Integrating Phylogenetic and Network Approaches to Study Gene Family Evolution: The Case of the *AGAMOUS* Family of Floral Genes

Evolutionary Bioinformatics
Volume 14: 1–11
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176934318764683

Daniel S Carvalho[1,2], James C Schnable[1,2] and Ana Maria R Almeida[3]

[1]Center for Plant Science Innovation, University of Nebraska–Lincoln, Lincoln, NE, USA.
[2]Department of Agronomy and Horticulture, University of Nebraska–Lincoln, Lincoln, NE, USA.
[3]Department of Biological Sciences, California State University East Bay, Hayward, CA, USA

ABSTRACT: The study of gene family evolution has benefited from the use of phylogenetic tools, which can greatly inform studies of both relationships within gene families and functional divergence. Here, we propose the use of a network-based approach that in combination with phylogenetic methods can provide additional support for models of gene family evolution. We dissect the contributions of each method to the improved understanding of relationships and functions within the well-characterized family of *AGAMOUS* floral development genes. The results obtained with the two methods largely agreed with one another. In particular, we show how network approaches can provide improved interpretations of branches with low support in a conventional gene tree. The network approach used here may also better reflect known and suspected patterns of functional divergence relative to phylogenetic methods. Overall, we believe that the combined use of phylogenetic and network tools provide a more robust assessment of gene family evolution.

KEYWORDS: gene families, phylogenetics, functional divergence, network analysis

RECEIVED: October 20, 2017. **ACCEPTED:** February 9, 2018.

TYPE: Special Collection: Algorithm Development for Evolutionary Biological Computation

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by CAPES (BJT 069/2013) and Instituto Nacional de Ciência e Tecnologia em Estudos Interdisciplinares e Transdisciplinares em Ecologia e Evolução (INCT-INTREE, 465767/2014-1) to A.M.R.A., a CNPq Science Without Borders scholarship (214038/2014-9) to D.S.C., and a Robert B. Daugherty Water for Food Institute research support award to J.C.S.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: James C Schnable, Center for Plant Science Innovation, University of Nebraska–Lincoln, Lincoln, NE 68588, USA. Email: schnable@unl.edu

Introduction

Advances in sequencing technology have lead to dramatic expansions in the number of sequenced genes within most gene families, both through the use of whole genome or whole transcriptome sequencing and through broader taxon sampling. Gene families are generally studied through the use of phylogenetic approaches to identify closely and distantly related sequences, as well as to classify divergence between gene copies into those resulting from speciation (orthology) or gene duplication (paralogy).[1,2] Thus, phylogenetic approaches are widely employed to study how sequence divergence can lead to divergence of structure and/or function.[3,4] When coupled with genome context information, this approach can provide insightful understanding of gene regulation and function.

For instance, it is well known that orthologous genes conserved at syntenic locations in the genome are more likely to exhibit conserved regulation[5] and function[6] than genes at non-syntenic locations. However, the prevalence of whole genome duplications in plants poses challenges to the study of gene family evolution using exclusively phylogeny-based methods[3] due to the diverse outcomes of duplicated genes. Whole genome duplications produce syntenic paralogs that can be reciprocally lost,[7,8] sub- or neofunctionalized,[9] or even retained in the same functional roles as a result of relative or absolute dosage constraints.[10]

A fundamental assumption of any phylogenetic reconstruction is that the observed changes occur exclusively through a hierarchical bifurcated branching process. This model is certainly a good representation of a major evolutionary force (ie, descent with modification)[11]; however, many will argue that it fails to capture the diversity of evolutionary processes which shape the gene content of extant species.[12,13]

One way to address the complexity of evolutionary processes is to apply network approaches to address questions related to cell organization and functioning,[14] human diseases relationships,[15] and plant gene function prediction.[16] Network approaches have also been successfully applied to study fungi evolution based on enzymes related to the chitin synthase pathway.[17] Recently, Carvalho et al[18] have used a network-based approach to address the origin of the mitochondria, providing a new perspective on the study mitochondrial evolution.

Network-based approaches can overcome some of the limitations of phylogenetic methods. For instance, these approaches do not require the assumption of a hierarchical bifurcating framework and therefore may be capable of dealing with more complex biological patterns and phenomena.[19–21] Networks are generally less precise in their ability to reconstruct the divergence points of different groups within a gene family; however, they may be able to capture additional insight into function evolution and divergence using information which might be lost in phylogenetic reconstructions.

In this study, we compare the information gained from conventional phylogenetic analysis and a network-based approach using a well-characterized subfamily of floral transcription factors, the *AGAMOUS* floral genes. The *AGAMOUS* gene subfamily comprises MADS-box transcription factors and is involved in important aspects of flower and fruit development.[22] Among angiosperms (flowering plants), the *AGAMOUS* subfamily is traditionally divided into the C and D lineages. C lineage genes include the closest relatives of the *Arabidopsis thaliana AGAMOUS* (*AG*) gene[23,24] in all angiosperms, as well as close relatives of the *SHATTERPROOF* (*SHP*) gene, present exclusively in core eudicots.

D lineage genes, on the other hand, include angiosperm *SEEDSTICK* (*STK*) genes.[25,26] The C/D split likely occurred after the split between gymnosperms and angiosperms and gymnosperms usually carry a single-gene copy of the *AGAMOUS* subfamily. While D lineage genes are usually related to ovule development, C lineage genes have been implicated in stamen and carpel development. Particularly, in core eudicots, *SHP* genes have also been shown to be involved in fruit development and ripening.[27–30]

This gene subfamily has been extensively studied and mutant characterization has provided insights into their functional roles in carpel, ovule, and fruit development as well as floral meristem termination. The *AGAMOUS* subfamily has undergone several instances of duplication followed by neo- and subfunctionalization throughout its evolutionary history in angiosperms,[26,31] and understanding the evolutionary history of this group has proven challenging as a result of low support for deep nodes on the tree.

Here, we investigate the contributions of a a similarity-based phylogenetic network approach to our understanding of AGAMOUS gene family evolution.[32–34] The phylogenetic network methods used here do not require the assumption of a scale-free topology, or the need to calculate gene correlation based on expression data,[16,35] which makes the approach used more straightforward. Also, the approach used here does not rely on an existing tree to generate the networks, as most phylogenetic networks. Overall, both the phylogeny and network results showed consistent clustering of the gene families. However, our results suggest that the network approach was less affected by sequence divergence. We demonstrate that a combination of both methods can provide additional insight into evolutionary events and functional divergence within gene families.

Methods

Sequence search and multiple sequence alignment

C and D lineage *AGAMOUS* nucleotide sequences were retrieved on Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) and NCBI (National Center for Biotechnology Information). Species of origin and accession numbers for each sequence included in this analysis are provided in Table 1. A multiple sequence alignment was performed using the ClustalW[36] alignment tool within Geneious v7.0.4,[37] based on translated nucleotides. Further refinements were made manually, using translated sequences as a way to guide manual curation. Manual curation of the multiple sequence alignment was performed using a codon-preserving approach and taking into account domains and motifs previously described in the literature.[25] Unalignable regions were removed prior to further analysis. The final multiple sequence alignment included 549 nucleotides. The alignment statistics obtained from HMMSTAT, from HMMER3 package,[38] were eff_nseq = 2.72, M = 531, relent = 0.45, info = 0.45, p relE = 0.31, and compKL = 0.02. jModelTest 2.1.1 was used to estimate the best-fit evolutionary model of nucleotide evolution.[39] A protein multiple sequence alignment was also performed with the same sequences and used in downstream phylogenetic analysis.

Phylogenetic analysis

Maximum likelihood analysis was performed using PhyML 3.0 (http://www.atgc-montpellier.fr/phyml/)[40,41] with the TN93 model,[42] a gamma distribution parameter of 1.107. Bootstrap support was calculated based on 100 iterations. The most likely tree was computed based on the PhyML-estimated parameters: transition/transversion ratio for purines of 2.541, transition/transversion ratio for pyrimidines of 4.342, and nucleotides frequencies of f(A) = 0.33406, f(C) = 0.20359, f(G) = 0.24537, and f(T) = 0.21698. An ML tree of the protein sequence multiple sequence alignment was also performed on PhyML 3.0 using the LG model of amino acid substitution.

Obtaining identity matrix

A pairwise distance matrix, based on a nucleotide multiple sequence alignment of the 93 sequences, was calculated using MEGA7. Even though the length of the final alignment obtained was 543 positions, removal of gaps and missing data was performed to calculate the distance matrix, resulting in a final set of 372 informative positions in the final filtered data set.[43] The number of base substitutions per site between sequences was calculated using the maximum composite likelihood model.[44] To obtain the identity value of the sequence pairs, we subtracted 1 from the distance value of every term of the distance matrix to finally obtain the identity matrix.

Network analysis

Once the gene identity matrix was generated, a set of 101 networks were created based on the identity threshold between sequence pairs (1 network for each threshold, 0% through 100%), which is represented by the parameter σ. In each network, each nucleotide sequence is represented by a single node. Two nodes (say *i* and *j*) are considered connected if the identity threshold is greater than σ. The networks were represented in the format of an adjacency matrix $M(\sigma)$, where the matrix elements M_{ij} (pairs of sequences) were either 1, if they were connected, or 0, if they were not connected.[45] Then, neighborhood matrices $\widehat{M}(\sigma)$ were built for each one of the $M(\sigma)$.[46,47] Each element $\widehat{m}_{ij}$ from $\widehat{M}(\sigma)$ represents the number of steps in the

Table 1. List of species and sequence identifiers used in this study.

GENE NAME	SPECIES NAME	CLADE NAME	ACCESSION OR GENE ID
GinbiMADS5	*Ginkgo biloba*	Ginkgoaceae	GU563899
PiabDAL2	*Picea abies*	Pinaceae	X79280.1
PiradAG	*Pinus radiata*	Pinaceae	AF023615
TbaccAG	*Taxus baccata*	Taxaceae	JF519754
CryjaMADS4	*Cryptomeria japonica*	Taxodiaceae	HM177453
ShenAG	*Saruma henryi*	Aristolochiaceae	AY464101
ChlspiSTK	*Chloranthus spicatus*	Chloranthaceae	AY464099
PeamAG1	*Persea americana*	Lauraceae	DQ398021
PeamAG2	*Persea americana*	Lauraceae	DQ398022
MafiAG1	*Magnolia figo*	Magnoliaceae	JQ326236
MapreSTK	*Magnolia praecossisima*	Magnoliaceae	AB050653
MialAG	*Michelia alba*	Magnoliaceae	JQ326219
ElguiAG1	*Elaeis guineensis*	Arecaceae	AY739698
ElguiAG2	*Elaeis guineensis*	Arecaceae	AY739699
ElguiSTK	*Elaeis guineensis*	Arecaceae	XP_010912706.1
*BdiAG1**	*Brachypodium distachyon*	Poaceae	Bradi2g06330.1
*BdiAG2**	*Brachypodium distachyon*	Poaceae	Bradi4g40350.1
*BdiAG3**	*Brachypodium distachyon*	Poaceae	Bradi2g25090.1
OsMADS3	*Oryza sativa*	Poaceae	L37528
OsMADS13	*Oryza sativa*	Poaceae	AF151693
OsMADS21	*Oryza sativa*	Poaceae	FJ750944
OsMADS58	*Oryza sativa*	Poaceae	AB232157
*SbAG1**	*Sorghum bicolor*	Poaceae	Sb03g002525
*SbAG2**	*Sorghum bicolor*	Poaceae	Sb08g006460 or Sobic.008G072900.1
*SbAG3**	*Sorghum bicolor*	Poaceae	Sb09g006360 or Sobic.009G075500.3
*ZAG2**	*Zea mays*	Poaceae	GRMZM2G160687-T03
*ZMM2**	*Zea mays*	Poaceae	GRMZM2G359952-T01
*ZAG1**	*Zea mays*	Poaceae	GRMZM2G052890-T01
LaschisAG	*Lacandonia schismatica*	Triuridaceae	GQ214163
LaschisSTK	*Lacandonia schismatica*	Triuridaceae	GQ214164
GongaSTK	*Gongora galeata*	Orchidaceae	AIU94767.1 or KF914206.1
GongaAG	*Gongora galeata*	Orchidaceae	AIU94768.1
HyvilSTK	*Hypoxis villosa*	Hypoxidaceae	AIU94766.1 or KF914205.1
HyvilAG	*Hypoxis villosa*	Hypoxidaceae	AIU94771.1
AspvirSTK	*Asparagus virgatus*	Asparagaceae	AB175825.1
AspvirAG	*Asparagus virgatus*	Asparagaceae	BAD18011.1

(Continued)

Table 1. (Continued)

GENE NAME	SPECIES NAME	CLADE NAME	ACCESSION OR GENE ID
BgilAG	*Berberis gilgiana*	Berberidaceae	AY464106
EupleAG1	*Euptelea pleiosperma*	Eupteleaceae	GU357452
EupleAG2	*Euptelea pleiosperma*	Eupteleaceae	GU357453
AkquiAG	*Akebia quinata*	Lardizabalaceae	AY464107
AktriAG	*Akebia trifoliata*	Lardizabalaceae	AY627635
AktriSTK	*Akebia trifoliata*	Lardizabalaceae	AY627629
HogrAG1	*Holboellia grandiflora*	Lardizabalaceae	JQ806406
HogrAG2	*Holboellia grandiflora*	Lardizabalaceae	JQ806407
EscaAG1	*Eschscholzia californica*	Papaveraceae	DQ088996
EscaAG2	*Eschscholzia californica*	Papaveraceae	DQ088997
EscaSTK	*Eschscholzia californica*	Papaveraceae	DQ088998
*AqAG1**	*Aquilegia coerulea*	Ranunculaceae	Aquca-136-00009.1
*AqAG2**	*Aquilegia coerulea*	Ranunculaceae	Aquca-022-00039.1
*AqAGL11**	*Aquilegia coerulea*	Ranunculaceae	Aquca-136-00010.1
ThathAG1	*Thalictrum thalictroides*	Ranunculaceae	JN887118
ThathAG2	*Thalictrum thalictroides*	Ranunculaceae	AY867879
MedilSTK	*Meliosma dilleniifolia*	Sabiaceae	AY464105
*AlyrAG**	*Arabidopsis lyrata*	Brassicaceae	946287
*AlyrSHP1**	*Arabidopsis lyrata*	Brassicaceae	486333
*AlyrSHP2**	*Arabidopsis lyrata*	Brassicaceae	321962
*AlyrSTK**	*Arabidopsis lyrata*	Brassicaceae	489841
ATSHP1	*Arabidopsis thaliana*	Brassicaceae	AT3G58780
ATSHP2	*Arabidopsis thaliana*	Brassicaceae	AT2G42830
ATSTK	*Arabidopsis thaliana*	Brassicaceae	AT4G09960 or NM_001203767.1
*BraAG**	*Brassica rapa*	Brassicaceae	Brara.K01743.1
*BraSHP1**	*Brassica rapa*	Brassicaceae	Brara.G01817.1
*BraSHP2**	*Brassica rapa*	Brassicaceae	Brara.E00310.1
*BraSTK**	*Brassica rapa*	Brassicaceae	Brara.C02624.1
*CaruAG**	*Capsella rubella*	Brassicaceae	Carubv10005558m
*CaruSHP1**	*Capsella rubella*	Brassicaceae	Carubv10019520m
*CaruSHP2**	*Capsella rubella*	Brassicaceae	Carubv10025002m
*CaruSTK**	*Capsella rubella*	Brassicaceae	Carubv10003771
*ThhSHP1**	*Thellungiella halophila*	Brassicaceae	Thhalv10006196
*ThhSHP2**	*Thellungiella halophila*	Brassicaceae	Thhalv10017047
*ThhSTK**	*Thellungiella halophila*	Brassicaceae	Thhalv10028938
*CapaSHP**	*Carica papaya*	Caricaceae	Evm. TU supercontig_50.73
*MetrAG**	*Medicago truncatula*	Fabaceae	Medtr8g087860.1

Table 1. (Continued)

GENE NAME	SPECIES NAME	CLADE NAME	ACCESSION OR GENE ID
MetrSHP	*Medicago truncatula*	Fabaceae	JX308825
*MetrSTK**	*Medicago truncatula*	Fabaceae	Medtr 3g 005530.1
*GoraAG1**	*Gossypium raimondii*	Malvaceae	Gorai.N017200.1
*GoraAG2**	*Gossypium raimondii*	Malvaceae	Gorai.011G035500.1
*GoraSHP**	*Gossypium raimondii*	Malvaceae	Gorai012G042600.1
*GoraSTK1**	*Gossypium raimondii*	Malvaceae	Gorai.009G265100.1
*GoraSTK2**	*Gossypium raimondii*	Malvaceae	Gorai.009G288000.1
*ThecAG**	*Theobroma cacao*	Malvaceae	Thecc1E6029596t1
*ThecSHP**	*Theobroma cacao*	Malvaceae	Thecc1EG001841t1
*ThecSTK**	*Theobroma cacao*	Malvaceae	Thecc1EG036541t1
*MguAG**	*Mimulus guttatus*	Phrymaceae	Mgv1a012605 or Migut.M00986.1
*MguSTK**	*Mimulus guttatus*	Phrymaceae	Mgv1a013047m or Migut.C01334.1
*PotriAG**	*Populus trichocarpa*	Salicaceae	Potri.011G075800.1
*PotriSTK**	*Populus trichocarpa*	Salicaceae	Potri.013G104900.1
*PotriSTK2**	*Populus trichocarpa*	Salicaceae	Potri.019G077200.1
TAG	*Solanum lycopersicum*	Solanaceae	L26295.1
TSHP	*Solanum lycopersicum*	Solanaceae	AY098735
TSTK	*Solanum lycopersicum*	Solanaceae	NM_001247265.2
*ViviSHP**	*Vitis vinifera*	Vitaceae	GSVIVG01000802001
*ViviAG**	*Vitis vinifera*	Vitaceae	GSVIVT01021303001

Genes retrieved from NCBI (National Center for Biotechnology Information) (genes with * were retrieved from Phytozome) (long).

shortest path connecting 2 nodes *i* and *j*. Whenever 2 nodes are not connected and belong in different clusters, $\widehat{m}_{ij} = 0$. A neighborhood matrix shows the number of edges connecting 2 nodes in the network. The neighborhood matrices were later used to calculate the network distance $\delta(\sigma, \sigma + \Delta\sigma)$ between the pairs of successive networks (in this case, $\Delta\sigma = 1$), to find the network with the most meaningful biological information, as previously described.[45] Further description of the symbols used here is in Table 2.

Gephi was used to visualize and further interrogate the networks.[48] The modularity calculation from Gephi, based on Blondel et al[49] and resolution from Lambiotte et al,[50] was used to classify individual nodes into communities.

To summarize the network approach applied here, we describe the main steps performed:

1. Alignment of gene sequences;
2. Calculation of genetic distances and generation of identity matrix;
3. Calculation of network distances;
4. Identification of best σ;
5. Network generation and analysis under most informative value.

The proposed approach used here requires less than 10 seconds to run on an Acer Intel Core i7-6700 CPU @ 3.40GHz for all data sets tested to date (<100 sequences). The scripts used here can be found on GitHub (https://github.com/deCarvalho90/network_analysis) and the software with a graphical interface is available in the work by Goés-Neto et al.[51]

Results

Phylogenetic analysis

The maximum likelihood phylogeny of *AGAMOUS* genes presented in Figure 1 is consistent with the topology previously published studies of the *AGAMOUS* gene family.[25,26,31] The most likely nucleotide tree had a log likelihood score of −20654.546986. The ML protein tree had poor support for main clades and therefore was not used in subsequent analysis (data not shown).

Table 2. Summary of symbols.

SYMBOL	DENOMINATION	DESCRIPTION
σ	Identity threshold	Threshold value used to build a network, based on similarity values ranging from 0% to 100%. Pairs of sequences that have an identity value greater than or equal to σ mean that they are connected
$M(\sigma)$	Adjacency matrix at σ	Adjacency matrix obtained at a certain value of σ, composed of 0 and 1, representing whether a pair of sequences is connected (represented by 1) or disconnected (represented by 0)
M_{ij}	Element of the adjacency matrix	Element of the adjacency matrix and represents the presence (1) or absence (0) of an edge between sequences *i* and *j* of an adjacency matrix *M*
$\widehat{M}(\sigma)$	Neighborhood matrix	Matrix composed of elements representing the least number of edges necessary to connect a pair of sequences
$\widehat{m}_{ij}$	Element of the neighborhood matrix	Element of the neighborhood matrix and represents the least number of edges connecting sequences *i* and *j*
$\Delta\sigma$	Increments of σ	Value incremented to σ, ie, $\Delta\sigma = 1$ means that the σ increases by 1
$\delta(\sigma, \sigma+\Delta\sigma)$	Network distance between 2 networks	Represents the network distance δ between the networks obtained at σ and $\sigma+\Delta\sigma$

Gymnosperm *AGAMOUS* genes (here termed C/D homologues) form a paraphyly at the base of the unrooted tree. An initial duplication event separates C and D lineage angiosperm genes and likely occurred in the common ancestor of angiosperms. Basal angiosperm C lineage homologues, although clustering with D lineage genes, exhibit expression patterns, and likely function, similar to that of core eudicot C lineage genes. D lineage genes form a monophyletic clade that includes all other angiosperm species included in this study.

Monocot D lineage genes appear as a paraphyly at the base of the D lineage clade; however, the relationships among D lineage genes otherwise are largely consistent with known species relationships. The relationships of C lineage genes are more convoluted. The base of this subtree is a polyphyly including monocot, basal eudicot, and core eudicot genes. At the base of the core eudicots, a second duplication event resulted in the split of the *AGAMOUS* and *PLENA/SHATTERPROOF* (*SHP*) lineages. A third duplication, likely at the base of the Brassicales, resulted in 2 copies of *SHP* genes in this group (*SHP1* and *SHP2*; Figure 1).

Basal angiosperm C lineage genes form a group that diverges after the gymnosperm C/D lineage, but before the angiosperm C/D lineage split. The artificial polyphyletic group of the *paleoAGAMOUS* includes monocot and basal eudicot sequences. While the basal eudicot genes group with other core eudicot *AGAMOUS* genes, monocot *paleoAGAMOUS* genes share a most recent common ancestor with D lineage genes. It is important to notice, however, that the low branch support in many areas of the *AGAMOUS* gene tree poses challenges to the interpretation of the evolutionary relationship between clades.

Network analysis

The network distance graph showed its highest peak at 75% identity, which means that the network generated at that peak is the most distant from the others (Figure 2A). Also, it means that the network presents a clear community structure with relevant evolutionary information. Despite the fact that the network with the biggest distance was obtained at 75% identity, the community structure was already too fragmented to answer questions about the evolution of the gene families analyzed in the phylogeny (Figure S1A). Even though the network obtained at 75% was too fragmented, the network still provided relevant information about the functional divergence of the genes. However, we wanted to see how the community structure would behave in a scenario closer to the phylogeny. To do so, we had to find the network where all sequences were connected in a way that it would still be possible to retrieve a community structure. A similar situation occurred in the work by Carvalho et al,[18] and the problem was solved by analyzing other networks in different peaks. Here, we attempted to solve this problem by analyzing the network at 51% to find the last network where all sequences were connected. However, it was not possible to see a clear community structure in this network due to the high degree of connectivity between nodes (Figure S1B). Finally, in this study, we focused mainly on the network obtained at the identity threshold 67, which meant that 2 sequences had to have an identity value of 67% or higher to be connected. The choice of the network threshold was based on the fact that all sequences in this study were connected, with exception of the out-group sequences, which reflected a scenario similar to the phylogeny.

After applying the modularity calculation (see section "Methods") in the 67% network, it was possible to see the emergence of the community structure of the network, containing 5 communities (C1-C5; Figure 2B). Each one of the communities mainly cluster genes that have similar functions. In C1, 3 out of the 5 nodes from gymnosperm C/D homologues are connected. Even though the 5 nodes are not connected, this result was expected due to the fact that they are

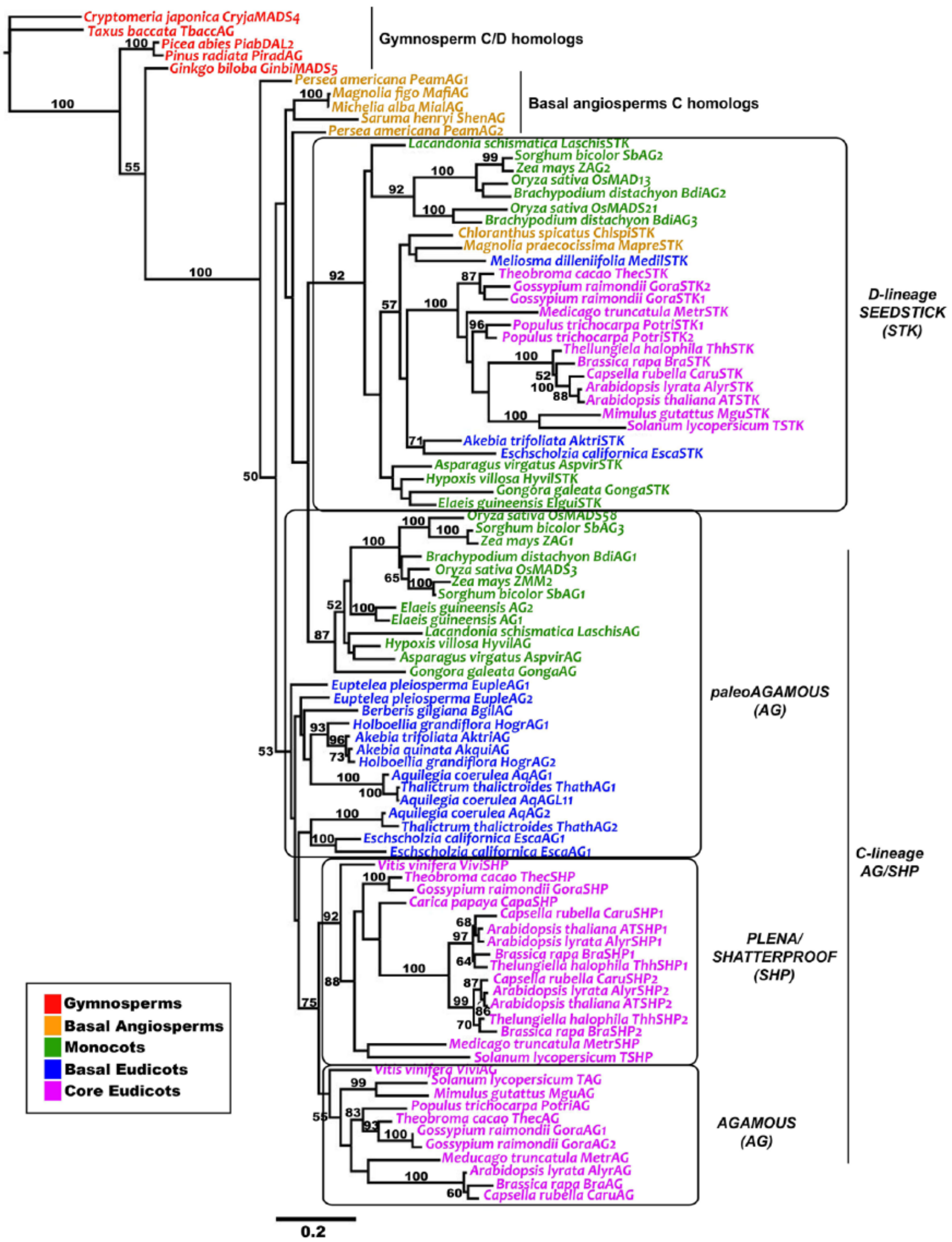

Cryptomeria japonica CryjaMADS4
Taxus baccata TbaccAG
Picea abies PiabDAL2
Pinus radiata PiradAG
Ginkgo biloba GinbiMADS5
Gymnosperm C/D homologs
Persea americana PeamAG1
Magnolia figo MafiAG
Michelia alba MialAG
Saruma henryi ShenAG
Persea americana PeamAG2
Basal angiosperms C homologs
Lacandonia schismatica LaschisSTK
Sorghum bicolor SbAG2
Zea mays ZAG2
Oryza sativa OsMAD13
Brachypodium distachyon BdiAG2
Oryza sativa OsMADS21
Brachypodium distachyon BdiAG3
Chloranthus spicatus ChlspiSTK
Magnolia praecocissima MapreSTK
Meliosma dilleniifolia MedilSTK
Theobroma cacao ThecSTK
Gossypium raimondii GoraSTK2
Gossypium raimondii GoraSTK1
Medicago truncatula MetrSTK
Populus trichocarpa PotriSTK1
Populus trichocarpa PotriSTK2
Thellungiela halophila ThhSTK
Brassica rapa BraSTK
Capsella rubella CaruSTK
Arabidopsis lyrata AlyrSTK
Arabidopsis thaliana ATSTK
Mimulus gutattus MguSTK
Solanum lycopersicum TSTK
Akebia trifoliata AktriSTK
Eschscholzia californica EscaSTK
Asparagus virgatus AspvirSTK
Hypoxis villosa HyvilSTK
Gongora galeata GongaSTK
Elaeis guineensis ElguiSTK
D-lineage SEEDSTICK (STK)
Oryza sativa OsMADS58
Sorghum bicolor SbAG3
Zea mays ZAG1
Brachypodium distachyon BdiAG1
Oryza sativa OsMADS3
Zea mays ZMM2
Sorghum bicolor SbAG1
Elaeis guineensis AG2
Elaeis guineensis AG1
Lacandonia schismatica LaschisAG
Hypoxis villosa HyvilAG
Asparagus virgatus AspvirAG
Gongora galeata GongaAG
Euptelea pleiosperma EupleAG1
Euptelea pleiosperma EupleAG2
Berberis gilgiana BgilAG
Holboellia grandiflora HogrAG1
Akebia trifoliata AktriAG
Akebia quinata AkquiAG
Holboellia grandiflora HogrAG2
Aquilegia coerulea AqAG1
Thalictrum thalictroides ThathAG1
Aquilegia coerulea AqAGL11
Aquilegia coerulea AqAG2
Thalictrum thalictroides ThathAG2
Eschscholzia californica EscaAG1
Eschscholzia californica EscaAG1
paleoAGAMOUS (AG)
Vitis vinifera ViviSHP
Theobroma cacao ThecSHP
Gossypium raimondii GoraSHP
Carica papaya CapaSHP
Capsella rubella CaruSHP1
Arabidopsis thaliana ATSHP1
Arabidopsis lyrata AlyrSHP1
Brassica rapa BraSHP1
Thelungiella halophila ThhSHP1
Capsella rubella CaruSHP2
Arabidopsis lyrata AlyrSHP2
Arabidopsis thaliana ATSHP2
Thelungiella halophila ThhSHP2
Brassica rapa BraSHP2
Medicago truncatula MetrSHP
Solanum lycopersicum TSHP
PLENA/ SHATTERPROOF (SHP)
Vitis vinifera ViviAG
Solanum lycopersicum TAG
Mimulus gutattus MguAG
Populus trichocarpa PotriAG
Theobroma cacao ThecAG
Gossypium raimondii GoraAG1
Gossypium raimondii GoraAG2
Meducago truncatula MetrAG
Arabidopsis lyrata AlyrAG
Brassica rapa BraAG
Capsella rubella CaruAG
AGAMOUS (AG)
C-lineage AG/SHP
Gymnosperms
Basal Angiosperms
Monocots
Basal Eudicots
Core Eudicots
0.2

Figure 1. Phylogenetic tree of the AGAMOUS family genes. Main functional groups are highlighted in black boxes along the tree.

part of the most distant out-group sequences as seen in Figure 1. In C2, however, the functions of the nodes are related to *AG*, *paleoAG* and basal angiosperm C homologues. This might suggest that the *AGAMOUS* genes have retained a function very similar to their basal angiosperm C homologues. In C3, the *SHP* genes are clustered together, but in a different community of the *AG* genes, also suggesting functional divergence. The genes clustered in C4 comprise the *STK* genes. Even though the communities were mostly composed of genes with similar functions, 3 genes exhibited unexpected placements. For instance, the *SHP* gene from *Vitis vinifera* (*ViviSHP*) clustered with other AG genes in C2, instead of with other *SHP* genes in C3. Similarly, *Sorghum bicolor SbAG2*, a *STK* gene, clustered in C5, instead of the expected C4, whereas *Sorghum bicolor SbAG3*, a *paleoAG* gene, clustered in C4, instead of the expected C5.

Finally, the genes clustered in C5 belong to the monocots *paleoAG*. This result might suggest that monocot *paleoAG* genes are evolving under different evolutionary forces than the *paleoAG* and *AG* lineages. Finally, we can notice that the grouping obtained by both methods were consistent with one another by comparing Figures 1 and 2D. Also, the results obtained at the 67% threshold is largely congruent with the one obtained for the protein network generated (Figure S3), obtained at the 60% threshold (highest peak). However, the protein network showed lower resolution because it clustered together *AG*, *eudicot paleoAG*, and *SHP* genes, whereas we see a clear separation of *SHP* from the other genes.

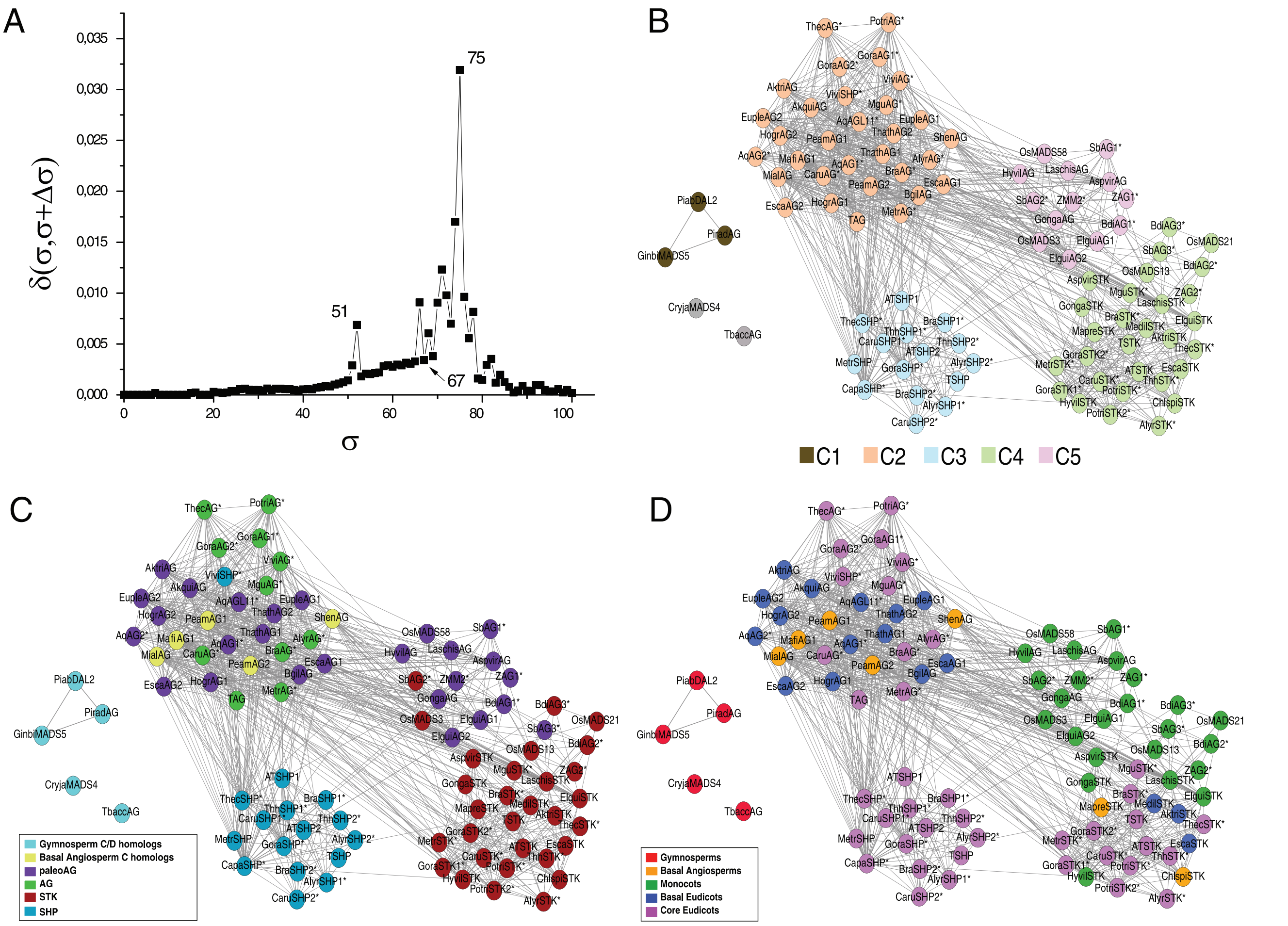
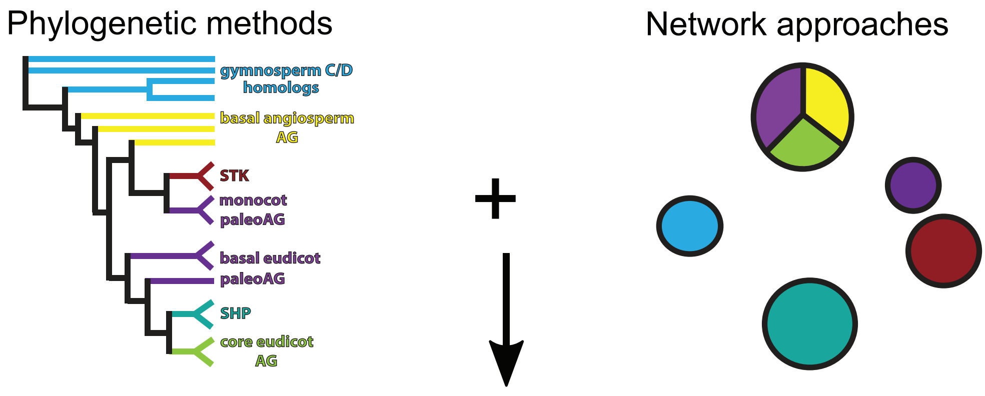


Figure 2. (A) Network distance graph based on the $\delta(\sigma, \sigma+\Delta\sigma)$ distance. The values for the analyzed networks obtained at 51%, 67%, and 75% are marked. (B) Network obtained at 67% identity. Nodes are colored based on the community they belong to (C1-C5), as result of the modularity algorithm (see methods). The sequences that do not belong to any community are represented as gray nodes. Network obtained at 67% identity, colored based on (C) gene function and (D) species phylogenetic placement.



Hypotheses of ancestral relationships

Easy identification of duplication events

Possibility of dating of evolutionary events

More robust account of gene family evolution:

- Hypotheses of functional diversification

- Hypotheses of distinct evolutionary process in action

Independent data to support gene clusters

Allows for graphical visualization of distinct evolutionary processes

Help tree interpretation for branches with low support

Figure 3. Schematic diagram of results based on phylogenetic (left) and network (right) analyses. Potential contributions of each approach, as well as benefits steaming from the combination of both methods are described below the diagrams.

Even though the 75% network showed a fragmented community structure for this study, we can notice that it shows that the *STK* sequences from maize, sorghum, rice, and brachypodium are in a separate community. This information might suggest that *STK* genes from grasses might be undergoing a functional divergence compared with the remaining *STK* genes; however, limitations in gene functional annotation do not enable us to further support this inference.

Both the phylogenetic and network-based analyses returned largely consistent sets of gene clusters. However, the grouping of monocots *paleoAG* sequences in a separate cluster (C5) than other C homologues from basal angiosperm, basal eudicot, and eudicot sequences (jointly clustered in C2) in the network-based analysis suggest 2 testable hypothesis: (1) monocot sequences are undergoing different and independent evolutionary processes when compared with other non-monocot *AG* homologues and (2) non-monocot *AG* sequences are clustered with *euAG* genes due to conservation of function.

Discussion

The use of phylogenetic methods to study gene family evolution has provided vast increases in the understanding of molecular evolution, and the utility of these methods for reconstructing ancestral relationships remains unparalleled. However, in many cases, complex evolutionary processes including neofunctionalization, repeated co-option into new biological roles, as has occurred in independent origins of C4 photosynthesis,[52] high birth/death gene families, and reciprocal gene loss following gene or genome duplication, reconstructing phylogenetic relationships may not be the most effective method for identifying genes with equivalent functional roles. Among the contributions of a network approach to gene family studies is the interpretation of the relationships among gene sequences that are not limited to a bifurcating pattern, which is often the case in a phylogenetic framework. A network approach allows for the emergence of patterns that are not seen otherwise. Here, we propose the use of network-based approach which has complementary sets of strengths and weaknesses to conventional phylogenetic methods and tested the contributions of these methods using data from the well-characterized *AGAMOUS* family of floral transcription factors.

For instance, in the phylogenetic tree, the non-monocot *euAG* and *paleoAG* genes are not clustered with the basal C homologues. Rather, in the networks we notice that these genes are clustered together suggesting a higher functional conservation between them, which is not seen in the tree. Also, from the tree alone we cannot infer whether either *euAG* or *SHP* genes neofunctionalized. However, because all the *euAG* and non-monocot *paleoAG* are clustered together with the ancestral C homologues, and apart from the *SHP* genes, we can infer that the *SHP* genes neofunctionalized, whereas the *euAG* and non-monocot *paleoAG* retained an ancestral function. We believe that a combined approach might help with discerning functional and structural evolution in a way that neither methods can provide on its own.

In agreement with the literature,[26,53] the network-based analysis recovered clusters of *paeloAG* and *AG* genes from basal angiosperms, basal eudicots, and core eudicots, potentially indicating conserved functional roles for the genes included in these clusters despite sequence divergence. In contrast, the position of the basal angiosperms' C lineage in the phylogenetic tree leads to uncertain interpretations of conserved or divergent function with respect to the D lineage. The network-based approach also separated the *STK* and *paleoAG* genes within the monocot lineage, despite the close phylogenetic relatedness of these 2 gene clades, consistent with reports of distinct functional roles for these 2 sets of genes in monocots.[54,55] For instance, *paleoAG* gene from maize has undergone a duplication event in the common ancestor of maize, wheat, and rice[25] which leads to subfunctionalization of these genes that perform functions still related to, but different from *Arabidopsis AG*.[56] A similar process also occurred in rice.[57] These differences may be the reason the monocot *paleoAG* clustered together in the network, but in a different community than the remaining *AG* gene sequences. Moreover, genetic networks of the inflorescence meristems can vary a lot between grasses and eudicots because several changes in these regulatory networks are either only present in grasses or perform a different function in eudicots.[58]

However, network-based approaches to studying gene families bring with them their own set of limitations. Some of these are inherent to the particular methodology used here, whereas others are a result of the relative immaturity of statistical and software tools for applying these methods to the analysis of gene family evolution. For example, a range of statistical methods are widely available for estimating the level of support for individual branches/clades within a given phylogeny, such as jackknife, bootstrap, and posterior probabilities.[59,60] In contrast, methods for calculation of cluster support in a biological context are far less mature, at least for the implementation employed here. The use of sequence identity as a measure of distance, while computationally tractable, also means discarding a great deal of information on the frequency of different types of substitutions at both the nucleotide and amino acid levels which can be incorporated into many modern phylogenetic algorithms.[61]

Figure 3 summarizes the contributions and relative strengths and weaknesses of phylogenetic and network-based approaches to the study of gene family evolution. We propose that the combination of both methods can provide more assessment of both functional and historical relationships between sequences than either approach alone.

Conclusions

Investigating the contributions of a particular network-based approach to the study of the evolution of a well-known family of transcription factor genes involved in floral development supports the idea that network-based approaches, when used in conjunction with phylogenetic methods, can be used to improve our understanding of functional conservation or divergence within gene family evolution. The network-based analysis of gene families used here currently lacks the robust ecosystem of computational tools and statistical approaches developed for phylogenetic analysis; however, it can already provide an independent assessment of relationship structures which can aid in the interpretation of phylogenetic data, especially in areas of the tree exhibiting low branch support. In particular, network analysis can be used to generate testable hypotheses regarding the conservation or divergence of gene function in cases of potential subfunctionalization or neofunctionalization. In combination, we believe that these methods provide a robust framework that expands the power of gene family evolution studies.

Author Contributions

DSC, JCS, and AMRA designed the analysis, analyzed the results and wrote the paper; DSC and AMRA conducted the analyses.

REFERENCES

1. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278:631–637.
2. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–338.
3. Guo YL. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *Plant J*. 2013;73:941–951.
4. Christin PA, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. Parallel recruitment of multiple genes into C4 photosynthesis. *Genome Biol Evol*. 2013;5:2174–2187.
5. Davidson RM, Gowda M, Moghe G, et al. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J*. 2012;71:492–502.
6. Dewey CN. Positional orthology: putting genomic evolutionary relationships into context. *Brief Bioinform*. 2011;12:401–412.
7. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A*. 2007;104:8397–8402.
8. Schnable JC, Freeling M, Lyons E. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol*. 2012;4:265–277.
9. Pophaly SD, Tellier A. Population level purifying selection and gene expression shape subgenome evolution in maize. *Mol Biol Evol*. 2015;32:3226–3235.
10. Bekaert M, Edger PP, Pires JC, Conant GC. Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell*. 2011;23:1719–1728.
11. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci U S A*. 2007;104:2043–2049.
12. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. 1999;96:3801–3806.
13. Ozkan H, Levy AA, Feldman M. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. *Plant Cell*. 2001;13:1735–1747.
14. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5:101–113.
15. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104:8685–8690.
16. Amrine KC, Blanco-Ulate B, Cantu D. Discovery of core biotic stress responsive genes in *Arabidopsis* by weighted gene co-expression network analysis. *PLoS ONE*. 2015;10:e0118731.
17. Góes-Neto A, Diniz MV, Santos LB, et al. Comparative protein analysis of the chitin metabolic pathway in extant organisms: A complex network approach. *Biosystems*. 2010;101:59–66.
18. Carvalho DS, Andrade RF, Pinho ST, et al. What are the evolutionary origins of mitochondria? a complex network approach. *PLoS ONE*. 2015;10:e0134988.
19. Han JDJ. Understanding biological functions through molecular networks. *Cell Res*. 2008;18:224–237.
20. Chasman D, Siahpirani AF, Roy S. Network-based approaches for analysis of complex biological systems. *Curr Opin Biotechnol*. 2016;39:157–166.
21. Fischer EK, Ghalambor CK, Hoke KL. Can a network approach resolve how adaptive vs nonadaptive plasticity impacts evolutionary trajectories? *Integr Comp Biol*. 2016;56:877–888.
22. Pinyopich A, Ditta GS, Savidge B, et al. Assessing the redundancy of MADS-box genes during carpel and ovule development. *Nature*. 2003;424:85–88.
23. Bowman JL, Smyth DR, Meyerowitz EM. Genes directing flower development in *Arabidopsis*. *Plant Cell*. 1989;1:37–52.
24. Bowman JL, Smyth DR, Meyerowitz EM. Genetic interactions among floral homeotic genes of *Arabidopsis*. *Development*. 1991;112:1–20.
25. Kramer EM, Jaramillo MA, Di Stilio VS. Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. *Genetics*. 2004;166:1011–1023.
26. Dreni L, Kater MM. MADS reloaded: evolution of the AGAMOUS subfamily genes. *New Phytol*. 2014;201:717–732.
27. Rutledge R, Regan S, Nicolas O, et al. Characterization of an AGAMOUS homologue from the conifer black spruce (*Picea mariana*) that produces floral homeotic conversions when expressed in Arabidopsis. *Plant J*. 1998;15:625–634.
28. Zhang P, Tan HT, Pwee KH, Kumar PP. Conservation of class C function of floral organ development during 300 million years of evolution from gymnosperms to angiosperms. *Plant J*. 2004;37:566–577.
29. Pan IL, McQuinn R, Giovannoni JJ, Irish VF. Functional diversification of AGAMOUS lineage genes in regulating tomato flower and fruit development. *J Exp Bot*. 2010;61:1795–1806.
30. Lovisetto A, Baldan B, Pavanello A, Casadoro G. Characterization of an AGAMOUS gene expressed throughout development of the fleshy fruit-like structure produced by *Ginkgo biloba* around its seeds. *BMC Evol Biol*. 2015;15:139.
31. Pabón-Mora N, Wong GKS, Ambrose BA. Evolution of fruit development genes in flowering plants. *Front Plant Sci*. 2014:5:300.
32. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2005;23:254–267.
33. Francis AR, Steel M. Which phylogenetic networks are merely trees with additional arcs? *Syst Biol*. 2015;64:768–777.
34. Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*. 2016;12:e1005896.
35. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
36. Goujon M, McWilliam H, Li W, et al. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*. 2010;38:W695–W699.
37. Kearse M, Moir R, Wilson A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–1649.
38. Eddy S. HMMER3: a new generation of sequence homology search software. http://hmmer.janelia.org. Published 2010.
39. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 2012;9:772–772.
40. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003;52:696–704.
41. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59:307–321.
42. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993;10:512–526.
43. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–1874.
44. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 2004;101:11030–11035.
45. Andrade RF, Rocha-Neto IC, Santos LB, et al. Detecting network communities: an application to phylogenetic analysis. *PLoS Comput Biol*. 2011;7:e1001131.
46. Andrade RF, Miranda JG, Lobão TP. Neighborhood properties of complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2006;73:046101.
47. Andrade RF, Pinho ST, Lobao TP. Identification of community structure in networks using higher order neighborhood concepts. *Int J Bifurcat Chaos*. 2009;19:2677–2685.
48. Bastian M, Heymann S, Jacomy M, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*. 2009;8:361–362.
49. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech: Theory E*. 2008;2008:P10008.
50. Lambiotte R, Delvenne JC, Barahona M. Laplacian dynamics and multiscale modular structure in networks (arXiv preprint arXiv:0812.1770). Published 2008.

51. Goés-Neto A, Diniz MV, Carvalho DS, et al. Comparison of complex networks and tree-based methods of phylogenetic analysis and proposal of a bootstrap method. *Peer J*. 2018;6:e4349.
52. Edwards E, Aliscioni S, Bell H, et al. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol*. 2011;193:304–312.
53. Kim S, Koh J, Yoo MJ, et al. Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators. *Plant J*. 2005; 43:724–744.
54. Favaro R, Immink R, Ferioli V, et al. Ovule-specific MADS-box proteins have conserved protein-protein interactions in monocot and dicot plants. *Mol Genet Genomics*. 2002;268:152–159.
55. Dreni L, Jacchia S, Fornara F, et al. The D-lineage MADS-box gene OsMADS13 controls ovule identity in rice. *Plant J*. 2007;52:690–699.
56. Mena M, Ambrose BA, Meeley RB, Briggs SP, Yanofsky MF, Schmidt RJ. Diversification of C-function activity in maize flower development. *Science*. 1996;274:1537–1540.
57. Yamaguchi T, Lee DY, Miyao A, Hirochika H, An G, Hirano HY. Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in Oryza sativa. *Plant Cell*. 2006;18:15–28.
58. Thompson BE, Hake S. Translational biology: from *Arabidopsis* flowers to grass inflorescence architecture. *Plant Physiol*. 2009;149:38–45.
59. Bremer K. Branch support and tree stability. *Cladistics*. 1994;10:295–304.
60. Sitnikova T, Rzhetsky A, Nei M. Interior-branch and bootstrap tests of phylogenetic trees. *Mol Biol Evol*. 1995;12:319–333.
61. Lio P, Goldman N. Models of molecular evolution and phylogeny. *Genome Res*. 1998;8:1233–1244.