Evolutionary diversification of the \textit{YABBY} gene family

Conny Bartholmes$^1$, Oriane Hidalgo$^1$, Stefan Gleissberg$^{1\S}$

$^1$Department of Environmental and Plant Biology, Ohio University, Porter Hall 500, Athens, Ohio 45701, USA

$^\S$Corresponding author

Email addresses:

CB: cb271507@ohio.edu

OH: hidalgo@ohio.edu

SG: gleissbe@ohio.edu
Abstract

Background

*YABBY* genes are seed plant specific transcriptional regulators that are involved in diverse aspects of leaf, shoot, and flower development. A series of duplications near the base of the angiosperms gave rise to five gene copies found throughout flowering plants. In *Arabidopsis* and other species, members of three gene groups, namely *FILAMENTOUS FLOWER*, *YABBY2*, and *YABBY5*, largely share leaf-specific expression, and have been termed "vegetative *YABBYs". In contrast, expression of two groups, *CRABS CLAW* and *INNER NO OUTER*, appears to be more restricted to floral parts. How the five copies evolved and how their expression and function diversified remains largely unknown, precluding the reconstruction of the natural history of this gene family. Here we report sequences from a basal eudicot and establish an improved phylogenetic framework of the *YABBY* gene family. We use this framework to evaluate sequence and expression evolution.

Results

A multilayered Bayesian analysis covering seed plants allows us to provide an improved resolution of *YABBY* gene clades. We establish that vegetative *YABBYs* do not form a monophyletic group, and that *CRABS CLAW* and *FILAMENTOUS FLOWER* arose from a common ancestor gene. *INNER NO OUTER* genes are sister to that ancestral gene. We identify several conserved motifs outside of known amino acid domains that further define all five angiosperm *YABBY* gene clades. In addition, we infer the evolution of gene expression and provide evidence for release of purifying constraint in certain branches of the gene family tree. Finally, we report the cloning and expression patterns of five vegetative *YABBY* genes from *Eschscholzia californica*, a basal eudicot.
Conclusions
Our study allows novel insights into the evolutionary history of this functionally diverse family of transcriptional regulators. The history of YABBY genes in flowering plants is characterized by pronounced changes of coding sequences and of gene expression in some clades, while other clades have retained more ancestral characteristics. Expression and functional studies in gymnosperms, basal angiosperms, and basal eudicots are needed to link sequence evolution to an understanding of the role of YABBY genes in the evolution of shoot development.
Background

The YABBY gene family is a small class of transcription factors with important roles in vegetative and reproductive shoot development in flowering plants. Unlike other transcription factor families, YABBY genes appear to be absent in non-seed plant lineages of embryophytes [1]. YABBYs are composed of an N-terminal zinc finger domain and a C-terminal helix-loop-helix motif similar to a high mobility group (HMG) box. The latter motif was termed the YABBY domain [2]. Angiosperms possess five distinct YABBY groups with characteristic expression patterns in leaves and/or floral organs. In Arabidopsis, three of these groups, FILAMENTOUS FLOWER (FIL)-like, YABBY2 (YAB2) and YAB5 are expressed in both leaf and floral organ primordia, whereas CRABS CLAW (CRC), and INNER NO OUTER (INO) are not expressed in vegetative tissues and are restricted to developing carpels and ovules, respectively [2-4]. Although YABBY genes have been found in gymnosperms, the relationships between gymnosperm and angiosperm genes are poorly understood. In addition, the diversification of the five angiosperm gene clades and their functions have remained unclear [5-8].

Loss-of-function studies of vegetative YABBY genes in core eudicots suggested that YABBY genes were involved in the evolution of leaf-specific pathways in the seed plants, where they were recruited to initiate lamina growth, cooperating with evolutionarily more ancestral genes [9]. In Arabidopsis and other core eudicots, YABBY genes integrate several leaf polarity pathways, contribute to leaf margin establishment and lamina growth, and promote the determinate fate of leaves by activating maturation processes and repressing shoot apical meristem genes [3,9-14]. In addition, some YABBYs also regulate shoot apical meristem development [15].
The importance of \textit{YABBY} genes in plant development is in contrast to the rather poor understanding of the phylogenetic history of this gene family. Phylogenetic studies have been unable to establish proper limits, taxonomical composition and relationships for most of the \textit{YABBY} gene groups. Limited sequence conservation outside the zinc finger and YABBY domains restrict the number of usable characters, which is reflected in weak phylogenetic frameworks obtained so far. To date, the only gene clades identified with significant statistical support are \textit{CRC} and \textit{INO} \cite{6,8}. \textit{YABBY} gene groups encompass sequences from all major angiosperm lineages \cite{6}, making it difficult to trace the history of duplications that underlies gene diversification, and to infer the putative ancestral functions of the genes as well as to detect events of sub- and/or neofunctionalization \cite{16}. However, a well-resolved phylogeny is crucial for understanding the evolutionary diversification of the \textit{YABBY} gene family \cite{16}.

We used a multi-layered approach to resolve the phylogeny of the \textit{YABBY} gene family, using sequences from all major angiosperm lineages as well as from gymnosperms, including several fully sequenced taxa. We partitioned the data set to analyze individual gene domains and codon positions separately and concatenated in an attempt to optimize the phylogenetic signal. Because \textit{YABBY}-like sequences from basal eudicots were underrepresented in public databases, we isolated additional sequences from the basal eudicot \textit{Eschscholzia californica} Cham. (California poppy; Papaveraceae). We address the origin of the five angiosperm \textit{YABBY} gene groups, the evolution of expression patterns, provide evidence for changes in selective pressure along branches, and describe additional gene clade specific conserved motifs.
Results

Five *YABBY*-like genes newly identified in *Eschscholzia californica*, including two pairs of paralogues

To increase the taxon sampling in basal eudicots, we newly isolated five *YABBY* genes from *Eschscholzia californica* (Papaveraceae), a member of the Ranunculales, the earliest branching eudicot clade [17]. NCBI BLAST searches of the sequences confirmed the presence of a zinc finger domain near the 5' end and a *YABBY* domain located further downstream, that together define *YABBY* genes. The consensus sequences were named *Eschscholzia californica YABBY1* (*EcYAB1*) to *EcYAB5* [GenBank HQ116795 to HQ116799]. Their sizes range from 498 bp (*EcYAB5*) to 684 bp (*EcYAB1*). Amino acid sequence comparisons revealed that two pairs of sequences shared a particularly high degree of sequence identity outside the two conserved domains (data not shown). The sequences of *EcYAB1* and *EcYAB2* (genetic divergence of 8.77%), as well as *EcYAB3* and *EcYAB4* (genetic divergence of 10.53%) were more similar to each other than to the other *Eschscholzia YABBY* sequences, where genetic divergence ranged between 25.26 and 36.83% (Additional file 1). This similarity suggested the presence of two sets of paralogues in *Eschscholzia*.

**Gymnosperm and angiosperm YABBY genes diversified separately**

To better resolve the phylogeny of the *YABBY* gene family, we established a phylogenetic framework resulting from 32 Bayesian analyses (Figure 1, Table 1). In these analyses, we compared how branch support was affected by taxon sample size, exclusion/inclusion of taxonomic groups, single and combined analyses of the zinc finger and *YABBY* domains and of codon positions with different data partitions, or the use of amino acid data (Table 1). Our analyses show that all available
gymnosperm YABBY genes are isolated from their angiosperm counterparts, thereby making them a valid outgroup for rooting the angiosperm gene trees. Topologies obtained in a genetic distance-based tree (Additional file 2) as well as several Bayesian inferences (analyses 1, 2, 7, 21, 28; Table 1) already indicated a segregation of gymnosperm and angiosperm YABBYs. However, significant support was only obtained in a Bayesian analysis in which we used the large taxon sampling and a dataset that was partitioned by domain and by codon position (PP = 0.96, analysis 3, Table 1).

The CRC, INO and YAB2 groups are each monophyletic with significant statistical support
All five angiosperm YABBY gene groups are discriminated in the phylogenetic framework (Figure 1, Table 1). Most of the 32 Bayesian inferences provide significant support for the CRC group, except analyses using either the second codon position of zinc finger domain alone or the third codon position of the YABBY domain (Table 1). The same was found for the INO group, whose monophyly is significantly supported in all analyses except those where individual codon positions were used (Table 1). The CRC and INO gene groups are also the only ones whose monophyly is supported in the neighbor-joining phylogram (BS = 100% for CRC, BS = 76% for INO; Additional file 2), the parsimony tree (BS = 100% for CRC, BS = 84% for INO; Additional file 3), and in previously published analyses [7,8]. In addition, we obtained significant support for the YAB2 gene group in a more restricted set of analyses using the large and medium taxon samplings (Table 1). Although not directly, the FIL and YAB5 groups are also identified by our phylogenetic framework, as described below.
Table 1 - Posterior probability support values of major YABBY gene family clades in 32 Bayesian analyses

1.Taxon samplings are indicated in Additional file 7. 2. A: angiosperms, CE: core eudicots, E: eudicots, G: gymnosperms, M: monocots. 3. Z: zinc finger domain, Y: YABBY domain, small numbers following domain symbol denote first, second, and third codon positions, aa: amino acids. Dash (-): clade not observed in the corresponding tree topology. NA: not applicable. Grey boxes indicate a significant support (PP ≥ 0.95).

<table>
<thead>
<tr>
<th></th>
<th>Sampling qualitative</th>
<th>Dataset (partition)</th>
<th>Analysis identification</th>
<th>INO</th>
<th>FIL</th>
<th>CRC</th>
<th>(FIL, CRC) in rooted analyses</th>
<th>(FIL, CRC) INO in rooted analyses</th>
<th>YAB2</th>
<th>YAB5 (without Cabomba and Maipurea)</th>
<th>(YAB2, YAB5) INO, FIL, CRC</th>
<th>(YAB2, YAB5) in unrooted analyses</th>
<th>(gymnosperms, angiosperms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>G + A</td>
<td>Z₁(2)Y₁(23)</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.85</td>
<td>0.94</td>
<td>0.94</td>
<td>1.00</td>
<td>0.78</td>
<td>NA</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₂(1)Y₁(2)</td>
<td>2</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.85</td>
<td>0.95</td>
<td>0.96</td>
<td>1.00</td>
<td>0.78</td>
<td>NA</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₃(1)Y₁(23)</td>
<td>3</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.70</td>
<td>0.95</td>
<td>0.87</td>
<td>1.00</td>
<td>0.74</td>
<td>NA</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₄(1)Y₁(23)</td>
<td>4</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₅(1)Y₁(23)</td>
<td>5</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.92</td>
<td>-</td>
<td>0.64</td>
<td>1.00</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₁(2)</td>
<td>6</td>
<td>0.99</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₂(1)(2)</td>
<td>7</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.63</td>
<td>-</td>
<td>1.00</td>
<td>0.80</td>
<td>NA</td>
<td>0.65</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₃(1)</td>
<td>8</td>
<td>0.93</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.89</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₄(1)</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>0.61</td>
<td>-</td>
<td>0.94</td>
<td>-</td>
<td>0.74</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₅(1)</td>
<td>10</td>
<td>0.68</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y₁(23)</td>
<td>11</td>
<td>0.99</td>
<td>-</td>
<td>1.00</td>
<td>0.93</td>
<td>-</td>
<td>-</td>
<td>0.64</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y₁(2)(23)</td>
<td>12</td>
<td>0.99</td>
<td>0.69</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y₂(1)</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>0.68</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y₃(1)</td>
<td>14</td>
<td>0.68</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Z₁(23)Y₁)(23)</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>0.73</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>Z₁(23)Y₁(23)</td>
<td>16</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₂(1)(23)</td>
<td>17</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.86</td>
<td>NA</td>
<td>0.92</td>
<td>1.00</td>
<td>NA</td>
<td>1.00</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₃(2)</td>
<td>18</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>NA</td>
<td>1.00</td>
<td>1.00</td>
<td>NA</td>
<td>0.99</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₄(2)</td>
<td>19</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.74</td>
<td>NA</td>
<td>1.00</td>
<td>1.00</td>
<td>NA</td>
<td>0.97</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₅(2)</td>
<td>20</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.94</td>
<td>NA</td>
<td>1.00</td>
<td>1.00</td>
<td>NA</td>
<td>0.83</td>
<td>NA</td>
</tr>
<tr>
<td>Medium</td>
<td>G + A</td>
<td>Z₁(12)Y₁(23)</td>
<td>21</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.92</td>
<td>0.71</td>
<td>0.97</td>
<td>1.00</td>
<td>0.73</td>
<td>NA</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₁(12)(3)Y₁(1)(23)</td>
<td>22</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.74</td>
<td>0.96</td>
<td>0.82</td>
<td>1.00</td>
<td>0.74</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₂(1)(23)</td>
<td>23</td>
<td>0.99</td>
<td>-</td>
<td>1.00</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₃(12)Y₁(23)</td>
<td>24</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Z₁(12)Y₁)(23)</td>
<td>25</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.62</td>
<td>NA</td>
<td>1.00</td>
<td>1.00</td>
<td>NA</td>
<td>0.91</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₁(2)</td>
<td>26</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>0.98</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td>Small</td>
<td>G + A</td>
<td>Z₁(123)Y₁(23)</td>
<td>27</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.88</td>
<td>0.86</td>
<td>0.82</td>
<td>1.00</td>
<td>0.51</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₁(123)(3)Y₁(1)(23)</td>
<td>28</td>
<td>1.00</td>
<td>0.70</td>
<td>1.00</td>
<td>0.71</td>
<td>-</td>
<td>0.92</td>
<td>1.00</td>
<td>0.58</td>
<td>NA</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₂(1)</td>
<td>29</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.90</td>
<td>-</td>
<td>0.88</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₃(123)Y₁(23)</td>
<td>30</td>
<td>1.00</td>
<td>0.51</td>
<td>1.00</td>
<td>0.71</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Z₁(123)Y₁)(23)</td>
<td>31</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.82</td>
<td>NA</td>
<td>0.84</td>
<td>1.00</td>
<td>NA</td>
<td>0.99</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Z₁(23)Y₁(23)</td>
<td>32</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>0.98</td>
<td>NA</td>
<td>0.87</td>
<td>0.99</td>
<td>NA</td>
<td>-</td>
<td>NA</td>
</tr>
</tbody>
</table>
**FIL-like YABBY genes are indirectly identified**
The *FIL* group remains largely unresolved in all analyses (Figure 1, Table 1).

However, *FIL*-like genes are well isolated from their closest paralogue group, *CRC*, and the earlier-diverged *INO* group, by a series of strongly supported branches. The *CRC* and *FIL* groups both contain representatives from basal angiosperms, monocots, basal eudicots, and core eudicots, indicating that *CRC* genes did not derive from within *FIL* genes, but rather represent a sister lineage.

**Only eudicot YAB5 genes constitute a monophyletic group that has a significant statistical support**
Monophyly of the *YAB5* group is well supported, with the exception of two *YAB5*-like sequences from the two basal angiosperms, *Cabomba caroliniana* and *Nuphar advena* (Figure 1, Table 1). Eudicot *YAB5*-like genes receive significant support in all analyses combining zinc finger and YABBY domains, as well as in analyses of the zinc finger domain alone (Table 1). It has been previously noted that no *YAB5* orthologue has been found in monocots [7]. Since *YAB5*-like sequences, including those of *Cabomba caroliniana* and *Nuphar advena*, are isolated from both the *YAB2* group and the group containing *CRC, FIL* and *INO* by branches with significant statistical support, and the fact that all these groups encompass taxa from basal angiosperms to core eudicots, allows the conclusion that also *YAB5* genes, including the two basal angiosperm genes, represent an early-angiosperm paralogue lineage. Taken together, all angiosperm *YABBY* gene groups can directly or indirectly be shown as being monophyletic; hence we refer to them as clades hereafter.
**Additional motifs characterize YABBY gene clades**

To investigate whether the variable amino acid regions outside of the zinc finger and YABBY domains show any sequence conservation at the level of gene clades, we conducted searches for conserved motifs in each of the five clades. Because these variable regions could not be aligned unambiguously, and were therefore excluded from phylogenetic analyses, we used the MEME program that does not require aligned sequences [18]. Most motifs detected by MEME were specific to one gene clade (Figure 2). Significantly, monophyly of the FIL-like genes is supported by seven motifs unique to this gene clade. Poales FIL-like genes, that show strong support in the Bayesian analyses (Figure 1), are also characterized by a longer coding region between the FIL-D motif and the YABBY domain, and alanine repeats following the FIL-F motif (Additional file 4). Further, the FIL-C motif was not detected in some grasses and basal angiosperms (Figure 2, Additional file 4).

CRC-like and INO-like genes, both well supported in the Bayesian analyses, have four and one unique motifs, respectively. CRC genes do not share motifs with FIL-like genes, further indicating that the FIL and CRC clades are monophyletic with respectively unique synapomorphies. YAB2-like and YAB5-like genes each contain four unique motifs that suggest that these clades diverged significantly. It is noteworthy that the two YAB5-like sequences from basal angiosperms whose affinities remained unresolved in the Bayesian analyses share some motifs with the well supported eudicot YAB5-like genes, providing additional evidence that these sequences are genuine YAB5-like. Specifically, the YAB5-B and YAB5-D motifs were detected in *Cabomba caroliniana*, and the YAB5-A and YAB5-C motifs were found in *Nuphar advena* (Figure 2, Additional files 4 and 5). Interestingly, two motifs are conserved between gymnosperm YABBY genes and angiosperm YAB2- and YAB5-like genes. These were accordingly named GY/YAB2/5-A and GY/YAB2/5-B,
respectively (Figure 2, Additional file 4 and 5). Furthermore, visual inspection of the 
amino acid alignment reveals that the FIL-B motif is somewhat similar to the amino 
end of the gymnosperm GY-A motif and to the YAB2-B and YAB5-B motifs of some 
basal angiosperms and monocots (Additional file 4). These cross-clade sequence 
similarities likely reflect preserved plesiomorphic characters in these angiosperm gene 
clades.

**Four gene duplications preceded diversification of angiosperms**

Since gymnosperm YABBY genes are separated from the five angiosperm YABBY 
clades, and because each angiosperm YABBY clade contains sequences from basal 
angiosperms, the YABBY clades must have originated through a series of four 
duplications that occurred just before the diversification of flowering plants. Although 
the angiosperm YABBY clades arose from a single gymnosperm ancestor gene, the 
precise result of the first duplication cannot be determined due to unresolved 
relationships between the YAB2 and YAB5 clades (Figure 1, Table 1). Three 
duplication scenarios are therefore possible (Figure 1C). In one scenario, YAB2 and 
YAB5 are sister clades, and both are sister to the clade containing INO, CRC and FIL. 
The other two scenarios display a basal grade of YAB2 and YAB5. In contrast, the 
subsequent duplications are well documented by the phylogenetic framework, 
dividing INO from the ancestor of CRC and FIL, and thereafter CRC and FIL. The 
clade composed of CRC, FIL and INO is resolved with significant support by 
partitioning the data (analyses 2, 3, 22; Table 1). Although the CRC and FIL clades 
are grouped together in most of the Bayesian inferences, their association is only 
supported in analyses of the YABBY domain alone with a more reduced sampling 
(analyses 23, 26, 32; Table 1). Loss of this support when adding the zinc finger
domain might indicate inconsistencies between the two domains. However, a partition homogeneity test showed no significant incompatibilities between the zinc finger and YABBY domain datasets ($p = 0.2670$). In addition, visual comparison of the trees does not suggest any significant conflicts in the datasets. Therefore, the loss of support is more likely due to the increase of homoplasy than to strong inconsistencies between the two domains.

**Different patterns of sequence variation amongst YABBY clades**

After the five *YABBY* genes had originated near the root of angiosperms, the duplicates show quite different evolutionary patterns. *CRC-* and *INO-*like genes accumulated a strong phylogenetic signal before angiosperm diversification, as is reflected by strong branch support in most analyses (Figure 1, Table 1). The same applies to eudicot representatives of *YAB5-*like genes. In contrast, *FIL-*like and *YAB2-*like genes, although likely monophyletic as well, have diverged much less and as a consequence have weak branch support. Genetic divergence of angiosperm clades from gymnosperm sequences also reflects this difference: it is higher for *CRC* (31.44 to 41.74%) and *INO* (29.47 to 41.40%), than for *FIL* (22.11-35.79%) and *YAB2* (22.11-36.84%). Genetic divergence between gymnosperms and *YAB5-*like genes however is also low (22.46-35.09%). Genetic divergence within clades appears not to be related with divergence from gymnosperms (27.02% for *YAB5*, 29.52% for *FIL*, 30.18% for *CRC*, 30.53% for *YAB2* and 34.79% for *INO*).

Interestingly, the most strongly diverged *CRC* and *INO* clades, and also *YAB5*, appear to have retained a single copy in angiosperms, while multiple copies of *FIL-*like and *YAB2-*like genes exist in several species (Figure 1). Although data for gymnosperms
are very scarce [1], *Picea sitchensis* has three *YABBYs* of genetic divergence comprised between 25.26 and 30.53%, which are likely to represent three paralogues. Therefore, retaining multiple gene copies may reflect an ancestral feature for *YABBY* genes.

To detect signature of molecular adaptation during *YABBYs* diversification, we analyzed the per site ratio of nonsynonymous to synonymous substitutions (ω) amongst lineages, using the codeml program implemented by PAML v.4.4 [19] (Figure 2). Comparison of the free-ratio model (allowing a different ω for each branch; –Ln L = 9771.11) against the one ratio model (attributing a unique ω for all branches; –Ln L = 9945.70) shows the free-ratio model as the best-fit for the data ($\chi^2 = 349.19$, 152 degrees of freedom, p = 0.00000). Most *YABBY* phylogeny branches have an ω value well below 1, indicating purifying selective constraint. Evidence of purifying constraint relaxation is found in some branches (Figure 2), but it remains unclear whether neutral or positive selection acted on these lineages (branch-site model leading to p > 0.05). All *YABBY* groups underwent purifying constraint relaxation at a given moment of their respective history or that of their ancestors, except for *FIL* that is separated from the gymnosperm *YABBY* ancestor by branches of very low ω (Figure 2).

**Isolated EcYABBYs are strongly expressed during vegetative and floral development**

The RT-PCR profiling of isolated *EcYABBYs* in a set of vegetative and floral tissues showed a strong expression in developing floral tissues and leaves (Figure 3).

Unsurprisingly, no *YABBY* transcripts were found in roots. The two *FIL*-like genes *EcYAB1* and *EcYAB2* show almost identical RT-PCR profiles, with a peak of
expression in young leaves of 2 mm in size, and a progressive down-regulation in older stages. Notably, EcYAB1 and EcYAB2 transcripts mainly concentrate in the leaf blade, are less abundant in the petiole, and are absent in the inflorescence stem. Expression of EcYAB5, a YAB5-like gene, is low in young seedlings, but stronger in the shoot tip of older plants. In addition, EcYAB5 expression is stronger in young leaves of 2-15 mm, and accumulates at lower levels in older leaves. The second pair of paralogous genes, EcYAB3 and EcYAB4, that belong to the YAB2 clade, shows very contrasting RT-PCR profiles in the tested tissues. Whereas strong expression of EcYAB3 is detected throughout almost all developmental stages, EcYAB4 transcripts are abundantly found only in two of the floral tissues but otherwise its expression is reduced or not detectable. EcYAB3 as well as EcYAB5 are expressed at considerable levels in the inflorescence stem and cauline leaves.

**YABBYs ancestral expression patterns**
To investigate how shifts in gene expression are associated with gene clade evolution, we attempted to infer the ancestral state of YABBY expression based on published data and results provided in this study. We used Mesquite v.2.73 [20] to distinguish between presence and absence of vegetative expression, since we found no case in which expression occurred only in vegetative tissue. The ancestral expression pattern for angiosperm YABBY genes included most likely vegetative tissue in addition to reproductive tissues (MP = 0.99; Figure 2). Our data indicate that vegetative expression was lost several times during YABBY gene clade diversification. Our analysis suggests that within the CRC-clade, vegetative expression was likely lost early on (MP = 0.85), as reported for the Amborella CRC copy [21], a condition that is also retained in eudicots (MP = 0.98). According to this assumption, CRC
expression in leaves was re-gained in monocots (MP = 0.92), where these genes assumed a novel role in leaf midrib differentiation that is distinct from the role of other YABBY genes in leaf development [22,23]. Another instance of loss of vegetative expression occurred for OsYABBY3, while two other rice FIL-like genes have retained leaf expression [7].

Information about YABBY expression is still spotty, but the currently available data suggest that FIL-, YAB2- and YAB5-like genes have retained a more ancestral expression pattern that includes both leaves and floral tissue. Hence, CRC- and INO-like genes exhibit markedly more derived characteristics that are reflected in both expression pattern and coding sequence evolution.

Discussion
The evolutionary framing of the YABBY gene family benefits from a multilayered phylogenetic approach
Accurate evolutionary histories of gene and taxon lineages constitute an essential necessity for comparing ortholog function and subsequently inferring gene function and developmental module changes [16,24,25]. This leads to a closer association of phylogenetic with evolutionary-developmental approaches [26]. However, establishing the phylogenetic history for a gene family can be challenging due to a combination of a limited number of alignable characters with a huge taxonomic sampling that covers five times the whole angiosperms (Figure 1). During the last decade, phylogenetic reconstructions of angiosperm relationships have been greatly advanced by a considerable increase of sequences for both genetic regions and taxonomic groups [17], allowing to reduce sampling bias effects. However, while the addition of molecular markers is driving major improvements for organism phylogenetic reconstructions [27], gene family phylogenies are often limited by the
number of reliably aligned characters. Here we have used a multilayered approach in an attempt to overcome this restriction and to better exploit the existing characters. In our analyses of 288 base pairs of \textit{YABBY} genes, Bayesian algorithms proved more powerful than a neighbor-joining and parsimony approaches (compare Figure 1 and Additional files 2 and 3). Further, varying the taxonomic sampling and/or sequence dataset helped to resolve certain relationships where the phylogenetic signal was otherwise overwritten by homoplasy, for example, \textit{FIL} and \textit{CRC} were only resolved in analyses of the \textit{YABBY} domain alone (Table 1). Partitioned analyses (Table 1) confirmed the suitability of incorporating various sequence evolution models to best explain the data [28 and references therein]. In addition, the third codon position, less relevant for amino acid identity and therefore more subject to saturation, was shown to contribute to the phylogenetic signal in our \textit{YABBY} analyses (Table 1).

Although the coding regions outside of the zinc finger and \textit{YABBY} domains were too variable between (and usually even within) gene clades to be included in Bayesian phylogenetic analyses, we found that members of each of the five gene clades share between one and seven unique amino acid motifs (Figure 2). These motifs represent synapomorphies that are valuable to underscore the monophyly of gene clades not significantly supported by Bayesian analyses of the conserved domains. Motif discovery as detected by the program MEME has been useful in the study of other plant transcription factor families [29-33].

\textbf{Duplication history of the major clades}
Since our analyses could not resolve the relationships between \textit{YAB2}, \textit{YAB5} and the remaining (\textit{INO}, \textit{CRC}, \textit{FIL}) clade, the duplication history of \textit{YABBY} genes remains
partially unclear. This leads to three possible duplication scenarios (Figure 1C). Paralogs of the five distinct *YABBY* clades are found in all extant angiosperms, including basal-most taxa, thus indicating that these duplications occurred before the diversification of angiosperms. Whether these gene duplications resulted from duplications of the whole genome or smaller genome fragments is more challenging to determine. Several studies point to a polyploidization event at the base of angiosperms between 101 and 168 million years ago [34], although it is not clear whether this event occurred before the split of the basal most angiosperm taxon *Amborella* [35]. It is tempting to speculate that this polyploidization could have been a driving force to acquire the traits linked to *YABBY* genes that are characteristic for angiosperms, such as carpel and outer integument morphogenesis [2,4,21,34,36-39]. The phylogenetic framework established here does not discriminate between the three possible duplication scenarios (Figure 1C). However, the first scenario in which *YAB2* and *YAB5* are sister clades (Figure 1C, top) is the only one that may account for a polyploidization event generating four clades simultaneously, namely *YAB2, YAB5, INO,* and the ancestor of *CRC* and *FIL*, hence reducing the number of duplication events to three. In the other two scenarios (Figure 1C, middle and bottom) four independent duplication events are necessary. It is likely that the *YABBY* gene family experienced additional rounds of smaller fractionated genome duplications, such as the one leading to the formation of the *CRC* and *FIL* clades. In fact, evoking polyploidization at this later stage of *YABBY* family diversification would imply the loss of many duplicates generated. All three scenarios are consistent with *YAB2* and *YAB5* sharing extra-domain motifs with gymnosperm ancestors.
Relationships of angiosperm YABBY gene clades

Only CRC and INO genes could be shown to be monophyletic clades in previous phylogenetic studies of the YABBY gene family [6-8], but their relationship to other YABBY genes has remained obscure. In some studies, CRC and INO genes tended to cluster together and appeared to be sister clades [6,7], suggesting that loss of vegetative expression occurred only once in the common ancestor of CRC and INO. Our study clarifies that CRC genes share a common ancestor with FIL genes, and that INO genes are sister to that ancestral CRC/FIL gene. This implies that the more specialized functions in floral development likely evolved independently in both CRC and INO. This is in line with our finding of purifying selective pressure in the last common ancestor of CRC and INO ($\omega = 0.0325$). Release of the purifying constraint is observed within the CRC clade at three instances (Figure 2) and in the branch leading to INO ($\omega = 0.5354$), likely implying changes in protein function.

Ancestral CRC genes functioned in carpel polarity, which represents a subfunctionalization of a general role of YABBY genes in leaf polarity for the carpel, a novel organ of angiosperms [8,21,39]. Our study indicates that loss of CRC expression in vegetative leaves coincides with the acquisition of the role in carpel polarity. In Arabidopsis, CRC coding regions have retained the ability to function in abaxial leaf identity when misexpressed in leaves, but FIL and YAB3, the sister genes, cannot replace CRC in gynoecium and nectary development [38,40]. This is in accordance with considerable modifications of the CRC protein, as opposed to changes in cis-regulatory regions, as evident from the strong phylogenetic signal in the zinc finger and YABBY domains, the lack of conservation of the last about 11 amino acids in the YABBY domain (Additional figure 4), and the acquisition of several additional conserved motifs in the coding region. CRC expression and function underwent considerable modifications during angiosperm diversification,
including a loss of the carpel polarity role and acquisition of carpel identity role in grasses, a novel role in grass leaf midrib formation, and a novel role in floral nectary development in core eudicots [2,6,8,22,23,39,41,]. Interestingly, none of these events were accompanied by retention of gene duplicates [8].

INO genes, like CRC, have strongly diverged in their coding sequence from putative ancestral YABBY genes, reflected in a strong phylogenetic signal (Figure 1). INO evolution might well be related to the evolution of bitegmic ovules, a synapomorphy of angiosperms [4,37], and more likely represents a neo-functionalization since the homology of integuments with leaves is questionable [42]. However, few sequence, expression, and functional data are available for INO compared to the other YABBY gene clades, impeding a clearer picture of its evolution at this point. For example, it remains unclear how widespread vegetative expression of INO, as reported for Oryza [7], may be among angiosperms.

Our results suggest that the single YABBY gene of the common ancestor of angiosperms was expressed in both reproductive and vegetative tissues (Figure 2). The "vegetative" angiosperm YABBY clades YAB2, YAB5, and FIL, have likely retained some ancestral functions. Our analyses clearly establish that these three clades do not form a monophyletic clade, although they show high degrees of expression and functional similarity [3,9,11,13]. We show here that YAB2 and YAB5, and to a lesser degree FIL genes, share motifs with gymnosperm YABBY genes, that are lost in the more derived CRC and INO proteins and can therefore be considered plesiomorphic. A more ancestral nature of the "vegetative" YABBYs is also reflected in the generally weaker branch support in Bayesian analyses of the zinc finger and
YABBY domains that are unable to resolve FIL genes as a clade (Figure 1, Table 1). The FIL amino acid sequence of the zinc finger and YABBY domains appear frozen by maintenance of strong purifying pressure in all branches between gymnosperms and FIL ($\omega$ of 0.0100, 0.0325, 0.0123 respectively). The "vegetative" angiosperm YABBY genes are exclusively expressed in leaf-homologous organs, both vegetative and floral [3,5,9,11-13,43-46]. Evidence from core eudicots have shown that vegetative YABBYs act in several aspects of leaf development, including the establishment and maintenance of leaf polarity, leaf margin establishment that guides lamina growth and leaflet initiation, activation of leaf maturation processes, and repression of shoot apical meristem genes [3,9-14,47]. Vegetative YABBYs also regulate SAM development, including phyllotaxy [15]. Why were representatives of the three vegetative YABBY clades retained in most lineages, given their largely redundant expression and function? It may be that redundancy in these essential developmental roles between the three vegetative YABBY clades may serve as a developmental buffer to stabilize these essential functions of shoot development, and might have been retained for that reason. Some vegetative YABBY functions may also depend on the formation of heterodimers between FIL and YAB5 as well as between YAB2 and YAB5-like genes in protein complexes, as reported in Arabidopsis [13]. These interactions would act to preserve the largely congruent expression in leaves of vegetative YABBYs and prevent the loss of one of the three genes during evolution. A similar case is seen in floral B-class MADS-box proteins that form multimeric complexes and are expressed in similar domains [48,49].

Nevertheless, YAB5 genes appear to have been lost in monocots [7]. It may be speculated that the higher number of FIL and YAB2 duplicates in monocots (grasses)
might replace the missing YAB5 in protein complexes. It is also interesting that eudicot YAB5 gene grouping was strongly supported in many of our analyses (Table 1), suggesting that the basal angiosperm YAB5 proteins may have not acquired all eudicot-specific features. Studies of basal angiosperm YAB5-like genes could help in understanding whether basal angiosperm YAB5 genes differ from their eudicot counterparts, and why this clade was lost in monocots.

Evolutionary trends of YABBY genes in the Poaceae lineage include several unique features, such as loss of one gene clade (YAB5), a strong release of purifying constraint acting on duplicates after polyploidization (YAB2), diversification of expression patterns (FIL), or the acquisition of a new function (CRC). Additionally, GC accumulation was detected at the third codon position of Poaceae FIL genes. Both domains are affected, although this tendency is stronger in the zinc finger domain (92.86 to 100% GC content). The bimodal distribution of GC content of grass genes has been previously documented, with some genes showing a GC content above 80% that is largely explained by GC accumulation at third codon positions [50]. Deciphering the origin and functional significance of this trend would help to shed light on evolutionary changes of FIL genes in the Poaceae family.

It will be interesting to see to what extent gymnosperm YABBY genes behave like their "vegetative" angiosperm counterparts, particularly whether they show leaf-specific expression, and function in both leaf and shoot development. While gene sampling in gymnosperms remains poor, multiple YABBY copies have been cloned from Picea sitchensis. To this point, nothing is known about the function or spatial and temporal expression pattern of YABBY genes in gymnosperms [1].
In several lineages, FIL and YAB2 clades show a trend toward retaining duplicates (Figure 1). This is contrasted by the apparent elimination of duplicates in the YAB5, INO, and CRC clades. Recent research indicates that a high percentage of genes encoding transcription factors are retained after whole genome duplications, whereas these duplicates are usually removed after small scale duplications [51,52]. This is explained by the gene balance hypothesis, which states that components of multi-subunit complexes have to be in stoichiometric balance for proper function [52].

Vegetative YABBY proteins form complexes with the transcriptional repressor LEUNIG [13]. Some YAB2 and FIL paralogues are found in lineages that have experienced recent polyploidization events [53,54], suggesting maintenance of additional copies due to preserved gene balance for example in Eschscholzia, core-eudicots, and Poaceae. On the contrary, the lack of duplicates in CRC, INO, and also YAB5 clades may indicate these proteins do not act in multi-component complexes, or that additional copies are deleterious or were otherwise removed.

Expression of EcYABBYs indicates conservation among eudicot

This is the first report on global YABBY gene expression from a basal eudicot, Eschscholzia californica. The five newly isolated "vegetative" YABBY genes are broadly expressed during leaf and flower development. EcYAB1, EcYAB2, EcYAB3, and EcYAB5 are strongly expressed in very small leaves, indicating a putative function of these genes in early stages of leaf development (Figure 3). Strong gene expression especially during initiation and outgrowth of lateral organs was also reported for Arabidopsis and Antirrhinum YABBY genes GRAM and PROL [3,9,11,12]. In Solanum and Tropaeolum majus, expression of FIL-like YABBY genes
occurs along the adaxial-abaxial boundary close to the margins, suggesting a potential role of YABBY genes in the formation of marginal structures typical of dissected and serrated leaves [45,47]. Studying tissue-specific expression of the FIL-like EcYAB1 and EcYAB2 using in situ hybridization, and especially functional studies of these and other vegetative YABBY genes in a species such as Eschscholzia may give valuable insights into the role of YABBY genes in leaf dissection. The YAB2-like EcYAB3 and the YAB5-like EcYAB5 were found to be weakly expressed in the stem, similar to the Arabidopsis YAB2 and YAB5 [9], indicating further conservation between core and basal eudicots.

Conclusions

A thorough understanding of the role of transcriptional regulators of development requires the elucidation of their evolutionary history. Phylogenetic studies of gene families can be challenged by short alignable coding regions and a rapid sequence of duplications. We demonstrate that the use of a multilayered Bayesian analysis allows to overcome some of these limitations, improving the resolution of YABBY gene evolution. We show that each of the five YABBY copies that emerged prior to angiosperm diversification underwent a distinct evolutionary path that is reflected in changes in selection pressure, the emergence of copy-specific sequence motifs, and changes in expression pattern. We conclude that model system based research on YABBY gene function should be complemented by expression and functional studies in phylogenetically distant and morphologically diverse lineages, including gymnosperms, basal angiosperms, and basal eudicots, in order to reveal the role of YABBY genes in the evolution of seed plant shoot development.
Methods

Plant material
Seeds of *Eschscholzia californica* (Cham.) sub. californica variety Aurantica Orange were purchased from Larner Seeds (Bolinas, CA, USA). This material was used to isolate sequences and address the expression pattern of the *Eschscholzia californica* homologues of the genes *FIL*, *YAB2* and *YAB5*. With the exception of these sequences, all the remaining ones included in the phylogenetic analyses were previously published [2-5,7,8,11,12,21-23,37,39,43,44,46,50,55-63] and/or deposited in GenBank.

Isolation of *Eschscholzia* YABBY sequences
3' RACE-PCR carried out on shoot derived cDNA with degenerated primers designed in the YABBY or zinc finger domains (Additional file 6), led to the amplification of the coding sequence of five *Eschscholzia* YABBY-like genes. The isolated sequences were blasted in NCBI, which confirmed the presence of the zinc finger/YABBY domain combination that is typical for YABBY genes. The sequences were then named *Eschscholzia californica* YABBY 1 (*EcYAB1*) to *EcYAB5* (GenBank accession numbers HQ116795 to HQ116799, respectively).

RNA was extracted from the shoot tip of 3 to 6-week-old *Eschscholzia californica* using TRI reagent (Sigma Aldrich, St. Louis, USA) following manufacturer’s instructions. First strand cDNA of 1µg total RNA was synthesized with M-MLV reverse transcriptase (Promega, Madison, USA), and the primer AB05 (Additional file 6). For the initial 3' RACE-PCR degenerated primers complementary to the zinc finger and YABBY domain were used. The primers EcY12F or Y1F527 were used together with the primer AB07 to amplify *FIL*-like sequences. *YAB2*-like sequences were amplified using primers EcY25F and EcY26R, and *YAB5*-like sequences were
amplified with the primers EcY28F and EcY28F. Products were amplified using 35 cycles of 30 s at 94°C, 30 s at 55°C, and 60 s at 72°C in a standard Bio-Rad Dyad thermocycler. Additional 5′ RACE-PCR was utilized using the primers EcY08R, EcY09R and EcY10R for EcYAB1, EcY54R, EcY53R and EcY40R for EcYAB3, EcY69R, EcY70R and EcY71R, and EcY44R, EcY45R and EcY50R for EcYAB4 and EcYAB5, respectively. All 5′ RACE-PCR were conducted using the 5′/3′ RACE kit, 2nd Generation (Roche Diagnostics, Mannheim, Germany) following manufacturer’s instructions. To verify correct assembly of partial sequences full length sequences were amplified. Full-length sequences of the five Eschscholzia YABBYs were obtained with primers EcY11F and ABO7 (EcYAB1), EcY11F and EcY21R (EcYAB2), EcY61F and ABO7 (EcYAB3), EcY72F and ABO7 (EcYAB4), and EcY62F and EcY63R (EcYAB5). Consensus sequences for all Eschscholzia genes were constructed with 15 to 25 independent sequence reads, and deposited in GenBank.

RT-PCR profile of EcYABBYs

For generating the expression profile, total RNA was isolated from Eschscholzia tissues and first strand cDNA was synthesized as stated above. In the subsequent PCR, 5μl of the 1:50 diluted cDNA were used in a 25μl reaction. The primers EcY37F and EcY13R were used for EcYAB1, and EcY37F and EcY14R for EcYAB2. The primers EcY31F and EcY39R were used to amplify EcYAB3, and EcY31F and EcY68R, and EcY28F and EcY43R for amplifying EcYAB4 and EcYAB5, respectively. ACTIN was amplified using primers Act2-fw and Act2-rev. Products were amplified using 27 cycles for ACTIN, and 30 cycles for EcYAB1, EcYAB2 and EcYAB3, respectively, 35 cycles for EcYAB4 and 37 cycles for EcYAB5 of 30 s at 94°C, 30 s at 55°C, and 60 s at 72°C. At least five technical replicates were run for
each gene. In addition, two biological replicates were used. cDNA synthesized without reverse transcriptase was used as control and resulted in no amplification.

**Phylogenetic analyses of YABBY and zinc finger domains**

*Eschscholzia* putative *YABBY* sequences were edited with other *YABBY* sequences from GenBank using MacClade 4.08 [64], and manually aligned with a codon-preserved strategy. *YABBY* sequences isolated from gymnosperms were used as outgroup after verifying that none of them clustered with any of the *YABBY* family gene groups found in angiosperms. Phylogenetic tree reconstruction was restricted to the zinc finger and YABBY domains to avoid alignment uncertainties (Additional file 7).

Bayesian inference (BI) was carried out with MrBayes version 3.1.2 [65]. Data sets corresponding to the different domains and the codon positions (cp) were analyzed both separately and simultaneously (Table 1). Concatenate analyses were carried out using data partition and without it (Table 1). The best-fit models of nucleotide substitutions for our datasets were selected by the AIC method with jModelTest 0.1 [66,67] (Additional file 8). Partitioned analyses were conducted with model parameters unlinked, tree topologies and branch lengths linked. Four Markov chains were run simultaneously for at least 10x10⁶ generations, and these were sampled every 1000 generations. Data from the first 1,000 generations were discarded as the *burn-in* period, after confirming that likelihood values were stabilized prior to the 1,000th generation. The 50% majority rule consensus trees and posterior probability (PP) of nodes were calculated from the pooled samples. Post-stationarity consensus trees from the two runs were verified to be identical, that is, the average standard deviation of split frequencies gets below 0.01 [68], and the two trees have similar
topology and estimation of DNA sequence change rate. Neighbor-joining analysis was carried out with PAUP* 4.0b10 [69], the support values for the branches obtained with 1000 bootstrap (BS) replicates. Searches for most parsimonious trees were conducted with PAUP* 4.0b10 [69], using an heuristic search of 1000 random sequence addition replicates, holding one tree at each step during the stepwise addition and no more than 500 trees of length \( L \geq 1 \) on each sequence addition replicate (following a similar approach as [70,71] and suggested in [72]). Starting trees were obtained via stepwise addition. All characters were unordered and equally weighted, and gaps were treated as missing data. Uninformative sites were excluded.

To ensure that no shorter trees exist, but also that the strict consensus tree obtained reflects all most parsimonious trees, a second search was then performed using the strict consensus as a constraint in a search of 10,000 random addition replicates, saving one tree of \( L \geq 1 \) on each replicate, and holding only not compatible trees (search strategy described in [73]).

**Congruence of zinc finger and YABBY domains**

Compatibility of the zinc finger and YABBY domain datasets was addressed by conducting a partition homogeneity test (incongruence length difference "ILD", [74]) as implemented by PAUP*, using 1000 homogeneity replicates and an heuristic search of 100 random sequence addition replicates, saving one tree at each step during the stepwise addition and no more than 1 tree of \( L \geq 1 \) on each sequence addition replicate. The ILD test depends on the length of the most parsimonious trees and not on recovering all of them [75], and therefore we adopted here a strategy that privileges the increased number of ILD replicates and thus a lower variance in ILD p-value. Calculation was conducted excluding uninformative sites [76].
**Motif discovery**

Full length *YABBY* amino acid sequences were searched for motifs outside the conserved zinc finger and YABBY domains, using the online program MEME version 4.1.0 [18]. MEME does not require an alignment of these variable regions. The initial search was done using all 93 sequences from the large data set (Additional file 7) with the search parameters set to 30 motifs of 5 to 50 amino acids in width. Motifs were expected to occur once or none per sequence. Because most identified motifs were found to be gene clade specific, further analyses were conducted with sequences of individual clades or sister clades. In these searches the parameters were set as before and only 10 motifs were calculated. The region between zinc finger and YABBY domain was additionally searched individually with parameters set to calculate 3 motifs of 5 to 15 amino acids in width. Motif occurrence and relative position within the amino acid sequence was mapped with relation to the zinc finger and YABBY domain (Figure 2, Additional files 4 and 5).

**Detection of positive selection on YABBY genes**

In order to identify branches of the *YABBY* gene tree along which release of purifying pressure and/or positive selection may have occurred, we used the codeml program implemented by PAML v.4.4 [19]. We first tested for evidence of released purifying pressure comparing the one ratio model M0 that assumes a constant dN/dS ratio (=ω, the per site ratio of nonsynonymous -dN- to synonymous -dS- substitution) along tree branches against the free ratio model that permits the ω of branches to vary. Branches with ω ≥ 1 were successively used as foreground for comparing the null and alternative version of modified branch-site model A (MA), to assess whether positive selection may have occurred on these branches. Likelihood ratio test (LRT) was used to compare the fit of model pairs. 2Δ values were calculated and posteriorly transformed into exact p-values using the chi-square function with a degree of
freedom determined as the difference of parameter numbers between the two compared models. The level of \( p \leq 0.05 \) was considered as significant.

**Inference of ancestral character state of expression profile**

Ancestral character state history was generated with Mesquite v.2.73 [20] under a maximum likelihood optimization, using the Markov one-rate (Mk1) parameter model and the tree corresponding to analysis (3) (Table 1). Marginal probabilities (MP) \( \geq 0.95 \) were considered as significant. Only the expression in vegetative tissues was inferred, because expression in flower was a permanent feature in the species studied.

**Authors' contributions**

SG and CB conceived and designed the general framework of the study, and OH the phylogenetic strategy. CB isolated the sequences and carried out the RT-PCR. OH and CB carried out the phylogenetic analyses. All authors analyzed the data, wrote the paper, and approved the final manuscript.

**Acknowledgements**

The authors thank John Bowman for providing some previously published *YABBY* sequences that were not available in the databases, Annette Becker and Sabrina Hutter for providing a partial *EcYAB1* sequence, Daryl W. Lam for installing PAML and Steven Romine for computer maintenance. Funding for this work was provided by Ohio University. CB received a Donald Clippinger fellowship from Ohio University and OH a MICINN postdoctoral grant from the Ministerio de Ciencia e Innovación, Spain.
References


36. Frohlich MW, Chase MW: After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature reviews* 2007, 450:1184-1189


42. Kelley DR, Gasser CS: **Ovule development: genetic trends and evolutionary considerations.** *Sex Plant Reprod* 2009, **22**:229-234

43. Juarez MT, Twigg RW, Timmermans MC: **Specification of adaxial cell fate during maize leaf development.** *Development* 2004, **131**:4533-4544

44. Zhao W, Su HY, Song J, Zhao XY, Zhang XS: **Ectopic expression of TaYAB1**, a member of the *YABBY* gene family in wheat, causes partial abaxialization of the adaxial epidermises of leaves and arrests the development of shoot apical meristem in *Arabidopsis*. *Plant Science* 2005, **170**:364-371


46. Tononi P, Möller M, Bencivenga S, Spada A: **GRAMINIFOLIA homolog expression in Streptocarpus rexii** is associated with the basal mersitem in phyllomorphs, a morphological novelty in Gesneriaceae. *Evolution & Development* 2010, **12**:61-73

47. Eshed Y, Izhaki A, Baum SF, Floyd SK, Bowman JL: **Asymmetric leaf development and blade expansion in Arabidopsis** are mediated by **KANADI and YABBY activities.** *Development* 2004, **131**:2997-3006

heterodimerization originated from homodimerization. *Mol Biol Evol*
2002, 19:587–596


55. Hirayama Y, Yamada T, Oya Y, Ito M, Kato M, Imaichi R: Expression patterns of class I *KNOX* and *YABBY* genes in *Ruscus aculeatus*
(Asparagaceae) with implications for phylloclade homology. *Dev Genes Evol* 2007, **217**:363-372


60. Ralph SG, Chun HJ, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA et al.: **A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis).** *BMC Genomics* 2008, **9**:484

61. Cong B, Barrero LS, Tanksley SD: **Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication.** *Nat Genet* 2008, **40**:800-804


73. Catalán P, Kellogg EA, Olmstead RG: **Phylogeny of Poaceae subfamily Pooideae based on chloroplast ndhF gene sequences.** *Mol Phyl Evol* 1997, **8**:150-166

74. Farris JS, Källersjö M, Kluge AG, Bult C: **Testing significance of incongruence.** *Cladistics* 1994, **10**:315-319

75. Berry PE, Hipp AL, Wurdack KJm Van Ee B, Riina R: **Molecular phylogenetics of the giant genus Croton and tribe Crotoneae (Euphorbiaceae sensu stricto) using ITS and trnL-trnF DNA sequences data.** *Am J Botany* 2005, **92**:1520-1534

76. Lee MSY: **Uninformative characters and apparent conflict between molecules and morphology.** *Molecular Biology and Evolution* 2001, **18**:676-680


Figures

Figure 1 - Phylogenetic framework of the YABBY gene family

(A) Bayesian phylogram resulting from the analyses of the zinc finger and YABBY domains for 93 homologous YABBY sequences. Topology corresponds to the analysis (3) with data partitioned by domain and by codon position (Table 1). Wider branches are those significantly supported (PP ≥ 0.95). Numbers in brackets above the branches indicate significant support of that branch in an analysis other than (3), the numbers referring to analyses listed in Table 1. G: gymnosperms, BA: basal angiosperms, M: monocots, BE: basal eudicots, CE: core eudicots. Black circles on the nodes represent the putative duplication events. GenBank accession numbers are indicated after the plant names.

(B) Unrooted Bayesian phylogram resulting from the analysis of the zinc finger and YABBY domains for 29 homologous YABBY sequences, corresponding to the analysis (31) (Table 1). Wider branches are those significantly supported (PP ≥ 0.95). Numbers in brackets above the branches indicate significant support of that branch in an analysis other than (31), the numbers referring to analyses listed in Table 1.

(C) The three possible duplication scenarios explaining the diversification of YABBYs. Black circles represent the putative duplication events.

Figure 2 – Gene expression, selection pressure and amino acid motifs of the YABBY family

Topology of the tree corresponds to the analysis (3) (Figure 1, Table 1). GenBank accession numbers are indicated after the plant names. To the left of sequence names are schematic representations of gene expression, when known [2-7,9,11,12,21-23,37,39,43,44,46,50,55-63,77,78]. No symbol indicates lack of data. Ancestral character state of gene expression is inferred along lineages. Pie charts at nodes
represent the proportional likelihoods (probability) of the ancestral state inferences. Expression in vegetative tissues is indicated in black, its absence in white, and unknown data in grey. Expression in flower was not considered because of its constancy. Width of the branches is proportional to the $\omega$ value calculated by PAML v.4.4 [19] with the free ratio model (that allows the $\omega$ of branches to vary) until reaching one. For $\omega \geq 1$ a box indicated the $\omega$, $d_N$ and $d_S$. To the right of the sequence names are schematic representations of YABBY protein structures. Black boxes indicate zinc finger and YABBY domain, respectively. Colored boxes represent different amino acid motifs shared by any sequences as calculated by MEME. The same color indicates the same motif. Motif length and position are depicted approximately in relation to box sizes, and position relative to zinc finger and YABBY domains. Motif sequences are listed in Additional file 5.

**Figure 3 - RT-PCR profile showing expression of *Eschscholzia californica* YABBY genes**

The five Eschscholzia vegetative *YABBY* genes are differentially expressed in vegetative and reproductive development. Actin was used as a reference. See text for details.
Additional files

Additional file 1.pdf

Genetic divergence between the homologues of *Eschscholzia californica* YABBY-like genes
Lower part shows uncorrected pairwise distances as calculated by PAUP* v.40b0 [69], upper part shows pairwise level of amino acid identity as calculated by MEGA. Only zinc finger and YABBY domains were analyzed, and values are expressed in percentages. ¹Indicates the values corresponding to the pairs of paralogues.

Additional file 2.pdf

Neighbor-joining tree based on the zinc finger and YABBY domain alignment of 93 homologous YABBY sequences.
Wider branches are those with bootstrap ≥ 70%, calculated with 1000 replicates using PAUP* v.40b0 [69]. GenBank accession numbers are indicated after the plant names.

Additional file 3.pdf

Bootstrap consensus tree generated by parsimony analysis, based on the zinc finger and YABBY domain alignment of 93 homologous YABBY sequences.
Bootstrap values ≥ 50% calculated with 114 replicates using PAUP* v.40b0 [69] are indicated on the branches. GenBank accession numbers are indicated after the plant names.

Additional file 4.pdf.

Amino acid alignment of YABBY coding regions
Gene groups are indicated on the left, and the position of conserved motifs and domains on the top. Blocks of conserved motifs are also marked in the alignment. GenBank accession numbers are indicated after the plant names.
Motifs identified in the amino acids sequence of YABBY genes
Motifs are named according to the schematic representation above the table. Sequence logos and regular expression were taken directly from MEME. In the sequence logos, the overall height of the letters at that position indicates sequence conservation. Heights of individual letters represent the probability of that letter to occur at that position and thereby are indicative of the relative frequency of the respective amino acid.

Primers used in this study

Nucleotide alignment of zinc finger and YABBY domains
Nucleotide alignment of the zinc finger and YABBY domains that were used in the phylogenetic analyses. Sequences used in different taxon samplings are specified. L: large sampling, M: medium sampling, S: small sampling. GenBank accession numbers are indicated after the plant names.

Numerical results for the 32 bayesian analyses
1 Taxon samplings are indicated in Additional file 7. 2 A: angiosperms, CE: core eudicots, E: eudicots, G: gymnosperms, M: monocots. 3 Z: zinc finger domain, Y: YABBY domain, small numbers following domain symbol denote first, second, and third codon positions, aa: amino acids.
Additional files provided with this submission:

Additional file 1: Additional file 1.pdf, 58K
Additional file 2: Additional file 2.pdf, 499K
http://www.biomedcentral.com/imedia/1333938114495032/supp2.pdf
Additional file 3: Additional file 3.pdf, 500K
Additional file 4: Additional file 4.pdf, 1138K
Additional file 5: Additional file 5.pdf, 2232K
Additional file 6: Additional file 6.pdf, 42K
Additional file 7: Additional file 7.pdf, 1145K
Additional file 8: Additional file 8.pdf, 78K