Available online at www.sciencedirect.com

ScienceDirect

Molecular Phylogenetics and Evolution 43 (2007) 225–239

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Phylogenetic utility of the *AP3/DEF* K-domain and its molecular evolution in *Impatiens* (Balsaminaceae)

Steven Janssens [a,*], Koen Geuten [a], Tom Viaene [a], Yong-Ming Yuan [b], Yi Song [b], Erik Smets [a,c]

[a] *Laboratory of Plant Systematics, Institute of Botany and Microbiology, Kasteelpark Arenberg 31, BE-3001 Leuven, Belgium*
[b] *Institut de Botanique, Université de Neuchâtel, Neuchâtel, Switzerland*
[c] *National Herbarium of the Netherlands, Leiden University Branch, PO Box 9514, NL-2300 RA Leiden, The Netherlands*

Received 24 March 2006; revised 15 November 2006; accepted 16 November 2006
Available online 26 November 2006

Abstract

APETALA3 (*AP3*)/*DEFICIENS* (*DEF*) is a MADS-box transcription factor that is involved in establishing the identity of petal and stamen floral organs. The *AP3/DEF* gene lineage has been extensively examined throughout the angiosperms in order to better understand its role in floral diversity and evolution. As a result, a large number of cloned *AP3/DEF* orthologues are available, which can be used for the design of taxon specific primers for phylogeny reconstruction of close relatives of the group of interest. Following this reasoning, we investigated the phylogenetic utility of the two *AP3/DEF* paralogues (*ImpDEF1* and *ImpDEF2*) that were recently identified in the genus *Impatiens* (Balsaminaceae). K-domain introns 4 and 5 of both *AP3/DEF* duplicates were amplified and sequenced for 59 *Impatiens* species. Phylogenetic analyses of the separated and combined *ImpDEF1* and *ImpDEF2* data sets result in highly congruent topologies with the previously obtained chloroplast *atpB-rbcL* data set. Combination of chloroplast and nuclear matrices results in a well-supported evolutionary hypothesis of *Impatiens*. Our results show that introns 4 and 5 in *AP3/DEF*-like genes are a valuable source of characters for phylogenetic studies at the infrageneric level.

Keywords: *APETALA3/DEFICIENS*; Gene duplication; *Impatiens*; *ImpDEF1*; *ImpDEF2*; K-domain; Phylogenetic utility

1. Introduction

During the last decade several attempts have been undertaken to apply low-copy nuclear genes for phylogenetic reconstruction (Bailey and Doyle, 1999; Bailey et al., 2002; Emschwiller and Doyle, 1999; Fan et al., 2004; Grob et al., 2004; Howarth and Baum, 2002; Lewis and Doyle, 2001; Mason-Gamer et al., 1998; Möller et al., 1999; Oh and Potter, 2003; Sang et al., 1997). Despite these efforts there is still a large demand for low-copy nuclear genes, which could be useful to generate or test existing phylogenetic hypotheses. Low-copy nuclear genes are considered to be useful at all taxonomical levels, but in particular at low systematic level when non-coding chloroplast regions and nuclear ribosomal markers are incapable in generating sufficient resolution. However, due to both practical problems (e.g., the necessity of cloning) and theoretical problems (e.g., paralogy vs. orthology) low-copy nuclear genes still remain undervalued in evolutionary studies.

An interesting group of low-copy nuclear genes, useful for phylogenetic reconstruction can be found in the MADS box gene family (Bailey and Doyle, 1999). Because of general interest in the molecular evolution of the MADS box gene family, many cloning efforts have been carried out on several members of this gene family (Aagaard et al., 2005; Kim et al., 2004; Kramer et al., 1998, 2003, 2006; Kramer and Irish, 2000; Litt and Irish, 2003; Munster et al., 1997; Rijpkema et al., 2006; Theissen et al., 1996;

[*] Corresponding author. Fax: +32 16 32 19 55.
E-mail address: Steven.Janssens@bio.kuleuven.be (S. Janssens).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.11.016

Vandenbussche et al., 2004). For example, by performing a BLAST search on an *AP3/DEF*-like gene of *Impatiens* (cut-off E value 2.00), 112 different hits of similar *AP3/DEF*-like genes were obtained in GenBank (November 2006). The public accessibility of these sequences and the emergence of many EST sequencing projects, reveal the potential of MADS box nuclear genes that can be used for molecular phylogenetics.

In addition, these MADS box gene studies also revealed the history of gene duplication and gene loss for several gene lineages of the MADS box gene family within numerous angiosperm taxa. As a result, the occurrence of duplication and gene loss events during the evolution of some of the MADS box gene lineages (e.g., *AP3/DEF*) is known to some extent. This information could make cases of mistaken paralogy less likely for these low-copy nuclear genes. In theory, paralogy problems tend to be more commonly present at higher taxonomical level and would therefore be less problematic at low taxonomic level (Sang, 2002).

Previously, several studies have focused on the molecular evolution of the *AP3/DEF* B-class genes in angiosperms (Zahn et al., 2005). *AP3/DEF* is responsible for encoding MADS domain containing transcription factors, which are required for the identity of petals and stamens in the developing floral meristem (Alvarez-Buylla et al., 2000a,b; Goto and Meyerowitz, 1994; Jack et al., 1992, 1994; Theissen et al., 2000). AP3/DEF belongs to the MIKC-type MADS-domain proteins generally consisting of four domains; a MADS domain (M), an Intervening region (I), a Keratin-like domain (K), and a C-terminus (C) (Munster et al., 1997; Fig. 1). The Keratin-like domain, which is located between I and C domain, is involved in mediating specific protein/protein interactions (Riechmann et al., 1996; Kaufmann et al., 2005). Like the MADS domain, the K-domain is a well-conserved region in most of the MADS box genes.

In the present study, we choose to sequence the *AP3/DEF* K-domain for the following reasons: At low phylogenetic level, the K-domain exons evolve fairly slow (in contrast to the C-terminal domain) allowing the design of specific primers, useful for all species of the genus. Furthermore, there is need for low-copy nuclear genes that could be used for phylogeny reconstruction in *Impatiens* and many other groups of flowering plants. According to Janssens et al. (2006) *Impatiens* can be regarded as a recently originated genus that was subject to an explosive speciation. Due to the rapid radiation of *Impatiens*, chloroplast markers provide insufficient resolution for the most diversified lineages of the genus. Therefore, we choose to utilize the low-copy nuclear *AP3/DEF*-like gene to improve the robustness and resolution of phylogenetic relationships at low taxonomic level, regarding the success of Bailey and Doyle (1999) who sequenced *PI* introns for phylogeny reconstruction.

Recently, Geuten and collaborators identified two copies of the *AP3/DEF* B-class gene in *Impatiens*, and one copy in *Marcgravia* (*MuDEF*). According to the phylogeny of the cloned *AP3/DEF*-like genes, *MuDEF* is sister of both *Impatiens* copies, suggesting the occurrence of a duplication in the *AP3/DEF* lineage subsequent to the divergence of Marcgraviaceae and Balsaminaceae. *Impatiens* is one of the two genera belonging to the Balsaminaceae family, which is considered to be closely related to the neotropical Marcgraviaceae (Anderberg et al., 2002; Bremer et al., 2002; Geuten et al., 2004; Schönenberger et al., 2005).

So far, most of the studies on *AP3/DEF* MADS box genes have been focusing on the evolutionary aspects of this B-class gene. The goal of this study is to investigate

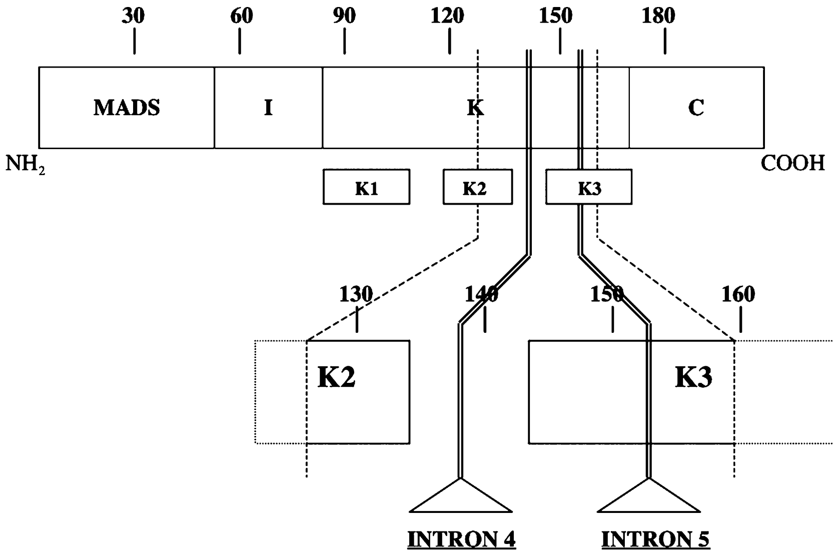

Fig. 1. Schematic representation of the MIKC-type domain structure of plant MADS proteins (MADS-, I-, K- and C-domain). The K domain, which is of specific interest for our study is hypothesized to form three amphipathic ∝-helices (K1, K2, and K3). The numbers at the top designate amino acid positions. Dashed vertical lines indicate the region that has been examined in this manuscript, while the position of intron 4 and 5 is marked as solid vertical lines.

the phylogenetic utility of *AP3/DEF* intron 4 and intron 5 at low taxonomic level and to examine the evolution of the two *AP3/DEF* introns in *Impatiens*.

2. Materials and methods

2.1. Molecular protocols

A list of 59 taxa with authorities, localities, voucher numbers and accession numbers is shown in Table 1. Total genomic DNA was extracted 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 two *AP3/DEF* homologues (*ImpDEF1* and *ImpDEF2*) in *Impatiens* were designed based on the results of Geuten et al. (2006): IhDEF1F (5′-GGGCTTGAGCAAGACATGGACAGT-3′), IhDEF1R (5′-CTTTTCCTGAAGTAGGTTTCTGTGTATCTG-3′), IhDEF2F (5′-AAACTCGAACAAGATGTGGATAGT-3′) and IhDEF2R (5′-AAATTGCTGGAGTAAGTTTCTATGTGTCTG-3′). The temperature profile for *ImpDEF1* consisted of 2 min initial denaturation at 94 °C and 30 cycles of 30 s denaturation at 94 °C, 30 s primer annealing at 57 °C and 1 min extension at 72 °C. Amplification of *ImpDEF2* was performed under the same conditions as described above except for an annealing temperature of 55.5 °C. Amplification reactions were carried out on a GeneAmp PCR system 9700 (Applied Biosystems). Cycle sequencing reactions were performed as in Geuten et al. (2004).

2.2. Data matrices and alignment

ImpDEF1 and *ImpDEF2* sequences were initially aligned with CLUSTALX applying the default parameters for gap opening and gap extension. Further adjustment of the preliminary aligned data matrix was performed manually using MacClade 4.05 (Maddison and Maddison, 2002). For parsimony-based analyses, we coded the indels in both duplicates as separate characters, following the ‘simple indel coding’ method (Simmons and Ochoterena, 2000). Non-informative or ambiguous gaps were not coded. Only the indel coded data matrices were used for phylogenetic analyses. Regarding the congruence between the data sets of both paralogues, we combined both matrices regardless of few species missing in one of the data sets. Therefore, a combined *AP3/DEF* matrix was obtained comprising 59 species. The same selection of species of the *atpB-rbcL* matrix of Janssens et al. (2006) was used for a combined nuclear-chloroplast analysis. All sequences were submitted to GenBank (Table 1), whereas data sets and representative trees were deposited in TreeBASE (accession number 1675).

2.3. Phylogenetic analyses

Data sets of *ImpDEF1* and *ImpDEF2* were analyzed separately and combined. Additionally, the *ImpDEF1/ImpDEF2* matrix was merged with the chloroplast *atpB-rbcL* data set. A partition homogeneity test (implemented in PAUP* 4.0b10a) was used to determine whether the data matrices were providing different signal in the combined analyses. 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.

Maximum parsimony (MP) analyses were conducted using PAUP* 4.0b10a (Swofford, 2002). Heuristic searches were conducted with tree-bisection reconnection (TBR) branch swapping on 10,000 random addition replicates, with five trees held at each step. Characters were equally weighted and character states were specified to be unordered. Non-parametric bootstrap analysis (MP-BS) was carried out to calculate the relative support for individual clades found in the parsimony analysis (Felsenstein, 1985). For each of the 500 bootstrap replicates, a heuristic search was conducted with identical settings as in the original heuristic analysis.

The best fitting substitution model for Bayesian analysis was selected using a series of likelihood ratio tests as implemented in ModelTest 3.06 (Posada and Crandall, 1998). Due to a large variation in substitution rate between exons and introns, *ImpDEF1* and *ImpDEF2* were analyzed using Bayesian analysis with mixed models. Both homologues were divided into five partitions according to the intron–exon boundaries present in the analyzed fragment, whereas ModelTest 3.06 (Posada and Crandall, 1998) chose an appropriate evolutionary model for each partition. Bayesian analyses were conducted with MrBayes 3.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Four chains (one cold, three heated), initiated from a random starting tree were run for 10 million generations. Every 500 generations, a tree was sampled from the chain for a total of 20,000 trees. Due to the burn-in, 10,000 sample points were discarded until stationarity was established among the chains. PAUP* 4.0b10a was used to calculate a 50% majority-rule consensus tree and to report the posterior probabilities for each clade. Posterior probabilities have been shown to overestimate branch support, so they will be interpreted with caution (Kolaczkowski and Thornton, 2006; Suzuki et al., 2002).

2.4. Data analyses

Only those species that were found in both data sets of *ImpDEF1* and *ImpDEF2* were used to investigate molecular evolution. GC-content and length variation were calculated by using PAUP* 4.0b10a (Swofford, 2002).

In order to investigate the possibility of saturation in the *ImpDEF1* and *ImpDEF2* dataset, saturation plots were used as initially described by Phillipe et al. (1994). Saturation for both datasets was examined by plotting observed distances (uncorrected P) as a function of ML-corrected distances.

Pairwise relative rate comparisons were conducted to analyze the degree of rate divergence between both *AP3/DEF* paralogues for each *Impatiens* species. Intron

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

Taxon	Origin	Voucher	Accession number *ImpDEF1*	Accession number *ImpDEF2*	Accession number *atp*B-*rbc*L
Impatiens apsotis Hook.f.	China, Sichuan	Yuan CN2k2-159 (NEU)	EF 133558	—	DQ147810
Impatiens aquatilis Hook.f.	China, Yunnan	Song CNY017 (NEU)	—	EF 133611	DQ147811
Impatiens arguta Hook.f. and Thoms.	China, Yunnan	Yuan CN2k-74 (NEU)	EF 133559	EF 133612	DQ147812
Impatiens aurea Muhl.	North American origin, cult. Holden arboretum	Janssens SJ008 (LV)	EF 133560	EF 133613	DQ147813
Impatiens aureliana Hook.f.	China, Yunnan	Yuan CN2k1-56 (NEU)	EF 133561	EF 133614	DQ147814
Impatiens auricoma Baill.	Comores origin, cult. Bot. Gard. Marburg	Janssens SJ001 (LV)	EF 133562	EF 133615	DQ147815
Impatiens balsamina L.	Indian origin, cult. Kruidtuin Leuven	Janssens SJ003 (LV)	EF 133563	EF 133617	DQ147816
Impatiens balfourii Hook.f.	Himalayan origin, cult. Denver Bot. Gard.	33051 (DBG)	EF 133564	EF 133616	DQ147817
Impatiens bicaudata H. Perrier	Madagascan origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133566	EF 133618	—
Impatiens begonifolia S. Akiyama and H. Ohba	China, Yunnan	Yuan CN2k1-51 (NEU)	EF 133565	EF 133619	DQ147819
Impatiens campanulata Wight	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133567	EF 133620	DQ147822
Impatiens capensis Meerb.	North American origin, cult. Holden arboretum	Janssens SJ009 (LV)	EF 133568	EF 133621	DQ147823
Impatiens chinensis L.	China, Yunnan	Yuan CN2k1-49 (NEU)	EF 133569	EF 133622	DQ147825
Impatiens chungtienensis Y.L. Chen	China, Yunnan	Yuan CN2k2-204 (NEU)	EF 133570	EF 133623	DQ147826
Impatiens clavicornu Turcz.	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133571	EF 133624	DQ147827
Impatiens columbaria J.J. Bos	African origin, cult. Nat. Bot. Gard. Meise	FB/S2966 (BR)	EF 133572	EF 133625	DQ147828
Impatiens conchibracteata Y.L. Chen	China, Yunnan	Hao 427 (NEU)	EF 133573	EF 133626	DQ147829
Impatiens congolensis G.M. Schulze and R. Wilczek	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE7 (NEU)	EF 133574	EF 133627	DQ147830
Impatiens cuspidata Wight and Arn.	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133575	EF 133628	DQ147832
Impatiens cyanantha Hook.f.	China, Yunnan	Yuan CN2k1-84 (NEU)	—	EF 133629	DQ147833
Impatiens davidi Franch.	China, Fujian	Yuan CN2k-09 (NEU)	EF 133576	EF 133630	DQ147835
Impatiens delavayi Franch.	China, Yunnan	Chassot and Yuan 99-154 (NEU)	EF 133577	EF 133631	DQ147836
Impatiens desmantha Hook.f.	China, Yunnan	Yuan CN2k-30 (NEU)	EF 133578	—	DQ147837
Impatiens eberhardtii Tardieu	Vietnam, Anam	Song s.n. (NEU)	EF 133579	EF 133632	DQ147839
Impatiens edgeworthii Hook.f.	Himalayan origin, cult. Univ. California Bot. Gard. Berkeley	89.2005 (UC)	EF 133580	EF 133633	DQ147840
Impatiens faberi Hook.f.	China, Sichuan	Song S007 (NEU)	EF 133581	EF 133634	DQ147841
Impatiens fenghwaiana Y.L. Chen	China, Guangxi	Yuan CN2k-41 (NEU)	—	EF 133635	DQ147842
Impatiens flaccida Arn.	South Indian origin, cult. Nat. Bot. Gard. Meise	FB/S3925 (BR)	EF 133582	EF 133636	DQ147845
Impatiens forrestii Hook.f. ex W.W. Smith	China, Yunnan	Yuan CN2k-79 (NEU)	EF 133583	EF 133637	DQ147847
Impatiens glandulifera Arn.	Belgium, Leuven	Janssens SJ002 (LV)	EF 133584	EF 133638	DQ147848
Impatiens hians Hook.f.	African origin, cult. Bot. Gard. Berlin	Schwerdtfeger 9492a (B)	EF 133585	EF 133639	DQ147849
Impatiens hawkeri W. Bull	New Guinean origin	Janssens SJ006 (LV)	EF 133586	EF 133640	DQ147850
Impatiens imbecilla Hook.f.	China, Sichuan	Hao 426 (NEU)	EF 133587	EF 133641	DQ147851
Impatiens kerriae Craib	Thailand, Qingmai	Chassot 99-238 (NEU)	—	EF 133642	DQ147853
Impatiens latifolia L.	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133588	EF 133643	DQ147854
Impatiens mengtseana Hook.f.	China, Yunnan	Yuan CN2k1-38 (NEU)	EF 133589	EF 133644	DQ147858
Impatiens meruensis Gilg.	Tanzanian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133590	EF 133645	DQ147859
Impatiens monticola Hook.f.	China, Sichuan	Hao 425 (NEU)	EF 133591	EF 133646	DQ147860
Impatiens omeiana Hook.f.	China, Sichuan, cult. Univ. California Bot. Gard. Berkeley	2002.0214 (UC)	EF 133592	EF 133647	DQ147864
Impatiens parviflora DC.	Belgium, Leuven	Janssens SJ004 (LV)	EF 133593	EF 133648	DQ147866

Impatiens platypetala Lindl.	Bali, Indonesian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133594	EF 133649	DQ147868
Impatiens poilanei Tardieu	Vietnam, Anam	Song s.n. (NEU)	EF 133595	EF 133650	DQ147869
Impatiens poculifer Hook.f.	China, Yunnan	Yuan CN2k2-209 (NEU)	EF 133596	EF 133651	DQ147870
Impatiens pseudoviola Gilg.	African origin, cult. Roy. Bot. Gard. Edinburgh	19680124 (E)	EF 133597	—	DQ147871
Impatiens purpurea Hand.-Mazz.	China, Yunnan	Song Y007 (NEU)	EF 133598	EF 133652	DQ147872
Impatiens rectangula Hand.-Mazz.	China, Yunnan	Yuan CN2k1-26 (NEU)	EF 133599	EF 133653	DQ147874
Impatiens rubrostriata Hook.f.	China, Yunnan	Yuan CN2k1-44 (NEU)	EF 133600	EF 133654	DQ147876
Impatiens scabrida DC.	Himalayan origin, cult. Holden arboretum	941314 (DBG)	EF 133601	EF 133655	DQ147877
Impatiens siculifer Hook. f.	China, Yunnan	Yuan CN2k-80 (NEU)	—	EF 133665	—
Impatiens taronensis Hand.-Mazz.	China, Yunnan	Yuan CN2k-57 (NEU)	EF 133602	EF 133656	DQ147882
Impatiens trichosepala Y.L. Chen	China, Yunnan	Yuan CN2k1-68 (NEU)	EF 133603	EF 133657	DQ147885
Impatiens tuberosa H. Perrier	Madagascan origin, cult. Bot. Gard. Univ. Kopenhagen	Janssens SJ005 (LV)	EF 133604	EF 133658	DQ147886
Impatiens uliginosa Franch.	China, Yunnan	Yuan CN2k2-173 (NEU)	—	EF 133659	DQ147887
Impatiens uniflora Hayata	China, Taiwan	Zhengyu Jiang T1 (NEU)	EF 133605	EF 133660	DQ147888
Impatiens usambarensis Grey-Wilson	African origin, cult. Bot. Gard. Koblenz Univ.	Fischer NE20 (NEU)	EF 133606	EF 133661	DQ147889
Impatiens usambarensis Grey-Wilson x *walleriana* Hook.f.	African origin, cult. Roy. Bot. Gard. Edinburgh	19821569 (E)	EF 133607	EF 133662	DQ147890
Impatiens viscida Wight	South Indian origin, cult. by Ray Morgan, U.K.	Ray Morgan s.n. (LV)	EF 133608	EF 133663	DQ147891
Impatiens walleriana Hook.f.	African origin, cult. Nat. Bot. Gard. Meise	S3926 (BR)	EF 133609	EF 133664	DQ147892
Impatiens xanthina Comber	China, Yunnan	Yuan CN2k1-15 (NEU)	EF 133610	—	DQ147893

sequences were ignored in this test. Tree taxon models were assembled for pairwise comparisons of the *AP3/DEF* paralogues in each *Impatiens* species. *Marcgravia umbellata* was chosen as outgroup (Geuten et al., 2006). We used the Likelihood Ratio test (LRT) method of Muse and Gaut (1994) as implemented in HyPhy (Kosakovsky Pond et al., 2005) to conduct the pairwise relative rate tests. The method of Muse and Gaut (1994) is a LRT of rate constancy between two lineages with reference to an outgroup lineage.

In order to identify positively selected amino acid sites in the K-domain of the *ImpDEF* duplicates, the ratio of non-synonymous to synonymous substitutions (d_N/d_S or ω) at 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 using discrete distribution models (M1, M2 and M3) and 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 beta (M7) and beta + ω model (M8).

3. Results

3.1. Gene copy number

A recent study in *Impatiens* revealed two copies of the *AP3/DEF* subfamily in *I. hawkeri* corresponding to the *euAP3* gene lineage (Geuten et al., 2006). Our results demonstrate that these paralogues are present in almost all other species of *Impatiens* examined. We were unable, however, to amplify the *ImpDEF2* paralogue in three species (*I. apsotis*, *I. pseudoviola* and *I. xanthina*). Earlier phylogenetic analyses (Janssens et al., 2006; Yuan et al., 2004) showed that these taxa have no common most recent ancestor. Additionally, we found five representatives of *Impatiens*, in which *ImpDEF1* was not amplifiable (*I. siculifer, I. fenghwaiana, I. uliginosa, I. aquatilis* and *I. cyanantha*). The last four species form a well-supported clade in previous chloroplast and nuclear phylogenies (Janssens et al., 2006; Yuan et al., 2004). The *ImpDEF2* paralogue in *I. desmantha* and the *ImpDEF2* paralogue in *I. kerriae* have been amplified in this study, but were removed from the data set since it was impossible to obtain a useful sequence for this locus. Due to a lack of fresh material to obtain mRNA from, it was impossible to find out whether some of these paralogues were not amplifiable because of a possible loss of one of the duplicates or due to difficulties with the PCR reaction. Despite this obscurity, our results certainly argue in favor of the presence of two copies of *AP3*/*DEF* since the origin of *Impatiens* or before.

3.2. Sequence variation

As with most plants, the AT content was significantly higher in the introns than in the exons. In intron 4 and 5 of *ImpDEF1* and intron 4 of *ImpDEF2*, thymine is the most common nucleotide with an average of 41.8%, while adenine is second most common with an average of 29.8%. However, intron 5 of *ImpDEF2* has an average amount of adenine (32.7%), which is roughly the same as that of thymine (33.4%). The saturation plots clearly showed that *ImpDEF1* and *ImpDEF2* introns investigated are not saturated in *Impatiens* (Fig. 2).

The average length of the introns investigated in *ImpDEF1* remains rather constant in the whole genus (76–132 bp in intron 4 and 64–95 bp in intron 5) except for one species (*I. flaccida*) with an enlarged intron 5 (249 bp). Also intron 4 of *ImpDEF2* shows a rather constant intron length (68–139 bp). On the contrary, the overall length of intron 5 in *ImpDEF2* varies strongly (69–474 bp) with a trend towards an increase of length in the more recently diversified lineages of the genus (Fig. 3). The length variation of intron 5 in *ImpDEF2* is generated by the occurrence of frequent deletions and insertions. We observed, however, that the deletion bias is rather weak, while the occurrence of long insertions (ranging from 126 nt to 266 nt in intron 5) is quite frequent. According to Comeron and Kreitmann (2000), insertions of this size (longer than 100 bp) do not occur very often in introns. Furthermore, Moriyama et al. (1998) assumed that sudden transitions from short towards long introns appear to be quite unusual. In *Arabidopsis*, the most favorable intron length is between 80 and 90bp (Hebsgaard et al., 1996). Goodall and Filipowicz (1990) hypothesized a minimum intron length of 70–73 nt in dicots, whereas Filipowicz et al. (1994) postulated a minimum of only 64 nt. In *ImpDEF2*, the intron length of intron 5 is close to 74 nt in most of the early diversified lineages of *Impatiens*, whereas the average intron length in the more derived lineages of *Impatiens* (Clades 7–15) is 275 nt on average. One can assume that the natural mutational bias towards deletions has been countered by strong selection in order to preserve the minimum intron length that is necessary for intron splicing (Comeron and Kreitmann, 2000).

3.3. Phylogenetics of ImpDEF1

The *ImpDEF1* region examined in our study contains 818 characters, from which 27 nt belong to exon 4, 304 nt to intron 4, 42 nt to exon 5, 430 nt to intron 5 and 15 nt to exon 6. Of the 273 variable positions, 176 are parsimony informative. MP analyses of *ImpDEF1* resulted in 2268 trees of 483 steps, a consistency index (CI) of 0.683 and a retention index (RI) of 0.804. The strict consensus tree is moderately resolved, but most of the clades identified by Janssens et al. (2006) were only supported by low bootstrap values (Fig. 4). The much better resolved Bayesian consensus tree is nearly similar to the MP consensus phylogeny. Most of the lineages have Bayesian posterior probabilities (BPP) of 95% or higher (Fig. 4). MP and Bayesian inference resolved the three main groups in *Impatiens* as described in Janssens et al. (2006). Although Clade I was recognized by parsimony analysis, this lineage shows no bootstrap support (Clade I—BS: <50/BPP: 100, Clade II—BS: 79/BPP: 100, Clade III—BS: 72/BPP: 100; Fig. 4).

3.4. Phylogenetics of ImpDEF2

The sequenced *ImpDEF2* region comprises 1438 aligned characters, of which 27 nt correspond to exon 4, 307 nt to intron 4, 42 nt to exon 5, 1047 nt to intron 5 and 15 nt to exon 6. Due to the fast evolving intron 5, the *ImpDEF2* fragment examined in our study is far more variable than the sequenced *ImpDEF1* fragment. Of the unambiguously aligned nucleotides, 435 are variable and 282 are parsimony informative. Parsimony analysis of *ImpDEF2* resulted in 374 trees of length 692 (CI: 0.752 and RI: 0.845). Despite the more variable nature of the intron structure in *ImpDEF2*, a much better resolved phylogeny was generated compared to the one *ImpDEF1*. Also bootstrap support values are seemingly higher in the *ImpDEF2* topology (Fig. 5). Bayesian analysis generated a consensus tree that was better resolved than the MP topology. Furthermore, most of the delimited lineages showed high Bayesian support percentages (Fig. 5).

In contrast with *ImpDEF1*, parsimony and Bayesian analysis of the *ImpDEF2* data matrix supported only two of the three main clades in *Impatiens*, as recognized in Janssens et al. (2006): Clade II — BS: 89/BPP: 100 and Clade III — BS: 87/BPP: 100. Although the earliest diverged lineages of *Impatiens* collapsed using the *ImpDEF2* data matrix, we noticed that the more recently diversified lineages are generally well resolved (Fig. 5).

3.5. Phylogenetics of combined ImpDEF1 and ImpDEF2

The combined data matrix contains 59 species and 2256 characters from which 708 are variable and 458 parsimony informative. The partition homogeneity test found no significant difference between both partitions of the combined data set ($p > 0.05$). Analysis of the combined data set yielded 1764 trees with length 1216 (CI: 0.718 and RI: 0.824).

Both Bayesian topology and parsimony consensus tree (Fig. 6) are completely in agreement with the consensus tree produced by the independent data sets. In comparison to *ImpDEF1* and *ImpDEF2*, we observed that the combined phylogeny is better resolved and that support values were generally higher for the combined data matrix. Most of the lineages have Bayesian support values of 95% or higher (Fig. 6). Once more, only two of the three major

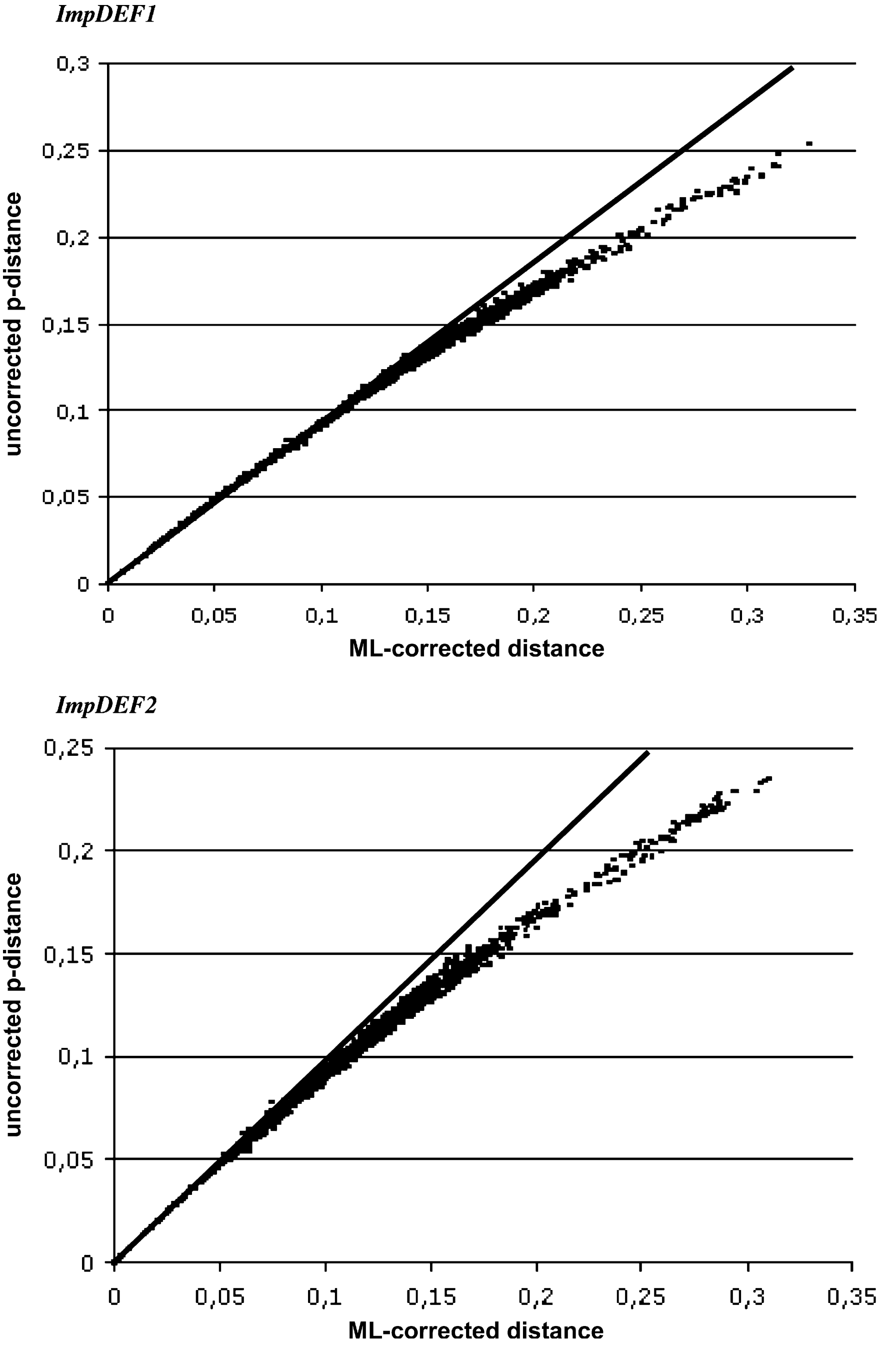

Fig. 2. Saturation plots of the *ImpDEF1* and *ImpDEF2* datasets in which ML distances were plotted against observed p-distances.

clades, described by Janssens et al. (2006) in *Impatiens* were resolved by MP and Bayesian analysis, but with moderate to high support (Clade II—BS: 80/BPP: 100 and Clade III—BS: 94/BPP: 100). The most early diverged lineages are also well resolved, but due to a collapsed position of *I. kerriae*, Clade I (Janssens et al., 2006) cannot be delimited. Nevertheless, many of the internal lineages, as recognized in Janssens et al. (2006) generally have moderate support values (Fig. 6): Himalayan-Eurasian clade 3 (BS: 77 and BPP: 100), South Chinese clade 4 (BS: 71 and BPP: 97), North American-Chinese clade 6 (BS: 67 and BPP: 100), South Chinese clade 7 (BS: 99 and BPP: 100), South Indian clade 8 (BS: 75 and BPP: 97), South Chinese clade 9 (BS: 99 and BPP: 100), West African clade 10 (BS: 90 and BPP: 100), Southeast Asian clade 11 (BS: 100 and BPP: 100), Madagascan clade 12 (BS: 100 and BPP: 100), East African clade 14 (BS: 91 and BPP: 100), South Indian-Vietnamese clade 15 (BS: 75 and BPP: 100).

3.6. Phylogenetics of combined *ImpDEF1*/*ImpDEF2* and *atpB*-*rbcL*

The resulting p value of the partition homogeneity test was not significant ($p > 0.05$), indicating that also the

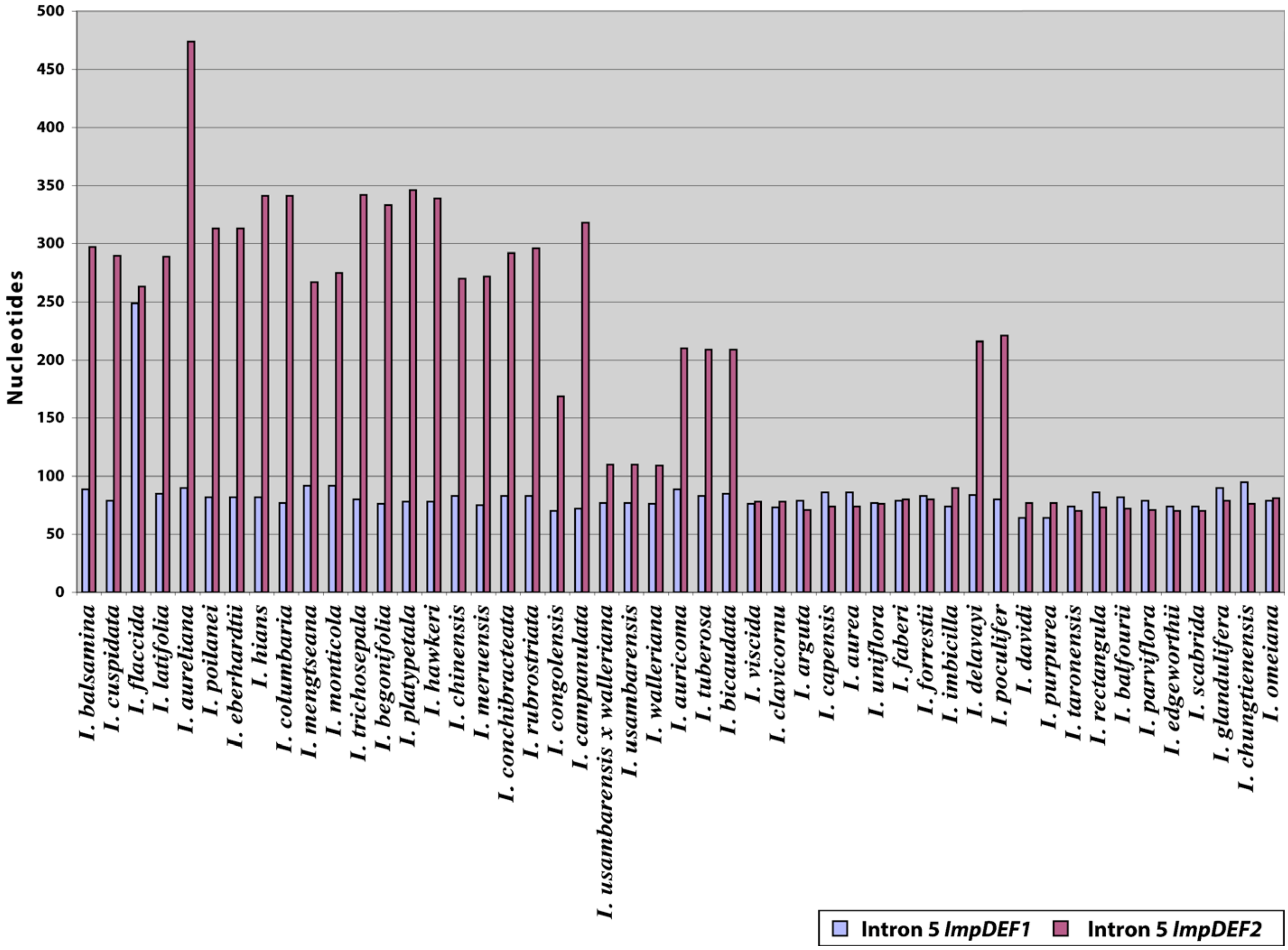

Fig. 3. Length variation of intron 5 in *ImpDEF1* and *ImpDEF2*.

difference between both partitions of the combined data are not in conflict. The combined nuclear-chloroplast data matrix contains 59 species and 3217 characters from which 985 are variable and 612 are parsimony informative. Maximum parsimony analysis of *ImpDEF1*, *ImpDEF2* and *atpB*-*rbcL* rendered 1680 trees of 1427 steps (CI: 0.724 and RI: 0.826). The highly resolved MP consensus tree (Fig. 7) is in agreement with the consensus tree produced by the independently analyzed and the combined *ImpDEF1*/*ImpDEF2* data sets, except for one discrepancy in clade 9. Within this clade, *I. xanthina* is sister to *I. begonifolia* and *I. trichosepala* in the combined *ImpDEF1*/*ImpDEF2* dataset, while in the combined chloroplast/nuclear dataset, *I. begonifolia* is sister to *I. xanthina* and *I. trichosepala*. Nevertheless, support for these taxa is generally low.

The Bayesian search yielded a well-resolved topology in which the overall relationships are depicted in the same way as in the parsimony-based one. We observed that Bayesian posterior probabilities and bootstrap support values for the combined nuclear/chloroplast data matrix were generally strong. All three major clades in *Impatiens* (Janssens et al., 2006) are resolved by MP and Bayesian inference, however, Clade I shows no bootstrap support (Clade I—BS: <50/BPP: 97, Clade II—BS: 89/BPP: 100 and Clade III—BS: 98/BPP: 100). Furthermore, the majority of the internal lineages are highly supported (Fig. 7): Himalayan-Eurasian clade 3 (BS: 99 and BPP: 100), South Chinese clade 4 (BS: 85 and BPP: 100), North American-Chinese clade 6 (BS: 76 and BPP: 100), South Chinese clade 7 (BS: 100 and BPP: 100), South Indian clade 8 (BS: 65 and BPP: 100), South Chinese clade 9 (BS: 99 and BPP: 100), West African clade 10 (BS: 98 and BPP: 100), Southeast Asian clade 11 (BS: 100 and BPP: 100), Madagascan clade 12 (BS: 100 and BPP: 100), East African clade 14 (BS: 100 and BPP: 100), South Indian-Vietnamese clade 15 (BS: 89 and BPP: 100).

3.7. Molecular evolution

In the present study, relative rate tests based on the three-taxon concept of Muse and Gaut (1994) were used to test for significance of rate differences between both paralogues. The results of the test demonstrated that there is in

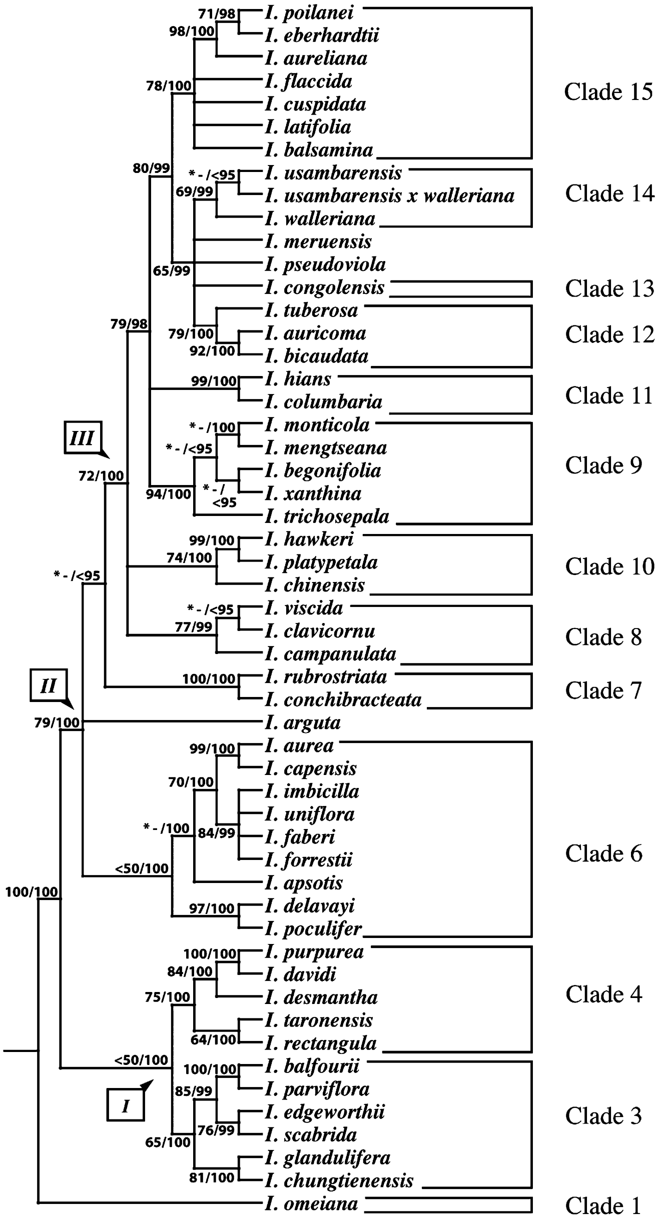

Fig. 4. Phylogenetic consensus tree based on *ImpDEF1* sequences. The first number on the branch represents bootstrap support of the MP analysis and the second number indicates Bayesian posterior probabilities. An asterisk indicates branches that collapse in the MP consensus tree. Deeper internal nodes (Roman numbers) or specific clades (numbers 3–15) are assigned according to Janssens et al. (2006).

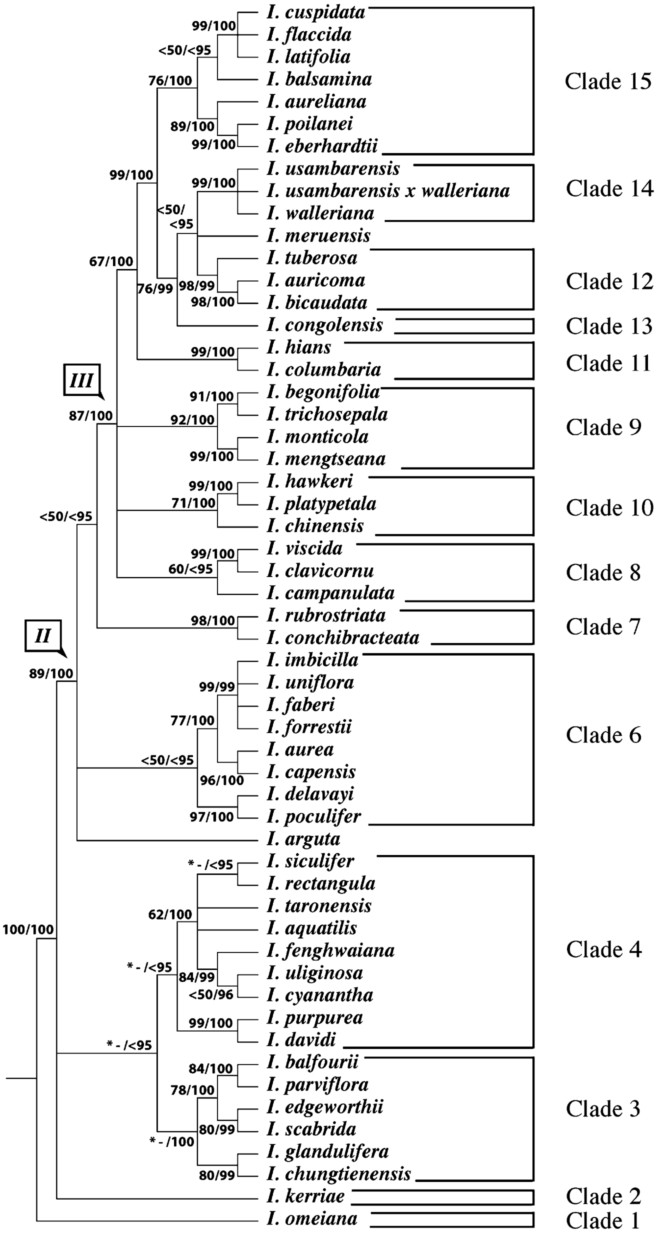

Fig. 5. Consensus phylogeny of the *ImpDEF2* data set. The first number on the branch represents bootstrap support of the MP analysis and the second number indicates Bayesian posterior probabilities. Branches that collapse in the MP consensus tree are indicated with an asterisk. Similar numbering is used as recognized in Janssens et al. (2006) to assign deeper internal nodes (Roman numbers) or specific clades (numbers 3–15).

general no significant rate difference between *ImpDEF1* and *ImpDEF2* in *Impatiens*, as the null hypothesis was almost never rejected in the different pairwise comparisons ($p > 0.05$). Nevertheless, for four species the null hypothesis was rejected ($p = 0.04 < 0.05$), suggesting a significant difference in rate between the K-domain exon sequences of the two *AP3/DEF* paralogues. Remarkably, these four species (*I. uniflora, I. faberi, I. imbicilla* and *I. forrestii*) all belong to the same clade. In this clade, the *ImpDEF2* coding sequence has evolved significantly further away from the outgroup sequence than *ImpDEF1*.

Furthermore, we tested whether there was detectable positive selection on codons for 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) mutations provided similar results for both *ImpDEF1* and *ImpDEF2* (Tables 2 and 3). In general, no positively selected sites could be identified, while all sites are characterized by a low omega value ($\omega = 0.1$) possibly suggesting that purifying selection has occurred.

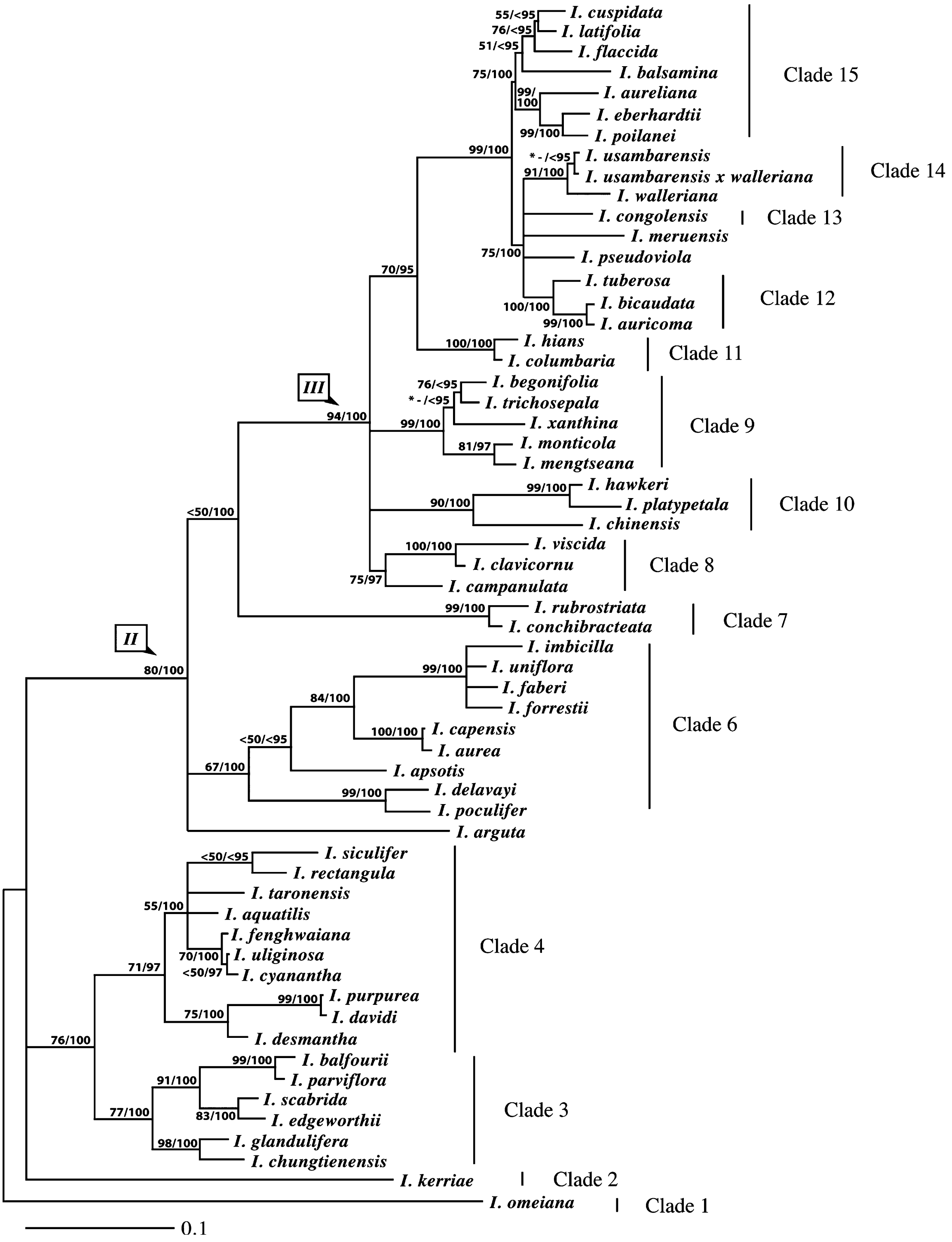

Fig. 6. Bayesian consensus phylogram based on the branch length of the combined *ImpDEF1* and *ImpDEF2* data set. The first number on the branch represents bootstrap support of the MP analysis and the second number indicates Bayesian posterior probabilities. An asterisk indicates branches that collapse in the MP consensus tree. Similar numeration is used as described in Janssens et al. (2006) to assign deeper internal nodes (Roman numbers) or specific clades (numbers 3–15).

4. Discussion

4.1. Phylogenetic usefulness of *AP3/DEF*

The use of large multigene family members for molecular phylogenetics is often considered to be problematic as paralogues loci could be mistaken for orthologues loci (Popp and Oxelman, 2004). Although *AP3/DEF*-like genes form a well-supported clade within angiosperms, several studies illustrated the occurrence of numerous duplication events at every taxonomic level within this *AP3/DEF* lineage (e.g., Di Stilio et al., 2005; Geuten et al., 2006; Kramer et al., 1998, 2003; Stellari et al., 2004; Zahn et al., 2005). Unmistakably, these duplication events could undermine

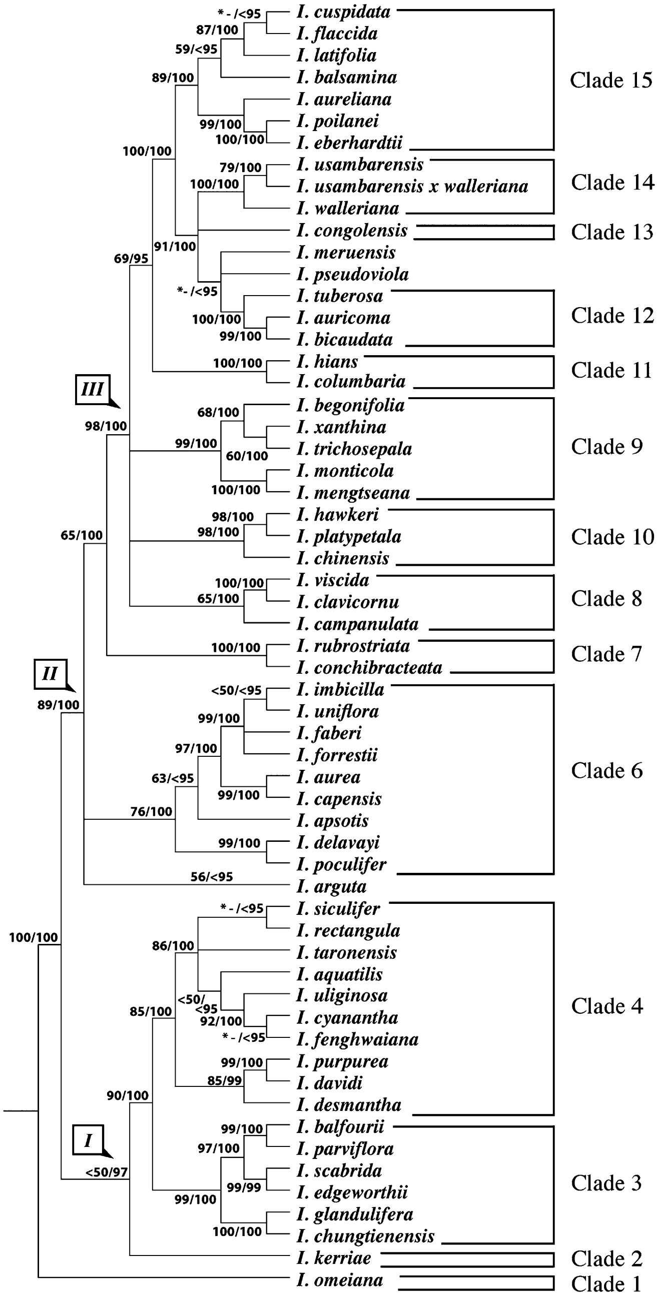

Fig. 7. Phylogenetic hypothesis based on the combined data set of *ImpDEF1*, *ImpDEF2*, and *atpB-rbcL*. The first number on the branch represents bootstrap support of the MP analysis and the second number indicates Bayesian posterior probabilities. An asterisk indicates branches that collapse in the MP consensus tree. Deeper internal nodes (Roman numbers) or specific clades (numbers 2–15) are assigned according to Janssens et al. (2006).

the phylogenetic utility of *AP3/DEF* like genes. However, because of the large number of studies that are carried out within the *AP3/DEF* gene lineage, several duplication events within this B-class gene lineage are mapped for angiosperm taxa. Moreover, the public availability of these sequences opens the possibility to design specific primers for those species and their paralogous gene loci that have been investigated in former studies.

In order to avoid any confusion within our group of interest, we independently investigated both paralogues of *AP3/DEF* that were identified in *Impatiens* by Geuten et al. (2006). By designing locus-specific primers for both *AP3/DEF* paralogues in *Impatiens*, we could amplify and sequence both duplicates without interference of the other copy. The use of these specific primers also eliminated the need for cloning to acquire unambiguous sequences. Furthermore it seems that we did not amplify any other locus that would be paralogues to *ImpDEF1* or *ImpDEF2*.

Our results show that *AP3/DEF* introns can be applied for phylogenetic analysis at a relatively low taxonomic level. Both paralogues of this gene (*ImpDEF1* and *ImpDEF2*) have been investigated separately and indicate that K-domain introns 4 and 5 in *AP3/DEF* contain useful phylogenetic information. Intron 4 and intron 5 are highly variable in *ImpDEF1* and *ImpDEF2*, resulting in a higher percentage of variable sites than the chloroplast *atpB-rbcL* spacer. Despite this high amount of variable sites in both *AP3/DEF* copies in *Impatiens*, the consistency and retention indices imply that the homoplasy content of introns 4 and 5 in *ImpDEF1* and *ImpDEF2* is low, and additionally intron 4 and intron 5 turn out to be almost unsaturated in both *AP3/DEF* duplicates. Furthermore, the phylogenies of *ImpDEF1* and *ImpDEF2* largely resemble the chloroplast *atpB-rbcL* phylogeny (Janssens et al., 2006; Figs. 4–6) and by combining both nuclear data sets, we obtained a phylogenetic hypothesis, which is better resolved than in separate analyses. By adding *atpB-rbcL* to the *ImpDEF1/ImpDEF2* data set, both resolution and support of the overall topology was improved once more.

The *AP3/DEF* locus in Balsaminaceae and in most angiosperms is characterized by six introns and five exons. Several studies have shown that introns of low-copy nuclear genes tend to diverge at high rate for both nucleotide and indel substitution (Gaut, 1998; Li, 1998; Sang, 2002). In contrast, chloroplast introns evolve at a significantly lower rate, frequently resulting in insufficient signal for phylogenetic studies at low taxonomic level. In *AP3/DEF* we found an i nteresting low copy nuclear gene, which can be used for the examination of relationships among closely related species, an aspect that certainly is becoming more important nowadays. The introns investigated within *ImpDEF1* and *ImpDEF2* can be characterized by a high rate of insertions, deletions and substitutions, providing sufficient phylogenetic data to resolve previously uncertain relationships within the rapidly radiated genus *Impatiens*. Due to the high rate of substitution and the extensive length variation within the *AP3/DEF* introns, it was sometimes difficult to align early-diversified species with species that are strongly diverged throughout the genus. However, with a relatively dense sampling that covers all major groups of the genus, alignment did not cause problems. Furthermore both *ImpDEF1* and *ImpDEF2* datasets show only little amount of homoplasy and produced almost identical phylogenetic hypotheses for *Impatiens*.

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

Model	l	ω-value	Proportion of sites	Positively selected sites
K-domain ImpDEF1				
M0 (single ratio)	−364.33	$\omega = 0.0512$		None
M1 (neutral)	−381.83	$\omega_0 = 0$	$p_0 = 0.7257$	Not allowed
		$\omega_1 = 1$	$p_1 = 0.2742$	
M2 (selection)	−364.33	$\omega_0 = 0$	$p_0 = 0.0000$	None
		$\omega_1 = 1$	$p_1 = 0.0000$	
		$\omega_2 = 0.0511$	$p_2 = 1.0000$	
M3 (discrete)	−365.77	$\omega_0 = 0.0511$	$p_0 = 1.0000$	None
		$\omega_1 = 0.0632$	$p_1 = 0.0000$	
		$\omega_2 = 0.1418$	$p_2 = 0.0000$	
M7 (beta)	−364.33			Not allowed
M8 (beta and ω)	−364.33	$\omega_0 = 0.0511$	$p_0 = 0.9999$	None
		$\omega_1 = 1.0000$	$p_1 = 0.0001$	

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

Model	l	ω-value	Proportion of sites	Positively selected sites
K-domain ImpDEF2				
M0 (single ratio)	−245.42	$\omega = 0.2739$		None
M1 (neutral)	−249.28	$\omega_0 = 0$	$p_0 = 0.2963$	Not allowed
		$\omega_1 = 1$	$p_1 = 0.7035$	
M2 (selection)	−249.28	$\omega_0 = 0$	$p_0 = 0.2962$	None
		$\omega_1 = 1$	$p_1 = 0.7037$	
		$\omega_2 = 1.2085$	$p_2 = 0.0000$	
M3 (discrete)	−245.42	$\omega_0 = 0.2739$	$p_0 = 1.0000$	None
		$\omega_1 = 0.4210$	$p_1 = 0.0000$	
		$\omega_2 = 0.8761$	$p_2 = 0.0000$	
M7 (beta)	−245.42			Not allowed
M8 (beta and ω)	−245.42	$\omega_0 = 0.2744$	$p_0 = 0.9999$	None
		$\omega_1 = 1.0000$	$p_1 = 0.0001$	

Although *AP3/DEF* introns are interesting for resolving phylogenetic relationships between relatively closely related species, they are probably not appropriate for phylogenetic analyses at high taxonomic levels. For example, it was not possible to align the introns of *M. umbellata* (Marcgraviaceae; Geuten et al., 2006), which is believed to be closely related to *Impatiens*, with the *ImpDEF1* or *ImpDEF2* introns of the most early diversified species in *Impatiens*. In fact, introns are occasionally considered to evolve too rapidly and too erratically to be used in phylogenetic reconstruction (Kupfermann et al., 1999). This is probably true when using introns to resolve relationships at inter- and intrafamilial level.

On the other hand, *AP3/DEF* exons are probably not useful for phylogenetics at low taxonomical level. Exons are generally considered to be under strong purifying selection. As a result, the exon substitution rate is restricted in order to avoid deleterious mutations within the coded sequence. Within *ImpDEF1* and *ImpDEF2*, the constrained substitution rate of the exons resulted in a relative high amount of homoplasy, making exons not useful at low phylogenetic level.

4.2. Molecular evolution of *AP3/DEF* in *Impatiens*

According to our results, we assume that the two copies of *AP3/DEF* in *Impatiens* are not redundant. Amino acid sequences of both copies differ considerably, especially in the C-domain (Geuten et al., 2006). Furthermore, both duplicates are present in nearly all representatives of *Impatiens* examined, and a strongly delimited Intron–Exon structure exists in both paralogues, suggesting that both genes are still functional. Relative rate tests showed generally no significant rate difference between both duplicates. Nevertheless, within the well-supported clade comprising *I. uniflora, I. faberi, I. fenghwaiana* and *I. forrestii*, we observed that the K-domain region of *ImpDEF1* and *ImpDEF2* evolved at a different evolutionary rate. We conclude that *ImpDEF1* and *ImpDEF2* in *Impatiens* avoided the process of gene loss, which is a typical fate in most of the gene duplicates. Nonetheless, it might be interesting to find out whether the incapability of obtaining both duplicates for some species is due to a loss of one of the *ImpDEF* copies or simply because of changes in particular primer sites.

An interesting case of molecular evolution within the *AP3/DEF* like genes in *Impatiens* can be found in the enlargement of intron 5 in *ImpDEF2*. Remarkably, this lengthening has only occurred in the most diversified species of *Impatiens* (Clades 7–15). Nevertheless, no morphological differences in petals and stamens have been observed so far, which could distinguish the most early diverged species (Clades 1–6) from the more recent diversified species (Clade 7–15). During the last decade, several studies tried to explain the importance of intron length divergence (Bell et al., 1999; Castillo-Davis et al., 2002; Comeron and Kreitmann, 2000; Parsch, 2003). Mount et al. (1992) emphasized the necessity of multiple compensatory mutations in order to converse a short intron into a long intron. A recent study demonstrated that short introns are mostly favored in highly expressed genes, e.g., genes encoding ribosomal proteins (Castillo-Davis et al., 2002). This way, the costs of transcription and other molecular processes like splicing could be reduced. Additionally, intron lengthening results in an increase of secondary structures within the pre-mRNA sequence. These structures are known to generate alternative splicing sites and might affect the original splicing of the intron–exon boundaries (Eperon et al., 1988; Lecharny et al., 2003). Possibly, the enlargement of intron 5 in *ImpDEF2* resulted in an increase of alternative splicing sites within the specific region. In case of alternative splicing in intron 5 of *ImpDEF2*, the translated products would become truncated due to the occurrence of in-frame stop codons, which are present in the close proximity of the 5′ and 3′ splicing site of intron 5 in most *Impatiens* species examined.

In both *AP3/DEF* paralogues, we observed that the protein coding exon structure in the K-domain is in general highly conservative, showing no positively selected changes. Within the *ImpDEF2* lineage, there is no species in which an amino acid substitution has taken place. In contrast, within the *ImpDEF1* lineage, four amino acid substitutions have occurred in different representatives throughout the whole genus. In *I. eberhardtii* we located an amino acid substitution in which a positively charged hydrophilic lysine is changed into an uncharged hydrophilic threonine in the region between the K2 and K3 subdomain at position 141 (Fig. 8). The other three amino acid substitutions in *ImpDEF1*, on the other hand, did not change in function compared to the amino acid they have replaced. According to Yang et al. (2003), the K3 subdomain of the *AP3/DEF* and *PI/GLO* subfamilies is not completely conserved in its hydrophobicity on positions **a** and **d**, sometimes causing a shift towards hydrophilic amino acids. However, when hydrophobic amino acids are present on positions **a** and **d** in the K3 subdomain, these specific amino acids remain unchanged in both *DEF* duplicates in *Impatiens*.

```
ImpDEF1
                defga          defgabcdefg   abcde
Impatiens       SLQRIRDRK)-(FKVLGNQIETHRKK)-(LRNVE
I. eberhardtii  .........)-(.T............)-(.K...
I. hians        .........)-(..............)-(.K...
I. columbaria   .........)-(..............)-(.K...
I. capensis     ......E..)-(..............)-(.....
I. aurea        ......E..)-(Y.............)-(.....
I. uniflora     ......E..)-(..............)-(.....
I. faberi       ......E..)-(..............)-(.....
I. forrestii    ......E..)-(..............)-(.....
I. imbicilla    ......E..)-(..............)-(.....
```

Fig. 8. Listed here are all species with inferred amino acid changes in *ImpDEF1* or *ImpDEF2*. Specific changes in amino acids are underlined and written in bold. Heptad repeats are present above each duplicate.

Acknowledgments

We thank the National Botanic Garden of Belgium (BR), National Herbarium of the Netherlands, Leiden University branch (L), Utrecht University branch (U) and Wageningen University branch (W), Beat Leuenberger of the Botanischer Garten Berlin-Dahlem (B), Holly Forbes of the University of California Botanical Garden at Berkeley (UC), the Botanical Garden of Marburg (MB), the Botanical Garden of the University of Copenhagen (C), the Royal Botanic Garden Edinburgh (E), South China Botanical Garden (IBSC), Ann Berthe of the Denver Botanic Gardens (DBG), Charlotte Chan of the Holden Arboretum, and Ray Morgan for providing plant material. This study was financially supported by research grants of the K.U. Leuven (OT/05/35) and the Fund for Scientific Research-Flanders (FWO, Belgium) (1.5.003.05). Steven Janssens holds a PhD research grant from the FWO, and Tom Viaene from the IWT. Koen Geuten was holding a personal research grant from the K.U.Leuven (OT/01/25) during the course of this study.

References

Aagaard, J.E., Olmstead, R.G., Willis, J.H., Phillips, P.C., 2005. Duplication of floral regulatory genes in the Lamiales. Am. J. Bot. 92, 1284–1293.

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., Yanofsky, M.F., 2000a. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. USA 97, 5328–5333.

Alvarez-Buylla, E.R., Liljegren, S.J., Pelaz, S., Gold, S.E., Burgeff, C., Ditta, G.S., Vergera-Silva, F., Yanofsky, M.F., 2000b. MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. Plant J. 24, 457–466.

Anderberg, A.A., Rydin, C., Källersjö, M., 2002. Phylogenetic relationships in the order Ericales s.l.: analyses of molecular data from five genes from the plastid and mitochondrial genomes. Am. J. Bot. 89, 677–687.

Bailey, C.D., Doyle, J.J., 1999. Potential phylogenetic utility of the low-copy nuclear gene *pistillata* in dictyledonous plants: comparison to nrDNA ITS and *trnL* intron in *Sphaerocardamum* and other Brassicaceae. Mol. Phylogenet. Evol. 13, 20–30.

Bailey, C.D., Price, R.A., Doyle, J.J., 2002. Systematics of the Halimolobine Brassicaceae: evidence from three loci and morphology. Syst. Bot. 27, 318–332.

Bell, M.V., Cowper, A.E., Lefranc, M.-P., Bell, J.I., Screaton, G.R., 1999. Influence of intron length on alternative splicing of CD44. Mol. Cell Biol. 18, 5930–5941.

Bremer, B., Bremer, K., Heidari, N., Erixon, P., Olmstead, R.G., Anderberg, A.A., Källersjö, M., Barkhordarian, E., 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Mol. Phylogenet. Evol. 24, 274–301.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., Kondrashov, F.A., 2002. Selection for short introns in highly expressed genes. Nat. Genet. 31, 415–418.

Comeron, J.M., Kreitmann, M., 2000. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics 156, 1175–1190.

Di Stilio, V.S., Kramer, E.M., Baum, D.A., 2005. Floral MADS box genes and homeotic gender dimorphism in *Thalictrum dioicum* (Ranunculaceae) – a new model for the study of dioecy. Plant J. 41, 755–766.

Emschwiller, E., Doyle, J.J., 1999. Chloroplast-expressed glutamine synthetase (ncpGS): potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). Mol. Phylogenet. Evol. 12, 310–319.

Eperon, L.P., Graham, I.R., Griffiths, A.D., Eperon, I.C., 1988. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? Cell 54, 393–401.

Fan, C., Purugganan, M.D., Thomas, D.T., Wiegmann, B.M., Xiang, Q.-Y., 2004. Heterogeneous evolution of the *Myc-like* Anthocyanin regulatory gene and its phylogenetic utility in *Cornus* L. (Cornaceae). Mol. Phylogenet. Evol. 33, 580–594.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Filipowicz, W., Gniadkowski, M., Klahre, U., Liu, H., 1994. Pre-mRNA splicing in plants. In: Lamond, A.I. (Ed.), Pre-mRNA Processing. R.G. Landes Company, Austin, pp. 65–77.

Gaut, B.S., 1998. Molecular clocks and nucleotide substitution rates in higher plants. Evol. Biol. 30, 93–120.

Geuten, K., Becker, A., Kaufmann, K., Caris, P., Janssens, S., Viaene, T., Theissen, G., Smets, E., 2006. Petaloidy in the balsaminoid genera *Impatiens* and *Marcgravia* as interpreted by changes in B-class MADS-box gene expression. Plant J. 47, 501–518.

Geuten, K., Smets, E., Schols, P., Yuan, Y.-M., Janssens, S., Küpfer, P., Pyck, N., 2004. Conflicting phylogenies of balsaminoid families and the polytomy in Ericales: combing data in a Bayesian framework. Mol. Phylogenet. Evol. 31, 711–729.

Goodall, G.J., Filipowicz, W., 1990. The minimum functional length of pre-mRNA introns in monocots and dicots. Plant Mol. Biol. 14, 727–733.

Goto, K., Meyerowitz, E.M., 1994. Function and regulation of the *Arabidopsis* floral homeotic gene PISTILLATA. Genes Dev. 8, 1548–1560.

Grob, G.B.J., Gravendeel, B., Eurlings, M.C.M., 2004. Phylogenetic utility of the nuclear *FLORICAULA/LEAFY* second intron: comparison with three chloroplast DNA regions in Amorphophallus (Araceae). Mol. Phylogenet. Evol. 30, 13–23.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., Brunak, S., 1996. Splice site prediction in Arabidopsis thaliana DNA by combining local and global sequence information. Nucl. Acids Res. 24, 3439–3452.

Howarth, D.G., Baum, D.A., 2002. Phylogenetic utility of a nuclear intron from nitrate reductase for the study of closely related plant species. Mol. Phylogenet. Evol. 23, 525–528.

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, G.L., Meyerowitz, E.M., 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.

Kaufmann, K., Melzer, R., Theissen, G., 2005. MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene 347, 183–198.

Kim, S., Yoo, M.-J., Albert, V.A., Farris, J.S., Soltis, P.S., Soltis, D.E., 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. Am. J. Bot. 91, 2102–2118.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Kramer, E.M., Di Stilio, V.S., Schlüter, P.M., 2003. Complex patterns of gene duplication in the APETALA3 and PISTILLATA lineages of the Ranunculaceae. Int. J. Plant Sci. 164, 1–11.

Kramer, E.M., Irish, V.F., 2000. Evolution of the petal and stamen developmental programs: Evidence from comparative studies of the lower eudicots and basal angiosperms. Int. J. Plant Sci. 161, 29–40.

Kramer, E.M., Dorit, R.L., Irish, V.F., 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, E.M., 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.

Kolaczkowski, B., Thornton, J.W., 2006. Is there a star tree paradox? Mol. Biol. Evol. 23, 1819–1823.

Kupfermann, H., Satta, Y., Takahata, N., Tichy, H., Klein, J., 1999. Evolution of *Mhc-DRB* introns: implications for the origin of primates. J. Mol. Evol. 48, 663–674.

Lecharny, A., Boudet, N., Gy, I., Aubourg, S., Kreis, M., 2003. Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. J. Struct. Funct. Genomics 3, 111–116.

Lewis, C.E., Doyle, J.J., 2001. Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). Mol. Phylogenet. Evol. 19, 409–420.

Li, W.-H., 1998. Molecular Evolution. Sinauer Associates, Sunderland.

Litt, A., Irish, V.F., 2003. Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. Genetics 165, 821–833.

Mason-Gamer, R.J., Weill, C.F., Kellogg, E.A., 1998. Granule-bound starch synthase: structure, function, and phylogenetic utility. Mol. Biol. Evol. 15, 1658–1673.

Maddison, D., Maddison, W., 2002. MacClade. Sinauer Associates, Sunderland.

Möller, M., Clokie, M., Cubas, P., Cronk, Q.C.B., 1999. Integrating molecular phylogenies and developmental genetics: a Gesneriaceae case study. In: Hollingsworth, P.M., Bateman, R.M., Gornall, R.J. (Eds.), Molecular Systematics and Plant Evolution. Taylor & Francis, London, pp. 375–402.

Moriyama, E.N., Petrov, D.A., Hartl, D.L., 1998. Genome size and intron size in *Drosophila*. Mol. Biol. Evol. 15, 770–773.

Mount, S.M., Burks, C., Hertz, G., Storme, G.D., White, O., Fields, C., 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucl. Acids Res. 20, 4255–4262.

Munster, T., Pahnke, J., Di Rosa, A., Kim, J.T., Martin, W., Saedler, H., Theisssen, 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.

Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

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.

Parsch, J., 2003. Selective constraints on intron evolution in *Drosophila*. Gen. Soc. Am. 165, 1843–1851.

Popp, M., Oxelman, B., 2004. Evolution of a RNA Polymerase Gene Family in *Silene* (Caryophyllaceae)—Incomplete Concerted Evolution and Topological Congruence Among Paralogues. Syst. Biol. 53, 914–932.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Phillipe, H., Sorhannus, U., Baroin, A., Perasso, R., Gasse, F., Adouette, A., 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. 7, 247–265.

Oh, S.-H., Potter, D., 2003. Phylogenetic utility of the second intron of *LEAFY* in *Neillia* and *Stephanandra* (Rosaceae) and implications for the origin of *Stephanandra*. Mol. Phylogenet. Evol. 29, 203–215.

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

Riechmann, J.L., Wang, M., Meyerowitz, E.M., 1996. DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. Nucl. Acids Res. 24, 3134–3141.

Rijpkema, A.S., Royaert, S., Zethof, J., vander 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.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 27, 121–147.

Sang, T., Donoghue, M.J., Zhang, D., 1997. Evolution of alcohol dehydrogenase genes in Peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. 14, 994–1007.

Schönenberger, J., Anderberg, A.A., Sytsma, K.J., 2005. Molecular phylogenetics and patterns of floral evolution in the Ericales. Int. J. Plant Sci. 166, 265–288.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

Stellari, G.M., Jaramillo, M.A., Kramer, E.M., 2004. Evolution of the APETALA3 and PISTILLATA lineages of MADS-box-containing genes in the basal angiosperms. Mol. Biol. Evol. 21, 506–519.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99, 16138–16143.

Swofford, D.L., 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland.

Theissen, G., Kim, J., Saedler, H., 1996. Classifiacation and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. Gene 156, 155–166.

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

Vandenbussche, M., Zethof, J., Royaert, S., Weterings, K., Gerats, T., 2004. The duplicated B-class heterodimer model: whorl-specific effects and complex genetic interactions in *Petunia hybrida* flower development. Plant Cell 16, 741–754.

Yang, Y., Fanning, L., Jack, T., 2003. The K domain mediates heterodimerization of the *Arabidopsis* floral organ identity proteins, APETALA3 and PISTILLATA. Plant J. 33, 47–60.

Yuan, Y.-M., Song, Y., Geuten, K., Rahelivololona, E., Wohlhauser, S., Fischer, E., Smets, E., Küpfer, P., 2004. Phylogeny and biogeography of Balsaminaceae inferred from ITS sequence data. Taxon 53, 391–403.

Zahn, L.M., Leebens-Mack, J., DePamphilis, C.W., 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.