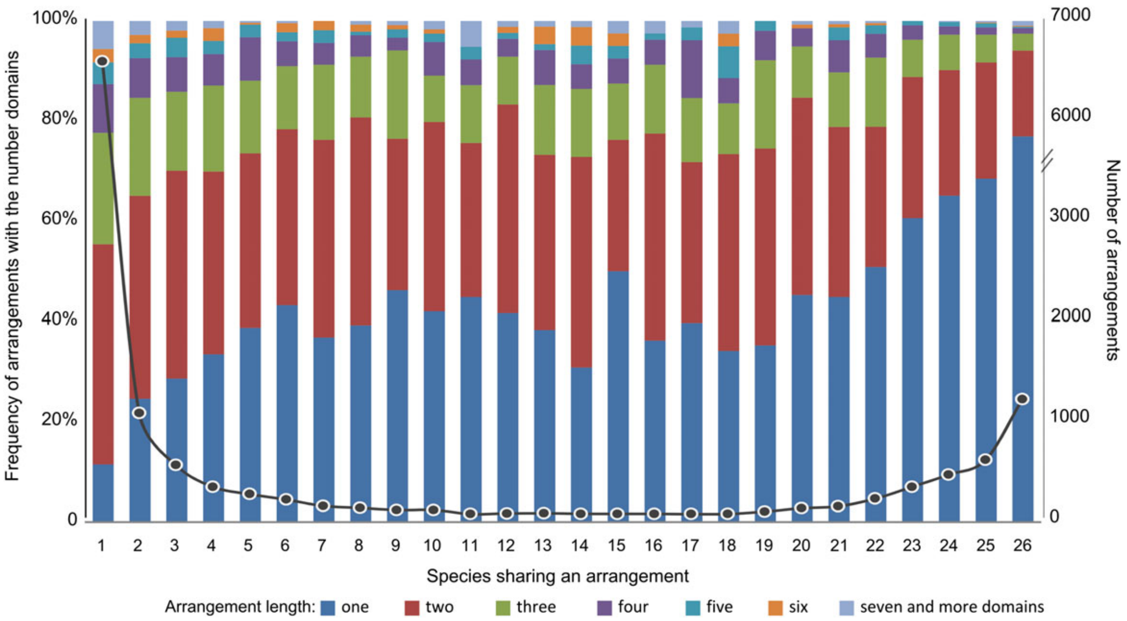

Fig. 3.—Arrangements shared between species. The dashed line represents the number of arrangements shared by the different numbers of species (right axis). The distribution of unique arrangements is roughly bimodal with the majority of arrangements shared by either few or all species. The left axis and barplots display the frequency of arrangements with a certain length (one, two, three, four, five, six, and seven or more domains). Although single-domain arrangements tend to occur in all species, longer arrangements are often species-specific.

Methods), we compared arrangement content at each node with the content at the corresponding parent node to determine arrangement gain and loss. As expected, both gain and loss rates correlate significantly with branch length (Spearman rank correlation, gain: $\rho = 0{:}56$, $P < 0.001$; loss: $\rho = 0.38$, $P = 0.003$, supplementary fig. 5, Supplementary Material online). Overall, arrangement gain rate is higher than arrangement loss rate. However, both rates correlate significantly with each other ($\rho = 0.56$, $P < 0.001$). By far, the largest amount of arrangement gain (2,814 arrangements) occurs along the branch to *M. domestica* followed by the branch to *R. communis* (1,018). Large amounts of arrangement loss can be found along the branches to *P. dactylifera* (1,028) and *L. japonica* (680); both plants also showed a high amount of domain loss. All values for arrangement gain and loss are given in supplementary table 4 (Supplementary Material online).

We investigated the amount of arrangements shared by different plants species (fig. 3). The distribution is bimodal, with the largest number of arrangements being either specific to one species (~7,000) or shared by all (~1,000); only a very small amount of arrangements is shared by 10–20 species. Although by far the largest proportion of arrangements shared by all species consists of single-domain proteins, the contrary is true for species-specific arrangements. Here, the largest number of arrangements tends to be composed of more than one domain, with a large proportion containing seven or more domains. This indicates that the longer an arrangement is, the more likely it is species-specific.

Modular Rearrangements

Using a simple model of modular rearrangement (for details, see Materials and Methods), we next explored the mechanisms that can facilitate the formation of novel arrangements. For this, we considered fusion, fission, terminal deletion, and domain addition. The results illustrate that 70% of all gained arrangements can be explained by exactly one solution (exact solutions). Of the gained arrangements, 14% can be explained by one particular mechanism, however, with a number of different possible solutions (nonambiguous solutions); only 4% have conflicting solutions (ambiguous solutions). The remaining 12% of all new arrangements are complex gains that likely arose by a chain of events (see Materials and Methods; fig. 4). The different events were found to occur with different frequencies (table 1). Fusion events makeup the largest proportion of exact solutions, followed by domain addition, fission, and terminal deletion. Fusion events occur with a frequency of 4.59/Myr, followed by fission with 1.98/Myr, and gain with 1.89/Myr. Domain deletion events can be split in C-terminal and N-terminal domain deletion; both events have a frequency of 0.7/Myr. All rates were averaged across all branches. We further explored event frequencies across different age bins. At the Embryophyta node, 68% of new arrangements are affected by domain addition and 26% by fusion. Domain deletion (4%) and fission (3%) are less prevalent at this node. Over time, the frequency of domain deletion and fission increases up to 13% and 21% in recent rearrangements, whereas domain additions

Table 1

Contribution of Fusion, Fission, C-Terminal Deletion, N-Terminal Deletion and Domain Addition to Simple Arrangement Gains

	Fusion	Fission	C-Del	N-Del	Add
Total number	9,669	4,073	1,283	1,424	4,848
Average number/Myr	4.59	1.98	0.7	0.7	1.89

NOTE.—Del, deletion; Add, addition.

decrease to a frequency of 24%. The largest fraction of recently gained arrangements (49%) can be explained by fusion events (fig. 4).

Discussion

Domain Emergence

The increasing availability of plant genomes has allowed us to conduct a comparative domain analysis between a set of diverse plant species. Here, we reconstruct the ancestral states of domain content and arrangement and investigate the functional impact of domain emergence and domain rearrangements across a comprehensive set of 29 genomes dating back ~800 Myr. However, the considered clade still contains a number of species for which genome sequences are missing, such as the gymnosperms or the charophyta. As these genomes become available, a more comprehensive picture of modular evolution in plants will emerge.

In contrast to animals, plants are sessile organisms that are unable to escape strong environmental shifts and must rather adapt to such variation. Hence, plants, more so than animals, are required to evolve mechanisms in order to deal with biotic and abiotic stresses. Here, we illustrate that the emergence of new domains can provide an important strategy for evolving stress response. More than 500 domains emerged within Viridiplantae of which more than 100 domains are unique for Tracheophyta (fig. 1). We recently assessed the impact of domain emergence in a set of insects, where only 30 domains emerged within 19 insect genomes spanning roughly 300 Myr of evolution (Moore and Bornberg-Bauer 2011). Hence, it would seem that plants exhibit a large amount of domain innovation. One might speculate that plants at least partly address the challenge of a sessile lifestyle by means of domain innovation. The investigation of GO terms of proteins containing emerged domains further supports this notion. A large number of terms are related to plant-specific processes, such as megagametogenesis and development of plant-specific organs. This is not surprising as the reproductive system and morphology of plants not only differ strongly from other kingdoms but are also highly variable between plant species (Endress 2001; Bennici 2005; Williams 2008; Kawakita and Kato 2009). Besides these plant-specific functions, a number of overrepresented GO terms correspond to response_to categories and to secondary metabolite pathways related to stress response, such as auxin and jasmonic acid. Such secondary metabolites are strongly related to the defense and response mechanisms in plants (Grace and Logan 2000; Pateraki and Kanellis 2010; Kerchev et al. 2012). As the composition of these compounds is variable between plant species and also within species (Kroymann 2011), such secondary metabolites may provide a strong flexible basis for improving adaptation and defense.

Functional links to photosynthesis are not found amongst emerged domains (fig. 2). This is likely explained by

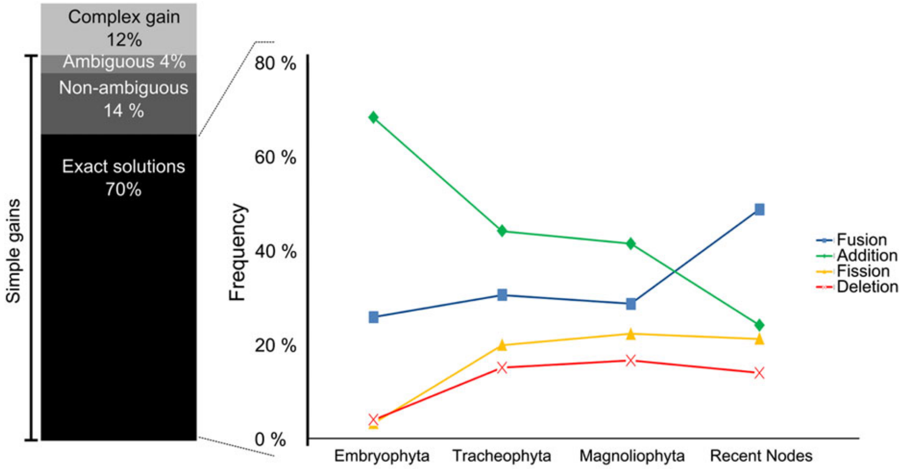

FIG. 4.—Mechanisms of rearrangement across different clades. We applied a search algorithm to assess the mechanisms that might account for newly gained arrangements (for details, see Materials and Methods). Only 12% of all gained arrangements cannot be explained by a one-step event (complex gains). The remaining 88% of simple gains can be further differentiated into exact solutions where only one particular mechanism (fusion, fission, terminal deletion, or domain addition) was necessary to explain the arrangement gain event (70%). All proteomes were divided into four different age bins: Embryophyta, Tracheophyta, Magnoliophyta, and Recent Nodes. The frequencies of fusion, fission, and terminal deletion increase over time, whereas the frequency of domain addition decreases.

photosynthesis not being unique to plants; photosynthetic processes can be found in algae and in many species of bacteria (Olson 1970, 2001). Indeed, photosynthesis-related GO terms can be detected by investigating gained domains which are absent in the outgroups (supplementary fig. 6, Supplementary Material online), as well as response_to terms and a number of plant-specific functionalities related to development, similar to those terms found in proteins containing emerged domains.

Emerged domains seem to be evolutionarily important as they have a high prevalence of 0.9–1, indicating that their occurrence is strongly conserved. Besides their widespread occurrence in nearly all leaves, such emerged domains often occur in high copy numbers (supplementary table 3, Supplementary Material online).

Investigating the most successful emerged domains uncovers connections to key functional categories such as transcription factors, binding-related processes, and secondary metabolites, including auxin and jasmonic acid (supplementary table 3, Supplementary Material online). Indeed, a burst of transcription factors and their constituent domains, which are assumed to be correlated with increasing complexity in plant evolution (Lang et al. 2010), has been found in angiosperms. The increase of plant complexity with duplication events (Freeling et al. 2006) may partly be the result of duplication facilitating increasingly complex regulatory networks (Veron et al. 2007).

Emerging domains exhibit an increased amount of intrinsic disorder; the more recent the emergence event, the more likely the domain in question exhibits intrinsic disorder. Disordered sequences may increase the binding affinity of proteins (Dyson and Wright 2005). High intrinsic disorder paired with the fact that emerged domains are significantly underrepresented in single-domain proteins (hypergeometric test, $P < 0.01$), leads us to the speculation that emerging domains may have higher interaction potential, which in turn may increase their viability and result in higher prevalence and frequency. Indeed, some of the most successful emerging domains have links to binding-related processes.

Arrangement Mechanisms

In plants, roughly 70% of the domain-containing proteins are single domain (supplementary fig. 4, Supplementary Material online). This high percentage of single-domain proteins can be an artifact of low domain coverage or "eroded-domains," which have diverged beyond detection (Weiner et al. 2006). Recent rearrangements can mostly be explained by the fusion of two single or two domain proteins. The relative rates of fusion and fission are similar to previously reported rates (Kummerfeld and Teichmann 2005). GO terms overrepresented in proteins, which arose via fusion, are stress-, defense-, and adaptation-related as well as related to the reproduction system (supplementary fig. 7, Supplementary Material online). In contrast, proteins formed by fission mainly play a role in metabolic and biosynthesis processes (supplementary fig. 7, Supplementary Material online). Proteins shaped by domain deletion are mainly related to basic functions such as the primary metabolism, and only a minor part of these proteins are stress–response related (supplementary fig. 7, Supplementary Material online).

Our results provide further evidence that duplication impacts rates of modular rearrangement (Buljan and Bateman 2009). We find that proteins affected by rearrangement events are overrepresented in duplicated genes (supplementary table 6, Supplementary Material online). Furthermore, we find indication that species with recent WGD have higher rates of fusion and fission in comparison to species without recent WGD (supplementary table 7, Supplementary Material online). In general, duplicates are thought to undergo one of three different scenarios: subfunctionalization, where the two duplicates share ancestral gene function; neofunctionalization, where one copy retains the ancestral function and the other copy diverges toward a novel function; and pseudogenization, where one copy is not expressed and is subsequently lost (Walsh 2003). One explanation for sub- or neofunctionalization is the loss or change of regulatory regions (Ganko et al. 2007). As the conservation of noncoding sequences follows an exponential decay rate (Reineke et al. 2011), the retention of both duplicates might be the result of the change of one of the gene's regulatory region under relaxed selectional constraints. The high retention rate of proteins that result from a fusion event might be explained by the conservation of at least one regulatory element in the upstream region, whereas after fission, one arising protein may lose a regulatory region and undergo pseudogenization followed by gene loss. A further reason for sub- and neofunctionalization after duplication might be domain rearrangements in one paralog or differential domain loss (Buljan et al. 2010).

We further illustrate the impact of protein domain rearrangements on an organism's protein repertoire (fig. 5). The emerging domains PAN_2 (emerged in the Tracheophyta) and S_locus_glycop (Embryophyta) often co-occur together with the B-lectin domain. Arrangements containing the two emerged domains S_locus_glycop and PAN_2 are frequently rearranged within paralogous genes (fig. 5) and obtain a catalytic function through the addition of kinase domains. Proteins that consist of arrangements with these two emerged domains have GO functions related to the recognition of pollen, protein phosphorylation, and cell recognition. Although we observed fusion events in tandemly duplicated genes in our case study, fusion events are not generally overrepresented in tandemly duplicated genes (supplementary table 5, Supplementary Material online). After fusion, duplicates might be difficult to recognize as paralogs. One might therefore speculate that in tandemly duplicated proteins fused arrangements are harder to detect. The increased

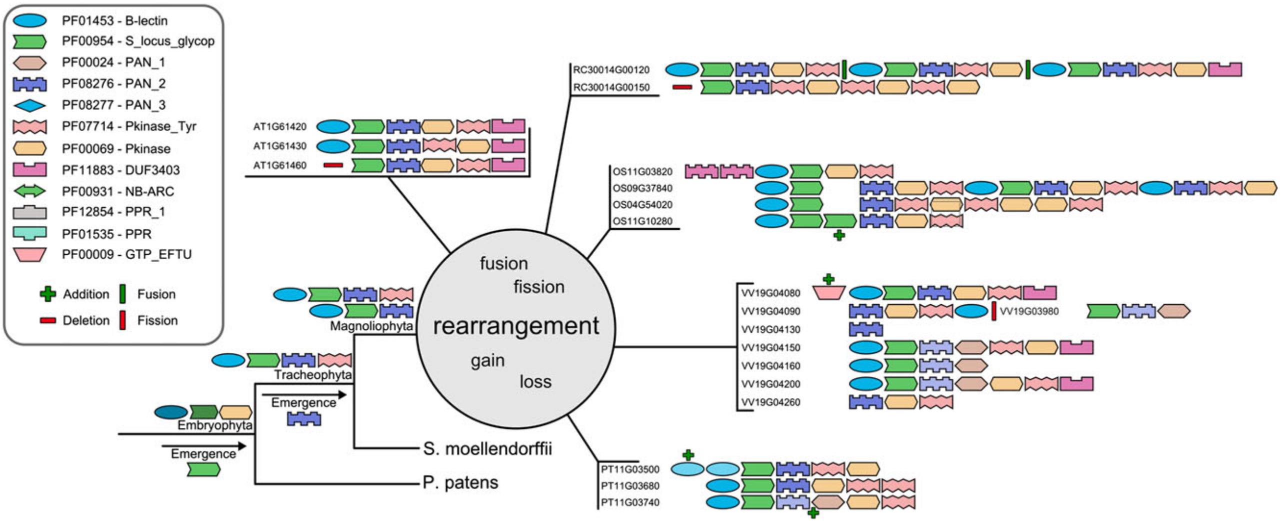

Fig. 5.—Example of two emergent domains at the Tracheophyta node (PAN_2) and Embryophyta node (S_locus_glycope). The evolution of example arrangements over time is shown in five different species (*Arabidopsis thaliana* [AT], *Oryza sativa* [OS], *Populus trichocarpa* [PT], *Ricinus communis* [RC], *Vitis vinifera* [VV]). The observable diversity in arrangements within this family is explainable by simple one-step events of fusion, fission, terminal deletion, and domain gain.

rates of events along more recent branches might be explained by WGD which have taken place in angiosperms (De Bodt et al. 2005; Freeling et al. 2006; Shoemaker et al. 2006; van de Peer et al. 2009; Paterson et al. 2010). Indeed, in a pairwise comparison of fusion and fission rates between plant pairs, which differ by one recent WGD, we find increased rates in plants with more recent WGD (supplementary table 7, Supplementary Material online). Roughly one-third of all vascular plants have undergone recent WGDs (Wood et al. 2009).

Arrangement Distribution

We investigated the distribution of shared arrangements among the plant species. The majority of domain arrangements are either species-specific or universal (fig. 3). This bimodal distribution is even stronger when we consider only a well-annotated subset of our species and exclude the green algae (supplementary fig. 8, Supplementary Material online). In particular, proteins with two or three domains are often species-specific. In combination with the observation that roughly 70% of all domain-containing proteins are single-domain proteins (supplementary fig. 4, Supplementary Material online), this can lead to the assumption that the fusion of single-domain proteins is a powerful mechanism to obtain species-specific proteins with new functionalities. This distribution suggests that only very few long arrangements are highly conserved; long arrangements are possibly more often affected by fission events. Proteins with arrangements shared by several but not all species are overrepresented in GO terms related to basic functions such as primary metabolism, cellulose biosynthetic process, and cell wall organization. In proteins with arrangements shared by a subset of between 5 and 24 proteomes, innate_immune_response is significantly overrepresented, suggesting that there might be different pathogens affecting different subclades. Proteins with GO terms related to reproduction, signal transduction, and prefixed with response_to are overrepresented in species-specific arrangements or those shared by only few species. The high number of species-specific arrangements observed here is in accordance with the observation that, within a set of five angiosperm species, around 20% of proteins do not align to an orthologous group (Paterson et al. 2010). The high amount of species-specific arrangements and genes might also be a consequence of frequent duplication events followed by lineage-specific retention (Paterson et al. 2010). This supports the hypothesis that plants have many flexible genetic mechanisms to produce species-specific adaptation (Bomblies 2010).

Gain and Loss of Domains and Arrangements

We investigated gain and loss at the levels of domains and domain arrangements by reconstructing the ancestral states based on maximum parsimony. We observe that gain and loss can frequently be found in all clades in plant evolution at both domain and arrangement levels. This is in agreement with Buljan and Bateman (2009), who found an equal event distribution after speciation and duplication within animals and a high amount of change in arrangements after duplication events. As we here do not conduct a direct comparison of paralogs, but instead compare presence/absence patterns of domains and their arrangements across proteomes, our results only support the notion that domain gain and loss can be found along all branches and that both have a significant correlation with each other and with branch length (fig. 1*A*). Branches with an increased loss rate have, on average, a higher domain gain rate. This high gain and loss rate, branch-specific variation and the large number of species-specific arrangements show the high variability and flexibility with which single-step mechanisms can create evolutionary novelties. One might speculate that the high gain rate of arrangements in *M. domestica* (supplementary table 4, Supplementary Material online) is caused by the recent polyploidy event or hybridization as consequence of domestication (Velasco et al. 2010). The large amount of domain loss in *P. dactylifera* might also be the consequence of low sequence coverage (Al-Dous et al. 2011). Differences in gain and loss between the different branches might also be a consequence of variation in generation time between plants. Evidence from studies in Fugu and Tetraodon suggests that intron loss is increased in species with shorter generation time (Loh et al. 2008). Similar patterns have been found in Arabidopsis and rice (Roy and Penny 2007).

Coverage

The average domain residue coverage is 50%. Protein coverage varies strongly even between closely related species (fig. 1*C*). Three plants belonging to the Fabaceae clade are included in this study, *G. max*, *L. japonica*, and *M. truncatula*. Their branches split around 50–60 Ma (Reineke et al. 2011). Several events of WGD have been found within the Fabaceae clade; all three species share a common WGD followed by additional independent WGDs (Blanc and Wolfe 2004). These WGDs in connection with different retention and pseudogenization rates might explain the variance in coverage within this clade. It is also possible that a number of plant-specific domains are still not yet described, as the number of sequenced plant genomes is still considerably lower than the currently available animal genomes. In the Fabaceae family, for example, a unique conserved disordered region has been described in sieve element occlusion genes (Ruping et al. 2010; Ernst et al. 2011). Many of these family-specific conserved functional sequences might be still not covered by Pfam. It should also be considered that genome quality between the investigated genomes varies, which might be the cause of differences in domain coverage; the most recently sequenced genomes exhibit low coverage in comparison to longer established genomes such as *M. truncatula* or *O. sativa* (fig. 1*C*).

Conclusions

The results presented here provide, from a phylogenomic perspective, multiple insights into the evolutionary dynamics of modular rearrangements and the potential adaptive benefits in plant genomes. Although around 70% of all proteins are single-domain proteins and a large fraction of these are shared by many species, we observe a very high volatility of novel domains and arrangements in general. Most strikingly, the majority of all arrangements is species-specific or restricted to a very small clade. Our phylogeny-based approach unravels that the majority of novel arrangements can be explained by single-step events such as fusion, fission, and terminal loss. Several events of accelerated activity of rearrangements and domain emergence could be associated to the respective changes in stress adaptation and morphogenesis. This is particularly pronounced for fusion in regulatory proteins. We thus observe a dominant effect of rearrangements on adaptation, which is partly driven by the high volatility of novel domains.

Taken together, this study illustrates another layer of complexity, which explains how modularity helps plants to both create and exploit their abundant genetic material in order to accomplish rapid adaptation in response to environmental challenges. We propose these results will fuel further large-scale experiments. Recent experiments in fungi using recombination of libraries of domains from signaling proteins (Peisajovich et al. 2010) and the expansion of domain repeats in self-recognition molecules (Chevanne et al. 2010) have already highlighted the enormous evolutionary potential of modularity in protein evolution. Along these lines, experiments on plant adaptation should be more explicitly geared at furthering our understanding of how protein modularity facilitates rapid adaptation.

Supplementary Material

Supplementary figures 1–9 and tables 1–7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Deutsche Forschungs Gemeinschaft, grant BO 2455/4-1 and by the Human Frontier Science Program (R6P0033/2006-C) to E.B.B.; A.D.M. acknowledges support by the WWU-PAS stipend. The authors would like to thank two anonymous reviewers for critical reading and helpful comments on the manuscript.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Akhunov ED, et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. 13:753–763.

Al-Dous EK, et al. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotechnol. 29:521–527.

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22:1600–1607.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Amoutzias GD, Weiner J, Bornberg-Bauer E. 2005. Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene 347:247–253.

Anderson CL, Janßen T. 2009. Monocots. In: Hedges S, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 203–212.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 310:311–325.

Argout X, et al. 2011. The genome of *Theobroma cacao*. Nat Genet. 43:101–108.

Ashburner M, Lewis S. 2002. On ontologies for biologists: the gene ontology—untangling the web. Novartis Found Symp. 247:66–80; discussion 80–90, 244–252.

Babushok D, Ostertag EE, Kazazian HH. 2007. Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci. 64:542–554.

Banks JA, et al. 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. Science 332: 960–963.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. 2009. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 19:243–254.

Baucom RS, et al. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the b73 maize genome. PLoS Genet. 5:e1000732.

Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. Plant Cell 23:1719–1728.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Biol Sci (Methodological). 57:289–300.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev. 15:621–627.

Bennici A. 2005. A fundamental plant evolutionary problem: the origin of land-plant sporophyte; is a new hypothesis possible? Riv Biol. 98:469–480.

Bhattacharya D, Yoon HS, Hedges SB, Hackett JD. 2009. Eukaryotes (eukaryota). In: Hedges S, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 116–120.

Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain rearrangements in protein evolution. J Mol Biol. 353:911–923.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Bomblies K. 2010. Doomed lovers: mechanisms of isolation and incompatibility in plants. Annu Rev Plant Biol. 61:109–124.

Bremer B, et al. 2009. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 161:105–121.

Buljan M, Bateman A. 2009. The evolution of protein domain families. Biochem Soc Trans. 37:751–755.

Buljan M, Frankish A, Bateman A. 2010. Quantifying the mechanisms of domain gain in animal proteins. Genome Biol. 11:R74.

Cartwright P, Collins A. 2007. Fossils and phylogenies: integrating multiple lines of evidence to investigate the origin of early major metazoan lineages. Integr Comp Biol. 47:744–751.

Chan AP, et al. 2010. Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol. 28:951–956.

Chevanne D, Saupe SJ, Clave C, Paoletti M. 2010. WD-repeat instability and diversification of the *Podospora anserina hnwd* non-self recognition gene family. BMC Evol Biol. 10:134.

Choi HK, et al. 2004. Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci U S A. 101:15289–15294.

Cohen-Gihon I, et al. 2011. Evolution of domain promiscuity in eukaryotic genomes—a perspective from the inferred ancestral domain architectures. Mol Biosyst. 7:784–792.

De Bodt S, Maere S, van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 20:591–597.

Dubchak I, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556.

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 6:197–208.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comp Biol. 7:e1002195.

Ekman D, Björklund AK, Elofsson A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol. 372:1337–1348.

Ekman D, Björklund AK, Frey-Skött J, Elofsson A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 348:231–243.

Endress PK. 2001. Origins of flower morphology. J Exp Zool. 291:105–115.

Enright AJ, Ouzounis CA. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol. 2:RESEARCH0034.

Ernst AM, et al. 2011. The sieve element occlusion gene family in dicotyledonous plants. Plant Signal Behav. 6:151–153.

Eucalyptus grandis Genome Project. 2010. [cited 2012 Jan 24]. Available from: http://www.phytozome.net/eucalyptus.

Finn RD, et al. 2008. The Pfam protein families database. Nucleic Acids Res. 38:D211–D222.

Fong JH, Geer LY, Panchenko AR, Bryant SH. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 366:307–315.

Forest F, Chase MW. 2009. Eurosid I. In: Hedges S, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 188–196.

Forslund K, Henricson A, Hollich V, Sonnhammer ELL. 2007. Domain tree based analysis of protein architecture evolution. Mol Biol Evol. 25:254–264.

Freeling M, Thomas BC, Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16:805–814.

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

Go SA, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp). Science 296:92–100.

Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420–3435.

Grace SC, Logan BA. 2000. Energy dissipation and radical scavenging by the plant phenylpropanoid pathway. Philos Trans R Soc Lond B Biol Sci. 355:1499–1510.

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.

Haploid Clementine Genome International Citrus Genome Consortium. 2011. [cited 2012 Jan 24]. Available from: http://www.phytozome.net/clementine.

Hedges SB, Dudley J, Kumar S. 2006. Timetree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. Proc Natl Acad Sci U S A. 106:3254–3258.

Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43:476–481.

Huang SW, et al. 2009. The genome of the cucumber, *Cucumis sativus*. Nat Genet. 41:1275–1281.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration. BMC Bioinformatics 11:24.

International Peach Genome Initiative. 2010. [cited 2012 Jan 24]. Available from: http://www.phytozome.net/peach.

Jaillon O, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.

Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Kawakita A, Kato M. 2009. Repeated independent evolution of obligate pollination mutualism in the phyllantheae-epicephala association. Proc Biol Sci. 276:417–426.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends Ecol Evol. 24:572–582.

Kerchev PI, Fenton B, Foyer CH, Hancock RD. 2012. Plant responses to insect herbivory: interactions between photosynthesis, reactive oxygen species and hormonal signalling pathways. Plant Cell Environ. 35:441–453.

Kersey P, et al. 2005. Integr8 and genome reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. 33:D297–D302.

Kroymann J. 2011. Natural diversity and adaptation in plant secondary metabolism. Curr Opin Plant Biol. 14:246–251.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21:25–30.

Lang D, et al. 2010. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol Evol. 2:488–503.

Levitt M. 2009. Nature of the protein universe. Proc Natl Acad Sci U S A. 106:11079–11084.

Loh YH, Brenner S, Venkatesh B. 2008. Investigation of loss and gain of introns in the compact genomes of pufferfishes (fugu and tetraodon). Mol Biol Evol. 25:526–535.

Ma LJ, et al. 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet. 5:e1000549.

Magallón SA, Sanderson MJ. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution 59:1653–1670.

Mayrose I, et al. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257.

Merchant SS, et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245–250.

Ming R, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya linnaeus*). Nature 452:991–996.

Misook H, Wen-Hsiung L, Chen CJ. 2007. External factors accelerate expression divergence between duplicate genes. Trends Genet. 23:162–166.

Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. Trends Biochem Sci. 33:444–451.

Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol. 29: 787–796.

Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61(Suppl 7):176–182.

Olson JM. 1970. The evolution of photosynthesis. Science 168:438–446.

Olson JM. 2001. 'Evolution of photosynthesis' (1970), re-examined thirty years later. Photosynth Res. 68:95–112.

Osborn TC, et al. 2003. Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19:141–147.

Oshima A, et al. 2010. Asymmetric configurations and N-terminal rearrangements in connexin26 gap junction channels. J Mol Biol. 405:724–735.

Palenik B, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 104:7705–7710.

Pasek S, Risler JL, Brézellec P. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics 22:1418–1423.

Pateraki I, Kanellis AK. 2010. Stress and developmental responses of terpenoid biosynthetic genes in *Cistus creticus* subsp. *creticus*. Plant Cell Rep. 29:629–641.

Paterson AH, Freeling M, Tang H, Wang X. 2010. Insights from the comparison of plant genome sequences. Annu Rev Plant Biol. 61:349–372.

Peisajovich SG, Garbarino JE, Wei P, Lim WA. 2010. Rapid diversification of cell signaling phenotypes by modular domain recombination. Science 328:368–372.

Proost S, et al. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell 21:3718–3731.

Reineke AR, Bornberg-Bauer E, Gu J. 2011. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res. 39:6029–6043.

Rensing SA, Lang D, Zimmer AD, et al. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69.

Rogers RL, Hartl DL. 2012. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. Mol Biol Evol. 29:517–529.

Roy SW, Penny D. 2007. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. Mol Biol Evol. 24:171–181.

Ruping B, et al. 2010. Molecular and phylogenetic characterization of the sieve element occlusion gene family in *Fabaceae* and non-*Fabaceae* plants. BMC Plant Biol. 10:219.

Schmutz J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Schnable PS, et al. 2009. The b73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.

Setaria italica Genome Sequencing Project. 2011. These sequence data were produced by the US Department of Energy Joint Genome Institute [Internet]. [cited 2012 Jan 24] Available from: http://www.phytozome.net/foxtailmillet.

Shan H, et al. 2009. Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. Mol Biol Evol. 26:2229–2244.

Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. Curr Opin Plant Biol. 9:104–109.

Siegel S, Castellan N Jr. 1988. Nonparametric statistics for the behavioral sciences, 2nd ed. Boston: McGraw-Hill Humanities.

Söding J, Lupas AN. 2003. More than the sum of their parts: on the evolution of proteins from peptides. Bioessays 25:837–846.

Soltis DE, Soltis PS, Zanis MJ. 2002. Phylogeny of seed plants based on evidence from eight genes. Am J Bot. 89:1670–1681.

Srivastava M, et al. 2008. The *Trichoplax* genome and the nature of placozoans. Nature 454:955–960.

Sweet Orange Genome Project. 2010. [cited 2012 Jan 24]. Available from: www.phytozome.net/citrus.

Tang H, et al. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 18:1944–1954.

The Arabidopsis Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Tuskan GA, et al. 2006. The genome of black cottonwood, *Populus trichcarpa* (Torr. & Gray). Science 313:1596–1604.

van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–732.

Velasco R, et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet. 42:833–839.

Veron AS, Kaufmann K, Bornberg-Bauer E. 2007. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. Mol Biol Evol. 24:670–678.

Vogel JP, et al. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768.

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

Wang M, Caetano-Anolles G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure 17:66–78.

Weiner J, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. FEBS J. 273:2037–2047.

Weiner J, Moore AD, Bornberg-Bauer E. 2008. Just how versatile are domains? BMC Evol Biol. 8:285.

Williams JH. 2008. Novelties of the flowering plant pollen tube underlie diversification of a key life history stage. Proc Natl Acad Sci U S A. 105:11259–11263.

Wood TE, et al. 2009. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A. 106:13875–13879.

Worden AZ, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324:268–272.

Yang X, Jawdy S, Tschaplinski TJ, Tuskan GA. 2009. Genome-wide identication of lineage-specific genes in *Arabidopsis, Oryza* and *Populus*. Genomics 93:473–480.

Young ND, et al. 2005. Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol. 137:1174–1181.

Zheng C, et al. 2009. Gene loss under neighborhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* genome. J Bioinform Comput Biol. 7:499–520.

Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12:R4.

Associate editor: Yves van de Peer