American Journal of Botany

American Journal of Botany 97(5): 874–892. 2010.

Migration of endpoints of two genes relative to boundaries between regions of the plastid genome in the grass family (Poaceae)[1]

Jerrold I. Davis[2,4] and Robert J. Soreng[3]

[2]L.H. Bailey Hortorium and Department of Plant Biology, Cornell University, 412 Mann Library, Ithaca, New York 14853-4301 USA; and [3]Department of Botany and United States National Herbarium, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20013-7012 USA

Overlapping genes occur widely in microorganisms and in some plastid genomes, but unique properties are observed when such genes span the boundaries between single-copy and repeat regions. The termini of *ndhH* and *ndhF*, situated near opposite ends of the small single-copy region (SSC) in the plastid genomes of grasses (Poaceae), have migrated repeatedly into and out of the adjacent inverted-repeat regions (IR). The two genes are transcribed in the same direction, and the 5′ terminus of *ndhH* extends into the IR in some species, while the 3′ terminus of *ndhF* extends into the IR in others. When both genes extend into the IR, portions of the genes overlap and are encoded by the same nucleotide positions. Fine-scale mapping of the SSC-IR junctions across a sample of 92 grasses and outgroups, integrated into a phylogenetic analysis, indicates that the earliest grasses resembled the related taxa *Joinvillea* (Joinvilleaceae) and *Ecdeiocolea* (Ecdeiocoleaceae), with ca. 180 nucleotides of *ndhH* extending into the IR, and with *ndhF* confined to the SSC. This structure is maintained in early-diverging grass lineages and in most species of the BEP clade. In the PACMAD clade, *ndhH* lies completely or nearly completely within the SSC, and ca. 20 nucleotides of *ndhF* extend into the IR. The nucleotide substitution rate has increased in the PACMAD clade in the portion of *ndhH* that has migrated into the SSC.

Key words: gene overlap; inverted-repeat region; *ndhF*; *ndhH*; nucleotide substitution rate; PACMAD clade; phylogenetics; plastid genome; Poaceae; small single-copy region.

In most embryophytes, the plastid genome is divided into four major regions, the large single-copy region (LSC), small single-copy region (SSC), and two intervening inverted repeat regions (IRs), which are identical in sequence but arranged in reverse order (reviewed by Palmer, 1985; Sugiura, 1992; Bock, 2007). If a gene spans a boundary (also known as a junction) between the IR and one of the single-copy regions, one copy of the fragment that lies within the IR is contiguous with the rest of the gene in the adjacent single-copy region, and the other copy is adjacent to whatever lies at the other end of the same single-copy region (e.g., Fig. 1). Two genes, *ndhF* and *ndhH*, are situated at opposite ends of the SSC region in most angiosperms, and each extends into the IR in some grass species. By sequencing across the two SSC-IR junctions, the positions of both genes, relative to the junctions, can be determined. Here we describe the varied occurrence of one or both of these genes spanning the SSC-IR junction in a range of species of the grass family and provide evidence for repeated migrations of portions of these genes across the junctions. Nucleotide substitution rates generally are slower in the IR than in the single-copy regions (e.g., Wolfe et al., 1987; Maier et al., 1995; Muse and Gaut, 1997; Yamane et al., 2006), and we document an acceleration in substitution rate that coincides with the migration of a portion of *ndhH* from the IR into the SSC in taxa of the PACMAD clade, a major lineage within the grass family, comprising subfamilies Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae.

[1] Manuscript received 4 August 2009; revision accepted 22 March 2010.

The authors thank K. Allred, C. Annable, P. Asimbaya, N. Barker, L. Clark, J. Conran, M. Crisp, J. Dransfield, G. Hui, S. Jacobs, S. Jones, E. Judziewicz, J. LaDuke, H. P. Linder, P. Peterson, E. Royl, P. Rudall, C. Schiers, N. Soreng, D. Stevenson, J. Wipff, W. Zhang, Fairchild Tropical Botanic Garden, Royal Botanic Gardens, Kew, and the USDA Plant Introduction Station for access to plant materials, M. Voionmaa for technical assistance, the U.S. NSF (grant #DEB-0318686) for funding in support of this research and Scot Kelchner and an anonymous referee for valuable comments.

[4] Author for correspondence (e-mail: jid1@cornell.edu)

doi:10.3732/ajb.0900228

Grass phylogenetics—The grass family (Poaceae) comprises ca. 10 000 species (Clayton and Renvoize, 1986; Watson and Dallwitz, 1992), most of them herbaceous annuals and perennials, though the family also includes the woody-textured bamboos. Phylogenetic analyses of the monocots, variously based on molecular and morphological characters, have converged in the placement of two small plant families, Joinvilleaceae and Ecdeiocoleaceae, as the closest relatives of the grasses (e.g., Doyle et al., 1992; Linder and Rudall, 1993; Chase et al., 1995, 2006; Linder and Kellogg, 1995; Kellogg and Linder, 1995; Stevenson and Loconte, 1995; Briggs et al., 2000; Bremer, 2002; Michelangeli et al., 2003; Davis et al., 2004; Graham et al., 2006; Marchant and Briggs, 2007; Bouchenak-Khelladi et al., 2008).

Analyses focusing on the grasses have provided an increasingly well-substantiated phylogenetic structure for the family. A major landmark in this effort was the combined analysis of several molecular character sets, plus a set of structural characters, yielding an overall phylogeny and a classification consistent with it as proposed by the Grass Phylogeny Working Group (GPWG, 2001, and prior analyses reviewed therein). The classification proposed by the GPWG recognized 12 subfamilies. Among them were four large subfamilies (Bambusoideae,

Chloridoideae, Panicoideae, and Pooideae, each including more than 1000 species) that collectively included ca. 90% of all grass species, plus eight smaller subfamilies and a few anomalous genera that were not assigned to subfamily. Three of the smaller subfamilies were placed as a series of lineages (Anomochlooideae, Pharoideae, and Puelioideae) diverging in succession from a major clade that includes the four large subfamilies and all other grasses. We use the phrase "early-diverging" to refer to the three small subfamilies that diverged early in the history of the family from the lineage that now includes most grass species. Within the major clade, all species fall into one or the other of two large subclades, one comprising subfamilies Panicoideae, Arundinoideae, Chloridoideae, Centothecoideae, Aristidoideae, and Danthonioideae, and designated the PACCAD clade, the other comprising subfamilies Bambusoideae, Ehrhartoideae, and Pooideae, and designated the BEP clade. Both of these clades had been observed in previous analyses, and prior acronyms for them are PACC (Davis and Soreng, 1993) and BOP (Clark et al., 1995), respectively.

Many additional phylogenetic analyses of the grass family have been conducted since, variously focusing on the overall structure of the family or on particular lineages within it. Within what had been designated the PACCAD clade, representatives of the small subfamily Centothecoideae long have been associated with Panicoideae (e.g., Clark et al., 1995; Soreng and Davis, 1998; Hilu et al., 1999; Mathews et al., 2000; GPWG, 2001; Duvall et al., 2007), and a recent analysis indicated that lineages from these two subfamilies are intermixed to the extent that neither is monophyletic (Sánchez-Ken and Clark, 2007), thus suggesting that taxa from the former Centothecoideae should be subsumed within a broadly defined Panicoideae. A formal classification along these lines had been published earlier (Zuloaga et al., 2003). Another recent development (Sánchez-Ken et al., 2007) has been the reinstatement of Micraroideae, which includes a putative monophyletic assemblage of *Micraira*, *Eriachne*, and other genera previously placed in isolated locations within the PACCAD clade, and sometimes left unassigned to subfamily. With the incorporation of Centothecoideae into Panicoideae and the adoption of Micraroideae, the PACCAD clade is unchanged in composition, but now is designated the PACMAD clade (Duvall et al., 2007). Thus, a modified version of the GPWG classification, also comprising 12 subfamilies, is recognized today, with the same three small early-diverging subfamilies, plus the nine subfamilies of the BEP and PACMAD clades.

Plastid genome structure—The plastid genome of embryophytes is double-stranded, usually ca. 120–160 kb long, and it usually includes more than 120 genes (Palmer, 1985; Bock, 2007; Jansen et al., 2007). Being divided into the LSC, SSC, and two equal and intervening IR regions, it falls within the category of amphimeric genomes (Rayko, 1997). Although one copy of the IR has been lost from some plant lineages (e.g., Lavin et al., 1990), most plastid genomes that have been examined retain both copies. The plastid genome has been employed extensively in plant phylogenetic studies, some of which have principally used variation in nucleotide sequences (e.g., Chase et al., 1993; Soltis et al., 2000; Leebens-Mack et al., 2005; Hansen et al., 2007), while others have examined structural features, such as insertions/deletions (indels), inversions, and variation in the positions of boundaries between the genomic regions (e.g., Doyle et al., 1992; Goulding et al., 1996; Plunkett and Downie, 2000; Perry et al., 2002; Cosner et al., 2004; Stefanović and Olmstead, 2005; Wang et al., 2008).

Although the plastid genome conventionally is mapped as a circular, double-stranded DNA molecule, with fixed relationships between the two LSC, SSC, and two IR regions, and with the latter usually labeled as separate structures (IR_A and IR_B), it actually exists in a variety of conformations within living cells, including interlinked aggregations of multiple copies (Palmer, 1983; Bendich, 2004; Oldenburg and Bendich, 2004; Bock, 2007). However, the conventional circular map does summarize many of the heritable structural features of the genome correctly. Among plant lineages that retain both copies of the IR, the sizes and boundaries of these regions, relative to the adjacent LSC and SSC, have been modified in various groups (e.g., Goulding et al., 1996; Aii et al., 1997; Plunkett and Downie, 2000; Perry et al., 2002; Stefanović and Olmstead, 2005; Wang et al., 2008).

Within the grass family, two genes that lie near opposite ends of the SSC region (*ndhF* and *ndhH*) sometimes extend into the IR regions (Maier et al., 1990; Ogihara et al., 2002; Davis and Soreng, 2007; Saski et al., 2007; Bortiri et al., 2008). These two genes are transcribed in the same direction within the SSC region (Fig. 1). Thus, the 5′ terminus of *ndhH* is situated near one SSC-IR junction (JSA), and the 3′ terminus of *ndhF* is situated near the other (JSB). In *Triticum aestivum* and *Oryza sativa*, *ndhH* extends across the junction, with the 5′ terminus and more than 150 nucleotides situated within the IR, while *ndhF* lies fully within the SSC. In *Zea mays*, just one nucleotide at the 5′ terminus of *ndhH* lies within the IR, while *ndhF* extends across the junction, with the 3′ terminus and 29 nucleotides situated within the IR. Ogihara et al. (2002) concluded from the similarities between *Triticum* and *Oryza* in these characteristics that these two taxa are more closely related to each other than either is to *Zea*, but in making this suggestion they did not provide evidence that the similarity is synapomorphic. *Triticum* (BEP clade, subfamily Pooideae) is, in fact, believed to be more closely related to *Oryza* (BEP clade, subfamily Ehrhartoideae) than either is to *Zea* (PACMAD clade, subfamily Panicoideae). However, one result of the current study was the demonstration that in most taxa of the sample, including grasses of early-diverging lineages, and outgroups, ca. 150–250 nucleotides (maximum 400) of *ndhH* extend into the IR region, the major exception being taxa of the PACMAD clade. Specifically, the occurrence of this portion of *ndhH* in the IR (as in *Triticum* and *Oryza*) is interpreted as the symplesiomorphic state in the grasses, and the migration of this gene region relative to the SSC-IR boundary, resulting in the transfer of most of these sites into the SSC (as in *Zea*), is interpreted as derived within the PACMAD clade. Meanwhile, the 3′ terminus of *ndhF* lies entirely within the SSC in the outgroups and most grasses and is interpreted as having migrated into the IR region in an ancestor of the PACMAD clade. The portion of *ndhH* that has migrated into the SSC from the IR in the PACMAD clade has experienced a corresponding nucleotide substitution rate acceleration.

When both genes extend into the IR, as in *Zea*, one or more nucleotide positions just within the IR encode portions of both genes. Within this region, the two genes are encoded on opposite strands, because the reading frame of *ndhH* extends from the IR into the SSC, while that of *ndhF* runs in the opposite direction. This situation, with two genes overlapping where they extend from different ends of a single-copy region into the same end of an IR, is a special case of the general phenomenon of gene overlap, which occurs widely in microorganisms (Fukuda et al., 2003) and some plastid genes (e.g., *psbD* and *psbC*). In this particular form of gene overlap, the region of overlap is

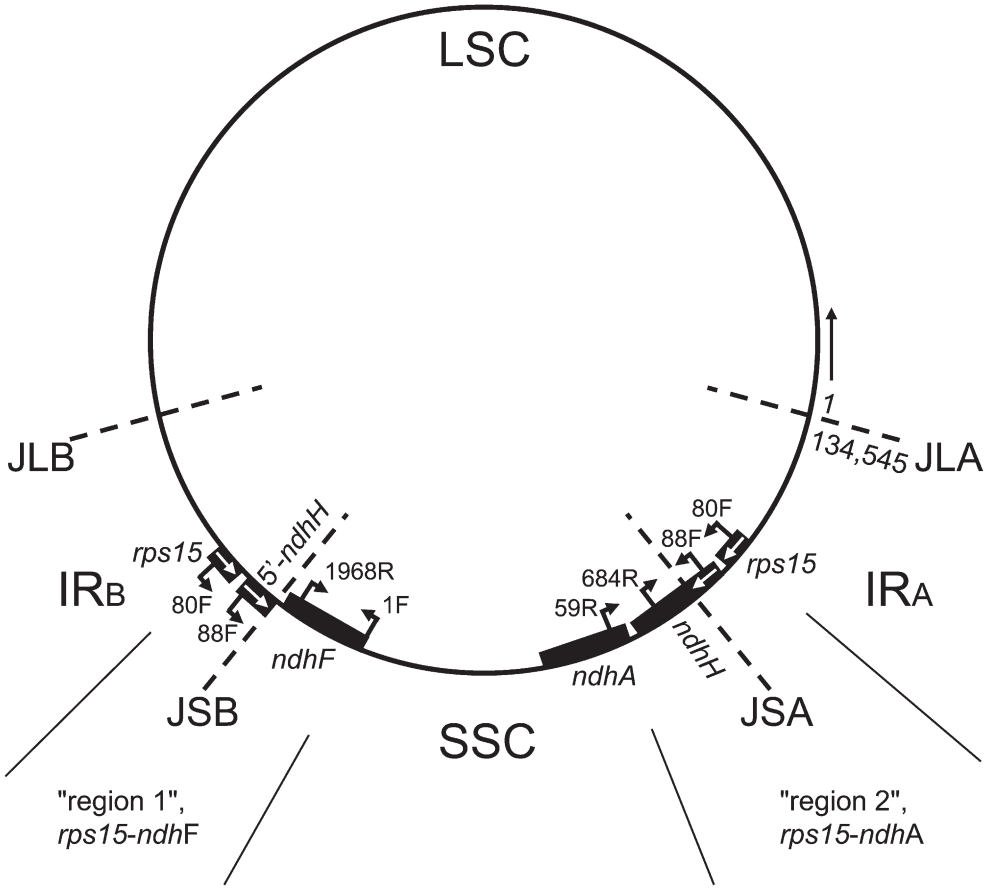

Fig. 1. Map of the plastid genome of *Triticum aestivum* (GenBank accession NC_002762), depicting the positions of two regions that were sequenced, and genes and gene fragments within these regions; lengths of genes and genomic regions are not drawn to scale. Junctions JLA, JLB, JSA, and JSB, identified by dashed lines, delimit large and small single-copy regions (LSC and SSC) and the two inverted repeat regions, IRA and IRB. Each of the sequenced regions spans one SSC-IR boundary. Numbering of nucleotides of the genome (1 through 134 545) begins at JLA and proceeds counterclockwise, as indicated. Genes depicted outside the circle are transcribed in counterclockwise sequence, and genes depicted inside the circle are transcribed in clockwise sequence; the direction of gene transcription in IR regions is indicated by white arrows. In *Triticum*, *ndhF*, *ndhA*, and the 3′ end of *ndhH* lie within the SSC region, while *rps15* and the 5′ end of *ndhH* lie within the IR regions, and thus are present as two copies. Positions and priming orientations (black arrows) are depicted for the primers used most frequently to amplify and sequence these regions; a complete list of primers is provided in Appendix 2.

coterminous with the portion situated in the IR of the gene that extends the shortest distance into the IR, and it includes only the terminal portion of the gene. Here we document four origins of gene overlap in the IR of grasses. In all but one case, the extent of the overlap is no more than four nucleotides in length, a pattern that suggests either physical instability or maladaptiveness of such an overlap. However, in one case, the overlap region is 43 nucleotides in length.

MATERIALS AND METHODS

Taxon sample—The taxon sample includes 90 representative species of Poaceae, plus one each of Joinvilleaceae and Ecdeiocoleaceae (Appendix 1). The taxonomic system adopted here for the grasses, at the subfamily level, is a modification of the 12-subfamily system proposed by the GPWG (2001), excluding Centothecoideae and including Micrairoideae, as described above. Of the 12 subfamilies, all are sampled except Puelioideae, for which a suitable DNA isolation was not available. Genera are assigned to subfamily and tribe (and to subtribe within Poeae) according to the Catalogue of New World Grasses (CNWG) (Judziewicz et al., 2000; Peterson et al., 2001; Soreng et al., 2003; Zuloaga et al., 2003 [and online revision of 18 June 2009 http://mobot.mobot.org/W3T/Search/nwgclass.html; archived web page available on request]). The CNWG classification is comprehensive at the generic level for all grass genera occurring in the Americas and includes miscellaneous additional genera for informational purposes. Most genera in the present analysis are classified in the CNWG system, either explicitly, or implicitly via nomenclatural linkage (e.g., by the inclusion of a tribe whose name is based on that of a genus that does not occur in the Americas). For four genera that were not assigned to subfamilies by the GPWG (2001), we follow the CNWG system (cf. Zhang, 2000; Sánchez-Ken and Clark, 2001, 2007; Duvall et al., 2007; Sánchez-Ken et al., 2007) in placing *Streptogyna* in Ehrhartoideae, *Gynerium* in Panicoideae, and *Micraira* and *Eriachne* in Micrairoideae. Four other taxa in the present analysis are not classified in the CNWG system, and they are assigned to subfamilies and tribes as follows (cf. Barker, 1997; Barker et al., 1999, 2007; GPWG, 2001; Pirie et al., 2008): *Amphipogon* (Arundinoideae, Arundineae); *Stipagrostis* (Aristidoideae, Aristideae); *Merxmuellera macowanii* (Danthonioideae, Danthonieae); and *Merxmuellera rangei* (Chloridoideae, no tribal assignment).

DNA methods—Two regions of the plastid genome were sequenced, each of them spanning one of the two SSC-IR boundaries (JSA and JSB; Fig. 1). One of the regions ("region 1", *rps15-ndhF*) extends from a point within *rps15*, in the IR region, across JSB to a point within and near the 5′ terminus of *ndhF*, in the SSC region (Fig. 1). The other region ("region 2", *rps15-ndhA*) extends from the same point within *rps15* across JSA to a point near the 5′ terminus of *ndhA*, at the other end of the SSC region. Nucleotide sequences were obtained from total genomic DNA isolations, using standard PCR and automated cycle sequencing methods. Amplification and sequencing primers are described in

Appendix 2, and positions of the most frequently used amplification primers are indicated in Fig. 1. Because both regions include identical portions of the IR, some primers (e.g., *rps15*-80F) were used in the amplification and sequencing of both regions. Also, for taxa in which a portion of *ndhH* extends into the IR region that is long enough to include the region corresponding to primer *ndhH*-88F (i.e., approximately the first 113 nucleotides of this gene), this primer sometimes was used in the sequencing of both regions. This condition is met in most taxa in the sample, the major exceptions being those of the PACMAD clade. Successful amplification and sequencing of the two regions provides complete sequences of *ndhH* (length of the complete gene is 1182 nucleotides in the reference plastid genome sequence of *Triticum aestivum*, GenBank accession NC_002762); 122 nucleotides of *rps15* (length of the complete gene is 273 nucleotides in the same reference sequence), and nearly complete sequences of *ndhF* (lacking ca. 50–70 nucleotides from the 5′ terminus; the complete length of *ndhF* is 2220 nucleotides in the same reference sequence).

Data structure and analysis—Sequences of *ndhH, ndhF,* and *rps15* were generated, aligned manually, and deposited in GenBank (Appendix I). Nucleotide sites within inferred gaps were encoded as "unknown" and thus treated as missing characters for taxa with deletions. Regions interpreted as only ambiguously alignable were excluded from analyses. Nine structural features of the three genes with parsimony-informative distributions, including two inversions, seven indels, and two characters representing presence/absence of portions of *ndhF* and *ndhH* in the IR regions, were identified and encoded as binary characters (Tables 1, 2). The indels were scored using simple gap coding (Simmons and Ochoterena, 2000). Nucleotides of *ndhF* within the two inversion regions were included in the analysis by replacing the observed sequence with its reverse complement for each taxon that was interpreted as being inverted (Graham et al., 2000; Soreng et al., 2007). For archival purposes, the data matrix includes the sequences as observed, but with the nucleotides in the inversion regions inactivated, and with these regions duplicated elsewhere in the matrix as active characters, with the observed sequences replaced by the reverse complements for the inverted taxa. The data matrix and explanatory text are available as supplemental materials (Appendices S1–S3, see Supplemental Data with online version of article).

Because both of the sequenced regions extend into *rps15* and because primer *rps15*-80F was used to amplify both of them, the portion of this gene that lies within the sequenced regions was sequenced twice from each taxon, as were all nucleotides between *rps15* and the SSC-IR boundaries. The portions of the two sequenced regions that extend into the IR region were aligned against each other to identify the locations of the two SSC-IR boundaries in each taxon (JSA and JSB, the points at which the sequences first differ from each other), and the positions of the 3′ terminus of *ndhF* and the 5′ terminus of *ndhH* relative to these boundaries (cf. Fig. 2). Contradictory nucleotide sequences were never observed within the twice-sequenced segment of the IR region for any taxon. In some cases, however, one or more sites within one or the other of the two fragments within the IR region was not read clearly, usually in the intergenic spacer region, but in some cases within *rps15*. In the latter cases, the reported *rps15* sequence for a taxon is a composite of readable portions of both sequences.

For most taxa, just one of the two genes (*ndhF* or *ndhH*) or neither of them extends into the IR region. In a few cases, however, both of these genes extend into the IR region, and in these cases there are nucleotide sites that lie within both genes and thus are homologous between the two genes (Fig. 2). The two genes are encoded in reverse order in the IR, so the sense strand of one is the antisense strand of the other, but nonetheless, any nucleotide that lies within the IR region in both genes is homologous between the two genes. To avoid including these nucleotides twice in the analysis (i.e., once within the *ndhF* sequence of a taxon, and once within the *ndhH* sequence of the same taxon), the duplicated nucleotides were excluded from the *ndhF* sequences prior to analysis. This exclusion affected a total of 70 cells of the data matrix, and of these, seven are in characters that are active and parsimony informative in the data matrix. To determine whether the exclusion of these seven cells affected the results of the principal analysis (see below), two additional analyses were conducted, one of them with these cells included in the *ndhF* sequences but excluded from the *ndhH* sequences, and the other with these cells included in the sequences of both genes (i.e., included twice in the data matrix). The sets of trees obtained by both of these analyses were identical to those obtained by the principal analysis, and no further reference to the other analyses is made. The archived data matrix includes 46 characters, for which 70 cells are nonempty, immediately following *ndhF*. These cells represent the sites interpreted as homologous between *ndhF* and *ndhH*, and the corresponding cells in the *ndhF* portion of the matrix are scored "N".

Parsimony analysis was conducted with the program TNT, version 1.1 (Goloboff et al., 2008), with all characters weighted equally and coded as nonadditive, missing nucleotides in indel regions coded as unknown, clades interpreted as resolved in individual trees only if supported under all possible optimizations (using the command "*collapse 3*"), and with *Joinvillea* specified as the outgroup. Parsimony-uninformative characters were removed from the matrix prior to analysis, and all reported tree lengths and consistency indices are based on only the parsimony-informative portion of the matrix. The data matrix was analyzed by invoking 1000 replicate search initiations with random taxon addition sequences, with 20 trees held per replicate and subjected to exhaustive TBR swapping and then to 200 ratchet iterations (Nixon, 1999), using 5% probabilities for character upweighting and downweighting. All trees obtained by these searches were pooled and subjected to exhaustive TBR swapping with up to 1 000 000 trees held in memory. A separate analysis was conducted using just the nucleotide sequence characters. Strict-consensus jackknife support (JS) (Farris et al., 1996; Soreng and Davis, 1998; Davis et al., 2004) was calculated from 10 000 replicates using a deletion frequency of 37%, with each jackknife replicate conducted using the same search procedures as used in the basic analysis, except that 10 search initiations were conducted for each jackknife replicate, rather than 1000, and up to 1000 trees were held during the final tree-bisection-reconnection (TBR) swapping phase, rather than 1 000 000. The program WinClada version 1.03 (Nixon, 2002) was used to edit data, determine tree lengths and related indices, examine character transformations on the resulting trees, and generate figures.

Optimizations of six structural characters are illustrated on a consensus tree (Fig. 3). The decision to present character optimizations on the consensus tree was made only after it was verified that the all transformations of these characters are identical in position and number among all most-parsimonious trees and the consensus tree. Optimization of one of these characters is ambiguous in one region of the tree (character 16, Tables 1, 2, near *Arundo* in Fig. 3; either two parallel gains or a gain and a loss), and this ambiguity is shared among all most-parsimonious trees and the consensus tree; all other optimizations of this and other characters mapped in Fig. 3 are unambiguous.

Nucleotide substitution rates—To determine whether the inferred shift in the relative positions of *ndhH* and the SSC-IR boundary is associated with a change in the nucleotide substitution rate in this region, we compared the number of inferred steps in this gene region among taxa of the PACMAD clade to the number of steps among other taxa. This comparison was conducted by enumeration of substitutions in various gene regions, with a focus on the portion of *ndhH* that migrated from the IR region into the SSC region in the PACMAD clade. This was conducted in a cladistic framework, without reference to a model or other method to account for unobserved steps. Steps were counted on one randomly selected most-parsimonious tree for the combined data set, as optimized under accelerated transformation, along all branches within the PACMAD clade, and compared to the corresponding sum for all other branches in the tree, except for the branch leading to the PACMAD clade. This branch was excluded from both of these categories, because transformations along it cannot be assigned either to the IR or SSC region (i.e., each substitution that is optimized on this branch could have occurred either before or after the migration of this portion of *ndhH*, which also lies on this branch). Corresponding comparisons also were made between numbers of steps within and outside of the PACMAD clade (always excluding the branch that leads to the PACMAD clade) for three other gene regions, each of which lies either within the IR region or the SSC region in all or nearly all taxa. Ambiguously aligned nucleotides, which had been excluded from cladistic analyses, also were excluded from these comparisons, but all other variable sites were included, regardless of whether they were parsimony informative, so that the sums would include autapomorphic steps. The four regions compared, and their aligned lengths, are as follows: (1) *ndhH* sites 5–154, 150 nucleotides—This region lies entirely within the SSC region in all taxa of the PACMAD clade and entirely within the IR region in all but seven taxa outside the PACMAD clade (entirely within the SSC region in *Celtica*, *Pleuropogon*, and *Olyra*, and partially within the SSC region in *Lithachne*, *Brachypodium*, *Brylkinia*, and *Molineriella*). (2) *ndhH*, from site 301 to the 3′ terminus, 894 nucleotides—This region lies entirely within the SSC region in all taxa of PACMAD clade and in all except four taxa outside the PACMAD clade (*Cynosurus* and the three species of *Bromus*); in each of the latter four taxa, no more than 100 nucleotides of this region lie within the IR region. (3) *ndhF*, including nine sites from the two inversions, and excluding unambiguously aligned regions and 17 additional sites near the 3′ terminus, which lie within the IR region in some taxa, 2105 nucleotides—This region lies entirely within the SSC region in all taxa examined. (4) *rps15*, from site 152 to the 3′ terminus, 122 nucleotides—This region lies entirely within the IR region in all taxa examined.

Fig. 2. Nucleotide sequences for both DNA strands at the two SSC-IR junctions of the plastid genomes of six grass taxa (cf. Appendix 1, Fig. 1). IR regions lie to the left of each junction (JSA and JSB), and the SSC region to the right. Boldface, capital letters signify encoding regions (sense strand) near the 5′ terminus of *ndhH*, also labeled at right as reading across the boundary into the SSC region, and underlined capital letters signify encoding regions (sense strand) near the 3′ terminus of *ndhF*, also labeled at right as reading across the boundary into the IR region.

RESULTS

Data structure—The combined sequences of the three genes sum to 3438 aligned sites, of which 907 (26.4%) are parsimony-informative. In addition to the nucleotide sequence characters, the inclusion of nine parsimony-informative structural characters (Tables 1–3) brought the total number of informative characters in the matrix to 916. Details concerning the distribution of states of these characters among the taxa in the sample, and on cladograms, are provided below.

Grass phylogenetics—Analysis of the combined matrix of three gene sequences and nine structural characters yielded 12 most-parsimonious trees of length 4024, CI 0.35, and RI 0.69 (Table 3, Fig. 3). Trees of this length were obtained by each of the 1000 searches that were conducted prior to the final phase of TBR swapping. The same set of trees was obtained when the analysis included only the nucleotide sequence characters; further discussion (e.g., support values) is based on results obtained with nucleotide and structural characters. Monophyly was not tested for Pharoideae and Aristidoideae, as each was sampled only once. Of the nine grass subfamilies represented by at least two species, eight are resolved as monophyletic with the current taxon sampling. In four of these cases, JS is between 98 and 100%, and among the other four, support for monophyly of Danthonioideae is 77%, and for each of the remaining, three it is less than 60%. In two of these cases, the support for a core group is strong, but an additional accession is weakly associated with the core group, so support for the subfamily itself is low. One of these cases is Chloridoideae, which has 51% JS, with *Merxmuellera rangei* placed as the sister of the rest of the subfamily, and with 100% support for the clade that includes all

TABLE 1. Descriptions, encoding rules, number of steps, and consistency of nine structural characters of the plastid genome and nine nucleotides within two inversion regions. Locations of inversions and indels refer to the *ndhF* and *ndhH* sequences of the reference sequences of *Triticum aestivum* (GenBank accession NC_002762) and may differ by a few nucleotides under alternative alignments.

Character no. (no. of steps, CI, and RI on most-parsimonious trees for structural characters), character description
Character 0 (6, 0.16, 0.44). presence/absence of one or more nucleotides of *ndhH* in the IR region (cf. Figs. 1, 2): 0 = absence, 1 = presence. Number of nucleotides in the IR region is provided in Table 2 and as the first of two numbers (hyphen = 0) after each taxon name in Fig. 3.
Character 1 (4, 0.25, 0.88): presence/absence of one or more nucleotides of *ndhF* in the IR region (cf. Figs. 1, 2): 0 = absence, 1 = presence. Number of nucleotides in the IR region is provided for each taxon in Table 2 and as the second of two numbers (hyphen = 0) after each taxon name in Fig. 3. Parenthesized numbers in Fig. 3 denote presence and number of nucleotides inserted (+) or deleted (−), relative to the reference sequence of *Triticum aestivum*, within the portion of *ndhF* that extends into the IR region. For example, in *Arundo* ["17(–3)"], 17 nucleotides of *ndhF* lie within the IR region, and there is a 3-bp deletion within this portion of the gene, relative to *Triticum*, so the portion of *ndhF* within the IR region in *Arundo* is interpreted as being homologous with a 20-nucleotide portion in *Triticum*.
Character 2 (2, 0.50, 0.95). *ndhF* inversion of sites 1918–1920: 0 (uninverted) = GTA (55 taxa) or a sequence differing from this at no more than one site, and from the inverted sequence at all three sites: ATA (1), CTA (4), GCA (1), GGA (1), and GTT (1). 1 (inverted) = TAC (23 taxa) or a sequence differing from this at no more than one site and from the inverted sequence at all three sites: TAT (1). Six taxa differ from one of the main sequences at one site and from the other at two sites, and they are scored unknown for the inversion and for the three nucleotides in the *ndhF* sequence: TTA (4), TTC (1), GAA (1).
Characters 3–5: sequence in region of *ndhF* inversion 1 or for taxa interpreted as having the inversion (state 1 of character 2), the reverse complement of the observed sequence. Taxa scored as unknown for character 2 are also scored as unknown for these three characters.
Character 6 (1, 1.00, 1.00): *ndhF* inversion of sites 1932–1937: 0 (uninverted) = GAAAAA (46 taxa) or a sequence differing from this at no more than three sites and from the inverted sequence at no fewer than five sites: AAAAAA (14), CAAAAA (13); CCAAAA (5); GAAAAG (1); GGAACA (1), TAAAAA (9), AAAAAG (1), 1 (inverted) = TTTTTC (2 taxa) or a sequence differing from this at no more than three sites and from the uninverted sequence at no fewer than five sites (none).
Characters 7–12. Sequence in region of *ndhF* inversion 2, or for taxa interpreted as having the inversion (state 1 of character 6), the reverse complement of the observed sequence.
Character 13 (2, 0.50, 0.00): Insertion of 3 nucleotides between *ndhF* sites 1414 and 1415. 0 = 3 nucleotides absent; 1 = 3 nucleotides present.
Character 14 (2, 0.50, 0.00): Insertion of 3 nucleotides between *ndhF* sites 1576 and 1577. 0 = 3 nucleotides absent; 1 = 3 nucleotides present.
Character 15 (4, 0.25, 0.66): One vs. two copies of 15 nucleotides between *ndhF* sites 1693 and 1719 (location and extent of indel duplication is approximate, as multiple reasonable alignments are possible; cf. Fig. 4). 0 = 15 nucleotides absent, 1 = 15 nucleotides present.
Character 16 (5, 0.20, 0.75): Deletion of *ndhF* nucleotides 2209–2211. 0 = 3 nucleotides absent; 1 = 3 nucleotides present.
Character 17 (2, 0.50, 0.00): Insertion of 6 nucleotides between *ndhH* sites 19 and 20. 0 = 6 nucleotides absent; 1 = 6 nucleotides present.

other elements of the subfamily. The other is Ehrhartoideae, which has 59% support, with *Streptogyna* placed as the sister of the rest of the subfamily and with 100% support for the clade that includes all other elements of the subfamily. The third subfamily with weak support is Anomochlooideae (49%). The only subfamily represented by more than one element and not resolved as monophyletic is Arundinoideae. In this case, a monophyletic Micrairoideae (with 100% support) is nested within a clade that also includes the three representatives of Arundinoideae. Support for this overall grouping is 66%, and it is less than 50% for each of the internal nodes within the paraphyletic arrangement of the three genera of Arundinoideae.

Higher-level relationships are as follows: The grass family is monophyletic in all 12 trees (JS = 100%; Fig. 3). Within the family, Anomochlooideae is resolved as the sister of a clade that includes all other grasses (JS = 78% for the latter), and within the latter group, Pharoideae is sister of a clade that includes the remaining grasses (JS = 100% for this clade), and in which the PACMAD and BEP clades are both monophyletic and placed as sister taxa. JS for the PACMAD clade is 100%; within it, three major subclades are detected, with relationships unresolved among the three. The first of these clades includes Micrairoideae and the three representatives of Arundinoideae (as described above), the second includes Aristidoideae, Danthonioideae, and Chloridoideae, and the third is the Panicoideae (including elements previously assigned to Centothecoideae). Support is generally weak for relationships among subfamilies within the PACMAD clade, with the strongest support (66%) for the clade that includes Micrairoideae and the three representatives of Arundinoideae. Within the PACMAD clade, six of the seven tribes that were sampled by more than one taxon are monophyletic; the seventh, Arundineae, the solitary tribe in subfamily Arundinoideae, is not. JS for the BEP clade is 76%. Within this clade, Ehrhartoideae (including *Streptogyna*) is resolved as the sister of Bambusoideae and Pooideae, with 57% support for the clade that includes the latter two subfamilies. Nine tribes within the BEP clade are each sampled more than once, and all except two are resolved as monophyletic. The exceptions are Bambuseae (paraphyletic, with Olyreae nested within) and Bromeae (paraphyletic, with Triticeae nested within). Further description of relationships within Pooideae is deferred to a forthcoming analysis in which this group is sampled in greater depth.

Structural features of the plastid genome—Examples of six representative structures of the two SSC-IR boundaries are illustrated in Fig. 2, which includes fine-scale maps of the 5′ terminus of *ndhH* and the 3′ terminus of *ndhF*, and positions of the endpoints of these genes relative to the two SSC-IR boundaries. The range of observed structures include presence of a portion of *ndhF* (*Sporobolus*), *ndhH* (*Triticum*), both (*Olyra*, *Eragrostis*, *Brachypodium*, *Ehrharta*), or neither (not depicted in Fig. 2) within the the IR region (Tables 1, 2). In *Ehrharta*, both genes extend into the IR region (272 nucleotides of *ndhH* and 43 nucleotides of *ndhF*), and thus 43 nucleotides of the two genes are homologous between the two genes. Among the 82 taxa in which *ndhH* extends into the IR region, the length of the portion of this gene that lies in the IR region ranges from one nucleotide to 400 (*Cynosurus*).

When presence/absence of a portion of *ndhH* in the IR region is optimized on trees obtained from the cladistic analysis, presence of the 5′ terminus of *ndhH* within the IR region is determined to be a plesiomorphy of the grasses (Fig. 3, character 0). Within the grasses, there are five independent transitions to the state in which *ndhH* lies entirely within the SSC region, plus one reversion to the plesiomorphic state. Three of the five transitions

TABLE 2. Scores for characters described in Table 1. Numbers of nucleotides of *ndhH* and *ndhF* that extend into the IR region are indicated for taxa with state 1 for characters 0 and 1, respectively (cf. Table 1). Inversion presence/absence characters (characters 2 and 6) and nucleotide sequences within inverted regions (characters 3–5 and 7–12) are in boldface for taxa scored as having the inversions; the sequences provided here for the inverted regions of these taxa are the reverse complements of the sequences actually observed in the coding strand of *ndhF*. Taxa are arranged by family, subfamily, and tribe (cf. Appendix I).

	Character numbers						
Taxon	0	1	2	345	6	1 1 1 7 8 9 0 1 2	11111 34567
Joinvillea	1 192	0	0	GTA	0	GAAAAA	10010
Ecdeiocolea	1 179	0	0	GCA	0	GGAACA	01010
Anomochloa	1 179	0	0	GTA	0	CCAAAA	00010
Streptochaeta	1 178	0	0	GTA	0	AAAAAA	00010
Pharus	1 281	0	0	GTA	0	GAAAAA	00010
Amphipogon	0	1 15	0	GTA	0	CCAAAA	00110
Arundo	1 1	1 17	?	???	0	CAAAAA	00101
Molinia	0	1 12	0	CTA	0	AAAAAA	00110
Eriachne mucronata	0	1 19	0	GTA	0	CAAAAA	00101
Eriachne pulchella	1 1	1 21	0	GTA	0	CAAAAA	00100
Micraira	0	1 16	0	GTA	0	CAAAAA	00100
Stipagrostis	1 1	1 20	?	???	0	CAAAAA	00100
Danthonia	1 1	1 32	0	ATA	0	CAAAAA	00100
Merxmuellera macowanii	1 1	1 20	0	GTA	0	CAAAAA	00100
Merxmuellera rangei	1 1	1 32	0	GTA	0	CAAAAA	00100
Distichlis	1 1	1 20	0	GTA	0	CCAAAA	00110
Eragrostis	1 4	1 23	0	GTA	0	CAAAAA	00110
Uniola	1 4	1 23	0	GTA	0	CAAAAA	00110
Spartina	0	1 19	0	GTA	0	CAAAAA	00110
Sporobolus	0	1 19	?	???	0	CAAAAA	00110
Zoysia	0	1 19	?	???	0	CCAAAA	00110
Chasmanthium	1 1	1 20	0	GTA	0	TAAAAA	00100
Thysanolaena	1 4	1 23	?	???	0	CCAAAA	00100
Gynerium	0	1 14	0	GTA	0	AAAAAA	00100
Panicum	1 1	1 32	0	GTA	0	AAAAAA	00100
Pennisetum	1 1	1 32	0	GTA	0	AAAAAA	00100
Saccharum	1 1	1 32	0	GTA	0	AAAAAA	00100
Sorghum	1 1	1 32	0	GTA	0	AAAAAA	00100
Zea	1 1	1 32	0	GTA	0	AAAAAA	00100
Chusquea	1 186	0 0	0	GTA	0	GAAAAA	00110
Guadua	1 186	0 0	0	GTA	0	GAAAAA	00110
Phyllostachys	1 187	0 0	0	GTA	0	GAAAAA	00110
Pseudosasa	1 187	0 0	0	GTA	0	GAAAAA	00110
Buergersiochloa	1 187	0 0	0	GTA	0	GAAAAA	00110
Eremitis	1 252	0 0	0	CTA	0	AAAAAA	00110
Lithachne	1 151	0 0	0	GTA	0	AAAAAA	00110
Olyra	1 1	1 17	0	GTA	0	AAAAAA	00100
Pariana	1 252	0 0	0	CTA	0	AAAAAA	00110
Streptogyna	1 181	0 0	0	GTA	0	GAAAAA	00110
Ehrharta	1 272	1 43	0	GTA	0	GAAAAA	00110
Leersia	1 181	0	0	GTA	0	GAAAAG	00010
Oryza nivara	1 163	0	0	GTA	0	GAAAAA	00010
Oryza sativa	1 163	0	0	GTA	0	GAAAAA	00010
Brachyelytrum	1 223	0	0	GTA	0	GAAAAA	00110
Nardus	1 174	0	?	???	0	TAAAAA	00010
Lygeum	1 212	0	0	GGA	0	GAAAAA	00110
Anisopogon	1 177	0	0	GTT	0	GAAAAA	00110
Duthiea	1 181	0	0	GTA	0	GAAAAA	01110
Phaenosperma	1 181	0	0	GTA	0	AAAAAG	00110
Sinochasea	1 181	0	0	GTA	0	GAAAAA	00110
Achnatherum	1 181	0	0	GTA	0	GAAAAA	00110
Ampelodesmos	1 181	0	0	GTA	0	GAAAAA	00110
Celtica	0	0	0	GTA	0	AAAAAA	00010
Hesperostipa	1 182	0	0	GTA	0	GAAAAA	00110
Nassella pulchra	1 181	0	0	GTA	0	GAAAAA	00110
Nassella viridula	1 192	0	0	GTA	0	GAAAAA	00110
Oryzopsis	1 182	0	0	GTA	0	TAAAAA	00110
Piptatherum	1 174	0	0	GTA	0	AAAAAA	00110
Stipa	1 192	0	0	GTA	0	GAAAAA	00110
Timouria	1 172	0	0	GTA	0	GAAAAA	00110
Trikeraia	1 181	0	0	GTA	0	GAAAAA	00110
Brylkinia	1 149	0	0	GTA	0	GAAAAA	00110

TABLE 2. Continued.

Taxon	Character numbers						
	0	1	2	345	6	111 7890 12	11111 34567
Glyceria	1 168	0	0	GTA	0	GAAAAA	00110
Melica	1 208	0	0	GTA	0	GAAAAA	00110
Pleuropogon	0	0	0	GTA	0	GAAAAA	00110
Schizachne	1 175	0	0	GTA	0	GAAAAA	00110
Diarrhena	1 181	0	0	GTA	0	GAAAAA	00110
Brachypodium	1 42	1 1	**1**	**GTA**	0	CAAAAA	00110
Bromus inermis	1 314	0	**1**	**GTA**	0	TAAAAA	00110
Bromus korotkiji	1 314	0	**1**	**GTA**	0	TAAAAA	00110
Bromus suksdorfii	1 317	0	**1**	**GTA**	0	TAAAAA	00110
Littledalea	1 187	0	**1**	**GTA**	0	GAAAAA	00110
Elymus	1 207	0	**1**	**GTA**	0	TAAAAA	00110
Hordeum	1 216	0	0	GTA	0	TAAAAA	00110
Triticum	1 207	0	**1**	**GTA**	0	TAAAAA	00110
Agrostis	1 174	0	**1**	**GTA**	0	GAAAAA	00110
Avena	1 181	0	**1**	**GTA**	**1**	**GAAAAA**	00110
Trisetum	1 181	0	**1**	**GTA**	**1**	**GAAAAA**	00110
Briza	1 181	0	**1**	**ATA**	0	GAAAAA	00110
Torreyochloa	1 181	0	**1**	**GTA**	0	GAAAAA	00110
Aira	1 181	0	**1**	**GTA**	0	GAAAAA	00110
Deschampsia	1 206	0	**1**	**GTA**	0	GAAAAA	00110
Molineriella	1 26	0	**1**	**GTA**	0	GAAAAA	00110
Phleum	1 175	0	**1**	**GTA**	0	GAAAAA	00110
Cynosurus	1 400	0	**1**	**GTA**	0	GAAAAA	10110
Dactylis	1 169	0	**1**	**GTA**	0	GAAAAA	00110
Holcus	1 157	0	**1**	**GTA**	0	GAAAAA	00110
Festuca	1 169	0	**1**	**GTA**	0	GAAAAA	00110
Parapholis	1 206	0	**1**	**GTA**	0	GAAAAA	00110
Poa	1 192	0	**1**	**GTA**	0	GAAAAA	00110
Puccinellia	1 170	0	**1**	**GTA**	0	GAAAAA	00110
Sclerochloa	1 156	0	**1**	**GTA**	0	GAAAAA	00110

in which the 5′ terminus of *ndhH* migrates out of the IR region occur in the PACMAD clade, each of them occurring once within each of the three major clades resolved in the consensus tree. Within one of these groups, the Arunidinoideae/Micrairoideae clade, the 5′ terminus of *ndhH* migrates out of the IR region, and then back into it, in *Eriachne pulchella*, which therefore differs in this feature from its sister taxon, *Eriachne mucronata*. The other two transitions to the state in which *ndhH* lies entirely within the SSC region are in *Celtica* and *Pleuropogon*, both of which are in the Pooideae.

Although there are only six transitions in the presence/absence of a portion of *ndhH* within the IR region, the size of the portion of *ndhH* that lies within the IR region varies widely among taxa (Table 2, Fig. 3). In the outgroups and in the first lineage to diverge within the grasses from the clade that includes all others (Anomochlooideae), 179–192 nucleotides of *ndhH* lie within the IR region. The number for *Ecdeiocolea* (179) is the same as that for *Anomochloa*. In the next-diverging grass lineage, *Pharus*, about 100 additional nucleotides lie within the IR region. Only 1–4 nucleotides of *ndhH* lie within the IR region in taxa of the PACMAD clade, and as already noted, these nucleotides migrate into the SSC region three times within this group. Within the BEP clade, numbers initially are similar to those of the early-diverging grasses, with occasional cases of increase (e.g., to 272 in *Ehrharta*, to more than 300 in the various species of *Bromus*, and to 400 in *Cynosurus*) and decrease (e.g., to 149 in *Brylkinia*, 42 in *Brachypodium*). In the two instances in the BEP clade in which *ndhH* migrates entirely out of the IR region (*Celtica* and *Pleuropogon*), the sister of the taxon in which this has occurred does not have an unusually

TABLE 3. Characteristics of partitions of the data matrix analyzed in this study. Aligned length of *ndhF* excludes regions in which alignment is ambiguous and includes nine nucleotide sites from two inversion regions, with sequences from inverted taxa reinverted for analysis (see Materials and Methods). The three gene partitions include nucleotides only, not structural features of the genes.

Character partition	Aligned length, nucleotides	No. of parsimony-informative characters (% of aligned length)	Trees from combined matrix		
			No. of steps	RI	CI
ndhH nucleotides	1194	279 (23.4%)	1282–1284	0.32	0.66
rps15 nucleotides	122	26 (21.3%)	49–50	0.58–0.59	0.90
ndhF nucleotides	2122	602 (28.4%)	2662–2665	0.35–0.36	0.70
Total nucleotides, three genes	3438	907 (26.4%)	3996	0.35	0.69
Structural characters	n/a	9 (n/a)	28	0.32	0.78
Total, all characters	n/a	916 (n/a)	4024	0.35	0.69

low number of sites in the IR region (174 in *Piptatherum*, 168 in *Glyceria*).

Among the 27 taxa in which a portion of *ndhF* lies within the IR region, the length of this portion is one nucleotide in *Brachypodium* (Table 2) and between 12 and 32 nucleotides in all other taxa except *Ehrharta* (Fig. 2, Table 2), in which 43 nucleotides of *ndhF* lie within the IR region. These nucleotides are the reverse complement of a region in *ndhH* (on the other strand) that is present in all sampled taxa and that also extends into the IR region in *Ehrharta*. The alignment of *ndhF* sequences, with the stop codons of *Ehrharta*, *Triticum*, and other taxa (usually TAA) treated as homologous, suggests that *Ehrharta* has a 36-nucleotide insertion relative to *Triticum*, just prior to the stop codon. However, the homology of the 43-nucleotide portion of *ndhF* of *Ehrharta* (comprising the stop codon, the preceding 36 nucleotides, and the four nucleotides that precede them) with a region in *ndhH* suggests that seven nucleotides at the 3′ end of *ndhF* actually were lost, and that the apparent insertion of 36 nucleotides of *ndhF* in *Ehrharta* represents an extension of the 3′ end of the coding region from the SSC-IR boundary to the first available stop codon inside the IR region.

When the presence/absence of a portion of *ndhF* in the IR region is optimized on trees obtained from the cladistic analysis, the absence of any portion of *ndhF* in the IR region is determined to be a plesiomorphy of the grasses (Fig. 3, character 1). Within the grasses, migration of the 3′ terminus of *ndhF* into the IR region is interpreted as a synapomorphy of the PACMAD clade, and no reversals are observed. Within the PACMAD clade, the length of the portion of *ndhF* that lies in the IR region ranges from 12 to 32 nucleotides. Three additional migrations of the 3′ terminus of *ndhF* into the IR region are inferred, in *Ehrharta*, *Olyra*, and *Brachypodium*. The lengths of the portions of *ndhF* that lie within the IR region in these taxa range from 1 to 43 nucleotides, and as noted, the 43 nucleotides of *ndhF* in *Ehrharta* include seven that are homologous with those in *Triticum*, plus an autapomorphic insertion of 36.

Two structural characters in the analysis are inversions within *ndhF*, one of them three nucleotides in length, the other six in length (Table 1, characters 2 and 6, respectively). Both inversion sites are surrounded by short inverted repeat sequences, indicative of a hairpin structure (Kelchner and Wendel, 1996). The inverted state of the three-nucleotide inversion arises once in all trees, in the sister group of *Diarrhena*, with a reversion to the plesiomorphic state occurring in *Hordeum* (Fig. 3, character 2). The six-nucleotide inversion arises once, as a synapomorphy of the clade that includes *Avena* and *Trisetum* (Fig. 3, character 6), and no reversions are observed.

The five remaining parsimony-informative structural characters are indels. Three of these (characters 13, 14, and 17; Tables 1 and 2; not mapped in Fig. 3) involve distantly related taxa (e.g., *Joinvillea* and *Cynosurus* for character 13) and are interpreted as parallelisms. The three-nucleotide indel near the 3′ terminus of *ndhF*, located within the portion of this gene that lies within the IR region in some taxa, was encoded as character 16 (Tables 1 and 2). As with the 36-nucleotide insertion in *Ehrharta*, variation in character 16 affects the length of the portion of *ndhF* that extends into the IR region (Fig. 3, numbers in parentheses). There are five steps in this character in all most-parsimonious trees and in the consensus tree (Fig. 3, character 16). The undeleted state is interpreted as plesiomorphic for the grasses. The deleted state arises independently in *Olyra* and at the origin of the PACMAD clade. Within the PACMAD clade there is a reversion to the undeleted state in Chloridoideae. Two additional steps occur within the Arundinoideae/Micrairoideae clade, either a reversion to the undeleted state followed by a secondary reversion to the deleted state (as mapped in Fig. 3, under accelerated optimization), or parallel reversions to the undeleted state in *Molinia* and *Amphipogon* (under delayed optimization).

A 15-nucleotide indel (character 15) has two forms, one of which appears to represent a tandem duplication of 15 nucleotides (Figs. 3, 4). The unduplicated state occurs in 10 taxa (Tables 1, 2), and one potential alignment of this region is depicted in Fig. 4 for these 10 taxa and several representative taxa with the putatively duplicated state. This region is one of the four regions of ambiguous alignment within *ndhF* that were excluded from the cladistic analysis on the basis of ambiguity of alignment. The 15-nucleotide segment that is apparently duplicated in some taxa can be recognized as consisting of two portions, the first nine nucleotides of which is a relatively constant motif (GATAATGGA and slight variants) with a second more divergent six-nucleotide motif (in taxa with two copies there appear to be two variants of this, one based on ATAATG and variants, the other on ATAGCG and variants). The six nucleotides that follow the nine-nucleotide motif in five of the taxa that have one copy of the 15 nucleotide region (*Joinvillea*, *Ecdeiocolea*, *Anomochloa*, *Streptochaeta*, and *Pharus*) resemble the version of the two six-nucleotide motifs closest to the 5′ end of the gene in duplicated taxa. In the other five taxa with single copies (*Leersia*, two species of *Oryza*, *Nardus*, and *Celtica*), the six nucleotides that follow the nine-nucleotide motif resemble the second of the two six-nucleotide motifs. Thus, it is possible to recognize two different forms of the unduplicated state without reference to a phylogeny. Because of the general ambiguity of alignment in this region, only two states were recognized for this character (unduplicated and duplicated), and all 10 taxa with a single 15-nucleotide copy in this region were scored identically. Optimization of this character on the 12 most-parsimonious cladograms (and consensus tree) implies four steps (Fig. 3, character 15, i.e., one duplication and three subsequent deletions in disjunct parts of the tree). The single-copy state is shared by the two outgroups and three early-diverging grasses and is interpreted as a plesiomorphy of the grasses. The five taxa that share this state correspond to one of the two groups that can be recognized a priori (Fig. 4). The tandem duplication state originates as a synapomorphy of the clade that includes the PACMAD and BEP clades (the leftmost instance of character 15 noted in Fig. 3), and the single-copy state then reoriginates independently as a deletion of one of the duplicated 15 nucleotide regions (i.e., the more 3′ copy) in three groups within this large clade (Oryzeae, *Nardus*, and *Celtica*). If the two groups of taxa with deletions alignable in different positions had been scored as having different states and the same trees were obtained, then the unduplicated state that is plesiomorphic in the grasses would still have been lost once because of duplication, and the deleted state that later arose within the grasses would have had three separate origins.

Nucleotide substitution rates—In the 894-nucleotide portion of *ndhH* that lies within the SSC region in nearly all sampled taxa, there are 1174 steps in the randomly selected tree that was used to compare nucleotide substitution rates or 1.31 steps per character (Table 4). For this region, the number of steps outside the PACMAD clade (0.88 steps per character) is about three times the number within the PACMAD clade (0.32). In

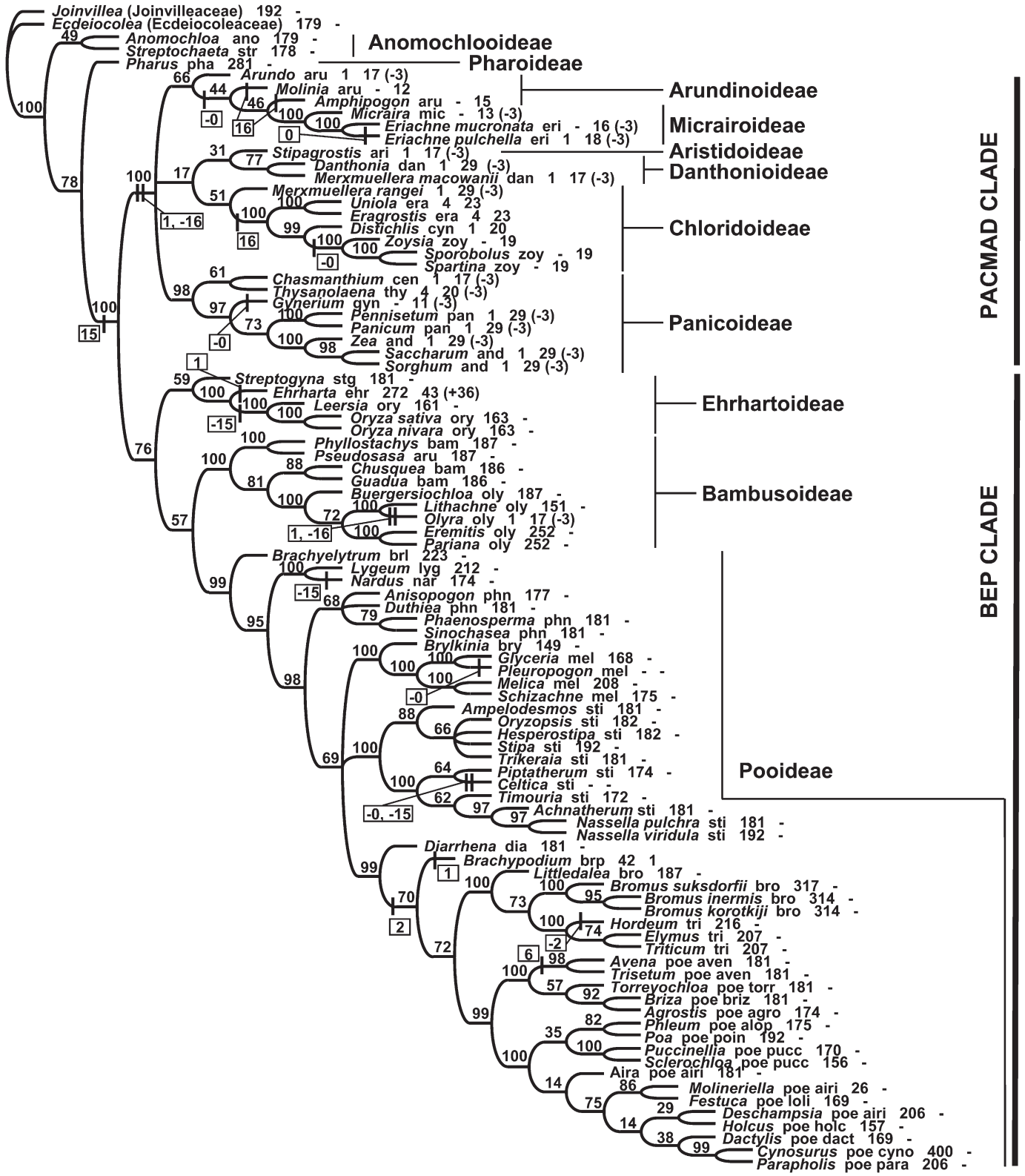

Fig. 3. Strict consensus of 12 most-parsimonious trees for 90 grass species and two nongrass outgroups, as resolved by combined analysis of three gene sequences and nine structural characters of the plastid genome (see text). Family names are indicated for the two outgroup taxa, and assignments of grasses to tribe and (for tribe Poeae) subtribe are signified by letter codes as specified in Appendix 1, where species names are provided. Within the grass family, two major clades (PACMAD and BEP) and 11 subfamilies are indicated by lines and labels at right. Two nonparenthesized numbers beside each taxon name signify the number of nucleotides of *ndhH* and *ndhF*, respectively, that extend into the IR region of the plastid genome (see text); hyphens denote 0 nucleotides. Parenthesized numbers denote the presence of an insertion (+) or deletion (−), relative to the reference sequence of *Triticum aestivum*,

the 2105-nucleotide portion of *ndhF* that lies within the SSC region in all taxa, there are 2944 steps in the tree, or 1.40 steps per character; as with the portion of *ndhH* that lies predominantly within the SSC region, the number of steps in this region outside the PACMAD clade is about three times the number within this clade (1.03 vs. 0.35 steps per character, respectively). In the sequenced portion of *rps15* (122 nucleotides in length), which lies within the IR region in all sampled taxa, there are 66 steps in the tree, or 0.54 steps per character, and the number of steps outside the PACMAD clade is about five times the number within (0.44 vs. 0.09 steps per character, respectively). Finally, in the 150-nucleotide region of *ndhH* that lies entirely or almost entirely within the IR region in most non-PACMAD taxa in the sample, and in the SSC region in all members of the PACMAD clade, there are 96 steps in the tree that was examined, an average of 0.64 steps per character, with about one-half of these steps outside the PACMAD clade, and the other half within the PACMAD clade (0.31 vs. 0.33 steps per character, respectively).

DISCUSSION

Structural features of the plastid genome—Positions of the endpoints of *ndhF* and *ndhH*, relative to the SSC-IR junction, previously inferred for a few species of the grass family from complete plastid genome sequences (e.g., Maier et al., 1990; Ogihara et al., 2002; Saski et al., 2007; Bortiri et al., 2008), were newly determined here for 84 grass species and two outgroups by targeted sequencing across the two SSC-IR junctions. Although characters representing the presence/absence of portions of *ndhF* and *ndhH* of any length within the IR region of the plastid genome were among the structural characters included in the cladistic analysis (characters 0 and 1, respectively [Tables 1, 2; Fig. 3]), lengths of the portions of these genes extending into the IR region were quite variable and were not treated as formal characters. However, general trends in the lengths of these portions are evident in the resulting phylogeny, as follows: It appears that in the earliest grasses, as in their closest relatives, *Joinvillea* and *Ecdeiocolea*, ca. 175–200 nucleotides at the 5′ end of *ndhH* extended into the IR region, while *ndhF* was confined to the SSC region; sampling of additional outgroups might alter this interpretation. These general features are retained in Anomochlooideae and Pharoideae, with the number of nucleotides of *ndhH* extending into the IR region increasing to nearly 300 in the latter. Following the divergence of the BEP and PACMAD clades from each other, these general structural features were retained within the BEP clade, and they occur in most sampled taxa of the three subfamilies in this clade, including, e.g., *Streptogyna* and *Oryza* (Ehrhartoideae), *Phyllostachys*, *Buergersiochloa*, and *Pariana* (Bambusoideae), and *Brachyelytrum*, *Lygeum*, *Stipa*, *Glyceria*, *Diarrhena*, *Triticum*, and *Poa* (Pooideae; Fig. 3). As in *Pharus*, the number of nucleotides of *ndhH* extending into the IR in taxa of the BEP clade occasionally increases substantially, as in *Ehrharta*, *Cynosurus*, and some elements of Olyreae and Bromeae. In other taxa (*Olyra* and *Brachypodium*), the number is substantially diminished, and it drops to zero in *Pleuropogon* and *Celtica*, in which *ndhH* lies entirely within the SSC. In most taxa of the BEP clade, *ndhF* remains entirely within the SSC, but portions of this gene, of lengths ranging from 1 to 43 nucleotides, lie within the IR region in three isolated taxa (*Ehrharta*, *Olyra*, and *Brachypodium*). Thus, the positions of the termini of *ndhH* and *ndhF* relative to the SSC-IR junctions, and the sizes of the portions of these genes that lie within the IR, vary widely within the BEP clade and bespeak a history of multiple parallel migration events.

In contrast to the BEP clade, the PACMAD clade, as a group, is marked by substantial changes in the positions of both *ndhH* and *ndhF* relative to the SSC-IR junctions (Table 2, Fig. 3), and following the establishment of these differences, additional changes have continued to occur in various lineages, as in the BEP clade. Most, but not all, of the portion of *ndhH* that had been situated within the IR in the earliest grasses migrated out of it early in the evolution of the PACMAD clade, leaving just one to four nucleotides remaining within the IR, while the 3′ terminus of *ndhF* migrated into the IR, resulting in the presence of ca. 12–30 nucleotides within this region in most taxa. Following the initial diversification of the PACMAD clade, the last few nucleotides of *ndhH* migrated out of the IR region in at least three different lineages, and in one case (*Eriachne pulchella*), one nucleotide later migrated back into the IR region. Also, a 3-nucleotide deletion event near the 3′ terminus of *ndhF* (now lying within the IR region) appears to have occurred prior to the diversification of major lineages within the PACMAD clade, and at least two subsequent reinsertions of three nucleotides appear to have occurred in this region and possibly a secondary deletion event as well (or three reinsertions and no secondary deletion, under an alternative optimization of the character).

Although most taxa of the BEP clade differ from most taxa of the PACMAD clade in these features, *Olyra* (of the BEP clade) is unusual in having a genomic structure that is typical of the PACMAD clade. In *Olyra*, 17 nucleotides at the 3′ terminus of *ndhF* lie within the IR region, one nucleotide at the 5′ terminus of *ndhH* lies within the IR (and thus is homologous with the 17th from the final nucleotide of *ndhF*), and a three-nucleotide deletion is present at the same location as in taxa of the PACMAD clade. All of these features vary within the PACMAD clade, and among taxa within this group, *Olyra* is identical to *Arundo* in all three characters (cf. Table 2, Fig. 3). In light of these similarities, the possibility was considered that a laboratory or clerical error might have led to the erroneous labeling of sequences from a representative of the PACMAD clade as having been collected from *Olyra*. Alternatively, the possibility was considered that the actual plastid sequence that occurs in *Olyra* might have been derived from a species of the PACMAD clade via horizontal gene transfer. A third possibility was that one or both of the sequences determined for *Olyra* might be a laboratory

←
in the portion of *ndhF* that extends into the IR (e.g., for *Arundo*, one nucleotide of *ndhH* and 17 nucleotides of *ndhF* extend into the IR region, and *Arundo* has a 3-bp deletion, relative to *Triticum*, in the portion of *ndhF* that extends into the IR region, so the 17 nucleotides of *ndhF* that extend into the IR, as aligned, correspond to a 20-nucleotide fragment in *Triticum*). Numbers above branches are jackknife support frequencies. State transformations are indicated for six structural characters (character numbers in rectangles; Tables 1, 2); transformations from state 1 to state 0 are signified by minus signs, and all others are from state 0 to state 1. Each structural character is an unambiguous synapomorphy of the same clade on all most-parsimonious trees and has the same number of steps and direction of transformation as indicated on the consensus tree, except that an alternative optimization exists for character 16 in all trees, within Arundinoideae, as a gain to the sister group of *Arundo* and a loss to the sister group of *Amphipogon*.

Fig. 4. Aligned sequences of a region of *ndhF* characterized by a 15-nucleotide indel, for 22 species of Poaceae, Joinvilleaceae, and Ecdeiocoleaceae (Appendix 1). Location is specified by nucleotide sites in the reference *ndhF* sequence of *Triticum aestivum* (GenBank accession NC_002762). A conserved nine-nucleotide motif (GATAATGGA and variant forms of this sequence), depicted in boldface, occurs twice in duplicated sequences (a duplication is required according to the interpretation in Fig. 3), with the initial nucleotides of the two copies separated by 15 nucleotides, and only once in unduplicated sequences.

chimera that combines a portion of the actual sequence from *Olyra* with a portion of a sequence from an element of the PACMAD clade.

To test for these possibilities, separate analyses of just the *ndhF* and *ndhH* sequences were conducted. The placement of *Olyra* was determined in each of these trees, as was the length of the terminal branch for *Olyra*. The significance of the terminal branch length lies in the possibility that an analysis might place a chimeric sequence within Olyreae, on the basis of nucleotides actually from *Olyra*, while the branch leading to *Olyra* might be inordinately long if another portion of the sequence was from a species in the PACMAD clade, since that portion of the sequence would exhibit autapomorphic origins of characters of that species and its relatives. The *ndhF* analysis included a second *Olyra* sequence obtained from GenBank (accession AM849172), in addition to those used in the combined analysis. This sequence is truncated ca. 50 nucleotides from the 3′ terminus of *ndhF*, so the state of indel character 16 cannot be determined. Also, there is no corresponding sequence for *ndhH* (or flanking regions of either gene), so the positions of the SSC-IR boundaries relative to the endpoints of these two genes also could not be determined for this sequence.

Analysis of the *ndhF* matrix yielded a consensus tree in which the two *Olyra* sequences were situated in a clade with *Lithachne*. This group was placed within a larger clade that included the other three sequences of Olyreae (as in the principal analysis, Fig. 3), and the terminal branch lengths of both sequences of *Olyra* were shorter than the average terminal branch lengths of the six sequences in this group. As in the principal analysis, the Olyreae was placed in a monophyletic Bambusoideae,

TABLE 4. Nucleotide substitution rates in four regions of the plastid genome, within and outside of the PACMAD clade. The first two gene regions in the table are in the SSC region of the plastid genome in all or nearly all sampled taxa; the third gene region is in the IR region of the plastid genome in all sampled taxa; the fourth gene region is in the SSC region in all taxa of the PACMAD clade and in the IR region in most other taxa.

Gene region; no. of aligned sites	Total steps (total in branch to PACMAD clade); ratio, no. of steps to no. of sites	No. of steps within PACMAD clade; ratio, no. of steps to no. of sites	No. of steps outside PACMAD clade; ratio, no. of steps to no. of sites	Ratio, no. of steps outside PACMAD clade to no. of. steps within it
ndhH site 301 to 3′ terminus; 894	1174 (100); 1.31	287; 0.32	787; 0.88	2.74
ndhF, in part (see text); 2105	2944 (24); 1.40	747; 0.35	2173; 1.03	2.91
rps15, sequenced portion; 122	66 (1); 0.54	11; 0.09	54; 0.44	4.91
ndhH sites 5–54; 150	96 (0); 0.64	49; 0.33	47; 0.31	0.96

and the Bambusoideae was placed as sister of Pooideae in a monophyletic BEP clade. Analysis of the 92 *ndhH* sequences yielded a consensus tree in which *Olyra* was the sister of *Lithachne* within a monophyletic Bambusoideae, and the terminal branch for *Olyra* was not unusually long. Hence, the available evidence is consistent with a conclusion that the *ndhF* and *ndhH* sequences obtained from *Olyra* for this study are not chimeras and that the phylogenetic affinities of each gene are with those of other taxa of the Olyreae and other elements of Bambusoideae, not with those of the PACMAD clade. Consequently, the structural similarities between the plastid genome of *Olyra* and those of *Arundo* and other taxa of the PACMAD clade are interpreted as having arisen independently.

The various migrations of gene termini relative to the two SSC-IR junctions are also of significance in terms of the phenomenon of gene overlap. When both *ndhF* and *ndhH* extend into the IR, one or more nucleotides in the IR simultaneously encode portions of both genes (Figs. 1, 2). As noted, the extent of this overlap is limited by the size of the smallest gene segment that extends into the IR. Within the present taxon sample, gene overlap arises at the point of origin of the PACMAD clade, where the 3′ terminus of *ndhF* enters the IR, which already includes the 5′ terminus of *ndhH*. At about this point in the history of the clade, most of the portion of *ndhH* previously included in the IR migrates into the SSC, leaving 1–4 nucleotides overlapping with *ndhF*. The overlap is lost multiple times within the clade, whenever the last few nucleotides of *ndhH* migrate into the SSC, and it is regained once in the PACMAD clade, in *Eriachne pulchella*.

Elsewhere in the taxon sample, there are three additional origins of gene overlap, in *Ehrharta*, *Olyra*, and *Brachypodium*. *Olyra* differs from its closest relatives in much the same way that taxa of the PACMAD clade differ from the early-diverging grass lineages and most of the BEP clade. As in the PACMAD clade, almost the entire portion of *ndhH* that once was situated in the IR has migrated into the SSC, and a portion of *ndhF*, less than 30 nucleotides in length, has migrated into the IR. As with the PACMAD clade, the occurrence of these events on the same branch of a cladogram does not indicate whether they occurred simultaneously or in succession, and in the latter case, which event may have occurred first. Thus, the co-occurrence of this pair of autapomorphic features in *Olyra* suggests either that a single mutational event mediated the origin of both features or that it is maladaptive for more than a few nucleotides of these two genes to overlap in this manner, possibly because it constrains the evolution of each. Under the latter interpretation, degrees of overlap of more than a few nucleotides, when they do arise, are soon eliminated by the migration of all or nearly all of one gene or the other out of the IR.

In *Brachypodium*, a similar pattern of overlap exists, but in this case it is *ndhF* that extends the shortest distance (one nucleotide) into the IR. The structure in *B. pinnatum*, as sampled here, differs substantially from those of its closest relatives in the Pooideae, and also from that of its close relative, *B. distachyon* (GenBank NC_011032; Bortiri et al., 2008), which also resembles other taxa of Pooideae, in having 209 nucleotides of *ndhH* in the IR, and *ndhF* confined to the SSC. Thus, both *B. distachyon* and *Olyra latifolia* exhibit substantially modified positions of the endpoints of *ndhH* and *ndhF*, with both genes extending into the IR, but with only one nucleotide of overlap. These patterns, along with the limited degree of overlap observed in taxa of the PACMAD clade, suggest that however an overlap may arise, it is maladaptive if too extensive and soon is lost as one gene or the other migrates back into the SSC.

The situation in *Ehrharta*, however, represents an exception to this pattern, with 43 nucleotides of the IR encoding portions of two different genes, on opposite DNA strands (and in opposite directions), including the termination codon of *ndhF*. Of these 43 nucleotides, 36 are interpreted as an insertion, but the entire 43-nucleotide region is homologous with the corresponding DNA strand in *ndhH* (Fig. 2), and no insertion is evident in that gene. In other words, the "insertion" in *ndhF* arose not by the expansion of a DNA region, but by the loss of a stop codon in *ndhF* (either through a point mutation or a small rearrangement that did not affect the length of the gene), resulting in the lengthening of the coding region to a point previously downstream from the 3′ terminus, where another stop codon was encountered.

Distributions of the three- and six-nucleotide inversions in *ndhF* (characters 2 and 6, respectively [Tables 1 and 2; Fig. 3]), as determined by the present analysis, correspond to those described previously (Davis and Soreng, 2007; Soreng et al., 2007). The three-nucleotide inversion is a synapomorphy of the clade within Pooideae that is the sister of *Diarrhena* (consisting of Bromeae, Triticeae, Poeae, and other small tribes), and there is a single reversion to the uninverted state, in *Hordeum*. Within the clade that has the three-nucleotide inversion, the six-nucleotide inversion is a synapomorphy of the two representatives of Poeae subtribe Aveninae.

The 15-nucleotide indel in *ndhF* (character 15 [Tables 1, 2, Fig. 3]) was determined to be homoplasious when first reported by Clark et al. (1995), who observed that the deleted state occurs in close relatives of the grasses and early-diverging grass lineages, as well as in Oryzeae. Thus, the authors interpreted the deleted state as plesiomorphic in the grasses, with an insertion marking the clade that includes taxa now conventionally placed in the PACMAD and BEP clades, and with a reversion to the deleted state marking the Oryzeae. As suggested by the alignment presented here (Fig. 4), it is reasonable to recognize differences on an a priori basis between two forms of the deleted state. Under this determination, the deleted state in taxa of the BEP clade could be scored as a different character or state than for taxa outside the BEP clade. However, results of the present phylogenetic analysis, consistent with those supported by other analyses, still would suggest that the state that occurs in the BEP clade has arisen three times in parallel within this group (Fig. 3). These repeated insertion/deletion events may be attributable to slipped-strand mispairing, as has been inferred in similar instances (e.g., Levinson and Gutman, 1987; Cummings et al., 1994; Kelchner, 2000; Dertien and Duvall, 2009).

The remaining structural characters in the present analysis are three additional indels (characters 13, 14, and 17), each of which differentiates two taxa from disparate regions of the tree from the rest of the sample and thus exhibits two steps and an RI of 0 in the analysis (Tables 1, 2; Fig. 3). These results, like those for the positions of endpoints of *ndhF* and *ndhH* relative to the SSC-IR junctions, as well as for the three-nucleotide inversion in *ndhF* and the 15-nucleotide indel in *ndhF*, tend to confirm the propensity of structural mutations to arise independently, in particular locations of the genome, often in identical positions, and to exhibit subsequent reversals to their plesiomorphic states (e.g., Kelchner and Wendel, 1996; Graham et al., 2000; Tsumura et al., 2000; Kim and Lee, 2005; Bain and Jansen, 2006). Thus, although structural characters of this sort often provide useful phylogenetic evidence (e.g., Luo et al.,

2006), they also are often individually homoplasious, like nucleotide sequence characters and even morphological characters, and are best interpreted in the context of a phylogeny based on a wide range of characters.

Nucleotide substitution rates—Comparison of the relative rates of nucleotide substitution within and outside the PACMAD clade, as determined for four gene regions, indicates that the two regions that lie within the SSC region in all or nearly all taxa, one of them a portion of *ndhF*, the other a portion of *ndhH*, evolve at similar rates (averages of 1.40 and 1.31 steps per nucleotide, respectively, across the same set of taxa; Table 4). Also, in both of these cases, the number of steps within the PACMAD clade is about one-third the number occurring elsewhere in the tree. The number of taxa in the PACMAD clade also is about one-third that of the number of taxa elsewhere in the tree, but the latter group is also a paraphyletic assemblage, and includes deeper branches in the tree, so these numbers do not provide absolute measures of evolutionary rates, but they do allow for comparisons of relative rates among gene regions across the same portions of the tree.

The portion of *rps15* that was examined lies entirely within the IR region in all taxa and exhibits about one-third the total number of steps per character as the two gene regions that lie in the SSC region of the genome. This general pattern is consistent with previous observations that substitution rates in the IR regions of the plastid genome are substantially lower than those in the single-copy regions (e.g., Wolfe et al., 1987; Maier et al., 1995; Muse and Gaut, 1997; Yamane et al., 2006). The observed number of steps in this portion of *rps15* within the PACMAD clade is only one-fifth that of the number outside this clade, while the corresponding ratio is about one-third for the two gene regions that lie in the SSC region of the genome. However, this region of *rps15* is only 122 nucleotides in length, so the relatively low observed rate for the PACMAD clade, relative to the rest of the tree, may not be an accurate indication of general substitution rates for genes in the IR region, or even for this gene. Leaving aside this matter of precision, the observation of substantially fewer steps within the PACMAD clade than outside of it is generally consistent with the patterns observed for the gene regions that lie within the SSC region of the genome.

A different pattern is observed for the portion of *ndhH* that lies in the SSC region in taxa of the PACMAD clade and in the IR region in taxa outside the PACMAD clade. Like the portion of *rps15* that was examined, this gene region is relatively small (150 nucleotides), so the observed rates may not be precise indicators of actual substitution rates. The number of steps in this gene region among taxa of the PACMAD clade is about equal to the number of steps among taxa outside the PACMAD clade. With about 0.33 steps per nucleotide site within the PACMAD clade, the substitution rate for this portion of *ndhH* is comparable to the rates observed within the PACMAD clade for the two other gene regions that lie in the SSC region of the genome (0.32 and 0.35). Conversely, with about 0.31 steps per nucleotide outside the PACMAD clade, the substitution rate for this portion of *ndhH* is comparable to the rate observed outside the PACMAD clade for the other gene region that lies in the IR region of the genome (0.44). Thus, relative substitution rates of the various gene regions correspond to a general pattern in which those that lie within the IR region of the genome evolve more slowly than those that lie within the SSC region of the genome. Muse and Gaut (1997) noted that the migration of a gene from the IR into a single-copy region was likely to lead to acceleration of the substitution rate in that gene. As demonstrated here, migration of even a small portion of a gene from the IR into the SSC can be associated with an increase in nucleotide substitution rate in the portion of the gene that migrated.

LITERATURE CITED

AII, J., Y. KISHIMA, T. MIKAMI, AND T. ADACHI. 1997. Expansion of the IR in the chloroplast genomes of buckwheat species is due to incorporation of an SSC sequence that could be mediated by an inversion. *Current Genetics* 31: 276–279.

BAIN, J. F., AND R. K. JANSEN. 2006. A chloroplast DNA hairpin structure provides useful phylogenetic data within tribe Senecioneae (Asteraceae). *Canadian Journal of Botany* 84: 862–868.

BARKER, N. P. 1997. The relationships of *Amphipogon, Elytrophorus* and *Cyperochloa* (Poaceae) as suggested by *rbcL* sequence data. *Telopea* 7: 205–213.

BARKER, N. P., C. GALLEY, G. A. VERBOOM, P. MAFA, M. GILBERT, AND H. P. LINDER. 2007. The phylogeny of the austral grass subfamily Danthonioideae: Evidence from multiple data sets. *Plant Systematics and Evolution* 264: 135–156.

BARKER, N. P., H. P. LINDER, AND E. H. HARLEY. 1999. Sequences of the grass-specific insert in the chloroplast *rpoC2* gene elucidate generic relationships of the Arundinoideae (Poaceae). *Systematic Botany* 23: 327–350.

BENDICH, A. J. 2004. Circular chloroplast chromosomes: The grand illusion. *Plant Cell* 16: 1661–1666.

BOCK, R. 2007. Structure, function, and inheritance of plastid genomes. *In* R. Bock [ed.], Topics in current genetics ,vol. 19, Cell and molecular biology of plastids, 29–63. Springer-Verlag, New York, New York, USA.

BORTIRI, E., D. COLEMAN-DERR, G. R. LAZO, O. D. ANDERSON, AND Y. Q. GU. 2008. The complete chloroplast genome sequence of *Brachypodium distachyon*: Sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Research Notes* 1: 61.

BOUCHENAK-KHELLADI, Y., N. SALAMIN, V. SAVOLAINEN, F. FOREST, M. VAN DER BANK, M. W. CHASE, AND T. R. HODKINSON. 2008. Large multigene phylogenetic trees of the grasses (Poaceae): Progress towards complete tribal and generic level sampling. *Molecular Phylogenetics and Evolution* 47: 488–505.

BREMER, K. 2002. Gondwanan evolution of the grass alliance of families (Poales). *Evolution* 56: 1374–1387.

BRIGGS, B. G., A. D. MARCHANT, S. GILMORE, AND C. L. PORTER. 2000. A molecular phylogeny of Restionaceae and allies. *In* K. L. Wilson and D. A. Morrison [eds.], Monocots: Systematics and evolution, 661–671. CSIRO, Collingwood, Australia.

CHASE, M. W., M. F. FAY, D. S. DEVEY, O. MAURIN, N. RØNSTED, T. J. DAVIES, Y. PILLON, ET AL. 2006. Multigene analyses of monocot relationships: A summary. *Aliso* 22: 63–75.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD, D. MORGAN, D. H. LES, B. D. MISHLER, M. R. DUVALL, ET AL. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL. Annals of the Missouri Botanical Garden* 80: 528–580.

CHASE, M. W., D. W. STEVENSON, P. WILKIN, AND P. J. RUDALL. 1995. Monocot systematics: A combined analysis. *In* P. J. Rudall, P. J. Cribb, D. F. Cutler, and C. J. Humphries [eds.], Monocotyledons: Systematics and evolution, 685–730. Royal Botanic Gardens, Kew, UK.

CLARK, L. G., W. ZHANG, AND J. F. WENDEL. 1995. A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Systematic Botany* 20: 436–460.

CLAYTON, W. D., AND S. A. RENVOIZE. 1986. Genera graminum: Grasses of the world. Her Majesty's Stationery Office, London, UK.

COSNER, M. E., L. A. RAUBESON, AND R. K. JANSEN. 2004. Chloroplast DNA rearrangements in Campanulaceae: Phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biology* 4: 27.

CUMMINGS, M. P., L. M. KING, AND E. A. KELLOGG. 1994. Slipped-strand mispairing in a plastid gene: *rpoC2* in grasses (Poaceae). *Molecular Biology and Evolution* 11: 1–8.

DAVIS, J. I., AND R. J. SORENG. 1993. Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. *American Journal of Botany* 80: 1444–1454.

DAVIS, J. I., AND R. J. SORENG. 2007. A phylogenetic analysis of the grasses (Poaceae), with attention to subfamily Pooideae and structural features of the plastid and nuclear genomes, including an intron loss in GBSSI. *Aliso* 23: 325–338.

DAVIS, J. I., D. W. STEVENSON, G. PETERSEN, O. SEBERG, L. M. CAMPBELL, J. V. FREUDENSTEIN, D. H. GOLDMAN, ET AL. 2004. A phylogeny of the monocots, as inferred from *rbcL* and *atpA* sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. *Systematic Botany* 29: 467–510.

DERTIEN, J. R., AND M. R. DUVALL. 2009. Biogeography and divergence in *Guaiacum sanctum* (Zygophyllaceae) revealed in chloroplast DNA: Implications for conservation in the Florida keys. *Biotropica* 41: 120–127.

DOYLE, J. J., J. I. DAVIS, R. J. SORENG, D. GARVIN, AND M. J. ANDERSON. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proceedings of the National Academy of Sciences, USA* 89: 7722–7726.

DUVALL, M. R., J. I. DAVIS, L. G. CLARK, J. D. NOLL, D. H. GOLDMAN, AND J. G. SÁNCHEZ-KEN. 2007. Phylogeny of the grasses (Poaceae) revisited. *Aliso* 23: 237–247.

FARRIS, J. S., V. A. ALBERT, M. KÄLLERSJÖ, D. LIPSCOMB, AND A. G. KLUGE. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* 12: 99–124.

FUKUDA, Y., Y. NAKAYAMA, AND M. TOMITA. 2003. On dynamics of overlapping genes in bacterial genomes. *Gene* 323: 181–187.

GOLOBOFF, P. A., J. S. FARRIS, AND K. C. NIXON. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* 24: 774–786 [version 1.1, published December, 2007].

GOULDING, S. E., R. G. OLMSTEAD, C. W. MORDEN, AND K. H. WOLFE. 1996. Ebb and flow of the chloroplast inverted repeat. *Molecular & General Genetics* 252: 195–206.

GRAHAM, S. W., P. A. REEVES, A. C. E. BURNS, AND R. G. OLMSTEAD. 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences* 161: S83–S96.

GRAHAM, S. W., J. M. ZGURSKI, M. A. MCPHERSON, D. M. CHERNIAWSKY, J. M. SAARELA, E. S. C. HORNE, S. Y. SMITH, ET AL. 2006. Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. *Aliso* 22: 3–20.

GPWG [Grass Phylogeny Working Group]. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden* 88: 373–457.

HANSEN, D. R., S. G. DASTIDAR, Z. CAI, C. PENAFLOR, J. V. KUEHL, J. L. BOORE, AND R. K. JANSEN. 2007. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution* 45: 547–563.

HILU, K. W., L. A. ALICE, AND H. LIANG. 1999. Phylogeny of Poaceae inferred from *matK* sequences. *Annals of the Missouri Botanical Garden* 86: 835–851.

HOLMGREN, P. K., AND N. H. HOLMGREN. 1998, Online edition of Index Herbariorum. New York Botanical Garden, Bronx, New York, USA. Website http://sweetgum.nybg.org/ih/ [accessed 10 February 2010].

JANSEN, R. K., Z. CAI, L. A. RAUBESON, H. DANIELL, C. W. DEPAMPHILIS, J. LEEBENS-MACK, K. F. MÜLLER, ET AL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369–19374.

JUDZIEWICZ, E. J., R. J. SORENG, G. DAVIDSE, P. M. PETERSON, T. S. FILGUEIRAS, AND F. O. ZULOAGA. 2000. Catalogue of New World grasses (Poaceae): I. Subfamilies Anomochlooideae, Bambusoideae, Ehrhartoideae, and Pharoideae. *Contributions from the United States National Herbarium* 39: 1–128.

KELCHNER, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87: 482–498.

KELCHNER, S. A., AND J. F. WENDEL. 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Current Genetics* 30: 259–262.

KELLOGG, E. A., AND H. P. LINDER. 1995. Phylogeny of Poales. *In* P. J. Rudall, P. J. Cribb, D. F. Cutler, and C. J. Humphries [eds.], Monocotyledons: Systematics and evolution, 511–542. Royal Botanic Gardens, Kew, UK.

KIM, K.-J., AND H.-L. LEE. 2005. Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Molecules and Cells* 19: 104–113.

LAVIN, M., J. J. DOYLE, AND J. D. PALMER. 1990. Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. *Evolution* 44: 390–402.

LEEBENS-MACK, J., L. A. RAUBESON, L. CUI, J. V. KUEHL, M. H. FOURCADE, T. W. CHUMLEY, J. L. BOORE, ET AL. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* 22: 1948–1963.

LEVINSON, G., AND G. A. GUTMAN. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203–221.

LINDER, H. P., AND E. A. KELLOGG. 1995. Phylogenetic patterns in the commelinid clade. *In* P. J. Rudall, P. J. Cribb, D. F. Cutler, and C. J. Humphries [eds.], Monocotyledons: Systematics and evolution, 473–496. Royal Botanic Gardens, Kew, UK.

LINDER, H. P., AND P. J. RUDALL. 1993. The megagametophyte in *Anarthria* (Anarthriaceae, Poales) and its implications for the phylogeny of the Poales. *American Journal of Botany* 80: 1455–1464.

LUO, Y., C. FU, D.-Y. ZHANG, AND K. LIN. 2006. Overlapping genes as rare genomic markers: The phylogeny of γ-proteobacteria as a case study. *Trends in Genetics* 22: 593–596.

MAIER, R. M., I. DÖRY, G. L. IGLOI, AND H. KÖSSEL. 1990. The *ndhH* genes of graminean plastomes are linked with the junctions between small single copy and inverted repeat regions. *Current Genetics* 18: 245–250.

MAIER, R. M., K. NECKERMANN, G. L. IGLOI, AND H. KÖSSEL. 1995. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *Journal of Molecular Biology* 251: 614–628.

MARCHANT, A. D., AND B. G. BRIGGS. 2007. Ecdeiocoleaceae and Joinvilleaceae, sisters of Poaceae (Poales): Evidence from *rbcL* and *matK* data. *Telopea* 11: 437–450.

MATHEWS, S., R. C. TSAI, AND E. A. KELLOGG. 2000. Phylogenetic structure in the grass family (Poaceae): Evidence from the nuclear gene phytochrome B. *American Journal of Botany* 87: 96–107.

MICHELANGELI, F. A., J. I. DAVIS, AND D. W. STEVENSON. 2003. Phylogenetic relationships among Poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from the mitochondrial and plastid genomes. *American Journal of Botany* 90: 93–106.

MUSE, S. A., AND B. S. GAUT. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* 146: 393–399.

NIXON, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15: 407–414.

NIXON, K. C. 2002. WinClada version 1.03. Computer program distributed by the author, Cornell University, Ithaca, New York, USA. Available at website http://www.cladistics.com/.

OGIHARA, Y., K. ISONO, T. KOJIMA, A. ENDO, M. HANAOKA, T. SHIINA, T. TERACHI, ET AL. 2002. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molecular Genetics and Genomics* 266: 740–746.

OLDENBURG, D. J., AND A. J. BENDICH. 2004. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *Journal of Molecular Biology* 335: 953–970.

OLMSTEAD, R. G., AND J. A. SWEERE. 1994. Combining data in phylogenetic systematics: An empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* 43: 467–481.

PALMER, J. D. 1983. Chloroplast DNA exists in two orientations. *Nature* 301: 92–93.

PALMER, J. D. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* 19: 325–354.

PERRY, A. S., S. BRENNAN, D. J. MURPHY, T. A. KAVANAGH, AND K. H. WOLFE. 2002. Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. *DNA Research* 9: 157–162.

PETERSON, P. M., R. J. SORENG, G. DAVIDSE, T. S. FILGUEIRAS, F. O. ZULOAGA, AND E. J. JUDZIEWICZ. 2001. Catalogue of New World grasses (Poaceae): II. Subfamily Chloridoideae. *Contributions from the United States National Herbarium* 41: 1–255.

PIRIE, M. D., A. M. HUMPHREYS, C. GALLEY, N. P. BARKER, G. A. VERBOOM, D. ORLOVICH, S. J. DRAFFIN, ET AL. 2008. A novel supermatrix approach improves resolution of phylogenetic relationships in a comprehensive sample of danthonioid grasses. *Molecular Phylogenetics and Evolution* 48: 1106–1119.

PLUNKETT, G. M., AND S. R. DOWNIE. 2000. Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. *Systematic Botany* 25: 648–667.

RAYKO, E. 1997. Organization, generation and replication of amphimeric genomes: A review. *Gene* 199: 1–18.

SÁNCHEZ-KEN, J. G., AND L. G. CLARK. 2001. Gynerieae, a new neotropical tribe of grasses (Poaceae). *Novon* 11: 350–352.

SÁNCHEZ-KEN, J. G., AND L. G. CLARK. 2007. Phylogenetic relationships within the Centothecoideae + Panicoideae clade (Poaceae) based on *ndhF* and *rpl16* intron sequences and structural data. *Aliso* 23: 487–502.

SÁNCHEZ-KEN, J. G., L. G. CLARK, E. A. KELLOGG, AND E. E. KAY. 2007. Reinstatement and emendation of subfamily Micrairoideae (Poaceae). *Systematic Botany* 32: 71–80.

SASKI, C., S.-B. LEE, S. FJELLHEIM, C. GUDA, R. K. JANSEN, H. LUO, J. TOMKINS, ET AL. 2007. Complete chloroplast genome sequences of *Hordeum vulgare, Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theoretical and Applied Genetics* 115: 571–590.

SIMMONS, M. P., AND H. OCHOTERENA. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* 49: 369–381.

SOLTIS, D. E., P. S. SOLTIS, M. W. CHASE, M. E. MORT, D. C. ALBACH, M. ZANIS, V. SAVOLAINEN, ET AL. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

SORENG, R. J., AND J. I. DAVIS. 1998. Phylogenetics and character evolution in the grass family (Poaceae): Simultaneous analysis of morphological and chloroplast DNA restriction site character sets. *Botanical Review* 64: 1–85.

SORENG, R. J., J. I. DAVIS, AND M. A. VOIONMAA. 2007. A phylogenetic analysis of Poaceae tribe Poeae sensu lato based on morphological characters and sequence data from three plastid-encoded genes: Evidence for reticulation, and a new classification for the tribe. *Kew Bulletin* 62: 425–454.

SORENG, R. J., P. M. PETERSON, G. DAVIDSE, E. J. JUDZIEWICZ, F. O. ZULOAGA, T. S. FILGUEIRAS, AND O. MORRONE. 2003. Catalogue of New World grasses (Poaceae): IV. Subfamily Pooideae. *Contributions from the United States National Herbarium* 48: 1–730.

STEFANOVIĆ, S., AND R. G. OLMSTEAD. 2005. Down the slippery slope: plastid genome evolution in Convolvulaceae. *Journal of Molecular Evolution* 61: 292–305.

STEVENSON, D. W., AND H. LOCONTE. 1995. Cladistic analysis of monocot families. *In* P. J. Rudall, P. J. Cribb, D. F. Cutler, and C. J. Humphries [eds.], Monocotyledons: Systematics and evolution, 543–578. Royal Botanic Gardens, Kew, UK.

SUGIURA, M. 1992. The chloroplast genome. *Plant Molecular Biology* 19: 149–168.

TSUMURA, Y., Y. SUYAMA, AND K. YOSHIMURA. 2000. Chloroplast DNA inversion polymorphism in populations of *Abies* and *Tsuga. Molecular Biology and Evolution* 17: 1302–1312.

WANG, R.-J., C.-L. CHENG, C.-C. CHANG, C.-L. WU, T.-M. SU, AND S.-M. CHAW. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology* 8: 36.

WATSON, L., AND M. J. DALLWITZ. 1992. The grass genera of the world. CAB International, Wallingford, UK.

WOLFE, K. H., W.-H. LI, AND P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.

YAMANE, K., K. YANO, AND T. KAWAHARA. 2006. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. *DNA Research* 13: 197–204.

ZHANG, W. 2000. Phylogeny of the grass family (Poaceae) from *rpl16* intron sequence data. *Molecular Phylogenetics and Evolution* 15: 135–146.

ZULOAGA, F. O., O. MORRONE, G. DAVIDSE, T. S. FILGUEIRAS, P. M. PETERSON, R. J. SORENG, AND E. J. JUDZIEWICZ. 2003. Catalogue of New World grasses (Poaceae): III. Subfamilies Panicoideae, Aristidoideae, Arundinoideae, and Danthonioideae. *Contributions from the United States National Herbarium* 46: 1–662.

APPENDIX 1. Taxa sampled for DNA sequences, plant accession information (with herbarium codes, Holmgren and Holmgren, 1998), and GenBank accession numbers for *ndhF*, *ndhH*, *rps15*. Taxonomic scheme (see section *Grass phylogenetics*) includes parenthesized codes for tribes of Poaceae (three letters), and subtribes of tribe Poeae (four letters). For eight taxa with an asterisk, gene sequences were obtained from published plastid genome sequences.

Taxonomic scheme	Species	Voucher (Herbarium)	*ndhF*	*ndhH*	*rps15*
Joinvilleaceae	*Joinvillea gaudichaudiana* Brongn. & Gris	J.I. Davis 751 (BH)	GU222696	GU222836	GU222752
Ecdeiocoleaceae	*Ecdeiocolea monostachya* F. Muell.	J.G. Conran et al. 938 (PERTH, ADU)	AY622313	GU222837	GU222753
Poaceae					
Anomochlooideae					
Anomochloeae (ano)	*Anomochloa marantoidea* Brongn.	J.I. Davis 753 (BH)	GU222697	GU222838	GU222754
Streptochaeteae (str)	*Streptochaeta sodiroana* Hack.	P.M. Peterson & E.J. Judziewicz 9525 (US)	AY622318	GU222839	GU222755
Pharoideae					
Phareae (pha)	*Pharus latifolius* L.	J.I. Davis; R.J. Soreng; no voucher	GU222698	GU222840	GU222756
Arundinoideae					
Arundineae (aru)	*Amphipogon strictus* R. Br.	H.P. Linder 5634 (BOL)	GU222717	GU222860	GU222776
	Arundo donax L.	M. Crisp 278 (CANB)	GU222718	GU222861	GU222777
	Molinia caerulea (L.) Moench	R.J. Soreng 3305; no voucher	GU222716	GU222859	GU222775
Micrairoideae					
Eriachneae (eri)	*Eriachne mucronata* R. Br.	S.W.L. Jacobs 8719 (NSW)	GU222714	GU222857	GU222773
	Eriachne pulchella Domin	S.W.L. Jacobs 8720 (NSW)	GU222715	GU222858	GU222774
Micraireae (mic)	*Micraira subulifolia* F. Muell.	S.W.L. Jacobs 8671 (NSW)	AY622316	GU222856	GU222772
Aristidoideae					
Aristideae (ari)	*Stipagrostis zeyheri* (Nees) DeWinter	N.P. Barker 1133 (BOL)	GU222711	GU222853	GU222769
Danthonioideae					
Danthonieae (dan)	*Danthonia californica* Bol.	grown from USDA Plant Intr. Sta. 232247; J.I. Davis 763 (BH)	GU222712	GU222854	GU222770
	Merxmuellera macowanii (Stapf) Conert	N.P. Barker 1008 (BOL)	GU222713	GU222855	GU222771
Chloridoideae					
No designated tribe	*Merxmuellera rangei* (Pilg.) Conert	N.P. Barker 960 (GRA)	GU222704	GU222846	GU222762
Cynodonteae (cyn)	*Distichlis spicata* (L.) E. Green subsp. *stricta* (Torr.) R.F. Thorne	K. Allred, 1992; no voucher	GU222709	GU222851	GU222767
Eragrostideae (era)	*Eragrostis tef* (Zucc.) Trotter	grown from commercial seed; J.I. Davis 771 (BH)	GU222708	GU222850	GU222766
	Uniola paniculata L.	J.I. Davis; no voucher	GU222707	GU222849	GU222765
Zoysieae (zoy)	*Spartina pectinata* Link	J.C. LaDuke; no voucher	GU222706	GU222848	GU222764
	Sporobolus giganteus Nash	P.M. Peterson et al. 10008 (US)	GU222705	GU222847	GU222763
	Zoysia Willd. sp.	J.I. Davis; no voucher	GU222710	GU222852	GU222768
Panicoideae					
Centotheceae (cen)	*Chasmanthium nitidum* (Baldwin) H.O. Yates	J.K. Wipff & S.D. Jones 2075 (TAES)	GU222699	GU222841	GU222757
Thysanolaeneae (thy)	*Thysanolaena maxima* (Roxb.) Kuntze	Fairchild Tropical Garden X-1-483 (FTG-81394, 81395)	GU222700	GU222842	GU222758
Gynerieae (gyn)	*Gynerium sagittatum* (Aubl.) P. Beauv.	L.G. Clark & P. Asimbaya 1472 (ISC)	GU222701	GU222843	GU222759
Paniceae (pan)	*Panicum virgatum* L.	grown from USDA Plant Intr. Sta. 421520; R.J. Soreng s.n. (BH)	GU222703	GU222845	GU222761
	Pennisetum alopecuroides (L.) Spreng.	R.J. Soreng s.n. (BH)	GU222702	GU222844	GU222760
Andropogoneae (and)	**Saccharum officinarum* L.	n/a	NC_006084	NC_006084	NC_006084
	**Sorghum bicolor* (L.) Moench	n/a	NC_008602	NC_008602	NC_008602
	**Zea mays* L.	n/a	NC_001666	NC_001666	NC_001666
Bambusoideae					
Bambuseae (bam)	*Phyllostachys* Siebold & Zucc. sp.	J.I. Davis 773 (BH)	GU222722	GU222865	GU222781
	Chusquea aff. *subulata* L.G. Clark	P.M. Peterson & E.J. Judziewicz 9499 (US)	GU222724	GU222867	GU222783
	Guadua angustifolia Kunth	P.M. Peterson & E.J. Judziewicz 9527 (US)	GU222725	GU222868	GU222784
Arundinarieae (arn)	*Pseudosasa japonica* (Steud.) Nakai	J.I. Davis 774 (BH)	GU222723	GU222866	GU222782
Olyreae (oly)	*Buergersiochloa bambusoides* Pilg.	J. Dransfield 1382 (K)	GU222726	GU222869	GU222785
	Eremitis Döll sp.	US National Herbarium Greenhouse 153, T.R. Soderstrom 2182 (US); or US National Herbarium Greenhouse 286; no voucher	GU222727	GU222870	GU222786
	Lithachne pauciflora (Sw.) P. Beauv.	L.G. Clark 1297 (ISC)	GU222729	GU222872	GU222788
	Olyra latifolia L.	P.M. Peterson & C.R. Annable 7311 (US)	GU222730	GU222873	GU222789
	Pariana radiciflora Döll	L.G. Clark & W. Zhang 1344 (ISC)	GU222728	GU222871	GU222787
Ehrhartoideae					
Streptogyneae (stg)	*Streptogyna americana* C.E. Hubb.	J.I. Davis; no voucher	GU222721	GU222864	GU222780
Ehrharteae (ehr)	*Ehrharta calycina* Sm.	grown from USDA Plant Intr. Sta. 208983; R.J. Soreng s.n. (BH)	GU222719	GU222862	GU222778

APPENDIX 1. Continued.

Taxonomic scheme	Species	Voucher (Herbarium)	*ndhF*	*ndhH*	*rps15*
Oryzeae (ory)	*Leersia virginica* Willd.	R.J. Soreng 3399a (BH)	GU222720	GU222863	GU222779
	**Oryza nivara* Sharma & Shastry	n/a	NC_005973	NC_005973	NC_005973
	**Oryza sativa* L.	n/a	NC_001320	NC_001320	NC_001320
Pooideae					
Brachyelytreae (brl)	*Brachyelytrum erectum* (Schreb.) P. Beauv.	R.J. Soreng 3427a (BH)	GU222731	GU222874	GU222790
Nardeae (nar)	*Nardus stricta* L.	E. Royl & C. Schiers s.n. (1988, B)	GU222733	GU222876	GU222792
Lygeeae (lyg)	*Lygeum spartum* L.	R.J. Soreng 3698 (BH)	GU222732	GU222875	GU222791
Phaenospermateae (phn)	*Anisopogon avenaceus* R. Br.	H.P. Linder 5590 (BOL)	GU222736	GU222879	GU222795
	Duthiea brachypodium (P. Candargy) Keng & Keng f.	R.J. Soreng 5358 (US)	GU222737	GU222880	GU222796
	Phaenosperma globosa Benth.	L.G. Clark 1292 (ISC)	GU222734	GU222877	GU222793
	Sinochasea trigyna Keng	R.J. Soreng 5644 (US)	GU222735	GU222878	GU222794
Stipeae (sti)	*Achnatherum occidentalis* (S. Watson) Barkworth subsp. *pubescens* (Vasey) Barkworth	R.J. Soreng 7418 (US)	GU222739	GU222882	GU222798
	Ampelodesmos mauritanica (Poir.) T. Durand & Schinz	R.J. Soreng & N.L. Soreng 4029 (BH)	GU222746	GU222890	GU222806
	Celtica gigantea (Link) F.M. Vázquez & Barkworth	R.J. Soreng 7443 (US)	GU222740	GU222884	GU222800
	Hesperostipa comata (Trin. & Rupr.) Barkworth	R.J. Soreng 7431 (US)	GU222744	GU222888	GU222804
	Nassella pulchra (Hitchc.) Barkworth	R.J. Soreng 7407 (US)	GU222741	GU222885	GU222801
	Nassella viridula (Trin.) Barkworth	grown from USDA Plant Intr. Sta. 387938; R.J. Soreng s.n. (BH)	GU222742	GU222886	GU222802
	Oryzopsis asperifolia Michx.	R.J. Soreng 5989 (US)	GU222743	GU222887	GU222803
	Piptatherum miliaceum (L.) Coss.	grown from USDA Plant Intr. Sta. 284145; J.I. Davis 767 (BH)	AY622317	GU222883	GU222799
	Stipa barbata Desf.	grown from USDA Plant Intr. Sta. 229468; J.I. Davis 768 (BH)	GU222745	GU222889	GU222805
	Timouria saposhnikovii Roshev.	R.J. Soreng 5448 (US)	GU222738	GU222881	GU222797
	Trikeraia pappiformis (Keng) P.C. Kuo & S.L. Lu	R.J. Soreng 5653 (US)	GU222747	GU222891	GU222807
Brylkinieae (bry)	*Brylkinia caudata* (Munro) F. Schmidt	Gao Hui 167 (SZ)	GU222750	GU222896	GU222812
Meliceae (mel)	*Glyceria grandis* S. Watson	J.I. Davis & R.J. Soreng; no voucher	AY622314	GU222894	GU222810
	Melica cupanii Guss.	grown from USDA Plant Intr. Sta. 383702; J.I. Davis 766 (BH)	AY622315	GU222892	GU222808
	Pleuropogon refractus (A. Gray) Benth.	R.J. Soreng 3381 (BH)	GU222749	GU222895	GU222811
	Schizachne purpurascens (Torr.) Swallen	R.J. Soreng 3348 (BH)	GU222748	GU222893	GU222809
Diarrheneae (dia)	*Diarrhena obovata* (Gleason) Brandenberg	J.I. Davis 756 (BH)	DQ786833	GU222897	GU222813
Brachypodieae (brp)	*Brachypodium pinnatum* (L.) P. Beauv.	grown from USDA Plant Intr. Sta. 440170; J.I. Davis 760 (BH)	AY622312	GU222898	GU222814
Bromeae (bro)	*Bromus inermis* Leyss.	grown from USDA Plant Intr. Sta. 314071; J.I. Davis 762 (BH)	DQ786821	GU222901	GU222817
	Bromus korotkiji Drobow	R.J. Soreng 5160 (US)	GU222751	GU222903	GU222819
	Bromus suksdorfii Vasey	R.J. Soreng 7412 (US)	DQ786822	GU222902	GU222818
	Littledalea tibetica Hemsl.	R.J. Soreng 5487; 5490; 5494 (US)	DQ786852	GU222899	GU222815
Triticeae (tri)	*Elymus trachycaulus* (Link) Shinners	R.J. Soreng 4291b (BH)	DQ786838	GU222900	GU222816
	**Hordeum vulgare* L. subsp. *vulgare*	n/a	NC_008590	NC_008590	NC_008590
	**Triticum aestivum* L.	n/a	NC_002762	NC_002762	NC_002762
Poeae (poe)					
Agrostidinae (agro)	**Agrostis stolonifera* L.	n/a	NC_008591	NC_008591	NC_008591
Aveninae (aven)	*Avena sativa* L. 'ASTRO'	grown from commercial seed; J.I. Davis 759 (BH)	DQ786814	GU222904	GU222820
	Trisetum cernuum subsp. *canescens* (Buckley) Calder & Roy L. Taylor	R.J. Soreng 3383a (BH)	DQ786874	GU222905	GU222821
Brizinae (briz)	*Briza minor* L.	grown from USDA Plant Intr. Sta. 378653; J.I. Davis 761 (BH)	DQ786820	GU222907	GU222823
Torreyochloinae (torr)	*Torreyochloa pauciflora* (J. Presl) G.L. Church	J.I. Davis 533 (BH)	DQ786872	GU222906	GU222822
Airinae (airi)	*Aira caryophyllea* L.	R.J. Soreng 5953b (US)	DQ786806	GU222908	GU222824
	Deschampsia cespitosa (L.) P. Beauv. subsp. *cespitosa*	R.J. Soreng 7417 (US)	DQ786831	GU222913	GU222829

APPENDIX 1. Continued.

Taxonomic scheme	Species	Voucher (Herbarium)	*ndhF*	*ndhH*	*rps15*
	Molineriella laevis (Brot.) Rouy	R.J. Soreng 3613 (BH)	DQ786857	GU222915	GU222831
Alopecurinae (alop)	*Phleum pratense* L.	R.J. Soreng 4293 (BH)	DQ786860	GU222911	GU222827
Cynosurinae (cyno)	*Cynosurus cristatus* L.	grown from RBG, Kew, seed bank 39006 (K)	DQ786829	GU222918	GU222834
Dactylidinae (dact)	*Dactylis glomerata* L. subsp. *hackelii* (Asch. & Graebn.) Cif. & Giacom.	R.J. Soreng 3692 (BH)	DQ786830	GU222917	GU222833
Holcinae (holc)	*Holcus annuus* C.A. Mey.	R.J. Soreng 3642 (BH)	DQ786849	GU222914	GU222830
Loliinae (loli)	*Festuca rubra* L.	R.J. Soreng 7424 (US)	DQ786839	GU222916	GU222832
Parapholiinae (para)	*Parapholis incurva* (L.) C.E. Hubb.	grown from RBG, Kew, seed bank 24867 (K)	DQ786859	GU222919	GU222835
Poinae (poin)	*Poa alpina* L.	R.J. Soreng 6115-1 (US)	DQ786861	GU222912	GU222828
Puccinelliinae (pucc)	*Puccinellia distans* (Jacq.) Parl.	J.I. Davis 755 (BH)	DQ786866	GU222909	GU222825
	Sclerochloa dura (L.) P. Beauv.	R.J. Soreng 3862 (BH)	DQ786869	GU222910	GU222826

APPENDIX 2. Primers used to amplify and sequence two regions of the plastid genome (cf. text and Fig. 1). Each primer name consists of (1) the name of the gene to which it is specific; (2) the numerical position within the primer binding region of the nucleotide closest to the 5′ end of the corresponding gene (regardless of the direction in which the primer reads), in the plastid genome sequence of *Triticum aestivum* (GenBank accession number NC_002762); and 3) either F (forward) or R (reverse), designating the direction in which the priming function proceeds, relative to the direction in which the gene is transcribed. Primers were developed by the authors, except as indicated.

Primer	Sequence
Region 1 (*ndhF* and *rps15*)	
ndhF-1F (Olmstead and Sweere, 1994)	5′ atg gaa caK aca tat Saa tat gc 3′
ndhF-45F	5′ act tcc agt tat tat gtc aat ggg Rtt t 3′
ndhF-274F (modified from Olmstead and Sweere, 1994)	5′ ctt act tct att atg tta ata cta at 3′
ndhF-309F	5′ Wgg aaY Yat ggt tct tat tta tag tga c 3′
ndhF-532F	5′ gcK ttt Dta act aat cgt gta ggg ga 3′
ndhF-818F	5′ gaa ttt ttc ttV tag ctc gag ttY ttc 3′
ndhF-933F	5′ tca Rag aga tat taa aag aag Ytt agc c 3′
ndhF-978F	5′ att ggg tta tat gat gtt agc tct agg t 3′
ndhF-1194F	5′ ttt att ggg tac act ttc tct ttg tg 3′
ndhF-1318F (modified from Olmstead and Sweere, 1994)	5′ gga tta act gcV ttt tat atg ttt cg 3′
ndhF-1421F	5′ att caa tat cSt tat ggg gaa aaa g 3′
ndhF-1811F	5′ atg caa ttt ctt ctg taa StY tag c 3′
ndhF-1969F	5′ tac agt tgg tca tat aat cgY ggt t 3′
ndhF-2101F	5′ ggt ctt SYt agt ttt tgt ata gga gaa g 3′
ndhF-972R (Olmstead and Sweere, 1994)	5′ cat cat ata acc caa ttg aga c 3′
ndhF-1117R	5′ cca tat tYt gac ttt tWt ctg gtg aat a 3′
ndhF-1373R	5′ act Rta atY ttg aaa atg aac acg ca 3′
ndhF-1968R	5′ ata acc Rcg att ata tga cca Rct gta t 3′
ndhF-2122F (Olmstead and Sweere, 1994)	5′ ccc cct aYa tat ttg ata cct tct cc 3′
ndhH-88F	5′ gtt act ctc gat ggt gaR gat gtt at 3′
rps15-80F	5′ ttc aag tat tca gtt tca cca ata aga t 3′
Region 2 (*ndhH* and *rps15*)	
ndhA-59R	5′ atc Sat atc agt cca tag act tct ttt a 3′
ndhH-684R	5′ atc taY ttt acg aag atc cca ttg tat t 3′
ndhH 88F	5′ gtt act ctc gat ggt gaR gat gtt at 3′
rps15-80F	5′ ttc aag tat tca gtt tca cca ata aga t 3′