J Mol Evol
DOI 10.1007/s00239-008-9083-7

# Selecting Single-Copy Nuclear Genes for Plant Phylogenetics: A Preliminary Analysis for the Senecioneae (Asteraceae)

**Inés Álvarez · Andrea Costa · Gonzalo Nieto Feliner**

Received: 11 March 2007 / Accepted: 25 January 2008


**Abstract** Compared to organelle genomes, the nuclear genome comprises a vast reservoir of genes that potentially harbor phylogenetic signal. Despite the valuable data that sequencing projects of model systems offer, relatively few single-copy nuclear genes are being used in systematics. In part this is due to the challenges inherent in generating orthologous sequences, a problem that is ameliorated when the gene family in question has been characterized in related organisms. Here we illustrate the utility of diverse sequence databases within the Asteraceae as a framework for developing single-copy nuclear genes useful for inferring phylogenies in the tribe Senecioneae. We highlight the process of searching for informative genes by using data from *Helianthus annuus*, *Lactuca sativa*, *Stevia rebaudiana*, *Zinnia elegans*, and *Gerbera* cultivar. Emerging from this process were several candidate genes; two of these were used for a phylogenetic assessment of the Senecioneae and were compared to other genes previously used in Asteraceae phylogenies. Based on the preliminary sampling used, one of the genes selected during the searching process was more useful than the two previously used in Asteraceae. The search strategy described is valid for any group of plants but its efficiency is dependent on the phylogenetic proximity of the study group to the species represented in sequence databases.

**Keywords** Single-copy nuclear genes · Phylogenetic markers · Cellulose synthase · Chalcone synthase · Deoxyhypusine synthase · Asteraceae · Senecioneae

## Introduction

Over the last two decades molecular data have become the most powerful and versatile source of information for revealing the evolutionary history among organisms (Van de Peer et al. 1990; Chase et al. 1993; Van de Peer and De Wachter 1997; Baldauf 1999; Mathews and Donoghue 1999; Soltis et al. 1999; Graham and Olmstead 2000; Brown 2001; Nozaki et al. 2003; Schlegel 2003; Hassanin 2006). In most cases, however, only a few molecular markers are employed for phylogeny reconstruction; in plants, for example, the predominant tools are chloroplast genes and multicopy rDNA genes and spacers such as ITS (Álvarez and Wendel 2003). Because of the limitations inherent in cpDNA and rDNA markers, and because of the enormous phylogenetic potential of single-copy nuclear genes, the latter are increasingly being used in systematic studies (Strand et al. 1997; Hare 2001; Sang 2002; Zhang and Hewitt 2003; Mort and Crawford 2004; Small et al. 2004; Schlüter et al. 2005). Among the main advantages of single-copy nuclear genes are (1) biparental inheritance; (2) co-occurrence of introns and exons within the same gene, yielding characters that evolve at different rates thus can provide phylogenetic signal at different levels; and (3) the very large number of independent markers. This potential has yet to be fully realized, in part because developing single-copy nuclear genes requires previously generated sequence information from related groups. When sequence availability is high (e.g., from genomic libraries or sequencing projects of closely related taxa), it may be

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-008-9083-7) contains supplementary material, which is available to authorized users.

I. Álvarez (✉) · A. Costa · G. N. Feliner
Real Jardín Botánico de Madrid, CSIC, Plaza de Murillo, 2, 28014 Madrid, Spain
e-mail: ines@rjb.csic.es

possible to screen thousands of sequences for potential use through comparisons with homologous sequences in other taxa (Fulton et al. 2002). Here we use this approach and the recommendations of Small et al. (2004) to design a selection strategy for identifying single-copy nuclear genes of potentially phylogenetic value in the tribe Senecioneae (Asteraceae). There is a recently published study pursuing similar aims, although it establishes different criteria for selection of genes (Wu et al. 2006).

Senecioneae is the largest tribe (~3000 species and ~150 genera) of one of the largest families of seed plants (Asteraceae), yet relative to the remaining tribes, it is rather poorly known from a systematic point of view. All molecular phylogenetic analyses of the Senecioneae are based on chloroplast markers (Jansen et al. 1990, 1991; Kim et al. 1992; Kim and Jansen 1995; Kadereit and Jeffrey 1996), and only a small portion of the tribe is represented. Currently several teams are collaborating to analyze available Senecioneae sequences (for about 600 species representing 115 genera) of several chloroplast markers (i.e., *ndhF*, *psbA-trnH*, *trnK*, *trnT-L*, *trnL*, and *trnL-F*) plus the ITS region of the nuclear ribosomal DNA, to generate a supertree of the tribe (Pelser et al. 2007; see also http://www.compositae.org/); this will provide an essential preliminary phylogenetic hypothesis of the tribe. Although supertrees are employed for phylogenetic analyses of large taxonomic groups (see http://www.tolweb.org/tree/), these methods are not devoid of criticisms (Bininda-Emonds 2004). In addition, in the Senecioneae, only the maternally inherited chloroplast genome and ITS, with its unpredictable evolutionary behavior (Álvarez and Wendel 2003), have been widely used. Thus, there is a need to employ additional independent nuclear markers, both to test previous phylogenetic hypotheses and to complement supertree datasets.

At present there are no genomic libraries available or sequencing projects for any member of the Senecioneae. However, two genomic libraries from model organisms belonging to different tribes, (*Helianthus annuus*, Heliantheae) and (*Lactuca sativa*, Lactuceae), provide a framework for selecting potentially homologous genes. Since *Lactuca* is relatively distant from *Helianthus* (see http://www.compositae.org/), comparisons of homologous sequences from these two genera may prove fruitful in designing tools for phylogenetic use in the Senecioneae. Thus, primers selected on conserved regions in *Helianthus* and *Lactuca* should also work for members of the Senecioneae and, presumably, for most members within the Asteraceae. These assumptions need to be tested, of course, as genes can vary in copy number or presence among taxa, and because primer sites for PCR amplification might be polymorphic. To minimize these problems it is helpful to compare as many sequence databases as possible. Within Asteraceae we had available for the present study sequence databases from genomic libraries of organisms from other genera, such as *Stevia* (Eupatorieae) and *Zinnia* (Heliantheae), thereby allowing us to use members from three different tribes (Eupatorieae, Heliantheae, and Lactuceae).

The approach we detail here is applicable to any group of organisms belonging to or related to taxonomic groups well represented in public nucleotide databases. While comparisons among sequence databases are relatively straightforward, the selection of the best candidate genes may be challenging due to (1) difficulty in diagnosing paralogy and (2) the need to assess variation and its phylogenetic utility. The latter, especially required at low taxonomic levels, can be ascertained only when a good representation of taxa and sequences (clones) is analyzed. Although some approaches, such as that of Wu et al. (2006), are successful for deep phylogenies, the lack of a phylogenetic analysis to assess all candidates selected during the search process might limit their usefulness at lower taxonomic levels. Polyploidy contributes additional complications, since multiple diverse sequences representing homeologues and paralogues may be present in the same genome (Fortune et al. 2007), but they are difficult to avoid in many plant groups, including the Senecioneae, where polyploidy is known to be prevalent in most lineages (Nordenstam 1977; Lawrence 1980; Knox and Kowal 1993; Liu 2004; López et al. 2005).

## Materials and Methods

### Plant Material

Fresh leaf tissue of plants from the living collection of Real Jardín Botánico in Madrid, collected in the field and preserved in silica gel or cultivated from seeds received from other botanical institutions (Table 1), were used to isolate total DNA with the Plant DNeasy kit (Qiagen), following the manufacturer's instructions. Since sampling was aimed at assessing the phylogenetic utility of the markers tested within the Senecioneae, we selected eight species from the main taxonomic groups within the tribe (Pelser et al. 2007) spanning different ploidy levels (from $x = 5$ to $2x$, $4x$, $6x$, and unknown) and distributed in different biogeographical areas. In addition, sequences from three species belonging to other tribes in Asteraceae were included as outgroups (see Table 1).

### DNA Sequence Databases

The main sources of DNA sequences used were the online databases (DDJB, EMBL, and GenBank). These databases are interconnected, making all data available at any of their

**Table 1** Plant materials used, indicating origin, voucher, geographic distribution of taxon, and chromosome numbers

| Taxon | Origin and voucher ID | Geographic distribution | 2*n* |
|---|---|---|---|
| *Cissampelopsis volubilis* (Blume) Miq. | Borneo: Sarawak, Batang Ai, Nanga Sumpa, March 2005, B. Nordenstam, BN 9450 | E & SE Asia (Indomalesia) | — |
| *Emilia sonchifolia* (L.) DC | Cultivated in RJB greenhouse from seeds collected in Tsukuba-Shi Ibaraki Botanic Garden (Japan), 31 May 2005, I. Álvarez, IA 1971 | Pantropical | 10 |
| *Euryops virgineus* (L. f.) DC | Spain: Madrid, RJB living collection, 23 May 2005, I. Álvarez, IA 1967 | South Africa | — |
| *Hertia cheirifolia* Kuntze | Spain: Madrid, RJB living collection, 13 Sept 2006, I. Álvarez, IA 1990 | South Africa | — |
| *Pericallis appendiculata* (L.f.) B. Nord. | Spain: Canary Islands, La Gomera, Vallehermoso, 16 Apr 2005, A. Herrero, J. Leralta & L. Medina, AH 2527 | Canary Islands | — |
| *Petasites fragrans* (Vill.) C. Presl. | Spain: Madrid, RJB living collection, 13 Sept 2006, I. Álvarez, IA 1991 | Central Mediterranean | 58, 59, 60, 61 |
| *Jacobaea maritima* (L.) Pelser & Meijden | Italy: Sicily, Parco della Madonie, Vallone Madonna degli Angeli, 2 June 2000, A. Herrero & al., AH 982 | Sicily | — |
| *Senecio vulgaris* L. | Spain, Madrid, spontaneous in RJB, 15 Apr 2005, I. Álvarez, IA 1966 | Cosmopolitan | 40 |
| *Echinacea angustifolia* DC (Heliantheae) | Spain, Madrid, RJB living collection, 22 June 2005, I. Álvarez, IA 1976 | North America | 11, 22 |
| *Lactuca sativa* L. (Lactuceae) | Spain, Madrid, RJB living collection, 22 June 2005, I. Álvarez, IA 1975 | Cultivated worldwide | 18 |

*Note.* Chromosome numbers were obtained from local floras and from the Index of Plant Chromosome Number database, available at http://www.mobot.mobot.org/W3T/Search/ipcn.html

web sites. Arbitrarily we choose the GenBank web site to do our searches (http://www.ncbi.nlm.nih.gov/). At present, about 63 million sequences are available for Eukaryota, of which 14 million are from plants. Focusing on single-copy nuclear genes for Senecioneae, and to accelerate searches within this database, we excluded sequences from plastid genomes as well as ribosomal DNA and microsatellites from the Asteraceae. A total of 180,747 sequences were downloaded in a file named "Asteraceae NCBI" that was the main database for our searches (Fig. 1).

Another database used is generated from genomic libraries from *Helianthus annuus* lines RHA801 and RHA280, *H. paradoxus*, *H. argophyllus*, *Lactuca sativa* L. cv. Salinas, and *Lactuca serriola* L. These libraries are from The Compositae Genome Project (available at http://www.compgenomics.ucdavis.edu/). From this web site we downloaded all assembled complementary DNAs (cDNAs) of *Lactuca* and *Helianthus* in files called "*Lactuca* CGP" and "*Helianthus* CGP," containing 8179 and 6760 sequences, respectively (Fig. 1).

In the same way and to compare sequences within the GenBank database, we independently downloaded other Asteraceae sequence databases with a relatively high number of sequences from GenBank. This yielded 17,633 sequences from *Zinnia* (Heliantheae, Asteraceae), 5574 sequences from *Stevia* (Eupatorieae, Asteraceae), and 697 sequences from other Senecioneae species, downloaded into the respective files "*Zinnia* NCBI," "*Stevia* NCBI," and "Senecioneae NCBI" (Fig. 1). Note that all of these sequences are already included in the main database "Asteraceae NCBI."

## Search Method

To compare sequences among all sets of sequences described above, we used BLAST (Altschul et al. 1990) with the program Blast-2.2.9, available at ftp://ftp.ncbi.nih.gov/blast/executables/. The use of this stand-alone version of the program allowed us to compare our database files against each other, obtaining output faster and in a form that is easier to analyze than the online version (Fig. 1). Output files were limited to those comparisons that at most have an expectation value ($E$) = 0.001 (i.e., 0.001 is the probability that matches between sequences are by chance).

The first BLAST was applied between the two largest files, *Lactuca* CGP and Asteraceae NCBI, excluding sequences of *Lactuca* from the latter to avoid redundancy. The second BLAST was done between *Helianthus* CGP and Asteraceae NCBI, excluding *Helianthus* and *Lactuca* sequences from the main database; thus the output file in this second search does not contain repeated comparisons with the first search (i.e., all *Lactuca* vs. *Helianthus* comparisons present in the first output are excluded in the second). Successively and in the same way, the remaining databases (*Zinnia* NCBI, *Stevia* NCBI, and Senecioneae NCBI) were compared to Asteraceae NCBI (Fig. 1).

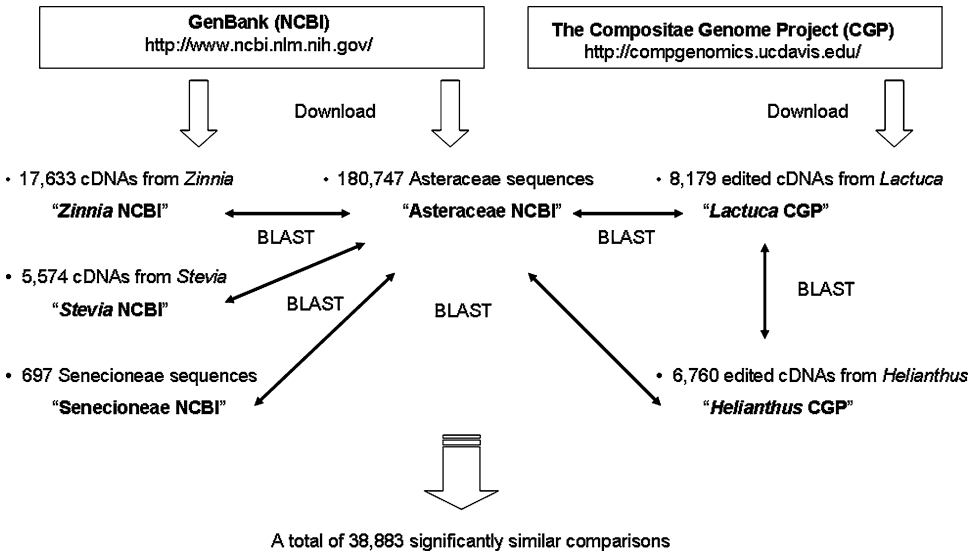


**Fig. 1** Scheme for the search method developed

*Selection of Candidate Genes* Due to the large number of comparisons obtained by BLAST (38,883), the first step in selection of candidate genes was to restrict our searches to those that obey the following constraint parameters: (1) percentage identity between 90% and 100%; (2) length of alignment ≥600 bp; (3) $E = 0$; and (4) presence in at least two Asteraceae tribes (Fig. 2). Using these constraints, nine candidate genes were selected for the next step (Table 2 and Supplementary Appendix 1) and compared by an online BLAST using both "blastn" and "blastx" search options. The former was used to find potentially homologous genomic sequences (i.e., including exons and introns) in other angiosperms, and the latter to estimate which protein (if any known) is similar to each candidate, allowing us to do a preliminary characterization and comparison to *Arabidopsis thaliana*, the closest organism to Asteraceae whose genome has been completely sequenced and assembled (Fig. 2).

The third step consisted of aligning each of the nine candidates with sequences found in Asteraceae NCBI plus genomic sequences (exons and introns) of its closest orthologous loci in other angiosperms. Candidate *QG_CA_Contig2080* was excluded due to its multiple significant alignments (>10 loci) in *Arabidopsis* genome (see Table 2). Each of the eight remaining candidates was aligned to design optimal primers (without ambiguous nucleotides) that presumably would amplify each marker in all Asteraceae for which the primer sites were conserved. The requirements for each candidate gene for the next step were (1) to have at least two highly conserved regions (perfect match through all sequences in the alignment) ~22 nucleotides long and located ~200 nucleotides of exon sequence apart from each other, and (2) sequence representation in the alignment of at least three different Asteraceae tribes. Finally, primers for five candidates that met these criteria (Table 3 and Supplementary Appendix 2) were designed and tested by PCR (Fig. 2).

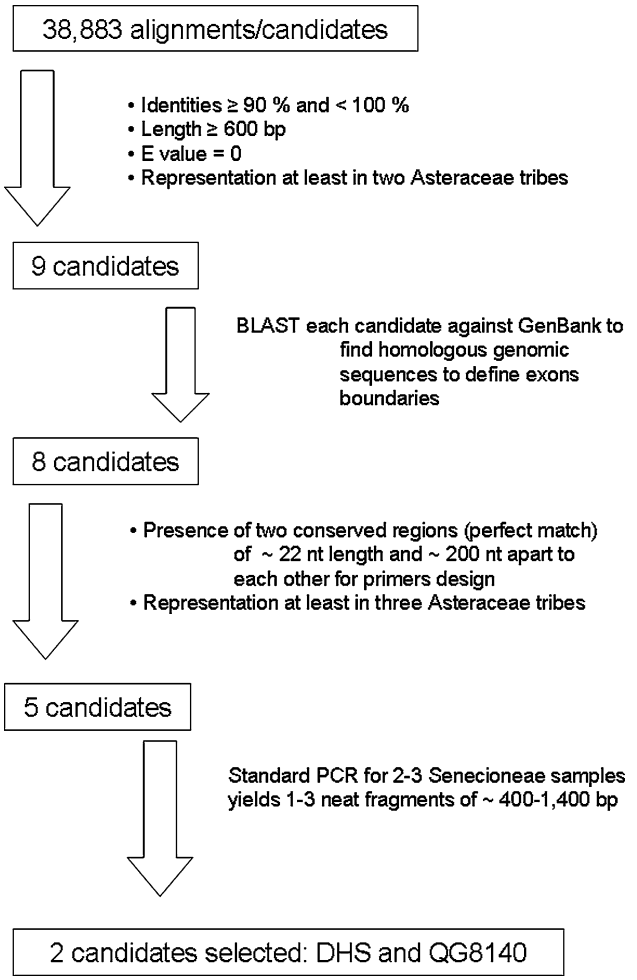


**Fig. 2** Scheme for the process of selection of candidate markers

**Table 2** BLAST results and identification of the nine preselected markers

| Preselected candidates | Origin | BLAST results using Asteraceae NCBI database | | | BLASTx results using GenBank protein database | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | First alignment[a] | Identity | Length | First alignment[a] | Identity | Length | E | Significant alignments with *Arabidopsis* genome loci |
| QG_CA_Contig2080 | 1 | BG524152 | 91% | 624 bp | AAD33072 | 87% | 305 aa | $2e-^{152}$ | At4g21960, At2g37130, At2g18150, At5g40150, At5g14130, At2g18140, At4g17690, At3g50990, At3g28200, At2g24800, and others |
| QG_CA_Contig2453 | 1 | BU024339 | 90% | 719 bp | AAO15916 | 78% | 538 aa | 0 | At4g30920, At2g24200, At4g30910 |
| QG_CA_Contig2630 | 1 | AY545660 | 90% | 641 bp | AAT45244 | 79% | 343 aa | $7e^{-128}$ | At2g45300, At1g48860 |
| QG_CA_Contig5271 | 1 | BU027516 | 90% | 682 bp | CAC00532 | 90% | 446 aa | 0 | At2g36530 |
| QG_CA_Contig5597 | 1 | BU028221 | 99% | 716 bp | CAC67501 | 72% | 232 aa | $3e^{-104}$ | At4g14030, At4g14040, At3g23800 |
| QG_CA_Contig8140 | 1 | CF088687 | 91% | 601 bp | AAT77289 | 99% | 181 aa | $2e^{-99}$ | At1g10630, At3g62290, At5g14670, At1g70490, At2g47170, At1g23490 |
| QH_CA_Contig1827 | 2 | BG525164 | 93% | 606 bp | NP_564985 | 85% | 251 aa | $1e^{-130}$ | At1g70160, At4g27020, At5g54870, At5g08610 |
| QH_CA_Contig5513 | 2 | BQ989463 | 92% | 681 bp | XP_472987 | 90% | 253 aa | $4e^{-119}$ | At5g36700, At5g47760 |
| SVE238622[a] | 3 | AJ704846 | 91% | 1086 bp | CAB65461 | 100% | 371 aa | 0 | At5g05920 |

*Note*. First alignments are not redundant. Origin of data: (1) *Lactuca* EST from the Compositae Genome Project Database; (2) *Helianthus* EST from the Compositae Genome Project Database; (3) GeneBank database. aa, amino acids

[a] For definition see Supplementary Appendix 1

As a test of the efficiency of our search strategy, we also explored the potential of two other single-copy nuclear genes formerly developed in the Asteraceae. One of these markers is a cellulose synthase gene that was used in *Gossypium* phylogenies (Cronn et al. 2002, 2003; Senchina et al. 2003; Álvarez et al. 2005) under the name *CesA1b* (called here *CesA*). Primers for this gene in Asteraceae were recently developed in a phylogeny of the genus *Echinacea* by one of us (I. Álvarez) based on sequences of *Gossypium* (Malvaceae) and *Zinnia* (Asteraceae) found in GenBank. Specific primers (Supplementary Appendixes 2 and 3) for a different region of this gene containing a longer exon sequence were designed, including sequences of *Echinacea* in our alignment. The second marker explored was a chalcone synthase (*CHS*) belonging to a gene family that was previously characterized in the Asteraceae (Helariutta et al. 1996). This allowed us to design specific primer

**Table 3** Candidates selected for testing by PCR and direct sequencing, sequences used for primer design, and primer combinations tested

| Candidate | Abbreviation | Other sequences used for primer design[a] | | Primer combinations[a] | | Exon length (bp) |
|---|---|---|---|---|---|---|
| | | cDNA | Genomic DNA | Forward | Reverse | |
| QG_CA_Contig2630 | QG2630 | AY545660, AY545668, AY545662, AY545659, AY545661 | AY545667 | 870F | 1472R | 354 |
| | | | | 968F | 1472R | 256 |
| QG_CA_Contig5271 | QG5271 | BU027516, CD850822, BQ916549, CF098760, BG525568, BG522060 | X58107 | 1291F | 1717R | 197 |
| QG_CA_Contig8140 | QG8140 | CF088687, CD854645, BG523226, CF091055, CD848661, BU026719, CD848409 | AL138651 (region: 79932..81190) | 72F | 1070R | 403 |
| QH_CA_Contig5513 | QG5513 | BQ989463, BQ847896, BQ988373, BG522222 | NC_003076 (region: 14438868..14441693) | 69F | 746R | 250 |
| | | | | 69F | 1002R | 374 |
| | | | | 69F | 1263R | 503 |
| SVE238622 | DHS | AJ704846, AJ704841, AY731231, AJ704847, AJ238623, AJ704842, AJ704850, AJ010120, AJ251500, AJ704849 | AB017060 (region: 10669..12587) | 142F | 1116R | 674 |

[a] For definitions of accession numbers see Supplementary Appendix 1, and for primer sequences see Supplementary Appendix 2

pairs for one of the copies (Supplementary Appendixes 2 and 3).

### Amplification, Cloning, and Sequencing

Seven candidate genes (*CesA*, *CHS*, *DHS*, *QG2630*, *QG5271*, *QG8140*, and *QH5513*) were tested for amplification, followed by direct sequencing using 13 Asteraceae-specific primer combinations and the PCR programs indicated (Supplementary Appendixes 2 and 3). All primer combinations of candidates *QG2630* and *QH5513* yielded complex fragment patterns (i.e., multiple fragments of different sizes depending on the sample) or failed to amplify; thus, they were excluded for the next step. For each of the remaining five candidates we selected the best primer combination (Supplementary Appendix 3) in terms of amplification simplicity and pattern obtained (one to three neat bands per sample) (Fig. 2).

Fragments of similar size across all samples were excised from 1.5% agarose gels and isolated using the Eppendorf Perfectprep Gel Cleanup kit following the manufacturer's instructions. Fragments were sequenced and checked for sequence identity using online BLAST. At this point we eliminated candidate *QG5271* due to the existence of an intron of unknown length plus one additional intron of 103 bp in *Hubertia ambavilla* (Senecioneae). With the remaining four candidates (*CesA*, *CHS*, *DHS*, and *QG8140*) we cloned and sequenced those fragments that match the target marker (one fragment per sample). Ligation and transformation reactions were performed with the Promega pGEM-T Easy Vector System II cloning kit as described in its instruction manual. A minimum of 10 colonies were picked for cloning, obtaining 5–10 cloned sequences per sample for each marker after excluding false positives. Growth of selected colonies, harvesting, and lysis by alkali were performed following by the protocol of Sambrook et al. (1989) with slight modifications. Sequencing reactions were carried out at the Sequencing Facility of Parque Científico de Madrid using the SP6 and T7 plasmid promoter primers as suggested in the cloning kit manual (see above).

### Data Analysis

Sequence alignments were performed manually using BioEdit v. 5.0.9 (Hall 1999). Exons were straightforward to align, while introns were mostly ambiguous among different genera and therefore were excluded from analyses.

Phylogenetic analyses were performed as a way to have a preliminary assessment of utility, i.e., signal of selected candidate genes. Phylogenies were reconstructed using maximum parsimony as implemented in PAUP*4.0b10 (Swofford 1999). Additionally, in order to depict distances among possible paralogues, neighbor-joining (NJ) analyses were performed. Searches for the most parsimonious trees (m.p.t.) were performed with the heuristic algorithm with the TBR option for searching optimal trees and ACCTRAN for character optimization. One hundred random addition sequences were performed, saving 1000 trees per replicate. Gaps were treated as missing data. NJ trees were based on a distance matrix derived from Nei and Li (1979) distances. Bootstrap analyses with 1000 replicates were performed to assess relative branch support.

## Results and Discussion

### Considerations on Search Methodology

The efficiency of the search method employed here will depend on the depth of sequence representation in databases. In our case, the fact that the tribe Senecioneae is included in a family (Asteraceae) that is well represented by sequence databases from several taxa (*Gerbera* cultivar, *Helianthus annuus*, *Lactuca sativa*, *Stevia rebaudiana*, and *Zinnia elegans*) makes it relatively easy to find highly similar sequences (i.e., putative orthologuous sequences). Thus, although it is likely that a more exhaustive search, comparing the main database to itself and also downloading *Helianthus* and *Lactuca* sequences from GenBank, would produce larger output files, it is unlikely that this would significantly increase the number of candidate genes. An additional consideration is that it is preferable to use the longer and nonredundant *Helianthus* and *Lactuca* cDNA sequences from the CGP than the shorter, often redundant sequences from public databases.

Similarly, selecting the stringency of parameters to use in analyzing BLAST results also influences the effectiveness of the general approach. To illustrate this point, of the initial comparisons obtained here (38,883), we used the following criteria for retention: (1) percentage identity between 85% and 100%, (2) length of alignment $\geq$200 bp, (3) $E < e^{-100}$, and (4) presence in at least two different Asteraceae tribes. This substantially reduced the number of comparisons, but to a number (272) that we still considered too large in terms of timing and costs of primer design and evaluation. With more stringent parameters (i.e., $\geq$ 90% and <100%, length of alignment $\geq$600 bp, $E = 0$), we reduced the number of candidate genes to nine (Table 2) for the next step.

Since our priority was to find conserved regions within orthologous genes, the question naturally arises as to why we performed nucleotide vs. nucleotide searches (blastn) instead of nucleotide vs. protein searches (btastx). First, the lower number of proteins vs. single-copy nuclear nucleotide sequences available (e.g., in the Asteraceae, this is

reduced to 4812 from 180,747) would restrict our searches noticeably. Second, and perhaps most importantly, protein searches may yield multiple equally similar comparisons that involve highly diverse nucleotide sequences, for example, in synonymous sites. Thus, it is possible to find relatively high variation in nucleotide alignment among different genera of Asteraceae, even with near-perfect protein identity. In fact, this kind of neutral mutation provides useful phylogenetic signal.

Once preselection of candidates is completed using blastn, we recommend further searches comparing candidates against the public protein databases and, also, against all the public nucleotide databases in an attempt to gain insight into the gene product and multigene family complexity, as well as to find related sequences from other organisms to include in the alignment. These analyses also are likely to reveal intron/exon boundaries along cDNA sequences of candidates, for which positions may be conserved even among rather distantly related organisms (Schlüter et al. 2005).

## Selection of Genes by Phylogenetic Analysis

There are several approaches that might help in orthology assessment (i.e., shared expression patterns, Southern hybridization analysis, and comparative genetic mapping [for reviews see Sang 2002; Small et al. 2004]), but phylogenetic analysis is the only one that can reveal orthologues. This analysis not only is important for providing evidence on orthology/paralogy relationships, but also provides insight into sequence similarity and relative levels of variation at different phylogenetic scales. Depending on the age of divergence and rates of molecular evolution, some regions of a gene, such as introns, may be too variable to align in some samples. Here we conducted phylogenetic analyses for the four markers selected (*CesA*, *CHS*, *DHS*, and *QG8140*).

*Analysis of CesA Sequences* PCRs for all samples resulted in one band of about 1.2–1.3 kb, except for *Petasites fragrans*, for which two bands ($\sim$1.3 and 1.5 kb) were recovered. After direct sequencing, we identified the shorter band ($\sim$1.3 kb) as the one that matches our target, and thus it was selected for cloning and sequencing.

A conserved structure in number and position of putative exons and introns is present in all samples. This includes complete sequences of three exons and four introns, plus partial sequences of two exons. Although it is possible to align introns, alignment ambiguities precluded confident phylogenetic analysis, so consequently only exons were considered further. Alignment of exons led to a matrix 853 nucleotides (nt) long and included 84 sequences as follows (see Supplementary Appendix 4 for GenBank accession numbers and TreeBase accession number "M3572"): 9 clones from *Cissampelopsis volubilis*, 8 clones from *Echinacea angustifolia*, 7 clones from *Emilia sonchifolia*, 8 clones from *Euryops virgineus*, 7 clones from *Hertia cheirifolia*, 7 clones from *Lactuca sativa*, 9 clones from *Pericallis appendiculata*, 8 clones from *Petasites fragrans*, 10 clones from *Jacobaea maritima*, 10 clones from *Senecio vulgaris*, and 1 cDNA sequence of *Zinnia elegans* downloaded from GenBank (AU288253).

Stop codons were found in five sequences (i.e., 1–8, 2–3, 6–10, 13–10, and 23–4). In addition, a few indels of 1–3 nt were detected in 12 sequences (i.e., 1–7, 2–9, 13–4, 14–8, 23–2, 25–1, 25–5, 35–1, 40–4, 40–6, 40–9, and 40–10). Therefore, we detected putative pseudogenes in 17 clones sequenced (20.5%), varying from none in *Cissampelopsis volubilis* to 50% of the sequences from *Echinacea angustifolia*. We also found two independent shifts (GT shifts to GC) on intron splicing sites in sequences 14–3 and 14–10 plus one shift occurring in the fourth intron in all clones of *Pericallis appendiculata* (sequences 35–1 thru 35–10) and clone 6–1 of *Lactuca sativa*. Within the ingroup, 286 sites were variable (33.5%), of which 190 (22.3%) were parsimony informative. Of the 190 parsimony-informative sites, 149 (17.5%) were synonymous and 41 (4.8%) were replacement changes. Percentages of polymorphic sites scarcely vary when the 17 putative pseudogenes are eliminated (i.e., 247 variable sites, 184 parsimony-informative sites); nor do numbers of synonymous (151) and parsimony-informative replacement changes (33). Analysis of replacement changes gives no clear pattern showing a scattered distribution throughout the sequences sampled.

Parsimony analyses including all sequences (functional and nonfunctional) and including only the putatively functional sequences were conducted to evaluate the level of coalescence of the putative pseudogenes and other paralogous sequences. For the *CesA* matrix using all sequences, only 1000 m.p.t. were saved and analyzed, with a length of 594 steps, consistency index (CI) excluding uninformative characters = 0.71, and retention index (RI) = 0.94. The strict consensus is shown in Fig. 3, and bootstrap values >50% are indicated above branches. Visual inspection indicates that most species, including the outgroup, present sequences (clones) belonging to different relatively well-supported clades in the cladogram, indicating the existence of several types of copies. The topology recovered does not show any discernible phylogenetic signal, probably due to the severity of this paralogy problem. A NJ analysis (not shown) resulted in a similar topology, in which branches for groups of terminals are relatively long. When pseudogenes are excluded from the analysis, a total of 576 m.p.t. are recovered, with a length of 480 steps, CI = 0.73, and RI = 0.94; the topology and

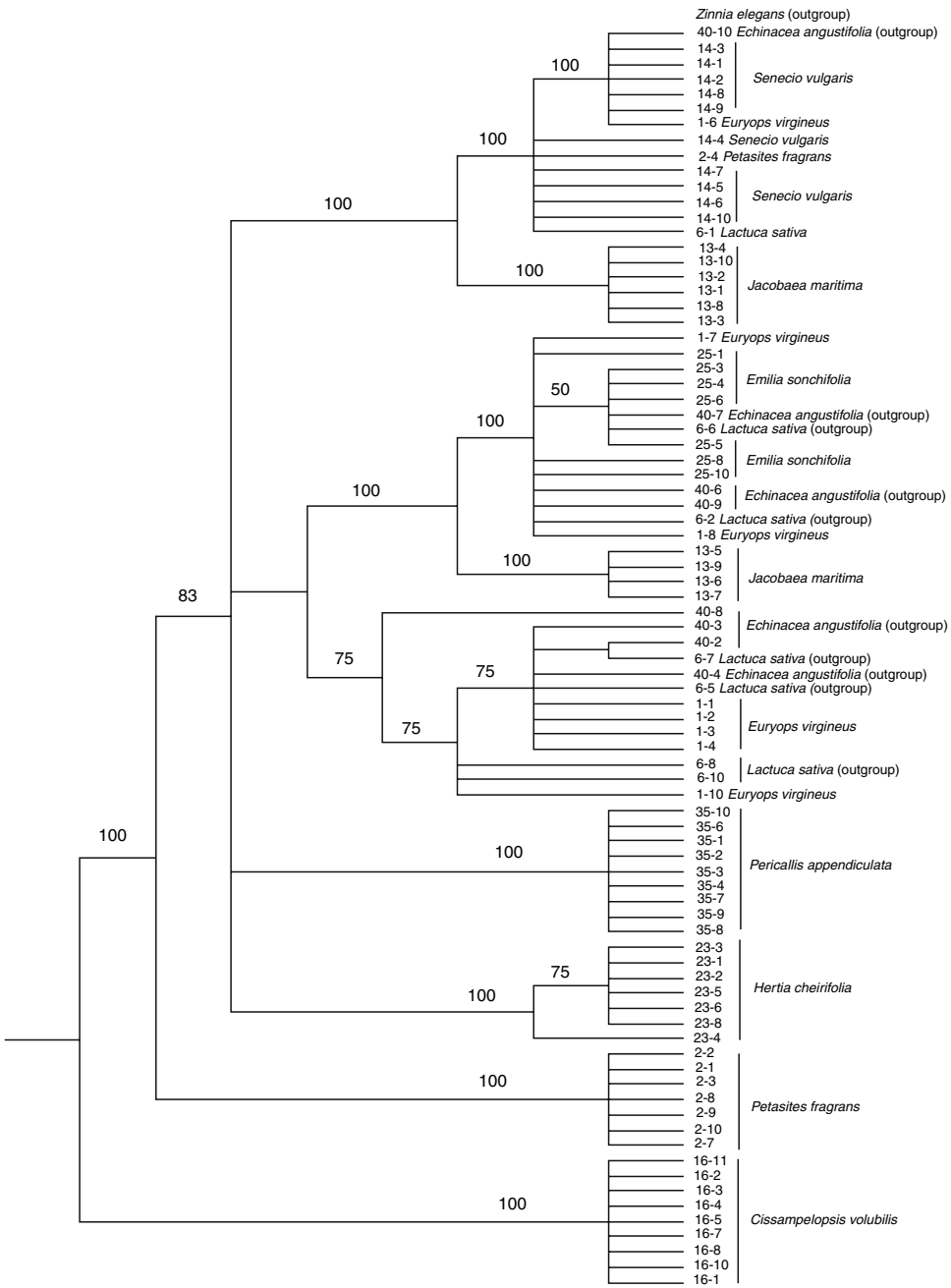


**Fig. 3** Strict consensus of 1000 m.p.t. obtained in the analysis of the *CesA* matrix (length = 594; CI = 0.71; RI = 0.94). Bootstrap values ≥50% are indicated above branches

support (not shown) structure of this tree are similar to those derived from the full dataset. Therefore, pseudogenes are not affecting the topology or the homoplasy level (i.e., CI and RI are similar and reasonably moderate). For this marker we conclude that although there is a great deal of sequence variation, more work is needed before *CesA* sequences will be useful in the Senecioneae, because of the apparent deep paralogy (perhaps prior to radiation of the tribe, or even the family).

*Analysis of CHS Sequences* All amplification products yielded one sharply resolved band of ~0.5 kb, except for *Echinacea angustifolia*, in which three bands of ~0.5, 0.8, and 2.0 kb were observed. After directly sequencing DNA isolated from several 0.5-kb bands to test homology with our target, we proceeded to isolate and sequence clones from all taxa included in the study. The aligned matrix has a total length of 518 nt, including a unique partial exon in 69 sequences (see Supplementary Appendix 4 for GenBank accession numbers and TreeBase accession number "M3571"): 1 cDNA sequence of *Callistephus chinensis* downloaded from GenBank (Z67988), 7 clones from *Cissampelopsis volubilis*, 5 clones from *Echinacea angustifolia*, 8 clones from *Emilia sonchifolia*, 8 clones

from *Euryops virgineus*, 1 cDNA sequence of *Gerbera* cultivar downloaded from GenBank (Z38096), 6 clones of *Hertia cheiriifolia*, 7 clones from *Lactuca sativa*, 5 clones of *Pericallis appendiculata*, 8 clones from *Petasites fragrans*, 6 clones from *Jacobaea maritima*, and 7 clones from *Senecio vulgaris*. Indels are present in three sequences: clones 40–1 and 40–5 have an insertion of 3 nt (ATT) plus one deletion of 3 nt (stop codons), and clone 25–2 has one deletion of one nucleotide. In addition, stop codons were found in sequences 2–3, 2–5, 23–3, and 35–3. Thus, a total of seven sequences are presumed to be pseudogenes. Within the ingroup, 245 sites were variable (47.3%), of which 210 (40.5%) were parsimony informative. Within the parsimony-informative sites, 121 (23.4%) were synonymous and 89 (17.2%) were replacement changes. Analysis revealed that most of the replacements are present in all clones of *Senecio vulgaris* (41 sites; 7.9%), all clones of *Euryops virgineus* (34 sites; 6.6%), clones 2–1, 2–5, and 2–6 of *Petasites fragrans* (8 sites; 1.5%), and all clones of *Cissampelopsis volubilis* (5 sites; 1%). A group of clones from the outgroup (40–3, 40–7, and 40–8 of *Echinacea angustifolia*) had 36 replacement changes (7%). Among the remaining samples, replacement sites are few (1–3; 0.2–0.6%) and scattered. Excluding pseudogenes (7 sequences) plus sequences with a high number of replacement changes (27 sequences), the number of variable sites within the ingroup decreases to 114 (22%), of which 91 (17.6%) are parsimony informative. Note that in this case the number of replacements in parsimony-informative sites decreases dramatically, to 9 (1.7%), where synonymous changes occur in the remaining 82 sites (15.8%).

In a parsimony analysis of a *CHS* complete matrix, 32 m.p.t. were obtained, with a length of 740 steps, CI = 0.61, and RI = 0.93. A strict consensus of these trees (Fig. 4) shows that except for *Echinacea*, all clones from the same individual form terminal clusters with high bootstrap support in almost all cases. Three main groupings are revealed (i.e., one formed by clones of *Petasites*, *Hertia*, *Emilia*, *Pericallis*, *Jacobaea maritima*, and the cDNA of *Callistephus*; a second group formed by clones of *Euryops*, *Senecio vulgaris*, *Lactuca*, and some clones of *Echinacea*; and a third group formed by clones from *Cissampelopsis* and some clones of *Echinacea*), although with low bootstrap support. The NJ analysis (not shown) presents an equivalent topology, with a noticeably long branch grouping the *Euryops* and *Senecio vulgaris* clones.

As noted above, all clones of *Euryops* and *Senecio vulgaris* displayed a high ratio of nonsynonymous changes, suggesting the presence of a paralogue in our amplifications and sequencing. In total we infer the presence of three different functional paralogues (one present in *Euryops* and *Senecio vulgaris*, another present in *Cissampelopsis* and some *Echinacea* clones, and a third present in the remaining samples). Pseudogenes are detected within the first and second types of paralogues, but they do not necessarily affect the topology recovered, since they coalesce with the remaining clones from each species. In an attempt to evaluate the level of phylogenetic signal in the major type of sequences recovered, a parsimony analysis with a reduced matrix (including the third type of sequences described above and excluding pseudogenes) was run. A strict consensus of 16 m.p.t. with a length of 282, CI = 0.8, and RI = 0.92 is shown in Fig. 5. Bootstrap support for species clades are high, as expected, although support for other clustering is low (around 60%), except for ingroup (85%). While this result is not incongruent with previous phylogenetic hypothesis, an ample sampling (in terms of both species and clones per species) is needed to allow further assessment of the utility of this gene.

*Analysis of DHS Sequences* Direct sequencing of the brightest band (1.2–1.4 kb) obtained by PCR for all samples indicated a high similarity to the exon sequence of our target gene. A total aligned length of 1463 nt includes three complete and two partial exons plus four introns of the *DHS* gene in 47 samples (see Supplementary Appendix 4 for GenBank accession numbers): 9 clones from *Echinacea angustifolia*, 10 clones from *Euryops virgineus*, 10 clones from *Lactuca sativa*, 9 clones from *Petasites fragrans*, and 9 clones from *Jacobaea maritima*. In addition, four cDNA sequences of the *DHS* gene from *Eupatorium cannabinum*, *Lactuca sativa*, *Petasites hybridus*, and *Senecio vernalis* found in GenBank (i.e., AJ704841, AY731231, AJ704846, and AJ238622 respectively) were included in the alignment for comparison. Intron/exon boundaries are conserved in all samples. During alignment we found four types of sequences (here called a, b, c, and d) easily distinguishable by intron similarity. While intron alignments within each type are unambiguous, introns among types are not confidently aligned. To a lesser extent, the identity of each main type of intron (a, b, c, and d) is also supported by exon sequence variation. Therefore, we considered each type of sequence to be independently analyzable.

Type a corresponds to clones 6–3 and 6–4 of *Lactuca sativa*, 13–1, 13–2, 13–4, 13–7, 13–8, and 13–9 of *Jacobaea maritima*, and 40–9 of *Echinacea angustifolia*. All sequences correspond to putative functional genes except clone 13–4, which presents a mutation (AG shifts to GG) in the third intron splicing site recognition. Within these sequences low variation is found; i.e., of 1388 sites, 37 (2.7%) were variable and 8 (0.6%) were parsimony informative. Type b includes sequences of clones 2–1, 2–3, 2–4, 2–7, 2–8, 2–9, and 2–10 of *Petasites fragrans*, 6–1, 6–2, 6–5, 6–6, 6–7, 6–8, and 6–9 of *Lactuca sativa*, and 40–1, 40–2, 40–6, 40–7, and 40–10 of *Echinacea angustifolia*. No stop codons were found, although there are deletions of

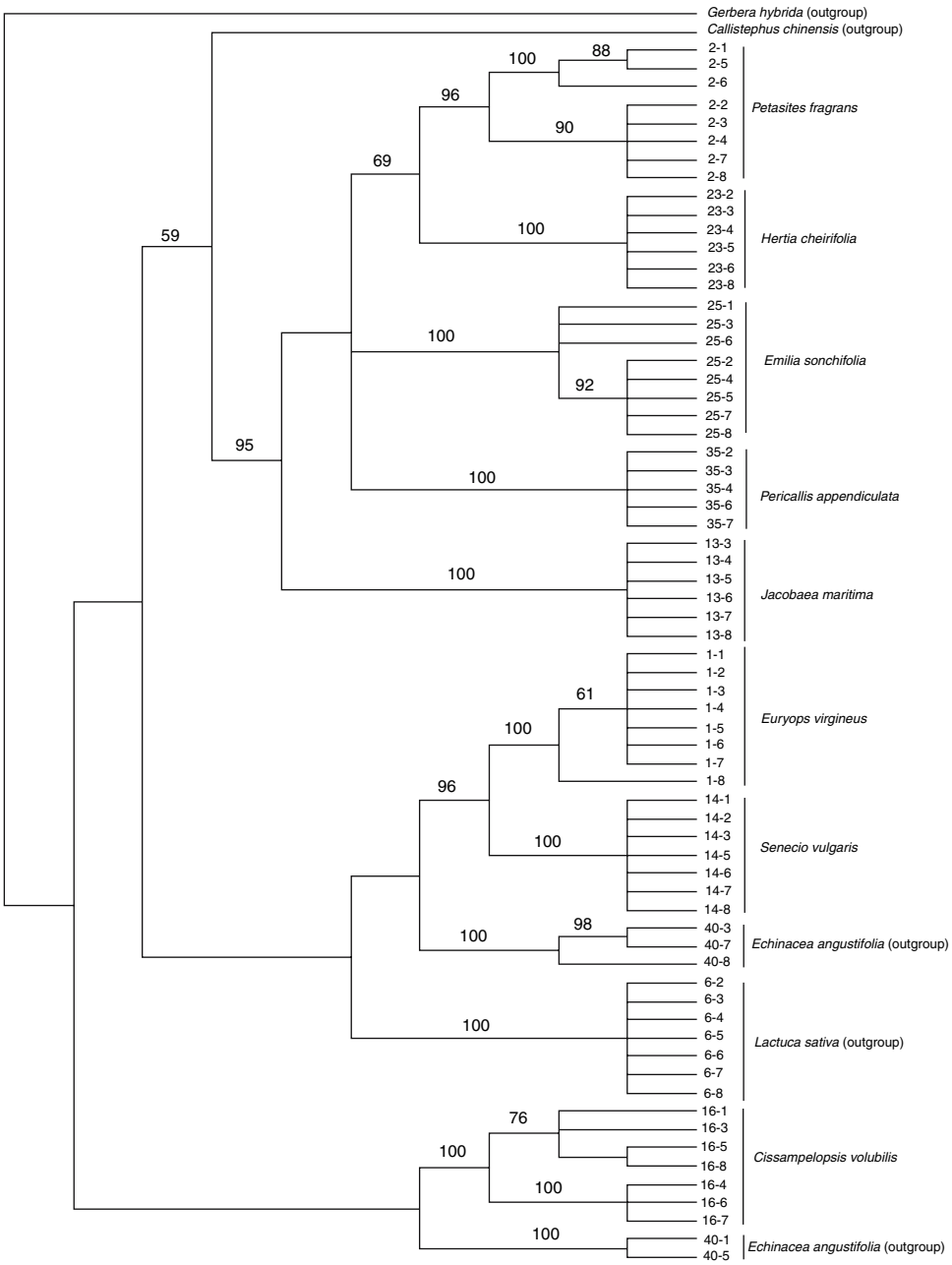


**Fig. 4** Strict consensus of the 32 m.p.t. obtained in the analysis of the *CHS* matrix (length = 740; CI = 0.61; RI = 0.93). Bootstrap values ≥50% are indicated above branches

one nucleotide in exon sequence and/or intron splicing site mutations in clones 2–9, 6–9, and 40–6. Analysis of polymorphic sites in a total alignment of 1456 nucleotides also reveals low levels of variation, with 87 (6%) variable sites, of which 4 (0.3%) are parsimony informative. The type c matrix is composed of clones 1–5, 1–7, and 1–8 of *Euryops virgineus*, 2–6 of *Petasites fragrans*, 6–10 of *Lactuca sativa*, 13–5 and 13–10 of *Jacobaea maritima*, and 40–4 and 40–8 of *Echinacea angustifolia*. All of these sequences are putative functional genes. As with previous types, levels of variation were low; of 1269 sites, 31 (2.4%) were variable and 6 (0.5%) were parsimony informative. Sequences of type d correspond to clones 1–1, 1–2, 1–3, 1–4, 1–6, 1–9, and 1–10 of *Euryops virgineus*, 2–2 of *Petasites fragrans*, 13–6 of *Jacobaea maritima*, and 40–3 of *Echinacea angustifolia*. All sequences analyzed were putative functional genes, and variation levels are comparable to the other types (i.e., of 1273 sites, 39 [3.1%] are variable and 5 [0.4%] parsimony informative). In all types of sequences, variation is too low to be useful for phylogenies at this level and consequently this marker was not further analyzed.

*Analysis of QG8140 Sequences* PCR products for most samples result in one unique bright band of about 0.8–1.2 kb. In a few cases, one to three faint bands of different sizes also

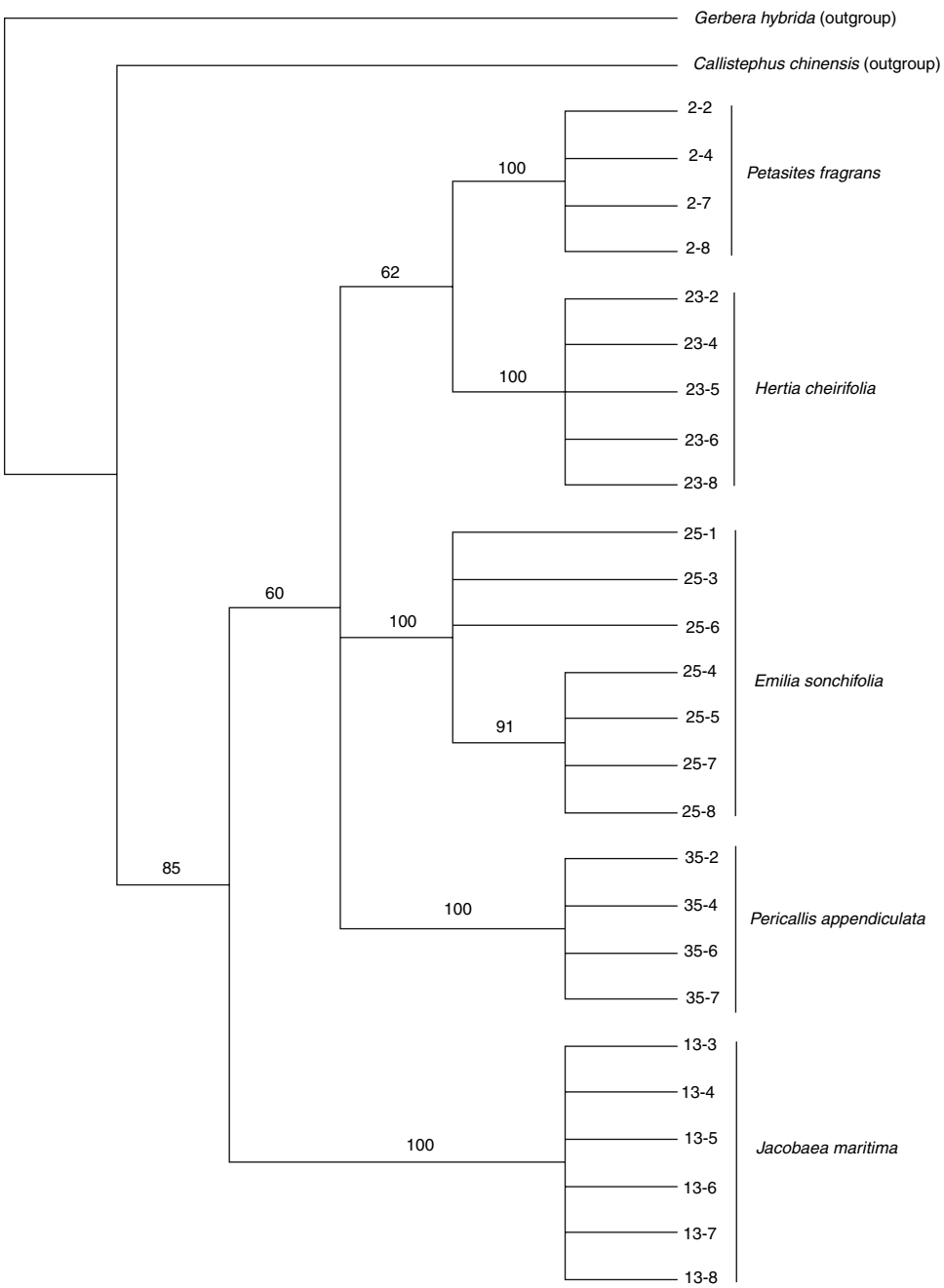


**Fig. 5** Strict consensus of the 16 m.p.t. obtained in the analysis of one type of copy sequences of the *CHS* matrix (length = 282; CI = 0.8; RI = 0.92). Bootstrap values ≥50% are indicated above branches

appear. Direct sequencing of the brightest band of each sample confirms putative exon homology with the target marker.

Alignment of exons (368 nt) for the 81 samples included 10 clones from *Echinacea angustifolia*, 7 clones from *Emilia sonchifolia*, 10 clones from *Euryops virgineus*, 8 clones from *Hertia cheiriifolia*, 10 clones from *Lactuca sativa*, 7 clones from *Pericallis appendiculata*, 9 clones from *Petasites fragrans*, 9 clones from *Jacobaea maritima*, 10 clones from *Senecio vulgaris*, and 1 cDNA of *Lactuca sativa* from the CGP (see Supplementary Appendix 4 for GenBank accession numbers and TreeBase accession number "M3570"). The alignment was unambiguous and without gaps, while introns were possible to align only among clones from the same sample, and thus they were excluded from the matrix. One stop codon was found in position 132 of the alignment in clone 23–5 of *Hertia cheirifolia*. Intron splicing site mutations were found in clones 6–8 of *Lactuca sativa*, 13–2 of *Jacobaea maritima*, and 35–4 of *Pericallis appendiculata*. Within the ingroup, 105 (28.5%) sites were variable and 68 (18.5%) were parsimony informative, of which 65 are synonymous changes and 3 are replacements.

Parsimony analysis of a *QG8140* complete matrix yielded 100 m.p.t. with a length of 213 steps, CI = 0.68,

and RI = 0.95. The strict consensus (Fig. 6) clusters all clones from the same individual together (bootstrap support, 89%–100%), with a few exceptions that jointly form one clade (100% bootstrap value), and for *Pericallis*, for which clones appear in three different clades. In a midpoint-rooted NJ analysis (Fig. 7), the two main groupings are one that includes a mixture of clones from different species and another group that includes the remaining samples clustering by individuals and species, including the outgroup. Therefore, sequences from the "mixed" group are more distant from other sequences of the same individual than from sequences from the outgroup, indicating that at least two types of copies are present in all these species, except for *Pericallis*. Parsimony analysis of a *QG8140* reduced matrix (excluding pseudogenes and sequences from the "mixed" group) was run in order to assess phylogenetic signal of the major copy type. The strict consensus of the 8 m.p.t. obtained, whose length is 145 steps, CI = 0.76, and RI = 0.96, is shown in Fig. 8. The ingroup forms a clade with 92% bootstrap support, where clones from the same individual form clades (with 80%–100% bootstrap support). Topology of

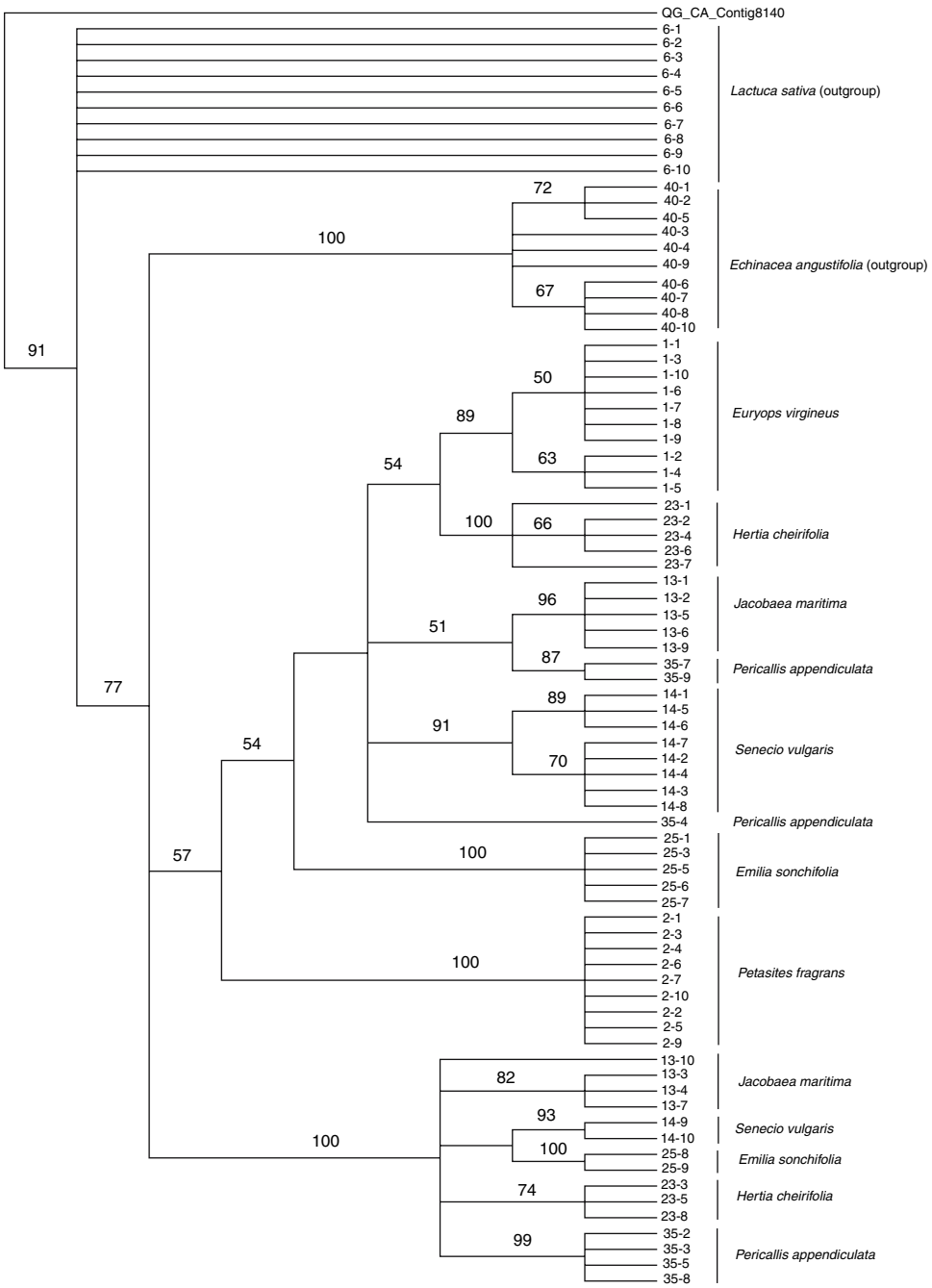


**Fig. 6** Strict consensus of the 100 m.p.t. obtained in the analysis of the *QG8140* matrix (length = 213; CI = 0.68; RI = 0.95). Bootstrap values ≥50% are indicated above branches

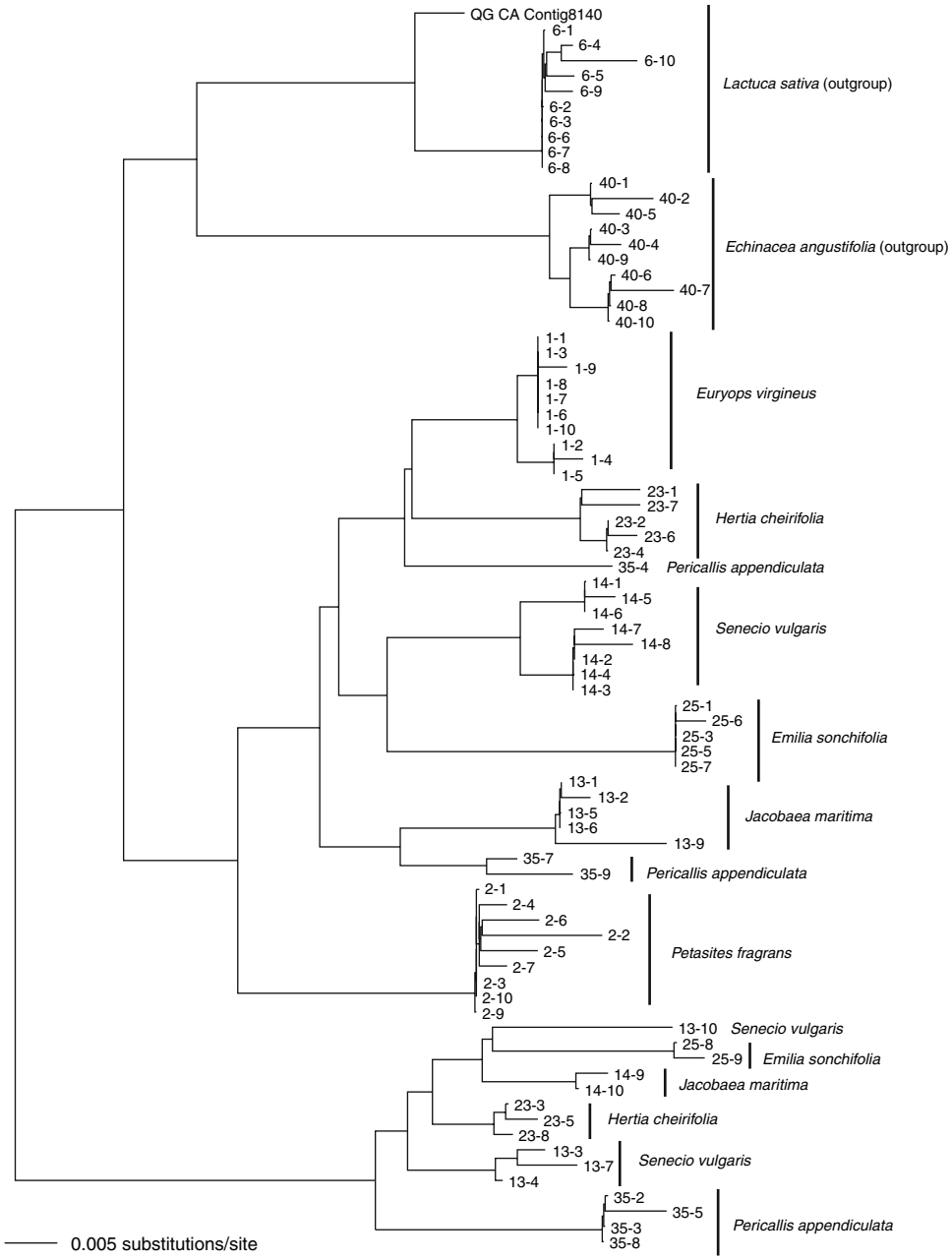


**Fig. 7** Midpoint rooted phylogram from neighbor-joining analysis of the *QG8140* matrix

the strict consensus of the reduced *QG8140* matrix is mostly congruent with the supertree of the tribe Senecioneae that uses sequences of ITS and several cpDNA markers (Pelser et al. 2007), with the exceptions of the positions of *Emilia* and *Pericallis* clones, both with low bootstrap support (<50% and 50%, respectively). It is expected that increasing the number of species sampled and the selection of a closely related outgroup will improve resolution and branch support. This preliminary analysis shows that the marker *QG8140* is a good candidate to develop for the Senecioneae phylogeny, although an increment of clones per individual (maybe at least 20) is recommended to increase the probability of picking orthologous copies.

## Conclusions

Systematics has now entered the era where there is widespread recognition of the immense potential value of nuclear genes for phylogeny reconstruction (Small et al. 2004). With the burgeoning databases of available sequences, it is now possible for highly useful markers to be developed toward this end (Small et al. 2004; Wu et al.

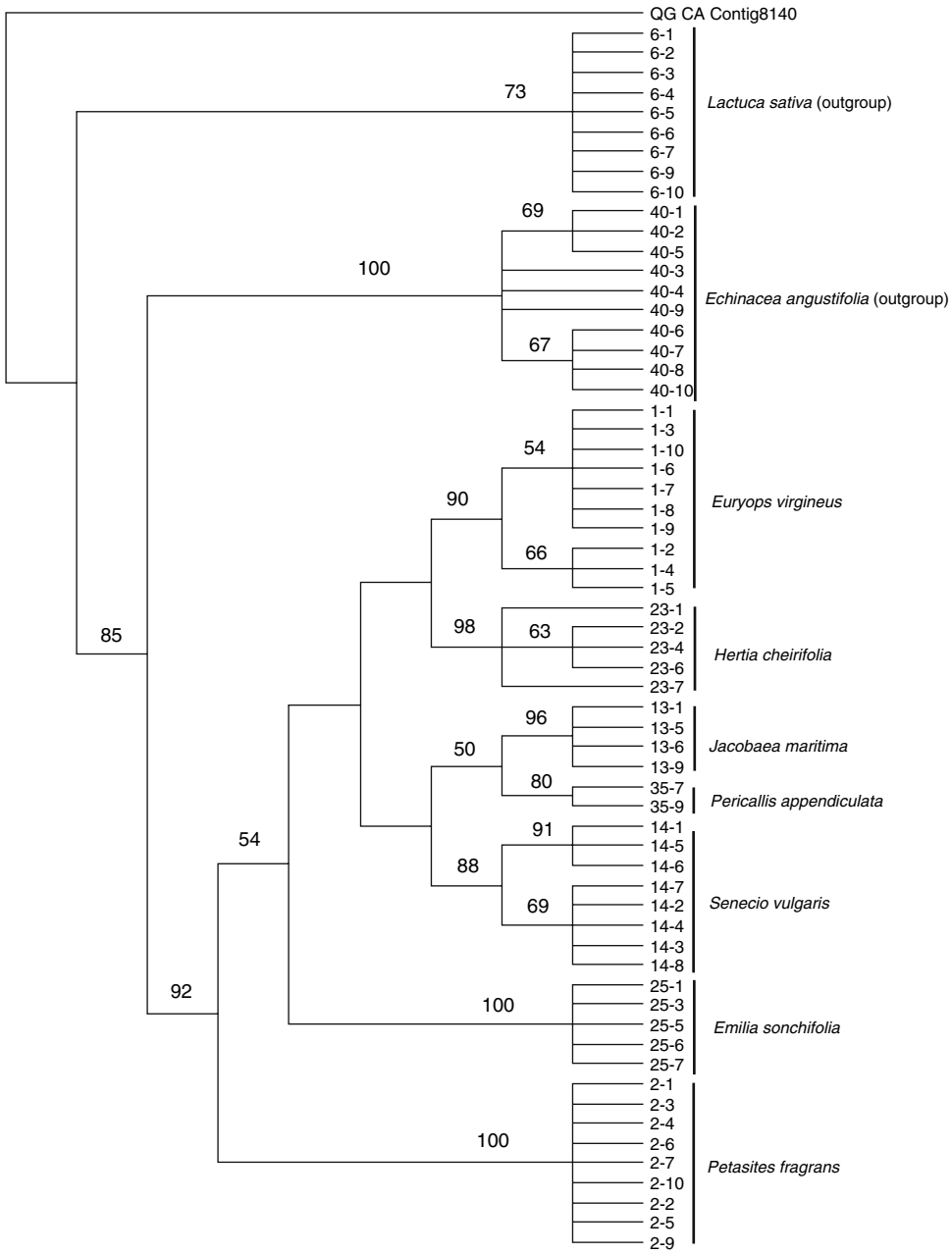


**Fig. 8** Strict consensus of the eight m.p.t. obtained in the analysis of the *QG8140* matrix excluding paralogues and pseudogenes (length = 145; CI = 0.76; RI = 0.96). Bootstrap values ≥50% are indicated above branches

2006). In the present work we have tested a protocol for searching for informative genes using the publicly available nucleotide databases and the BLAST tool, and illustrated the application of this approach with exemplar sampling from the tribe Senecioneae (Asteraceae). The search method was shown to be quite successful, resulting in several potentially useful single-copy nuclear genes; further analysis, however, demonstrated that of the initial candidates, two (*DHS* and *QG8140*) were recommended as phylogenetically most promising. Selection of candidate genes is a challenging process, in that (1) it must balance the number of candidates to test (in our case several hundred) with the laboratory costs and investment of time, and (2) although our strategy is designed to explicitly minimize amplification of paralogues, their presence and the ultimate phylogenetic value of any particular candidate can be confidently assessed only after phylogenetic analysis using some level of exemplar sampling in the group of interest. Here, in addition to the two new genes (*DHS* and *QG8140*), two additional markers previously known within the Asteraceae were tested (*CesA1* and *CHS*). For all these markers different paralogues were identified by phylogenetic analyses. In some cases the presence of several nonsynonymous changes defines a group of paralogues,

although sometimes only a few synonymous changes characterize these sequences. Putative pseudogenes were identified on the basis of stop codons and nucleotide deletions that alter exon structure, and confirmed by phylogenetic analyses. The inclusion/exclusion of such pseudogenes in the phylogenetic analyses does not seem to alter the topology (e.g., by causing long-branch attraction problems) or homoplasy levels. This finding adds to the recently realized minor impact or even utility of pseudogenes in phylogenetic analysis (Razafimandimbison et al. 2004), provided that they are identified (Mayol and Rosselló 2001). Despite characterization of paralogues, specific primer design for each kind was not possible due to low levels of sequence variation in conserved regions. Based on the preliminary sampling used, one of the genes selected during the searching process (*QG8140*) was found to be more useful than the two previously used in Asteraceae (*CesA1*, *CHS*). After independent analyses of these four markers for the samples included, only *QG8140* gives a phylogenetic signal mostly congruent with previous hypothesis (Jansen et al. 1990, 1991; Kim et al. 1992; Kim and Jansen 1995; Kadereit and Jeffrey 1996; Pelser et al. 2007), suggesting that this is a useful gene for phylogenetic purposes in the Senecioneae. In general, and even when strictly or mostly orthologous sequences are amplified and sequenced, it will remain necessary to be cognizant of issues of deep coalescence of alleles, PCR-mediated or in vivo allelic recombination, and many other phenomena that can impact apparent phylogenetic signal with nuclear markers. Although using single-copy nuclear genes for phylogenetic analysis remains challenging, it is hoped that the approach described here will be broadly useful in efforts to implement these powerful tools in other groups.

**Acknowledgments** We thank J. F. Wendel for valuable comments on the manuscript, B. Nordenstam and P. B. Pelser for criticism of the experimental design and sampling, R. W. Michelmore for permission to use sequences from the Compositae Genome Project, C. Cotti for lab work, P. B. Pelser for DNA samples, J. Castresana for advice on search protocols, and A. Herrero, J. Leralta, L. Medina, and B. Nordenstam for plant material. This work was funded by the Spanish Ministry of Education and Science (CGL2004-03872).

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410

Álvarez I, Cronn R, Wendel JF (2005) Phylogeny of the New World diploid cottons (*Gossypium* L., Malvaceae) based on sequences of three low-copy nuclear genes. Plant Syst Evol 252:199–214

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogenet Evol 29:417–434

Baldauf SL (1999) A search for the origins of animals and fungi: comparing and combining molecular data. Am Nat 154:S178–S188

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Brown JR (2001) Genomic and phylogenetic perspectives on the evolution of prokaryotes. Syst Biol 50:497–512

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim KJ, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993) Phylogenetics of seed plants—an analysis of nucleotide-sequences from the plastid gene rbcl. Ann Mo Bot Gard 80:528–580

Cronn R, Small RL, Haselkorn T, Wendel JF (2002) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot 89:707–725

Cronn R, Small RL, Haselkorn T, Wendel JF (2003) Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. Evolution 57:2475–2489

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML (2007) Evolutionary dynamics of *Waxy* and the origin of hexaploid *Spartina* species (Poaceae). Mol Phylogenet Evol 43:1040–1055

Graham SW, Olmstead RG (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am J Bot 87:1712–1730

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hare MP (2001) Prospects for nuclear gene phylogeography. Trends Ecol Evol 16:700–706

Hassanin A (2006) Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol Phylogenet Evol 38:100–116

Helariutta Y, Kotilainen M, Elomaa P, Kalkkinen N, Bremer K, Teeri TH, Albert VA (1996) Duplication and functional divergence in the chalcone synthase gene family of Asteraceae: evolution with substrate change and catalytic simplification. Proc Natl Acad Sci USA 93:9033–9038

Jansen RK, Holsinger KE, Michaels HJ, Palmer JD (1990) Phylogenetic analysis of chloroplast DNA restriction site data at higher taxonomic levels - an example from the Asteraceae. Evolution 44:2089–2105

Jansen RK, Michaels HJ, Palmer JD (1991) Phylogeny and character evolution in the Asteraceae based on chloroplast DNA restriction site mapping. Syst Bot 16:98–115

Kadereit JW, Jeffrey C (1996) A preliminary analysis of cpDNA variation in the tribe Senecioneae (Compositae). In: Hind DJN, Beentje HJ (eds) Compositae: systematics. Proceedings of the International Compositae Conference, Kew, 1994. Royal Botanic Gardens, Kew, pp 349–360

Kim KJ, Jansen RK (1995) *Ndhf* sequence evolution and the major clades in the sunflower family. Proc Natl Acad Sci USA 92:10379–10383

Kim KJ, Jansen RK, Wallace RS, Michaels HJ, Palmer JD (1992) Phylogenetic implications of rbcL sequence variation in the Asteraceae. Ann Mo Bot Gard 79:428–445

Knox EB, Kowal RR (1993) Chromosome-numbers of the East-African giant Senecios and giant Lobelias and their evolutionary significance. Am J Bot 80:847–853

Lawrence ME (1980) *Senecio* L (Asteraceae) in Australia—chromosome numbers and the occurrence of polyploidy. Aust J Bot 28:151–165

Liu JQ (2004) Uniformity of karyotypes in *Ligularia* (Asteraceae: Senecioneae), a highly diversified genus of the eastern Qinghai-Tibet Plateau highlands and adjacent areas. Bot J Linn Soc 144:329–342

López MG, Wulff AF, Poggio L, Xifreda CC (2005) Chromosome numbers and meiotic studies in species of *Senecio* (Asteraceae) from Argentina. Bot J Linn Soc 148:465–474

Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286:947–950

Mayol M, Rosselló JA (2001) Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. Mol Phylogenet Evol 19:167–176

Mort ME, Crawford DJ (2004) The continuing search: low-copy nuclear sequences for lower-level plant molecular phylogenetic studies. Taxon 53:257–261

Nei M, Li WH (1979) Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Nordenstam B (1977) Senecioneae and Liabeae—systematic review. In: Harborne JB, Turner BL (eds) The biology and chemistry of the Compositae. Academic Press, London, pp 799–830

Nozaki H, Matsuzaki M, Takahara M, Misumi O, Kuroiwa H, Hasegawa M, Shin-i T, Kohara Y, Ogasawara N, Kuroiwa T (2003). The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J Mol Evol 56:485–497

Pelser PB, Nordenstam B, Kadereit JW, Watson LE (2007) An ITS phylogeny of the tribe Senecioneae (Asteraceae) and a new delimitation of *Senecio* L. Taxon 56:1077–1104

Razafimandimbison SG, Kellogg EA, Bremer B (2004) Recent origin and phylogenetic utility of divergent ITS putative pseudogenes: a case study from Naucleeae (Rubiaceae). Syst Biol 53:177–192

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: Aa laboratory manual. Cold Spring Harbor Press, Cold Spring Harbor, NY

Sang T (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit Rev Biochem Mol 37:121–147

Schlegel M (2003) Phylogeny of eukaryotes recovered with molecular data: highlights and pitfalls. Eur J Protistol 39:113–122

Schlüter PM, Stuessy T, Paulus HF (2005) Making the first step: practical considerations for the isolation of low-copy nuclear sequence markers. Taxon 54:766–770

Senchina DS, Álvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol 20:633–643

Small RL, Cronn R, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. Aust Syst Bot 17:145–170

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404

Strand AE, LeebensMack J, Milligan BG (1997) Nuclear DNA-based markers for plant evolutionary biology. Mol Ecol 6:113–118

Swofford DL (1999) PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.02b. Sinauer, Sunderland, MA

Van de Peer Y, De Wachter R (1997) Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. J Mol Evol 45:619–630

Van de Peer Y, Neefs JM, De Wachter R (1990) Small ribosomal subunit RNA sequences, evolutionary relationships among different life forms, and mitochondrial origins. J Mol Evol 30:463–476

Wu F, Mueller LA, Crouzillat D, Pétiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single copy, orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid plant clade. Genetics 174(3):1407–1420

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol 12:563–584