



Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses

Shaoxing Huang^{1,+}, Anchalee Sirikhachornkit^{1,+}, Justin D. Faris^{2,3}, Xiujuan Su¹, Bikram S. Gill², Robert Haselkorn¹ and Piotr Gornicki^{1,*}

¹Department of Molecular Genetics and Cell Biology, University of Chicago, 920 East 58th Street, Chicago, IL 60637, USA (*author for correspondence; e-mail pg13@midway.uchicago.edu); ²Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA; ³current address: USDA/ARS Cereal Crops Research Unit, Northern Crop Science Laboratory, Fargo, ND 58105, USA; ⁺these authors contributed equally to the project

Received 6 March 2001; accepted in revised form 28 August 2001

Key words: evolution, gene sequence, grass, plant, Poaceae, Pooideae, Triticeae

Abstract

We have applied a two-gene system based on the sequences of nuclear genes encoding multi-domain plastid acetyl-CoA carboxylase (ACCase) and plastid 3-phosphoglycerate kinase (PGK) to study grass evolution. Our analysis revealed that these genes are single-copy in most of the grass species studied, allowing the establishment of orthologous relationships between them. These relationships are consistent with the known facts of their evolution: the eukaryotic origin of the plastid ACCase, created by duplication of a gene encoding the cytosolic multi-domain ACCase gene early in grass evolution, and the prokaryotic (endosymbiont) origin of the plastid PGK. The major phylogenetic relationships among grasses deduced from the nucleotide sequence comparisons of ACCase and PGK genes are consistent with each other and with the milestones of grass evolution revealed by other methods. Nucleotide substitution rates were calculated based on multiple pairwise sequence comparisons. On a relative basis, with the divergence of the Pooideae and Panicoideae subfamilies set at 60 million years ago (MYA), events leading to the *Triticum/Aegilops* complex occurred at the following intervals: divergence of *Lolium* (*Lolium rigidum*) at 35 MYA, divergence of *Hordeum* (*Hordeum vulgare*) at 11 MYA and divergence of *Secale* (*Secale cereale*) at 7 MYA. On the same scale, gene duplication leading to the multi-domain plastid ACCase in grasses occurred at 129 MYA, divergence of grass and dicot plastid PGK genes at 137 MYA, and divergence of grass and dicot cytosolic PGK genes at 155 MYA. The ACCase and PGK genes provide a well-understood two-locus system to study grass phylogeny, evolution and systematics.

Abbreviations: ACCase, acetyl-CoA carboxylase; *Acc-1* and *Acc-2*, genes encoding plastid and cytosolic ACCase, respectively; Ψ -*Acc-2*, *Acc-2*, related partially processed pseudogene; PGK, 3-phosphoglycerate kinase; *Pgk-1* and *Pgk-2*, genes encoding chloroplast and cytosolic PGK, respectively; MYA, million years ago

Introduction

Our goal is to address various questions about evolution of grasses at different levels of genetic relatedness using a multi-gene DNA sequence comparison. In this

The nucleotide sequence data reported will appear in the GenBank database under the accession numbers AF343496–AF343536, AF343459–AF343473, AF343454–AF343456, AF343457, AF343458, AF343474–AF343495, and AF343447–AF343453.

paper we describe an analysis of the origin and evolution of the multi-domain acetyl-CoA carboxylase (ACCase) and its gene family (*Acc*) as well as plastid and cytosolic 3-phosphoglycerate kinase (PGK) and their genes (*Pgk*) in plants, focusing on the grass family. A large set of new *Acc* and *Pgk* gene sequences from selected species were analyzed and several milestone events in grass evolution leading to the *Triticum* and *Aegilops* lineage were examined.

The grass family

The grass family (Poaceae) includes over 10 000 species belonging to the major subfamilies Pooideae, Bambusoideae, Chloridoideae, and Panicoideae (Devos and Gale, 1997; Kellogg, 1998). The family includes some of the world's major food and forage crop plants. The tribes of the subfamily Pooideae include Triticeae (wheat, barley and rye), Aveneae (oats) and Poeae (*Lolium* and *Festuca*). The tribes of the subfamily Panicoideae include Paniceae (*Panicum*, *Setaria*, *Pennisetum*), Maydeae (maize) and Andropogoeae (*Saccharum*, *Sorghum*). The subfamily Bambusoideae includes the tribe Oryzeae (rice) and the subfamily Chloridoideae includes the tribe Chloridoideae (*Eleusine*). The various taxa range from annuals to perennials, from herbaceous to woody trees, and they vary in their habit, physiology (C3 and C4 photosynthesis) and reproductive systems. The genome size in grasses varies from one of the smallest in rice (440 Mb) to among the largest in wheat (16 000 Mb). Some of the variation in genome size is due to ploidy varying from diploid to highly polyploid, for example in sugarcane.

Wheat and its relatives

The main genetic characteristic of the Triticeae tribe, including wheat (*Triticum*), *Aegilops*, rye (*Secale*) and barley (*Hordeum*), is the basic chromosome number of 7 and a variable ploidy level. The 31 species of the *Triticum* and *Aegilops* genera include 13 diploids and 18 polyploids. Genetic relationships among bread wheat (*Triticum aestivum*) and its wild and domesticated diploid and polyploid relatives of the *Triticum* and *Aegilops* complex have been studied extensively (Dvorak and Zhang, 1990; Tsunewaki, 1991; Dvorak *et al.*, 1993; Wang *et al.*, 1997; Dvorak, 1998), prompted by its economic importance (reviewed in Cox, 1998). The basic genomes of diploid species of the *Triticum/Aegilops* complex have letter names: A, B, D, G, and S. The designation of the genomes in polyploid species reflect their origin. The hexaploid bread wheat (*T. aestivum*) has the genome composition AABBDD. It arose by spontaneous hybridization between the AABB-genome tetraploid *T. turgidum* and the DD-genome diploid *A. tauschii*. The A genome of the tetraploid was donated by the AA-genome diploid *T. urartu*. The origin of the B genome remains controversial but *A. speltoides* is the most likely living relative of an extinct or yet to be discovered B genome donor species.

Timeline of grass evolution

The early evolution of plants has been discussed recently (Soltis and Soltis, 2000; Pryer *et al.*, 2001). Fossil records suggest that the grass family probably originated in the Upper Cretaceous period (Crepet and Feldman, 1991) after the divergence of monocots from dicots. The oldest monocot fossil was recovered from Upper Cretaceous sediments, documenting the existence of this plant lineage earlier than 90 MYA, with some putative monocot fossils dated even earlier (Gandolfo *et al.*, 1998). The oldest fossils of species representing different subfamilies were dated, as discussed (Wolfe *et al.*, 1989; Clark *et al.*, 1995), at ca. 40 MYA suggesting that the radiation of the grass subfamilies may have occurred some 50–70 MYA. On this scale, hexaploid bread wheat is a young species established by hybridization between tetraploid *T. turgidum* and diploid *A. tauschii* about 8000 years ago (McFadden and Sears, 1946).

In the absence of reliable fossil records, establishing the timeline of the early events leading to the diverse lineages of the grass family is difficult. This lack of fossil records is reflected in difficulties in estimating the time of monocot-dicot divergence based on the molecular clock concept (Wolfe *et al.*, 1989). The divergence time of the grass subfamilies, 50–70 MYA, has been frequently used to calibrate the molecular clock and to calculate absolute nucleotide substitution rates (Wolfe *et al.*, 1987, 1989; Gaut and Doebley, 1997; Gaut, 1998). These calculations, however inaccurate and tentative, proved useful in providing some perspective on the sequence of evolutionary events and allowing comparisons of results from independent studies.

Molecular tools

Classical studies on grass classification and phylogeny based on morphological and physiological characters have been extended recently to analysis at the DNA sequence level (e.g. Clark *et al.*, 1995, 2000; Hsiao *et al.*, 1995; Catalan *et al.*, 1997; Peterson and Seberg, 1997; Hsiao *et al.*, 1998; Mason-Gamer *et al.*, 1998; Mathews *et al.*, 2000; Zhang, 2000). Questions concerning selection, genetic diversity, domestication, nucleotide substitution rates in nuclear and chloroplast genes, molecular clocks, and divergence times were analyzed for some grass lineages based on the DNA sequence of several genes (e.g. Wolfe *et al.*, 1987, 1989; Gaut *et al.*, 1996, 1997; Muse and Gaut, 1997; Cummings

and Clegg, 1998; Eyre-Walker *et al.*, 1998). Comparative analysis was applied to multi-gene chromosome regions (Chen *et al.*, 1998) and to retrotransposons (SanMiguel *et al.*, 1999) aiming at a better understanding of evolutionary processes in intergenic regions and their effect on the neighboring functional genes.

Molecular tools based on DNA sequence comparisons, their successful applications and potential problems for plants have been discussed recently (Kellogg *et al.*, 1996; Clegg, 1997; Clegg *et al.*, 1997; Gaut, 1998; Doyle and Gaut, 2000; Muse, 2000; Soltis and Soltis, 2000; Pryer *et al.*, 2001). This approach can provide much needed information if it is based on sequences of a sufficiently large DNA fragment with a significant number of variable characters. Its success depends on proper selection and good understanding of the origin, evolution, copy number, and structure and function of the genes (or other DNA segments) being analyzed. Different DNA segments (e.g. exons or introns) can be used to analyze groups of species of different genetic relatedness. Establishing orthologous relationships between the genes is one of the key issues. Gene- and chromosome locus-specific effects on the nucleotide substitution rates and rate heterogeneity in different lineages are some others. Multi-gene analysis and substantial sampling of taxa is highly recommended.

ACCase and PGK and their functions in plants

The key steps in the evolution of plant ACCases revealed by earlier analysis based on protein subunit structure and sequence are summarized in Figure 1. Plants have two forms of ACCase. The cytosolic isoform is a large multi-domain enzyme of eukaryotic origin. It provides malonyl-CoA for the synthesis of very-long-chain fatty acids and a variety of important secondary metabolites, and for malonylation. A different ACCase isoform found in the plastids catalyzes the first step in *de novo* fatty acid biosynthesis. In grasses, this ACCase is a multi-domain enzyme of eukaryotic origin, which arose by duplication of an ancestral cytosolic ACCase gene. In most other plants, it is a multi-subunit enzyme of prokaryotic (endosymbiont) origin. The multi-domain ACCase functions as a dimer consisting of ca. 250 kDa subunits. The multi-subunit ACCase contains four different subunits each 30–50 kDa in size.

Plant cytosolic and plastid PGK isozymes participate in glycolysis and the Calvin cycle, respectively. Plant PGKs are small (ca. 44 kDa) monomeric pro-

teins. They are both encoded by nuclear genes that arose by duplication of a gene of prokaryotic (endosymbiont) origin. The origin, evolution and function of plant PGK have been discussed recently (Martin and Schnarrenberger, 1997).

ACCase and PGK and their genes followed two different scenarios for the evolution of metabolic functions in organelles and the flow of genetic material between plastid and nuclear genomes (Martin *et al.*, 1998).

ACCase and PGK and their genes in wheat

Genes encoding wheat cytosolic and plastid ACCase, both nuclear, are ca. 15 kb in size and have some 30 introns which are variable in size but located at equivalent positions (Podkowinski *et al.*, 1996; Gornicki *et al.*, 1997), except for introns in the leader and plastid transit peptide coding region which are unique in sequence and location (Figure 2). The cytosolic gene has two introns fewer. All known plant genes encoding the multi-domain ACCase have the same intron-exon structure. Proteins encoded by the two wheat genes are 2260 and 2311 amino acids long. Their sequences are 67% identical. A putative plastid transit peptide is present at the N-terminus of the plastid isoform (Gornicki *et al.*, 1997). Multiple copies of the cytosolic ACCase gene (*Acc-2*) exist in wheat (Gornicki *et al.*, 1994; Podkowinski *et al.*, 1996; Faris *et al.*, 2001) and they are clustered on the long arms of chromosomes 3A, 3B, 3D, and 5D (Faris *et al.*, 2001). An *Acc-2*-related partially processed pseudogene (Ψ -*Acc-2*) is present in the *Acc-2* locus on chromosomes 3A, 3B and 3D (Faris *et al.*, 2001). A single copy of the plastid ACCase gene (*Acc-1*) is present on the short arms of each of the homoeologous group 2 chromosomes 2A, 2B, and 2D (Gornicki *et al.*, 1997).

The genes encoding wheat PGK isozymes, both nuclear, are ca. 3 kb in size. The wheat plastid gene (*Pgk-1*) has 5 introns. The wheat cytosolic gene (*Pgk-2*) has at least four introns (this work) all located at equivalent positions. Proteins encoded by plastid and cytosolic PGK genes contain 480 and 401 amino acids, respectively. Their sequences are 82% identical. A putative plastid transit peptide is present at the N-terminus of the plastid isoform. The chromosome location of the *Pgk* genes in wheat is not known.

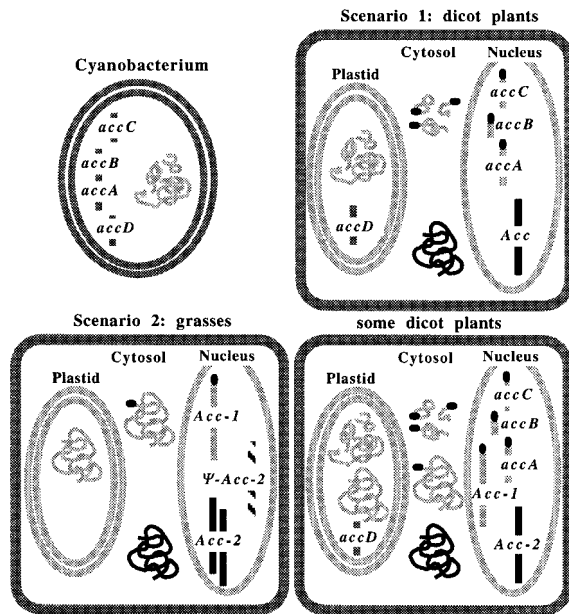


Figure 1. The origin of plant ACCases was deduced from their subunit structure and amino acid sequence comparisons (Sasaki *et al.*, 1995; Podkowinski *et al.*, 1996; Gornicki *et al.*, 1997; Inledon and Hall, 1997). Endosymbiont of cyanobacterial origin took over fatty acid biosynthesis using its four-subunit ACCase to provide malonyl-CoA. Three genes (*accA*, *B* and *C*, in bacterial nomenclature, encoding α subunit of carboxyltransferase, biotin carboxyl carrier protein, and biotin carboxylase, respectively) were then transferred to the nucleus. The gene encoding the β subunit of carboxyltransferase (*accD*) remained in the chloroplast genome. ACCase of the ancestral host eukaryote was retained and functions in the cytosol in fatty acid elongation and secondary metabolism. Duplication of the cytosolic ACCase gene (*Acc*) led to the creation of a new eukaryotic-type gene encoding a multi-domain plastid ACCase (*Acc-1*). The multi-domain plastid ACCase appeared in the plant kingdom independently in grasses and in some dicots, probably in multiple lineages (Schulte *et al.*, 1997; Gimenez-Espinosa *et al.*, 1999; Christopher and Holtum, 2000). In *Brassica napus*, both multi-subunit (prokaryotic-type) and multi-domain (eukaryotic-type) ACCases exist in plastids. In grasses, both plastid and cytosolic isozymes are of the multi-domain type and are encoded by nuclear genes *Acc-1* and *Acc-2*, respectively. No bacterial-type multi-subunit ACCase has been found in grasses, although plastid genomes of some grasses contain a remnant of the *accD* gene. In the wheat lineage, more recent gene duplication event(s) created new copies of the *Acc-2* gene. A partially processed pseudogene (Ψ -*Acc-2*) is also present in at least some species of the Triticeae tribe. Subunits of plastid and cytosolic ACCases and their genes are shown in light and dark gray as coils and rectangles, respectively. The plastid transit peptide and its coding sequence is indicated with a black oval.

Materials and methods

Plant material

Seeds used in this study were obtained from the Wheat Genetics Resource Center (Kansas State University), except for *Lolium* (*Lolium rigidum*, AUS92) which was provided by T. Niderman (Novartis, Basel). Plant material used in this study is listed in Table 1.

PCR cloning and sequencing

The PCR cloning experiment was designed to allow amplification and identification of multiple copies of each gene from species of different ploidy and different genetic relatedness. A set of universal primers targeted to highly conserved exon sequences was designed for each gene (Table 2). *Acc*-specific PCR primers were designed based on the available coding sequences of ACCases from wheat and maize (GenBank AF029896, AF029897, AF305201, U39321, AF305205 and U19183). *Pgk*-specific PCR primers were designed based on the sequence of cDNA encoding wheat cytosolic (GenBank X15233) and plastid PGK (GenBank X15232), and of a genomic clone encoding wheat plastid PGK (GenBank X73528). Length of the primers, conditions of the PCR amplification step, utilization of several primer combinations for each gene, cloning of multiple amplification products were all designed to overcome potential PCR bias.

DNA from plant material was extracted as described before (Faris *et al.*, 2000). High-fidelity PCR was carried out according to the manufacturer's protocol (Roche). All components of the PCR except for the DNA polymerase were incubated for 2–3 min at 94 °C. The PCR was then initiated by the addition of polymerase. Amplification was for 35 cycles, each 0.5 min at 94 °C, 0.5 min at 52–56 °C and 4–6 min at 68 °C, using 0.5–1.0 μ g of genomic DNA as template in a 50 μ l reaction. The PCR products were cloned and multiple clones were sequenced.

Several quality control measures were employed to eliminate all cloning and sequencing artifacts, such as chimeric DNA molecules (Faris *et al.*, 2001). This was achieved by sequencing multiple clones representing each gene. With a few exceptions, those cases in which the PCR error rate was negligible in comparison to the nucleotide substitution rates analyzed, all the sequences used in the phylogenetic analysis were derived from at least three independent clones.

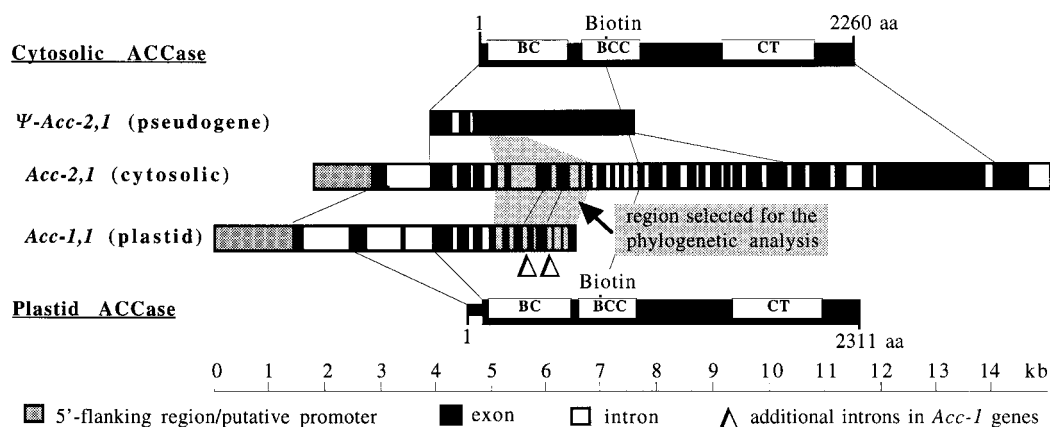


Figure 2. Structure of wheat cytosolic and plastid ACCases and their genes. ACCase functional domains – biotin carboxylase (BC), biotin carboxyl carrier containing covalently attached biotin (BCC), carboxyltransferase (CT) – were identified by sequence comparison as described before (Gornicki *et al.*, 1994, 1997; Podkowinski *et al.*, 1996). Gene fragment selected for the phylogenetic analysis of *Acc* genes is shaded in gray. The diagram is based on partial sequence of *Acc-1,1* (GenBank AF029897), full-length sequence of *Acc-2,1* (GenBank AF305204 and U39321) and partial sequence of Ψ -*Acc-2,1* (GenBank AF305209).

Sequence and phylogenetic analysis software

Sequencher (Gene Codes Corporation, Ann Arbor, MI) was used to manage the sequencing project. ClustalX (Thompson *et al.*, 1994) was used to create multiple sequence alignments. MacClade (W. P. Maddison and D. R. Maddison, Sinauer Associates, www.sinauer.com) was used for analysis of multiple sequence alignments. PAUP (D. L. Swofford, Sinauer Associates, www.sinauer.com) and MEGA (S. Kumar, K. Tamura, I. Jakobsen and M. Nei, www.megasoftware.net) were used to calculate genetic distances and phylogenetic trees.

Phylogenetic analysis

All multiple sequence alignments were created using ClustalX. Alignments of exon and intron sequences were used separately to create phylogenetic trees and to calculate nucleotide substitution rates. The alignment of *Acc* exon sequences and the corresponding 'intronless' pseudogene sequences was unambiguous with only a few gaps (all multiples of three nucleotides). This alignment included exons of the human *ACCI* gene as an outgroup. Multiple alignment of *Acc-1* intron sequences was obtained in three steps. First, gene sequences (exons and introns) from Triticeae species were aligned. At this level, no manual adjustments were required. In the next step, *Lolium* sequences were added. Alignment of *Lolium* sequences was manually adjusted in a segment of one intron. Multiple alignments of all *Acc-1* sequences at once did

not provide any clear suggestions for possible further improvements. Intron sequences of maize genes could not be aligned in a meaningful way. Finally, exon sequences were removed from the alignment. Alignment of the *Acc-2* gene sequences was obtained by the same step-wise approach beginning with independent alignments of sequences of clades *Acc-2I* and *Acc-2II*, aligning them together, adding *Lolium* sequences and finally separating exons and introns (Faris *et al.*, 2001). A short stretch of sequence found only in an *Acc-2,1* intron was manually adjusted to be a single insertion rather than aligning it with other sequences in this region, as suggested by the outcome of the ClustalX analysis. As described before (Faris *et al.*, 2001), alignment of intron sequences of the two major *Acc-2* clades (I and II) required only minor manual adjustments. However, the sequences of some regions of *Lolium Acc-2* introns were too divergent to produce reliable alignment. Alignments of *Pgk-1* and *Pgk-2* sequences from Triticeae species were created separately and then separated into exon and intron parts. Exon sequences of all plant *Pgk* genes, including maize, plus the *Synechocystis Pgk* gene as an outgroup, were then aligned together. This alignment was unambiguous, with only one 3 nt gap. The alignment of *Pgk* intron sequences from Triticeae did not require any manual adjustments, except for a single segment found in *Pgk-1* from rye but not in any other species. This segment was treated as one insertion rather than aligning it with some other sequences in this region, as suggested by the outcome of ClustalX analysis. Intron sequences

Table 1. Plant material

Species	Genome	Accession/cultivar	Origin
<i>Triticum urartu</i>	A	TA763	Lebanon
<i>Triticum monococcum</i>	A ^m		
ssp. <i>monococcum</i>		TA2025	Turkey
		TA138	Unknown
		TA142	Bosnia
ssp. <i>aegilopoides</i>		TA291	Iraq
		TA183	Iran
<i>Triticum turgidum</i>	AB		
ssp. <i>dicoccoides</i>		TA73	Lebanon
		TA1392	Israel
		TA84	Turkey
		TA51	Israel
<i>Triticum timopheevii</i>	AG		
spp. <i>timopheevii</i>		TA103	Yugoslavia
spp. <i>armeniicum</i>		TA2	Armenia
<i>Triticum aestivum</i>	ABD	TAM107	USA
<i>Aegilops tauschii</i>	D		
ssp. <i>tauschii</i>		TA1704	Tadjikistan
ssp. <i>tauschii</i>		TA1691	Unknown
<i>Aegilops speltoides</i>			
ssp. <i>speltoides</i>	S	TA2368	Turkey
		TA2780	Israel
		TA1789	Iraq
		TA1793	Syria
ssp. <i>ligustica</i>	S	TA2645	Turkey
		TA2779	Israel
		TA1770	Iraq
<i>Aegilops searsii</i>	S ^s	TA2343	Syria
		TA2355	Israel
		TA1837	Jordan
		TA1840	Israel
<i>Aegilops longissima</i>	S ^l	TA1912	Israel
		TA1921	Jordan
<i>Aegilops sharonensis</i>	S ^{sh}	TA1966	Israel
		TA2065	Turkey
<i>Aegilops bicornis</i>	S ^b	TA1954	Egypt
Rye (<i>Secale cereale</i> , Triticeae)		Imperial	–
Barley (<i>Hordeum vulgare</i> , Triticeae)		Betzes	–
Ryegrass (<i>Lolium rigidum</i> , Poaceae)		AUS 92	–

of maize *Pgk* genes could not be aligned to the other sequences in a meaningful way.

Phylogenetic trees were created in several different ways. First, phylogenetic trees based on all available exon sequences of plant multi-domain ACCases and all plant PGKs were created by the neighbor-joining method. These trees were calculated including

all exon positions without correction for multiple substitutions and were characterized by very good support for major clades indicated by bootstrap values higher than 80% of 1000 replicates. Second, neighbor-joining trees based on synonymous sites only, with or without correction for multiple substitutions (Jukes-Cantor and other methods), were calculated. The number of

Table 2. Gene-specific primers used for PCR.

Forward primers		Reverse primers
	<i>Acc-1</i>	
GTTCTGGCTCCCAATATTATC		TTCAAGAGATCAACTGTGTAATCA
CCCAATATTATCATGAGACTTGCA		CAACATTTGAATGAATHCTCCACG
	<i>Pgk-1</i>	
CACCTGGGTCGTCTAAGGGTGTT		ACCACCAGTTGAGATGTGGCTCAT
TCGTCTAAGGGTGTTACTCTAA		AAGCTCGCGCCACCACCAGTTGAG
	<i>Pgk-2</i>	
CATCTGGGCGCCAAAAGGTGTC		CCGCCAAAAGGTGTCACCCCAA
CACCGCGGTGGAAATGTGGCTCA		AAGCTGGCGCCACCAGCGGTGGAA

nucleotide differences between some groups of sequences was too high to calculate corrected distances. Trees including subsets of sequences were then evaluated. These trees had the same topology at those well-supported nodes but differed at some of the nodes with low bootstrap support. Finally, the same conclusion was reached for trees generated by the heuristic maximum parsimony search (equally weighted characters and nucleotide transformations, 1000 random-addition replicates, tree bisection-reconnection branch swapping). Twelve best trees (length 1045) were found for the *Acc* gene based on 291 informative characters. A single best tree (length 1486) was found for the *Pgk* gene based on 376 informative characters. Parsimony bootstrap analysis followed the same scheme with 1000 replicates each with 10 random-addition replicates. A subset of 40 sequences representing all major clades was included in the parsimony bootstrap analysis of the *Acc* gene.

Nucleotide substitution rates at nonsynonymous, synonymous and intron positions were calculated for major clades. Average values and standard deviations were calculated for all feasible pairwise comparisons of each type. Synonymous and intron substitution rates were corrected for multiple substitutions using the Jukes-Cantor method as implemented by MEGA. Divergence times were calculated as described (Gaut, 1998). Substitution rate heterogeneity was assessed for major lineages by a simple relative-rate test (Li and Bousquet, 1992).

Results

Gene sequences for phylogenetic analysis

Obtaining reliable nucleotide sequence information is a key issue in phylogenetic analysis. To assure proper sampling, multiple species of different genetic relatedness need to be included. The PCR cloning approach is well suited for this task but it has to work reliably for diverse species for which no gene sequence is available. PCR errors and cloning artifacts need to be eliminated. Quality control becomes especially critical for analysis of very closely related species. These problems are compounded by plant polyploidy and by the multicopy character of some genes, so multiple copies of each gene must be analyzed.

We have developed a PCR-based method to obtain DNA sequence information for phylogenetic analysis of four different genes from grasses (this work, unpublished results and Faris *et al.*, 2001): a 1.5–1.8 kb fragment of the *Acc* genes and ca. 1.5 kb fragment of the *Pgk* genes. The *Acc-1* gene fragment encoding the plastid enzyme spans 8 exons and 7 introns and the corresponding fragment of the *Acc-2* gene encoding the cytosolic ACCase spans 6 exons and 5 introns (Figure 2). This fragment of the *Acc* genes encodes 236 amino acids of the biotin carboxylase domain (approximately one-tenth of the mature ACCase). The *Pgk-1* (plastid enzyme) and *Pgk-2* (cytosolic enzyme) gene fragments span 5 exons and 4 introns and encode 298 amino acids (approximately three-quarters of the mature PGK). The cloning method was tested on multiple *Triticum* and *Aegilops* species as well as rye, barley and *Lolium* (*Lolium rigidum*) (Faris *et al.*, 2001, and this work), maize and switchgrass (*Panicum virgatum*) (this work and unpublished results).

Acc-2 gene sequences (Faris *et al.*, 2001), maize and switchgrass *Acc-1* sequences (unpublished results), as well as coding sequences of several plant multi-domain ACCase and several PGKs available from GenBank were also included in the analysis.

How many copies of ACCase and PGK genes exist in grass genomes?

Southern analysis and chromosome mapping indicate that a single copy of the *Acc-1* gene is present in each of the group 2 homoeologous chromosomes in hexaploid wheat (Gornicki *et al.*, 1997). Sequence analysis of the *Acc-1* gene from most species of the Triticeae tribe showed a single copy of the gene per diploid chromosome. A single copy was found in barley and rye. All *Triticum* and most *Aegilops* species probed extensively have only one copy of the *Acc-1* gene. The exceptions were *A. speltoides* species accessions 1789, 2779 and 1780, a diploid habitually out-pollinated species. The presence of two copies of the *Acc-1* gene in this species may be explained by heterozygosity or by the heterogeneous nature of the populations. However, the existence of duplicated genes cannot be ruled out. Two copies were detected in maize, consistent with its chromosome constitution, and in *Lolium* despite its diploid nature. These two species, however, were not probed extensively.

Analysis of the *Acc-2* gene in hexaploid wheat (Faris *et al.*, 2001) suggested that it is present in a small number of copies clustered at two chromosome loci. Two major clades, *Acc-2I* and *Acc-2II*, which most likely arose by gene duplication, as well as additional possible gene duplication events which occurred at a later stage in some *Triticum* and *Aegilops* lineages, were identified by sequence analysis (Faris *et al.*, 2001). The exact copy number and strict orthologous relationships among the *Acc-2* genes of the Triticeae tribe species have not been established. No evidence of multiple copies of the pseudogene Ψ -*Acc-2* was found.

The analysis of the *Pgk-1* gene showed that, similar to *Acc-1*, it is present in a single copy per diploid chromosome. The only exception is *A. speltoides* accession 1789, for which two sequences were detected, with the same possible explanation as for the *Acc-1* gene in this species. The *Pgk-2* gene copy number remains unknown. Too few genes were analyzed to reach any conclusion in this regard.

Phylogenetic inferences based on the Acc and Pkg genes

The phylogenetic tree based on exon positions (synonymous and nonsynonymous) of multi-domain ACCases genes was created by the neighbor-joining method with distances not corrected for multiple substitutions and gaps excluded only from pairwise comparisons (Figure 3). The dicot clade was easily distinguished from the grass clade. The revealed relationships of taxa within the dicot clade are consistent with their taxonomic relationships. Within grasses, the phylogeny of the plastid *Acc-1* gene showed clearly two clades corresponding to the Panicoideae and Pooideae subfamilies. It revealed distinctiveness of the tribes Paniceae (*Panicum*) and Maydeae (maize) within Panicoideae, and tribes Poeae (*Lolium*) and Triticeae (barley, rye, wheat) within Pooideae. It showed the close relationship of *Triticum/Aegilops* species with rye, and the *Triticum/Aegilops*/rye clade with barley. Despite a loss of resolution, various species of *Triticum* and *Aegilops* could be distinguished and the relationships of some diploid donor species with polyploid wheat were revealed. For example, the A genome of diploids and polyploids formed one clade, and *A. tauschii* ssp. *meyeri* (TA1691) clustered with the D genome of hexaploid wheat.

The Pooideae *Acc-2* clade includes four major branches: Poeae (*Lolium*), two Triticeae clades (I and II) and a clade consisting of the *Acc-2*-related processed pseudogene (Ψ -*Acc-2*) (Figure 3). It has been suggested previously that the *Acc-2* small gene family was formed by recurring gene duplications and possible gene loss, as well as intron loss (Faris *et al.*, 2001). These gene duplications occurred at different times, from the early event that gave rise to the *Acc-1* gene to more recent events in *Lolium*, barley and some *Triticum* and *Aegilops* lineages. Two introns were lost in all grass *Acc-2* genes, additional introns were lost in one barley gene, and most introns are not present in the pseudogene (Faris *et al.*, 2001). Despite the unresolved question of *Acc-2* gene copy number and difficulties in establishing orthology, some phylogenetic relationships among Pooideae species were clearly revealed and consistent with the relationships deduced based on the *Acc-1* phylogeny. This includes relationships among the *Lolium*, barley, rye and *Triticum/Aegilops* lineages as well as some relationships among *Triticum* and *Aegilops* species. The phylogeny based on the pseudogene sequences seems to reflect correctly these relationships as well.

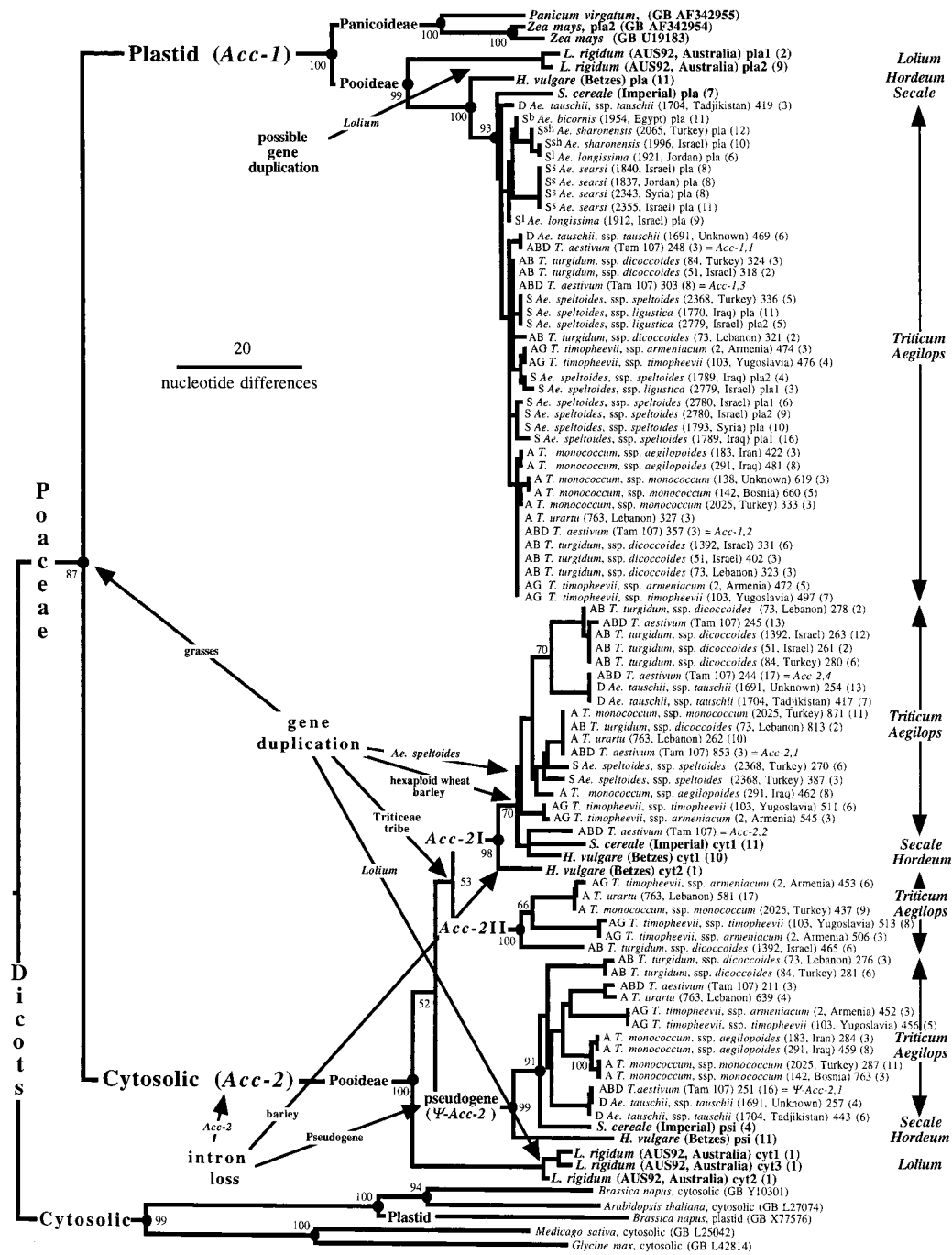


Figure 3. Phylogenetic relationships among plant multi-domain ACCases. A phylogenetic tree was created by the neighbor-joining method with distances not corrected for multiple substitutions and gaps excluded only from pairwise comparisons. All exons and all 'intronless' pseudogene positions were included in the calculation. Alignment length: 696 nucleotides (232 codons); outgroup: exons of human *ACC1* gene (GenBank X68968); invariant positions: 290 (376 without outgroup); bootstrap values: % of 1000 trials (shown for selected nodes only). The sequence alignment unambiguous with a single 3-nucleotide gap in the *B.napus* plastid sequence, and two gaps (one of 3 and one of 30 nucleotides) in the human outgroup sequence. Taxon names include genome composition (*Triticum/Aegilops* species), species name, accession number/cultivar name, geographical origin (*Triticum/Aegilops* species and *Lolium rigidum*), sequence name/number and number of clones analyzed. Taxon names of grass species other than *Triticum* and *Aegilops* are shown in bold. Genome composition of *Triticum/Aegilops* species is shown using abbreviated names listed in Table 1. *Acc-2* genes (GenBank AF306803–306029) were described in Faris *et al.* (2001). Other sequences obtained from GenBank are identified by their accession numbers (GB). New gene sequences were deposited in GenBank under the following accession numbers: *Acc-1* genes, AF3434960–AF343536; Ψ -*Acc-2* (pseudogene), AF343459–AF343473; *L. rigidum* *Acc-1* genes, AF34357 and AF343458; and *L. rigidum* *Acc-2* genes, AF343454–AF343456. Dots indicate nodes relevant for the analysis presented in this paper which showed strong statistical support by the neighbor-joining method including synonymous sites only and by the maximum parsimony method.

Table 3. Substitution rates and relative timeline of grass evolution leading to the wheat lineage. Average values of all available pairwise distances of each type are shown in nucleotide substitutions per site with standard deviations. Sequences included in the calculation and exon sequence alignment are described in Figures 3 and 4. Distances shown in the ‘intron/all’ column included intron positions for active genes and all positions for the ‘intronless’ pseudogene (Ψ -*Acc-2*). Rates at ‘nonsynonymous’ and ‘synonymous’ positions of the pseudogene calculated as if it was an active gene are shown in parenthesis for comparison. Distances at nonsynonymous positions were not corrected for multiple substitutions and distances at synonymous positions and in introns were corrected for multiple substitutions using the Jukes-Cantor method as implemented by MEGA. Gaps were excluded only from pairwise comparisons. Relative rates for *Acc-1*, *Pgk-1* and *Pgk-2* were calculated assuming Pooideae-Panicoideae diverged 60 MYA (Introduction). Relative rates for *Acc-2* and Ψ -*Acc-2* were calculated based on the calculated rate for barley-*Triticum/Aegilops* (11.4 MYA). Relative synonymous *Acc-1*-*Acc-2* rate was calculated based on an average barley-*Triticum/Aegilops* rate for *Acc-1* and *Acc-2*. Values in the ‘Timeline’ column are averages of the relative rates calculated for individual genes (where available) excluding, for reasons discussed in the text, data points shown in brackets. Assumed (set) values of the relative rates are shown in brackets. The relative rates are expressed in million years. N.d., not determined.

		Nonsynonymous per site	Synonymous		Intron/all		Timeline relative
			per site	relative	per site	relative	
<i>Triticum</i>							
Rye	<i>Acc-1</i>	N.d.	0.065 ± 0.007	8.9	0.056 ± 0.003	8.0	7.2±1.6
<i>Secale</i>	<i>Acc-2I</i>	N.d.	0.094 ± 0.019	(11.5)	0.047 ± 0.003	(8.6)	
	Ψ - <i>Acc-2</i>	N.d.	(0.105 ± 0.025)	(7.2)	0.028 ± 0.005	6.9	
	<i>Pgk-1</i>	N.d.	0.064 ± 0.007	4.7	0.077 ± 0.005	7.6	
<i>Triticum/Aegilops</i>							
Barley (<i>Hordeum</i>)	<i>Acc-1</i>	0.003 ± 0.001	0.081 ± 0.010	11.1	0.080 ± 0.004	[11.4]	11.4±0.6
	<i>Acc-2I</i>	0.005 ± 0.001	0.093 ± 0.029	[11.4]	0.062 ± 0.006	[11.4]	[11.4]
	Ψ - <i>Acc-2</i>	(0.013 ± 0.002)	(0.166 ± 0.019)	[11.4]	0.046 ± 0.004	[11.4]	[11.4]
	<i>Pgk-1</i>	0.008 ± 0.002	0.165 ± 0.008	12.1	0.116 ± 0.007	[11.4]	[11.4]
	<i>Pgk-2</i>	0.005 ± 0.000	0.104 ± 0.012	10.9	0.161 ± 0.002	[11.4]	[11.4]
<i>L. rigidum</i> (<i>Lolium</i>)	<i>Acc-1</i>	0.011 ± 0.001	0.252 ± 0.010	34	0.243 ± 0.006	35	35±1
	<i>Acc-2I</i>	0.011 ± 0.001	0.297 ± 0.030	36	0.259 ± 0.005	(48)	
Pooideae							
Panicoideae	<i>Acc-1</i>	0.012 ± 0.002	0.439 ± 0.019	60	N.d.	N.d.	[60]
	<i>Pgk-1</i>	0.071 ± 0.001	0.820 ± 0.027	60	N.d.	N.d.	(set)
	<i>Pgk-2</i>	0.042 ± 0.001	0.572 ± 0.031	60	N.d.	N.d.	
Other rates							
	<i>Acc-2I</i> - <i>Acc-2II</i>		0.211 ± 0.021	26	0.189 ± 0.008	35	31±5
	<i>Acc-1</i> - <i>Acc-2</i>		0.981 ± 0.089	129	N.d.	N.d.	129
	grass <i>Pgk-1</i> -dicot <i>Pgk-1</i>		1.869 ± 0.266	137	N.d.	N.d.	146 ± 9
	grass <i>Pgk-2</i> -dicot <i>Pgk-2</i>		1.480 ± 0.145	155	N.d.	N.d.	

The phylogenetic tree based on exon positions (synonymous and nonsynonymous) of the *Pgk* genes was created by the neighbor-joining method with distances not corrected for multiple substitutions and gaps excluded only from pairwise comparisons (Figure 4). This analysis included fewer species than the *Acc*-based phylogenetic reconstruction. It has been suggested previously that duplication of the bacterial-type PGK gene occurred before the dicot-monocot separation (Martin and Schnarrenberger, 1997). The topology of this part of the *Pgk* tree is consistent with this suggestion, although it is not unambiguous

as indicated by low bootstrap values. Resolution of further questions about the sequence of early events in the evolution of plant PGK would require a better sampling that includes more sequences from dicots as well as sequences from monocots other than grasses. The relationships among the major lineages, Pooideae versus Panicoideae, barley versus rye and rye versus *Triticum/Aegilops* species are well supported. Several clades of *Triticum/Aegilops* species could be distinguished and they were similar to the *Triticum/Aegilops* clades found on the *Acc* tree (Figure 3). As discussed above, the copy number of *Pgk-2* genes in Triticeae is

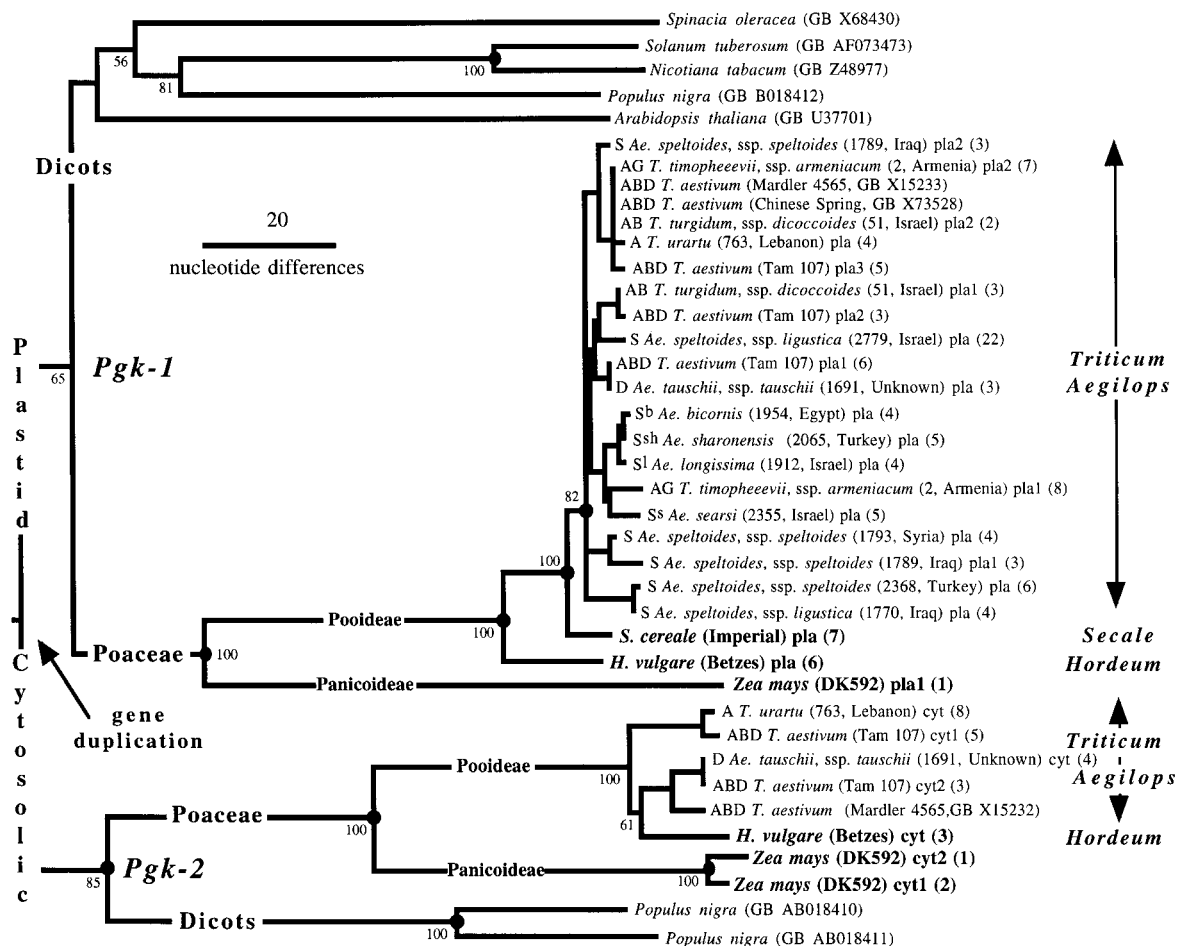


Figure 4. Phylogenetic relationships among plant PGKs. A phylogenetic tree was created by the neighbor-joining method with distances not corrected for multiple substitutions and gaps excluded only from pairwise comparisons. All exon positions were included in the calculation. Alignment length: 897 nucleotides (299 codons); outgroup: *Synechocystis* PGK gene (GenBank D90915); invariant positions: 398 (477 without outgroup); bootstrap values: % of 1000 trials (shown for selected nodes only). The sequence alignment was unambiguous with a single 3-nucleotide insertion in the maize cytosolic sequences. Taxon names include genome composition (*Triticum/Aegilops* species), species name, accession number/cultivar name, geographical origin (*Triticum/Aegilops* species), sequence name/number and number of clones analyzed. Taxon names of grass species other than *Triticum* and *Aegilops* are shown in bold. The genome composition of *Triticum/Aegilops* species is shown with abbreviated names listed in Table 1. Sequences obtained from GenBank are identified by their accession numbers (GB). New gene sequences were deposited in GenBank under the following accession numbers: *Pgk-1* genes, AF343474–AF343495; *Pgk-2* genes, AF343447–AF343453. Dots indicate nodes relevant for the analysis presented in this paper which showed strong statistical support by the neighbor-joining method including synonymous sites only and by the maximum parsimony method.

uncertain, preventing the establishment of orthologous relationships among the sequences of the *Pgk-2* clade.

It is important to emphasize that, as indicated in Figures 3 and 4, the topology of the neighbor-joining tree at some major nodes is not only supported by statistical analysis (high bootstrap values) but is congruent with the topologies of phylogenetic trees reconstructed by heuristic maximum parsimony searches. Reconstructing phylogenetic relationships among *Triticum* and *Aegilops* species, based on both

exon and intron sequence analysis of *Acc* and *Pgk* genes, will be presented elsewhere.

Nucleotide substitution rates at *Acc* and *Pgk* loci

Average nucleotide substitution rates at synonymous, nonsynonymous and intron sites between several major plant lineages for *Acc-1*, *Acc-2*, *Pgk-1* and *Pgk-2* genes, as well as the nucleotide substitution rate for the Ψ -*Acc-2* pseudogene, were calculated (Table 3) using multiple pairwise distances between sequences

of well-defined and statistically supported clades identified in Figures 3 and 4. Distances at nonsynonymous positions were not corrected for multiple substitutions and distances at synonymous positions and in introns were corrected for multiple substitutions using the Jukes-Cantor method. In these calculations, gaps were excluded only from pairwise comparisons. Rates at 'nonsynonymous' and 'synonymous' positions of the pseudogene (Ψ -*Acc-2*), calculated as if it was an active gene, are shown in parentheses for comparison. Finally, standard deviations were calculated for each type of pairwise distances. With a few exceptions, these standard deviations did not exceed 20%.

The nucleotide substitution rates show a consistent pattern. First, nonsynonymous rates are 10 to 40 times lower than the corresponding synonymous and intron rates. This is in agreement with earlier observations for other nuclear genes in grasses (Gaut, 1998). Second, for each gene, synonymous rates are very similar to substitution rates at all intron sites. Third, synonymous rates as well as intron substitution rates between grass lineages are very similar for both *Acc* and *Pgk* genes.

Substitution rates at all sites of the pseudogene are 2–3 times lower than synonymous and intron rates for *Acc* and *Pgk* genes. At the same time, 'nonsynonymous' substitution rates calculated for the pseudogene, as if it was an active gene without introns, are 10-fold lower than substitution rates at synonymous sites, a correlation typical of active genes such as *Acc* and *Pgk*.

The nucleotide substitution rates were also compared on a relative basis (Table 3), setting the clock for the divergence of Pooideae and Panicoideae subfamilies at 60 MYA to reflect approximately what is known about the historical timeline of grass evolution (Introduction) and to enable comparisons with the results of other studies for which similar calculations have been made (Wolfe *et al.*, 1987; Wolfe *et al.*, 1989; Gaut, 1998). Most importantly, such relative rates allow comparisons between multiple genes and lineages by eliminating some locus-specific effects. They enable establishment of a relative sequence of evolutionary events based on analysis of multiple loci. The relative rates discussed in this paper are expressed in million years.

In general, the relative rates calculated for synonymous and intron positions of *Acc* and *Pgk* genes are very similar. The relative rate comparison also indicates that the pseudogene sequence can be used for the phylogenetic analysis. Some of these calculations are, however, less reliable (they are shown

in parenthesis in Table 3). First, there is a significant variation of synonymous rates calculated between rye and *Triticum* species. Second, calculations for the *Acc-2* gene of rye and *Triticum* species are not reliable because of the unresolved questions of the gene copy number and orthology (discussed above). Third, the *Acc-2* intron substitution rate between *Lolium* and *Triticum/Aegilops* lineages is questionable because of uncertainties in the nucleotide sequence alignment (Faris *et al.*, 2001).

The relative timeline of evolutionary events presented in Table 3 derived from sequence comparisons at synonymous and intron sites of two or three different genes was very consistent, except for the rye-*Triticum* divergence based on synonymous substitution rates. This divergence time was 7.2 ± 1.6 when all available single values were averaged. When the highly variable synonymous rates were excluded, the average divergence time, 7.5 ± 0.6 , was very similar but had a lower standard deviation. The calculated divergence times of barley and *Lolium* from *Triticum/Aegilops* were 11.4 ± 0.6 and 35 ± 1 , respectively. An average grass-monocot divergence time derived from *Pgk* synonymous substitution rates was 146 ± 9 . Our estimate of the divergence time of barley is very similar to an earlier estimate of 12 MYA based on the molecular clock set exactly as in this study but on the sequences of different genes (Wolfe *et al.*, 1989). Our estimate of the divergence of grasses and dicots is lower than an earlier estimate of some 200 MYA (Wolfe *et al.*, 1989).

On an absolute basis, the nucleotide substitution rates presented in Table 3 are within the range of such rates calculated for other plant nuclear genes (Wolfe, *et al.*, 1987; Gaut, 1998). The rates at synonymous positions and in introns are $2-8 \times 10^{-9}$ nucleotide substitutions per site per year, assuming the divergence time between Pooideae and Panicoideae to be 60 MYA.

No dramatic synonymous rate heterogeneity between the major grass lineages, outside of the *Triticum/Aegilops* clade, was detected using a simple relative-rate test (Li and Bousquet, 1992). Some heterogeneity was observed for the *Acc* genes for major Triticeae lineages.

Acc- and Pgk-based analysis of plant phylogenies

The phylogenetic analysis presented in this paper focused on the Triticeae tribe but also included other grasses as well as some dicot species. These results

indicate that the approach can be applied to study various groups of plants, monocots including grasses, as well as dicots.

The *Lolium* example shows that the analysis based on rates at synonymous positions of the *Acc* genes can be extended to analyze evolution of grasses above the tribe level. Including the *Pgk* genes in this analysis would make it even more significant. For species at these levels of genetic relatedness, intron sequences are too divergent to allow reliable multiple-sequence alignments to be created. Exon-based phylogenetic analysis above the family level can be expected to be less reliable, as the number of nucleotide changes per site at synonymous positions reaches values above 0.7 and calculations of rates corrected for multiple substitutions becomes difficult. Some results of phylogenetic analysis covering longer evolutionary distances are shown in Figures 3 and 4 and in Table 3. Phylogenetic analysis over long evolutionary distances based on the *Acc* and *Pgk* sequences and for plants other than grasses was beyond the scope of this study.

The barley and rye example proves that the two-gene approach (*Acc-1* and *Pgk-1*) is suitable to study evolution below the tribe level. For more divergent species, synonymous positions, intron sites, and pseudogene sites can be used. Intron sequences will be more useful for closely related species as more nucleotide changes can be analyzed. The observed number of nucleotide differences in cloned fragments of *Acc* and *Pgk* genes was about 2 and 5 changes per million years at synonymous positions and in introns, respectively, with the clock calibrated as described above. The high consistency of the intron data is especially encouraging, for example, for the more detailed analysis of the species of the *Triticum/Aegilops* complex which will be presented elsewhere. However, even the exon-based phylogenies (Figures 3 and 4) provide important information on the pattern of evolution of these species, such as the origin of the A genome in tetraploid and hexaploid wheats (*T. urartu*) and the origin of the D genome in hexaploid wheat (*A. tauschii*).

Several potential sources of error in gene sequence-based phylogenetic analysis were identified, including difficulties in establishing orthologous relationships and in creating meaningful sequence alignments, greater variability in substitution rates for some lineages and some genes, and too few or too many nucleotide changes for species at the two ends of the genetic relatedness scale. To overcome these problems, a multi-gene approach was needed.

Discussion

We have established a two-gene system to study grass evolution based on nuclear genes encoding the eukaryotic-type (multi-domain) plastid ACCase (*Acc-1* genes) and the prokaryotic (endosymbiont)-type plastid PGK (*Pgk-1* genes). The single-copy nature of these genes observed in most grass species analyzed so far is essential for the establishment of orthologous relationships between the genes. This is critical for the use of gene sequences to reveal phylogenetic relationships between species with a high confidence level. Genes encoding cytosolic ACCase and PGK (*Acc-2* and *Pgk-2*, respectively) can serve as an additional source of information in cases where the multiple-copy character of these genes is understood and does not affect the interpretation. The *Acc-2*-related pseudogene also serves as a source of phylogenetic information for the Triticeae tribe.

The ACCase and PGK genes included in this study complement each other because of their different evolutionary origin, the dual subcellular localization of their products, their involvement in different metabolic processes and their presumed different chromosome location. Corroborating evidence from independent data sets obtained for genes with such different characteristics adds significantly to the confidence level of any conclusions reached. The method has inherent limitations as to the relatedness of species that can be analyzed. For more divergent species, from the tribe to the subfamily level, reliable alignment of intron sequences becomes a problem. Exon sequences can then be used to address some of the questions by the analysis of nucleotide substitution rates at synonymous positions. Phylogenetic analysis over even longer evolutionary distances requires careful assessment of synonymous and nonsynonymous nucleotide substitution rates and their heterogeneity in different lineages. At the other end of the spectrum, the method will fail for very closely related species, for different accessions of the same species or populations, because there will not be enough nucleotide changes to count, even in introns. The method is suitable for analysis within the window defined by these two extremes.

Our data are consistent with the known facts on the origin of ACCase and PGK in grasses, eukaryotic versus prokaryotic, respectively (see Introduction). Furthermore, the *Acc*- and *Pgk*-based reconstruction of the phylogenetic relationships between some grass lineages is in full agreement with the known milestones of grass evolution. Finally, nucleotide substitution

rates at synonymous and nonsynonymous positions, and in introns of *Acc* and *Pgk* genes, are similar to each other and are within the range found for other nuclear genes. These are all important check points.

Neighbor-joining trees, calculated including all exon positions (synonymous and nonsynonymous), were selected for the presentation to provide an overview of all available information on plant multi-domain ACCase genes (Figure 3) and all plant PGK genes (Figure 4). The key steps in evolution of these two enzymes in plants, revealed by our exon sequence-based phylogenetic analysis, are consistent with earlier findings (Introduction, Figure 1). Gene duplication, giving rise to the multi-domain plastid ACCase isozyme in grasses, occurred early in the evolution of the family, well before the divergence of the major grass subfamilies. A similar gene duplication occurred independently in *Brassica napus*, leading to a new plastid ACCase isozyme in this species.

It was suggested that plant PGKs (cytosolic and chloroplast) are more closely related to the cyanobacterial PGK than to PGKs from other eukaryotes (Martin and Schnarrenberger, 1997). Contrary to the history of the eukaryotic-type ACCase gene, duplication of the bacterial-type PGK gene must have occurred before the dicot-monocot separation. This is in agreement with the phylogeny shown in Figure 4 although the topology of this part of the tree is not unambiguous.

The phylogenetic relationships among major grass lineages deduced from nucleotide sequence comparisons of the two genes are consistent with each other, with grass taxonomy and with the milestones of grass evolution revealed by other methods (Wolfe *et al.*, 1989; Hsiao *et al.*, 1994; Barker and Linder, 1995; Clark *et al.*, 1955; Bennetzen and Kellogg, 1997; Kellogg, 1998; Soreng and Davis, 1998). These milestones include divergence of the Pooideae and Panicoideae subfamilies, divergence of *Panicum* and *Zea* in the Panicoideae subfamily, and a series of events leading to the *Triticum/Aegilops* complex, and divergence of *Lolium*, *Hordeum* and *Secale*. These relationships are supported by phylogenies derived by different methods. The results of the phylogenetic analysis of rye and barley based on *Acc-2* and *Pgk-2* genes are difficult to interpret because gene copy number and orthologous relationships in these cases have not been established.

Other well-resolved clades include the two major *Acc-2* clades identified previously (Faris *et al.*, 2001) and the partially processed pseudogene (Ψ -*Acc-2*)

clade. The topology of the phylogenetic tree (Figure 3) suggests that the divergence of the two major *Acc-2* clades occurred after the divergence of *Lolium* from the *Triticum/Aegilops* lineage. The pseudogene was probably formed at approximately the same time, although this conclusion may not be accurate because of the different nature of the genes being compared.

On a relative basis, with the divergence of the Pooideae and Panicoideae subfamilies set at 60 MYA, events leading to the *Triticum/Aegilops* complex occurred at the following intervals: divergence of *Lolium* at 35 MYA, divergence of *Hordeum* at 11 MYA and divergence of *Secale* at 7 MYA. On the same scale, gene duplication leading to the multi-domain plastid ACCase in grasses occurred at 129 MYA, divergence of grass and dicot plastid PGK genes at 137 MYA, and divergence of grass and dicot cytosolic PGK genes at 155 MYA. Our estimate of the divergence time of *Hordeum* is very similar to an earlier estimate of 12 MYA based on the molecular clock set exactly as in this study but on sequences of different genes (Wolfe *et al.*, 1989). On the other hand, if the divergence of grass and dicot *Pgk* genes reflects divergence of monocots and dicots, our estimate for the timing of this event, 137–155 MYA, is below an earlier estimate of about 200 MYA (Wolfe *et al.*, 1989) but could still be reconciled with the sparse fossil records (Introduction) and with the proposed timing of the *Acc* gene duplication in grasses at 129 MYA. The absolute divergence times calculated from a very crude fossil-based estimate of the divergence time of the major grass subfamilies (Introduction) used to calibrate the molecular clock have to be treated with caution. On a relative basis, however, they probably reflect quite well the sequence of events leading to the wheat lineage.

The significant amount of information for the Triticeae tribe will allow us to revisit the problem of phylogenetic relationships among wheat and its relatives, and some unresolved problems of the origin of their genomes. Good understanding of the evolution of the *Acc* and *Pgk* gene families and the congruence of different phylogenetic trees at major branch points makes the approach suitable for the phylogenetic analysis of the grass family as well as other plants.

Acknowledgements

We thank W. J. Buikema for his help with the sequencing part of the project and J. Wendel for comments

on the manuscript. DNA sequencing was performed by the University of Chicago Cancer Center DNA Sequencing Facility. This work was supported by a gift from Monsanto.

References

- Barker, N.P. and Linder, H.P. 1995. Polyphyly of Arundinoideae (Poaceae): evidence from *rbcl* sequence data. *Syst. Bot.* 20: 423–435.
- Bennetzen, J.L. and Kellogg, E.A. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9: 1509–1514.
- Catalan, P., Kellogg, E.A. and Olmsted, R.G. 1997. Phylogeny of Poaceae subfamily Pooideae based on chloroplast *ndhF* gene sequence. *Mol. Phylogenet. Evol.* 8: 150–166.
- Chen, M., SanMiguel, P., DeOliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A. and Bennetzen, J.L. 1998. Microcolinearity in *sh2*-homologous regions of maize, rice and sorghum genomes. *Proc. Natl. Acad. Sci. USA* 95: 3431–3435.
- Christopher, J.T. and Holtum, J.A.M. 2000. Dicotyledons lacking the multisubunit form of the herbicide-target enzyme acetyl coenzyme A carboxylase may be restricted to the family Geraniaceae. *Aust. J. Plant Physiol.* 27: 845–850.
- Clark, L.G., Kobayashi, M., Mathews, S., Spangler, R.E. and Kellogg, E.A. 2000. The Puelioideae, a new subfamily of Poaceae. *Syst. Bot.* 25: 181–187.
- Clark, L.G., Zhang, W. and Wendel, J.F. 1995. A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Syst. Bot.* 20: 436–460.
- Clegg, M.T. 1997. Plant genetic diversity and the struggle to measure selection. *J. Hered.* 88: 1–7.
- Clegg, M.T., Cummings, M.P. and Durbin, M.L. 1997. The evolution of plant nuclear genes. *Proc. Natl. Acad. Sci. USA* 94: 7791–7798.
- Cox, T.S. 1998. Deepening the wheat gene pool. *J. Crop Prod.* 1: 1–25.
- Crepet, W.L. and Feldman, G.D. 1991. The earliest remains of grasses in the fossil record. *Am. J. Bot.* 78: 1010–1–14.
- Cummings, M.P. and Clegg, M.T. 1998. Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. *Proc. Natl. Acad. Sci. USA* 95: 5637–5642.
- Devos, K.M. and Gale, M.D. 1997. Comparative genetics in the grasses. *Plant Mol. Biol.* 35: 3–15.
- Doyle, J.J. and Gaut, B.S. 2000. Evolution of genes and taxa: a primer. *Plant Mol. Biol.* 42: 1–23.
- Dvorak, J. 1998. Genome analysis in the *Triticum-Aegilops* alliance. *Cytogenet. Evol.* 1: 8–11.
- Dvorak, J., DiTerlizzi, P., Zhang, H.-B. and Resta, P. 1993. The evolution of polyploid wheats: identification of the A genome donor species. *Genome* 36: 21–31.
- Dvorak, J., Luo, M.-C., Yang, Z.-L. and Zhang, H.-B. 1998. The structure of the *Aegilops tauschii* gene pool and the evolution of hexaploid wheat. *Theor. Appl. Genet.* 97: 657–670.
- Dvorak, J. and Zhang, H.B. 1990. Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. *Proc. Natl. Acad. Sci. USA* 87: 9640–9644.
- Eyre-Walker, A., Gaut, R.L., Hilton, H., Feldman, D.L. and Gaut, B.S. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* 95: 4441–1116.
- Faris, J., Sirikhachornkit, A., Haselkorn, R., Gill, B. and Gornicki, P. 2001. Chromosome mapping and phylogenetic analysis of the cytosolic acetyl-CoA carboxylase loci in wheat. *Mol. Biol. Evol.* 18, 1720–1733.
- Faris, J.D., Haen, K.M. and Gill, B.S. 2000. Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* 154: 823–835.
- Gandolfo, M.A., Nixon, K.C., Crepet, W.L., Stevenson, D.W. and Friss, E.M. 1998. Oldest known fossils of monocotyledons. *Nature* 394: 532–533.
- Gaut, B.S. 1998. Molecular clocks and nucleotide substitution rates in higher plants. *Evol. Biol.* 30: 93–120.
- Gaut, B.S., Clark, L.G., Wendel, J.F. and Muse, S.V. 1997. Comparison of the molecular evolutionary process at *rbcl* and *ndhF* in the grass family (Poaceae). *Mol. Biol. Evol.* 14: 769–777.
- Gaut, B.S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* 94: 6809–6814.
- Gaut, B.S., Morton, B.R., McCaig, B.C. and Clegg, M.T. 1996. Substitution rate comparison between grasses and palms: synonymous rate difference at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcl*. *Proc. Natl. Acad. Sci. USA* 93: 10274–10278.
- Gaut, B.S., Tredway, L.P., Kubik, C., Gaut, R.L. and Meyer, W. 2000. Phylogenetic relationship and genetic diversity among members of the *Festuca-Lolium* complex (Poaceae) based on ITS sequence data. *Plant Syst. Evol.* 224: 33–53.
- Gimenez-Espinosa, R., Plaisance, K.L., Plank, D.W., Gronwald, J.W. and De Prado, R. 1999. Propaquizafop absorption, translocation, metabolism and effect on acetyl-CoA carboxylase isoforms in chickpea (*Cicer arietinum* L.). *Pestic. Biochem. Physiol.* 65: 140–150.
- Gornicki, P., Fans, J., King, I., Podkowinski, J., Gill, B. and Haselkorn, R. 1997. Plastid-localized acetyl-CoA carboxylase of bread wheat is encoded by a single gene on each of the three ancestral chromosome sets. *Proc. Natl. Acad. Sci. USA* 94: 1417–14185.
- Gornicki, P., Podkowinski, J., Scappino, L.A., DiMaio, J., Ward, E. and Haselkorn, R. 1994. Wheat acetyl-CoA carboxylase: cDNA and protein structure. *Proc. Natl. Acad. Sci. USA* 91: 6860–6864.
- Hilton, H. and Gaut, B.S. 1998. Speciation and domestication in maize and its wild relatives: evidence from the *Globulin-1* gene. *Genetics* 150: 863–872.
- Hsiao, C., Chatterton, N.J., Asay, K.H. and Jensen, K.B. 1994. Phylogenetic relationships of 10 grass species: an assessment of phylogenetic utility of the internal transcribed spacer region in nuclear ribosomal DNA in monocots. *Genome* 37: 112–120.
- Hsiao, C., Chatterton, N.J., Asay, K.H. and Jensen, K.B. 1995. Molecular phylogeny of the Pooideae (Poaceae) based on nuclear rDNA (ITS) sequences. *Theor. Appl. Genet.* 90: 389–398.
- Hsiao, C., Jacobs, S.W.L., Barker, N.P. and Chatterton, N.J. 1998. A molecular phylogeny of the subfamily Arundinoideae (Poaceae) based on sequences of rDNA. *Aust. Syst. Bot.* 11: 41–52.
- Inclendon, B.J. and Hall, J.C. 1997. Acetyl-coenzyme A carboxylase: quaternary structure and inhibition by graminicidal herbicides. *Pestic. Biochem. Physiol.* 57: 255–271.
- Kellogg, E.A., Appels, R. and Mason-Gamer, R.J. 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae). *Syst. Bot.* 21: 321–347.
- Kellogg, E.A. 1998. Relationships of cereal crops and other grasses. *Proc. Natl. Acad. Sci. USA* 95: 2005–2010.
- Li, P. and Bousquet, J. 1992. Relative-rate test for nucleotide substitutions between two lineages. *Mol. Biol. Evol.* 9: 1185–1189.

- Martin, W. and Schnarrenberger, C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* 32: 1–18.
- Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. and Kowallik, K.V. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393: 162–165.
- Mason-Gamer, R.J., Weil, C.F. and Kellog, E.A. 1998. Granule-bound starch synthase: structure, function and phylogenetic utility. *Mol. Biol. Evol.* 15: 1658–1673.
- Mathews, S., Tsai, R.C. and Kellog, E.A. 2000. Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. *Am. J. Bot.* 87: 96–107.
- McFadden, E.S. and Sears, E.R. 1946. The origin of *Triticum* speltoïdes and its free-threshing hexaploid relatives. *J. Hered.* 37: 81–89.
- Muse, S.V. 2000. Examining rates and patterns of nucleotide substitutions in plants. *Plant Mol. Biol.* 42: 25–43.
- Muse, S.V. and Gaut, B.S. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* 146: 393–399.
- Petersen, G. and Seberg, O. 1997. Phylogenetic analysis of the Triticeae (Poaceae) based on *rpoA* sequence data. *Mol. Phylog. Evol.* 7: 217–230.
- Podkowinski, J., Sroga, G.E., Haselkorn, R. and Gornicki, P. 1996. Structure of a gene encoding a cytosolic acetyl-CoA carboxylase of hexaploid wheat. *Proc. Natl. Acad. Sci. USA* 93: 1870–1874.
- Pryer, K.M., Schneider, H., Smith, A.R., Cranfill, R., Wolf, P.G., Hunt, J.S. and Sipes, S.D. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* 409: 618–622.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. 1999. The paleontology of intergene retrotransposons of maize. *Nature Genet* 20: 43–45.
- Sasaki, Y., Konishi, T. and Nagano, Y. 1995. The compartmentation of acetyl-coenzyme a carboxylase in plants. *Plant Physiol.* 108: 445–449.
- Schulte, W., Topfer, R., Stracke, R., Schell, J. and Martini, N. 1997. Multi-functional acetyl-CoA carboxylase from *Brassica napus* is encoded by a multi-gene family: indication for plastidic localization of at least one isoform. *Proc. Natl. Acad. Sci. USA* 94: 3465–3470.
- Soltis, E.D. and Soltis, P.S. 2000. Contributions of plant molecular systematics to studies of molecular evolution. *Plant Mol. Biol.* 42: 45–75.
- Soreng, R.J. and Davis, J.I. 1998. Phylogenetics and character evolution in the grass family (Poaceae): simultaneous analysis of morphological and chloroplast DNA restriction site data sets. *Bot. Rev.* 64: 1–85.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22: 4673–4680.
- Tsunewaki, K. 1991. A historical review of cytoplasmic studies in wheat. In: T. Sasakuma and T. Kinoshita (Eds) *Nuclear and Organelle Genomes of Wheat Species*, Kihara Memorial Foundation, Yokohama) pp. 16–28.
- Wang, G.-Z., Miyashita, N.T. and Tsunewaki, K. 1997. Plasmon analyses of *Triticum* (wheat) and *Aegilops*: PCR-single-strand conformational polymorphism (PCR-SSCP) analyses of organellar DNAs. *Proc. Natl. Acad. Sci. USA* 94: 14570–14577.
- Wolfe, K.H., Gouy, M.L., Yang, Y.W., Sharp, P.M. and Li, W.H. 1989. Date of the monocot dicot divergence estimated from chloroplast DNA-sequence. *Proc. Natl. Acad. Sci. USA* 86: 6201–6205.
- Wolfe, K.H., Li, W.H. and Sharp, P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* 84: 9054–9058.
- Zhang, W.P. 2000. Phylogeny of the grass family (Poaceae) from rpl16 intron sequence data. *Mol. Phylogenet. Evol.* 15: 135–146.