J Mol Evol
DOI 10.1007/s00239-016-9750-z

ORIGINAL ARTICLE

Diversification of R2R3-MYB Transcription Factors in the Tomato Family Solanaceae

Daniel J. Gates[1,2] · Susan R. Strickler[3] · Lukas A. Mueller[3] · Bradley J. S. C. Olson[4] · Stacey D. Smith[2]

Received: 1 November 2015 / Accepted: 15 June 2016

Abstract MYB transcription factors play an important role in regulating key plant developmental processes involving defense, cell shape, pigmentation, and root formation. Within this gene family, sequences containing an R2R3 MYB domain are the most abundant type and exhibit a wide diversity of functions. In this study, we identify 559 R2R3 MYB genes using whole genome data from four species of Solanaceae and reconstruct their evolutionary relationships. We compare the Solanaceae R2R3 MYBs to the well-characterized *Arabidopsis thaliana* sequences to estimate functional diversity and to identify gains and losses of MYB clades in the Solanaceae. We identify numerous R2R3 MYBs that do not appear closely related to *Arabidopsis* MYBs, and thus may represent clades of genes that have been lost along the *Arabidopsis* lineage or gained after the divergence of Rosid and Asterid lineages. Despite differences in the distribution of R2R3 MYBs across functional subgroups and species, the overall size of the R2R3 subfamily has changed relatively little over the roughly 50 million-year history of Solanaceae. We added our information regarding R2R3 MYBs in Solanaceae to other data and performed a meta-analysis to trace the evolution of subfamily size across land plants. The results reveal many shifts in the number of R2R3 genes, including a 54 % increase along the angiosperm stem lineage. The variation in R2R3 subfamily size across land plants is weakly positively correlated with genome size and strongly positively correlated with total number of genes. The retention of such a large number of R2R3 copies over long evolutionary time periods suggests that they have acquired new functions and been maintained by selection. Discovering the nature of this functional diversity will require integrating forward and reverse genetic approaches on an -omics scale.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-016-9750-z) contains supplementary material, which is available to authorized users.

✉ Daniel J. Gates
dgates@huskers.unl.edu

Stacey D. Smith
Stacey.D.Smith@colorado.edu

[1] School of Biological Sciences, University of Nebraska, Lincoln 68588, USA

[2] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder 80309, USA

[3] Boyce Thompson Institute for Plant Research, Ithaca, NY 14853, USA

[4] Division of Molecular, Cellular and Developmental Biology, Kansas State University, Manhattan,, KS 66506, USA

Introduction

In plants, the majority of genes belong to multigene families, which can vary up to three orders of magnitude in size (Zhang 2003; Guo 2013). For example, the anthocyanin pathway enzyme, dihydroflavonol-reductase, comprises a small family of up to three copies, whereas the F-Box proteins involved in substrate recognition possess hundreds of copies, with over 600 in rice and *Arabidopsis* (Yang et al. 2008). While some of the differences in family size may reflect methodological approaches to gene discovery and classification (Frech and Chen 2010), much of this variation is likely attributable to historical differences in rates of gene duplication and retention of duplicate copies across families (Clegg et al. 1997; Adams and Wendel 2005).

This study focuses on the R2R3 subfamily of MYB transcription factors, a group particularly notable for its

Published online: 30 June 2016

expansion in plants. MYBs are a large class of transcription factors found in all eukaryotic organisms and characterized by one or more repeats of the MYB domain (Lipsick 1996; Kranz et al. 2000). Each repeat forms a helix-turn-helix structure, and studies in the human c-MYB indicate that the C-terminal α-helix directly binds to the major DNA groove (Ogata et al. 1996). In many cases, MYBs interact with other proteins, e.g., WD40 and bHLH proteins, in order to regulate their target genes (Grotewold et al. 1994; Ramsay and Glover 2005).

The R2R3 MYBs form the largest subfamily of MYB transcription factors in plants, although the number of R2R3 copies varies fivefold across taxa (Feller et al. 2011; Du et al. 2012). R2R3 MYBs are diagnosable by their two imperfect MYB repeats that follow the R2 and R3 structure of the c-MYB (Kranz et al. 2000). In *Arabidopsis*, this group of MYB genes performs a wide array of functions, including specification of epidermal cell fate (Oppenheimer et al. 1991; Wada et al. 1997), regulation of flavonoid biosynthesis (Mehrtens et al. 2005), and response to environmental and hormonal cues (Urao et al. 1996; Abe et al. 1997). Given the diversity of R2R3s, the subfamily has been subdivided into 23 subgroups based on phylogenetic relationships and function in *Arabidopsis* (Romero et al. 1998; Kranz et al. 2000; Stracke et al. 2001; Dubos et al. 2010).

Beyond *Arabidopsis*, the phylogenetic and functional diversity of R2R3 MYBs has begun to be explored in a range of taxa across angiosperms (cucumber, Li et al. 2012; apple, Cao et al. 2013; salvia, Li and Lu 2014; popular, rice, maize, switchgrass, Zhao and Bartley 2014). The tomato family offers a particularly interesting system for tracing the diversification of R2R3 MYBs because of the relatively large number of available genomes (Potato Genome Sequencing Consortium 2011; Bombarely et al. 2012; Tomato Genome Consortium 2012; Kim et al. 2014) and knowledge of the key roles of these MYBs in morphological and biochemical phenotypes (Borovsky et al. 2004; Pattanaik et al. 2010). For example, R2R3 MYBs are important regulators of shoot-branching development in tomato (*Solanum lycopersicum*) (Busch et al. 2011). In *Nicotiana benthamiana*, an R2R3 MYB induces production of phenylpropanoid-polyamine conjugates that provide defense against herbivory (Kaur et al. 2010). Multiple R2R3 MYBs are responsible for differences in flower pigmentation across species of *Petunia*, *Iochroma*, and *Nicotiana*, and play a role in the evolution of plant–pollinator interactions (Quattrocchio et al. 1999; Smith and Rausher 2011; Hermann et al. 2013).

Existing research suggests that the overall number of R2R3 MYBs present in *Solanum* is similar to other Asterid angiosperms (Zhao et al. 2014). However, little is known about the size of the R2R3 subfamily in other important Solanaceae species. In the present study, we apply bioinformatic and statistical phylogenetic approaches to reconstruct the expansion of the R2R3 MYB subfamily across angiosperms as a whole and within the Solanaceae in particular. This work builds on publicly available genomes in Solanaceae (Potato Genome Sequencing Consortium 2011; Bombarely et al. 2012; Tomato Genome Consortium 2012; Kim et al. 2014), an unpublished draft genome for *Iochroma cyaneum* (Gates et al. unpublished) as well as studies of R2R3 MYBs in other taxa (Matus et al. 2008; Cao et al. 2013). Specifically, we aim to address three major questions: First, how does the size and diversity of functional groups in Solanaceae compare to that present in *Arabidopsis*? Second, after combining our Solanaceae data with information from other plant genomes, does the R2R3 MYB subfamily show consistent increases in copy number (i.e., are new duplicates of R2R3s generally retained) across the phylogeny? Finally, how much of the variation in R2R3 gene subfamily size across land plants is explained by differences in genome size? To our knowledge, this study represents the first attempt to reconstruct ancestral R2R3 gene subfamily sizes and statistically estimate MYB gains and losses across the phylogeny. Collectively, these analyses will provide both a broad-scale picture of the evolution of this gene subfamily during land plant history as well as a detailed look at the shifts in functional diversity of R2R3 MYBs within the economically important Solanaceae.

Methods

Taxon Sampling Within Solanaceae

The Solanaceae includes many species cultivated as crops (e.g., tomato, potato, chili pepper, eggplant, tobacco) and ornamentals (e.g., *Petunia*, *Nicotiana*, *Iochroma*). Currently, there are approximately 2700 recognized species in the family, nearly half of which fall into the genus *Solanum* (Hunziker 2001; Särkinen et al. 2013). With the recent addition of three hot pepper genomes and three more resequenced *Nicotiana* genomes, there are nine publicly available genomes within the Solanaceae: *S. lycopersicum* (tomato) (Tomato Genome Consortium 2012), *Solanum tuberosum* (potato) (Potato Genome Sequencing Consortium 2011), *N. benthamiana* (Bombarely et al. 2012), three *Nicotiana tabacum* (tobacco) (Sierro et al. 2014), and three different varieties of *Capsicum annuum* (chili pepper) (Kim et al. 2014; Albert and Chang 2014). For identification of Solanaceae R2R3 MYBs, we used the *S. lycopersicum*, *S. tuberosum*, and *N. benthamiana* genomes available as well as a draft assembly of the *I. cyaneum* genome (described below). These species are distributed across the two subfamilies (Solanoideae and Nicotianoideae) of the large

$X = 12$ clade (Olmstead et al. 2008) and thus span both shallow and relatively deep divergences (8 million years ago (Ma) for the two *Solanum* spp. to 24 Ma for *Solanum* versus *Nicotiana* (Särkinen et al. 2013)). In addition to being a sister to the rest of the selected Solanaceae species, *N. benthamiana* is a polyploid (Knapp et al. 2004; Bombarely et al. 2012; Wang and Bennetzen 2015) while *I. cyaneum*, *S. lycopersicum*, and *S. tuberosum* are diploids. Although *C. annuum* was not included in our phylogenetic analyses because its genome became available only recently, we did estimate the number of R2R3 MYBs in this species to trace changes in gene subfamily size across Solanaceae and land plants overall (see below).

Genome Assembly for *Iochroma cyaneum*

As part of efforts to expand our knowledge about genomic diversity across the Solanaceae (including noncrop species), we constructed a low-coverage genome of *I. cyaneum*, an Andean shrub that is being developed as a model for floral evolution (Smith and Baum 2006; Smith and Rausher 2011). We sampled a single individual of an accession first cultivated by William D'Arcy at the Missouri Botanical gardens (voucher: Smith, 265 (WIS)). For this genome, we sequenced four lanes of standard genomic libraries of 400-bp fragments, and two lanes of mate pair sequencing of 2- and 5-kb libraries, respectively. All lanes were sequenced on an Illumina Hi-Seq 2000 through the Weill Cornell Genomics Facility (http://corefacilities.weill.cornell.edu/genomics.html). All Illumina libraries were sequenced as 100-bp paired end reads. All sequences were first quality checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and we removed contaminating sequences and bases with quality scores less than $q = 30$ with fastq-mcf (https://code.google.com/p/ea-utils/wiki/FastqMcf). For trimming of the internal junction adapters in the mate pair library, we used a custom Perl script that is available upon request. We used SOAP2 (Li et al. 2008) for de novo assembly and gap closing using sequences from four standard libraries and two mate pair libraries for scaffolding of the assembled contigs. We varied kmer sizes in assembly from 21–63, with 63 giving the best assembly statistics. As assembly errors and artifacts were a concern, especially in such a large gene family with multiple highly conserved motifs, we supplemented our low-copy genome with an *I. cyaneum* transcriptome. To construct the transcriptome, we used a Trinity (Grabherr et al. 2011) de novo assembly on floral RNA sequences that were sequenced on ½ lane of an Illumina Hi-Seq 2000 at the University of Missouri's genomics core facility (https://web.rnet.missouri.edu/biotech/dnacore/). Because the *Iochroma* sequences came from two independent raw datasets (genome and transcriptome), we ensured that we were not including technical duplicates as separate genes by collapsing any sister sequences (*I.cyaneum* sequences more closely related to another *I. cyaneum* sequence than nearest *N. benthamiana*, *S. lycopersicum*, *S. tuberosum* sequence) into a single sequence unless both of those sequences originated from the genome build.

Identification of R2R3 MYBs in Solanaceae

In order to assess MYB divergence within the Solanaceae, we used sequences from three publicly available genomes as well as our *I. cyaneum* genomic resources. We downloaded protein sequences and coding sequences for *N. benthamiana*, *S. tuberosum* and *S. lycopersicum* from solgenomics.net (Bombarely et al. 2011). We used an ab initio genomic scan approach to identify coding sequences and putative gene-containing regions as well as their respective amino acid translations for the *I. cyaneum* low-coverage genome and transcriptome using Augustus 2.7 with default settings (Stanke and Morgenstern 2005). We filtered these complete coding sequence datasets using a tailored bioinformatic pipeline. First, we trained an HMM profile for the 126 published *A. thaliana* R2R3 MYBs by aligning the amino acid sequences with ClustalX (Larkin et al. 2007) and then constructed the HMM profile using the program HMMER (http://hmmer.janelia.org) with default parameter values. We used an e-value cutoff of 1e-60 as it represented a discrete break in the bimodally distributed e-values output by HMMer. We tested the accuracy of this method by applying our same HMM profile to the *Mimulus guttatus* amino acid sequences that should contain approximately 119 R2R3 MYBs (Feller et al. 2011). Without changing any settings or the criteria for inclusion based on e-value output from our above HMM search, *M. guttatus* amino acid sequences were estimated to contain 114 R2R3 MYBs. Thus, our pipeline produces similar estimates of gene subfamily members as previous studies and is suitable for the downstream comparative genomic analyses.

Phylogenetic Inference for Solanaceae R2R3 MYBs

We created a phylogenetic tree for R2R3 MYBs from Solanaceae and *A. thaliana* (*At*) in order to identify clades that are closely related to the established *At* functional subgroups (Stracke et al. 2001). We aligned amino acid sequences using both ClustalX (Larkin et al. 2007) and MAFFT (Katoh et al. 2002), and chose the ClustalX alignment as it gave a better reconstruction of the conserved R2R3 domain and a less gappy alignment. We also checked the alignment by eye to remove any likely pseudogenes with substantial insertions or deletions within the R2R3 MYB domains. We used PAL2NAL (Suyama et al.

2006) to align the CDS sequences exactly following the aligned protein sequences. We constructed a maximum likelihood tree based on the nucleotide alignment using RAxML (Stamatakis 2006) with a GTR+G+I model and estimated clade support using 100 rapid bootstrap replicates. We also inferred amino acid phylogeny using RAxML with the PROTGAMMAJTT model and estimated clade support with 100 rapid bootstrap replicates. Since there is no clear outgroup to the R2R3 MYB gene family, both trees were midpoint rooted using the phangorn R package (Schliep 2011). We assigned Solanaceae sequences and clades to the R2R3 MYB subgroups defined by Stracke et al. (2001) if they were more closely related to *At* sequences than to other Solanaceae sequences and formed a clade with high support (>70 %). Solanaceae sequences with no closely related *At* sequences were not placed in any functional subgroup; these may represent cases in which the ancestral gene lineage was lost in *Arabidopsis* or new lineages were gained along the Asterid lineage that contains Solanaceae.

Gene Family Evolution Across Angiosperms

In order to place R2R3 gene diversity in Solanaceae into a broader context, we traced the evolution of subfamily size across a sample of 14 other land plant lineages. We constructed our dataset by adding our counts of gene subfamily size to the counts published by Feller et al. (2011) as well as counts from *Cucumis sativus*, *Malus x. domestica*, and *Salvia mitorhiza* (Li et al. 2012; Cao et al. 2013; Li and Lu 2014). We used the same HMM approach described above to estimate the number of R2R3 MYBs in *C. annuum*. The species tree topology and the depths of internal nodes (in millions of years) above the family level were taken from Bell et al. (2010). We used Paterson et al. (2004) to determine the dates of Poaceae nodes (*Zea mays*, *Sorghum bicolor*, and *Oryza sativa*), and the divergence times of major Solanaceae lineages were taken from Wu and Tanksley (2010). We used the program CAFE (De Bie et al. 2006) to identify gains and losses of R2R3 MYB gene copies along the branches of the plant phylogeny. CAFE estimates changes in gene subfamily size by simulating branches using a birth/death model and identifies significant expansions or contractions in subfamily size (Hahn et al. 2005). In addition, we used CAFE to identify the single internal branch with the greatest shift in diversification rate. We created a custom script to move across each internal branch, in each case, fitting a two-rate model (one background diversification rate and a second rate for the selected branch and its descendants). The internal branch resulting in the greatest increase in likelihood was thus considered to correspond to the strongest shift in diversification rate across the phylogeny.

To assess whether the number of R2R3 MYBs may be related to overall changes in genome size or total gene number (e.g., following polyploidization or segmental duplications), we performed phylogenetic generalized least squares (PGLS) (Grafen 1989; Hansen and Martins 1996). This approach allows us to test whether increases in overall genome size or gene number explain the large numbers of MYBs in some taxa, after accounting for phylogenetic relatedness. For all species except for *Salvia miltiorrhiza* and *I. cyaneum,* we gathered c-values from the Kew database (data.kew.org/cvalues/). There is currently no c-value for *S. miltiorrhiza* in the Kew database. To approximate the genome size, we used the average c-value size of the ten other *Salvia* species that were entered in the database since there were only modest size differences between the accessions (mean 0.62; SD 0.19). The c-value for *I. cyaneum* was measured with flow cytometry using *S. lycopersicum* as the reference. We validated this result by comparison with the estimated genome size in our draft genome assembly. We gathered total number of genes from assembly statistics for all species except *I. cyaneum* and *S. miltiorrhiza* as these species lack a completed reference genome. After obtaining c-values and gene numbers for all species, we conducted phylogenetic regressions with PGLS using the number of R2R3 MYBs as a response variable and the c-values or gene number as the independent variable. PGLS, implemented in the nlme package (Pinheiro et al. 2011), generalizes independent contrasts and allows for a wider range of models of trait evolution. For both PGLS analyses, we used the Brownian motion model, as model selection with AIC indicated that the more complex Ornstein–Uhlenbeck model (Butler and King 2004) did not provide a significantly better fit.

Results

Phylogenetic Analysis and R2R3 Diversity in Solanaceae

Our bioinformatics pipeline recovered 559 R2R3 MYBs across the four Solanaceae genomes analyzed. The nucleotide alignment used for phylogenetic inference and raw tree data is available on Dryad (datadryad.org) at doi:10.5061/dryad.d63t5. The raw *I. cyaneum* reads will also be available on Solgenomics.net (Bombarely et al. 2011) upon publication. For tomato, our pipeline recovered very similar estimates of R2R3 MYBs as in a previous study (Zhao et al. 2014). Using the same tomato genome build, Zhao et al. found 121 R2R3 MYBs, 119 of which corresponded to full-length coding sequences. Our pipeline also found 119 complete sequences. Even though the *Iochroma* draft genome has a lower N50 score and lower

coverage than the published genomes (Table 1), we recovered a similar number of R2R3 genes (110) as in other $X = 12$ clade Solanaceae (range 111–119). *N. benthamiana* is a recent polyploid (Leitch et al. 2008), but the number of copies (171) is much less than twice the number in the other Solanaceae, which are all diploids. It is possible that the count for *Nicotiana* may represent an underestimate if homeologous copies were collapsed during genome assembly or were not included in gene models due to loss of expression (Coate and Doyle 2010). It is more likely, however, that copies have been lost through diploidization. Most *Nicotiana* polyploids are $n = 48$, but *N. benthamiana* is $n = 38$, suggesting that it has lost ten chromosomes. Also, the *N. benthamiana* genome is 40 % smaller than other polyploids like *N. tabacum* and *N. rustica* (Wang and Bennetzen 2015).

The maximum likelihood phylogeny of the 559 R2R3 MYBs from the four Solanaceae genomes resolves many major clades within the gene subfamily (Fig. 1). Across the tree, 81 % of nodes had 70 % or greater bootstrap support (S1 Fig.). Most clades recover a similar topology to the species tree, which places the two *Solanum* species (potato and tomato) together, and *I. cyaneum* and *N. benthamiana* as successive sister species. We also observed apparent Solanaceae-specific and species-specific duplications within functional groups. For example, AtMYB37 and AtMYB38, members of the S14 subgroup, are sister to 16 Solanaceae sequences, which are grouped into three clades, each showing the expected species tree relationships.

By including *At* sequences, we putatively assigned 236 of the 559 Solanaceae sequences to 14 of the 23 functional subgroups. For example, subgroup S16 comprises three copies in *Arabidopsis* that appear to be involved in light signaling and hypocotyl elongation (Dubos et al. 2010). These *At* S16 sequences fall into a well-supported clade with seven Solanaceae sequences, which were thus putatively assigned to this functional group (S1 Fig., Table S1). Two *At* subgroups (S12 and S15) did not have any Solanaceae orthologs. For two *At* subgroups, there are likely Solanaceae orthologs, but we could not propose any assignments because of low bootstrap support. In addition, several subgroups (S10/S11/S24 and S18/S20) did not appear as monophyletic groups (are shown merged in Fig. 1). Thus we outlined the larger clade, but did not provide clade-by-clade assignments for Solanaceae sequences within the larger clades.

Subgroup assignments were largely robust to the choice of protein or DNA sequence in phylogeny reconstruction. All well-supported functional group clades (BS > 70 %) that we identified in the DNA sequence phylogeny were also present in the protein phylogeny (S2 Fig.), and nine of these 14 clades received strong support. Backbone relationships varied, but this was anticipated given the relatively low support for these nodes. One relationship that is well supported in the DNA phylogeny but not present in the protein phylogeny is the sister group relationship between subgroups S5 and S6. These two subgroups are highly similar in function as they regulate different upstream (S5) and downstream (S6) sections of the anthocyanin/flavonol pathway. In the protein tree, S5 and S6 form a paraphyletic grade, while in the DNA phylogeny they appear as a clade. Previous phylogenies support the latter pattern (Kranz et al. 1998; Stracke et al. 2001; Li and Lu 2014), suggesting that DNA phylogeny may be more reliable for resolving these shallow intersubgroup relationships.

Despite the lack of resolution in some parts of the phylogeny, we detected interesting patterns where it was possible to associate Solanaceae sequences with functionally categorized *At* sequences. Among the well-supported 14 subgroups, subgroup 14, involved in regulating growth and organ formation (Dubos et al. 2010), contained the largest number of total copies in Solanaceae (Table S1, Fig. 2). This subgroup is also one of the largest subgroups in *At* with six genes. With the highest total number of R2R3 MYBs, *Nicotiana* also had the greatest number of genes in all of the subgroups except subgroups S6, S7, and S18, where it had fewer sequences than one or more of the other sampled taxa (Fig. 2, Table S1). Looking across subgroups, we observed wide variation in functional content across species. While the numbers of genes in some subgroups are relatively constant across the species (e.g., S9, S23; Fig. 2), others are enriched in particular taxa (e.g., S14), and others are entirely absent from some taxa (e.g., S5, S6 in potato).

Table 1 Assembly descriptions and sources for four genomes used to characterize R2R3 diversity in Solanaceae

Taxon	Total Length (GB)	Scaffolds and Contigs	N50 length	Coverage	Citation
Solanum lycopersicum	0.74	12	–	NA	Tomato Genome Consortium (2012)
Solanum tuberosum	0.73	649	1,318,000	84×	Potato Genome Sequencing Consortium (2011)
Nicotiana benthamiana	2.46	602,802	29,049	63×	Bombarely et al. (2012)
Iochroma cyaneum	3.23	1,029,317	17,952	77×	Gates et al. unpubl.

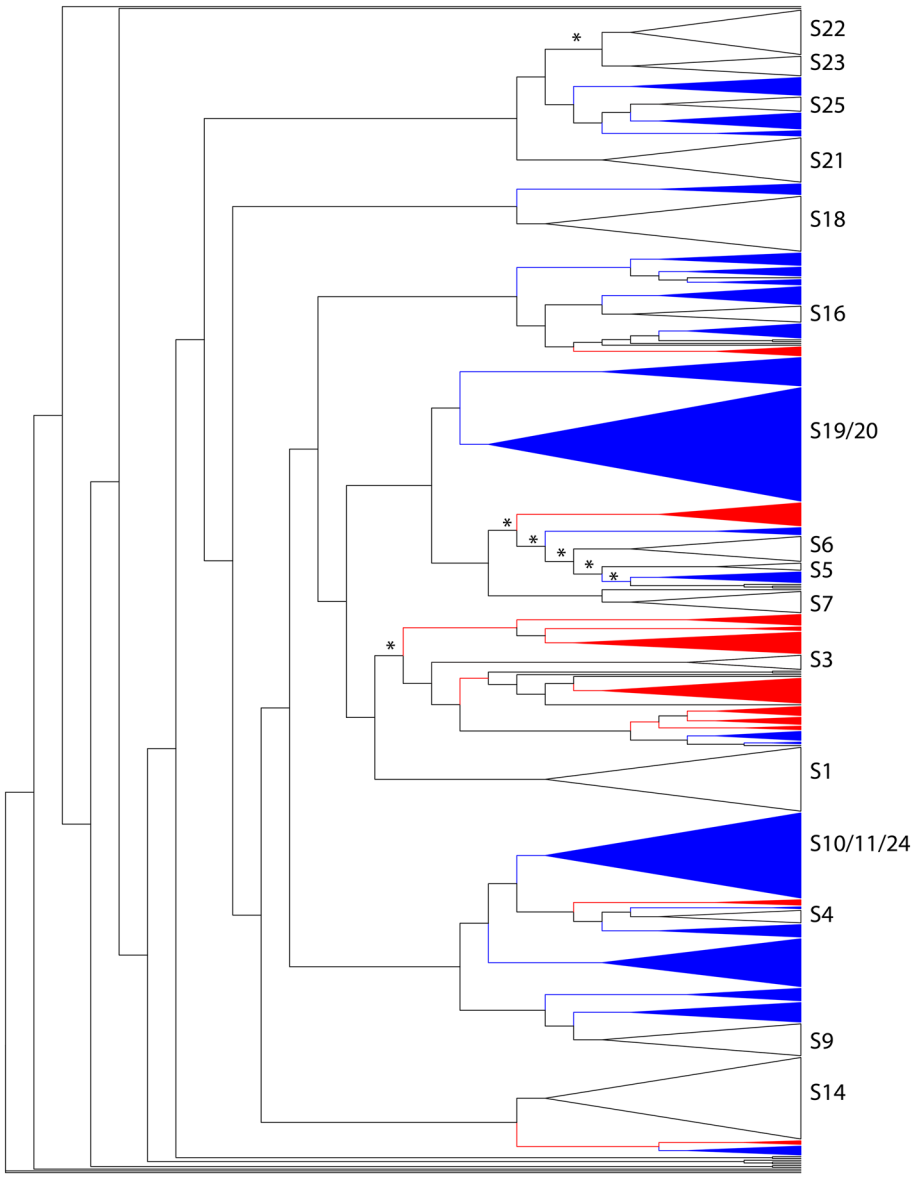

Fig. 1 Summary of maximum likelihood tree for Solanaceae R2R3 MYBs. All clades collapsed to triangles represent strongly supported nodes (BS > 70 %). *White triangles* contain Solanaceae sequences that are closely related to *A. thaliana* functional subgroup members (S1 Fig) and are putatively assigned to those subgroups (Table S1). *Blue clades* contain Solanaceae sequences that are closely related to some *At* sequences, but were not assigned to a functional group either because those clades have low support (S1 Fig.), have *At* sequences whose function is unclassified (Dubos et al. 2010), or comprise a mixture of sequences from multiple subgroups (e.g., S10/11/24 in the figure). *Red clades* did not contain any *At* sequences (see also S1 Fig.). Outside of the collapsed clades, any branches with greater than 70 % *bootstrap* support are indicated with an asterisk

Analysis of R2R3 MYB Subfamily Evolution Across Terrestrial Plants

The CAFE analyses indicated that 20 branches (7 internal and 13 terminal) have experienced significant shifts in R2R3 gene subfamily size when compared to null distributions generated under a constant birth–death process (Fig. 3). Estimates of ancestral gene subfamily size revealed that the largest expansion occurred on the branch leading to *Glycine max* where gene subfamily size increases from 156 to 288 (Fig. 3). The greatest reduction in gene subfamily occurs on the branch leading to *Cucumis sativus* where the number of copies shifts from 156 to 55. Overall, of the 20 branches with significant changes, 11 are expansions and nine are contractions (Fig. 3). We also iterated a partitioned model across all internal branches that allowed a single shift in diversification rate. The partitioned analysis placed the optimal position for the single-rate shift at the branch leading to eudicots (χ^2 test = 378, $p < 0.0001$). The background rate (outside of eudicots) and rate within eudicots were 0.0024 and 0.0087 gains+losses/gene/million years, respectively.

As a random birth/death process appears to be a poor predictor of gene family size, we used phylogenetic comparative methods (PGLS) to investigate whether differences in genome size or total gene number predict the variation. The PGLS analyses supported a significant positive relationship between the number of R2R3 MYBs and genome size ($T_{18} = 2.12$, $p = 0.0496$) and with the total number of genes ($T_{16} = 3.76$, $p = 0.0021$) (Fig. 4). These positive relationships are consistent with the presence of lineages where repeated shifts in genome size are associated with

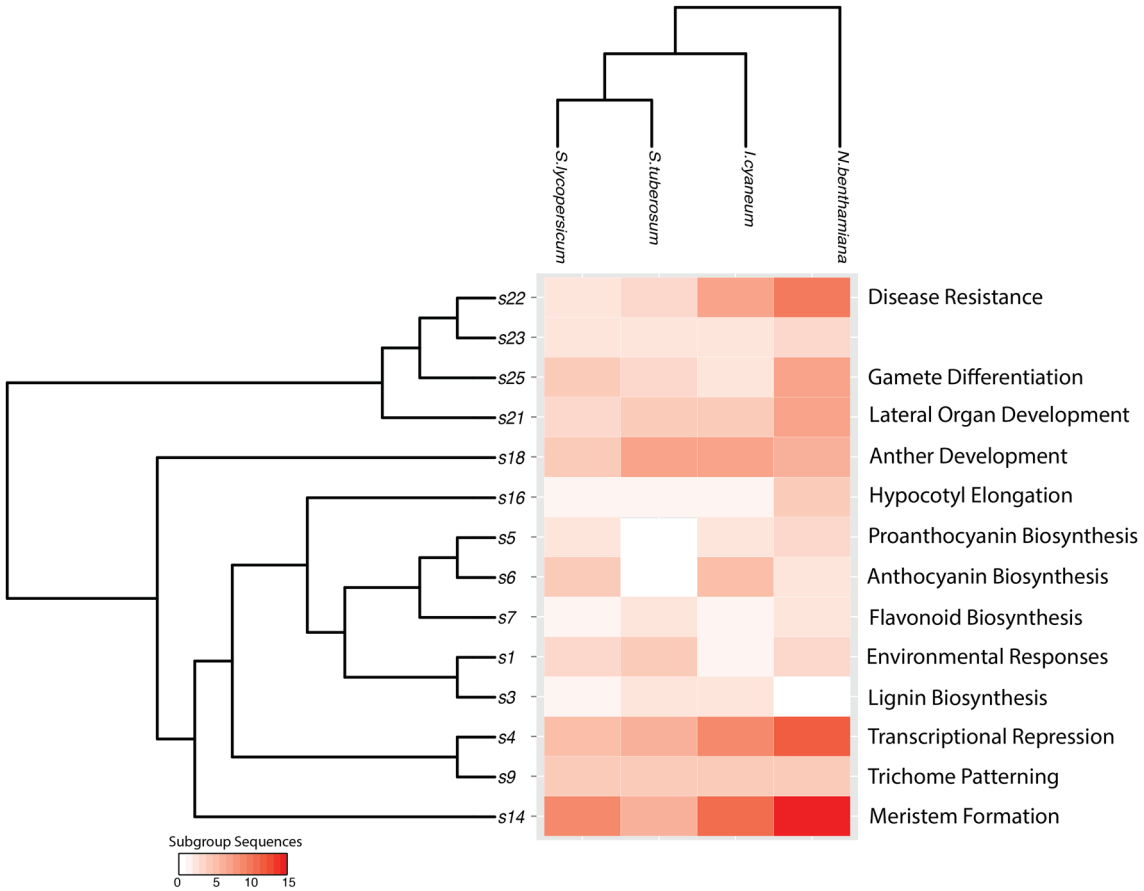

Fig. 2 R2R3 MYB subgroup content of four Solanaceae species. The phylogenetic relationships among the MYB subgroups (tree on the *left*) follow Fig. 1. The relationships among the four taxa are shown at the *top*. The subgroup names follow the naming scheme illustrated in Fig. 2 of Stracke et al. (2001). Some subgroups that could not definitively be delimited (e.g., S10/11/24) are not included since specific clade members could not be delimited and counted

concordant changes in the number of R2R3 MYBs (e.g., monocots: *Oryza*, *Sorghum*, and *Zea,* Fig. 3). The tighter relationship between total gene number and R2R3 MYB family size may reflect that some variation in genome size (e.g., the large genome of *Z. mays*) is due to fluctuations in repetitive content. Examination of a scatterplot of the data (Fig. 4) also reveals a number of outliers (e.g., *Populus trichocarpa*, *Glycine max, Physcomitrella patens*), which contain far more or far fewer R2R3 MYBs than expected, given their genome size and their total number of genes. These provide interesting candidates for follow-up studies examining the possible functional changes associated with expansions and contractions of R2R3 MYB diversity.

Discussion

Functional R2R3 Diversity in Solanaceae

Phylogenetic analyses allowed us to pinpoint taxon-specific gains and losses of MYB copies in Solanaceae, which may be tied to the changes in functional diversity. We found two *A. thaliana* clades with no Solanaceae orthologs: subgroups S12 and S15. Of these two subgroups, subgroup S15 is of interest because it contains three important transcriptional activators of root hair formation and trichome development: WER, GL1, and MYB23 (Lee and Schiefelbein 1999; Kirik et al. 2005). Functional analyses using heterologous expression suggest that S15 MYBs from *A. thaliana* fail to elicit the same trichome responses outside of the Brassicaceae (Payne et al. 1999). Thus, our findings support the hypothesis that trichome formation is analogous between Brassicaceae and Solanaceae and involves different transcriptional elements in these taxa (Serna and Martin 2006).

By comparing the evolutionary relationships of Solanaceae MYBs to sequences from *A. thaliana*, we can also identify clades that lack *A. thaliana* relatives and therefore serve currently unknown functional roles. For instance, we find a well-supported clade of 14 Solanaceae sequences that is closely related to the S5, S6, and S7 subgroups, but contains no *A. thaliana* members (S1 Fig.). The S5, S6, and

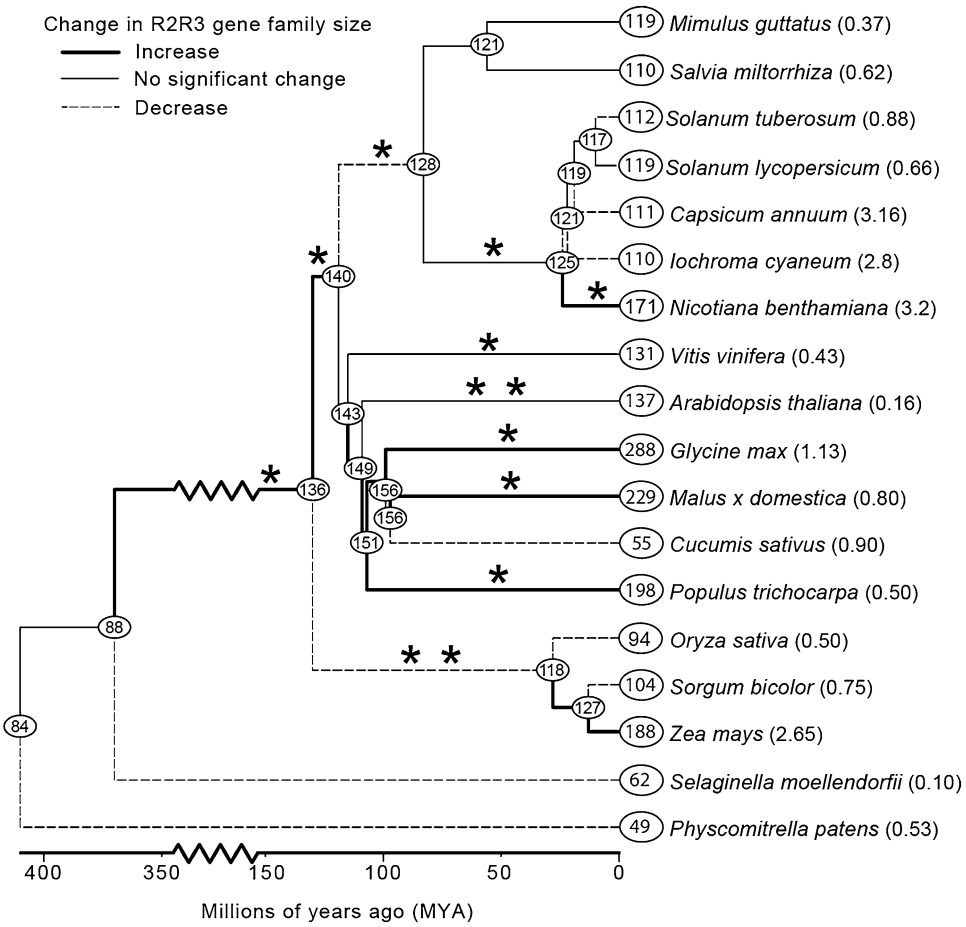

Fig. 3 Evolution of R2R3 gene subfamily size across land plants. Number of R2R3 MYB genes in each sampled taxon is shown at the tips; genome sizes appear in parentheses after species names. Branch lengths are in units of time, but compressed (*jagged line*) between 150 and 350 MYA. Inferred ancestral gene subfamily sizes are indicated at nodes. Significant increases or decreases in R2R3 subfamily size according to the CAFE analysis are indicated with *bold* and *dashed lines*, respectively. Asterisks indicate polyploidy events from, except for *N. benthamiana* (Bombarely et al. 2012) and *M.x domestica* (Jung et al. 2012)

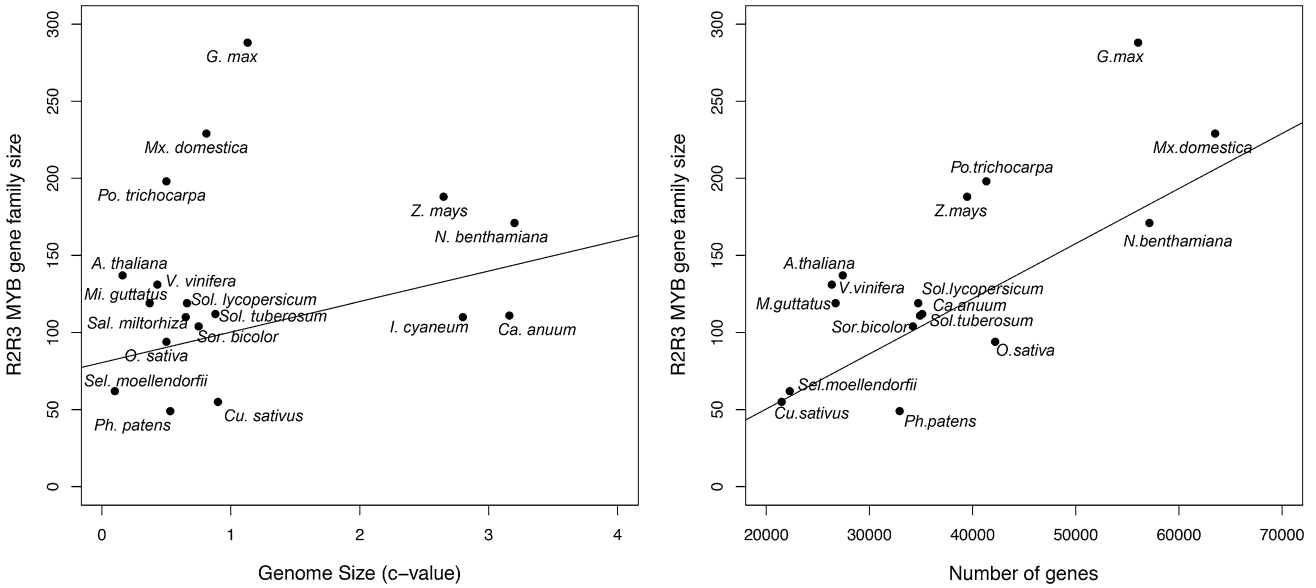

Fig. 4 Relationship between the number of R2R3 MYBs and aspects of genome size. Gene family size is regressed against genome size (c-values) on the *left* and against the number of genes in the genome on the *right*. Each point is a species in the analysis. Genus names are abbreviated from Fig. 3 (either first letter if unique or multiple letters if not). Trendlines are from the PGLS regression analysis (see text)

S7 genes regulate the synthesis of flavonoids, a group of specialized metabolites including anthocyanin pigments, tannin precursors, and other stress-response compounds (Winkel-Shirley 2002). Although additional functional studies would be required, the uncategorized clade of Solanaceae sequences may serve a similar role in flavonoid regulation. Phylogenetic affinities of this and other Solanaceae R2R3 MYBs in clades without *A. thaliana* members (Fig. 1, S1 Fig.) can serve as a starting point for examining patterns of expression and mutant phenotypes.

We also document a relatively stable gene subfamily size of between 110 and 119 members within diploid Solanaceae species that are over 20 million years diverged (Särkinen et al. 2013; Ng and Smith 2016). Despite the relative stability in the presence or absence of subgroup members across species, our survey found evidence of fluctuation in content across R2R3 subgroups (Fig. 2) that may be related to patterns of phenotypic diversification within lineages. *S. tuberosum* and *N. benthamiana* are the only Solanaceae lineage that have no sequences for entire subgroups (in S5, S6 for *S.tuberosum* and S3 for *N. benthamiana*; Fig. 2). Not only is subgroup S6 lacking potato sequences in our analyses, it is also one of the three subgroups where the polyploid *N. benthamiana* does not have the most sequences out of the four surveyed Solanaceae species. Instead, *Iochroma* has the largest number of subgroup 6 sequences. This subgroup shares a common c-terminal motif that is characteristic of genes related to anthocyanin pigment biosynthesis (Paz-Ares et al. 1987; Quattrocchio et al. 1998; Nakatsuka et al. 2008; Jung et al. 2009; Shang et al. 2011; Takahashi et al. 2013). In this context, it is notable that *Iochroma* also produces a wide diversity of pigment types and patterns (Smith and Baum 2006; Smith and Rausher 2011). This presents the interesting possibility that the radiation of flower colors within the *Iochroma* may be related to a diversification of transcriptional regulators associated with anthocyanin production. Similar lineage-specific effects may explain the expansion and contraction of other subgroups within different Solanaceae species, despite overall maintenance of gene subfamily size.

Expansions and Contractions of the R2R3 MYB Subfamily in Angiosperms

Our results indicate that the size of the R2R3 MYB subfamily varies dramatically across terrestrial plants, which is consistent with previous studies (Kranz et al. 2000; Feller et al. 2011). On mapping R2R3 gene subfamily size onto the phylogeny, we observed many large expansions and contractions (Fig. 3). Based on the CAFE analysis, we inferred that 66 % of the branches show significant changes in gene subfamily size relative to a constant birth–death model (Fig. 3). The single strongest shift in diversification rate of the R2R3 subfamily, a nearly fourfold increase, occurred along the eudicot stem lineage. This diversification shift may relate to the specialization of R2R3 MYBs for novel functions in angiosperms (Stracke et al. 2001), similar to other gene families associated with angiosperm-specific organ development like the APETALA2-like (Kim et al. 2006) and SEPALATA MADS-Box subfamily (Zahn et al. 2005). For example, different MYB copies are specialized for different cell fates, such as trichomes and petal cells (Ramsay and Glover 2005), and for different locations, such as petal veins or corolla lobes (Schwinn et al. 2006; Albert et al. 2011). Characterizing R2R3 diversity in other disparate lineages, such as the magnoliids and ranunculids, would allow us to pinpoint more exactly the timing of subfamily expansion and its potential importance for morphological innovation.

Since the results of the CAFE analyses largely rejected the ability of a stochastic birth–death process to explain R2R3 MYB family size fluctuations, we investigated the possibility that this variation is related to overall changes in genome size, e.g., due to segmental duplications or changes in ploidy (Huynen and Van Nimwegen 1998). Our analyses showed significant relationship between genome size (c-values) and MYB gene subfamily size using PGLS with a BM model. This result appears to be driven by clades like the monocots, where increases in genome size are consistently associated with increases in the number of R2R3 MYBs (Figs. 3, 4). We also analyzed the relationship between number of genes and the MYB gene subfamily size. In this analysis, we see a strong positive trend between gene number and MYB family size and a tighter correlation than that observed for overall genome size (Fig. 4). This contrast may be attributed to factors such as repetitive element proliferation (Hawkins et al. 2006; Vitte and Bennetzen 2006), which can increase genome size without increasing gene number.

This study contributes to a growing body of literature that examines how genome content and size vary across both deep and shallow evolutionary timescales. Previous studies suggest that the dynamics of gene family evolution may vary depending on the type of gene (e.g., cell cycle vs. metabolic pathway genes, (Molina and van Nimwegen 2009)), and the size of other families in the genome (Huynen and Van Nimwegen 1998). For example, in prokaryotes, families of transcription factors (TFs) have been shown to evolve following power laws, where the number of TFs increases exponentially with the number of genes in the genome (Molina and van Nimwegen 2009). Comparative analyses at more recent timescales in plants, however, show significant stochasticity in transcription factor family evolution following changes in genome size

(Schranz and Mitchell-Olds 2006; Barker et al. 2008). These differences may relate to the selective pressures acting on the genes in different taxa (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Chapman et al. 2006). While the number of plant taxa analyzed here ($n = 18$) is too small to test laws of genomic evolution (Koonin 2011), the increasing number of published plant genomes and transcriptomes (Goodstein et al. 2012; Michael and Jackson 2013; Wickett et al. 2014) opens the possibility for extending such broad-scale genomic analyses to plants in the near future.

Conclusions

Our comparative analysis of the R2R3 MYBs across Solanaceae and other plants underscores the dynamic history of this gene subfamily, both in terms of functional diversity and size. Within Solanaceae, we documented significant fluctuations in both the total R2R3 copy number and subgroup composition across the family. We observed similar repeated expansions and contractions across land plants, but with a general trend of higher numbers in angiosperms. These shifts in subfamily size mirror, to some extent, the changes in overall genome size, although the notable exceptions to this pattern (e.g., *Cucumis*) suggest that other factors are also at play.

The continuing emergence of new plant genomes will offer greater opportunities to trace the evolutionary history of gene families, like MYBs, and to target taxa and gene copies for further investigation (Brockington et al. 2013). As demonstrated above, statistical comparative methods allow us to estimate the number of gains and losses, the rates of gene subfamily diversification, and the extent to which these changes are due to fluctuations in overall genome size. We expect that future studies surveying a wider range of taxa in a phylogenetic framework will uncover additional cases of lineage-specific gene subfamily expansion and contraction.

One major limitation to a phyloinformatic approach to tracing gene subfamily evolution is the lack of functional information. The well-studied roles of R2R3 MYBs in *Arabidopsis* and other model systems provide starting points for determining function, and thus far, such predictions have held for some subgroups, such as the anthocyanin-regulating R2R3s (e.g., Borovsky et al. 2004; Yamagishi et al. 2010). In Solanaceae, forward genetic screens of MYB mutants can be used to confirm function, as has been done with other classes of genes (Quattrocchio et al. 1999; Borovsky et al. 2004; Kaur et al. 2010; Pattanaik et al. 2010; Busch et al. 2011; Hermann et al. 2013). Reverse genetic screens will also be increasingly useful as transformations, and silencing protocols are readily available for many Solanaceae crops and model systems (tomato, tobacco, and petunia) (Huang et al. 2015). The advances in functional annotations and the growing availability of publicly available gene ontology databases should add a new dimension to bioinformatics studies such as this by allowing for more precise testing of hypotheses regarding the evolutionary and functional diversification of R2R3 MYB transcription factors.

Acknowledgments The authors would like to thank E. Braun for advice regarding alignments and using HMMer and J. Storz for providing useful references. We would also like to thank two anonymous reviewers for providing helpful feedback.

References

Abe H, Yamaguchi-Shinozaki K, Urao T et al (1997) Role of arabidopsis MYC and MYB homologs in drought-and abscisic acid-regulated gene expression. The Plant Cell 9:1859–1868

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141

Albert VA, Chang T-H (2014) Evolution of a hot genome. Proc Natl Acad Sci 111:5069–5070

Albert NW, Lewis DH, Zhang H et al (2011) Members of an R2R3-MYB transcription factor family in Petunia are developmentally and environmentally regulated to control complex floral and vegetative pigmentation patterning. Plant J 65:771–784

Barker MS, Kane NC, Matvienko M et al (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 25:2445–2455

Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296–1303

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691

Bombarely A, Menda N, Tecle IY et al (2011) The sol genomics network (solgenomics. net): growing tomatoes using Perl. Nucleic Acids Res 39:D1149–D1155

Bombarely A, Rosli HG, Vrebalov J et al (2012) A draft genome sequence of nicotiana benthamiana to enhance molecular plant-microbe biology research. Mol Plant Microbe Interact 25:1523–1530

Borovsky Y, Oren-Shamir M, Ovadia R et al (2004) The A locus that controls anthocyanin accumulation in pepper encodes a MYB transcription factor homologous to Anthocyanin2 of Petunia. Theor Appl Genet 109:23–29

Brockington SF, Alvarez-Fernandez R, Landis JB et al (2013) Evolutionary analysis of the MIXTA gene family highlights potential targets for the study of cellular differentiation. Mol Biol Evol 30:526–540

Busch BL, Schmitz G, Rossmann S et al (2011) Shoot branching and leaf dissection in tomato are regulated by homologous gene modules. The Plant Cell 23:3595–3609

Butler MA, King AA (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683–695

Cao Z-H, Zhang S-Z, Wang R-K et al (2013) Genome wide analysis of the apple MYB transcription factor family allows the identification of MdoMYB121 gene confering abiotic stress tolerance in plants. PLoS One 8:e69955

Chapman BA, Bowers JE, Feltus FA, Paterson AH (2006) Buffering of crucial functions by paleologous duplicated genes may

contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci 103:2730–2735
Clegg MT, Cummings MP, Durbin ML (1997) The evolution of plant nuclear genes. Proc Natl Acad Sci 94:7791–7798
Coate JE, Doyle JJ (2010) Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol 2:534–546
De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271
Du H, Feng B-R, Yang S-S et al (2012) The R2R3-MYB transcription factor gene family in maize. PLoS One 7:e37463
Dubos C, Stracke R, Grotewold E et al (2010) MYB transcription factors in Arabidopsis. Trends Plant Sci 15:573–581
Feller A, Machemer K, Braun EL, Grotewold E (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66:94–116
Frech C, Chen N (2010) Genome-wide comparative gene family classification. PLoS One 5:e13409
Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186
Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
Grafen A (1989) The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci 326:119–157
Grotewold E, Drummond BJ, Bowen B, Peterson T (1994) The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 76:543–553
Guo Y (2013) Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. Plant J 73:941–951
Hahn MW, De Bie T, Stajich JE et al (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15:1153–1160
Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417
Hawkins JS, Kim H, Nason JD et al (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261
Hermann K, Klahre U, Moser M et al (2013) Tight genetic linkage of prezygotic barrier loci creates a multifunctional speciation island in Petunia. Curr Biol 23:873–877
Huang X, Yue Y, Sun J et al (2015) Characterization of a fertility-related SANT/MYB gene (PhRL) from petunia. Sci Hortic 183:152–159
Hunziker AT (2001) The genera of Solanaceae. ARG Gantner Verlag, KG, Ruggell
Huynen MA, Van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15:583–589
Jung CS, Griffiths HM, De Jong DM et al (2009) The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. Theor Appl Genet 120:45–57
Jung S, Cestaro A, Troggio M et al (2012) Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies. BMC Genom 13:129
Katoh K, Misawa K, Kuma K-I, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066
Kaur H, Heinzel N, Schöttner M et al (2010) R2R3-NaMYB8 regulates the accumulation of phenylpropanoid-polyamine conjugates, which are essential for local and systemic defense against insect herbivores in Nicotiana attenuata. Plant Physiol 152:1731–1747
Kim S, Soltis PS, Wall K, Soltis DE (2006) Phylogeny and domain evolution in the APETALA2-like gene family. Mol Biol Evol 23:107–120
Kim S, Park M, Yeom S-I et al (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet 46:270–278
Kirik V, Lee MM, Wester K et al (2005) Functional diversification of MYB23 and GL1 genes in trichome morphogenesis and initiation. Development 132:1477–1485
Knapp S, Chase MW, Clarkson JJ (2004) Nomenclatural changes and a new sectional classification in Nicotiana (Solanaceae). Taxon 53:73–82
Koonin EV (2011) Are there laws of genome evolution. PLoS Comput Biol 7:e1002173
Kranz HD, Denekamp M, Greco R et al (1998) Towards functional characterisation of the members of the R2R3-MYB gene family from Arabidopsis thaliana. Plant J 16:263–276
Kranz H, Scholz K, Weisshaar B (2000) c-MYB oncogene-like genes encoding three MYB repeats occur in all major plant lineages. Plant J 21:231–235
Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
Lee MM, Schiefelbein J (1999) WEREWOLF, a MYB-related protein in arabidopsis, is a position-dependent regulator of epidermal cell patterning. Cell 99:473–483
Leitch IJ, Hanson L, Lim KY et al (2008) The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). Ann Bot 101:805–814
Li C, Lu S (2014) Genome-wide characterization and comparative analysis of R2R3-MYB transcription factors shows the complexity of MYB-associated regulatory networks in Salvia miltiorrhiza. BMC Genom 15:277
Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714
Li Q, Zhang C, Li J et al (2012) Genome-wide identification and characterization of R2R3MYB family in Cucumis sativus. PLoS One 7:e47576
Lipsick JS (1996) One billion years of Myb. Oncogene 13:223–235
Matus JT, Aquea F, Arce-Johnson P (2008) Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. BMC Plant Biol 8:83
Mehrtens F, Kranz H, Bednarek P, Weisshaar B (2005) The Arabidopsis transcription factor MYB12 is a flavonol-specific regulator of phenylpropanoid biosynthesis. Plant Physiol 138:1083–1096
Michael TP, Jackson S (2013) The first 50 plant genomes. The Plant Genome. doi:10.3835/plantgenome2013.03.0001in
Molina N, van Nimwegen E (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet 25:243–247
Nakatsuka T, Haruta KS, Pitaksutheepong C et al (2008) Identification and characterization of R2R3-MYB and bHLH transcription factors regulating anthocyanin biosynthesis in gentian flowers. Plant Cell Physiol 49:1818–1829
Ng J, Smith SD (2016) Widespread flower color convergence in Solanaceae via alternate biochemical pathways. New Phytol 209:407–417
Ogata K, Kanei-Ishii C, Sasaki M et al (1996) The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and trans-activation. Nat Struct Mol Biol 3:178–187

Olmstead RG, Bohs L, Migid HA et al (2008) A molecular phylogeny of the Solanaceae. Taxon 57:1159–1181

Oppenheimer DG, Herman PL, Sivakumaran S et al (1991) A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. Cell 67:483–493

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci 101:9903–9908

Pattanaik S, Kong Q, Zaitlin D et al (2010) Isolation and functional characterization of a floral tissue-specific R2R3 MYB regulator from tobacco. Planta 231:1061–1076

Payne T, Clement J, Arnold D, Lloyd A (1999) Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in Nicotiana tabacum. Development 126:671–682

Paz-Ares J, Ghosal D, Wienand U et al (1987) The regulatory c1 locus of Zea mays encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. EMBO J 6:3553

Pinheiro J, Bates D, DebRoy S, Sarkar D (2011) R Development Core Team. 2010. nlme: linear and nonlinear mixed effects models. R package version 3.1-97

Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Quattrocchio F, Wing JF, Va K et al (1998) Analysis of bHLH and MYB domain proteins: species-specific regulatory differences are caused by divergent evolution of target anthocyanin genes. Plant J 13:475–488

Quattrocchio F, Wing J, van der Woude K et al (1999) Molecular analysis of the anthocyanin2 gene of petunia and its role in the evolution of flower color. The Plant Cell 11:1433–1444

Ramsay NA, Glover BJ (2005) MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. Trends Plant Sci 10:63–70

Romero I, Fuertes A, Benito MJ et al (1998) More than 80R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. Plant J 14:273–284

Särkinen T, Bohs L, Olmstead RG, Knapp S (2013) A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol Biol 13:214

Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593

Schranz ME, Mitchell-Olds T (2006) Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell 18:1152–1165

Schwinn K, Venail J, Shang Y et al (2006) A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus Antirrhinum. The Plant Cell Online 18:831–851

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet 20:461–464

Serna L, Martin C (2006) Trichomes: different regulatory networks lead to convergent structures. Trends Plant Sci 11:274–280

Shang Y, Venail J, Mackay S et al (2011) The molecular basis for venation patterning of pigmentation and its effect on pollinator attraction in flowers of Antirrhinum. New Phytol 189:602–615

Sierro N, Battey JND, Ouadi S et al (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5:3833

Smith SD, Baum DA (2006) Phylogenetics of the florally diverse Andean clade Iochrominae (Solanaceae). Am J Bot 93:1140–1153

Smith SD, Rausher MD (2011) Gene loss and parallel evolution contribute to species difference in flower color. Mol Biol Evol 28:2799–2810

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467

Stracke R, Werber M, Weisshaar B (2001) The R2R3-MYB gene family in Arabidopsis thaliana. Curr Opin Plant Biol 4:447–456

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

Takahashi R, Yamagishi N, Yoshikawa N (2013) A MYB transcription factor controls flower color in soybean. J Hered 104:149–153

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

Urao T, Noji M, Yamaguchi-Shinozaki K, Shinozaki K (1996) A transcriptional activation domain of ATMYB2, a drought-inducible Arabidopsis Myb-related protein. Plant J 10:1145–1148

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Wada T, Tachibana T, Shimura Y, Okada K (1997) Epidermal cell differentiation in Arabidopsis determined by a Myb homolog, CPC. Science 277:1113–1116

Wang X, Bennetzen JL (2015) Current status and prospects for the study of Nicotiana genomics, genetics, and nicotine biosynthesis genes. Mol Genet Genomics 290:1–11

Wickett NJ, Mirarab S, Nguyen N et al (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci 111:E4859–E4868

Winkel-Shirley B (2002) Biosynthesis of flavonoids and effects of stress. Curr Opin Plant Biol 5:218–223

Wu F, Tanksley SD (2010) Chromosomal evolution in the plant family Solanaceae. BMC Genom 11:182

Yamagishi M, Shimoyamada Y, Nakatsuka T, Masuda K (2010) Two R2R3-MYB genes, homologs of petunia AN2, regulate anthocyanin biosyntheses in flower tepals, tepal spots and leaves of Asiatic hybrid lily. Plant Cell Physiol 51:463–474

Yang X, Kalluri UC, Jawdy S et al (2008) The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. Plant Physiol 148:1189–1200

Zahn LM, Kong H, Leebens-Mack JH et al (2005) The evolution of the SEPALLATA subfamily of MADS-Box genes a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics 169:2209–2223

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298

Zhao K, Bartley LE (2014) Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, poplar, rice, maize, and switchgrass. BMC Plant Biol 14:135

Zhao P, Li Q, Li J et al (2014) Genome-wide identification and characterization of R2R3MYB family in Solanum lycopersicum. Mol Genet Genomics 289:1183–1207