Resolving phylogenetic relationships within the *Arracacia* clade (Apiaceae subfamily Apioideae) using cpDNA sequence data

Clark A. Danderson[1], Michael Hermann[2], and Stephen R. Downie[1]
[1]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA
[2]Bioversity International, Cali, Colombia

INTRODUCTION & OBJECTIVES

Previous molecular systematic studies into the higher-level relationships of Apiaceae subfamily Apioideae have revealed a well-supported clade comprising a morphologically heterogeneous assemblage of ten genera and 107 species (*Arracacia, Coaxana*, *Coulterophytum*, *Dahliaphyllum*, *Donnellsmithia*, *Enantiophylla*, *Mathiasella*, *Myrrhidendron*, *Prionosciadium* and *Rhodosciadium*). This clade is named after its largest and earliest described genus, *Arracacia* (Downie *et al.,* 2000, 2001). Recent additions to the *Arracacia* clade, based on analysis of nrDNA ITS sequences, include the Central and South American genera *Neonelsonia* (2 spp.) and *Ottoa* (1 sp.). ITS data have also suggested that *Niphogeton* (18 spp.), *Perissocoeleum* (4 spp.) and *Cotopaxia* (1 sp.) may form a sister group to the clade or comprise its earliest diverging lineages. No obvious morphological synapomorphies are known for the group, although many genera are characterized by polyploid members, the presence of petals with inflexed apices and a distribution in high montane temperate or sub-alpine habitats of Central and South America. The results of phylogenetic analysis of ITS sequences show that the largest genera of the clade, *Arracacia*, *Prionosciadium* and *Rhodosciadium,* are each highly polyphyletic. Additionally, previous efforts to delimit the genus *Arracacia* to the exclusion of other genera have been unsuccessful, resulting in a taxonomy best described as provisional.

In this study, our main objective is to investigate the efficacy of 18 non-coding loci from the chloroplast genome in resolving relationships within the taxonomically difficult *Arracacia* clade, as previous and concurrent studies using ITS sequences result in poorly resolved and weakly supported trees. Using an ITS-derived phylogeny for the group, eight taxa from throughout the tree and one outgroup are chosen for inclusion in this pilot study. The results will indicate which cpDNA loci will be most useful for further investigation of relationships within the *Arracacia* clade.

MATERIALS & METHODS

170 accessions representing 101 taxa were examined for nrDNA ITS sequence variation and analyzed using maximum parsimony in PAUP* (Swofford, 2002). ITS data (excluding 5.8S) for 35 accessions of the *Arracacia* clade and outgroup taxa were obtained previously (Downie and Katz-Downie, 1996; Downie *et al.,* 1998, 2002; Katz-Downie *et al.,* 1999; Sun *et al.,* 2004; C. Calviño, unpublished data); data for all remaining accessions were obtained specifically for this study. For the cpDNA study, eight taxa representing major lineages within the *Arracacia* clade as inferred through ITS (*Arracacia ebracteata*, *A. xanthorrhiza*, *Coaxana purpurea*, *Enantiophylla heydeana*, *Mathiasella bupleuroides*, *Myrrhidendron donnell-smithii*, *Ottoa oenanthoides* and *Rhodosciadium argutum*) and the outgroup taxon *Aethusa cynapium* were chosen. Attempts to obtain all 18 non-coding cpDNA loci for each taxon were made. For five of these regions, sequences of one or two of the taxa are missing (and are noted in Table 1). Partitioned and combined data matrices were analyzed using maximum parsimony. Heuristic searches were implemented using random stepwise addition of taxa and tree-bisection-reconnection (TBR) branch swapping. One hundred bootstrap replicates were performed using the full heuristic search option, with TBR branch swapping, random stepwise addition of taxa and MULTREES options in effect. The effectiveness of each locus for resolving phylogeny was determined by its number of potentially informative characters (PICs = # of substitutions + # of indels) based on Shaw *et al.* (2005, 2007), the number of parsimony informative substitutions and its ability to recover nodes of a fully resolved tree.

RESULTS

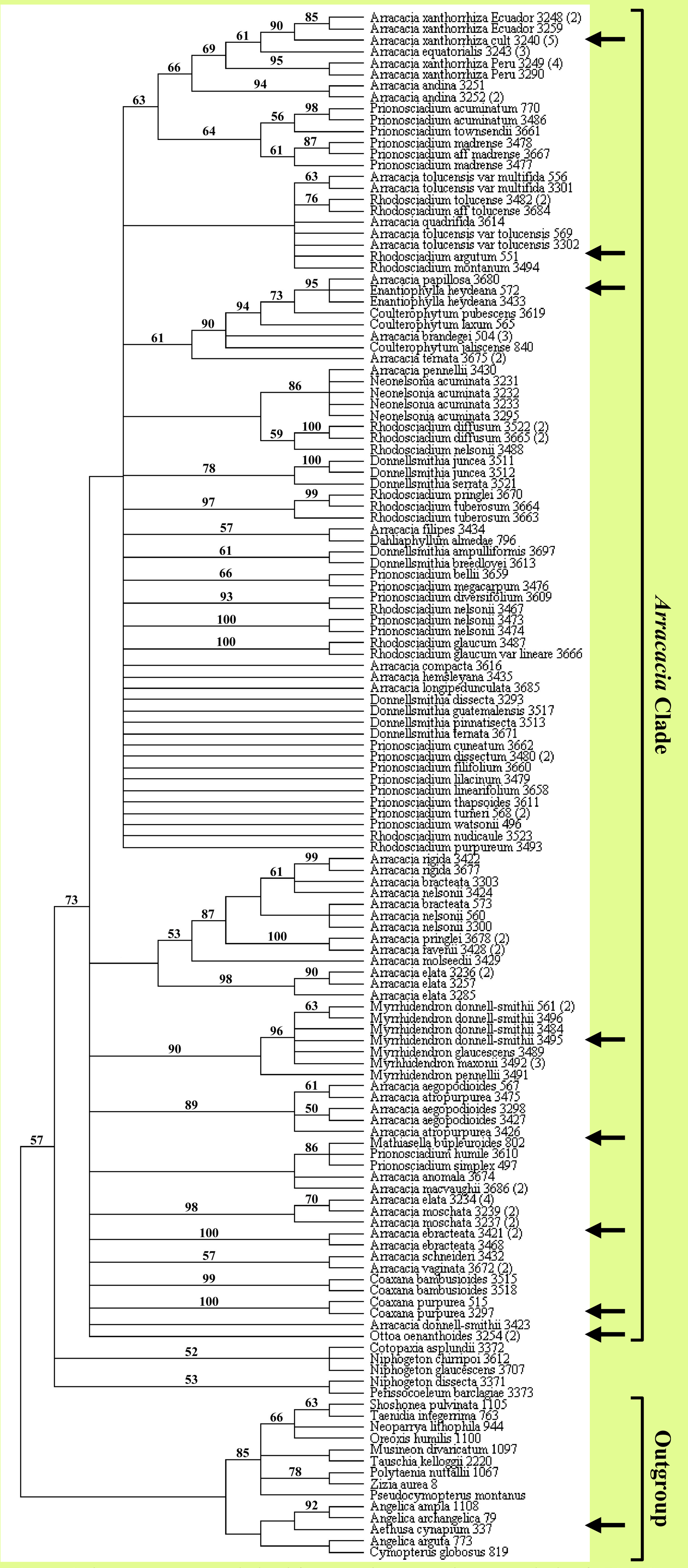

Fig. 1. Strict consensus tree derived from maximum parsimony analysis of 170 ITS sequences from the *Arracacia* clade and outgroups. Numbers above branches are bootstrap values; values <50% are not indicated. The arrows point to the nine taxa used in the cpDNA pilot study.

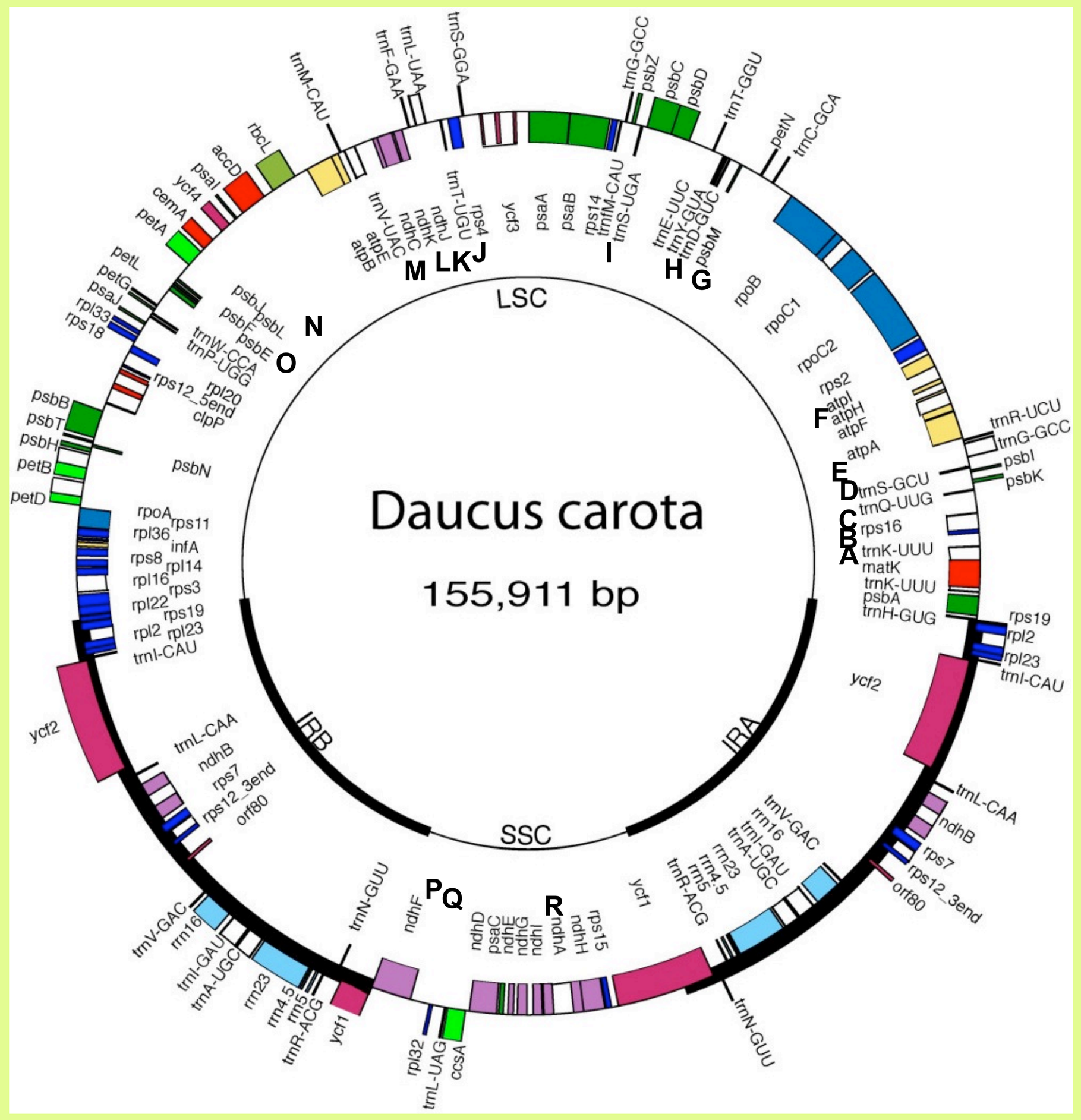

Fig. 2. Map of the chloroplast genome of *Daucus carota* (Scandiceae subtribe Daucinae), modified from Ruhlman *et al.* (2006). Letters (A–R) adjacent to loci correspond with the 18 regions compared in Table 1.

Table 1. Metrics and tree information resulting from sequence comparisons and maximum parsimony analyses of the 18 cpDNA regions investigated (A-R, Fig. 2) plus ITS. Abbreviations: L = length; in = ingroup; out = outgroup; PI = parsimony informative; subst = substitutions; PICs = potentially informative characters; Seq Diver in = pairwise sequence divergence in the ingroup; BS = bootstrap support for the indicated node in a fully resolved tree (Fig. 3a). Nodes absent in the resultant strict consensus trees are indicated by "–" and nodes absent as a result of incomplete sampling are indicated by "N/A". The CI values do not include uninformative characters.

	3' *rps*16–5' *trn*K (A)	*rps*16 intron (B)	*trn*Q–5' *rps*16 (C)	*trn*S–5' *trn*G (D)	*trn*G intron (E)	*atp*I–*atp*H (F)	*ycf*6–*psb*M (G)	*trn*D–*trn*T (H)	*trn*S–*trnf*M (I)	*trn*T–5' *trn*L (J)	*trn*L intron (K)	*ndh*J–3' *trn*L (L)	3' *trn*V–*ndh*C (M)	*psb*J–*pet*A (N)	*pet*L–*psb*E (O)	*ndh*F–*rpl*32 (P)	*rpl*32–*trn*L (Q)	*ndh*A intron (R)	ITS
# Taxa	9	9	9	7	7	9	9	9	7	9	9	9	8	9	9	8	9	9	9
Aligned L (bp)	840	862	1384	573	733	1143	1155	1181	1104	798	518	872	1091	1012	1009	1057	1038	1087	448
Indels in/out	10/6	7/2	10/10	12/4	8/0	5/3	7/6	7/7	7/7	3/2	1/3	7/4	3/5	11/7	5/2	20/7	16/13	8/5	6/6
Subst in/out	13/14	14/14	25/20	15/5	12/8	14/9	18/10	30/12	18/11	13/12	4/1	16/10	14/12	10/20	17/10	30/25	37/17	19/14	98/10
# PI gaps	1	0	4	0	1	1	1	2	0	0	0	2	1	0	1	1	4	2	0
# PI subst	5	3	6	5	1	5	3	7	2	2	2	1	4	5	5	5	18	5	23
PICs	43	37	65	36	28	31	41	56	43	30	9	37	34	48	34	82	83	46	120
Seq Diver in	0–1.29%	0.12–0.82%	0.08–0.88%	0.38–1.72%	0–1.13%	0.27–1.07%	0.18–0.97%	0.37–1.29%	0.37–1.02%	0.13–0.76%	0–0.60%	0.12–1.08%	0.38–1.72%	0–1.09%	0–1.00%	0.31–1.50%	0.21–2.32%	0.09–0.94%	4.35–11.67%
BS Node 1	94	63	88	N/A	N/A	69	–	53	N/A	–	66	–	–	99	–	N/A	100	98	82
BS Node 2	–	–	–	N/A	N/A	–	–	–	N/A	–	–	–	N/A	–	–	–	–	94	–
BS Node 3	–	–	–	–	–	–	–	–	–	–	–	–	–	64	–	–	76	63	–
BS Node 4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	55
BS Node 5	–	–	75	–	–	–	64	86	–	–	66	–	–	–	73	–	66	–	72
BS Node 6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
# of Trees	2	6	39	4	1	9	3	5	1	47	1	18	6	18	6	29	4	2	3
Tree L	28	31	47	19	20	24	28	45	28	28	5	26	28	30	29	46	53	35	136
CI	1.000	0.968	0.957	0.895	1.000	0.917	0.964	0.956	1.000	0.964	1.000	1.000	0.929	1.000	0.966	0.978	0.943	1.000	0.882
RI	1.000	0.800	0.800	0.500	1.000	0.750	0.857	0.800	1.000	0.667	1.000	1.000	0.667	1.000	0.909	0.667	0.880	1.000	0.543

Table 2. Tree information resulting from maximum parsimony analyses of various partitioned and combined data sets. BS = bootstrap support for the indicated node in a fully resolved tree (Fig. 3a). The CI values do not include uninformative characters. Abbreviations: cpDNA all = all 18 examined cpDNA loci; LSC and SSC = all loci from the large and small single copy regions, respectively; Comb. #1 = *psb*J–*pet*A, *rpl*32–*trn*L & *ndh*A intron; Comb. #2 = *trn*Q–5' *rps*16, *psb*J–*pet*A, *rpl*32–*trn*L & *ndh*A intron; Comb. #3 = *trn*Q–5' *rps*16, *trn*D–*trn*T, *psb*J–*pet*A, *rpl*32–*trn*L & *ndh*A intron; Comb. #4 = *trn*D–*trn*T, *psb*J–*pet*A, *rpl*32–*trn*L & *ndh*A intron; Comb. #5 = *trn*Q–5'*rps*16, *trn*D–*trn*T, *rpl*32–*trn*L & *ndh*A intron.

	cpDNA all	cpDNA all w/gaps	LSC region	SSC region	Intergenic spacers	Introns	Gaps only	Comb. #1	Comb. #2	Comb. #3	Comb. #4	Comb. #5	Comb. #3 & ITS	Comb. #5 & ITS
BS Node 1	100	100	100	100	100	100	99	100	100	100	100	100	100	100
BS Node 2	59	–	–	73	–	86	–	73	71	69	72	72	64	64
BS Node 3	97	95	64	95	97	–	–	96	97	99	94	94	95	90
BS Node 4	74	82	80	–	79	–	51	–	–	92	93	93	98	97
BS Node 5	100	100	99	62	99	78	95	77	94	94	74	88	100	98
BS Node 6	–	–	–	55	–	–	–	55	61	67	53	61	88	83
# of Trees	3	3	6	1	2	6	1	3	3	1	1	1	1	1
Tree Length	565	596	429	135	470	93	29	119	166	213	166	183	351	321
CI	0.942	0.926	0.937	0.963	0.940	0.968	0.690	0.966	0.964	0.953	0.952	0.945	0.920	0.913
RI	0.736	0.723	0.690	0.868	0.733	0.850	0.735	0.905	0.885	0.839	0.846	0.818	0.711	0.689

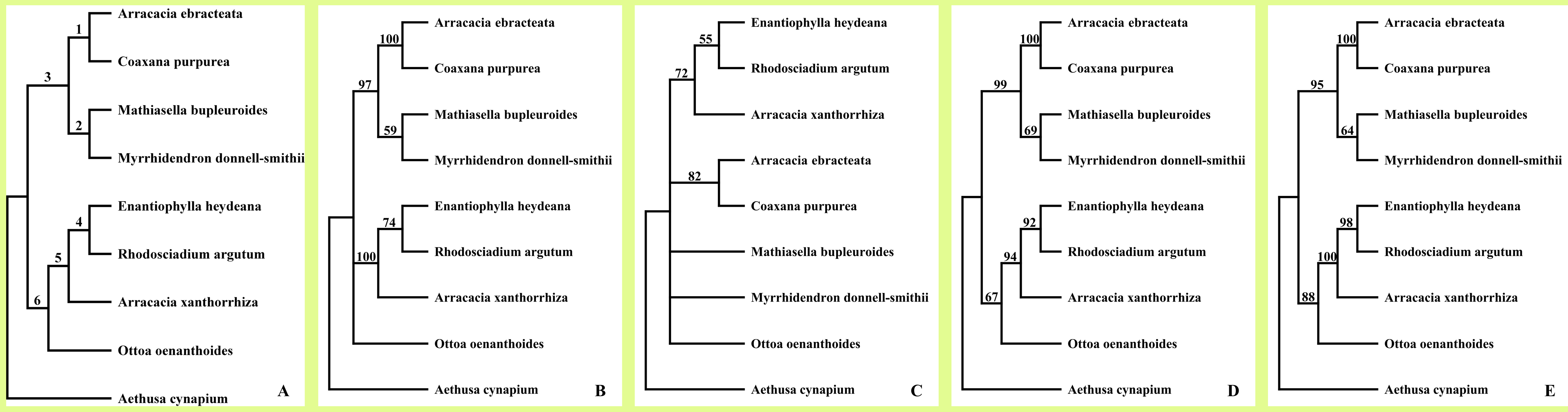

Fig. 3. Strict consensus trees of (A) hypothetical relationships of a fully resolved tree with nodes labeled, (B) the combined 18 cpDNA loci (cpDNA all), (C) ITS region, (D) the *trn*Q–5' *rps*16, *trn*D–*trn*T, *psb*J–*pet*A, *rpl*32–*trn*L, and *ndh*A intron loci (Comb. #3) and (E) Comb. #3 & ITS. Numbers at nodes are bootstrap values.

DISCUSSION AND CONCLUSIONS

Of the 18 cpDNA loci examined, the five regions producing the greatest number of parsimony informative substitutions, highest PICs values and greatest efficacy at resolving the relationships of the taxa (Table 1) were the *trn*Q–5' *rps*16 (C), *trn*D–*trn*T (H) and *psb*J–*pet*A (N) loci of the LSC region and the *rpl*32–*trn*L (Q) and *ndh*A intron (R) loci of the SSC region. Although the *ndh*F–*rpl*32 (P) marker of the SSC region had the second highest PICs value and five parsimony informative substitutions, this region was unable to resolve the relationships within the clade. The most variable and potentially useful cpDNA locus was determined to be the *rpl*32–*trn*L region. This region had 18 parsimony informative substitutions and 83 PICs and was able to recover three of the six nodes in a fully resolved tree (Fig. 3A). This is in agreement with the findings of Shaw *et al.* (2007) in which this region demonstrated the highest number of PICs. In comparison with ITS, these regions are considerably less variable. The ITS data produced 23 parsimony informative substitutions and 120 PICs and the pairwise sequence divergence within the ingroup taxa ranged from 4.35–11.67%; however, this region was only able to recover three nodes (see Fig. 3C) and had a higher incidence of homoplasy. As none of the loci were able to individually recover all six nodes, different combinations of the loci were examined (Table 2). Of the LSC and SSC combined loci, the latter recovered more nodes (five out of six), likely due to the region containing two highly variable loci (*rpl*32–*trn*L and *ndh*A). It is also noted that the more variable intergenic spacer regions, when combined, produced a tree with more well-supported nodes than the less variable intron regions. These results are also in agreement with Shaw *et al.* (2007) as their study suggested that intergenic spacers had a greater average percentage variability than introns. When gaps were included in the analysis of the combined cpDNA loci, the resulting tree was less resolved. This was likely due to the high incidence of homoplasy in the gaps data (see Gaps only in Table 2). The analysis of the combined 18 cpDNA loci (Fig. 3B) was able to recover only five of the six hypothetical nodes; this may be due, however, to the incomplete sampling of six of the loci. The tree topology did not conflict with the tree recovered from ITS (Fig. 3C); however, the ITS tree was less resolved and recovered three of the nodes. There is no discordance between the ITS and cpDNA data sets. When the five most variable cpDNA loci were combined (Fig. 3D), a fully resolved tree was produced with high BS support (>90%) for four of the nodes and moderately weak support (67 and 69%) for the remaining two. When *psb*J–*pet*A is removed from the analysis (Table 2), the same topology is produced with similar BS values, suggesting that the combined loci *trn*Q–5' *rps*16, *trn*D–*trn*T, *rpl*32–*trn*L and *ndh*A intron will equally resolve relationships. When the five most variable loci were combined with ITS (Fig. 3E), a tree of identical topology and greater node support was obtained. This further suggests congruence of the cpDNA and nrDNA data sets. Similar results were obtained when *psb*J–*pet*A is excluded (Table 2).

In conclusion, four regions have been identified as being potentially useful at resolving the relationships of the *Arracacia* clade: *trn*Q–5' *rps*16, *trn*D–*trn*T, *rpl*32–*trn*L and *ndh*A intron.

REFERENCES

Downie, S.R., and D.S. Katz-Downie. 1996. A molecular phylogeny of Apiaceae subfamily Apioideae: evidence from nuclear ribosomal DNA internal transcribed spacer sequences. American Journal of Botany 83: 234–251.

Downie, S.R., S. Ramanath, D.S. Katz-Downie, and E. Llanas. 1998. Molecular systematics of Apiaceae subfamily Apioideae: phylogenetic analyses of nuclear ribosomal DNA internal transcribed spacer and plastid *rpoC1* intron sequences. American Journal of Botany 85: 563–591.

Downie, S.R., D.S. Katz-Downie, and M.F. Watson. 2000. A phylogeny of the flowering plant family Apiaceae based on chloroplast DNA *rpl16* and *rpoC1* intron sequences: towards a suprageneric classification of subfamily Apioideae. American Journal of Botany 87: 273–292.

Downie, S.R., G.M. Plunkett, M.F. Watson, K. Spalik, D.S. Katz-Downie, C.M. Valiejo-Roman, E.I. Terentieva, A.V. Troitsky, B.-Y. Lee, J. Lahham, and A. El-Oqlah. 2001. Tribes and clades within Apiaceae subfamily Apioideae: the contribution of molecular data. Edinburgh Journal of Botany 58: 301–330.

Downie, S.R., R.L. Hartman, F.-J. Sun, and D.S. Katz-Downie. 2002. Polyphyly of the spring-parsleys (*Cymopterus*): molecular and morphological evidence suggests complex relationships among the perennial endemic genera of western North American Apiaceae. Canadian Journal of Botany 80: 1295–1324.

Katz-Downie, D.S., C.M. Valiejo-Roman, E.I. Terentieva, A.V. Troitsky, M.G. Pimenov, B.-Y. Lee, and S.R. Downie. 1999. Towards a molecular phylogeny of Apiaceae subfamily Apioideae: additional information from nuclear ribosomal DNA ITS sequences. Plant Systematics and Evolution 216: 167–195.

Ruhlman, T., S.B. Lee, R.K. Jansen, J.B. Hostetler, L.J. Tallon, C.D. Town, and H. Daniell. 2006. Complete plastid genome sequence of *Daucus carota*: implications for biotechnology and phylogeny of angiosperms. BMC Genomics 7: 222.

Shaw, J., E.B. Lickey, J.T. Beck, S.B. Farmer, W. Liu, J. Miller, K.C. Siripun, C.T. Winder, E.E. Schilling, and R.L. Small. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92: 142–166.

Shaw, J., E.B. Lickey, E.E. Schilling, and R.L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany 93: 275–288.

Sun, F.-J., S.R. Downie, and R.L. Hartman. 2004. An ITS-based phylogenetic analysis of the perennial, endemic Apiaceae subfamily Apioideae of western North America. Systematic Botany 29: 419–431.

Swofford, D.L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10 for 32-bit Microsoft Windows. Sunderland: Sinauer Associates.

ACKNOWLEDGMENTS

We thank the University of Illinois Graduate College for the Conference Travel Grant which provided support to attend this conference. We thank our collaborator C. Calviño for use of her unpublished sequences as well as the curators of those herbaria who provided access to specimens. We would like to specially thank D. Katz-Downie, C. Calviño and J. Cordes for their assistance in the lab. Funding for this project was provided through an NSF grant (DEB 0089452), the Herbert H. Ross Award (Illinois Natural History Survey), the Francis M. and Harlie M. Clark Research Support Grant (School of Integrative Biology, University of Illinois at Urbana-Champaign), and the Harold C. and Sonja L. Labinsky Award (School of Integrative Biology, University of Illinois at Urbana-Champaign).