Molecular Evolution Constraints in the Floral Organ Specification Gene Regulatory Network Module across 18 Angiosperm Genomes

Jose Davila-Velderrain,[1,2] Andres Servin-Marquez,[3] and Elena R. Alvarez-Buylla*[,1,2]

[1]Instituto de Ecología, Universidad Nacional Autónoma de México, México, D.F., México

[2]Centro de Ciencias de la Complejidad, C3, Universidad Nacional Autónoma de México, México, D.F., México

[3]Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, México

***Corresponding author:** E-mail: eabuylla@gmail.com.

Associate editor: Michael Purugganan

Abstract

The gene regulatory network of floral organ cell fate specification of *Arabidopsis thaliana* is a robust developmental regulatory module. Although such finding was proposed to explain the overall conservation of floral organ types and organization among angiosperms, it has not been confirmed that the network components are conserved at the molecular level among flowering plants. Using the genomic data that have accumulated, we address the conservation of the genes involved in this network and the forces that have shaped its evolution during the divergence of angiosperms. We recovered the network gene homologs for 18 species of flowering plants spanning nine families. We found that all the genes are highly conserved with no evidence of positive selection. We studied the sequence conservation features of the genes in the context of their known biological function and the strength of the purifying selection acting upon them in relation to their placement within the network. Our results suggest an association between protein length and sequence conservation, evolutionary rates, and functional category. On the other hand, we found no significant correlation between the strength of purifying selection and gene placement. Our results confirm that the studied robust developmental regulatory module has been subjected to strong functional constraints. However, unlike previous studies, our results do not support the notion that network topology plays a major role in constraining evolutionary rates. We speculate that the dynamical functional role of genes within the network and not just its connectivity could play an important role in constraining evolution.

Key words: **gene regulatory network, flower development, molecular evolution, functional constraint.**

Article

Introduction

An outstanding goal in molecular evolution is to bridge the gap between the study of individual molecules and the study of systems on higher levels of biological organization. In modern evolutionary studies, the limitations of considering genes as individual entities upon which evolutionary forces act independently are becoming generally accepted. The emerging picture is that in which evolutionary forces, functional constraints, and molecular interactions are conditionally dependent on the systems level (Cork and Purugganan 2004). Following this line of research, several studies have analyzed molecular evolution at the pathway or network level (see, e.g., Hahn et al. 2004; Alvarez-Ponce et al. 2009; Jovelin and Phillips 2009; Yang et al. 2009; Montanucci et al. 2011; Alvarez-Ponce 2012). Most studies support the idea that evolutionary forces acting on genes are in close relation with the structure/topology of their functional network.

Previous network-based molecular evolutionary studies have focus on investigating networks in relation to the evolutionary rates of their genes based on large-scale molecular networks (Fraser et al. 2002; Agrafioti et al. 2005; Hahn and Kern 2005; Lemos et al. 2005; Alvarez-Ponce and Fares 2012). Recently, similar analysis have been applied to well-characterized, relatively small pathways (Alvarez-Ponce et al. 2009, 2011; Casals et al. 2011; Fitzpatrick and O'Halloran 2012; Lavagnino et al. 2012; Invergo et al. 2013). Both approaches have uncovered interesting yet preliminary patterns (see Montanucci et al. 2011 and references therein). The conclusion, so far, appears to be that evolutionary pressures acting on genes are in close relation with the structure of their functional network. But contrasting results have been found in several cases, and when considering the latter, there is no general consensus for the relationship between network properties and the molecular evolution of its components: different patterns have been found for different interacting systems and different species sets. Thus, the need for resolution of contrasting results and the search of robust evolutionary patterns call for new studies. It has been suggested that the analysis of new pathways might help to uncover general patterns and to disentangle topological restrictions of networks from the biological properties and functions (Montanucci et al. 2011). Here, we argue that the study of the molecular evolution of the genes involved in regulatory modules that have been uncovered with dynamical gene regulatory network (GRN) models could help uncover general evolutionary principles, given that such models allow a

rigorous distinction between structure and function. In contrast to schematic representations that depict gene regulatory interactions, dynamic models may consider the nonlinear aspects of regulation and explore the way gene expression changes in time, both in wild-type and perturbed simulated systems (Alvarez-Buylla et al. 2010). Nevertheless, to the best of our knowledge, a network-based molecular evolutionary study is lacking for the case of experimentally grounded and functionally validated dynamic GRN models.

It is generally accepted that GRNs are underlying molecular systems orchestrating developmental processes (Huang and Kauffman 2009; Alvarez-Buylla et al. 2010). On the other hand, it has been suggested that the specific nature of evolutionary forces acting on the component genes depends largely on the function of the interacting system (Cork and Purugganan 2004). In this work, we follow a similar approach to that of previous network-level evolutionary studies; but instead of analyzing a new metabolic pathway, we focus on the molecular evolution and network properties of a well-studied GRN module: the experimentally grounded floral organ cell fate specification determination GRN (FOS-GRN) (see Espinosa-Soto et al. 2004; Alvarez-Buylla et al. 2010 for updates).

The FOS-GRN (fig. 1) integrates molecular genetic data for the ABC genes and their main interactors in *A. thaliana*. This GRN includes key regulators underlying the transition from the shoot apical meristem once it produces the apical inflorescence meristem with the flower primordia in its flanks (flowering locus t [*FT*], terminal flower1 [*TFL*], embryonic flower1 [*EMF1*], LEAFY [*LFY*], APETALA1 [*AP1*], fruitfull [*FUL*]), the ABCs and some of their interacting genes (APETALA1 [*AP1*], APETALA3 [*AP3*], PISTILLATA [*PI*], APETALA2 [*AP2*], AGAMOUS [*AG*], SEPALLATA [*SEP*]), as well as some genes that link floral organ specification to other modules regulating primordia formation and homeostasis (*AG* and *WUS*) and to some regulators of organ boundaries (*UFO*). From the 15 genes, 6 are members of the MADS-box protein family (*AG*, *AP1*, *AP3*, *PI*, *SEP*, *FUL*) and belong to five different subfamilies (AG, SQUA, GLO, DEF, and AGL2) within the clades of MADS-box genes (Becker and Theissen 2003).

The model was proposed on the basis of experimental data for these 15 genes in the model plant *A. thaliana*. Among the 15 genes, 5 are grouped into three classes (A-type, B-type, and C-type) whose combinations, described by the ABC model, are necessary for floral organ cell specification (Coen and Meyerowitz 1991). A-type genes (*AP1* and *AP2*) are necessary for sepal specification, A-type together with B-type (*AP3* and *PI*) for petal specification, B-type and C-type (*AGAMOUS*) for stamen specification, and the C-type gene (*AG*) alone for carpel primordia cell specification. Although the ABC model of flower development was published more than 20 years ago, it was just recently that the model of the FOS-GRN provided a sufficient explanation for the observed ABC patterns and the stable gene expression configurations observed during early flower development in *Arabidopsis* (Mendoza and Alvarez-Buylla 1998; Espinosa-Soto et al. 2004; and updates and review in Alvarez-Buylla et al. 2010). The network

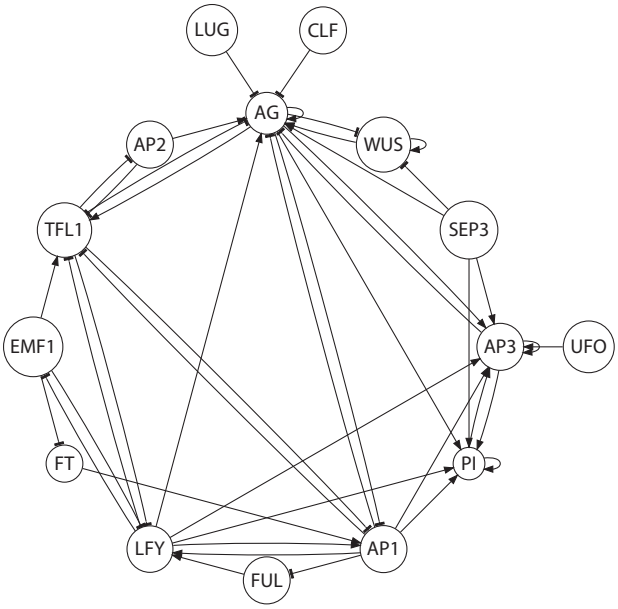


Fig. 1. Graph representation of the FOS-GRN. Arrows and blunt-ended edges correspond to activating and repressing interactions, respectively.

has been studied from different perspectives (Alvarez-Buylla et al. 2008; Sanchez-Corrales et al. 2010; Villarreal et al. 2012), and the results of multiple studies have shown that its dynamical behavior is robust enough as to predict the observed phenotypes both in wild-type and several mutant conditions. In other words, there is enough evidence to sustain the claim that the 15 genes involved in the network form a core regulatory module responsible for primordial cell fate determination during early stages of flower development. We reasoned that such a functional constraint could play a strong role in constraining evolutionary rates at the molecular level. Based on this idea, here we addressed whether orthologous genes of the FOS-GRN were found and conserved in distantly related angiosperm species, and then we addressed the evolutionary forces that could have shaped its evolution under the hypothesis that positive Darwinian selection would not be a prevailing force.

A large number of the genes involved in floral development belong to the eukaryotic MADS-box gene family (Riechmann et al. 1997). Most studies on the molecular basis of floral development focus on these genes, particularly floral homeotic genes such as *AGAMOUS* (*AG*), *APETALA3* (*AP3*), *PISTILLATA* (*PI*), and several *AGAMOUS*-like genes (Lawton-Rauh et al. 2000). Background information on genetic and expression analyses indicate that members of a floral homeotic gene group tend to share similar developmental functions in flower and inflorescence morphogenesis (Purugganan et al. 1995; Purugganan 1997), thus reflecting high conservation among evolutionarily related regulatory genes. Previous studies on the evolutionary forces acting on some of the genes involved in flower development have focused on intraspecific population genetics data (Purugganan and Suddith 1999) or data from two closely related species (Yang et al. 2011). These studies have shown that although most floral genes have evolved under strong purifying

selection, some show elevated nonsynonymous substitution rates and/or positively selected sites. However, given that these molecular evolutionary studies have focused mostly on closely related species, it is not known whether the complete set of genes conforming the FOS-GRN are globally conserved among flowering plants. In order to first explore this possibility, here we follow a comparative genomics approach, and, unlike previous work, we study the molecular evolution of the network over a broad taxonomic distance involving monocots and dicots; the recent completion, annotation, and analysis of the genomes of several flowering plant species has provided the opportunity to do so.

In summary, the aim of this work was 3-fold: 1) to explore the degree of conservation of the genes involved in the FOS-GRN, 2) to uncover the prevailing molecular evolutionary forces acting upon its genes, and 3) to study the evolutionary constraints that its network properties and known biological function impose to the molecular evolution of its components. With this in mind, we first searched for the homologs of the genes in the *A. thaliana* FOS-GRN in all the flowering species with a sequenced and annotated genome available (a total of 18; see fig. 2 for the species used and their placement in angiosperm phylogeny). With the sequence data for the FOS-GRN genes, we measured the action of selective pressures on individual protein-coding genes through the estimation of synonymous and nonsynonymous substitution rates (dS and dN, respectively) when comparing among species. The ratio dN/dS measures the strength and nature of the evolutionary forces indicating positive selection, neutral evolution, or purifying selection when it is higher, equal, or lower than 1, respectively. Both an overall ratio for the entire coding sequence of a gene and estimates considering variation of the ratio among sites were calculated (Yang and Bielawski 2000). We then calculated molecular conservation features other than evolutionary rates for each gene and asked whether these features in addition to the evolutionary parameters (dN, dS, dN/dS) show a pattern of association with the known biological functions of the genes. Finally, we addressed whether the forces that have shaped the evolution of the genes during the divergence of angiosperms were correlated to the placement of each gene within the FOS-GRN.

Results

Identification of the FOS-GRN Genes in Flowering Plant Genomes

The experimentally grounded FOS-GRN proposed by Espinosa-Soto et al. was used as a reference (Espinosa-Soto et al. 2004; and updated in Alvarez-Buylla et al. 2010). The original network proposed for *A. thaliana* has 15 genes and their regulatory (activating or inhibitory) interactions (supplementary table S1, Supplementary Material online). In order to study the conservation of the genes in the network across species, we conducted homology analysis using the Plaza Comparative Genomics Platform (Proost et al. 2009) (see Materials and Methods). For each gene in the network, a total of 418 putative homologs (orthologs and in-paralogs) of the 15 *A. thaliana* (Ath) FOS-GRN genes were identified in the genomes of the other 17 flowering plant species: *Arabidopsis lyrata* (Aly), *Brachypodium distachyon* (Bdi), *Carica papaya* (Cpa), *Fragaria vesca* (Fve), *Glycine max* (Gma), *Lotus japonicus* (Lja), *Malus domestica* (Mdo), *Manihot esculenta* (Mes), *Medicago truncatula* (Mtr), *Oryza sativa japonica* (Osj), *Oryza sativa indica* (Osi), *Populus trichocarpa* (Ptr), *Ricinus communis* (Rco), *Sorghum bicolor* (Sbi), *Theobroma cacao* (Tca), *Vitis vinifera* (Vvi), and *Zea mays* (Zma) (see fig. 2). These results correspond to the preliminary network conservation data and were organized in the form of a conservation matrix (also called phylogenetic profile) where each row represents a gene vector composed by a set of characters {0, 1, 2, 3, 4} representing the absence (0), presence (1), or the total number of in-paralogs (2, 3, 4) of each gene; and each column represents a species (supplementary table S2, Supplementary Material online). All FOS-GRN genes studied, with the exception of *EMF1*, have orthologs in all 18 genomes. The gene *EMF1* was not found as an ortholog of the *EMF1* gene in *A. thaliana* (AT5G11530) among the monocot plants: *B. distachyon*, *O. sativa japonica*, *O. sativa indica*, *S. bicolor*, and *Z. mays*. However, following the same methodology, but using instead the corresponding protein sequence of the gene *EMF1* reported for *O. sativa* (OS01G12890) as query, putative orthologs were found in all four cases. For the only case of this gene (*EMF1*), it was discovered that there exists one orthologous group for dicots and a different group for monocots. The relationship between both groups is not clear and will be studied in subsequent studies.

Manual Curation of Putative In-paralogs

The preliminary conservation data of the proteins in the FOS-GRN of *A. thaliana* were manually curated to produce the final conservation data of the proteins in the FOS-GRN reported here in the form of a conservation matrix (fig. 2) and the corresponding list of gene IDs (supplementary table S4, Supplementary Material online). Certain proteins were eliminated from the list due to evidence of partial gene copies or annotation errors (see Materials and Methods). We found that all the FOS-GRN genes have homologs in all 18 genomes searched. Results also show that all the genes underwent a number of duplication and/or loss events. The detailed evolutionary processes (e.g., duplication, loss, and pseudogenization) leading to the expansion of the network across angiosperms will be explored in a future study.

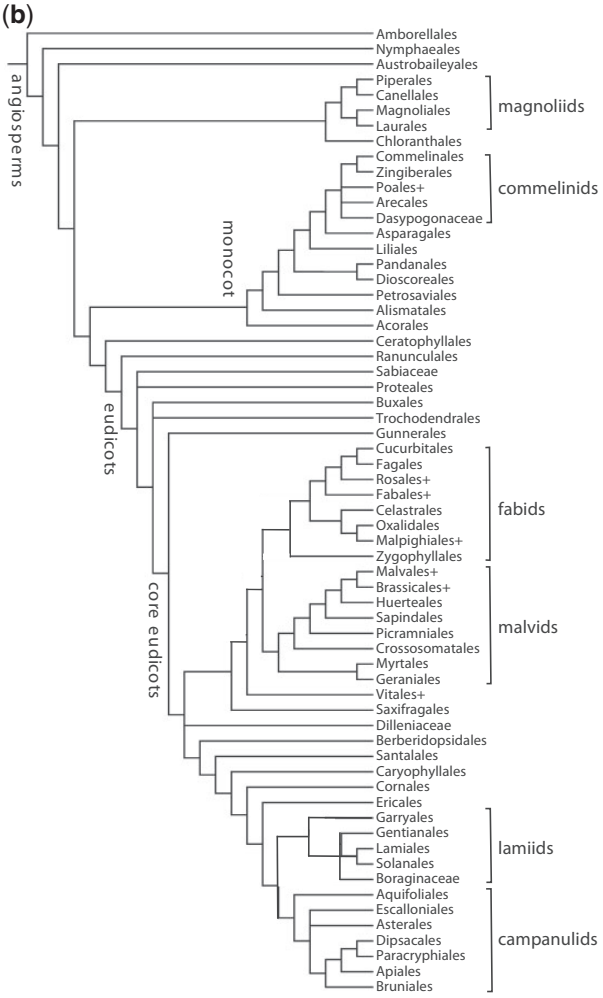
Molecular Evolutionary Analysis of the FOS-GRN

The nonsynonymous (dN) to synonymous (dS) substitution rate ratio (dN/dS) was calculated in order to infer the impact of natural selection on the FOS-GRN. The values of the overall ratio dN/dS range from 0.05936 for *PI* to 0.39577 for *EMF1*, suggesting that purifying selection or selection constraint best explains the evolution of the genes in the FOS-GRN (table 1). Given that the estimation of an overall dN/dS for the whole coding sequence is a very conservative measure of positive selection (Yang and Bielawski 2000), estimates that account for variation in dN/dS among sites in order to detect specific sites that could have been fixed by positive selection were also

(a)

Gene	Aly	Ath	Bdi	Fve	Gma	Lja	Mdo	Mes	Mtr	Osj	Osi	Ptr	Rco	Sbi	Tca	Vvi
AG	1	1	2	1	2	2	2	1	1	2	1	1	1	2	1	1
AP1	1	1	2	1	2	1	2	2	2	2	1	2	1	2	1	1
AP2	1	1	1	1	4	1	4	2	1	1	1	2	2	1	2	2
AP3	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
CLF	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1
EMF1	1	1	1*	1	1	1	3	1	2	1*	1*	2	1	1*	1	1
FT	1	1	2	1	2	1	2	1	1	3	2	3	4	2	1	1
FUL	1	1	1	1	3	1	2	2	1	1	1	1	1	1	1	1
LFY	1	1	1	3	2	1	2	2	1	1	1	1	1	1	1	1
LUG	1	1	3	2	4	1	1	1	2	1	2	2	1	2	2	1
PI	1	1	2	1	4	1	1	2	1	2	1	2	1	2	1	1
SEP	1	1	1	1	2	1	2	1	2	1	1	2	1	1	1	1
TFL1	1	1	1	1	2	1	2	3	1	2	2	2	1	2	2	1
UFO	1	1	1	2	2	1	2	1	1	1	1	2	1	1	1	1
WUS	1	1	1	3	2	2	5	3	2	1	1	4	1	1	2	3

(b)



(c)

Species	APG III
Carica papaya (Cpa)	Brassicales
Arabidops is thaliana (Ath)	Brassicales
Arabidops is lyrata (Aly)	Brassicales
Manihot es culenta (Mes)	Malpighiales
Glycine max (Gma)	Fabales
Lotus japonicus (Lja)	Fabales
Medicago truncatula (Mtr)	Fabales
Populus trichocarpa (Ptr)	Malpighiales
Ricinus communis (Rco)	Malpighiales
Theobroma cacao (Tca)	Malvales
Brachypodium distachyon (Bdi)	Poales
Oryza sativa japonica (Osj)	Poales
Oryza sativa indica (Osj)	Poales
Sorghum bicolor (Sbi)	Poales
Zeamays (Zma)	Poales
Fragaria vesca (Fve)	Rosales
Malus domestica (Mdo)	Rosales
Vitis vinifera (Vvi)	Vitales

Fig. 2. Gene conservation data, species used, and their placement in Angiosperm phylogeny. (*a*) Conservation matrix of the genes involved in the FOS-GRN across Angiosperm species (*Genes were identified using *Oryza sativa* EMF1 protein (OS01G12890) for homology search; $^{+}$Families considered in the analysis). (*b*) Angiosperms phylogeny APG III according to Bremer et al. (2009). (*c*) Species used in the analysis.

calculated. Results showed that the genes *UFO*, *FT*, and *CLF* yielded a marginal significant *P* value when comparing the model M8 assuming positive selection with the null model M7 of the program CODEML (see Materials and Methods). However, the test was no longer significant after correcting for multiple comparisons. For all 15 genes, the models M2a was not significantly better than the null model M1a (supplementary table S5, Supplementary Material online). The overall dN/dS, dN, and dS were computed for each gene under the M0 model (table 1). The genes of the FOS-GRN are subject to strong purifying selection with an overall mean dN/dS of 0.124. Overall dN/dS values are plotted in figure 3; from the 15 genes, 13 (86.66%) have a dN/dS value <0.15.

Table 1. Evolutionary Parameters of the FOS-GRN Genes.

Gene	Locus	Protein Length	Percent of Analized Codons	dN	dS	dN/dS
AP1	AT1G69120	256	89	0.7683	6.1525	0.12487
AP2	AT4G36920	432	80	0.6095	5.6578	0.10773
AP3	AT3G54340	232	93	0.7713	8.0723	0.09555
CLF	AT2G23380	902	67	0.6369	5.386	0.11824
EMF1	AT5G11530	1096	66	3.8105	9.6281	0.39577
FT	AT1G65480	175	99	0.4509	5.9891	0.07529
FUL	AT5G60910	242	95	0.8261	7.0456	0.11725
LFY	AT5G61850	420	83	0.715	8.5201	0.08392
LUG	AT4G32551	931	78	0.5995	5.2033	0.11522
PI	AT5G20240	208	58	0.602	10.1413	0.05936
SEP	AT1G24260	250	90	0.6172	7.9816	0.07733
TFL1	AT5G03840	177	97	0.5	5.973	0.0837
UFO	AT1G30950	442	84	0.9109	8.4945	0.10723
WUS	AT2G17950	292	51	2.615	12.799	0.20431

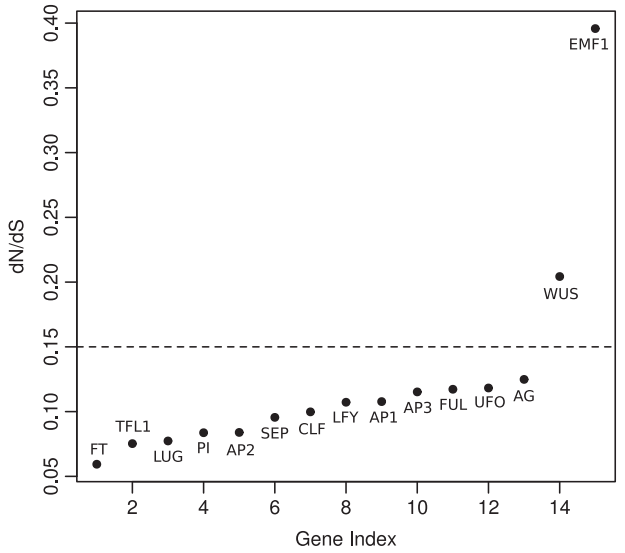


Fig. 3. Calculated dN/dS values sorted in increasing order. The horizontal dotted line is plotted to show that, from the 15 genes, 13 (86.66%) have a dN/dS value <0.15. Plotted values were calculated using the M0 model.

Analysis of the Classes of Genes

To test whether the measures of dN/dS, dN, or dS were statistically different between two gene classes, the ABC genes and the additional genes in the network, a Kruskal–Wallis test was performed. Although the genes *EMF1* and *WUS* showed higher dN/dS values than the 86.66% of the genes, the test gave no significant differences in dN/dS, dN, or dS between the classes. Means and *P* values are shown in supplementary tables S6 and S7, Supplementary Material online.

Model-Based Clustering of Sequence Conservation Features

During initial exploratory data analysis, it was observed that protein and DNA coding sequences of some of the genes in the FOS-GRN across the angiosperm species show interesting patterns in measures of conservation other than the evolutionary parameters (supplementary figs. S1–S7, Supplementary Material online). The following conservation features were calculated (see Materials and Methods): the degree of variability in protein size of each protein over all species (measured by the coefficient of variation), mean protein pairwise sequence distances, mean protein sequence distance, and mean DNA sequence distance (table 2). Given such data, the following question raised: is there an association between such conservation patterns and the functional classification of the proteins in the network?

In order to explore this possibility, a model-based clustering analysis was applied. Clustering is the process of grouping similar objects together. Here, a feature-based clustering approach was used, in which an $N \times D$ feature matrix is used as input (Murphy 2012). A feature matrix was assembled where each of the N rows represents a particular gene and the D columns corresponded to the conservation features listed above, together with an additional column corresponding to the dN/dS data (table 2). In other words, each row represents a conservation feature vector for each gene. This analysis does not make any assumption about the prior known functional category of the genes. Instead, it divides the genes into clusters according to the similarity among their feature vectors. The analysis was restricted to include all but the *EMF1* and *WUS* genes: when all the genes were included, an additional cluster was invariably obtained for each of the two genes (*EMF1* and *WUS*) given their high dN/dS and interspecies sequence distances (data not shown). Interestingly, the methodology uncovered three clusters (fig. 4): one corresponding the genes *AG*, *AP1*, *AP2*, and *PI* (circles); one for the genes *FUL*, *LFY*, *UFO*, and *AP3* (triangles); and the last one to the additional genes *CLF*, *FT*, *LUG*, *SEP*, and *TFL1* (squares). The four genes in the first cluster correspond to ABC floral organ identity genes. While the genes in the second cluster, except *AP3*, floral meristem identity genes (Krizek and Fletcher 2005). These results suggest an association between molecular size and sequence conservation features, evolutionary rates, and functional category. Those genes with a well-characterized function

Table 2. Gene Conservation Features.

Gene	No. Protein Sequences	Protein Size (CV)	Protein Mean Distance	DNA Codon Mean Distance	dN/dS
AG	30	0.241130802	0.437267	0.4316755	0.09981
AP1	30	0.210350092	0.4575694	0.4514316	0.12487
AP2	30	0.095096262	0.4422398	0.4184613	0.10773
AP3	22	0.099160315	0.4942555	0.4455772	0.09555
CLF	22	0.354842355	0.4084511	0.3709341	0.11824
EMF1	28	0.323552388	1.742475	1.2504943	0.39577
FT	31	0.254844679	0.2390374	0.3548233	0.07529
FUL	22	0.18187526	0.464033	0.4358368	0.11725
LFY	25	0.182826509	0.4513709	0.4390543	0.08392
LUG	37	0.236524789	0.3276286	0.3408934	0.11522
PI	29	0.184321532	0.4537865	0.3875463	0.05936
SEP	25	0.19888967	0.3333186	0.3837757	0.07733
TFL1	27	0.009807079	0.2475236	0.324812	0.0837
UFO	23	0.037201769	0.6089123	0.5777019	0.10723
WUS	36	0.17482874	1.1723443	0.8080616	0.20431

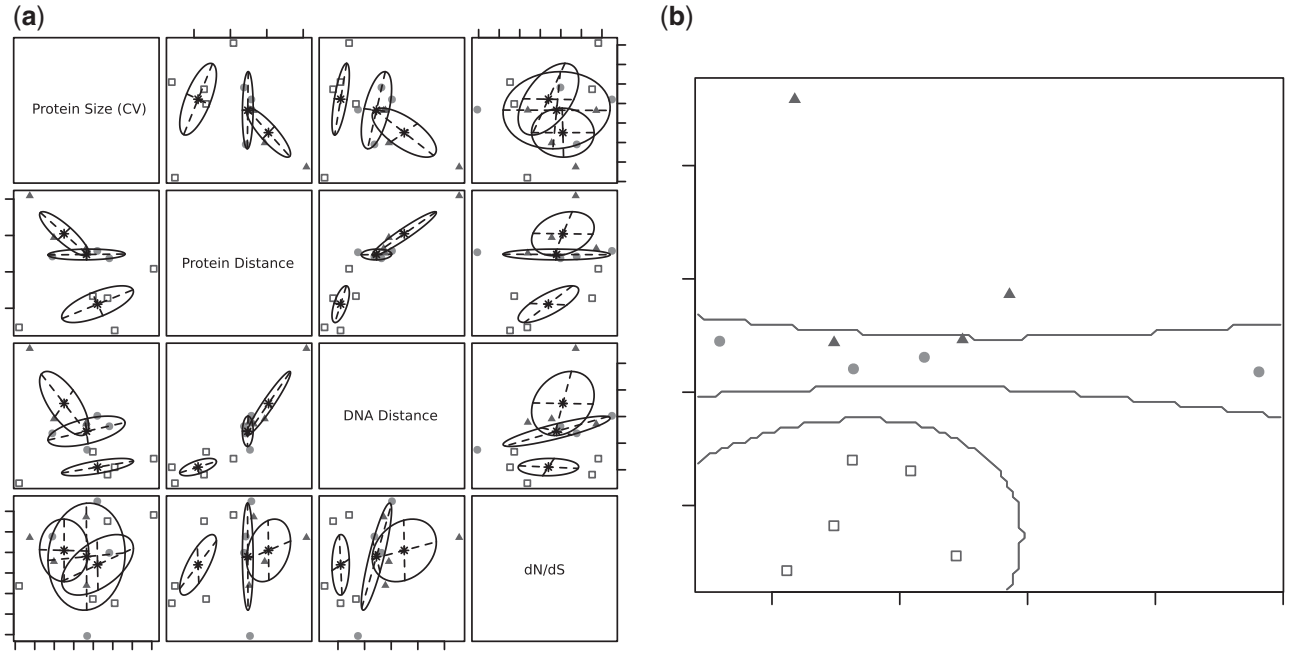


Fig. 4. Output from the model-based clustering analysis. (*a*) Scatterplot matrix for conservation features with points (genes) marked according to the corresponding cluster; the ellipses shown are the multivariate analogs of the standard deviations for each mixture component. (*b*) Data projection on a dimension reduced subspace. Clustering structure and boundaries are shown; genes are marked according to the corresponding cluster.

(e.g., a direct involvement in the processes of floral or meristem identity) share more similar conservation features among them than with the additional interacting genes which are known to be involved in various processes. Genes in the last cluster are known to integrate the flowering process with upstream signaling mechanisms and either promote (e.g., *FT*) or inhibit (*TFL1*) flower organ development. Figure 4*b* shows a two-dimensional projection of the feature vector along with the corresponding classification boundaries; it is interesting to note that the boundaries between the meristem (triangles) and flower identity (circles) clusters merge and are clearly separated from the third cluster (squares). This is consistent with the known biological mechanisms where genes such as *AP1* participate as both meristem and floral organ identity genes.

Given that clustering is an unsupervised learning technique, it is hard to evaluate the quality of the output on any given method. One way to do so is to rely on some external form of data with which to validate the method. In the case in point, labels representing functional categories can be assigned to each gene. Each gene was labeled with one of the three categories: floral organ identity, floral meristem identity, and other. The clustering was then compared with the labels using a standard metric: the Rand index (see Materials and Methods). This metric was calculated for the output of the clustering. Then, its statistical significance was

assessed through their frequency sampling distribution computed using a bootstrap resampling method (Murphy 2012). The observed clustering decisions are highly significant (P value = 0.0002). Thus, there is statistical support for an association between the molecular conservation features, the evolutionary rates, and the functional category of the genes in the FOS-GRN.

The Strength of the Purifying Selection and Network Structure

Each node in the network was characterized by a set of features including the molecular evolutionary parameters (dN, dS, and dN/dS) and its placement within the network topology, using measures such as centrality, degree, closeness, betweenness, and eccentricity (see Materials and Methods). GRNs contain directed interactions with either an inhibitory or an activating character. Given that the dynamical behavior of GRNs is associated with the type of interactions within the network, the topological network properties, out-degree, in-degree, activating in-degree, and inhibitory in-degree, were also included as features (supplementary table S8, Supplementary Material online). Once the evolutionary parameters and the network topological features were calculated, the goal was to answer the following questions: 1) Is there a relationship between the evolutionary parameters and the network nodes topological location within the FOS-GRN? 2) How strong is the relationship found, if any? 3) Which network topological features contribute the most to evolutionary rates?

A relationship between each of the evolutionary parameters and each of the node's topological features within the FOS-GRN was tested. Assuming an approximately linear relationship, model coefficients were estimated independently for each of the networks' topological features as single predictor variables of the evolutionary parameters. Hypothesis tests on the coefficients were performed in order to test whether or not there is a relationship between the variables in each case. Mathematically, this corresponds to testing whether the corresponding coefficient is equal to 0 or not. Details of the least squares models for the regression of dN/dS on each of the topological features used are provided in supplementary table S10, Supplementary Material online. Interestingly, the null hypothesis that the coefficient is equal to 0 could not be rejected for any case; consequently, a relationship between the dN/dS and any of the networks topological features tested could not be declared to exist, given the available data. The same analysis was applied individually to dN and dS as response variables. Only a marginal significant relationship (P value ~0.05) was found between dS and closeness. In a preliminary analysis, Spearman's rank correlation coefficients between the evolutionary parameters and the topological network properties were also calculated and are reported in supplementary table S9, Supplementary Material online. No significant correlation was found between the measures of centrality and the evolutionary estimates.

Similarity in Evolutionary Parameters of Interacting Genes

It has been suggested that interacting elements within a network share more similar values of evolutionary parameters within themselves than with noninteracting components (Alvarez-Ponce et al. 2009). In order to test whether this pattern is present in the FOS-GRN, two different approaches were applied: 1) the average absolute difference (AAD) of the value of the evolutionary parameters between interacting components in the networks was used as an statistic and compared with its null distribution in an ensemble of similar but random networks (Alvarez-Ponce et al. 2009), and 2) a matrix of pairwise shortest path distances between the genes in the network was compared with the matrices of pairwise absolute differences in evolutionary parameters (Montanucci et al. 2011). Using the former approach, an AAD of dN/dS of 0.0567 was calculated for the FOS-GRN. The histogram of the corresponding statistic on an ensemble of 100,000 random networks with the same number of nodes and interactions is shown in supplementary figure S8, Supplementary Material online. The simulated data follow closely a Gaussian distribution. The obtained data were used to estimate the probability of observing such a small value. Two approaches were followed: 1) calculating the fraction of random networks showing an AAD value $\leq$0.0567 and 2) calculating the probability of such a value using a Gaussian density function with an empirically estimated mean and standard deviation (supplementary fig. S8, Supplementary Material online). The resulting probabilities were 0.12768 and 0.12852, respectively.

For the second approach, a Mantel test comparing a matrix of pairwise distances between genes in the network and matrices of pairwise absolute differences in evolutionary parameters was applied for dN/dS, dN, and dS. The test found no significant correlation between distance and difference in any evolutionary parameter (supplementary table S11, Supplementary Material online). The results of both approaches do not support the hypothesis that neighboring genes share similar evolutionary constraints in the case of the FOS-GRN.

Discussion

The question of whether the role of regulators involved in the control of floral initiation is conserved across flowering plants has been raised recently in the literature (Wellmer and Riechmann 2010). Of particular interest is the situation of grass-like plants and other monocots, which are distantly related to *A. thaliana* and its relatives. Based on the identification of homologs of the main regulators involved in the control of floral initiation of *A. thaliana* in monocots as well as observations of expression patterns in different species, it has been suggested that many aspects of the topology of the floral transition network seem to be conserved between dicots and monocots (Wellmer and Riechmann 2010). However, empirical gene conservation data based on whole-genome analysis were lacking. Given the availability of multiple genomes of angiosperms—both monocots and dicots—a comparative genomics approach was possible and

enabled us to uncover a clearer picture of the conservation status of the regulators known to be involved in the control of floral initiation and floral organ specification in *A. thaliana* across angiosperms. We focused specifically on the regulators participating in the FOS-GRN model (see Espinosa-Soto et al. 2004; Alvarez-Buylla et al. 2010 for updates). Our results show that all the FOS-GRN genes have representatives in the 18 angiosperm species used in this study. The existence of all the genes in all the surveyed species, together with the high selective constraint level found in this study (mean dN/dS = 0.124), suggests that the FOS-GRN is functionally constrained across all these species belonging to nine families, nine orders, and both monocot and dicot species. This is consistent with what we might expect given the robustness of the FOS-GRN as a developmental regulatory module and the observed expression patterns of some of the genes of this GRN documented for different species (Espinosa-Soto et al. 2004; Alvarez-Buylla et al 2010). These results, however, do not provide information of whether or not there are considerable differences in network circuitry among species. The empirical data obtained here may serve, nonetheless, as a basis to explore the dynamical behavior of the corresponding FOS-GRN in different species under the assumption of conserved interactions among network components. Indeed, further model refinements as well as phenotypic validations and testable predictions could be generated following such a theoretical approach.

Our results also show that the genes in the FOS-GRN have undergone a number of duplication and/or loss events. The evolutionary history of MADS-box genes involved in flowering has been extensively studied with phylogenetic approaches (see, e.g., Alvarez-Buylla et al. 2000). A complex history of gene duplications within the AP1/FUL clade during angiosperm evolution is well documented (Preston and Kellogg 2007). The results of gene conservation obtained in this work suggest a similar complex history for most of the genes of the other gene families in the FOS-GRN. Furthermore, it is well known that some of the species included in the study have shared whole-genome duplication (WGD) events. For example, *A. thaliana* has experienced at least three WGD events—two recent events since its divergence from other members of the Brassicales clade and a more ancient event shared with most, if not all, eudicots (Bowers et al. 2003). A WGD event occurred more than once before the split between *A. thaliana* and *A. arenosa* (Ha et al. 2009); consequently, the two *Arabidopsis* species included in the analysis have shared WGD events which are not shared with the other species. This evolutionary scenario may partially account for the complex pattern of duplications observed in the conservation data; unfortunately, it also makes it difficult to establish clear relationships of orthology. The empirical conservation data reported herein thus serve as a basis for further phylogenetic studies which are needed in order to better explain the processes leading to the conservation and expansion of the FOS-GRN across angiosperms. The data concerning the overall conservation of the FOS-GRN genes obtained here suggest interesting questions for future investigation in diverse angiosperm species, such as addressing whether the interactions of the flower organ identity genes and their interacting partners are conserved among monocots and dicots or not. What is the role of the duplicated genes in the dynamics of the FOS-GRN? Does such gene redundancy increase the robustness of the process at the level of the GRN dynamics? These and similar questions can be explored starting from the conservation data reported here and following a combination of theoretical and experimental approaches. A first approach to the role of duplications in the FOS-GRN can be found in Espinosa-Soto et al. (2004) for the case of the B-function genes in *Petunia*. This study showed that the FOS-GRN is dynamically robust to duplications.

In a study based on a comparative genomics approach, the quality of genome annotation is of major concern. The fact that putative annotation errors were detected recurrently in the same species gives support to the curational process followed, but it also suggests the need of more careful annotations in the genomes of *L. japonicus*, *O. sativa indica*, *P. trichocarpa*, *M. esculenta*, and *R. communis*. Future improvements in annotation quality may help the curational process in gene network conservation studies. Here, we report the conservation data for the FOS-GRN both before (supplementary table S5, Supplementary Material online) and after manual curation (fig. 2).

Selective Constraints in the FOS-GRN

It has been suggested that additional plant species, other than the experimental model species, should be included in molecular evolutionary studies to completely appreciate the conservation and evolvability of the regulatory network for flower development (Yang et al. 2011). Here, we show that the whole GRN controlling cell specification during early stages of flower development, when primordial floral organ cells are specified, has evolved under purifying selection. Unlike previous studies, we considered a wider range of angiosperms including both monocot and dicot species. Our results agree with previous conclusions: floral organ identity genes evolved under strong purifying selection. The evolution of the genes considered in the FOS-GRN is functionally constrained, as evidenced by the dN/dS ratios. We calculated an overall mean dN/dS of 0.124. From the 15 genes, 13 (86.66%) have a dN/dS value <0.15. Yang et al. recently reported the molecular evolutionary analysis of a group of 58 genes involved in flower development that includes all the genes that were analyzed in the present work, with the exception of EMF1 (Yang et al. 2011). Their analysis included only the species *A. thaliana* and *A. lyrata*. In their study, the authors report an average dN/dS value of 0.17 and interpret this result as evidence suggesting that these genes have overall evolved under purifying selection. On the other hand, the smaller average dN/dS value that we calculated for the 15 genes of the FOS-GRN (0.124) is based on a much wider range of species; and some of them are more distantly related than the two compared in the study of Yang et al. Furthermore, the calculated average value is highly influenced by the high dN/dS value corresponding to *EMF1* (0.39577). If we omit *EMF1* in the calculation, the average dN/dS is 0.1049864. This observation supports the conclusion

that the calculated dN/dS values are small and suggest that the FOS-GRN is functionally constrained. In order to find further support for our interpretation, we analyzed the dN and dS values previously reported for the whole-genome set of orthologous between *A. thaliana* and *A. lyrata* and calculated the average dN/dS value over the complete data set (see Materials and Methods). The calculated average dN/dS is 0.29; the complete empirical distribution is shown in supplementary figure S9*a*, Supplementary Material online. Using this whole-genome data set, we conducted a resampling experiment in order to calculate the likelihood of observing an average dN/dS value over a group of 15 genes equal or smaller to the one we report (0.124). The fraction of values from this distribution with a value equal or less than 0.124 was 0.00038. Hence, the encountered small value could be found in a random sample of the same size with a very small probability (P value = 0.00038). Supplementary figure S9*b*, Supplementary Material online, shows the distribution of simulated average dN/dS values. Considering that our dN/dS calculations are based on a set including more distant species, this empirical evidence strongly supports our claim that the reported average dN/dS of 0.124 is small.

When testing for evidence of positive selection as a force which could have fixed specific sites, using models that account for site class variability in dN/dS, we found that sites with a dN/dS >1 may exist only in *UFO*, *FT*, and *CLF* as evidenced by a marginal significant P value (before controlling for multiple tests) when comparing model M8 assuming positive selection with the null model M7 (see Materials and Methods). On the other hand, no single site in these proteins showed a high posterior probability when the Bayes' theorem was applied in order to identify potential targets of diversifying selection. Thus, in this study, both global and site varying models failed to detect any signature of positive selection for any codon of the FOS-GRN genes.

Unlike the above results, previous studies have found evidence of adaptive evolution acting at particular sites in some of the genes included in the FOS-GRN. Olsen et al. found evidence that suggests an adaptive mechanism behind the patterns of variation found on *TFL1* and *LFY*. These and similar studies (see, e.g., Olsen et al. 2002; Moore et al. 2005) are, in contrast to the present study, based on population genetic tests and data. Hence, these studies have captured the patterns of variation in these genes resulting from recent divergent evolution. Future studies should further investigate the microevolutionary process at play among the FOS-GRN genes. Some evidence at hand suggests that even for more recent divergences, floral organ identity genes will show evidence of strong purifying selection (Yang et al. 2011), but other flower transition genes seem to have been prone to positive selection as well (Martínez-Castilla and Alvarez-Buylla 2003); however, both selective forces are not mutually exclusive in any given gene.

Martínez-Castilla and Alvarez-Buylla (2003) focused on the Arabidopsis MADS-box gene family and found several sites within the MADS and K boxes, with high probabilities of having been fixed under positive selection, suggesting that these boxes may have played important roles in the acquisition of novel functions during recent events of MADS-box diversification. Here, through the analysis of alignments constructed on the basis of 1–1 orthologous relationships for distantly related angiosperm species, we did not find evidence of positive selection on such sites. Our result suggest that although adaptive evolution probably plays an important role during recent diversification events of the MADS-box gene family, a constrained evolution have prevailed upon the functionally established orthologous members across species which diverged more years ago. The question of whether or not the MADS-box gene family shows similar signs of adaptive evolution in species other than *A. thaliana* is open. This question, and its relevance for the phenotypic evolution of plants, is interesting given the complexity of the duplication events that have shaped the MADS-box gene family in angiosperms, as evidenced by the presence of multiple copies of flowering MADS-box genes found in several angiosperm species.

Selective Constraints and Functional Categories

Previous studies on floral genes in different populations of *A. thaliana* or different Arabidopsis species have also shown that floral organ identity genes evolved under strong purifying selection, but some flowering-time genes experienced relatively relaxed purifying selection and positive selection (Olsen et al. 2002; Moore et al. 2005). It has been suggested that selective constraints acting on genes of the same family are closely associated with their functions (Yang et al. 2011). The FOS-GRN includes genes which have been shown to be functionally associated with the promotion of flower meristem identity (*LFY*, *AP1*, *UFO*) or with floral organ identity (the ABC genes *AP1*, *AP2*, *AP3*, *PI*, *AG*). For historical and empirical reasons, the ABC genes have been qualified as having a prominent role in the process of cell fate and organ type specification during early flower development. Given this background information, the presence of a stronger functional constraint upon such genes in relation with the other interacting genes would be a reasonable hypothesis. Our results show that there is no significant difference between the molecular evolutionary parameters of these genes and the other genes in the FOS-GRN (supplementary table S7, Supplementary Material online), however. This suggests that the ABC genes have not been subject to a stronger functional constraint than the rest of the FOS-GRN genes, at least as evidenced by the differential rate of evolution analyses that we performed in this study. Instead, it seems that it is the whole regulatory module which is under a strong evolutionary constraint.

In contrast to the previous result, when molecular size and sequence conservation features were considered in addition to the dN/dS, it was possible to cluster the proteins into groups consistent with their functional roles. Specifically, an unsupervised model-based clustering analysis grouped the FOS-GRN proteins into three clusters consistent with their associated functions during inflorescence and flower development; and this consistency was assessed statistically

(see Results). Our results show that meristem and flower identity genes share similar molecular conservation features among them, whereas these are quite different from those observed in genes known to be involved in several other mechanisms with no apparent single prominent function. We interpret these results as evidence suggesting a constraint associated with the functional role of the genes. Although it is complicated to define rigorously a specific function for the individual components of complex molecular systems such as GRNs, given that no gene acts independently of their interacting partners or in a context-specific manner, our multivariate clustering approach uncovered a nontrivial pattern. Without any prior assumption about differences among the proteins, the methodology separated the genes in groups in a way consistent with the empirically known functions. Furthermore, the classification boundaries separating the clusters only merge in the case of the two groups in which some of their components are known to be associated with both functions (e.g., *AP1* is both a meristem and floral organ identity protein). Interestingly, it is only possible to uncover such a pattern when conservation measures other than evolutionary rates or sequence similarity were considered. The degree of conservation in sequence length seems to be relevant and closely associated with the molecular function. Finally, it is worth mentioning that the uncovered pattern is only obtained when considering several conservation features and not just a single evolutionary parameter or similarity measure.

Molecular Evolutionary Parameters and Network Architecture

Previous studies have suggested several approaches to test whether there is a relationship between network architecture and the molecular evolutionary parameters of the network's components (dN, dS, dN/dS): 1) the calculation of correlation coefficients between network topological measures of centrality and molecular evolutionary parameters (Montanucci et al. 2011), 2) the calculation of whether interacting nodes within a network have more similar values of the evolutionary parameters than noninteracting nodes (Alvarez-Ponce et al. 2009), and 3) the comparison of a matrix of pairwise shortest path distances between genes in the network and matrices of pairwise absolute differences in evolutionary parameters (Montanucci et al. 2011). Here, the three approaches were applied to the FOS-GRN, in addition to a regression-based modeling approach. Most of the above approaches assume that the architecture or topology of the network affects the molecular evolution of its nodes, and they implicitly assume then that such static network structure somehow is correlated to dynamical or functional modularity. Unlike previous network-level molecular evolutionary studies, we did not find a significant relationship between network architecture and the evolutionary parameters: 1) no significant correlation was found between the evolutionary parameters and the measures of centrality of the nodes, 2) analyses did not support the hypothesis that neighboring genes in the network share similar evolutionary constraints, and 3) regression coefficients did not support a relationship between the molecular evolutionary parameters and any of the nodes' topological features tested. This result suggests that the proteins of the FOS-GRN, although subject to purifying evolutionary forces, do not show any discernible pattern of association between the strength of constraint and the local structural properties within the network. This implies that the whole module is subject to similar molecular evolutionary constraints and/or the structural considerations do not have a functional or dynamical relevance that might have been important for the evolutionary constraints experienced by different nodes within the FOS-GRN. These results should be interpreted with caution, however, because of the small sample size. Statistical analysis has two goals that directly conflict. First is to find patterns in data. The second goal is a fight against apophenia, the human tendency to invent patterns in random data (Klemens 2008). In the context of GRNs, care should be taken when testing for the existence of relationships (or lack thereof) between node features and evolutionary patterns based on statistical analysis. The identification of "real patterns" could be limited by the size of the data set analyzed. Nonetheless, it is noteworthy that previous studies for small pathways/networks with a similar number of nodes as in the GRN analyzed here ($\leq$20 nodes) have found significant trends between topological and evolutionary parameters (see, e.g., Alvarez-Ponce et al. 2009; Fitzpatrick and O'Halloran 2012).

Given that we did find an association between conservation features of the genes—including evolutionary rates—and their functional role during flower development, and considering that the role of specific genes in the specification of meristem and floral identity has been probed during the analysis of the FOS-GRN as a dynamical system (Espinosa-Soto et al. 2004), we speculate that functional (dynamical), instead of topological, network properties, such as those associated with robustness, could be significantly associated with the molecular evolutionary constraints of the genes in the FOS-GRN reported here.

Overall, our results depict a general picture of the evolutionary pattern of the FOS-GRN where functional constraint better explains the evolution of its genes. The approach followed here provided new data relevant for the study of the evolution of the mechanisms at the molecular level that are behind organ identity during early flower development. Specifically, we have shown that 1) the FOS-GRN genes are conserved among 18 Angiosperm species; 2) a complex history of gene duplications seems to have been involved in the expansion of the network across angiosperms; 3) the whole FOS-GRN has evolved under purifying selection; 4) ABC floral organ identity genes do not show a significantly stronger evolutionary constraint than the other genes in the FOS-GRN; 5) an association between protein length and sequence conservation features, evolutionary rates, and functional category seems to prevail among the genes in the FOS-GRN; and 6) the FOS-GRN does not show any significant relationship between network architecture and the evolutionary parameters of its genes.

Materials and Methods

Sequence Data

The FOS-GRN described in Espinosa-Soto et al. (2004) and updated in Alvarez-Buylla et al. (2010) was used as study system; the corresponding genes are reported in supplementary table S1, Supplementary Material online. The identifiers of the genes involved in this network were obtained from the TAIR database (http://www.arabidopsis.org, last accessed November 24, 2013) and integrated into the workbench tool of the Plaza Comparative Genomics Platform (http://bioinformatics.psb.ugent.be/plaza/, last accessed November 24, 2013) (Proost et al. 2009).

After applying the PLAZA integrative method of orthologous genes finding (discussed later), both the sequence data of the genes of *A. thaliana* and the sequence data of the corresponding homologous genes were retrieved using the export functionality of the PLAZA'S workbench tool. This first data set corresponds to the FOS-GRN preliminary gene conservation set which includes those species with a sequenced and annotated genome and is represented as a conservation matrix and a list of corresponding gene identifiers in supplementary tables S2 and S3, Supplementary Material online, respectively. In order to reduce the probability of reporting the conservation of nonfunctional proteins, the preliminary data set was manually curated. For this purpose, erroneous automatic orthology designations were discarded, and those groups of adjacent gene annotations actually corresponding to different regions of a single gene were merged (discussed later). The final and corrected conservation data of the FOS-GRN proteins across angiosperms are reported in the form of a conservation matrix (fig. 1*a*) and its corresponding list of gene IDs (supplementary table S4, Supplementary Material online).

Homology Search

The PLAZA Comparative Genomics Platform offers an access point for plant comparative genomics centralizing genomic data produced by different genome sequencing initiatives (Proost et al. 2009). The PLAZA integrative method of orthologous genes integrates a complementary set of data types and methodologies in order to infer orthologous gene relationships based on the following sources of evidence: Orthologous gene families (ORTHO) inferred using OrthoMCL, Tree-based orthologs (TROG) inferred using tree reconciliation of the phylogenetic tree of a gene family, Best-Hits-and-Inparalogs (BHI) inferred from Blast hits against the PLAZA protein database, and Anchor points refer to gene-based colinearity between species. Using this tool, different homology relationship types can be considered: when a gene has no paralogs and only 1 ortholog (1–1), when a gene has 1 or more paralogs and only 1 ortholog (N–1), and the corresponding combinations for a total of four different orthology relationship: 1–1, N–1, 1–N, and M–N. In this work, the PLAZA integrative method was used to infer homology gene relationships for each protein in the FOS-GRN. The following settings were used: all orthologous relationship types were allowed, all evidence types were taken into account, and 18 plant species corresponding to the Phylum Angiospermae were included (see Results).

Manual Curation of Putative In-paralogs

As the degree and quality of annotation of whole-genome projects varies considerably among species, it is not adequate to rely only on automatic procedures, and instead, careful data set cleaning is necessary. Further manual curation to the reported gene groups after a homology analysis should be considered in order to reduce the likelihood of including nonfunctional proteins in other analyses. For each gene in the preliminary conservation data list (supplementary table S3, Supplementary Material online), the following information was extracted from PLAZA Comparative Genomics Platform: CDS sequence, protein sequence, chromosome, location (e.g., start, stop), length, and InterPro annotated protein domains. Given these data, some putative in-paralog genes were manually eliminated from the preliminary conservation data. On the other hand, the homology status of some genes was updated based on one or more of the following criteria: partial proteins (small size), lack of any of the protein domains of the orthologous gene in *A. thaliana*, neighboring genomic location, or low sequence alignment quality. The preliminary status of certain genes in the conservation data as multiple single paralogous copies in the same genome was modified to single copy orthologous genes, once it was realized that in many cases different boxes of the same open reading frame were sometimes annotated as different genes. Details of the manual curation process and sequence selection criteria are described in the supplementary text, Supplementary Material online.

Multiple Alignments and Phylogenetic Inference

All protein multiple sequence alignments (MSAs) were generated using the software CLUSTALW version 2.1 (Larkin et al. 2007). The software PAL2NAL (Suyama et al. 2006) was used to generate multiple codon alignments from the corresponding aligned protein sequences and the corresponding DNA coding sequences. For each orthologous group, a maximum likelihood phylogeny estimation was conducted using the software Phylm (Guindon and Gascuel 2003; Guindon et al. 2010) applying the nucleotide substitution model that best fits the data according to the Akaike information criterion. Details of the selected substitution models are provided on supplementary table S12, Supplementary Material online. Both phylogeny estimation and substitution model selection were conducted using the function phymltest of the package ape (Paradis et al. 2004) in the R statistical programming environment (www.R-project.org, last accessed November 24, 2013) as described in Paradis (2012).

Analysis of the Evolutionary Rates

The evolutionary parameters dN, dS, and dN/dS were estimated following a maximum likelihood procedure as implemented in the software codeml of the PAML package version

4.5 (Yang 2007). Due to the broad range of species considered for the conservation study, it was not possible to obtain reliable alignments for all 18 species for molecular evolutionary analysis. This analysis was then restricted to a representative group of species (*A. lyrata*, *A. thaliana*, *B. distachyon*, *G. max*, *M. esculenta*, *O. sativa*, *S. bicolor*, *T. cacao*, and *Z. mays*) to avoid bias in the dN/dS values—this decision was based on the manual inspection of the resulting alignments. All alignments are publicly available upon request. Only MSAs based on (putative) 1:1 ortholog sets were used. In the cases in which there were more than one gene copy in a given species, the gene with the most complete sequence or the one without any homogenization features (stop codons or frameshift mutations) was used. For each codon alignment, two tests of positive selection were performed. In order to test whether the assumption of positive selection fits better the data than the assumption of nearly neutral evolution, the model M2a was compared against the null model M1a through a likelihood ratio test (LRT). In a second test of positive selection, the models M7 (null model of neutral evolution), which assumes that dN/dS follows a (discrete) beta distribution among sites and M8 (positive selection model), which adds a class of dN/dS which can be greater than 1, were compared through an LRT. The false discovery rate and Bonferroni corrections for the multiple tests of positive selection were conducted using the function p.adjust of the stats package in the R statistical programming environment. In all the analyses, the F3×4 codon frequency model was used. Details of the LRT for each comparison are provided in supplementary table S5, Supplementary Material online. The strength of purifying selection was measured using the dN/dS values computed through the M0 model, which calculates rates encompassing all the branches of the tree and for the entire length of the sequence.

The dN and dS values reported for the whole-genome orthologous pairs between *A. thaliana* and *A. lyrata* were downloaded from the Ensemble Plant website (http://plants.ensembl.org/index.html, last accessed November 24, 2013) using the BioMart platform for data retrieval. The corresponding dN/dS values and their statistics were calculated over the complete data set, omitting missing data (a total of 22,531 values). The empirical distribution is shown in supplementary figure S9*a*, Supplementary Material online. A resampling experiment was conducted using the complete set of dN/dS values as follows: a large number of gene groups of size 15 (100,000) were randomly generated, the dN/dS average value was calculated for each group, and the distribution obtained values was used to estimate the likelihood of observing an average value equal or smaller than the one calculated for the FOS-GRN (0.124). The simulated distribution is shown in supplementary figure S9*b*, Supplementary Material online.

Gene Conservation Features

The pairwise distances from protein MSAs were calculated using the function dist.ml of the package phangorn (Schliep 2011) with the default parameters. In the case of DNA codon MSAs, a matrix of pairwise distances was computed using the dist.dna function of the package ape (Paradis et al. 2004) with the default parameters. To obtain a final scalar conservation feature, the corresponding means were calculated and used as a summary statistics. The coefficient of variation in protein size of each protein over all species was calculated as a measure of the degree of conservation (variation) in molecular size. All the calculations discussed in this section were conducted using the R statistical programming environment.

Genes Clustering and Function

Hypothesis tests of statistical difference of the evolutionary parameters between the ABC floral organ identity genes and the other genes in the FOS-GRN were conducted following a nonparametric method (Kruskal-Wallis test). A model-based clustering analysis was conducted using the molecular and sequence conservation features in table 2 (last four columns) as an input feature matrix. Intuitively, the goal of clustering is to assign points that are similar to the same cluster and to ensure that points that are dissimilar are in different clusters. The analysis was conducted as implemented in the function Mclust of the mclust package version 4.1 (Fraley et al. 2012). This procedure fits a Gaussian finite mixture model to the data through an EM algorithm. The best model is selected according to the Bayesian information criterion. The clustering procedure was evaluated using the functional categories of the genes as an external form of data for validation. The clustering was then compared with the labels using as summary statistic the Rand index, which measures the fraction of clustering decisions that are correct (Murphy 2012). The Rand index was calculated using the function cluster_similarity of the package clusteval (http://cran.r-project.org/web/packages/clusteval/, last accessed November 24, 2013). In order to assess the statistical significance of the clustering, the frequentist sampling distribution of a standard summary statistic that quantifies the fraction of clustering decision that are correct was computed using a bootstrap method. The Rand index was used as a summary statistic (Murphy 2012). Specifically, a character vector corresponding to the clustering output was permuted a large number of times (n = 1,000,000) and compared each time with the labels vector using the Rand index. The obtained sampling distribution was used to calculate the probability of observing a Rand index value equal or greater than the one observed when comparing the original output of the clustering analysis with the labels vector. Both model-based clustering analysis and clustering evaluation were conducted in the R statistical programming environment.

Evolutionary Rates and Network Architecture

The measures of centrality describe numerically the topological importance of a node in a graph, given its structure. For each gene (node) in the FOS-GRN, the following measures of centrality were calculated: degree (number of nodes it is connected to), closeness (reciprocal of the average distance to all other nodes), betweenness (fraction of all shortest paths that pass through it), and eccentricity (maximum distance

from it to all other nodes). All network topological computations were conducted using the igraph package (Csardi and Nepusz 2006). Two analyses were conducted in order to test for the association of the evolutionary parameters of the genes and their topological features within the network. 1) Spearman correlation coefficients were calculated between each evolutionary parameter given by the model M0 (dN, dS, and dN/dS) and each topological features. 2) Simple linear regression models were fitted using each evolutionary parameter as response variable and each topological feature as predictor.

It was also investigated whether genes that are interacting in the FOS-GRN have related values of the evolutionary parameters. For this purpose, two additional analyses were conducted. In the first analysis, following Alvarez-Ponce et al. (2009), the average absolute difference of the value of the evolutionary parameters between interacting components in the network was calculated and then used as an statistic in a simulation (sampling) procedure in order to assess how frequently it is expected to observe this or a smaller value in an ensemble of similar but random networks. Specifically, 100,000 networks each with the same number of nodes and interactions were generated, and the statistic was calculated for each of these networks. The estimated distribution of the statistic over the ensemble of networks was then used to calculate the probability of observing a value equal or smaller than that calculated in the FOS-GRN. A Gaussian density function with parameters estimated from the data (mean = 0.0713 and standard deviation = 0.0128) was also fitted from the observed simulated data and used for probability calculations. In the second analysis, following Montanucci et al. (2011), a matrix of pairwise shortest path distances between the nodes (path distance matrix) and three matrices of absolute pairwise gene differences in each of the evolutionary parameters were computed. Each of these last matrices was then compared with the path distance matrix through standardized Mantel tests using the ecodist package. All the analyses discussed in this section were conducted using the R statistical programming environment.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S12 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

E.R.A.B. acknowledges the support of the Miller Institute for Basic Research in Science, University of California, Berkeley, while spending a sabbatical leave in the lab of Chelsea Specht. The authors acknowledge technical support of Rigoberto V. Pérez-Ruiz and logistical and administrative help of Diana Romo. This article constitutes a partial fulfillment of the graduate program Doctorado en Ciencias Biomédicas of the Universidad Nacional Autónoma de México, UNAM in which J.D.-V. developed this project. This work was supported by grants from CONACYT, Mexico: 180098 (to E.R.A.B.) from PAPIIT-UNAM, IN203113-3 (to E.R.A.B.).

References

Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MP. 2005. Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol Biol.* 5:23.

Alvarez-Buylla ER, Azpeitia E, Barrio R, Benitez M, Padilla-Longoria P. 2010. From ABC genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. *Semin Cell Dev Biol.* 21:108–117.

Alvarez-Buylla ER, Chaos A, Aldana M, et al. (11 co-authors). 2008. Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS One* 3:e3626.

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, De Pouplana LR, Marťınez-Castilla L, Yanofsky MF. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc Natl Acad Sci U S A.* 97:5328–5333.

Alvarez-Ponce D. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. *BMC Evol Biol.* 12:192.

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19:234–242.

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. *Genome Biol Evol.* 3: 87–101.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. *Genome Biol Evol.* 4:1263–1274.

Becker A, Theissen G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet Evol.* 29:464–489.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438.

Bremer B, Bremer K, Chase M, Fay M, Reveal J, Soltis D, Soltis P, Stevens P. 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.

Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R, Calafell F, Awadalla P, Netea MG, Bertranpetit J. 2011. Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol Biol.* 11:202.

Coen ES, Meyerowitz EM. 1991. The war of the whorls: genetic interactions controlling flower development. *Nature* 353:31–37.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26:479–484.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695:5.

Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER. 2004. A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell* 16: 2923–2939.

Fraley C, Raftery AE, Murphy TB, Scrucca L. 2012. MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report no. 597, Department of Statistics, University of Washington.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.

Fitzpatrick DA, O'Halloran DM. 2012. Investigating the relationship between topology and evolution in a dynamic nematode odor genetic network. *Int J Evol Biol.* 2012(2012):548081.

Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.
Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A.* 106:2295–2300.
Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? *J Mol Evol.* 58: 203–211.
Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 22(4):803–806.
Huang S, Kauffman S. 2009. Complex gene regulatory networks-from structure to biological observables: cell fate determination. In: Meyers RA, editor. Encyclopedia of complexity and systems science. Berlin: Springer. p. 1180–1293.
Invergo BM, Montanucci L, Laayouni H, Bertranpetit J. 2013. A system-level, molecular evolutionary analysis of mammalian phototransduction. *BMC Evol Biol.* 13:52.
Jovelin R, Phillips PC. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* 10:R35.
Klemens B. 2008. Modeling with data: tools and techniques for scientific computing. Princeton (NJ): Princeton University Press.
Krizek BA, Fletcher JC. 2005. Molecular mechanisms of flower development: an armchair guide. *Nat Rev Genet.* 6:688–698.
Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
Lavagnino N, Serra F, Arbiza L, Dopazo H, Hasson E. 2012. Evolutionary genomics of genes involved in olfactory behavior in the *Drosophila melanogaster* species group. *Evol Bioinform Online.* 8:89–104.
Lawton-Rauh AL, Alvarez-Buylla ER, Purugganan MD. 2000. Molecular evolution of flower development. *Trends Ecol Evol.* 15:144–149.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol.* 22(5):1345–1354.
Martínez-Castilla L, Alvarez-Buylla ER. 2003. Adaptative evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. *Proc Natl Acad Sci U S A.* 100(23):13407–13412.
Mendoza L, Alvarez-Buylla ER. 1998. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *J Theor Biol.* 193(2):307–319.
Montanucci L, Laayouni H, Dall'Olio GM, Bertranpetit J. 2011. Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. *Mol Biol Evol.* 28:813–823.
Moore RC, Grant SR, Purugganan MD. 2005. Molecular population genetics of redundant floral-regulatory genes in *Arabidopsis thaliana*. *Mol Biol Evol.* 22:91–103.
Murphy K. 2012. Machine learning: a probabilistic approach. Cambridge (MA): MIT Press.
Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. *Genetics* 160:1641–1650.
Paradis E. 2012. Analysis of phylogenetics and evolution with R. New York: Springer.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
Preston JC, Kellogg EA. 2007. Conservation and divergence of APETALA1/FRUITFULL-like gene function in grasses: evidence from gene expression analyses. *Plant J.* 52:69–81.
Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718–3731.
Purugganan MD. 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. *J Mol Evol.* 45:392–396.
Purugganan MD, Rounsley SD, Schmidt RJ, Yanofsky MF. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* 140:345–356.
Purugganan MD, Suddith JI. 1999. Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. *Genetics* 151:839–848.
Riechmann JL, Meyerowitz EM. 1997. MADS domain proteins in plant development. *J Biol Chem.* 378:1079.
Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L. 2010. The *Arabidopsis thaliana* flower organ specification gene regulatory network determines a robust differentiation process. *J Theor Biol.* 264: 971–983.
Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.
Villarreal C, Padilla-Longoria P, Alvarez-Buylla ER. 2012. General theory of genotype to phenotype mapping: derivation of epigenetic landscapes from N-node complex gene regulatory networks. *Phys Rev Lett.* 109:118102.
Wellmer F, Riechmann JL. 2010. Gene networks controlling the initiation of flower development. *Trends Genet.* 26:519–527.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang L, Chun-Ce G, Gui-Xia H, Hong-Yan S, Hong-Zhi K. 2011. Evolutionary pattern of the regulatory network for flower development: insights gained from a comparison of two *Arabidopsis* species. *J Syst Evol.* 49:528–538.
Yang Y, Zhang F, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. *BMC Evol Biol.* 9:206.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 15:496–503.