Selection on Length Mutations After Frameshift Can Explain the Origin and Retention of the *AP3/DEF*-Like Paralogues in *Impatiens*

Steven B. Janssens · Tom Viaene · Suzy Huysmans · Erik F. Smets · Koen P. Geuten

Abstract Evolution of class B genes through gene duplication has been proposed as an evolutionary mechanism that contributed to the enormous floral diversity. Frameshift mutations are a likely mechanism to explain the divergent C-terminal sequences of MIKC gene subfamilies. So far, the inferences for frameshifts and selective pressures on the C-terminal domain are made for old duplications for which the exact selective pressures are obscured by evolutionary time. This motivated us to study an example of a recent duplication, which allows us to consider in more detail the selective pressures that are involved after duplication. We find that after duplication and frameshift of *Impatiens* class B genes, the individual codons show no evidence for adaptive selection. It is rather the length of the C-terminal domain that either is strictly conserved or varies strongly. This suggests a role for the length of the C-terminal domain in the retention of duplicated genes.

S. B. Janssens (✉) · T. Viaene · S. Huysmans · E. F. Smets
Laboratory of Plant Systematics, Institute of Botany and Microbiology, K.U. Leuven, P.O. Box 2437,
BE-3001 Leuven, Belgium
e-mail: steven.janssens@bio.kuleuven.be

E. F. Smets
National Herbarium of the Netherlands, Leiden University Branch, P.O. Box 9514, NL-2300 RA Leiden, The Netherlands

K. P. Geuten
Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA

Introduction

Members of the MADS-box gene family function as key regulators throughout plant development (for a review see Becker and Theissen 2003; Hileman et al. 2006). In recent years, major progress has been made in understanding the evolutionary history of this gene family. From a phylogenetic point of view, the evolutionary history of the MADS-box gene family is characterized by frequent gene duplication and gene loss (Shinozuka et al. 1999; Parenicova et al. 2003; Hileman et al. 2006; Leseberg et al. 2006). Gene loss is regarded as the most likely fate for one of the duplicated copies. However, following a period of redundancy, retention of both copies can occur through neo- or subfunctionalization (Moore and Purugganan 2005). Neofunctionalization, indicating a novel function acquired by one of the two duplicates, can occur through changes in either the protein sequence or in the cis-regulatory sequences, while subfunctionalization is the process in which each duplicate implements different elements of the original function of the ancestral gene, and occurs most likely in cis-regulatory sequences (Ohno 1970; Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000; Zhang 2003; Moore and Purugganan 2005; Preston and Kellogg 2006). The occurrence of neofunctionalization through changes in coding sequences is considered to be detectable by comparing synonymous and nonsynonymous substitutions. Several studies in the MADS-domain family, however, were not able to find convincing evidence for this type of duplicate retention mechanism (Aagaard et al. 2005; Kellogg 2006; Preston and Kellogg 2006). Since differences in protein structure of type II MADS box genes are most obviously present in the C-terminal domain (Kramer et al. 1998, 2004; Kramer and Irish 2000; Vandenbussche et al. 2003) and frameshifts resulting from

length mutations can explain the origin of the characteristic C-termini of the different MADS-domain subfamilies (Vandenbussche et al. 2003), the length mutations occurring in the C-terminal domain could be positively selected for, a hypothesis which so far has not been investigated.

Within the MADS-box gene family, class B genes are one of the best-studied subfamilies (Kramer et al. 1998; Stellari et al. 2004; Aagaard et al. 2005; Zahn et al. 2005; Janssens et al. 2007). Floral homeotic class B genes, which comprise *AP3/DEF* and *PI/GLO*, are required for petal identity in the second whorl and stamen identity in the third whorl of eudicot flowers (Jack et al. 1992, 1994; Goto and Meyerowitz 1994; Theissen et al. 2000; Alvarez-Buylla et al. 2000). *AP3/DEF* belong to the type II class MADS-box genes, consisting of four distinct regions: a MADS domain (M), an intervening region (I), a keratine-like domain (K), and a C-terminus (C) (Munster et al. 1997). The MADS domain is essential for binding DNA to CarG promoter sequences. The intervening region is considered to be critical to mediate dimerisation with other MIKC-type proteins (Riechmann et al. 1996). In addition, the keratin-like domain, which is located between domain I and the C-terminus, is involved in mediating specific protein/protein interactions (Riechmann et al. 1996). The C-terminal region, which is most variable in length and sequence, contributes to the formation of higher-order protein complexes between dimers of MIKC proteins (Egea-Cortines et al. 1999; Honma and Goto 2001). The evolution of the C-terminal domain in *AP3/DEF* can be understood by inferring a frameshift mutation in the ancestral *paleoAP3*-motif. This frameshift gave rise to the present *euAP3*-motif in core eudicots (Vandenbussche et al. 2003; Kramer et al. 2006). Although these frameshift mutations are usually expected to cause loss of function due to the formation of premature stop codons or a fundamental change of the downstream protein sequence, they sometimes result in a functional fragment that will remain conserved in the genome (e.g., the *euAP3*-motif). Additionally, other C-terminally truncated class B genes have been found in diverse angiosperm species (e.g. Kramer et al. 2006).

In *Impatiens* (Balsaminaceae), a recent study identified two *AP3/DEF*-like genes with strongly divergent C-terminal domains in the species *I. hawkeri* (Geuten et al. 2006). One of these duplicates is characterized by a truncated C-terminal domain with loss of the *euAP3*- and *PI*-derived motifs (*IhDEF2*), while the other is unusually long but with *euAP3*- and *PI*-derived motifs present (*IhDEF1*). From a phylogenetic point of view, an *AP3/DEF*-like gene in *Marcgravia* (*MuDEF*; Marcgraviaceae) is sister of both *Impatiens AP3/DEF*-like genes, from which it can be derived that the duplication event that gave rise to the two paraloguous *ImpDEF* copies must have happened

subsequent to the Balsaminaceae-Marcgraviaceae split. In addition, this duplication most likely occurred prior to the origin of the genus *Impatiens*, because we were able to sequence both copies from several representatives of the genus (Janssens et al. 2007). Both genes in *I. hawkeri* (*IhDEF1* and *IhDEF2*) seem to have retained a rather similar expression pattern, occurring in particular parts of the corolla and the stamens (Geuten et al. 2006). In addition, we know that the two *AP3/DEF* paralogues in *Impatiens* show differences in protein interaction (Geuten et al. 2006).

The case of *Impatiens* is interesting because it concerns a relatively recent duplication, which seemingly coincides with the origin of the genus. This allows us to trace the events subsequent to duplication and frameshift with time being less of an obscuring factor.

In this paper we show that recent duplications and subsequent frameshift mutations that give rise to evolutionary retained duplicates can be considered an ongoing process in MADS-box gene diversification. Furthermore, we demonstrate that the diversification of such a recent duplicate pair is likely not because of positive selection on individual codons, but because of positive selection on the length of the C-terminal domain. The latter is a completely new and previously uninvestigated finding that helps us understand how these duplicates survive after their origin.

Materials and Methods

Molecular Protocols

DNA was extracted for 62 species (Table 1) using a modified version of the hot CTAB protocol (Geuten et al. 2006; Janssens et al. 2006). Specific primers used for the amplification of the C-terminal ends of the two *AP3/DEF* homologues in *Impatiens* (*ImpDEF1* and *ImpDEF2*) were designed using the cloned cDNA sequences of Geuten et al. (2006): IhCT-DEF1F (5′-GCAGATACACAGAAA CCTACTTCAG-3′), IhCT-DEF1R (5′-TTATAGTTAAA GCAAAGCATAGGTTGTGAG-3′), IhCT-DEF2F (5′-AC AGACACATAGAAACTTACTCCAGC-3′), and IhCT-DEF2R (5′-CCATCACCATCTTCATCTTCAATCC-3′). The temperature profile used to amplify the C-terminal end of *ImpDEF1* consisted of 2 min of initial denaturation at 94°C and 30 cycles of 30 s of denaturation at 94°C, 30 s of primer annealing at 52.5°C, and 1 min of extension at 72°C. Amplification of the C-terminal domain of *ImpDEF2* was executed under the same conditions as described above except for an annealing temperature of 53°C. Amplification reactions were executed with a GeneAmp PCR system 9700 (Applied Biosystems, Fostercity, CA, USA).

Table 1 Accession numbers, voucher data, and origin of plant material for taxa included in the DNA analyses

Taxon	Origin	Voucher	Accession no., C-terminus *ImpDEF1*	Accession no., C-terminus *ImpDEF2*
Impatiens apsotis Hook. f.	China, Sichuan	Yuan CN2k2-159 (NEU)	EU373212	—
Impatiens arguta Hook. f. & Thoms.	China, Yunnan	Yuan CN2k-74 (NEU)	EU373213	EU373273
Impatiens aurea Muhl.	North American origin, cult. Holden arboretum	Janssens SJ008 (LV)	EU373214	—
Impatiens aureliana Hook. f.	China, Yunnan	Yuan CN2k1-56 (NEU)	EU373215	EU373274
Impatiens auricoma Baill.	Comores origin, cult. Bot. Gard. Marburg	Janssens SJ001 (LV)	EU373216	EU373275
Impatiens balsamina L.	Indian origin, cult. Kruidtuin Leuven	Janssens SJ003 (LV)	EU373217	EU373276
Impatiens bequartii De Wild	Congo origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE1 (NEU)	EU373218	EU373277
Impatiens bicaudata H. Perrier	Madagascan origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373219	EU373278
Impatiens burtonii Hook. f.	African origin, cult. by Paul Shirley, the Netherlands	Paul Shirley s.n. (LV)	EU373220	EU373279
Impatiens campanulata Wight	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373221	EU373280
Impatiens capensis Meerb.	North American origin, cult. Holden arboretum	Janssens SJ009 (LV)	EU373222	—
Impatiens chinensis L.	China, Yunnan	Yuan CN2k1-49 (NEU)	EU373223	EU373281
Impatiens chimiliensis Comber	China, Yunnan	Yuan CN2k-51 (NEU)	EU373224	—
Impatiens chungtienensis Y.L. Chen	China, Yunnan	Yuan CN2k2-204 (NEU)	EU373225	EU373282
Impatiens columbaria J.J. Bos	African origin, cult. Nat. Bot. Gard. Meise	FB/S2966 (BR)	EU373226	EU373283
Impatiens conchibracteata Y.L. Chen	China, Yunnan	Hao 427 (NEU)	EU373227	EU373284
Impatiens congolensis G.M. Schulze & R. Wilczek	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE7 (NEU)	EU373228	EU373285
Impatiens cuspidata Wight & Arn.	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373229	EU373286
Impatiens davidi Franch.	China, Fujian	Yuan CN2k-09 (NEU)	EU373230	—
Impatiens delavayi Franch.	China, Yunnan	Chassot and Yuan 99–154 (NEU)	EU373231	—
Impatiens drepanophora Hook. f.	China, Yunnan	Yuan CN2k1-41 (NEU)	EU373232	—
Impatiens eberhardtii Tardieu	Vietnam, Anam	Song s.n. (NEU)	EU373233	EU373287
Impatiens edgeworthii Hook. f.	Himalayan origin, cult. Univ. California Bot. Gard. Berkeley	89.2005 (UC)	EU373234	EU373288
Impatiens engleri Gilg	Tanzania	Knox 3627 (LV)	—	EU373289
Impatiens flaccida Arn.	South Indian origin, cult. Nat. Bot. Gard. Meise	FB/S3925 (BR)	EU373235	EU373290
Impatiens flanaganae Hemsl.	African origin, cult. Roy. Bot. Gard. Edinburgh	19860179 (E)	EU373236	—
Impatiens forrestii Hook. f. ex W.W. Smith	China, Yunnan	Yuan CN2k-79 (NEU)	EU373237	—
Impatiens glandulifera Arn.	Belgium, Leuven	Janssens SJ002 (LV)	EU373238	—
Impatiens hawkeri W. Bull	New Guinean origin	Janssens SJ006 (LV)	EU373239	EU373291
Impatiens hians Hook. f.	African origin, cult. Bot. Gard. Berlin	Schwerdtfeger 9492a (B)	EU373240	—
Impatiens hochstetteri Warb.	Ethiopia	De Wilde 5610 (BR)	—	EU373292
Impatiens imbecilla Hook. f.	China, Sichuan	Hao 426 (NEU)	EU373241	—
Impatiens kamerunensis Warb.	Cameroon	VM s.n. (LV)	EU373242	EU373293

Table 1 continued

Taxon	Origin	Voucher	Accession no., C-terminus *ImpDEF1*	Accession no., C-terminus *ImpDEF2*
Impatiens keillii Gilg.	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE11 (NEU)	EU373243	—
Impatiens langbianensis Tardieu	China, Yunnan	Yuan s.n. (NEU)	EU373244	EU373294
Impatiens lecomtei Hook. f.	China, Yunnan	Yuan CN2k2-202 (NEU)	EU373245	—
Impatiens mackeyana Hook. f.	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer EF21 (NEU)	EU373246	EU373295
Impatiens mannii Hook. f.	East Congo	Geuven Carvalho 4534 (BR)	EU373247	EU373296
Impatiens mengtseana Hook. f.	China, Yunnan	Yuan CN2k1-38 (NEU)	EU373248	—
Impatiens meruensis Gilg.	Tanzanian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373249	EU373297
Impatiens monticola Hook. f.	China, Sichuan	Hao 425 (NEU)	EU373250	—
Impatiens napoensis Y.L. Chen	China, Yunnan	Yuan CN2k1-61 (NEU)	EU373251	—
Impatiens omeiana Hook. f.	China, Sichuan, cult. Univ. California Bot. Gard. Berkeley	2002.0214 (UC)	EU373252	—
Impatiens oxyanthera Hook. f.	China, Sichuan	Song S008 (NEU)	EU373253	—
Impatiens platypetala Lindl.	Bali, Indonesian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373254	—
Impatiens poculifer Hook. f.	China, Yunnan	Yuan CN2k2-209 (NEU)	EU373255	EU373298
Impatiens poilanei Tardieu	Vietnam, Anam	Song s.n. (NEU)	EU373256	—
Impatiens pseudoviola Gilg.	African origin, cult. Roy. Bot. Gard. Edinburgh	19680124 (E)	EU373257	—
Impatiens rubrostriata Hook. f.	China, Yunnan	Yuan CN2k1-44 (NEU)	EU373258	—
Impatiens sakeriana Hook. f.	Cameroon	VM 125 (LV)	EU373259	EU373299
Impatiens scutisepala Hook. f.	China, Yunnan	Yuan CN2k-56 (NEU)	EU373260	—
Impatiens soulieana Hook. f.	China, Sichuan	Yuan CN2k2-163 (NEU)	EU373261	—
Impatiens scabrida DC.	Himalayan origin, cult. Holden arboretum	941314 (DBG)	EU373262	—
Impatiens stenosepala Pritz. ex Diels	China, Shaanxi	Wang SH-001 (NEU)	EU373263	—
Impatiens tinctoria A. Rich	African origin, cult. Nat. Bot. Gard. Meise	2002-1249-61 (BR)	EU373264	—
Impatiens trichosepala Y.L. Chen	China, Yunnan	Yuan CN2k1-68 (NEU)	EU373265	—
Impatiens tuberosa H. Perrier	Madagascan origin, cult. Bot. Gard. Univ. Copenhagen	Janssens SJ005 (LV)	EU373266	EU373300
Impatiens uliginosa Franch.	China, Yunnan	Yuan CN2k2-173 (NEU)	EU373267	EU373301
Impatiens usambarensis Grey-Wilson	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE20 (NEU)	EU373268	EU373302
Impatiens usambarensis Grey-Wilson × *walleriana* Hook. f.	African origin, cult. Roy. Bot. Gard. Edinburgh	19821569 (E)	EU373269	EU373303
Impatiens viscida Wight	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EU373270	—
Impatiens xanthina Comber	China, Yunnan	Yuan CN2k1-15 (NEU)	EU373271	—
Impatiens yingjiangensis S. Akiyama & H. Ohba	China, Yunnan	Yuan CN2k1-55 (NEU)	EU373272	—

Sequencing reactions were performed on an ABI310 capillary sequencer (Applied Biosystems). All accessions surveyed were homozygous in length for both paralogues, thus displaying no sequence polymorphism.

Phylogenetic Analysis

Initial alignment of the DNA sequences was carried out with CLUSTALX and manually adjusted following the

protein alignment using MacClade 4.05 (Maddison and Maddison 2002). We used MrBayes 3.1.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) to conduct Bayesian analysis. The best-fitting substitution model for Bayesian analysis (GTR+G) was selected using a series of likelihood-ratio tests as implemented in Model-Test 3.06 (Posada and Crandall 1998). According to Janssens et al. (2006), *I. omeiana* is sister to all other *Impatiens* species and was, therefore, used as outgroup for all phylogenetic analyses. All sequences were submitted to GenBank (Table 1).

Tests for Selection on Amino Acids

In order to identify positively selected amino acid sites in the C-terminal domain of the *ImpDEF* duplicates, the ratio of nonsynonymous-to-synonymous substitutions (d_N/d_S or ω) at the codon level has been examined. d_N/d_S values for amino acid sites in *ImpDEF* were estimated using the approach of Nielsen and Yang (1998) as implemented in HyPhy (Kosakovsky Pond et al. 2005). This site-specific method, which uses a likelihood-based approach to identify selection, assumes variable selective pressures among sites under the assumption of different classes of sites in the gene with different d_N/d_S (ω) ratios. Both *ImpDEF1* and *ImpDEF2* data sets were investigated by comparing discrete distribution models (M1, M2, and M3) and comparing continuous distribution models (M7 and M8). Likelihood-ratio tests were carried out for the comparison of following evolutionary models: single rate model (M0) and discrete model (M3), neutral (M1) and selection model (M2), and β (M7) and $\beta + \omega$ model (M8).

In order to examine differences in selective pressure along the branches to either duplicate clade, the likelihood of branch-site models (model A and model B [Yang and Nielsen 2002]) was tested with *codeml*, a software program of the PAML package (Yang 1997). By comparing the likelihood of model A (MA) and model B (MB) with the likelihood of model 1a (M1a) and model 3 (M3), respectively, we were able to test for selection on either of the branches leading to *ImpDEF1* and *ImpDEF2*.

Indel Substitution Rates and Selective Pressures

Several complementary approaches were applied to investigate the evolutionary rates and selective pressures in the *Impatiens AP3/DEF* C-terminal domain sequences. First, we tested whether the strong sequence diversity of the *ImpDEF1* C-terminal domain in *Impatiens* shows an increased rate compared to the *AP3/DEF* C-terminal domain sequences of Brassicaceae and Solanaceae. To do this, we used the approach of Podlaha and Zhang (2003), in which the rate of indel mutations was calculated per site

per year. Indels in the C-terminal end of *ImpDEF1* were separately mapped following the parsimony principle onto the *ImpDEF1* C-terminal domain-based phylogeny of *Impatiens* and calculated for each branch of the tree. This tree was chosen because it had resulted in the best-resolved and supported phylogenetic hypothesis for our species sampling.

Subsequently, we tested for possible selection on indels using an approach that is described by Podlaha et al. (2005). This method hypothesizes that the indel mutation rate can be computed per unit of d_S distance and is in contrast to the first method not affected by inaccurate time estimates. Using this approach, the possibility of selection on indel substitutions is examined by separately comparing the indel substitution rate of the C-termini of *ImpDEF1* and *ImpDEF2* with the indel mutation rate of *ImpDEF* introns 4 and 5.

Results

The C-Terminal Domain of *ImpDEF2* Is Conserved in Length and Sequence, Whereas the C-Terminal Domain of *ImpDEF1* Shows Strong Length Variation

The cloned *AP3/DEF* sequences from *I. hawkeri* (*IhDEF1* and *IhDEF2*) are remarkably different, with *IhDEF2* terminated by a premature stop codon, positioned immediately before the normally expected *PI*-derived motif. To confirm that the sequences did not represent experimental artifacts, the partial cDNA sequences were checked by sequencing the corresponding genomic locus (Geuten et al. 2006). Our data show that the truncation in the *ImpDEF2* C-terminus is coded by the genome of almost all *Impatiens* species examined, suggesting that this derived protein does have a specific function throughout the genus. We present an alignment of the protein sequence of the C-terminus of *ImpDEF1* and *ImpDEF2* in Figs. 1 and 2, respectively. The length of these C-terminal ends was calculated following the initial mapping of the protein regions in *AP3* (Krizek and Meyerowitz 1996). The average length of the *ImpDEF2* C-terminus is 20.95 ± 0.25 codons, showing almost no length variation. Furthermore, within the C-terminal domain of *ImpDEF2*, only two independent indel substitutions were observed. One mutation was found in *I. flaccida*, while another was located in *I. edgeworthii*. Both species are not closely related, as they belong to two different clades in the *Impatiens* phylogeny (Janssens et al. 2006).

In contrast to *ImpDEF2*, *ImpDEF1* contains both characteristic *PI*-derived and *euAP3*-motifs and is characterized by a considerable length divergence throughout the whole genus (Fig. 1). To make PCR amplification of the

C-terminal domain in *ImpDEF1* possible using a single primer pair, we designed the reverse primer in the *euAP3*-motif. As a result, we only used the sequenced segment of the *ImpDEF1* C-terminal domain, which is the equivalent to the complete C-terminal domain minus the length of the *euAP3*-motif (27 nucleotides [nt] or 9 amino acids [aa]). Hence, the average length of the C-terminal domain (minus the *euAP3*-motif) in *ImpDEF1* is 93.95 ± 5.3 aa, with a maximum length of 101 codons for *I. xanthina* and a minimum length of 74 codons for *I. omeiana*. Only the sequenced segment of the *ImpDEF1* C-terminal ends has been used for further analyses in our study. Indel substitutions that occurred within the C-termini of *ImpDEF1* were mapped onto the *ImpDEF1* C-terminal domain phylogeny. In total, 118 indel substitutions were found following the parsimony principle (Fig. 3).

The Origin of *ImpDEF2* and *ImpDEF1* Can be Explained by Inferring a Frameshift Mutation

To infer the mutational event(s) that resulted in the divergent C-terminal sequences, we compared the C-terminus of *ImpDEF1* of *I. omeiana* with every available *ImpDEF2* C-terminal domain present in our dataset. Unfortunately, we were unable to sequence the *ImpDEF2* C-terminus for *I. omeiana*, despite the fact that the *Imp-DEF2* K-domain had already been successfully sequenced for this species (Janssens et al. 2007). Therefore we could not compare both C-terminal ends in *I. omeiana*, which is considered to be sister to all other species of *Impatiens* (Janssens et al. 2006).

Comparing the two *ImpDEF* paralogues resulted in a simple explanation in which a deletion of 11 nt gave rise to the truncated C-terminal end in *ImpDEF2*. This deletion caused a shift in the original reading frame which resulted in a premature stop codon, located 10 aa before the *PI*-derived motif (Fig. 4).

Figure 4 shows the alignment of both *ImpDEF* C-terminal domains with the *AP3/DEF* C-termini of *Hydrangea macrophylla* and *Marcgravia umbellata*. It is interesting to note that the differences in nucleotides between the shifted *ImpDEF2* region and the paralogous *ImpDEF1* region resulted from substitutions that occurred most likely before the frameshift event (Fig. 4). This alignment illustrates that three of five nucleotide substitutions were present as silent mutations in the original amino acid order. Apparently, two
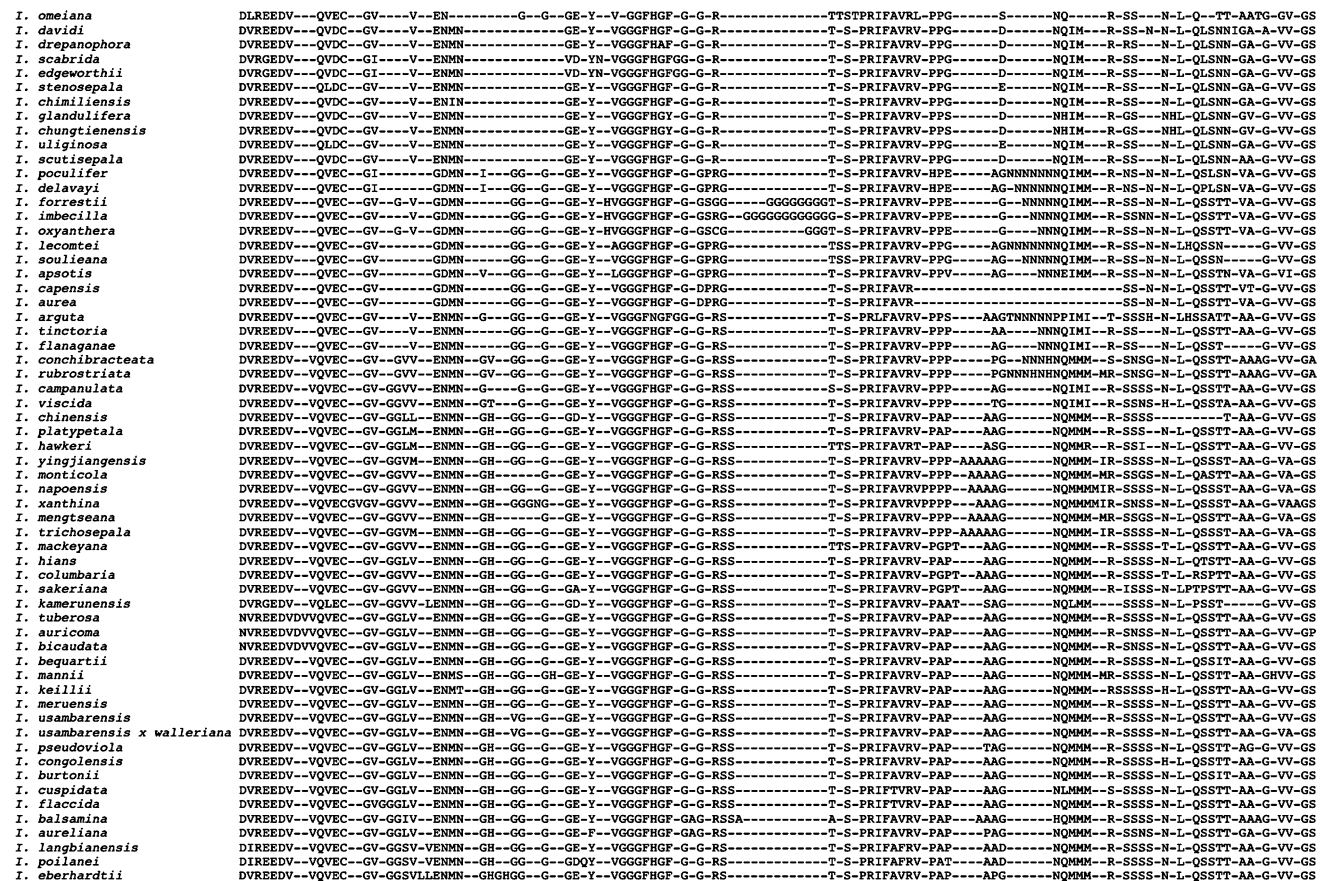
```
I. omeiana               DLREEDV---QVEC--GV----V--EN---------G--G--GE-Y--V-GGFHGF-G-G-R---------------TTSTPRIFAVRL-PPG------S------NQ-----R-SS---N-L-Q--TT-AATG-GV-GS
I. davidi                DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------D------NQIM---R-SS-N-N-L-QLSNNIGA-A-VV-GS
I. drepanophora          DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------D------NQIM---R-RS----N-L-QLSNN-GA-G-VV-GS
I. scabrida              DVRGEDV---QVDC--GI----V--ENMN------------VD-YN-VGGGFHGFGG-G-R-------------T-S-PRIFAVRV-PPG------D------NQIM---R-SS----N-L-QLSNN-GA-G-VV-GS
I. edgeworthii           DVRGEDV---QVDC--GI----V--ENMN------------VD-YN-VGGGFHGFGFG-G-R------------T-S-PRIFAVRV-PPG------D------NQIM---R-SS----N-L-QLSNN-GA-G-VV-GS
I. stenosepala           DVREEDV---QLDC--GV----V--ENMN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------E------NQIM---R-SS----N-L-QLSNN-GA-G-VV-GS
I. chimiliensis          DVREEDV---QVDC--GV----V--ENIN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------D------NQIM---R-SS----N-L-QLSNN-GA-G-VV-GS
I. glandulifera          DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGY-G-G-R-------------T-S-PRIFAVRV-PPS------D------NHIM---R-GS----NHL-QLSNN-GV-G-VV-GS
I. chungtienensis        DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGY-G-G-R-------------T-S-PRIFAVRV-PPS------D------NHIM---R-GS----NHL-QLSNN-GV-G-VV-GS
I. uliginosa             DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------E------NQIM---R-SS----N-L-QLSNN-AA-G-VV-GS
I. scutisepala           DVREEDV---QVDC--GV----V--ENMN------------GE-Y--VGGGFHGF-G-G-R-------------T-S-PRIFAVRV-PPG------D------NQIM---R-SS---N-L-QLSNN-AA-G-VV-GS
I. poculifer             DVREEDV---QVEC--GI-------GDMN--I---GG--G--GE-Y--VGGGFHGF-G-GPRG-----------T-S-PRIFAVRV-HPE-----AGNNNNNNNQIMM--R-NS-N-N-L-QSLSN-VA-G-VV-GS
I. delavayi              DVREEDV---QVEC--GI-------GDMN--I---GG--G--GE-Y--VGGGFHGF-G-GPRG-----------T-S-PRIFAVRV-HPE-----AG-NNNNNNQIMM--R-NS-N-N-L-QPLSN-VA-G-VV-GS
I. forrestii             DVREEDV---QVEC--GV--G-V--GDMN------GG--G--GE-Y--HVGGGFHGF-G-GSGG------GGGGGGGGT-S-PRIFAVRV-PPE------G---NNNNQIMM--R-SS-N-N-L-QSSTT-VA-G-VV-GS
I. imbecilla             DVREEDV---QVEC--GV----V--GDMN------GG--G--GE-Y--HVGGGFHGF-G-GSRG--GGGGGGGGGGGGG-S-PRIFAVRV-PPE------G---NNNNQIMM--R-SSNN-N-L-QSSTT-VA-G-VV-GS
I. oxyanthera            DVREEDV---QVEC--GV--G-V--GDMN------GG--G--GE-Y--HVGGGFHGF-G-GSCG---------GGGT-S-PRIFAVRV-PPE------G----NNNQIMM--R-SS-N-N-L-QSSTT-VA-G-VV-GS
I. lecomtei              DVREEDV---QVEC--------GDMN------GG--G--GE-Y--AGGGFHGF-G-GPRG-----------TSS-PRIFAVRV-PPG-----AGNNNNNNNQIMM--R-SS-N-N-LHQSSN-----G-VV-GS
I. soulieana             DVREEDV---QVEC--------GDMN------GG--G--GE-Y--VGGGFHGF-G-GPRG-----------TSS-PRIFAVRV-PPG------AG--NNNNNQIMM--R-SS-N-N-L-QSSN------G-VV-GS
I. apsotis               DVREEDV---QVEC--------GDMN--V---GG--G--GE-Y--LGGGFHGF-G-GPRG-----------T-S-PRIFAVRV-PPV-----AG-----NNNEIMM--R-SS-N-N-L-QSSTN-VA-G-VI-GS
I. capensis              DVREEDV---QVEC--------GDMN------GG--G--GE-Y--VGGGFHGF-G-DPRG-----------T-S-PRIFAVR----------------------SS-N-N-L-QSST-VT-G-VV-GS
I. aurea                 DVREEDV---QVEC--GV------GDMN------GG--G--GE-Y--VGGGFHGF-G-DPRG---------T-S-PRIFAVR--------------------------SS-N-N-L-QSST-VA-G-VV-GS
I. arguta                DVREEDV---QVEC--GV----V--GDMN--G---GG--G--GE-Y--VGGGFNGFGG-G-RS----------T-S-PRLFAVRV-PPS-----AAGTNNNNNPPIMI--T-SSSH-N-LHSSATT-AA-G-VV-GS
I. tinctoria             DVREEDV---QVEC--GV----V--ENMN------GG--G--GE-Y--VGGGFHGF-G-G-RS-----------T-S-PRIFAVRV-PPP-----AA----NNNQIMI--R-SS----N-L-QSSTT-AA-G-VV-GS
I. flanaganae            DVREEDV---QVEC--GV----V--ENMN------GG--G--GE-Y--VGGGFHGF-G-G-RS----------T-S-PRIFAVRV-PPP------AG----NNNQIMI--R-SS----N-L-QSST-----G-VV-GS
I. conchibracteata       DVREEDV---VQVEC--GV--GVV--ENMN--GV--GG--G--GE-Y--VGGGFHGF-G-G-RSS---------T-S-PRIFAVRV-PPP-----PG--NNNHHQMMM--S-SNSG-N-L-QSSTT-AAAG-VV-GA
I. rubrostriata          DVREEDV---VQVEC--GV--GVV--ENMN--GV--GG--G--GE-Y--VGGGFHGF-G-G-RSS--------T-S-PRIFAVRV-PPP-----PGNNNHHHNQMMM-MR-SNSG-N-L-QSSTT-AAAG-VV-GA
I. campanulata           DVREEDV---VQVEC--GV--GGVV--ENMN--G-----G--G--GE-Y--VGGGFHGF-G-G-RSS-------S-S-PRIFAVRV-PPP-----AG----NQIMI--R-SSNS-N-L-QSSTT-AA-G-VV-GS
I. viscida               DVREEDV---VQVEC--GV--GGLV--ENMN--GT----G--G--GE-Y--VGGGFHGF-G-G-RSS------T-S-PRIFAVRV-PPP-----TG------NQIMI--R-SNS-H-L-QSSTA-AA-G-VV-GS
I. chinensis             DVREEDV---VQVEC--GV-GGLL--ENMN--GH--GG--G--GD-Y--VGGGFHGF-G-G-RSS--------T-S-PRIFAVRV-PAP-----AG-------NQMMM--R-SSSS---------T-AA-G-VV-GS
I. platypetala           DVREEDV---VQVEC--GV-GGLM--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AG-------NQMMM--R-SSS--N-L-QSSTT-AA-G-VV-GS
I. hawkeri               DVREEDV---VQVEC--GV-GGLM--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------TTS-PRIFAVRT-PAP-----ASG------NQMMR--R-SSI--N-L-QSSTT-AA-G-VV-GS
I. yingjiangensis        DVREEDV---VQVEC--GV-GGVM--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PPP-AAAAAG------NQMMM-IR-SSSS-N-L-QSSST-AA-G-VA-GS
I. monticola             DVREEDV---VQVEC--GV-GGVV--ENMN--GH------G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PPP--AAAAAG------NQMMM-MR-SSGS-N-L-QASTT-AA-G-VA-GS
I. napoensis             DVREEDV---VQVEC--GV-GGVV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRVPPPP--AAAAG------NQMMMMIR-SNS-N-L-QSSST-AA-G-VV-GS
I. xanthina              DVREEDV---VQVECGVGV-GGVV--ENMN--GH--GGGNG--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRVPPPP---AAAG------NQMMMMIR-SNSS-N-L-QSSST-AA-G-VAAGS
I. mengtseana            DVREEDV---VQVEC--GV-GGVV--ENMN--GH------G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PPP--AAAAG------NQMMM-MR-SSGS-N-L-QSSST-AA-G-VA-GS
I. trichosepala          DVREEDV---VQVEC--GV-GGVM--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PPP-AAAAAG------NQMMM-IR-SSSS-N-L-QSSST-AA-G-VA-GS
I. mackeyana             DVREEDV---VQVEC--GV-GGVV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------TTS-PRIFAVRV-PGPT---AAG------NQMMM--R-SSSS-N-L-QTSTT-AA-G-VV-GS
I. hians                 DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QTSTT-AA-G-VV-GS
I. columbaria            DVREEDV---VQVEC--GV-GGVV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PGPT---AAAG-----NQMMM--R-SSSS-T-L-RSPTT-AA-G-VV-GS
I. sakeriana             DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GA-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PGPT----AAG------NQMMM--R-ISSS-N-LPTPSTT-AA-G-VV-GS
I. kamerunensis          DVRGEDV---VQLEC--GV-GGVV--LENMN--GH--GG--G--GD-Y--VGGGFHGF-G-G-RSS------T-S-PRIFAVRV-PAAT----SAG-----NQLMM------SSS-N-L-PSST-----G-VV-GS
I. tuberosa              NVREEDVDVVQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. auricoma              NVREEDVDVVQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SNSS-N-L-QSSTT-AA-G-VV-GP
I. bicaudata             NVREEDVDVVQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SNSS-N-L-QSSTT-AA-G-VV-GS
I. bequartii             DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSIT-AA-G-VV-GS
I. mannii                DVREEDV---VQVEC--GV-GGLV--ENMS--GH--GG--GH-GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSST-AA-GHVV-GS
I. keillii               DVREEDV---VQVEC--GV-GGLV--ENMT--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--RSSSSS-H-L-QSSTT-AA-G-VV-GS
I. meruensis             DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. usambarensis          DVREEDV---VQVEC--GV-GGLV--ENMN--GH--VG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. usambarensis x walleriana DVREEDV---VQVEC--GV-GGLV--ENMN--GH--VG--G--GE-Y--VGGGFHGF-G-G-RSS----T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VA-GS
I. pseudoviola           DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP----TAG------NQMMM--R-SSSS-N-L-QSSTT-AG-G-VV-GS
I. congolensis           DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-H-L-QSSTT-AA-G-VV-GS
I. burtonii              DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSIT-AA-G-VV-GS
I. cuspidata             DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS-------T-S-PRIFTVRV-PAP-----AAG------NQMMM--S-SSSS-N-L-QSSTT-AA-G-VV-GS
I. flaccida              DVREEDV---VQVEC--GVGGGLV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RSS------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. balsamina             DVREEDV---VQVEC--GV-GGIV--ENMN--GH--GG--G--GE-Y--VGGGFHGF-GAG-RSSA---------A-S-PRIFAVRV-PAP----AAAG-----NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. aureliana             DVREEDV---VQVEC--GV-GGLV--ENMN--GH--GG--G--GE-F--VGGGFHGF-GAG-RS--------T-S-PRIFAVRV-PAP-----PAG------NQMMM--R-SSNS-N-L-QSST-GA-G-VV-GS
I. langbianensis         DIREEDV---VQVEC--GV-GGSV-VENMN--GH--GG--G--GE-Y--VGGGFHGF-G-G-RS--------T-S-PRIFAVRV-PAP-----AAG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. poilanei              DIREEDV---VQVEC--GV-GGSV-VENMN--GH--GG--G--GDQY--VGGGFHGF-G-G-RS--------T-S-PRIFAFRV-PAT----AAD------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
I. eberhardtii           DVREEDV---VQVEC--GV-GGSVLLENMN--GHGHGG--G--GE-Y--VGGGFHGF-G-G-RS--------T-S-PRIFAVRV-PAP----APG------NQMMM--R-SSSS-N-L-QSSTT-AA-G-VV-GS
```

Fig. 1 Amino acid sequence alignment of the C-terminal domain of *ImpDEF1* from 61 *Impatiens* species. Dots represent amino acids identical to the first sequence, while dashes represent aligned gaps

```
I. arguta                      GFREEEEAQYGLVXNEGGWL
I. aureliana                   GFREEEEAHYGLAENDGGWL
I. auricoma                    GFREEEAAHYGLAENEGGWL
I. balsamina                   GFREEEEAHYGLAENEGGWL
I. bequartii                   GFREEEVAHYGLAENEGGWL
I. bicaudata                   GFREEEEAHYGLAENEGGWL
I. burtonii                    GFREEEEAHYGLAENEGGWL
I. campanulata                 GFREEEEAQYGLAENEGGWL
I. chinensis                   GFREEEETQYGLVENEGGWL
I. chungtienensis              GFREEEEAQYGLVENEGGWL
I. columbaria                  GFREEEEAQYGLAENEGGWL
I. conchibracteata             GFREEEEAQYGLAENEGGWL
I. congolensis                 GFREEEEAHYGLAENEGGWL
I. cuspidata                   GFREEEEAHYGLAENEGGWL
I. eberhardtii                 GFREEEEAHYGLAENEGGWL
I. edgeworthii                 DFRGEE-AQYGLVENEGGWL
I. engleri                     GFREEEEPHYGLAENEGGWL
I. flaccida                    GFREEE-AHYGLAENEGGWL
I. hawkeri                     GFREEEETQYALAENEGGWL
I. hochstetteri                GFREEEEAHYGLAENEGGWL
I. kamerunensis                GFREEEEAQYGLPENEGGWL
I. langbianensis               GFREEEEAHYGLAENEGGWL
I. mackeyana                   GFREEEEAQYGLAENEGGWL
I. mannii                      GFREEEEAHYGLAENEGGWL
I. meruensis                   GFREEEEANYGLAENEGGWL
I. poculifer                   GFREEEEAQYGLADNEGGWL
I. sakeriana                   GFREEEEAQYGLAENEGGWL
I. tuberosa                    GFREEEEAHYGLAENEGGWL
I. uliginosa                   GFREEEEAQYGLVENEGGWL
I. usambarensis                GFREEEEAHYGLAENEGGWL
I. usambarensis x walleriana   GFREEEEAHYGLAENEGGWL
```

Fig. 2 Amino acid sequence alignment of the truncated C-terminal end of *ImpDEF2* from 31 *Impatiens* species. Dots represent amino acids that are identical to the first sequence, while dashes represent aligned gaps

of the three substitutions were nonsynonymous mutations in *ImpDEF2* until their reading frame was moved two positions to the left and they became synonymous. Based on the comparison of the *Impatiens* paralogues with *M. umbellata* and *H. macrophylla*, it became clear that one of the three silent substitutions occurred in *ImpDEF1* and not in *ImpDEF2*. Furthermore, one substitution is present in the second position of the original amino acid arrangement, suggesting that the nucleotide at this position must have changed after the deletion event. The shifted fragment also contained one substitution in the first position of the original codon order. However, when we compared this codon with the same amino acids in *M. umbellata* and *H. macrophylla*, we noticed that this nucleotide alteration occurred in *ImpDEF1* instead of *ImpDEF2*.

Codon Mutations in the C-Terminal Domain of *ImpDEF1* and *ImpDEF2* Evolve Under Purifying Selective Pressure

In order to examine the molecular evolution of the C-terminal region, we investigated which selective pressures act on nonsynonymous substitutions in each of the two paralogues. The comparison of different models of evolution that allow for relative fixation rates of synonymous (silent) and nonsynonymous (amino acid-altering)
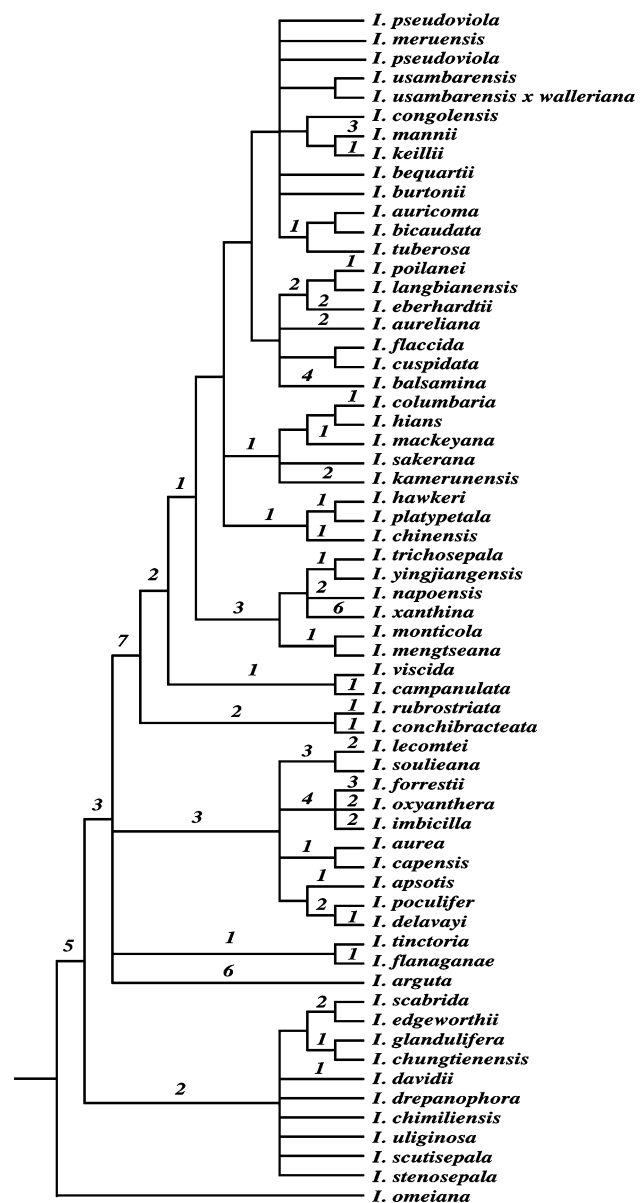
Fig. 3 Bayesian consensus cladogram based on the C-terminal domain sequences of *ImpDEF1*. The numbers on the branches indicate the number of indel substitutions in the C-terminal end of *ImpDEF1* based on the alignment shown in Fig. 1

mutations provided similar results for both *ImpDEF1* and *ImpDEF2* (Tables 2 and 3, respectively). For both paralogues, the estimated ω-ratios were significantly lower than 1.0, suggesting that purifying selection has taken place instead of positively selected evolution.

In addition, the branch-site models (models A and B) specifying the branches toward *ImpDEF1* and *ImpDEF2* did not explain the data significantly better than the models they were compared with (MA vs. M1a, $2\Delta_L = 0.02$, $p = 0.41$; MB vs. M3, $2\Delta_L = 0.04$, $p = 0.63$), suggesting that selection on the branches did not change subsequent to duplication.
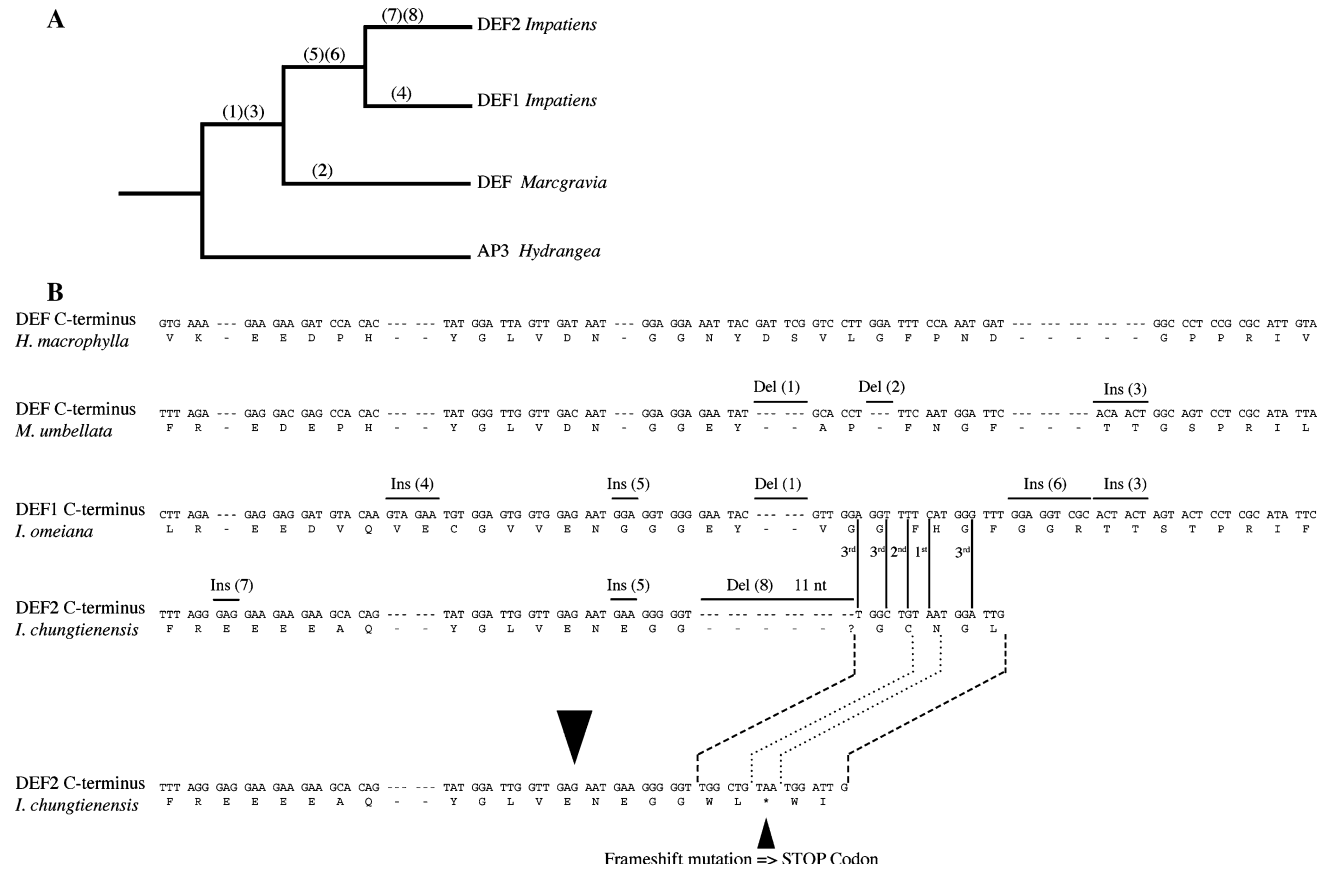
Fig. 4 A Phylogeny of the basal asterid *AP3/DEF*-like proteins including *Hydrangea*, *Marcgravia*, and *Impatiens* (Geuten et al. 2006). Insertions (Ins) and deletions (Del) that occurred in the C-terminal ends of *AP3/DEF* in *Marcgravia* and/or in one of the two *AP3/DEF* paralogues of *Impatiens* are indicated with a number in parentheses (see B) and are highlighted on the basal asterid *AP3/DEF* phylogeny. **B** Alignment of the *AP3/DEF*-like C-terminal domains of *Hydrangea macrophylla*, *Marcgravia umbellata*, *Impatiens omeiana* (ImpDEF1), and *I. chungtienensis* (ImpDEF2). The deletion causing the frameshift mutation is indicated on the initial *ImpDEF2* sequence. The result of the deletion is shown in the frameshifted *ImpDEF2* sequence. Differences between the frameshifted fragment in *Imp-DEF2* and the similar fragment in *ImpDEF1* are highlighted with a line between both fragments

Table 2 Sites under positive selection, parameter estimates, likelihood values, and ω-values for the *ImpDEF1* C-terminal domain as estimated for several discrete distribution and continuous distribution models

ImpDEF1 C-terminus

Model	P	L	ω-value	Proportion of sites	Positively selected sites
M0 (single ratio)	1	−2719.22	$\omega = 0.2685$		None
M1 (neutral)	1	−2760.51	$\omega 0 = 0.0000$	$p0 = 0.2311$	Not allowed
			$\omega 1 = 1.0000$	$p1 = 0.7688$	
M2 (selection)	3	−2760.51	$\omega 0 = 0.0000$	$p0 = 0.2306$	None
			$\omega 1 = 1.0000$	$p1 = 0.7693$	
			$\omega 2 = 3.046$	$p2 = 0.0000$	
M3 (discrete)	5	−2703.54	$\omega 0 = 0.0708$	$p0 = 0.4159$	None
			$\omega 1 = 0.2501$	$p1 = 0.0000$	
			$\omega 2 = 0.4437$	$p2 = 0.5841$	
M7 (β)	2	−2703.89			Not allowed
M8 (β and ω)	4	−2703.89	$\omega 0 = 0.2819$	$p0 = 0.9999$	None
			$\omega 1 = 1.0000$	$p1 = 0.0001$	

Table 3 Sites under positive selection, parameter estimates, likelihood values, and ω-values for the *ImpDEF2* C-terminal domain as estimated for several discrete distribution and continuous distribution models

ImpDEF2 C-terminus

Model	P	L	ω-value	Proportion of sites	Positively selected sites
M0 (single ratio)	1	−261.07	$\omega = 0.3224$		None
M1 (neutral)	1	−260.53	$\omega0 = 0.0000$	$p0 = 0.3749$	Not allowed
			$\omega1 = 1.0000$	$p1 = 0.6251$	
M2 (selection)	3	−260.53	$\omega0 = 0.0000$	$p0 = 0.3736$	None
			$\omega1 = 1.0000$	$p1 = 0.6263$	
			$\omega2 = 1.7159$	$p2 = 0.0000$	
M3 (discrete)	5	−258.77	$\omega0 = 0.0000$	$p0 = 0.3297$	None
			$\omega1 = 0.2500$	$p1 = 0.0000$	
			$\omega2 = 0.5137$	$p2 = 0.6703$	
M7 (β)	2	−259.45			Not allowed
M8 (β and ω)	4	−259.45	$\omega0 = 0.3288$	$p0 = 0.9999$	None
			$\omega1 = 1.0000$	$p1 = 0.0001$	

Accelerated 3*n*-Indel Substitution Rate in the *ImpDEF1* C-Terminal Domain and Low 3*n*-Indel Substitution Rate in the *ImpDEF2* C-Terminal Domain

Despite the evidence for negatively selected codons within the relatively long C-terminal domains of *ImpDEF1*, we noticed that there is a remarkable length divergence between C-termini of paralogues from different species. Therefore we were interested to know (a) if this is a remarkable observation in comparison to other taxonomic groups and (b) which selective pressures act on these length mutations in *ImpDEF1*. In addition, we tested whether the truncated C-termini of *ImpDEF2* might also have evolved under selective pressure.

For comparison to other species, Brassicaceae and Solanaceae were selected. For these core eudicot plant families, it is reasonable to assume that the *euAP3* lineage is represented by one member in the genome. For Solanaceae, we used the data on MADS-box class B genes presented by Rijpkema et al. (2006). This dataset contains eight species: *Brunfelsia uniflora, Cestrum elegans, Juanulloa aurantiaca, Mandragora autumnalis, Petunia hybrida, Solandra maxima, Solanum lycopersicum*, and *Solanum pseudolulo*. Indel substitution events that occurred within the C-terminal domain of this group were mapped onto the Solanaceae tree of Olmstead et al. (1999). In this phylogeny, the Cestroideae containing *Cestrum elegans* are sister to the Solanaceae species that were present in the dataset of Rijpkema et al. (2006). Consequently, we separately compared *Cestrum elegans* with each species from the Solanaceae dataset and counted the number of indel substitutions that had occurred between the outgroup and the examined species. On average, three

indel substitutions were found for a common C-terminus in Solanaceae with a length of 183.9 nt. Moreover, insertions or deletions that occurred within the Solanaceae C-terminal domain are always in multiples of three nucleotides.

According to Wikström et al. (2001) the crown group in Solanaceae, in which the Cestroideae is one of the most basal lineages, has been dated at an age of 41 to 36 million years (± 38.5 my on average). The method of Podlaha and Zhang (2003) was used to calculate the rate of indel substitution in Solanaceae. Consequently, the substitution rate for 3*n*-indels in the C-termini of Solanaceae is $3/183.86/(38.5 \times 10^7 \times 2) = (2.11 \pm 0.15) \times 10^{-10}$ per site per year (Table 4).

In order to obtain a second estimate of the average substitution rate for insertions and deletions in the C-terminal domain of core eudicots, we used the same approach for a distinct dataset. We chose to compare the C-terminal domain of *Arabidopsis thaliana* with that of *Brassica napus* (Jack et al. 1992; Pylatuik et al. 2003). Recent studies in Brassicaceae dated the most recent common ancestor of *Arabidopsis* and *Brassica* between 20.4 and 14.5 million years ago (Bowers et al. 2003; Schranz and Mitchell-Olds 2006). By comparing the C-termini of both species, we identified only one 3*n*-indel over an alignment of 189 nt. As a result, the neutral rate of indel substitution for these data is $1/189/(17.45 \times 10^7 \times 2) = (1.52 \pm 0.2) \times 10^{-10}$ per site per year (Table 4).

The estimated substitution rates can now be used to calculate the expected number of indels in the C-termini of *ImpDEF1* and *ImpDEF2* considering that the crown group in *Impatiens* originated 25.9 million years ago (Janssens et al., submitted for publication). The observed number of 3*n*-indels in the C-terminal domain of *ImpDEF1* is 15.99,

Table 4 Comparison of substitution rates of 3*n*-indels for *ImpDEF1* and *ImpDEF2*

Indel rate (per site per 10^{10} yr) of C-terminal domain in *AP3/DEF*-like genes in			
Solanaceae	*Arabidopsis-Brassica*	*ImpDEF1*	*ImpDEF2*
2.11	1.52	12.30	0.18
No of indels in C-terminus of *ImpDEF1*			
Expected from Solanaceae data	Expected from *Arabidopsis-Brassica* data	Observed in *ImpDEF1*	
3.21	2.15	15.99	
No of indels in C-terminus of *ImpDEF2*			
Expected from Solanaceae data	Expected from *Arabidopsis-Brassica* data	Observed in *ImpDEF2*	
0.66	0.47	0.06	

which is significantly higher than the indel substitution rate for Solanaceae ($p < 0.001$; Poisson test). In contrast, the observed number of indels in the C-terminal region of *ImpDEF2* is 0.06, which is significantly lower than the indel substitution rate for Solanaceae ($p < 0.001$). Similar results were obtained using the estimated indel substitution rate based on the comparison between *Arabidopsis* and *Brassica*. With this substitution rate, we expected on average 2.15 indels in the C-terminal end of *ImpDEF1* and 0.47 indel in the C-terminus of *ImpDEF2* (Table 4). The significant difference between the observed and the expected amount of indels is even higher for both *Impatiens* paralogues when using the *Arabidopsis*/*Brassica*-based substitution rate ($p < 0.001$; Poisson test). These results indicate that indels in the C-terminal domain of *ImpDEF1* evolve at a significantly faster rate. In contrast, indel substitution events in *ImpDEF2* evolve at a significantly slower rate.

Because of the observed rate differences, we wanted to test if they could be explained by positive and/or purifying selection on indel substitutions. Similarly to Podlaha et al. (2005), we used the general indel mutation rate in introns to evolve under neutral selection. Following this approach, we calculated the number of 3*n*-indel substitutions that occurred within intron 4 and intron 5 of both *ImpDEF1* and *ImpDEF2* (Janssens et al. 2007). An average of 14.72 3*n*-indels/kb was found for the two introns in both paralogues. Comparing this value with the average number of 3*n*-indels/kb in the C-terminal domain of *ImpDEF1* (54.47 3*n*-indels/kb) and *ImpDEF2* (0.91 3*n*-indel/kb), we noticed a significant difference between the neutral value and the calculated value for each paralogue (Poisson test: *ImpDEF1*, $p < 0.001$; *ImpDEF2*, $p < 0.001$). This result

suggests that indel substitutions in the C-terminal domain of *ImpDEF1* genes evolve under positive selection, while these mutations are strongly selected against in the *ImpDEF2* gene.

Discussion

Frameshift mutations, which often result in the early termination of the transcript by coding a premature stop codon, are generally regarded as detrimental for protein function and have therefore been considered of little evolutionary significance. More recently, however, the combination of gene duplications and frameshift mutations is assumed to be an important evolutionary mechanism for the emergence of new biological functions (Kopelman et al. 2005; Raes and Van de Peer 2005; Taylor and Raes 2004). Vandenbussche et al. (2003) identified the importance of frameshifts in flowering plant MADS-box gene-family evolution. Recently, Kramer et al. (2006) raised the question of how often these frameshift mutations subsequent to duplication would have occurred within the *AP3/DEF* lineage. These authors identified three other frameshifts in the *AP3/DEF* lineage. Although we describe an additional frameshift mutation in the *AP3/DEF* lineage, the case of *Impatiens* is of specific interest, as it involves a recent duplication event that probably coincides with the origin of the genus.

By comparing the C-terminal domains of *ImpDEF1* with those of *ImpDEF2*, not only did we observe a conserved dissimilarity in length between the two duplicates, but also we noticed a significant difference in the molecular evolution of both C-termini. As for *ImpDEF1*, the C-terminal domain is characterized by numerous indel substitutions. The fact that no out-of-frame indels were detected within the *ImpDEF1* gene lineage suggests that this C-terminus is still a functional domain. The indel substitution rate of the C-terminal end is significantly higher than that of the *AP3/DEF* C-terminus in Brassicaceae or Solanaceae. Furthermore, by comparison to what can be considered a "neutral rate" of 3*n* substitutions in the K-domain introns, our results reveal an adaptive selective pressure on the length of the *ImpDEF1* C-terminus. Consequently we can consider the functional mechanism responsible for this mode of evolution. Knowing that the K-domain sequence of both paralogues is very similar (Janssens et al. 2007), we assume that the protein interaction specificity of each gene copy is probably derived from their different C-terminal domains. Some clues come from the yeast-two-hybrid data presented by Geuten et al. (2006). According to their results IhDEF1 and IhDEF2 neither form homodimers nor heterodimerize with each other in vitro. Both paralogues, however, dimerize in vitro with IhGLO. In addition,

IhDEF1 does not interact, while IhDEF2 does interact with IhSEP3. This means that the *ImpDEF1* C-terminal domain can be involved in establishing a specific configuration of the higher order transcriptional regulatory complex, possibly by recruiting other specific factors to the complex. So far, no earlier studies have suggested that the C-terminus would be involved in such an interaction, yet it would be interesting to test for such a novel function.

In contrast to *ImpDEF1*, the C-terminal domain in *ImpDEF2* is characterized by an indel substitution rate that is significantly lower than estimated for the C-terminus of a standard Brassicaceae or Solanaceae *AP3/DEF*-like gene. The observed indel substitution rate of the C-terminal end in *ImpDEF2* is significantly lower than was expected based on a neutral indel rate, indicating that the length of this domain is under strong purifying selection. Similarly to *ImpDEF1*, though, the truncated C-terminal domain could be involved in allowing protein-protein interactions that influence the configuration of the higher-order complex regulating transcription of *AP3/DEF*-like downstream genes. Likewise, the very conservative evolution of the *ImpDEF2* paralogue is probably related to a new function. Together, their retention throughout the genus, their mode of evolution, and the molecular biological evidence suggest that the gene products of both *ImpDEF1* and *ImpDEF2* act in functional transcriptional regulatory complexes. Recently, Piwarzyk et al. (2007) demonstrated that both euAP3- and PI-derived motifs are not required for the *AP3/DEF* floral organ identity function, and thus it is possible that both *AP3/DEF* paralogues in *Impatiens* are still functional in floral identity specification. The specific role of these paralogues and the complexes they form remains to be determined, however.

The several subfamilies of MADS-box genes have characteristic C-terminal domains. Establishing these diverse C-terminal domains is proposed to have happened by frameshift mutations for several of these subfamilies (Vandenbussche et al. 2003). Here we have demonstrated that positive and negative selection on length mutations can explain the diverse sequences of the *Impatiens AP3/DEF*-like genes. In doing so, we identified a novel selective pressure for MADS-box gene C-terminal domains. *Impatiens* might be an exception, but could as well be an example of a more general phenomenon. The retention of frameshift mutations in MADS-box gene evolution, resulting in length difference in addition to a different sequence, seems to emphasize the importance of length for function. Functionally interpreting adaptive evolution by length mutations would require a better understanding of the specific function of the C-terminal domains of MADS-box genes. Conversely, taking length into account as a variable could help to better understand C-terminal domain function.

References

Aagaard JE, Olmstead RG, Willis JH, Phillips PC (2005) Duplication of floral regulatory genes in the Lamiales. Am J Bot 92:1284–1293

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Ribas de Pouplana L, Martinez-Castilla L, Yanofsky MF (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci USA 97:5328–5333

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

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

Egea-Cortines M, Saedler H, Sommer H (1999) Ternary complex formation between the MADS-box proteins SQUAMOSA, DEF-ICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus. EMBO J 18:5370–5379

Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwait J (1999) Preservation of duplicate genes by complementary degenerate mutations. Genetics 151:1531–1545

Geuten K, Becker A, Kaufmann K, Caris P, Janssens S, Viaene T, Theissen G, Smets E (2006) Petaloidy and petal identity MADS-box genes in the balsaminoid genera Impatiens and Marcgravia. Plant J 47:501–518

Goto K, Meyerowitz EM (1994) Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. Genes Dev 8:1548–1560

Hileman LC, Sundstrom JF, Litt A, Chen M, Shumba T, Irish VF (2006) Molecular and phylogenetic analyses of the MADS-Box gene family in Tomato. Mol Biol Evol 23:2245–2258

Honma T, Goto K (2001) Complexes of MADS-box proteins are insufficient to convert leaves into floral organs. Nature 409:525–529

Huelsenbeck J, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Jack T, Brockman L, Meyerowitz E (1992) The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. Cell 68:683–697

Jack T, Fox GL, Meyerowitz EM (1994) Arabidopsis homeotic gene APETALA3 ectopic expression: transcriptional and

posttranscriptional regulation determine floral organ identity. Cell 76:703–716

Janssens S, Geuten K, Yuan Y-M, Song Y, Küpfer P, Smets E (2006) Phylogenetics of *Impatiens* and *Hydrocera* using chloroplast *atpB-rbcL* spacer sequences. Syst Bot 31:171–180

Janssens S, Geuten K, Viaene T, Yuan Y-M, Song Y, Smets E (2007) Phylogenetic utility of the *AP3/DEF* K-domain and its molecular evolution in *Impatiens* (Balsaminaceae). Mol Phylogenet Evol 43:225–239

Kellogg EA (2006) Progress and challenges in studies of the evolution of development. J Exp Bot 57:3505–3516

Kopelman NM, Lancet D, Yanai I (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet 37:588–589

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679

Kramer EM, Irish VF (2000) Evolution of the petal and stamen developmental programs: evidence from comparative studies of the basal angiosperms. Int J Plant Sci 161:S29–S40

Kramer EM, Dorit RL, Irish VF (1998) Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics 149:765–783

Kramer EM, Jaramillo MA, Di Stilio VS (2004) Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. Genetics 166:1011–1023

Kramer EM, Su H-J, Wu C-C, Hu J-H (2006) A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the *APETALA3* gene lineage. BMC Evol Biol 6:30–47

Krizek BA, Meyerowitz EM (1996) Mapping the protein regions responsible for the functional specificities of the *Arabidopsis* MADS domain organ-identity proteins. Proc Natl Acad Sci USA 93:4063–4070

Leseberg CH, Li A, Kang H, Duvall M, Mao L (2006) Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. Gene 378:84–94

Lynch M, Conery J (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M, Force A (2000) The probability of duplicate-gene preservation by subfunctionalization. Genetics 154:459–473

Maddison D, Maddison W (2002) MacClade. Sinauer Associates, Sunderland, MA

Moore RC, Purugganan M (2005) The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 100:15682–15687

Munster T, Pahnke J, Di Rosa A, Kim JT, Martin W, Saedler H, Theissen G (1997) Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. Proc Natl Acad Sci USA 94:2415–2420

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the *HIV*-1 envelope gene. Genetics 148:929–936

Ohno S (1970) Evolution by gene duplication. Springer, New York

Olmstead RG, Sweere JA, Spangler RE, Bohs L, Palmer JD (1999) Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. In: Nee M, Symon DE, Lester RN, Jessop JP (eds) Solanaceae IV: advances in biology and utilization. Royal Botanic Gardens Kew, pp 111–137

Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. Plant Cell 15:1538–1551

Piwarzyk E, Yang Y, Jack T (2007) The conserved C-terminal motifs of the *Arabidopsis* proteins *APETALA3* and *PISTILLATA* are dispensable for floral organ identity function. Plant Physiol 145:1495–1505

Podlaha O, Zhang J (2003) Positive selection on protein-length in the evolution of a primate ion sperm channel. Proc Natl Acad Sci USA 100:122241–12246

Podlaha O, Webb DM, Tucker PK, Zhang J (2005) Positive selection for indel substitutions in the rodent sperm protein catsper1. Mol Biol Evol 22:1845–1852

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818

Preston JC, Kellogg EA (2006) Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL*-like genes in grasses (Poaceae). Genetics 174:421–437

Pylatuik JD, Lindsay DL, Davis AR, Bonham-Smith PC (2003) Isolation and characterization of a *Brassica napus* cDNA corresponding to a B-class floral development gene. J Exp Bot 54:2385–2387

Raes J, Van de Peer Y (2005) Functional divergence of proteins through frameshift mutations. Trends Genet 21:428–431

Riechmann JL, Wang M, Meyerowitz EM (1996) DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA* and *AGAMOUS*. Nucleic Acids Res 24:3134–3141

Rijpkema AS, Royaert S, Zethof J, van der Weerden G, Gerats T, Vandenbussche M (2006) Analysis of the *Petunia TM6 MADS* box gene reveals functional divergence within the *DEF/AP3* lineage. Plant Cell 18:1819–1832

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

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

Shinozuka Y, Kojima S, Shomura A, Ichimura H, Yano M, Yamamoto K, Sasaki T (1999) Isolation and characterization of rice MADS-box gene homologues and their RFLP mapping. DNA Res 6:123–129

Stellari GM, Jaramillo MA, Kramer EM (2004) Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. Mol Biol Evol 21:506–519

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615–643

Theissen G, Becker A, Di Rosa A, Kanno A, Kim JT, Munster T, Winter KU, Saedler H (2000) A short history of MADS-box genes in plants. Plant Mol Biol 42:115–149

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T (2003) Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Res 31:4401–4409

Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc Roy Soc London B 268:2211–2220

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl BioSci 13:555–556

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Zahn LM, Leebens-Mack J, DePamphilis CW, Ma H, Theissen G (2005) To B or not to B a flower: the role of *DEFICIENS* and *GLOBOSA* orthologs in the evolution of the angiosperms. J Hered 96:225–240

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