Plant Syst Evol (2009) 277:75–84
DOI 10.1007/s00606-008-0105-0

ORIGINAL ARTICLE

Phylogenetic utility of *ycf*1 in orchids: a plastid gene more variable than *mat*K

Kurt M. Neubig · W. Mark Whitten · Barbara S. Carlsward · Mario A. Blanco · Lorena Endara · Norris H. Williams · Michael Moore

Received: 18 April 2008 / Accepted: 23 September 2008 / Published online: 12 November 2008

Abstract Plastid DNA sequences have been widely used by systematists for reconstructing plant phylogenies. The utility of any DNA region for phylogenetic analysis is determined by ease of amplification and sequencing, confidence of assessment in phylogenetic character alignment, and by variability across broad taxon sampling. Often, a compromise must be made between using relatively highly conserved coding regions or highly variable introns and intergenic spacers. Analyses of a combination of these types of DNA regions yield phylogenetic structure at various levels of a tree (i.e., along the spine and at the tips of the branches). Here, we demonstrate the phylogenetic utility of a heretofore unused portion of a plastid protein-coding gene, hypothetical chloroplast open reading frame 1 (*ycf*1), in orchids. All portions of *ycf*1 examined are highly variable, yet alignable across Orchidaceae, and are phylogenetically informative at the level of species. In Orchidaceae, *ycf*1 is more variable than *mat*K both in total number of parsimony informative characters and in percent variability. The nrITS region is more variable than *ycf*1, but is more difficult to align. Although we only demonstrate the phylogenetic utility of *ycf*1 in orchids, it is likely to be similarly useful among other plant taxa.

Keywords Chloroplast · nrITS · *mat*K · Orchidaceae · Phylogeny · Molecular systematics · *ycf*1

K. M. Neubig (✉) · M. A. Blanco · L. Endara
Department of Botany, University of Florida, Gainesville, FL 32611-8526, USA
e-mail: kneubig@flmnh.ufl.edu

K. M. Neubig · W. M. Whitten · L. Endara · N. H. Williams
Florida Museum of Natural History, University of Florida, P.O. Box 117800, Gainesville, FL 32611-7800, USA
e-mail: whitten@flmnh.ufl.edu

L. Endara
e-mail: lendara@flmnh.ufl.edu

N. H. Williams
e-mail: orchid@flmnh.ufl.edu

B. S. Carlsward
Department of Biological Sciences, Eastern Illinois University, 600 Lincoln Avenue, Charleston, IL 61920, USA
e-mail: bscarlsward@eiu.edu

M. A. Blanco
Jardín Botánico Lankester, Universidad de Costa Rica, Apdo. 1031-7050, Cartago, Costa Rica
e-mail: mblanco@flmnh.ufl.edu

M. Moore
Biology Department, Oberlin College, Science Center K111, 119 Woodland Street, Oberlin, OH 44074-1097, USA
e-mail: Michael.Moore@oberlin.edu

Introduction

Chloroplast DNA (cpDNA) sequences have been widely utilized by systematists for reconstructing plant phylogenies because of their ease of amplification and sequencing and because of their range of variability, providing useful phylogenetic characters (Soltis and Soltis 1998). However, relatively few chloroplast regions are commonly used for phylogenetic studies, although efforts have been made to discover more variable ones (Shaw et al. 2005, 2007). Often, a compromise must be made between using relatively conserved coding regions that are easily aligned versus highly variable introns or intergenic spacers that are more variable but often difficult to align. Combined analyses of these types of DNA regions frequently yield

phylogenetic structure at various levels of a tree. The numerous indels (insertions/deletions) in noncoding cpDNA make alignment challenging and subjective, especially at higher phylogenetic levels, with resultant problems of homology of nucleotide characters. Protein-coding genes are often easily aligned, but are usually more conserved and lack sufficient variation to resolve inter- and intraspecific relationships. For example, the most variable of the widely used plastid protein-coding genes, *mat*K, often provides few or no parsimony-informative sites between closely related species within orchid genera (personal observation). The variability, combined with the fact that *mat*K does not always maintain reading frame indicates that *mat*K is a pseudogene, at least in some orchid taxa (Whitten et al. 2000; Kocyan et al. 2008).

Comparative genomic studies have suggested that one putative protein-coding plastid gene, hypothetical *c*hloroplast open reading *f*rame 1 (*ycf*1) may be more variable than *mat*K (Timme et al. 2007). At approximately 5,500 bp, *ycf*1 represents the second longest reading frame in the plastid genome (only *ycf*2 is longer), and is present in nearly all plant plastid genomes sequenced to date (Raubeson and Jansen 2005). The function of the putative *ycf*1 protein is unknown. Nevertheless, Drescher et al. (2000) have demonstrated that *ycf*1 is essential to plant survival. The *ycf*1 reading frame is unusual among plastid genes in that it usually spans the boundary of the inverted repeat (IR) and the small-single copy (SSC) regions of the plastid genome (Raubeson and Jansen 2005). However, in the orchid genus *Phalaenopsis*, the entirety of *ycf*1 is found in the SSC region (Chang et al. 2006). The phylogenetic utility of *ycf*1 has only recently begun to be explored. The less variable IR portion of *ycf*1 has been included in phylogenetic analyses in one recent study (Jian et al. 2008), but the SSC portion of the gene has never been utilized phylogenetically, to our knowledge. Preliminary observations suggested that the SSC portion of *ycf*1 may be more variable than *mat*K, and thus potentially more valuable as a low-level phylogenetic marker. To test whether *ycf*1 could provide better resolution and support at higher taxonomic levels than *mat*K, we sequenced about 1,500 bp of the 3′ portion of *ycf*1 for 62 species of orchids. We then compared the phylogenetic resolution and clade support for *ycf*1-derived trees at multiple taxonomic levels of Orchidaceae to two other commonly used gene regions, the plastid *mat*K gene and the nuclear ribosomal internal transcribed spacer (ITS) region. Our results demonstrate that portions of *ycf*1 are relatively easy to amplify and align because of its conserved reading frame. Moreover, *ycf*1 possesses a high level of variability similar to or just below that of ITS, and thus provides superior resolution and support at lower taxonomic levels in Orchidaceae compared to *mat*K.

Materials and methods

Taxon sampling

Specimens were obtained from wild-collected and cultivated plants (Table 1). Taxa were chosen to represent a broad sampling at three different taxonomic levels of orchids: subfamily, genus, and species. For subfamily analyses, representatives of subfamilies Cypripedioideae, Orchidoideae, Epidendroideae, and Vanilloideae were used (sensu Chase et al. 2003). Vandeae (a tribe of Epidendroideae) were chosen to show relationships among closely related genera. *Sobralia* and *Elleanthus* (tribe Sobralieae) were chosen to show relationships among closely related species.

Extractions, amplification, and sequencing

Methods for DNA extraction and amplification of nrITS 1&2 and *mat*K are presented by Whitten et al. (2000). In *Phalaenopsis* (GenBank AY916449), the *ycf*1 open reading frame (ORF) is 5,451 bp in length. Because of its length, we did not attempt amplification of the entire region; instead, we sequenced an approximately 1,500-base pair (bp) portion from the 3′ end (Fig. 1) and a approximately 1,200-bp portion from the 5′ end. Primers were designed based on an alignment of complete *ycf*1 sequences from GenBank of *Phalaenopsis* and *Acorus*; initial primers were refined, as partial sequences of various Orchidaceae were obtained to find primers that amplified broadly across epidendroid orchids. Reaction components were as follows: 0.5–1.0 μL template DNA (~10–100 ng), 16.0–17.5 μL water, 2.5 μL 10× buffer, 2.0 μL of 25 mM $MgCl_2$, 0.5 μL of 10 μM dNTPs, 0.5 μL each of 10 μM primers and 0.5 units *Taq*. This region was amplified using a "touchdown" protocol with the following parameters: 94°C, 3 min; 8× (94°C, 30 s; 60–51°C, reducing 1°C per cycle, 1 min; 72°C, 3 min); 30× (94°C, 30 s; 50°C, 1 min; 72°C, 3 min); 72°C, 3 min, with amplimers 3720F (TAC GTA TGT AAT GAA CGA ATG G) and 5500R (GCT GTT ATT GGC ATC AAA CCA ATA GCG). Additional internal primers IntF (GAT CTG GAC CAA TGC ACA TAT T) and IntR (TTT GAT TGG GAT GAT CCA AGG) were also required for sequencing. Primers 1F (ATG ATT TTT AAA TCT TTT CTA CTA G) and 1200R (TTG TGA CAT TTC ATT GCG TAA AGC CTT) were used for the 5′ portion of *ycf*1 under the same PCR conditions.

Data analysis

Sequence data were edited and assembled using Sequencher 4.6™ (GeneCodes, Ann Arbor, MI, USA). All sequences were deposited in GenBank (Table 1) and data matrices are available upon request. Some data for nrITS and *mat*K were

Table 1 Species names, voucher information, and GenBank accession numbers for all taxa used in this study

Species	Voucher number	ITS	*mat*K	*ycf*1
Subfamily Cypripedioideae				
Paphiopedilum armeniacum S.C. Chen & F.Y. Liu	Whitten 3315 (FLAS)	None	EU490698	EU490759
Paphiopedilum delenatii Guillaumin	Whitten 3316 (FLAS)	None	EU490699	EU490760
Phragmipedium besseae Dodson & Kuhn	Whitten 2864 (FLAS)	None	EU490701	EU490764
Phragmipedium ecuadorense Garay	Whitten 2803 (FLAS)	None	AY918832	EU490765
Phragmipedium longifolium (Warsz. & Rchb. f.) Rolfe	Whitten 2802 (FLAS)	None	AY918831	EU490766
Phragmipedium schlimii (Linden ex Rchb. f.) Rolfe	Whitten 2865 (FLAS)	None	EU490702	EU490767
Selenipedium aequinoctiale Garay	Blanco 2475 (FLAS)	None	EU490707	EU490779
Subfamily Epidendroideae				
Aerangis citrata (Thouars) Schltr.	Whitten 1788 (FLAS)	DQ091600	DQ091337	EU490715
Aeranthes grandiflora Lindl.	Carlsward 238 (FLAS)	DQ091760	DQ091412	EU490716
Ancistrochilus rothschildianus O'Brien	Whitten 2847 (FLAS)	None	EU490675	EU490717
Ascocentrum christensonianum Haager	TBG145826 (*)	None	AB217708	None
Ascocentrum miniatum (Lindl.) Schltr.	Carlsward 273 (SEL)	DQ091678	None	EU490718
Basiphyllaea hamiltoniana J.D. Ackerman & W.M. Whitten	Whitten 99108 (FLAS)	None	EU490676	EU490720
Bifrenaria tyrianthina (Loudon) Rchb. f.	Whitten 3008 (FLAS)	None	DQ210752	EU490721
Bletia purpurea (Lam.) DC.	Whitten 3359 (FLAS)	None	EU490678	EU490722
Bletilla striata (Thunb. ex Murray) Rchb. f.	Neubig 192 (FLAS)	None	EU490679	EU490723
Bulbophyllum lobbii Lindl.	Chase 89007 (K)	None	AY121740	None
Bulbophyllum scaberulum (Rolfe) Bolus	Whitten 2925 (FLAS)	None	None	EU490724
Campylocentrum micranthum (Lindl.) Rolfe	Carlsward 180 (FLAS)	AF506298	AF506347	EU490725
Ceratostylis incognita J.T. Atwood & J. Beckner	Whitten 1993 (FLAS)	None	EU490680	EU490726
Chiloschista parishii Seidenf.	Carlsward 222 (FLAS)	DQ091733	None	EU490727
Chiloschista viridiflava Seidenf.	OR-2392002239 (*)	None	AB217719	None
Cryptopus paniculatus H. Perrier	Hermans 5392 (K)	DQ091588	DQ091327	EU490728
Dendrophylax sallei (Rchb. f.) Benth. ex Rolfe	Whitten 1945 (JBSD)	AY147225	AY147239	EU490730
Dichaea eligulata Folsom	Pupulin 1094 (USJ-L)	None	EU123625	EU123747
Dressleria dilecta (Rchb. f.) Dodson	Whitten 1019 (FLAS)	None	AF239507	EU490731
Elleanthus ampliflorus Schltr.	Blanco 2949 (FLAS)	EU490663	EU490682	EU490732
Elleanthus aurantiacus (Lindl.) Rchb. f.	Whitten 1611 (FLAS)	EU490664	EU490683	EU490733
Elleanthus caricoides Nash	Blanco 3106 (FLAS)	EU490665	EU490684	EU490734
Elleanthus conifer (Rchb. f. & Warsz.) Rchb. f.	Blanco 2527 (FLAS)	EU490666	EU490685	EU490735
Elleanthus cynarocephalus (Rchb. f.) Rchb. f.	Blanco 3105 (FLAS)	None	EU490686	EU490736
Elleanthus lancifolius C. Presl	Whitten 1575 (FLAS)	EU490667	EU490687	EU490737
Elleanthus oliganthus (Poepp. & Endl.) Rchb. f.	Whitten 1502 (FLAS)	EU490668	EU490688	EU490738
Elleanthus poiformis Schltr.	Blanco 3075 (FLAS)	EU490669	EU490689	EU490739
Elleanthus tricallosus Ames & C. Schweinf.	Blanco 2961 (FLAS)	EU490670	EU490690	EU490740
Encyclia guatemalensis (Klotzsch) Dressler & G.E. Pollard	Whitten 3372 (FLAS)	None	EU490691	EU490741
Epipactis helleborine (L.) Crantz	Whitten 3326 (FLAS)	None	EU490692	EU490742
Eriopsis biloba Lindl.	Whitten 3327 (FLAS)	None	EU490693	EU490743
Erycina hyalinobulbon (La Llave & Lex.) N.H. Williams & M.W. Chase	Chase 83395 (K)	None	AF350615	EU490744
Eulophia guineensis Lindl.	Whitten 99029 (FLAS)	None	AF239509	EU490745
Govenia sodiroi Schltr.	Whitten 2682 (FLAS)	None	EU490695	EU490747
Inti chartacifolia (Ames & C. Schweinf.) M.A. Blanco	Whitten 1597 (FLAS)	None	DQ209942	EU490750
Isochilus major Schltdl. & Cham.	Whitten 3320 (FLAS)	None	EU490696	EU490749
Microcoelia aphylla (Thouars) Summerh.	Carlsward 341 (FLAS)	DQ091651	DQ091400	EU490751
Microcoelia exilis Lindl.	Whitten 1937 (FLAS)	DQ091658	DQ091406	EU490752
Mystacidium aliceae Bolus	Whitten 1787 (FLAS)	DQ091571	DQ091360	EU490753

Table 1 continued

Species	Voucher number	ITS	*mat*K	*ycf*1
Neomoorea wallisii (Rchb. f.) Schltr.	Whitten 3010 (FLAS)	None	DQ210743	EU490754
Odontoglossum harryanum Rchb. f.	Chase 86165 (K)	None	AF350648	EU490755
Oeoniella polystachys (Thouars) Schltr.	Carlsward 221 (FLAS)	DQ091736	DQ091432	EU490756
Palmorchis powellii (Ames) C. Schweinf. & Correll	Vargas 2115 (INB)	None	EU490697	EU490757
Paphinia clausula Dressler	Whitten 3600 (FLAS)	None	None	EU490758
Paphinia neudeckeri Jenny	Whitten 88041 (FLAS)	None	AF239471	None
Peristeria elata Hook.	Whitten 90158 (FLAS)	None	AF239442	EU490761
Phaius tankervilliae (Banks ex L'Hér.) Blume	Neubig 193 (FLAS)	None	EU490700	EU490762
Phalaenopsis wilsonii Rolfe	Carlsward 331 (FLAS)	DQ091672	None	EU490763
Phalaenopsis wilsonii Rolfe	TBG144214 (*)	None	AB217751	None
Pleione formosana Hayata	Whitten 3364 (FLAS)	None	EU490703	EU490768
Polycycnis gratiosa Endres & Rchb. f.	Whitten 93178 (FLAS)	None	AF239469	EU490769
Polystachya modesta Rchb. f.	Carlsward 219 (SEL)	DQ091562	DQ091313	EU490770
Rangaeris muscicola (Rchb. f.) Summerh.	Carlsward 169 (SEL)	DQ091630	DQ091387	EU490774
Rhipidoglossum xanthopollinium (Rchb. f.) Schltr.	Carlsward 384 (FLAS)	DQ091582	DQ091370	EU490775
Rudolfiella saxicola (Schltr.) Hoehne	Whitten 97020 (FLAS)	None	AY870011	EU490776
Scaphosepalum rapax Luer	Endara 1502 (FLAS)	None	EU490705	EU490777
Scaphyglottis amparoana (Schltr.) Dressler	Whitten 2640 (FLAS)	None	EU490706	EU490778
Sobennikoffia humbertiana H. Perrier	Carlsward 304 (FLAS)	DQ091750	DQ091433	EU490780
Sobralia bouchei Ames & C. Schweinf.	Blanco 3000 (FLAS)	EU490671	EU490708	EU490781
Sobralia crocea (Poepp. & Endl.) Rchb. f.	Whitten 1578 (FLAS)	EU490672	EU490709	EU490782
Sobralia warszewiczii Rchb. f.	Blanco 2676 (FLAS)	EU490673	EU490710	EU490783
Soterosanthus shepheardii (Rolfe) Jenny	Dodson 18580-3 (FLAS)	None	AF239457	EU490784
Stanhopea annulata Mansf.	Whitten 87242 (FLAS)	None	AF239444	EU490786
Stanhopea tigrina Bateman ex Lindl.	Whitten 93122 (FLAS)	None	AF239448	EU490787
Tipularia discolor (Pursh) Nutt.	Whitten 3288 (FLAS)	None	EU490712	EU490789
Trichocentrum tigrinum Linden & Rchb. f.	Chase 83439 (K)	None	EU490713	EU490790
Trichoglottis atropurpurea Rchb. f.	Carlsward 173 (FLAS)	DQ091713	DQ091316	EU490791
Trichopilia sanguinolenta (Lindl.) Rchb. f.	Chase 84547 (K)	None	AF350659	EU490792
Tropidia polystachya (Sw.) Ames	Whitten 2830 (FLAS)	EU490674	EU490714	EU490793
Warczewiczella marginata Rchb. f.	Whitten 1865 (FLAS)	None	AY869958	EU490794
Warrea warreana (Lodd. ex Lindl.) C. Schweinf.	Whitten 1752 (FLAS)	None	EU123675	EU123798
Xylobium pallidiflorum (Hook.) G. Nicholson	Whitten 1876 (FLAS)	None	AF239434	EU490795
Zygopetalum maxillare Lodd.	Whitten 94103 (FLAS)	None	EU123676	EU123799
Subfamily Orchidoideae				
Baskervilla sp.	Whitten 2783 (FLAS)	None	EU490677	EU490719
Cyclopogon sp.	Trujillo 388 (HURP)	None	EU490681	EU490729
Gomphichis sp.	Trujillo 379 (HURP)	None	EU490694	EU490746
Habenaria repens Nutt.	Neubig 217 (FLAS)	None	None	EU490748
Habenaria repens Nutt.	Chase 89124 (K)	None	AJ310036	None
Ponthieva racemosa (Walter) C. Mohr	Salazar 6049 (MEXU)	None	AJ543936	None
Ponthieva sp.	Trujillo 332 (HURP)	None	None	EU490771
Prescottia aff. *oligantha* (Sw.) Lindl.	da Silva 861 (*)	None	AJ519449	None
Prescottia oligantha (Sw.) Lindl.	Whitten 3314 (FLAS)	None	None	EU490772
Pterichis sp.	Trujillo 386 (HURP)	None	EU490704	EU490773
Spiranthes vernalis Engelm. & A. Gray	Neubig 194 (FLAS)	None	EU490711	EU490785
Stenoptera ecuadorana Dodson & C. Vargas	Salazar 6357 (K)	None	AJ543940	None

Table 1 continued

Species	Voucher number	ITS	*mat*K	*ycf*1
Stenoptera sp.	Trujillo 389 (HURP)	None	None	EU490788

Vouchers are deposited at the following herbaria: Florida Museum of Natural History Herbarium (FLAS); Instituto Nacional de Biodiversidad, Costa Rica (INB); Herbario Jardin Botanico Nacional Dr. Rafael M. Moscoso, Dominican Republic, (JBSD); Royal Botanic Garden, Kew, UK (K); Herbario Universidad Nacional Autonoma de Mexico (MEXU); Herbarium Marie Selby Botanical Garden, Florida, USA (SEL); Herbario Universidad Ricardo Palma, Peru (HURP); Herbarium Jardin Botanico Lankester, Costa Rica (USJ-L)

Voucher information is unavailable for sequences downloaded from GenBank and is indicated by an asterisk (*)

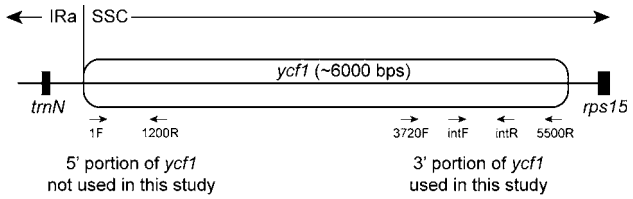

Fig. 1 Relative position of *ycf*1 in the small single copy (SSC) region to the inverted repeat (IRa) in the chloroplast as found in *Phalaenopsis aphrodite*. Only the downstream (3′) portion of this gene was used in this study. Primers are indicated with *small arrows*

compiled from sequences deposited in GenBank from previous phylogenetic studies, supplemented with a few new sequences. Sequence data were automatically aligned using ClustalX in MacClade (Maddison and Maddison 2000) and then manually aligned using Se-Al v2.0a11 (Rambaut 1996). All characters were unordered and weighted equally. Missing data were coded as "?," gaps were coded as "-," and nucleotides of ambiguous identity were coded as "N." No sequence data were excluded from analyses. Analyses were performed using PAUP*4.0b10 (Swofford 1999) with Fitch parsimony (Fitch 1971). A heuristic search strategy consisted of branch swapping by tree bisection and reconnection (TBR), stepwise addition with 5,000 random-addition replicates holding five trees at each step, and saving multiple trees (MULTREES). Levels of support were assessed using bootstrap values, estimated with 1,000 bootstrap replicates, using TBR algorithm for branch swapping for five random-addition replicates per bootstrap replicate. Parsimony searches were used in lieu of other methods (e.g., maximum likelihood, Bayesian, or distance) to provide simple comparisons of sequence variability and branch lengths.

Gaps in the *ycf*1 and *mat*K subfamilial-level matrices were coded using PAUPGAP (Cox 1997) with simple gap coding (Simmons and Ochoterena 2000). Matrices of other regions and other taxa contained too few gaps to be phylogenetically useful.

Results

Amplification of the 3′ portion of *ycf*1 was highly consistent and reliable among taxa with the exception of two species of *Vanilla* (*V. barbellata* Rchb. f. and *V. odorata* C. Presl). *Pogonia ophioglossoides* (L.) Ker Gawl., also a member of subfamily Vanilloideae, was amplified and sequenced successfully (data not included in these analyses). The *ycf*1 sequence for *Pogonia* was significantly shorter than other orchids examined (~380 bp), but still gave congruent phylogenetic signal with *mat*K (results not shown). Bootstrap consensus trees and phylograms for subfamily-level analyses of *mat*K and *ycf*1 are presented in Fig. 2. Phylograms comparing ITS, *mat*K, and *ycf*1 for tribes Vandeae (genus-level analyses) and Sobralieae (species-level analyses) are presented in Fig. 3. We used gaps as phylogenetic characters (for the subfamilial-level analyses only) to examine their utility. Gap characters in *ycf*1 were highly informative [94 total gaps, of which 51 were parsimony-informative; consistency index (CI) = 0.65, retention index (RI) = 0.87, tree length (L) = 144; tree not shown] compared to *mat*K (12 gaps total, of which three were parsimony-informative; CI = 1, RI = 1, L = 12; tree not shown). Substitution rates for the three codon positions in *ycf*1 parallel those of *mat*K (Whitten et al. 2000) as nonsynonymous substitutions are surprisingly high (Table 2).

All analyses show that *ycf*1 is more variable than *mat*K, one of the most widely used plastid coding regions (Table 3). Variability in *ycf*1 ranges approximately from two to four times that of *mat*K in terms of parsimony-informative characters. In the intrafamilial analysis of orchids, *ycf*1 was substantially more variable than *mat*K both in total number of parsimony-informative characters (PICs) and percent variability. The ITS region is more variable and yielded more PICs than either *mat*K or *ycf*1 in the species-level analysis of *Elleanthus* and *Sobralia*. However, in the analysis of tribe Vandeae, *ycf*1 yielded more PICs and a longer tree than ITS and *mat*K. Minor incongruence exists among data sets in our genus-level (Carlsward et al. 2006) and species-level (*Sobralia* and *Elleanthus*) analyses, but lack strong bootstrap support. Incongruence is common when comparing multiple datasets and can be caused by many different biological, experimental, or analytical reasons (Johnson and Soltis 1998; Buckley et al. 2001).

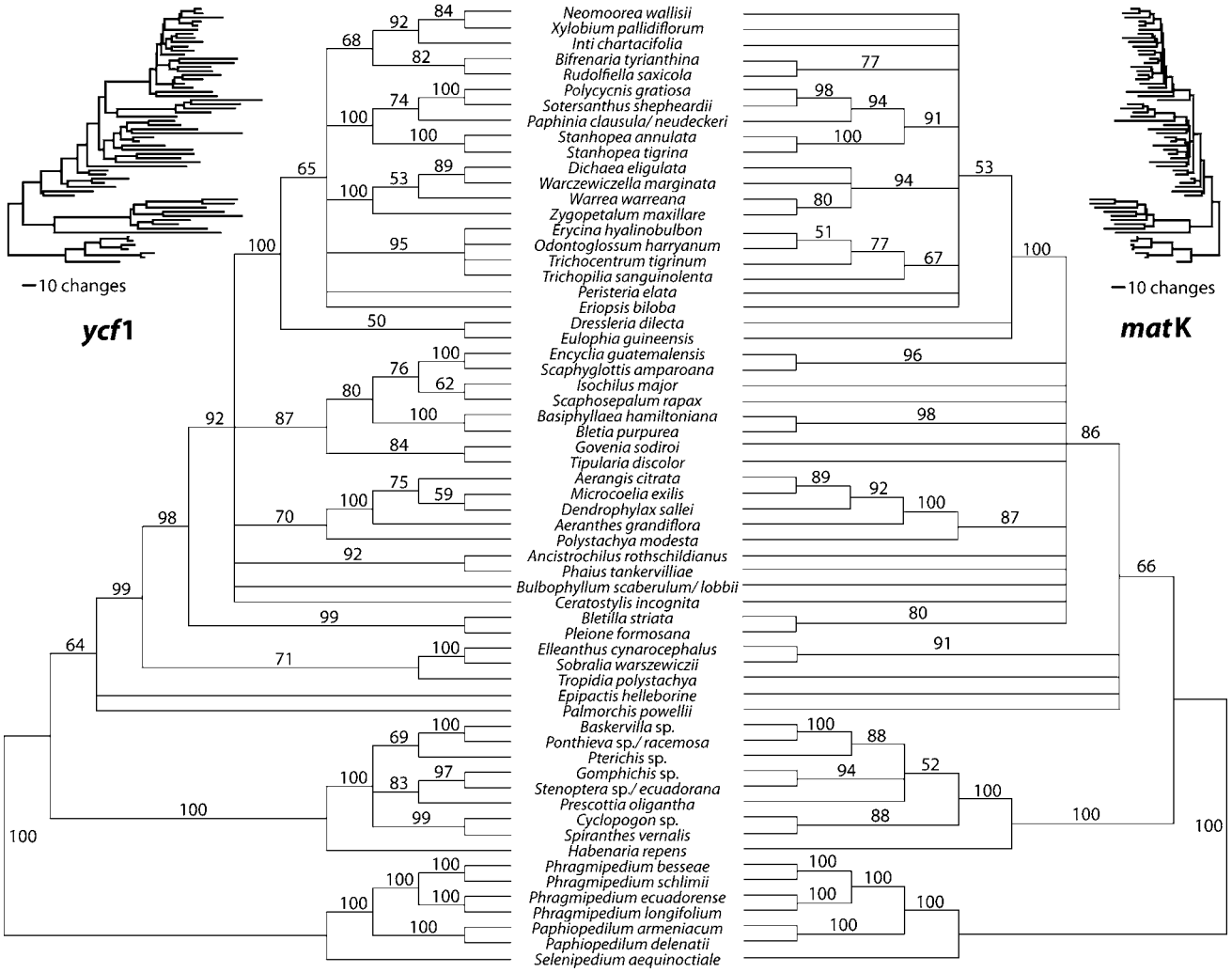

Fig. 2 Comparison of bootstrap consensus trees obtained with the analysis of *ycf*1 (*left*) and *mat*K (*right*) for a broad sampling of orchid taxa; bootstrap support values higher than 50% are indicated above branches. Scaled phylograms obtained from parsimony searches are shown in the *upper corners*, demonstrating the relative branch lengths for each

Discussion

Subfamilial-level analysis

Many publications have assessed taxonomic relationships within the orchid family using various DNA regions (Chase et al. 2003; Cameron 2004; Freudenstein et al. 2004). However, these data sets have produced phylogenetic trees with low resolution in part, because their phylogenetic markers have low divergence rates (e.g., *rbc*L, *atp*B, *psa*B, *ndh*F, and to a lesser extent *mat*K).

Direct comparison of *ycf*1 to *mat*K shows that *ycf*1 is substantially more variable in orchids (Fig. 2). A similar result was obtained when comparing sequence regions between the plastid genomes of *Helianthus* and *Lactuca* (Asteraceae); *ycf*1 was almost twice as variable as *mat*K (Timme et al. 2007). The *mat*K region has been shown to be among the most variable protein-coding plastid DNA regions (providing the most phylogenetic characters) and thus has frequently been used in phylogenetic analyses. Sequence divergence has been demonstrated to be greater in *mat*K than in many other coding regions, such as *rbc*L, with more strongly supported relationships at deeper taxonomic levels (Muller et al. 2006).

The higher variation in *ycf*1 allows recovery of several topologies in orchids that previously have only been resolved when multiple plastid gene regions have been combined (Cameron 2002). For example, the sister relationship of Neottieae (including *Palmorchis* and *Epipactis*) to the rest of Epidendroideae, followed by Tropidieae (including *Tropidia*) and Sobralieae (including *Sobralia* and *Elleanthus*), has only been recovered when multiple gene regions are combined. The sister relationship of Arethuseae (including *Bletilla* and *Pleione*) to the

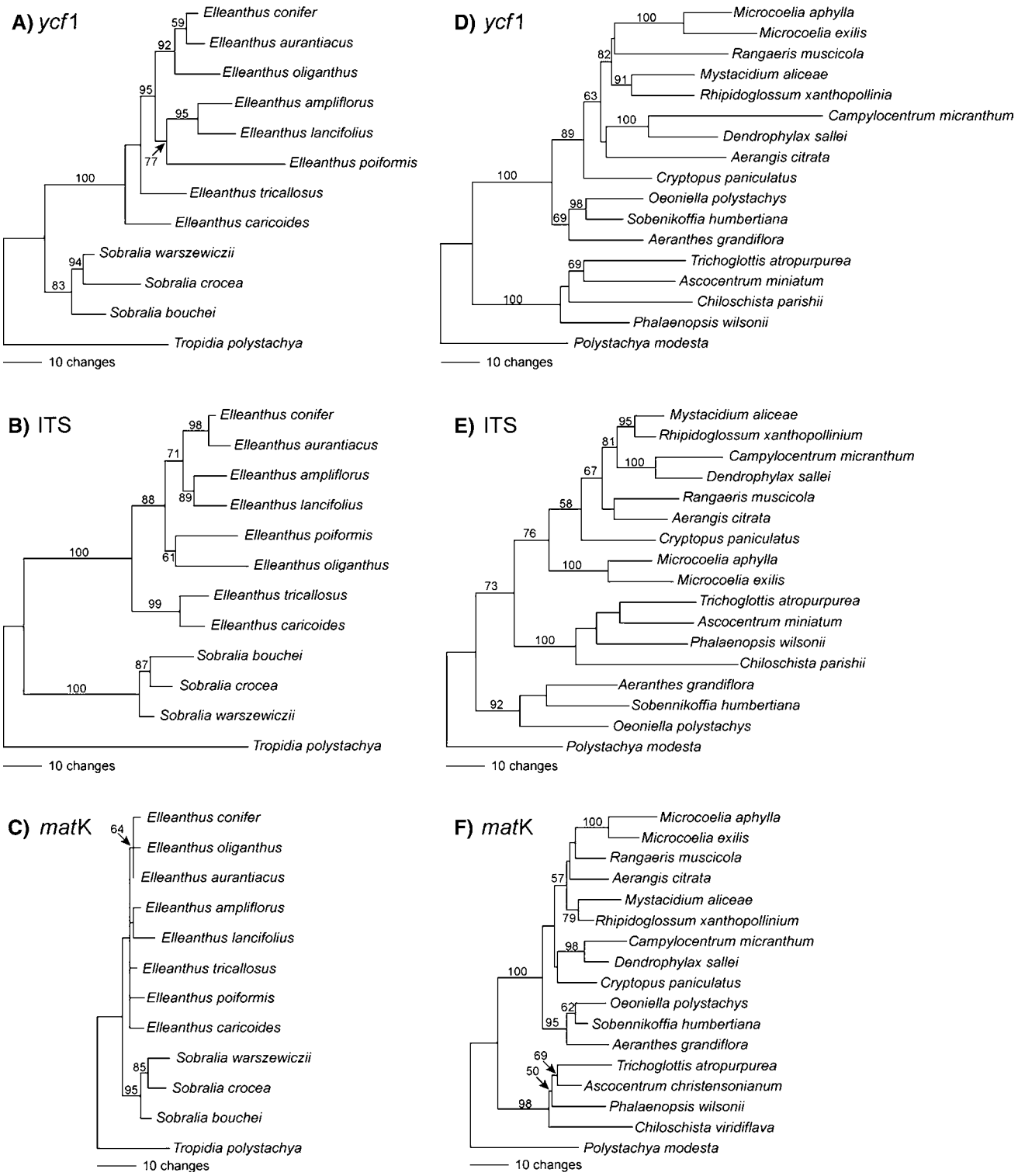

Fig. 3 Comparison of phylograms (with bootstrap percentages higher than 50% indicated) using three gene regions for tribes Sobralieae (*left*) and Vandeae (*right*); *ycf*1 (*upper row*), ITS nrDNA (*middle row*), and *mat*K (*bottom row*)

remainder of Epidendroideae (to the exclusion of the previously mentioned taxa) also illustrates the power of *ycf*1 compared to previously published phylogenies using other gene regions. Additionally, *ycf*1 recovers relationships among Epidendreae (including *Encyclia*, *Scaphyglottis*, *Isochilus*, *Scaphosepalum*, *Basiphyllaea*, and *Bletia*), a taxonomic group with notoriously poor sequence divergence (van den Berg et al. 2005). The monophyly of

Table 2 Statistical information on molecular change (substitutions) for each of the data sets used in this study

Data set	First codon position	Second codon position	Third codon position	Transitions/ transversions	A	C	G	T
Subfamily-level *ycf*1	980	737	1,034	1,140/1,186	0.427	0.132	0.138	0.303
Subfamily-level *mat*K	438	391	658	668/675	0.308	0.161	0.152	0.379
Genus-level (Vandeae) ITS	–	–	–	304/149	0.198	0.295	0.341	0.166
Genus-level (Vandeae) *mat*K	124	106	97	100/111	0.306	0.162	0.149	0.384
Genus-level (Vandeae) *ycf*1	232	186	257	268/288	0.428	0.127	0.142	0.303
Species-level (*Elleanthus*) ITS	–	–	–	174/56	0.238	0.258	0.308	0.197
Species-level (*Elleanthus*) *mat*K	21	15	38	20/30	0.300	0.171	0.154	0.375
Species-level (*Elleanthus*) *ycf*1	84	72	75	82/124	0.421	0.136	0.147	0.296

Nucleotide composition is based on all characters (with missing data and gaps excluded)

Table 3 Quantitative data collected in this study on the parsimony analyses performed

Data set	Aligned length (bp)	Total parsimony-informative characters (PICs)	% Variability	Tree length	CI	RI	Total number of most parsimonious trees (MPTs)	Number of strongly supported clades (>79% bootstrap)
Subfamily-level *ycf*1	1,908	630	53.5	2,751	0.541	0.720	48	33
Subfamily-level *mat*K	1,341	351	43.1	1,487	0.531	0.696	19	28
Genus-level (Vandeae) ITS	735	153	40.3	571	0.662	0.566	2	6
Genus-level (Vandeae) *mat*K	1,349	85	17.1	327	0.807	0.734	12	5
Genus-level (Vandeae) *ycf*1	1,761	174	25.9	675	0.806	0.702	2	8
Species-level (*Elleanthus*) ITS	842	102	25.1	277	0.845	0.800	1	7
Species-level (*Elleanthus*) *mat*K	1,342	16	4.8	74	0.905	0.720	6	2
Species-level (*Elleanthus*) *ycf*1	1,650	68	11.3	231	0.866	0.791	3	6

Calypsoeae (including *Tipularia* and *Govenia*) and the sister relationship of that tribe to the aforementioned Epidendreae have only been recovered with extensive combined gene analyses, but is also recovered by *ycf*1 alone. Within Cymbidieae (top of Fig. 2, from *Neomoorea* down to *Eulophia*), *ycf*1 also indicates the monophyly of subtribes Oncidiinae (represented by *Erycina*, *Odontoglossum*, *Trichocentrum*, and *Trichopilia*), Zygopetalinae (represented by *Zygopetalum*, *Warrea*, *Warczewiczella*, and *Dichaea*), Stanhopeinae (represented by *Stanhopea*, *Paphinia*, *Soterosanthus*, and *Polycycnis*), and (to a lesser degree) Maxillariinae (represented by *Rudolfiella*, *Bifrenaria*, *Inti*, *Xylobium*, and *Neomoorea*); however, the relationships among these subtribes remain poorly resolved (Whitten et al. 2000).

Genus- and species-level analyses

One of the most challenging aspects of plant molecular systematics is finding DNA markers that are variable enough to provide resolution among genera and species. For various historical and practical reasons, *mat*K and ITS are among the most commonly used DNA markers. However, *mat*K is often not variable enough to provide a satisfactory number of phylogenetically informative characters, especially at lower taxonomic levels. Our data demonstrate that *ycf*1 performs better than *mat*K at the genus and species level in terms of both variability and strongly supported topologies. In contrast, *ycf*1 is not more variable (in percentage) than ITS in any data set. However, the ease of alignment and the higher number of characters afforded by *ycf*1 may outweigh the higher percentage of variable characters in ITS.

In the analysis of Vandeae (Table 3, Fig. 3), *ycf*1 produced more PICs, and more strongly supported clades than either ITS or *mat*K. All markers give a well-supported Aeridinae (*Ascocentrum*, *Chilochista*, *Phalaenopsis*, *Trichoglottis*; bootstrap of 98–100%), but the subtribe's position differs between the nrDNA and cpDNA data sets, perhaps because of paralogy. Of the chloroplast data sets, *ycf*1 shows greater sequence divergence and a better-supported spine than in most of the *mat*K tree.

The monophyly of *Sobralia* and *Elleanthus* is strongly supported by both ITS and *ycf*1 (Fig. 3). In contrast, *mat*K has remarkably low sequence divergence with very poor support throughout the tree, but does support the monophyly of *Sobralia*. Among species of *Elleanthus*, morphological features of inflorescence structure support

the topology recovered in *ycf*1 over that of ITS (unpublished data). The regions *ycf*1 and ITS produced similar numbers of strongly supported clades, despite *ycf*1 having slightly fewer PICs. Additional analyses of relationships within *Dichaea* and *Scaphosepalum*, and various genera of subtribe Oncidiinae show similar trends of variability in the *ycf*1 gene (unpublished data).

Implications of this study

Levels of variation in first, second, and third codon positions are nearly equal in *ycf*1, as in *mat*K (Table 2). As a result, there is no synonymous substitution bias as is found in most protein-coding DNA regions. This is surprising, because *ycf*1 is an essential gene for many plants (Drescher et al. 2000), as supported by the presence of *ycf*1 in almost all plant lineages (Raubeson and Jansen 2005), except in some grasses, which are known to lack both *ycf*1 and *ycf*2 in their plastid genomes (Asano et al. 2004; Chang et al. 2006). Although levels of variation are not equal among every nucleotide position in *ycf*1 in orchids (Fig. 4), there are no distinct regions of hypervariability such as those seen in ITS (Baldwin et al. 1995; Whitten et al. 2000). In *Panax*, *ycf*1 exhibits relatively long indels associated with short direct repeats (Kim and Lee 2004) resulting from illegitimate recombination events that have been observed in several plastid genomes (Ogihara et al. 1988; Milligan et al. 1989; Shimada and Sugiura 1989). Many indels were found in *ycf*1 of orchids, but they were dissimilar in that the indels were usually relatively short repeats of adjacent nucleotides.

Other portions of *ycf*1, other than the 3′ portion shown in this study, may also hold promise for orchid phylogenetics. Preliminary (unpublished) data using ~1,200 bp of the 5′ portion of the *ycf*1 gene (Fig. 1) show some potential for resolving orchid relationships. However, with limited sampling, we have found mixed phylogenetic results. In members of the Oncidiinae, the 5′ portion of *ycf*1 seems highly variable as in the 3′ portion presented in this article. However, broader phylogenetic sampling among orchids has shown lower variability in the 5′ portion of *ycf*1, which is consistent with the usual position of this region of the gene within the inverted repeat of many nonorchid plant groups. The lower variability of the 5′ IR portion of *ycf*1 in other plant groups enables relatively easy alignment across angiosperms (including *Phalaenopsis*), whereas in the SSC portion of *ycf*1 (including the 3′ portion used in this study), alignment of many regions of the gene is impossible across angiosperms (M. Moore et al., unpublished data). Although the entirety of *ycf*1 in orchids lies within the SSC region (Chang et al. 2006), our data suggest that the 5′ region of *ycf*1 retains this lower level of variation in orchids, thus reducing its usefulness as a marker at family-level phylogenetic analysis.

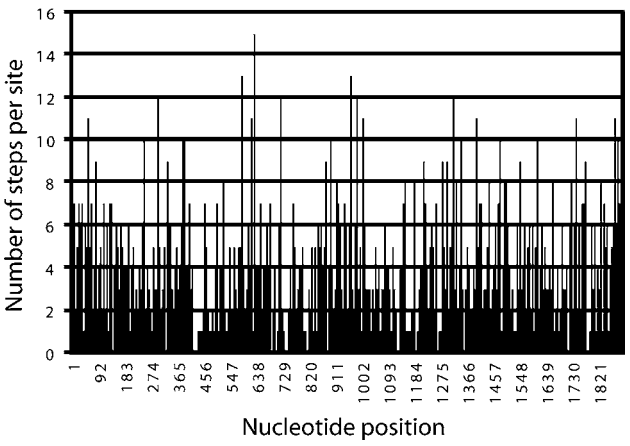

Fig. 4 Histogram showing number of steps per site for the 3′ portion of *ycf*1 (see Fig. 2) based on a single, randomly chosen most parsimonious tree for subfamily analyses

Our results indicate that *ycf*1 has great phylogenetic utility in orchids and potentially in other plant groups. It is variable at very low and high taxonomic levels, but alignment difficulties may preclude its use in extensive interfamilial phylogenetic analyses. In orchids, *ycf*1 amplifies and sequences reliably (with the exception of the two species of *Vanilla* assayed in this study). Although primer design for *ycf*1 can be challenging due to the large number of indels, it appears to be an optimal choice as a phylogenetic marker among orchids and probably other groups of higher plants. The entire coding portion of *ycf*1 is 5,451 bp in *Phalaenopsis aphrodite* (Chang et al. 2006); so, sequencing of the entire gene for large numbers of species may prove difficult due to numerous indels and homopolymer stutter regions. However, the growing number of entire chloroplast genome DNA sequences may allow identification of conserved regions that will be useful for primer design. Primer design and subsequent PCR is likely to be most successful when customized within families.

Acknowledgments Portions of this research were funded by the 11th World Orchid Conference Fellowship (University of Florida, for K.N.), by the US National Science Foundation grant no. DEB-234064 for the project *Systematics of Maxillariinae (Orchidaceae): generic delimitation, pollinator rewards, and pollination*, and by grant no. IOB-0543659 for the project *Mechanisms of the evolutionary origins of Crassulacean acid metabolism in Tropical Orchids*. We also thank the American Orchid Society for funding of *Molecular systematics of the neotropical Sobralieae: parting the reeds of Sobralia and relatives*.

References

Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K (2004) Complete nucleotide sequence of the sugarcane (*Saccharum*

officinarum) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA Research 11:93–99

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann Missouri Bot Gard 82:247–277

Buckley TR, Simon C, Shimodaira H, Chambers GK (2001) Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. Molec Biol Evol 18:223–234

Cameron K (2002) Molecular systematics of Orchidaceae: a literature review and an example using five plastid genes. In: Nair H (ed) Proceedings of the 17th World Orchid Conference. Shah Alam, Malaysia

Cameron KM (2004) Utility of plastid *psa*B gene sequences for investigating intrafamilial relationships within Orchidaceae. Molec Phylogenet Evol 31:1157–1180

Carlsward BS, Whitten WM, Williams NH, Bytebier B (2006) Molecular phylogenetics of Vandeae (Orchidaceae) and the evolution of leaflessness. Amer J Bot 93:770–786

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Molec Biol Evol 23:279–291

Chase MW, Freudenstein JV, Cameron KM, Barrett RL (2003) DNA data and Orchidaceae systematics: a new phylogenetic classification. In: Dixon KW, Kell SP, Barrett RL, Cribb PJ (eds) Orchid conservation. Natural History Publications, Kota Kinabalu, pp 69–89

Cox AV (1997) PaupGap version 1.0: program and documentation. Royal Botanical Gardens, Kew

Drescher A, Ruf S, Calsa T Jr, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Pl J 22:97–104

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

Freudenstein JV, van den Berg C, Goldman DH, Kores PJ, Molvray M, Chase MW (2004) An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. Amer J Bot 91:149–157

Jian S, Soltis PS, Gitzendanner MA, Moore MJ, Li R, Hendry TA, Qiu Y-L, Dhingra A, Bell CD, Soltis DE (2008) Resolving an ancient, rapid radiation in Saxifragales. Syst Biol 57:38–57

Johnson LA, Soltis DE (1998) Assessing congruence: empirical examples from molecular data. In: Soltis DE, Soltis PS, Doyle JJ (eds) Molecular systematics of plants II: DNA sequencing. Kluwer, Boston, pp 297–348

Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Research 11:247–261

Kocyan A, de Vogel EF, Gravendeel B (2008) Molecular phylogeny of *Aerides* (Orchidaceae) based on one nuclear and two plastid markers: a step forward in understanding the evolution of the Aeridinae. Molec Phylogenet Evol 48:422–443

Maddison DR, Maddison WP (2000) MacClade 4: analysis of phylogeny and character evolution. Version 4.06. Sinauer Associates, Sunderland

Milligan BG, Hampton JN, Palmer JD (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molec Biol Evol 6:355–368

Muller KF, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *mat*K, *trn*T-F, and *rbc*L in basal angiosperms. Molec Phylogenet Evol 41:99–117

Ogihara Y, Terachi T, Sasakuma T (1988) Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc Natl Acad Sci USA 85:8573–8577

Rambaut A (1996) Se-Al: sequence alignment editor, v2.0a11. University of Oxford, Oxford

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ (ed) Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CABI Publishing, Cambridge, pp 45–68

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Amer J Bot 92:142–166

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Amer J Bot 94:275–288

Shimada H, Sugiura M (1989) Pseudogenes and short repeated sequences in the rice chloroplast genome. Curr Genet 16:293–301

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. Syst Biol 49:369–381

Soltis DE, Soltis PS (1998) Choosing an approach and an appropriate gene for phylogenetic analysis. In: Soltis DE, Soltis PS, Doyle JJ (eds) Molecular systematics of plants II: DNA sequencing. Kluwer, Boston, pp 1–42

Swofford DL (1999) PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland

Timme RE, Kuehl JV, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Amer J Bot 94:302–312

van den Berg C, Goldman DH, Freudenstein JV, Pridgeon AM, Cameron KM, Chase M (2005) An overview of the phylogenetic relationships within Epidendroideae inferred from multiple DNA regions and recircumscription of Epidendreae and Arethuseae (Orchidaceae). Amer J Bot 92:613–624

Whitten MW, Williams NH, Chase MW (2000) Subtribal and generic relationships of Maxillarieae (Orchidaceae) with emphasis on Stanhopeinae: combined molecular evidence. Amer J Bot 87:1842–1856