Cladistics

Cladistics 24 (2008) 727–745

10.1111/j.1096-0031.2008.00207.x

Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species

Mingai Li[a,*†], Jörg Wunder[a,†], Gaetano Bissoli[a,b], Eliana Scarponi[a], Silvia Gazzani[a], Enrico Barbaro[a], Heinz Saedler[a,c] and Claudio Varotto[a]

[a]*Center for the Study of Biodiversity –Trentino, Institute for Agriculture Research of San Michele all'Adige, Trento, Italy;* [b]*Instituto de Biología Molecular y Celular de Plantas, Universidad Politécnica de Valencia-CSIC, Valencia, Spain;* [c]*Department of Molecular Plant Genetics, Max-Planck-Institute for Plant Breeding Research, Cologne, Germany*

Accepted 26 November 2007

Abstract

With the aim of developing widely applicable gene markers for phylogenetic reconstructions at low taxonomic level, we tested the low copy nuclear Conserved Ortholog Set (COS) genes. Most of the 15 genes tested provided good amplification efficiency (as compared with *rbcL*) from a set of 67 representative angiosperm families. Nine selected COS markers were further characterized at both intra- and interfamilial level on a test set, including 25 species representative of 15 different families. While four of the COS led to incongruent results, the remaining five improved the phylogenetic reconstructions of closely related species as illustrated in the case of Orobanchaceae species. They were found to be highly informative in phylogenetic reconstruction of congeneric species, where introns provide a higher proportion of parsimony informative sites in comparison with traditional phylogenetic markers such as *ITS* and *matK*. At higher phylogenetic distance, where only coding regions could be aligned, the polymorphism levels of the COS ranged between those of *ndhF* and *matK*.

On the basis of these results, the success rate in developing universally amplifiable low copy nuclear markers based on COS genes is about 30%. We report the successful development of five pCOS that, together with a few other well characterized genes, such as *Rpb2* and *GbssI*, can be considered the closest approximation to low-copy "universally" amplifiable markers for phylogeny in plants at present. The possible pitfalls of universally amplifiable COS marker development and their range of applicability at different taxonomic levels in comparison with traditional phylogenetic molecular markers are discussed.

Owing to the availability of universal primers to reliably amplify them from a wide array of taxa and to the fact that they are well characterized, chloroplast and ribosomal genes (e.g., *matK*, Hilu et al., 2003; *ITS*, Soltis et al., 1997, 2000) account for nearly 90% of 345 data matrices from 136 species-level papers surveyed in a recent study (Hughes et al., 2006). For studies dealing with hybridization, polyploidy, character evolution, rapid evolutive radiations, speciation and domestication of closely related species (Cronn et al., 2002b; Doyle et al., 2003; Bailey et al., 2004; Linder and Rieseberg, 2004; Alvarez et al., 2005), however, chloroplast DNA (cpDNA) and nuclear ribosomal genes (nrDNA) may not be sufficiently variable to provide good resolution (Despres et al., 2003; Pelser et al., 2003; Hughes et al., 2006) or are affected by variable levels of concerted evolution (Baldwin et al., 1995; Wendel et al., 1995; Doyle et al., 2004). Low-copy nuclear gene (LCNG) loci have thus received the most attention to aid in phylogenetic reconstruction in such cases or to solve incongruence between cpDNA or nrDNA.

Examples of LCNGs used for phylogenetic reconstructions at the intergeneric level are *Ccr* (cinnamoyl CoA reductase; Poke et al., 2006) and *ncpGS* (chloroplast-expressed glutamine synthetase; Doyle et al., 2003); examples of LCNGs for phylogenetic reconstructions at the interspecific or intraspecific level include *Sam* (S-adenosyl methionine synthetase; Londo et al.,

*Corresponding author:
E-mail address: claudio.varotto@iasma.it
†These authors equally contributed to this work.

2006) and *PgiC* (cytosolic phosphoglucose isomerase; Ford et al., 2006).

Although still relatively scarce, during the last few years a growing number of studies has been published that used multiple independent nuclear loci for phylogenetic reconstruction (for a comprehensive discussion see reviews by Sang, 2002; Small et al., 2004; Schlueter et al., 2005; Hughes et al., 2006). Ideally, more than one sequential marker should be used to take into account the fact that gene trees do not necessarily represent the species tree because of random sorting of polymorphic alleles in different lineages (Maddison, 1997). The markers should also be easily amplifiable from a wide array of taxonomic groups ("universal") and at the same time evolve fast enough to be informative at the desired taxonomic level (Sang, 2002). The current progress of large genomic projects offers new possibilities to approach this problem (see, e.g., Scherson et al., 2005; Choi et al., 2006). Through a comparative approach using a comprehensive collection of tomato ESTs and the Arabidopsis genomic sequence, Fulton et al. (2002) pioneered the identification of genome-wide sets of conserved orthologous genes in plants. These markers, named COS (Conserved Ortholog Set), were demonstrated to be single or low copy in both genomes and to have remained relatively stable in sequence since the early radiation of dicotyledonous plants. Further refinements of sets of COS markers have been independently carried out by Kozik and Michelmore on a set of six EST collections (http://cgpdb.ucdavis.edu/COS_Markers/COS_Markers.html) and by Rudd et al. (2005) on sequence data from over 50 plant species. More recently, Wu et al. (2006) developed a second set of COS markers (COSII) and demonstrated their utility for both comparative mapping and phylogenetic reconstruction in euasterids I. In this study we describe the selection, development and testing across various eudicot orders of 15 candidate phylogeny COS markers (pCOS) selected from the set developed by Kozik and Michelmore. We report the successful development of five pCOS that, together with a few other well characterized genes such as *Rpb2* and *GbssI* (Denton et al., 1998), can be considered the closest approximation to low-copy "universally" amplifiable markers for phylogeny in plants at present. The possible pitfalls in the development of such molecular markers and their range of applicability at different taxonomic levels in comparison with traditional phylogenetic molecular markers are discussed.

Materials and methods

Plant material

Leaf material was collected from natural populations on the south-eastern range of the Alps (Trentino-Alto Adige) with the exception of *Antirrhinum* species (provided by the Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany). DNA extraction was performed either with standard CTAB extraction (Doyle and Doyle, 1987) or by means of the Plant DNEasy extraction kit (Qiagen, Hilden, Germany). The complete list of species can be found in Appendix 1.

Selection of COS genes and primer design

The original set of COS markers as elaborated by Kozik and Michelmore consisted of 2343 Arabidopsis genes having a single blast hit (*e*-value = 1E − 20) to at least one EST in one of the six species analyzed (maize, rice, soybean, tomato, sunflower and lettuce; http://cgpdb.ucdavis.edu/COS_Arabidopsis). This data set was reduced to 95 genes that had: (1) a blast hit to each maize, rice, soybean and tomato, and (2) a literature reference defining their function. The 20 top-ranking genes according to, in order of priority, blast hit length, percentage identity, standard deviation and mutant phenotypes in higher plants or attribution of clear biological function were used for primer design. Blast searches were used to query the NCBI databases and retrieve homologous sequences from angiosperm species. ClustalW multiple sequence alignments (Thompson et al., 1994) were manually edited using the program BioEdit (Hall, 1999). Owing to the different representation and degree of nucleotidic conservation of homologs of this set of genes, degenerate primers were designed for 15 of the 20 genes (see in Appendix 2 the list of the selected 15 COS markers, their functions and primers used; Fig. 1a shows a schematic graphic representation of structures for the five markers that were characterized in depth in this study). Primer design was carried out manually according to the general guidelines described in the literature (Strand et al., 1997).

Polymerase chain reaction amplification and sequencing

Polymerase chain reaction (PCR) amplifications were performed according to two different basic protocols. In the first protocol, cycling conditions were: initial template denaturation 95 °C for 2 sec, followed by 33–35 cycles of 94 °C for 40 sec, 50–57 °C for 30 sec, 72 °C for 40 sec, 72 °C for 5 min. In the second protocol, used in case of failure or low amplification by means of the former one, cycling conditions were the same, except that the number of cycles was reduced to 25. A 10-fold dilution was used as the template for a second round of PCR with all of the cycling parameters identical to those used in the first round except for the cycling number ranging from 25 to 30.

PCR products were cloned in the pGEM-T vector (Promega Corp., Madison, WI) and sequenced from both directions. Sequences used in this study have been

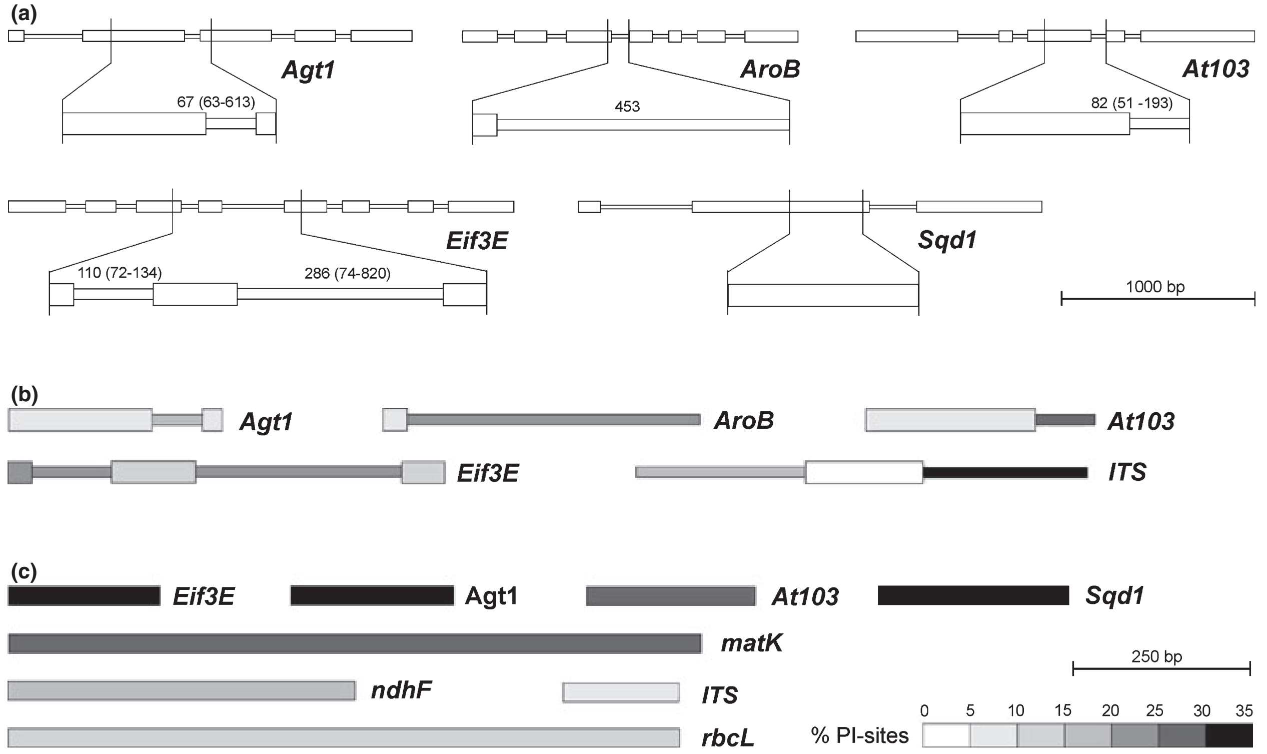


Fig. 1. Structure and variability of pCOS. (a) Position of amplified pCOS sequences (blow-up) with respect to *Arabidopsis thaliana* genomic DNA. Numbers indicate the intron length corresponding to the alignment of seven *Melampyrum* species, in brackets the length range according to the "Class-Set" of 25 different angiosperm species. Scale of 1000 bp refers to the Arabidopsis sequences. (b) Variability of pCOS genes compared with ITS corresponding to the alignment of seven *Melampyrum* species. Exon and intron boxes are shaded according to the level of polymorphisms (% parsimony informative sites, values for the variability of the entire pCOS amplicons are shown in Table 2). (c) Variability of pCOS genes compared with traditional markers corresponding to the alignment of 12 angiosperm species (Class-Set_Ang-12). For *Agt1*, *Eif3E* and *At103* only the exons were used. The scale of 250 bp and the shading-scale of percentage PI sites refers to both (b) and (c).

deposited in GenBank with accession numbers from AM503637 to AM503893.

Data analysis

Multiple sequence alignments obtained with ClustalW (Thompson et al., 1994) were manually refined using the program BioEdit (Hall, 1999) in particular for insertion/deletion (indel) structure of the intron regions. To decide when alignments of divergent sequences (order and above) became ambiguous, ClustalW alignments with an increasing number of species were performed. Alignments were visually inspected to identify the taxonomic distances after which the addition of a new species significantly decreased alignment quality despite using different gap opening and extension penalties and manual editing. To provide a quantitatively uniform criterium for assessment of alignment quality, the three alignments at lower and higher taxonomic distance compared with the one considered critical were subjected to a sliding window analysis of polymorphism level with DnaSP v4.10 (Rozas et al., 2003). A 10 bp sliding window was moved in 1 bp steps along the alignment. As the average polymorphism level for alignment of random sequences was about 70%, alignments with more than 50% of their length displaying polymorphism levels above this threshold were considered ambiguous. Alignments being ambiguous based on this criterium are depicted in gray in Fig. 2.

The analysis of the indel structure resulting from the alignment procedure was carried out with the help of GapCoder (Young and Healy, 2003). DnaSP was used to perform polymorphism analysis and other gene and alignment characterizations like G + C content.

For phylogenetic analysis and tree-building using maximum parsimony (MP) and maximum likelihood (ML) PAUP* version 4.0b10 (Swofford, 2003) was used. Trees were calculated with swap = TBR, addition = random, hsearch replicates = 1000, trees hold at each step = 1, collapse = MaxBrLen, ti/tv ratio 2 : 1, gaps were treated as missing; for bootstrap analysis: replicates = 1000, hsearch = 100. ML was performed according to the best substitution model obtained by applying hierarchical likelihood ratio tests implemented in Modeltest 3.7 (Posada and Crandall, 1998). ML settings were as shown above for MP bootstrap analysis with starting tree(s) = stepwise addition, number of replicates = 1000. For ML bootstrap analysis: analysis

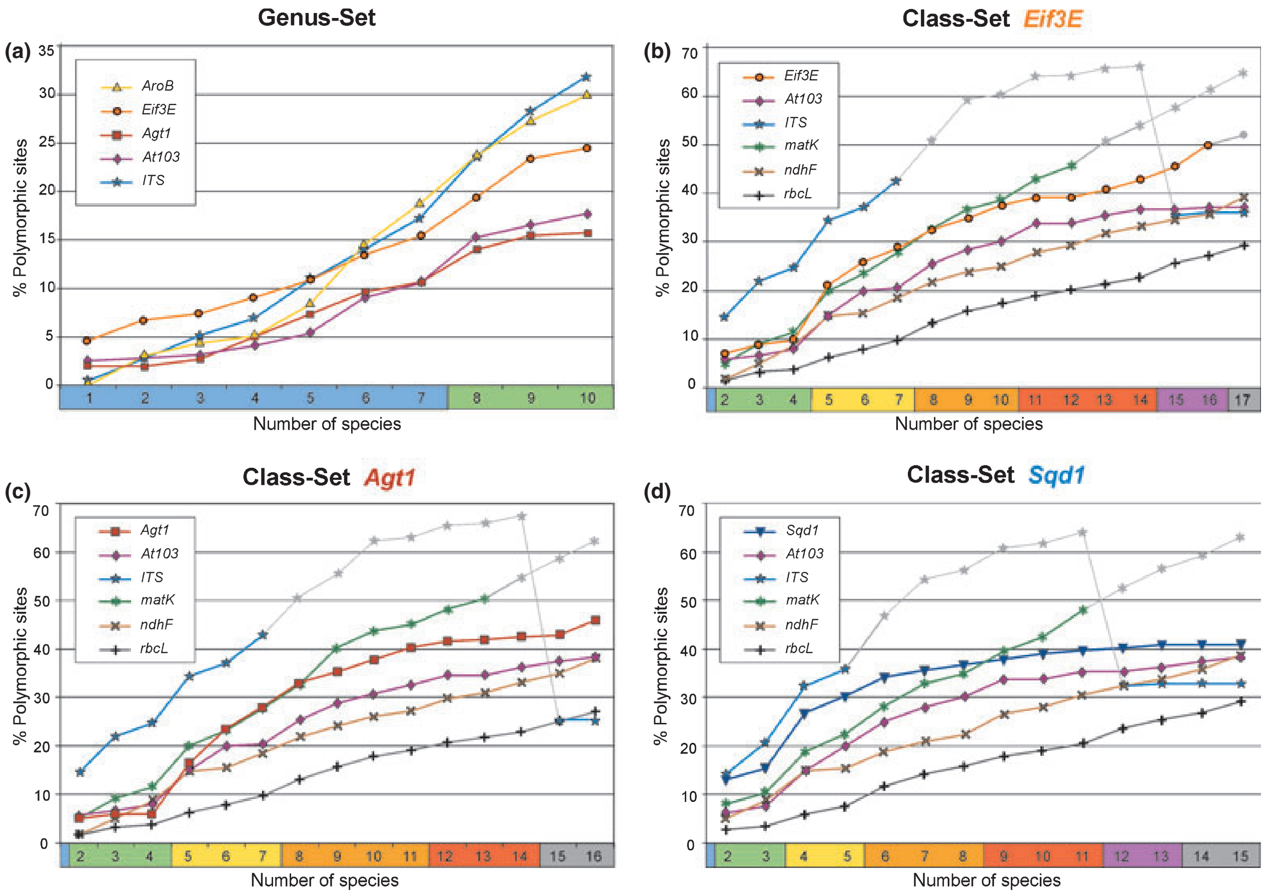


Fig. 2. Variability of pCOS and traditional genes at different taxonomic levels. The curves indicate the level of polymorphism in alignments encompassing species at increasing taxonomic distances. The numbers on the *x*-axis refer to the number of species included in the alignment. Each sequence added represents another species, genus, family, order or class, which leads to a cumulation of taxonomic distances from left to right. The *y*-axis indicates the level of polymorphism within each alignment (% polymorphic sites) in relation to the alignment length. The color bars below the *x*-axis refer to different taxonomic levels: [(*blue* = *Melampyrum*) + *green*] = Orobanchaceae + *yellow* = Lamiales (Lamiales + *orange*) = Asterids, *red* = Rosids, *purple* = basal Eudicots, *gray* = Monocots. (a) "Genus-Set" includes seven *Melampyrum* species each represented by three populations (1–7) and three outgroup species (8–10), see also Fig. 1(b). (b–d) These denote different subsets of the "Class-Set" representing 25 different species, genera, families, orders and classes throughout the angiosperms. The genus *Melampyrum* is only represented by *M. sylvaticum*. *At103*, *ITS*, *ndhF* and *rbcL* are present in all three figures. Curves in gray indicate that the alignment of the sequences for this marker was ambiguous at this taxonomic level. The drop of the *ITS* curve at high taxonomic levels is due to the restriction of the alignment to the *5.8S* ribosomal DNA.

replicates = 100, search replicates = 10. Neighbor-Joining analysis was performed using TreeCon Vers. 1.3b (Van de Peer and De Wachter, 1997), distance = k2p (Kimura, 1980), indels were considered as a single SNP not taking into account the length of the inserted/deleted fragment, bootstrap replicates = 1000. Bayesian inference was calculated using MrBayes 3 (Ronquist and Huelsenbeck, 2003) using the default settings and the same model as for ML with 100 000 generations.

Results

Amplification efficiency of COS markers

The primers developed for 15 selected COS markers were prescreened on an initial set of 10 Orobanchaceae and seven Plantaginaceae (Olmstead et al., 2001; Bennett and Mathews, 2006) species (Test-Set-1; Appendix 1) and the corresponding products sequenced. The nine primer combinations amplifying from more than 50% of these species were then used to test amplification from a total of 87 species, representing 67 families mostly distributed across dicotyledonous angiosperms, with a few representatives from monocotyledonous angiosperms, gymnosperms, bryophytes and pteridophytes (Test-Set-2; Appendix 1). The amplification results were compared with those obtained for *rbcL*, one of the most commonly used universally amplifiable phylogenetic markers (Table 1). The numbers reported in Table 1 refer to the percentage of families (total 67) and species (total 87) from which PCR products resulted on agarose gels either as single bands (columns "Class I",

Table 1
Summary of PCR amplification for nine selected COS markers across 67 angiosperm families. Amplification efficiencies are reported as percentage with respect to total number of families and species tested

			Families (%)		Species (%)		
Name	AGI-ID	Function	Class I	Class II	Class I	Class II	Reference
*Agt1**	AT2G13360	Encodes a peroxisomal photorespiratory enzyme that catalyzes transamination reactions with multiple substrates. It is involved in photorespiration	58	64	52	59	Liepman and Olsen (2003)
Apg1	AT3G63410	Responsible the methylation step of plastoquinone biosynthesis. The gene product is also involved in tocopherol (vitamin E) biosynthesis	42	51	40	47	Cheng et al. (2003)
*AroB**	AT5G66120	Putative 3-dehydroquinate synthase activity, aromatic amino acid family biosynthesis	31	42	28	37	Barten and Meyer (1998)
*At103**	AT3G56940	Chlorophyll biosynthesis: magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase activity	60	69	57	66	Rzeznicka et al. (2005)
ChlP	AT1G74470	Multifunctional protein involved in the biosynthesis of chlorophyll geranylgeranyl moiety. It catalyzes the reduction of prenylated geranylgeranyl-chlorophyll *a* to phytyl-chlorophyll *a* (chlorophyll *a*) and free geranylgeranyl pyrophosphate to phytyl pyrophosphate	58	79	54	74	Keller et al. (1998)
*Eif3E**	AT3G57290	Translation initiation factor activity, transcription initiation, part of the eukaryotic translation initiation factor 3 complex and associated with subunits of the COP9 signalosome	36	45	30	39	Yahalom et al. (2001)
Gi	AT1G22770	GIGANTEA locus. Response to cold, hydrogen peroxide, regulation of circadian rhythm, positive regulation of long-day photoperiodism, flowering, flower development	69	79	61	74	Oliverio et al. (2007)
Hmgs	AT4G11820	Acetyl-CoA C-acetyltransferase activity, hydroxymethylglutaryl-CoA synthase activity, isopentenyl diphosphate biosynthetic process, mevalonate pathway	34	48	32	44	Montamat et al. (1995)
*Sqd1**	AT4G33030	Chloroplast, cellular response to phosphate starvation, UDP sulfoquinovose synthase activity, sulfolipid biosynthesis	87	91	80	86	Essigmann et al. (1999)
rbcL	ATCG00490	Ribulose-bisphosphate carboxylase activity, fixation of carbon dioxide	76	85	80	87	
Average			53	63	48	58	
SD			18	18	17	18	
% relative to *rbcL*			69	74	60	67	

Class I, single band on agarose gels; Class II, one major band with a maximum contamination of about 30% from secondary products.
*Markers that after phylogenetic validation have been collectively called pCOS.

amenable for direct sequencing) or as a major band with a maximum contamination of about 30% from secondary products (columns "Class II", amenable for sequencing upon cloning). The average amplification efficiency from the different families tested was 69% and 74% as compared with *rbcL* for class I or class II, respectively. The same comparison at the level of number of species yielded slightly lower values (60% and 67%, respectively), thus indicating a more homogeneous amplification of *rbcL* among congeneric species as compared with COS. Low levels of linear correlation ($r \leqslant 0.394$), were observed among COS markers on the basis of the binary coding (1/0) of amplifiable/non-amplifiable species across the whole test data set, thus indicating low family specific amplification by the primers.

Sequence characterization of COS markers compared with traditional phylogenetic markers

The nine COS markers selected were further characterized and compared with the traditional markers *ITS*, *matK*, *rbcL* and *ndhF*. Five COS markers (see below) were amplified and sequenced for a set of 25 species sampled from selected orders and families throughout the angiosperms (hereafter called "Class-Set"; see Appendix 1). *At103*, *Eif3E*, *Agt1*, *Sqd1* and *ChlP* were successfully confirmed by sequencing from, respectively, 25, 23, 22, 19 and 22 species of the Class-Set. The remaining four genes (*Apg1*, *AroB*, *Gi* and *Hmgs*) were sequenced for a subset of the Class-Set encompassing seven *Melampyrum* species, each represented by three populations, and three outgroup species (Orobanchaceae) represented by one population each ("Genus-Set"; Appendix 3). Sequencing confirmed the successful isolation of these genes from all the species with the exception of *Sqd1* from two *Melampyrum* species.

Preliminary phylogenetic analyses led to the exclusion of four genes from further analysis due to paralogy problems. The remaining five genes (*Agt1*, *AroB*, *At103*, *Eif3E* and *Sqd1*) were characterized in depth, in particular with respect to their level of polymorphism, from sequence alignments of the individuals and species within the *Melampyrum* genus (Intra-Generic-Set; Table 2) and from those in the Class-Set in common to all markers (Class-Set_Ang-12; Table 3). A schematic representation of the genomic organization of these five genes in *A. thaliana* is shown in Fig. 1(a), while Fig. 1(b,c) show the nucleotide variability of coding and non-coding regions for each of the genes analyzed at genus and class level, respectively.

The COS markers developed in this study were shorter than traditional markers, ranging from 302 bp to 540 bp in the genus used as reference. In other genera and families, however, the length that resulted was sometimes dramatically higher due to a higher proportion of non-coding regions (Tables 2 and 3). The COS markers varied from a minimum of 7% for *AroB* to a maximum of 100% of coding regions for *Sqd1*.

The variability of the COS genes in terms of polymorphic sites estimated as a percentage of variable or parsimony informative sites and indels for the alignments obtained for either the Intra-Generic-Set or the Class-Set is summarized in Tables 2 and 3, respectively.

In both the Intra-Generic-Set and the Class-Set, COS markers displayed a lower number of polymorphic sites, in absolute terms, compared with that of the traditional marker *ITS*, as expected from their lower absolute sequence lengths (Table 2). However, the proportions of polymorphic sites of *Eif3E* and *AroB* in the Intra-Generic-Set are similar to that of *ITS* (89% and 109%, respectively). The number of polymorphisms or parsimony informative sites for the other two COS (*Agt1* and *At103*) in comparison with that of *ITS* was about 60%.

In the Class-Set the amount of nucleotide variation per site of COS averaged from 66% of *At103* to 86% of *Eif3E* compared with *matK* (*matK* was used because *ITS* sequences were not alignable at high taxonomic level). Notably, the fraction of parsimony informative sites in the Class-Set was higher for all COS in comparison with that of *matK*.

Table 2
Variability of selected COS genes at the intrageneric level. Data refer to the Intra-Generic-Set (21 individuals, seven *Melampyrum* species, three populations/species)

	Length				Polymorphic sites		PI-sites		Substitutions	
Gene	Sequences	Alignment	Coding (%)	Indels	Absolute	Relative to *ITS*	Absolute	Relative to *ITS*	Non-synonymous	Synonymous
Agt1	274-301	301	45-87	2	32	0.62	28	0.66	0.017	0.059
AroB	302-438	448	7-11	15	84	1.09	75	1.19	0.030	0.030
At103	291-321	322	75-86	13	34	0.61	26	0.57	0.000	0.079
Eif3E	482-540	618	42-47	26	95	0.89	79	0.91	0.022	0.111
ITS	575-599	619	27-29	23	107	1.00	87	1.00	0.018	0.006

PI sites, parsimony informative sites; substitutions were calculated per site.

Table 3
Variability of selected COS genes at the family order level. Data refer to the Class-Set_Ang-12 (12 angiosperm families)

	Length				Polymorphic sites		PI sites		Substitutions	
Gene	Sequences	Alignment	Coding (%)	Indels	Absolute	Relative to matK	Absolute	Relative to matK	Non-synonymous	Synonymous
Eif3E	481–1140	227	20–59	0	105	0.86	73	1.20	0101	0.357
Agt1	300–972	238	28–79	0	106	0.83	81	1.27	0044	0.461
At103	298–429	240	55–79	0	86	0.66	67	1.04	0032	0.380
Sqd1	267	267	100	0	107	0.74	91	1.28	0045	0.468
ITS	558–710	158	24–100	0	33	0.39	10	0.24	0120	0.073
matK	906–939	980	100	20	529	1.00	262	1.00	0226	0.190
ndhF	482–492	493	100	5	172	0.64	96	0.73	0088	0.248
rbcL	976	976	100	0	236	0.45	119	0.46	0027	0.217

PI sites, parsimony informative sites; substitutions were calculated per site. "Coding (%)" refers to the full sequence length, while the other values to coding regions.

The curves shown in Fig. 2 indicate the change in the proportion of polymorphic sites by stepwise addition of a single species to the alignment of each data set (Fig. 2a: Genus-Set; Fig. 2b–d: Class-Set). Segments in gray indicate that for the corresponding species alignment was ambiguous (see Materials and methods). For the genus-level comparison, both coding and non-coding regions from COS were used (Fig. 2a), while at the class level only exons were used (Fig. 2b–d). In general, COS markers with the highest nucleotide diversity such as *AroB* and *Eif3E* resulted as polymorphic (*AroB*) or more polymorphic (*Eif3E*) than *ITS* for closely related congeneric species (Genus-Set; Fig. 2a). At moderate phylogenetic distance (different families within an order), *Eif3E* and *Agt1* displayed a lower percentage of polymorphic sites as compared with *ITS*, with *Sqd1* following closely the *ITS* trend (Fig. 2b–d). *At103* displayed relative polymorphism levels higher than those of *ndhF*, but lower than those of *matK*. At higher taxonomic distances (different orders within a subclass), where first *ITS* and then *matK* alignments became ambiguous, *Agt1*, *At103* and *Sqd1* tended to reach a plateau at about 40% of polymorphic sites, while the *Eif3E* curve, after a flexion of about the 40% value, continued to increase steadily.

Within the *Melampyrum* genus, the number of indels, despite being correlated to the absolute number of polymorphic sites, did not correlate to the fraction of non-coding regions present in the alignment of each gene (Table 2). For the Class-Set the introns were excluded from the alignment and no deletions were observed in any of the COS.

Phylogenetic reconstruction with COS markers at different phylogenetic levels

The nine COS markers selected were further characterized to test their applicability for phylogenetic studies compared with the traditional markers *ITS*, *matK*, *rbcL* and *ndhF*.

Phylogenetic reconstructions were carried out for the Class-Set (and in some cases for considerably larger species sets of Orobanchaceae, Plantaginaceae and Solanaceae) with *At103*, *Eif3E*, *Agt1*, *Sqd1* and *ChlP* sequences. *ChlP* had to be excluded due to paralogy problems. For the remaining genes, phylogenetic reconstruction for taxa within families was in agreement with previous studies (data not shown).

The other four COS genes (*Apg1*, *AroB*, *Gi* and *Hmgs*) were sequenced only for the Genus-Set because of their large intron fraction, which was unalignable at higher taxonomic distance. This choice was motivated by the aim of testing the COS markers as extensively as possible on closely related, congeneric species, where they could potentially be most useful.

Three of these genes (*Apg1*, *Gi* and *Hmgs*) were not suitable for phylogenetic reconstruction due to amplification of paralogous sequences (data not shown). On the other hand, *AroB*, like *Agt1*, *Eif3E* and *At103* provided consistent, even though partly incongruent, phylogenetic reconstructions on this set of species, as assessed by the consistency of tree topology for sequences derived from different populations within the same species (Figs 3 and 4). The levels of incongruence and the resolution among congeneric species within the Genus-Set was similar to that of *ITS* for *AroB* and *Eif3E*, while it was lower for *Agt1* and *At103*. Topological incongruence correlated with the absolute amount of parsimony informative sites.

After their experimental validation, we refer to the five COS markers useful in phylogenetic reconstruction (*Agt1*, *AroB*, *Eif3E*, *At103* and *Sqd1*) as pCOS (phylogeny-COS) to distinguish them from COS markers that were not validated on this basis.

Phylogenetic reconstruction at low taxonomic levels

On the basis of the low resolution at higher taxonomic distances, we concentrated on testing the newly developed pCOS on a test case of closely related species within the genus *Melampyrum*. In order to assess the robustness of the phylogenetic reconstruction at low evolutionary distances, intraspecific variability was estimated by analyzing three different populations for each species. The phylogenetic reconstructions based on two of the most widely used marker at this taxonomic level (*ITS* and *matK*) were compared with the tree obtained from the concatenated pCOS markers. In the absence of a clearly resolved phylogeny with *ITS* and *matK* for the species considered, in fact, we considered a total evidence approach best suited to summarize the data. The analysis of the phylogenetic reconstructions resulting from all the different data partitions confirmed that concatenation provided, overall, the most resolved phylogeny for the species used as a test case (data not shown). In contrast to *ITS* alone, the pCOS data set showed a highly resolved phylogeny with the exception of the terminal clade encompassing *M. italicum*, *M. velebiticum* and *M. sylvaticum*. In particular, this clade, not significantly supported by *ITS*, was clearly supported in the concatenated data set. The placement of the most basal species considered, *M. cristatum*, *M. pratense* and *M. arvense*, not resolved by *ITS* alone, showed high to moderate bootstrapping values in the trees originating from the combined data set.

The complete data set originated from the concatenation of all the markers (four pCOS and *ITS*) provided a highly supported phylogenetic reconstruction for all the species with all the algorithms used, with the exception of one *M. italicum* population grouped with

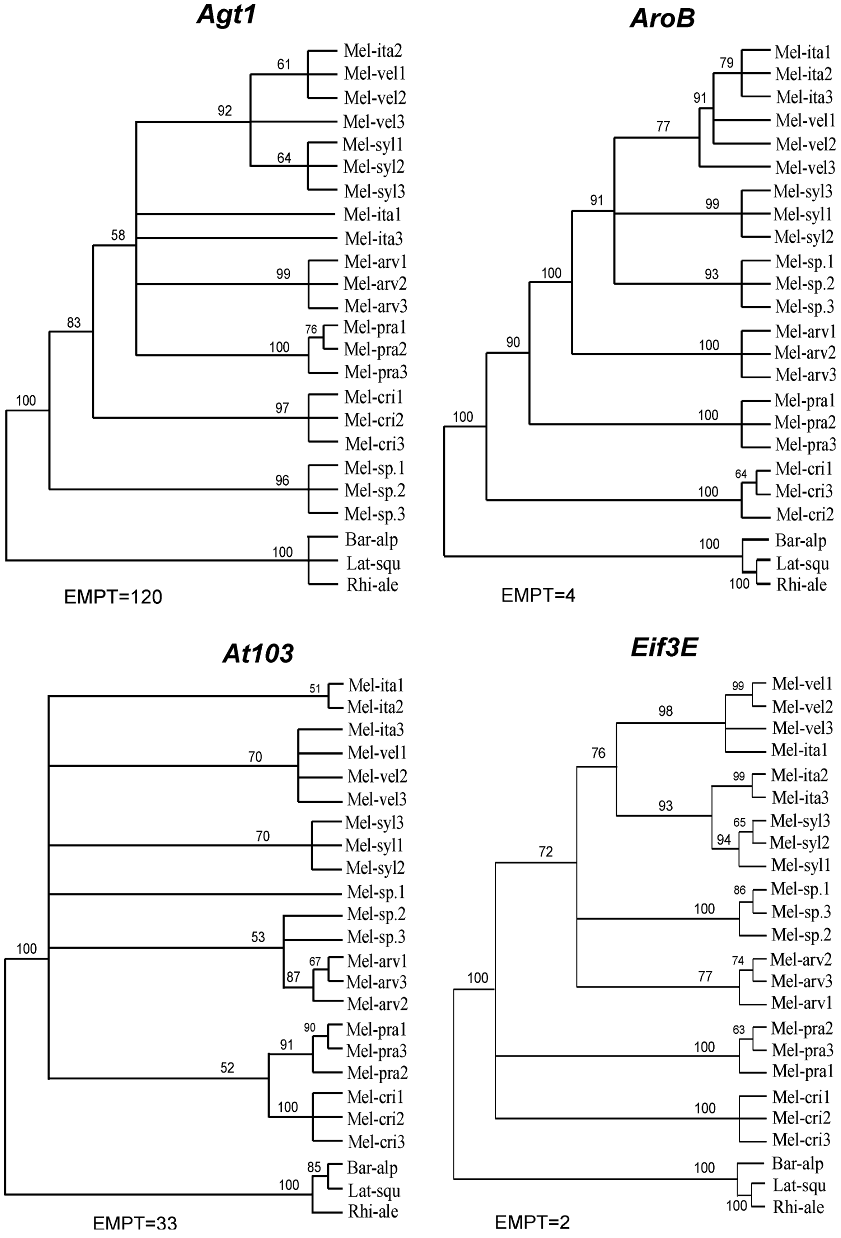


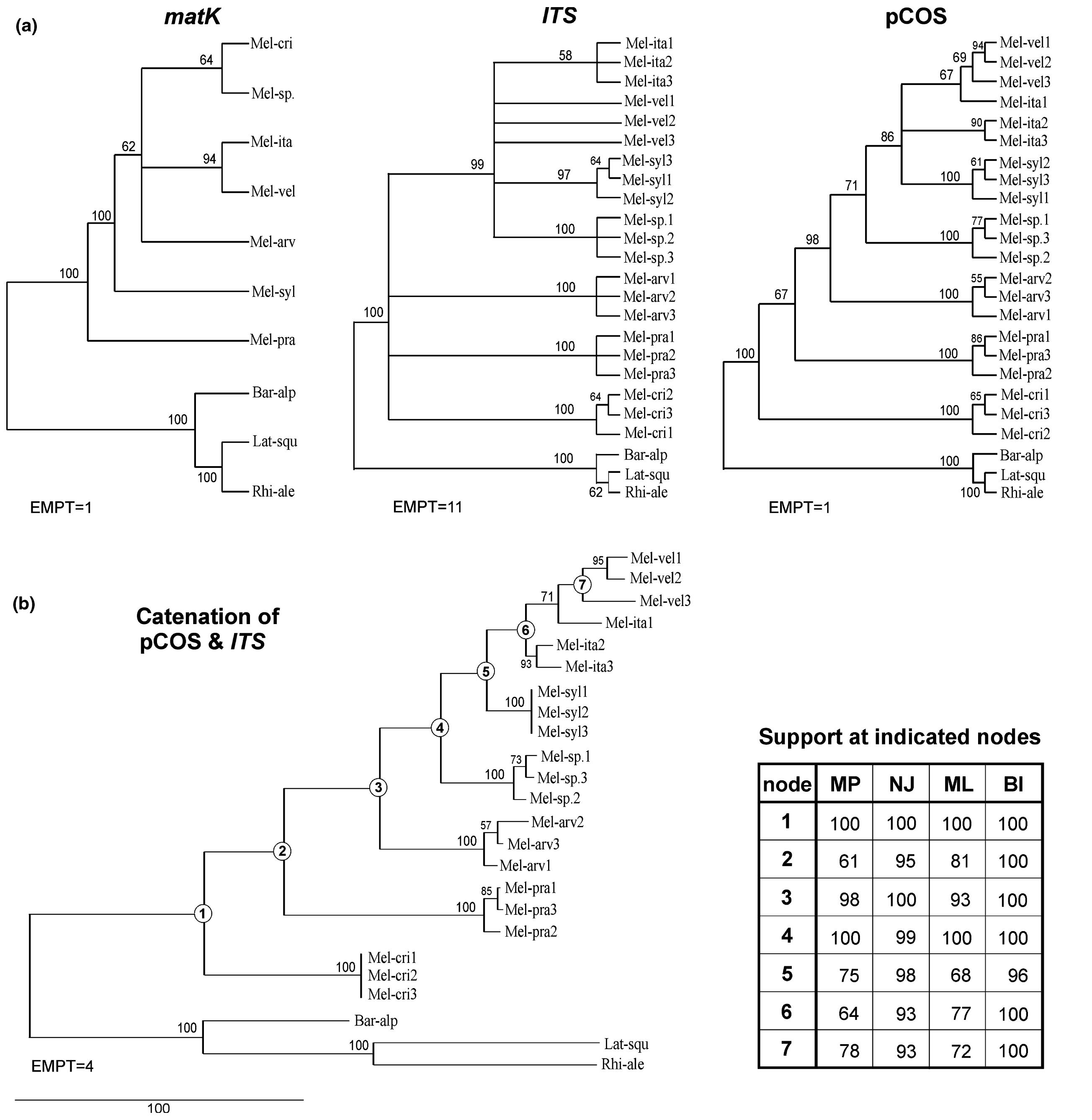
Fig. 3. Cladograms of single genes representing seven *Melampyrum* species, rooted by three outgroup species (Genus-Set). The genes and cladograms (MP) are characterized by the following features: *AroB*: 486 bp, PI = 106, TL = 176, CFI = 0.4; *Agt1*: 305 bp, PI = 35, TL = 60, CFI = 0,1; *At103*: 322 bp, PI = 43, TL = 82, CFI = 0.4; *Eif3E*: 623 bp, PI = 111, TL = 217, CFI = 0.6; *ITS*: 642 bp, PI = 124, TL = 322, CFI = 0.5. For all trees, branches with less than 50% support are drawn as unresolved. MP, maximum parsimony; PI, parsimony informative; TL, tree length; CFI, consensus fork index; EMPT, equally most parsimonious trees. Mel-ita, *Melampyrum italicum*; Mel-vel, *M. velebiticum*; Mel-syl, *M. sylvaticum*; Mel-arv, *M. arvense*; Mel-pra, *M. pratense*; Mel-cri, *M. cristatum*; Mel-sp., M. spec.nov., the number 1, 2, 3 behind each species refers to different populations, outgroup: Bar-alp, *Bartsia alpina*; Lat-squ, *Lathraea squamaria*; Rhi-ale, *Rhinanthus alectorolophus*.

M. velebiticum likely due to an hybridization event (Fig. 4).

The increased bootstrap values for single nodes resulted from concatenation of different genes and suggested that two to three pCOS could provide a similar phylogenetic reconstruction as that observed with *ITS* (data not shown).

Discussion

While various LCNG loci have been identified that are highly variable and can be applied to phylogenetic reconstruction of selected plant groups (e.g., NADP-dependent isocitrate dehydrogenase, *idhB*, Weese and Johnson, 2005; *PI*, Bailey and Doyle, 1999; malate



node	MP	NJ	ML	BI
1	100	100	100	100
2	61	95	81	100
3	98	100	93	100
4	100	99	100	100
5	75	98	68	96
6	64	93	77	100
7	78	93	72	100

Fig. 4. Phylogenetic trees of seven *Melampyrum* species, rooted by an outgroup of three species (Genus-Set). For all trees, branches with less than 50% support are drawn as unresolved. Unless otherwise specified, parsimony informative (PI) sites, tree length (TL) and consensus fork index (CFI) below refer to *Melampyrum* species excluding the outgroup. Numbers in brackets refer to the value corresponding to only one population per *Melampyrum* species. (a) Cladograms (MP) of 50% majority-rule consensus trees obtained by 1000 bootstrap replicates; inference of the *matK* tree, obtained with one population/species, is based on a total of 59 PI sites. For *Melampyrum*: PI sites = 5, TL = 36, CFI = 0.67. The *ITS* tree is based on a total of 124 PI sites. For *Melampyrum*: PI sites = 87 (35), TL = 147 (123), CFI = 0.25. The pCOS tree is based on 295 PI sites obtained by concatenating the pCOS genes *Eif3E*, *Agt1*, *AroB* and *At103*. For *Melampyrum*: PI sites = 208 (59),TL = 344 (260), CFI = 1.0. (b) Phylogram (MP) of a 50% majority-rule consensus tree obtained by 1000 bootstrap replicates; the scale at the bottom represents 100 substitutions. Inference of the tree is based on 419 PI sites. For *Melampyrum*: PI sites = 295 (94),TL = 496 (384), CFI = 0,8. The support corresponding to bootstrap replicates (MP, NJ, ML) and the posterior probabilities (BI) is shown in the table on the right (NJ, 1000 replicates; ML, 100 replicates, BI, 100 000 generations). Mel-ita, *Melampyrum italicum*; Mel-vel, *M. velebiticum*; Mel-syl, *M. sylvaticum*; Mel-arv, *M. arvense*; Mel-pra, *M. pratense*; Mel-cri, *M. cristatum*; Mel-sp. *M. spec.nov.* (description in preparation), the number 1, 2, 3 behind each species abbreviation refers to different populations, outgroup: Bar-alp, *Bartsia alpina*; Lat-squ, *Lathraea squamaria*; Rhi-ale, *Rhinanthus alectorolophus*; EMPT, equally most parsimonious trees; MP, maximum parsimony; NJ, neighbor joining; ML, maximum likelihood; BI, Bayesian inference.

synthase, *Ms*, and phosphoribulokinase, *Pry*, Lewis and Doyle, 2002; chalcone synthase, *Chs*, Lihova et al., 2006), less frequently LCNG genes have been investigated for their phylogenetic utility on wide taxonomic ranges. Taking the number of different taxa (families and above) as a criterium, to date the closest approximations to universally amplifiable LCNG are provided by *Rpb2* (Oxelman et al., 2004), *GbssI* (Mason-Gamer et al., 1998), and phytochrome genes (e.g., Mathews and Donoghue, 2000). The existing universally amplifiable markers are generally first developed on a limited number of species, followed by the generalization to more families and orders. *GbssI*, for instance, was first developed for Poaceae phylogeny and then extended by different authors to representatives of Convolvulaceae, Solanaceae, Rosaceae, Araliaceae, Malvaceae, Adoxaceae and Proteaceae (e.g., Mason-Gamer et al., 1998; Walsh and Hoot, 2001; Winkworth and Donoghue, 2004; Mast et al., 2005). This, on the one hand, slows down their development and, on the other hand, can hinder the extension of a potentially useful, universally amplifiable nuclear marker, at least in a first phase, to a wider number of taxa. In the present study, we tried to systematically address this problem by developing and testing phylogenetic markers based on Conserved Ortholog Set (COS) genes (Fulton et al., 2002). Wu et al. (2006) identified a second data set of COS markers (COSII) and demonstrated their utility for phylogeny and comparative mapping for euasterid I species. Our approach differs from that taken by Wu et al. (2006) for the fact that, being COS genes by definition low copy in different families, we focused on the obtainment of universally amplifiable nuclear markers that are *bona fide* low copy, leaving to the single investigators the task of characterizing the best markers (based on ease of amplification, copy number and polymorphism level) for the taxonomic group of interest and question at hand. As polyploidization and genomic rearrangements such as segmental duplications are widespread in plants (Vision et al., 2000; Cui et al., 2006), in fact, we considered the ability to quickly obtain sequencing data from virtually any species of interest more relevant than ascertaining *a priori* the uniqueness of the loci being amplified, provided that their copy number should be reasonably low. We instead focused on the characterization of the variability of the different genes at various taxonomic levels and, whenever no obvious paralogy problems were apparent, on their information content for phylogenetic reconstruction of congeneric species. The choice of the genes used for this study was done before the publication of the COSII data set. We could therefore not take advantage of the improvements with respect to the work of Fulton et al. (2002) and Kozik and Michelmore (in particular of the reciprocal blast match triangulation) used to obtain the COSII. However, we did a less biased choice of the candidate COS, that were obtained from the comparison of both monocots and dicots instead of euasterid I species only (Kozik and Michelmore; http://cgpdb.ucdavis.edu/COS_Arabidopsis). Of the 15 genes considered in our analyses, eight are COSII. Only two of them and three of the COS as determined by Kozik and Michelmore resulted suitable for phylogenetic analysis in the group of species used by us as test case. This confirms that the validation of the markers to use must be carried out on a case-by-case basis, as the complex patterns of gene duplication and loss in plants render virtually impossible the identification of universal LCNG markers.

Because pCOS were developed as a complement to traditional markers at low taxonomic levels, we characterized them with respect to: (1) ease of amplification; (2) information content (level of polymorphism); and (3) information quality (with respect to paralogy). This three aspects are going to be discussed separately.

Ease of amplification

Despite being of paramount importance, only a few papers have been published specifically addressing the design of primers for universally amplifiable LCNG markers from angiosperms (reviewed in Schlueter et al., 2005). The different choices in gene selection, taxon sampling and product validation hinder the possibility to compare the results obtained on COS markers with most of the publications based on broad taxonomic sampling. For instance, the eight primer sets developed by Strand et al. (1997) were tested on seven different species encompassing each dicot and one monocot subclasses. Amplification efficiency higher than 85% was reported for four of the markers developed, but most of the primer combinations produced multiple amplification bands and only a total of five bands from two species were confirmed by sequencing. In another study Xu et al. (2004) used a whole-genome Arabidopsis-rice comparison with 13 418 putative primer pairs among which 15 were selected for experimental characterization in one monocot and one dicot genera (six species per genus). Two of the 15 primer pairs tested (13%) resulted in successful amplification from each of the 12 species tested. However, no sequence validation of the amplification products and phylogenetic reconstructions were carried out.

More detailed examples of primer design for specific groups of species are better suited for comparison with the results of our study. One recent example in this regard was provided by Wu et al. (2006), who reported an efficiency of 40% for transfer of COSII markers from the euasterid species used for primer design to *Physalis*, another euasterid I species. Choi et al. (2006), with the aim of developing Fabaceae-specific markers, reported an 8.8% efficiency of amplification from each of 15 genotypes from an initial set of six species (Choi et al., 2006). Similar results can be inferred in the case of the genera *Pinus* (Syring et al., 2005) and *Astragalus* (Scherson et al., 2005) and of Cupressaceae (Kusumi et al., 2002).

The results obtained for the pCOS markers developed, despite being based on amplifications from phylogenetically distant taxa spanning different orders and subclasses of angiosperms, are in line with those of the intrafamilial studies mentioned above. The amplification of *At103*, *Eif3E* and *Agt1* pCOS markers was confirmed for, respectively, 25, 23 and 22 species of the Class-Set, ranging from monocots to dicots (i.e., 20% of the markers developed amplified with a success rate of ≥ 88%). By comparing these results with the amplification tests carried out on the whole set composed by 67 different families one can expect to attain high amplification rates even from markers that amplified from less than 50% of the families (e.g., *Eif3E*; Table 2). The pCOS therefore are a good starting point to rapidly obtain family- or genus-specific sequences.

Even primers performing well on a limited set of species can encounter a consistent drop in amplification efficiency when tested on a wider array of species, a fact that can substantially decrease the chance to obtain a full data set for more than one marker (about 50% for a set of eight primers on 95 species; Choi et al., 2006). The pCOS developed do not seem to be strongly affected by this problem, but more extensive intrafamilial sampling will be needed to confirm this observation.

Information content

As already observed in other studies involving LCNGs (e.g., Kusumi et al., 2002; Senchina et al., 2003; Scherson et al., 2005; Choi et al., 2006) different nuclear markers evolve at different rates. Also the pCOS markers present various degrees of polymorphism, with about half of them slowly evolving, while the others evolve at a rate similar to that of *ITS*.

The heterogeneity in the sampling of taxa and in the metrics used to characterize the different markers hinders an accurate comparison. Despite these difficulties, the pCOS show levels of sequence divergence higher than most of the previously characterized LCNG markers at intraspecific and intrageneric level. The percentage of parsimony informative sites of pCOS developed in this study ranged between 8.1% and 16.7% for the *Melampyrum* ingroup, about one order of magnitude higher than that for the 16 LCNG characterized in diploid *Gossypium* species, characterized by a rapid diversification (Cronn et al., 2002b). On a different set of New World diploid cottons, three of these markers (*A1341*, *Ces*A1b and *AdhC*) were reported to have a percentage of parsimony informative sites of 1.1%, 2.7% and 9.1% (Alvarez et al., 2005), thus indicating a significant variation from set to set of species. With the exception of *AdhC*, these values are similar to those observed in *Gaertnera*, another genus characterized by a rapid radiation (Malcomber, 2002). In analogous works, the divergence rates of different LCNG was determined for *Pinus* (Syring et al., 2005), *Saltugilia* (Weese and Johnson, 2005) and *Sphaerocardamum*/Brassicaceae (Bailey and Doyle, 1999). In all these cases, the evolutionary rates of pCOS at the genus and family level were either higher or equal to those of the LCNG mentioned above. The comparison of pCOS variability with that of other LCNG at higher phylogenetic distances is difficult because of the very limited number of studies addressing this topic above the family level. Possibly the most complete study to this regard is that of Oxelman et al. (2004), where the *Rpb2* gene has been characterized in a very broad set of taxa ranging from bryophytes to angiosperms. The value of 45% of PI sites for the alignment obtained likely represents the upper limit to the 27.9–34.1% range obtained for pCOS due to the difference in taxa sampling (data not shown).

Currently one of the biggest limitations of pCOS is their length. This in turn limits the absolute amount of polymorphic sites that can contribute to phylogenetic reconstruction. As in the case of other markers such as *Rpb2*, the isolation of cDNA sequences from phylogenetic distant taxa was used to increase the sampling breadth (Denton et al., 1998), RACE or comparable methods could be used to extend the length of pCOS. In light of the continuous increase in the amount of sequence information in public databases, in fact, it is possible that the design of universal primers for the development of additional pCOS will become more straightforward by using existing tools (see, e.g., Rose et al., 1998; Gadberry et al., 2005; Fredslund et al., 2006) or through the development of new software.

Phylogenetic information quality

The ease of PCR amplification and an adequate amount of polymorphisms alone do not, of course, guarantee that a genomic locus is useful for phylogenetic reconstruction. The genomic copy number and the specificity in amplification of orthologous sequences are two equally relevant features.

The use of COS markers as a starting point for marker development was motivated by their expected low copy number in the genomes of various taxa. The *in silico* analysis of two fully sequenced genomes (rice and poplar) in addition to Arabidopsis confirms that the representative COS loci used in this study are indeed low copy in these distantly related plant taxa as well (Table 4). This confirms that pCOS are a convenient starting point to develop broadly applicable LCNG markers, provided that their copy number in the species of interest will be case by case experimentally

Table 4
Number and location in the fully sequenced *Oryza sativa* (rice) and *Populus trichocarpa* (poplar) genomes of homologs of COS genes

Marker	No. of copies in rice	Rice chromosome no.	No. of copies in poplar	Poplar location
Agt1	1	8	2	1, 9
Apg1	2	7, 12	3	2, 5, 8
AroB	1	9	2	scaffold_57, scaffold_4978
At103	1	1	2	6, 16
Bio2	1	8	3	7, scaffold_14504, scaffold_64
ChlP	2	1, 2	5	9, 12, scaffold_66, scaffold_129, scaffold_4339
Det3	1	5	2	1, scaffold_88
Eif3E	2	7	2	6, 16
Gi	1	1	2	2, 5
Hcf136	1	6	2	5, 7
Hmgs	3	3, 8, 9	2	1, 3
Psy	3	6, 9, 12	4	1, 2, 3, 5
Rml1	2	5, 7	2	1, 3
Sqd1	1	5	1	7
Sqd2	3	1, 3, 7	2	6, 16
Average	1.67		2.40	
SD	0.82		0.99	

ascertained according to well established methods (reviewed in Small et al., 2004).

In the present study orthology assessment was carried out through consistency of phylogenetic reconstruction in comparison with traditional markers such as *ITS*, being a taxon-specific characterization of single pCOS out of the scope of this work. Particularly for different orders and subclasses, the low support of the phylogenetic reconstruction due to excessive sequence divergence does not allow to exclude that in some cases paralogous sequences were sampled. At the genus and family level, however, the overall agreement or disagreement between pCOS and traditional markers clearly distinguished cases of paralogy (e.g., *ChlP*) from orthology (e.g., *Eif3E*).

Among the markers that passed this first selection, the more polymorphic pCOS and *ITS* displayed a higher consistency in the grouping of conspecific individuals from different populations. The lower intraspecific resolution observed for *Agt1* and *At103* could indicate paralogous sequence sampling. However, the analysis of the level of polymorphism for the single genes indicated that the most polymorphic genes were those more congruent to *ITS* (Fig. 3). The incongruence among topologies produced by different pCOS was therefore likely not due to paralogy, but to insufficient level of PI characters for single loci. On this basis, we decided to take a total evidence approach to data analysis. The addition of a progressively higher number of phylogenetically informative sites, indeed, provided congruent and supported phylogenies. The whole concatenated data set achieved a level of support better than any other combination of markers independently on the method used for phylogenetic reconstruction. However, in the example shown, the two pCOS loci *AroB* and *Eif3E* provided the majority of the phylogenetic information in the combined data set. A broader choice of universally amplifiable markers with equally high polymorphism level could, on the one hand, contribute to avoid the overweighing of topologies provided by more informative markers, at the risk of positively misleading the phylogeny (Bull et al., 1993; Kubatko and Degnan, 2007). On the other hand, the development of new, longer pCOS would improve the sequencing efficiency. We conclude that, on the average, two to three pCOS of the length and level of polymorphism similar to those developed in this study could provide a phylogenetic reconstruction with a resolution higher than that provided by *ITS*. This estimation, possibly conservative as both *AroB* and *Eif3E* alone provided already a more resolved phylogeny than *ITS* (Figs 3 and 4), should be robust to genus to genus variations of species divergence wider than those observed in the case study presented.

We further noticed a direct correlation between amplification efficiency and presence of paralogs in the species from which sequences were obtained. This trend is consistent with the observation that *ChlP* underwent repeated duplication events to attain the copy number of 5 as in poplar (Table 4). On the other hand, primers with lower amplification efficiencies provided in general more robust markers for phylogenetic reconstruction.

When to use pCOS markers?

The use of LCNG has some drawbacks as compared with cpDNA and nrDNA. Lower amplification efficiency, mixtures of paralogous sequences, possible chimerisms of amplicons are all possible concerns in PCR amplifications of LCNG (Emshwiller and Doyle, 1999; Cronn et al., 2002a). The orthology of the amplified products has to be tested through sometimes lengthy and labor-intensive procedures (see, e.g., Doyle and Doyle, 1999; Small and Wendel, 2000). Moreover, the use of several markers makes it more difficult to obtain complete data sets in cases of high numbers of taxa (e.g., Choi et al., 2006). In light of these limitations, it is probably still convenient to rely on cpDNA or nrDNA for standard phylogenetic issues, whenever possible.

In a growing number of cases, however, LCNG loci may provide the best or the only alternative to traditional molecular markers (see, e.g., Marhold and Lihova, 2006) or in cases where morphological characters are insufficient or homoplasious (e.g., Bailey et al., 2006). In such cases, even though detailed protocols helping in the establishment of group-specific LCNG markers have been recently published (Small et al., 2004; Hughes et al., 2006) and large collections of high quality LCNG sets have been developed for selected taxa (e.g., Choi et al., 2006; Wu et al., 2006), the development of highly informative LCNG phylogenetic markers may still be an extremely time-consuming task, sometimes fraught with low success rate (Small et al., 2004). From the cases reported in the literature and from the present study, in fact, one can expect to attain the development of a LCNG marker that is informative for the taxonomic question at hand from roughly 5 to 20% of the starting candidates. The development of family and, even more, genus-specific markers therefore is very uneconomical as one can expect that 80–95% of the effort will be wasted to start the quest for the "right" marker(s) all over. Markers developed from genomic or cDNA-based multilocus data (e.g., Bailey et al., 2004; Whittall et al., 2006) are readily obtained, but the transferrability of such markers to other taxa is expected to be low. On the other hand, anchor-based methods such as COS and similar approaches (e.g., Lyons et al., 1997; Syring et al., 2005) are better suited to generate widely amplifiable, prescreened LCNG markers. By using COS genes as a starting point we demonstrated that it is possible to develop informative LCNG markers having the potential to find application, after a case-by-case evaluation, in a broad range of taxa. This could be particularly useful in light of the plethora of species still awaiting precise classification in the phylogenetic twilight zone between too closely and too distantly related. The possibility of developing a higher number of improved universally amplifiable pCOS markers as demonstrated in this work holds the promise to substantially increase the steadily growing but still limited toolbox available to plant researchers having an interest in phylogenetic reconstructions of closely related species.

Acknowledgments

The authors thank F. Prosser and M. Stech for helpful discussion about morphological traits of the *Melampyrum aggregate* and editors and anonymous referees for significantly contributing to improve this paper. This work was financially supported by the "CSBT" project (GP664-210303; GP3299-301204) funded by the Autonomous Province of Trento.

References

Alvarez, I., Cronn, R., Wendel, J.F., 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.

Angiosperm Phylogeny Group, II, 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141, 399–436.

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

Bailey, C.D., Hughes, C.E., Harris, S.A., 2004. Using RAPDs to identify DNA sequence loci for species level phylogeny reconstruction: an example from *Leucaena* (Fabaceae). Syst. Bot. 29, 4–14.

Bailey, C.D., Koch, M.A., Mayer, M., Mummenhoff, K., O'Kane, S.L., Warwick, S.I., Windham, M.D., Al-Shehbaz, I.A., 2006. Toward a global phylogeny of the Brassicaceae. Mol. Biol. Evol. 23, 2142–2160.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S., Donoghue, M.J., 1995. The *ITS* region of nuclear ribosomal DNA—a valuable source of evidence on angiosperm phylogeny. Ann. Mo. Bot. Gard. 82, 247–277.

Barten, R., Meyer, T.F., 1998. Cloning and characterisation of the *Neisseria gonorrhoeae* aroB gene. Mol. Gen. Genet, 258, 34–44.

Bennett, J.R., Mathews, S., 2006. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. Am. J. Bot. 93, 1039–1051.

Bull., J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384–397.

Cheng, Z., Sattler, S., Maeda, H., Sakuragi, Y., Bryant, D.A., DellaPenna, D., 2003. Highly divergent methyltransferases catalyze a conserved reaction in tocopherol and plastoquinone synthesis in cyanobacteria and photosynthetic eukaryotes. Plant Cell 15, 2343–2356.

Choi, H.K., Luckow, M.A., Doyle, J., Cook, D.R., 2006. Development of nuclear gene-derived molecular markers linked to legume genetic maps. Mol. Genet. Genom. 276, 56–70.

Cronn, R., Cedroni, M., Haselkorn, T., Grover, C., Wendel, J.F., 2002a. PCR-mediated recombination in amplification products derived from polyploid cotton. Theor. Appl. Genet. 104, 482–489.

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

Cui, L.Y., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., Depamphilis, C.W., 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16, 738–749.

Denton, A.L., Mcconaughy, B.L., Hall, B.D., 1998. Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. Mol. Biol. Evol. 15, 1082–1085.

Despres, L., Gielly, L., Redoutet, W., Taberlet, P., 2003. Using AFLP to resolve phylogenetic relationships in a morphologically diversified plant species complex when nuclear and chloroplast sequences fail to reveal variability. Mol. Phylogenet. Evol. 27, 185–196.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of leaf tissue. Phytochem. Bull. 19, 11–15.

Doyle, J.J., Doyle, J.L., 1999. Nuclear protein-coding genes in phylogeny reconstruction and homology assessment: some examples from Leguminosae. In: Hollingsworth, P.M., Bateman, R.M., Gornall, R.J. (Eds.), Molecular Systematics and Plant Evolution, pp. 229–254. Taylor & Francis, London.

Doyle, J.J., Doyle, J.L., Harbison, C., 2003. Chloroplast-expressed glutamine synthetase in *Glycine* and related Leguminosae: phylogeny, gene duplication, and ancient polyploidy. Syst. Bot. 28, 567–577.

Doyle, J.J., Doyle, J.L., Rauscher, J.T., Brown, A.H.D., 2004. Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): a study of contrasts. Biol. J. Linn. Soc. 82, 583–597.

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

Essigmann, B., Hespenheide, B.M., Kuhn, L.A., Benning, C., 1999. Prediction of the active-site structure and NAD (+) binding in SQD1, a protein essential for sulfolipid biosynthesis in Arabidopsis. Arch. Biochem. Biophys. 369, 30–41.

Ford, V.S., Lee, J., Baldwin, B.G., Gottlieb, L.D., 2006. Species divergence and relationships in *Stephanomeria* (Compositae): *PGIC* phylogeny compared to prior biosystematic studies. Am. J. Bot. 93, 480–490.

Fredslund, J., Madsen, L.H., Hougaard, B.K., Sandal, N., Stougaard, J., Bertioli, D., Schauser, L., 2006. GEMprospector—online design of cross-species genetic marker candidates in legumes and grasses. Nucleic Acids Res. 34, W670–W675.

Fulton, T.M., Van der Hoeven, R., Eannetta, N.T., Tanksley, S.D., 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Pl. Cell 14, 1457–1467.

Gadberry, M.D., Malcomber, S.T., Doust, A.N., Kellogg, E.A., 2005. Primaclade—a flexible tool to find conserved PCR primers across multiple species. Bioinformatics 21, 1263–1264.

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

Hilu, K.W., Borsch, T., Muller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Rohwer, J.G., Campbell, C.S., Chatrou, L.W., 2003. Angiosperm phylogeny based on *matK* sequence information. Am. J. Bot. 90, 1758–1776.

Hughes, C.E., Eastwood, R.J., Bailey, C.D., 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philos. Trans. R. Soc. B Biol. Sci. 361, 211–225.

Keller, Y., Bouvier, F., d'Harlingue, A., Camara, B., 1998. Metabolic compartmentation of plastid prenyllipid biosynthesis—evidence for the involvement of a multifunctional geranylgeranyl reductase. Eur J. Biochem. 251, 413–417.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J. Mol. Evol. 16, 111–120.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.

Kusumi, J., Tsumura, Y., Yoshimaru, H., Tachida, H., 2002. Molecular evolution of nuclear genes in Cupressaceae, a group of conifer trees. Mol. Biol. Evol. 19, 736–747.

Lewis, C.E., Doyle, J.J., 2002. A phylogenetic analysis of tribe Areceae (Arecaceae) using two low-copy nuclear genes. Plant Syst. Evol. 236, 1–17.

Liepman, A.H., Olsen, L.J., 2003. Alanine aminotransferase homologs catalyze the glutamate: glyoxylate aminotransferase reaction in peroxisomes of Arabidopsis. Plant Physiol. 131, 215–227.

Lihova, J., Shimizu, K.K., Marhold, K., 2006. Allopolyploid origin of *Cardamine asarifolia* (Brassicaceae): Incongruence between plastid and nuclear ribosomal DNA sequences solved by a single-copy nuclear gene. Mol. Phylogenet. Evol. 39, 759–786.

Linder, C.R., Rieseberg, L.H., 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91, 1700–1708.

Londo, J.P., Chiang, Y.C., Hung, K.H., Chiang, T.Y., Schaal, B.A., 2006. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. Proc. Natl Acad. Sci. USA 103, 9578–9583.

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., Obrien, S.J., 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nature Genet. 15, 47–56.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Malcomber, S.T., 2002. Phylogeny of *Gaertnera* Lam. (Rubiaceae) based on multiple DNA markers: evidence of a rapid radiation in a widespread, morphologically diverse genus. Evol. Int. J. Org. Evol, 2002, 42–57.

Marhold, K., Lihova, J., 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst. Evol. 259, 143–174.

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

Mast, A.R., Jones, E.H., Havery, S.P., 2005. An assessment of old and new DNA sequence evidence for the paraphyly of *Banksia* with respect to *Dryandra* (Proteaceae). Aust. Syst. Bot. 18, 75–88.

Mathews, S., Donoghue, M.J., 2000. Basal angiosperm phylogeny inferred from duplicate phytochromes *a* and *c*. Int. J. Plant Sci. 161, S41–S55.

Montamat, F., Guilloton, M., Karst, F., Delrot, S., 1995. Isolation and characterization of a cDNA encoding *Arabidopsis thaliana* 3-hydroxy-3-methylglutaryl-coenzyme A synthase. Gene 167, 197–201.

Oliverio, K.A., Crepy, M., Martin-Tryon, E.L., Milich, R., Harmer, S.L., Putterill, J., Yanovsky, M.J., Casal, J.J., 2007. GIGANTEA regulates phytochrome A-mediated photomorphogenesis independently of its role in the circadian clock. Plant Physiol. 144, 495–502.

Olmstead, R.G., dePamphilis, C.W., Wolfe, A.D., Young, N.D., Elisons, W.J., Reeves, P.A., 2001. Disintegration of the Scrophulariaceae. Am. J. Bot. 88, 348–361.

Oxelman, B., Yoshikawa, N., Mcconaughy, B.L., Luo, J., Denton, A.L., Hall, B.D., 2004. *Rpb2* gene phylogeny in flowering plants, with particular emphasis on Asterids. Mol. Phylogenet. Evol. 32, 462–479.

Pelser, P.B., Gravendeel, B., Van Der Meijden, R., 2003. Phylogeny reconstruction in the gap between too little and too much divergence: the closest relatives of *Senecio jacobaea* (Asteraceae) according to DNA sequences and AFLPs. Mol. Phylogenet. Evol. 29, 613–628.

Poke, F.S., Martin, D.P., Steane, D.A., Vaillancourt, R.E., Reid, J.B., 2006. The impact of intragenic recombination on phylogenetic reconstruction at the sectional level in Eucalyptus when using a single copy nuclear gene (cinnamoyl CoA reductase). Mol. Phylogenet. Evol. 39, 160–170.

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

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

Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., Mccallum, C.M., Henikoff, S., 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res. 26, 1628–1635.

Rozas, J., Sanchez-Delbarrio, J.C., Messeguer, X., Rozas, R., 2003. Dnasp, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496–2497.

Rudd, S., Schoof, H., Mayer, K., 2005. Plantmarkers—a database of predicted molecular markers from plants. Nucleic Acids Res. 33, D628–D632.

Rzeznicka, K., Walker, C.J., Westergren, T., Kannangara, C.G., von Wettstein, D., Merchant, S., Gough, S.P., Hansson, M., 2005. Xantha-1 encodes a membrane subunit of the aerobic Mg-protoporphyrin IX monomethyl ester cyclase involved in chlorophyll biosynthesis. Proc. Natl Acad. Sci. USA 102, 5886–5891.

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

Scherson, R.A., Choi, H.K., Cook, D.R., Sanderson, M.J., 2005. Phylogenetics of New World *Astragalus*: screening of novel nuclear loci for the reconstruction of phylogenies at low taxonomic levels. Brittonia 57, 354–366.

Schlueter, P.M., Stuessy, T.F., Paulus, H.F., 2005. Making the first step: practical considerations for the isolation of low-copy nuclear sequence markers. Taxon 54, 766–770.

Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J.K., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A., Wendel, J.F., 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol. Biol. Evol. 20, 633–643.

Small, R.L., Wendel, J.F., 2000. Copy number lability and evolutionary dynamics of the *adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155, 1913–1926.

Small, R.L., Cronn, R.C., Wendel, J.F., 2004. Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot. 17, 145–170.

Soltis, D.E., Soltis, P.S., Nickrent, D.L., Johnson, L.A., Hahn, W.J., Hoot, S.B., Sweere, J.A., Kuzoff, R.K., Kron, K.A., Chase, M.W., Swensen, S.M., Zimmer, E.A., Chaw, S.M., Gillespie, L.J., Kress, W.J., Sytsma, K.J., 1997. Angiosperm phylogeny inferred from 18s ribosomal DNA sequences. Ann. Mo. Bot. Gard. 84, 1–49.

Soltis, D.E., Soltis, P.S., Chase, M.W., Mort, M.E., Albach, D.C., Zanis, M., Savolainen, V., Hahn, W.H., Hoot, S.B., Fay, M.F., Axtell, M., Swensen, S.M., Prince, L.M., Kress, W.J., Nixon, K.C., Farris, J.S., 2000. Angiosperm phylogeny inferred from 18s rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linn. Soc. 133, 381–461.

Strand, A.E., Leebensmack, J., Milligan, B.G., 1997. Nuclear DNA-based markers for plant evolutionary biology. Mol. Ecol. 6, 113–118.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.

Syring, J., Willyard, A., Cronn, R., Liston, A., 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot. 92, 2086–2100.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Van de Peer, Y., De Wachter, R., 1997. Construction of evolutionary distance trees with Treecon for Windows: accounting for variation in nucleotide substitution rate among sites. Comput. Appl. Biosci. 13, 227–230.

Vision, T.J., Brown, D.G., Tanksley, S.D., 2000. The origins of genomic duplications in Arabidopsis. Science 290, 2114–2117.

Walsh, B.M., Hoot, S.B., 2001. Phylogenetic relationships of *Capsicum* (Solanaceae) using DNA sequences from two noncoding regions: the chloroplast *atpB-rbcL* spacer region and nuclear *waxy* introns. Int. J. Plant Sci. 162, 1409–1418.

Weese, T.L., Johnson, L.A., 2005. Utility of NADP-dependent isocitrate dehydrogenase for species-level evolutionary inference in angiosperm phylogeny: a case study in *Saltugilia*. Mol. Phylogenet. Evol. 36, 24–41.

Wendel, J.F., Schnabel, A., Seelanan, T., 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl Acad. Sci. USA 92, 280–284.

Whittall, J.B., Medina-Marino, A., Zimmer, E.A., Hodges, S.A., 2006. Generating single-copy nuclear gene data for a recent adaptive radiation. Mol. Phylogenet. Evol. 39, 124–134.

Winkworth, R.C., Donoghue, M.J., 2004. *Viburnum* phylogeny: evidence from the duplicated nuclear gene *GbssI*. Mol. Phylogenet. Evol. 33, 109–126.

Wu, F., Mueller, L.A., Crouzillat, D., Petiard, V., Tanksley, S.D., 2006. Combining bioinformatics and phylogenetics to identify large sets of single copy, orthologous genes (COS II) for comparative, evolutionary and systematics studies: a test case in the euasterid plant clade. Genetics. 174, 1407–1420.

Xu, W., Briggs, W.J., Padolina, J., Timme, R.E., Liu, W., Linder, C.R., Miranker, D.P., 2004. Using MoBIoS' scalable genome join to find conserved primer pair candidates between two genomes. Bioinformatics 20, i355–362.

Yahalom, A., Kim, T.H., Winter, E., Karniol, B., von Arnim, A.G., Chamovitz, D.A., 2001. Arabidopsis eIF3e (INT-6) associates with both eIF3c and the COP9 signalosome subunit CSN7. J. Biol. Chem. 276, 334–340.

Young, N.D., Healy, J., 2003. Gapcoder automates the use of indel characters in phylogenetic analysis. Bmc Bioinformatics 4, 6.

Appendix 1

Species used in this study. The first three columns indicate for which of the test sets described in the paper each species was used. T1 (Test-Set-1): 17 species, two families (Plantaginaceae, Orobanchaceae); T2 (Test-Set-2): 87 species, 67 families (angiosperms, one fern and one moss); C (Class-Set): 25 species, 19 families (angiosperms). The numbers in square brackets are GenBank accession numbers for sequences obtained from GenBank. Species names and authors were obtained from the Provisional Global Plant Checklist (GPC) managed by IOPI (International Organization for Plant Information) (accessed 20 December 2006) or, if not listed there, from The International Plant Names Index (2004) at http://www.ipni.org. Family classification is according to the Angiosperm Phylogeny Group (APG, 2003).

Set		Genus	Species	Author	Family	Comments
T2		*Acer*	*pseudoplatanus*	L.	Aceraceae	
T2	C	*Aesculus*	*hippocastanum*	L.	Hippocastanaceae	for *matK* [AY968630], for ITS *A. wangii* Hu. [AF406968], for *rbcL A. pavia* L. [U39277] from GenBank were used
T2		*Ailanthus*	*altissima*	(Mill.) Swingle	Simaroubaceae	
T2		*Alcea*	*rosea*	L.	Malvaceae	
T2		*Allium*	*senescens* ssp. *montanum*	(Fr.) Holub	Alliaceae	

Appendix 1

(Continued)

Set			Genus	Species	Author	Family	Comments
			Antirrhinum	*barrelieri*	Boreau	Plantaginaceae	
			Antirrhinum	*meonanthum*	Lange	Plantaginaceae	
			Antirrhinum	*molle*	L.	Plantaginaceae	
			Antirrhinum	*nuttallianus*	(Benth. ex A. DC.) D.A. Sutton	Plantaginaceae	
			Antirrhinum	*siculum*	Miller	Plantaginaceae	
	T2	C	*Arabidopsis*	*thaliana*	(L.) Heynh.	Brassicaceae	
	T2		*Aristolochia*	*clematitis*	L.	Aristolochiaceae	
	T2		*Artemisia*	*vulgaris*	L.	Asteraceae	
	T2		*Asplenium*	*trichomanes*	L.	Aspleniaceae*	
T1		C	*Bartsia*	*alpina*	L.	Orobanchaceae	for *ndhF* [AF123678], for *rbcL* [AF190903] from GenBank were used
	T2		*Berberis*	*vulgaris*	L.	Berberidaceae	
	T2		*Bergenia*	*crassifolia*	(L.) Fritsch	Saxifragaceae	
	T2	C	*Betula*	*pendula*	Roth	Betulaceae	for *rbcL B. nigra* L. [L12634] from GenBank was used
	T2		*Bryonia*	*dioica*	Jacq.	Cucurbitaceae	
	T2		*Buxus*	*sempervirens*	L.	Buxaceae	
	T2		*Campanula*	*trachelium*	L.	Campanulaceae	
	T2		*Capsella*	*bursa-pastoris*	(L.) Medical.	Brassicaceae	
	T2		*Cardamine*	*hirsuta*	L.	Brassicaceae	
	T2		*Carex*	*umbrosa*	Host	Cyperaceae	
	T2		*Cedrus*	*deodara*	(D. Don) G. Don fil.	Pinaceae	
	T2		*Celtis*	*australis*	L.	Ulmaceae	
	T2		*Centranthus*	*ruber*	(L.) DC.	Valerianaceae	
T1			*Chaenorhinum*	*minus*	(L.) Lange	Plantaginaceae	
	T2		*Chelidonium*	*majus*	L.	Papaveraceae	
	T2		*Convallaria*	*majalis*	L.	Convallariaceae	
	T2		*Cornus*	*mas*	L.	Cornaceae	
	T2		*Corydalis*	*solida*	(L.) Clairv.	Fumaraceae	
	T2	C	*Crocus*	*vernus*	(L.) Hill	Iridaceae	for *matK C. pulchellus* Herbert [AJ579941], for *rbcL C. pulchellus* Herbert [AJ309668] from GenBank were used
			Cuscuta	*spec.*		Solanaceae	
	T2		*Cyclamen*	*purpurascens*	Mill.	Primulaceae	
T1			*Cymbalaria*	*muralis*	P. Gaertn., B. Mey. & Scherb.	Plantaginaceae	
	T2	C	*Dactylis*	*glomerata*	L.	Poaceae	for *rbcL* [AY395535], for *matK Brachypodium sylvaticum* (Huds.)P.Beauv. [AF1644007] from GenBank were used
	T2		*Daphne*	*mezereum*	L.	Thymelaceae	
T1			*Digitalis*	*lutea*	L.	Plantaginaceae	
	T2		*Erica*	*carnea*	L.	Ericaceae	
	T2		*Erodium*	*cicutarium*	(L.) L'Hér.	Geraniaceae	
	T2		*Euonymus*	*europaeus*	L.	Celastraceae	
	T2		*Euphorbia*	*cyparissias*	L.	Euphorbiaceae	
	T2		*Euphorbia*	*helioscopia*	L.	Euphorbiaceae	
	T2	C	*Fallopia*	*dumetorum*	(L.) Holub	Polygonaceae	for *rbcL F. sachalinense* (F. Schmidt) Ronse Decr. [AF297125] from GenBank was used
	T2		*Ficus*	*carica*	L.	Moraceae	
	T2		*Forsythia*	x *intermedia*	Zabel	Oleaceae	
	T2		*Fraxinus*	*ornus*	L.	Oleaceae	
	T2		*Gagea*	*villosa*	(M. Bieb.) Sweet	Liliaceae	
	T2	C	*Galium*	*mollugo*	L.	Rubiaceae	for *rbcL* [AY395538] from GenBank was used
	T2		*Geranium*	*rotundifolium*	L.	Geraniaceae	
	T2		*Globularia*	*cordifolia*	L.	Plantaginaceae	
	T2	C	*Hedera*	*helix*	L.	Araliaceae	for *rbcL* [L01924] was taken from GenBank
	T2		*Helianthemum*	*nummularium* s.l.	(L.) Mill.	Cistaceae	
	T2		*Humulus*	*lupulus*	L.	Cannabaceae	
	T2		*Iris*	*germanica*	L.	Iridaceae	
	T2		*Jasminum*	*nudiflorum*	Lindl.	Oleaceae	
	T2		*Juglans*	*regia*	L.	Juglandaceae	
	T2		*Koelreuteria*	*paniculata*	Laxm.	Sapindaceae	

Appendix 1

(Continued)

Set			Genus	Species	Author	Family	Comments
	T2	C	*Lamium*	*purpureum*	L.	Lamiaceae	for *rbcL* [Z37403] from GenBank was used
T1		C	*Lathraea*	*squamaria*	L.	Orobanchaceae	for *rbcL L. clandestina* L. [AF026833] from GenBank was used
	T2		*Lathyrus*	*vernus*	(L.) Bernh.	Fabaceae	
	T2		*Leucojum*	*vernum*	L.	Amaryllidaceae	
T1		C	*Linaria*	*alpina*	(L.) Mill.	Plantaginaceae	for *rbcL Antirrhinum majus* L. [L11688] from GenBank was used
			Linaria	*vulgaris*	Mill.	Plantaginaceae	
	T2		*Listera*	*ovata*	(L.) R. Br.	Orchidaceae	
	T2		*Lithospermum*	*arvense* s.str.	L.	Boraginaceae	
	T2	C	*Lonicera*	*xylosteum*	L.	Caprifoliaceae	for *Agt1* and *Eif3E L. caprifolium* L. was used
	T2		*Magnolia*	spec.		Magnoliaceae	
		C	*Melampyrum*	*arvense*	L.	Orobanchaceae	
T1		C	*Melampyrum*	*cristatum*	L.	Orobanchaceae	
T1		C	*Melampyrum*	*italicum*	Soó	Orobanchaceae	
T1		C	*Melampyrum*	*pratense*	L.	Orobanchaceae	
		C	*Melampyrum*	spec.‡		Orobanchaceae	
		C	*Melampyrum*	*sylvaticum*	L.	Orobanchaceae	
		C	*Melampyrum*	*velebiticum*	Borbás	Orobanchaceae	
	T2		*Mercurialis*	*perennis*	L.	Euphorbiaceae	
T1			*Odontites*	*luteus*	(L.) Clairv.	Orobanchaceae	
T1			*Orobanche*	*gracilis*	Sm.	Orobanchaceae	
	T2		*Ostrya*	*carpinifolia*	Scop.	Corylaceae	
	T2		*Oxalis*	*acetosella*	L.	Oxalidaceae	
	T2		*Paeonia*	*arborea*	Donn.	Paeoniaceae	
	T2		*Parietaria*	*officinalis*	L.	Urticaceae	
	T2		*Paris*	*quadrifolia*	L.	Melanthiaceae	
	T2		*Pastinaca*	*sativa* s.l.	L.	Apiaceae	
T1			*Pedicularis*	*elongata*	Kern.	Orobanchaceae	
T1			*Pedicularis*	*verticillata*	L.	Orobanchaceae	
		C	*Physalis*	*alkekengi*	L.	Solanaceae	for *rbcL* [U08617] from GenBank was used
	T2		*Plantago*	*lanceolata*	L.	Plantaginaceae	
	T2		*Polygala*	*chamaebuxus*	L.	Polygalaceae	
	T2		*Polygonatum*	*multiflorium*	(L.) All.	Ruscaceae	
	T2		*Potentilla*	*tabernaemontani*	Asch.	Rosaceae	
	T2		*Primula*	*vulgaris*	Huds.	Primulaceae	
	T2	C	*Prunus*	*avium*	L.	Rosaceae	for *rbcL P. domestica* ssp. *intermedia* Röder [AF227901] from GenBank was used
	T2		*Pulmonaria*	*officinalis*	L.	Boraginaceae	
	T2	C	*Ranunculus*	*bulbosus*	L.	Ranunculaceae	for *matK* [AY954188], for *rbcL R. acris* L. [AY395557] from GenBank was used
T1		C	*Rhinanthus*	*alectorolophus*	(Scop.) Pollich	Orobanchaceae	for *ndhF Pedicularis foliosa* L. [AF123689], for *rbcL R. minor* L. [AY395558] from GenBank were used
	T2		*Rhytidiadelphus*	*triquetrus*	(Hedw.) Warnst.	Hylocomiaceae†	
	T2		*Ruta*	*graveolens*	L.	Rutaceae	
	T2		*Scabiosa*	spec.		Dipsacaceae	
	T2	C	*Scrophularia*	*canina*	L.	Scrophulariaceae	for *rbcL Scrophularia* spec. L. [L36449] from GenBank was used
	T2		*Sedum*	cf. *acre*	L.	Crassulaceae	
	T2		*Selaginella*	*selaginoides*	(L.) P. Beauv.	Selaginellaceae*	
	T2		*Senecio*	*vulgaris*	L.	Asteraceae	
	T2		*Silene*	*vulgaris*	(Moench) Garcke	Caryophyllaceae	
	T2		*Stellaria*	*media*	(L.) Vill.	Caryophyllaceae	
	T2		*Tamus*	*communis*	L.	Dioscoreaceae	
	T2		*Taraxacum*	*officinale*	Weber	Asteraceae	
	T2		*Tussilago*	*farfara*	L.	Asteraceae	
	T2		*Valerianella*	*locusta*	(L.) Laterr.	Valerianaceae	
T1			*Veronica*	*beccabunga*	L.	Plantaginaceae	
T1			*Veronica*	*urticifolia*	Jacq.	Plantaginaceae	
	T2		*Vinca*	*major*	L.	Apocynaceae	

Appendix 1

(Continued)

Set	Genus	Species	Author	Family	Comments
T2	*Viola*	*arvensis*	Murray	Violaceae	
T2	*Viola*	*suavis*	M.Bieb.	Violaceae	
	Withania	*somnifera*	(L.) Dunal	Solanaceae	

*Pteridophyta.

†Bryophyta.

‡*Melampyrum spec*. refers to a new species so far misclassified as *M. italicum*.

Appendix 2

Complete list of COS markers tested in this study. The Arabidopsis Genomic Initiative Identifier (AGI-ID) for the *A. thaliana* orthologs is provided. Testing level abbreviations: A = primer prescreening (Test Set 1: Orobanchaceae + Plantaginaceae); B = sequencing (Genus-set); C = sequencing (Class-Set). Application range: G = genus; F = family; O = order; ND = not determined; question marks indicate possible application ranges to be tested in other taxonomic groups.

Name	AGI-ID	Function	Forward primer	Reverse primer	Testing level	Application
Det3	AT1G12840	Vacuolar ATP synthase subunit C-related	tgggatgaggcaaagtacccnacnatgtc	gttacagtgtaragrgcataytcatt	C	F/O
Gi	AT1G22770	*Gigantea* protein -related	tgggctacagatgcacttga	cacgcaagaaatgcaratgcatcca	B	G?
ChlP	AT1G74470	Geranylgeranyl reductase	cggcgagtgacsaaratgaagatgat	ggcgagacgtcrtcrccgacrtacat	C	F/O
Agt1	AT2G13360	Alanine-glyoxylate aminotransferase	gatttccghatggatgantgggg	ccaytcctccttctghgtgcagtt	C	G?
Bio2	AT2G43360	Biotin synthase	ggvtgcagygaagaytgttc	gtgcaacaracytccatbcccat	C	F/O
At103	AT3G56940	Mg-protoporphyrin IX monomethyl ester cyclase	cttcaagccmaagttcatcttcta	ttggcaatcattgaggtacatngtmacata	C	S/G/F
Eif3E	AT3G57290	Translation initiation factor 3-related protein	tttgaatgtggcaactaytctrgtgctgc	acctcttcacactcyytcatctt	B	G?
Apg1	AT3G63410	Chloroplast inner envelope membrane protein, putative	ggaccagtcgccncaycagct	aaccactcaatrtactcttc	A	ND
Hmgs	AT4G11820	Hydroxymethylglutaryl-CoA synthase	gctttgttnaaytgtgtbaattgggt	aagtcatagacatgdgmcatrtg	A	ND
Rml1	AT4G23100	Gamma-glutamylcysteine synthetase	cctggtggtcagttygarcttagtgg	ttgtcagtatctgtccdtatrtgg	B	G?
Sqd1	AT4G33030	Sulfite:UDP-glucose sulfotransferase	cttgggacsatgggtgartatgg	ccwacagcagcytgmacacagaacc	A	ND
Sqd2	AT5G01220	UDP-sulfoquinovose:DAG sulfoquinovosyltransferase	gtcccaatagtratgtcdtaycacac	ctttggaataagggtgthgattc	A	ND
Psy	AT5G17230	Geranylgeranyl-diphosphate geranylgeranyl transferase	gagacatgatygaaggratg	atgtcttcatctganagvccsgc	A	ND
Hcf136	AT5G23120	Photosystem II stability/ assembly factor	tcagctgargangaagatttcaa	cagcactgaaagttcchgtrtagta	B	G?
AroB	AT5G66120	3-dehydroquinate synthase, putative	gcattctaccaarcwcartgtgt	gctttgttttcacatgawckcttdatagca	A	ND

Appendix 3

Species sets used for sequencing and phylogenetic reconstruction in this study.

Order	Family	Genus	Species	Set							
Lamiales	Orobanchaceae	*Melampyrum*	*italicum*	A	B1	C	D				
Lamiales	Orobanchaceae	*Melampyrum*	*velebiticum*	A	B2	C	D				
Lamiales	Orobanchaceae	*Melampyrum*	*spec.*	A	B3	C	D				
Lamiales	Orobanchaceae	*Melampyrum*	*sylvaticum*	A	B4	C	D	E1	F1	G1	H

Appendix 3

(Continued)

Order	Family	Genus	Species	Set							
Lamiales	Orobanchaceae	*Melampyrum*	*pratense*	A	B5	C	D				
Lamiales	Orobanchaceae	*Melampyrum*	*arvense*	A	B6	C	D				
Lamiales	Orobanchaceae	*Melampyrum*	*cristatum*	A	B7	C	D				
Lamiales	Orobanchaceae	*Bartsia*	*alpina*		B8	C	D	E2	F2		
Lamiales	Orobanchaceae	*Lathraea*	*squamaria*		B9	C	D	E3	F3	G2	H
Lamiales	Orobanchaceae	*Rhinanthus*	*alectorolophus*		B10	C	D	E4	F4	G3	H
Lamiales	Plantaginaceae	*Linaria*	*alpina*				D	E5	F5	G4	H
Lamiales	Scrophulariaceae	*Scrophularia*	*canina*				D	E6	F6	G5	H
Lamiales	Lamiaceae	*Lamium*	*purpureum*				D	E7	F7		
Solanales	Solanaceae	*Physalis*	*alkekengi*				D	E8	F8	G6	H
Gentianales	Rubiaceae	*Galium*	*mollugo*				D		F9		
Dipsacales	Caprifoliaceae	*Lonicera*	*xylosteum*				D	E9	F10	G7	H
Apiales	Araliaceae	*Hedera*	*helix*				D	E10	F11	G8	H
Rosales	Rosaceae	*Prunus*	*avium*				D	E11	F12	G9	H
Fagales	Betulaceae	*Betula*	*pendula*				D	E12	F13	G10	H
Brassicales	Brassicaceae	*Arabidopsis*	*thaliana*				D	E13	F14	G11	H
Sapindales	Sapindaceae	*Aesculus*	*hippocastanum*				D	E14			
Caryophyllales	Polygonaceae	*Fallopia*	*dumetorum*				D	E15		G12	
Ranunculales	Ranunculaceae	*Ranunculus*	*bulbosus*				D	E16		G13	
Asparagales	Iridaceae	*Crocus*	*vernus*				D		F15	G14	
Poales	Poaceae	*Dactylis*	*glomerata*				D	E17	F16	G15	H
Total no. of species				7	10	10	25	17	16	15	12
Total no. of individuals				21	24	10	25	17	16	15	12

The abbreviations of the species sets are: A = Intra-Generic-Set (seven *Melampyrum* species, three populations per species); B = Genus-Set (seven *Melampyrum* species, three populations per species, and three outgroups); C = M1 + 3 out (seven *Melampyrum* species, one population per species); D = Class-Set (complete); E = Class-Set *Eif3E* (Class-Set species for *Eif3E*); F = Class-Set *Agt1* (Class-Set species for *Agt1*); G = Class-Set *Sqd1* (Class-Set species for *Sqd1*); H = Class-Set_Ang-12 (Class-Set species in common to all markers). Numbers for the sets B, E, F and G correspond to the species used in Fig. 2.