New Phytologist

Methods

The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies

Yao-Wu Yuan, Chang Liu, Hannah E. Marx and Richard G. Olmstead

Department of Biology, University of Washington, Seattle, WA 98195, USA

Author for correspondence:
Yao-Wu Yuan
Tel: +1 206 616 7156
Fax: +1 206 685 1728
Email: colreeze@u.washington.edu

Received: *9 October 2008*
Accepted: *25 November 2008*

New Phytologist (2009) **182**: 272–283
doi: 10.1111/j.1469-8137.2008.02739.x

Key words: rapidly evolving, intronless, nuclear gene loci, plant phylogenetics, pentatricopeptide repeat (PPR) genes.

Summary

• Despite the paramount importance of nuclear gene data in plant phylogenetics, the search for candidate loci is believed to be challenging and time-consuming. Here we report that the pentatricopeptide repeat (PPR) gene family, containing hundreds of members in plant genomes, holds tremendous potential as nuclear gene markers.

• We compiled a list of 127 PPR loci that are all intronless and have a single orthologue in both rice (*Oryza sativa*) and *Arabidopsis thaliana*. The uncorrected p-distances were calculated for these loci between two *Arabidopsis* species and among three Poaceae genera. We also selected 13 loci to evaluate their phylogenetic utility in resolving relationships among six Poaceae genera and nine diploid *Oryza* species.

• PPR genes have a rapid rate of evolution and can be best used at intergeneric and interspecific levels. Although with substantial amounts of missing data, almost all individual data sets from the 13 loci generate well-resolved gene trees.

• With the unique combination of three characteristics (having a large number of loci with established orthology assessment, being intronless, and being rapidly evolving), the PPR genes have many advantages as phylogenetic markers (e.g. straightforward alignment, minimal effort in generating sequence data, and versatile utilities). We perceive that these loci will play an important role in plant phylogenetics.

Introduction

The paramount importance of single- or low-copy nuclear gene sequence data in plant phylogenetic studies has been elaborated extensively in reviews (Sang, 2002; Small *et al.*, 2004; Hughes *et al.*, 2006) and has resonated frequently in empirical studies (e.g. Alvarez *et al.*, 2008; Yuan & Olmstead, 2008; Steele *et al.*, 2008) and commentaries (Mort & Crawford, 2004; Crawford & Mort, 2004). However, nuclear loci that can be routinely employed for most angiosperm groups are scarce and the search for candidate nuclear loci is still a challenging and continuing endeavour (Mort & Crawford, 2004). Earlier studies (Strand *et al.*, 1997; Bailey & Doyle, 1999; Olsen & Schaal, 1999; Small & Wendel, 2000; Tank & Sang, 2001; Howarth & Baum, 2002) mainly took the 'low-copy nuclear gene approach' (Hughes *et al.*, 2006) by selecting a well-characterized gene or small gene family and testing the phylogenetic utility of these selected loci in a particular study group. This approach has often proved effective, but it is restricted to a single locus or a small number of loci, which are insufficient to resolve many plant phylogenetic problems, especially at lower taxonomic levels.

With the rapid development of complete genome sequence and expressed sequence tag (EST) databases emerges the conserved orthologue set (COS) approach (Fulton *et al.*, 2002; Wu *et al.*, 2006; Padolina, 2006; Chapman *et al.*, 2007; Alvarez *et al.*, 2008). By comparisons of EST and/or complete genome sequences between model organisms, COS markers can be identified and used to develop sets of primers that amplify putative orthologue sequences across the taxa of interest for phylogenetic investigations. As whole genome sequence and EST databases continue growing on a daily basis, the COS approach holds great promise in screening a large number of nuclear loci for phylogenetic studies. However, there is one major problem with the COS approach. Given the vast number of putative COS markers this approach often produces (e.g.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Wu *et al.* (2006) found 2869 single-copy orthologues shared by euasterids) and little prior knowledge about these loci, it often requires labour-intensive preliminary work to screen loci for their appropriate phylogenetic utility in a specific study group. For instance, after examining 141 nuclear primer combinations designed from such COS markers (Padolina, 2006), Steele *et al.* (2008) found only three phylogenetically informative loci for resolving interspecies relationships in the genus *Psiguria* (Cucurbitaceae) and two loci for resolving intergeneric relationships in the family Geraniaceae. In the end these authors concluded, 'In any case, identifying phylogenetically informative LCN [low copy nuclear] markers remains a time-consuming endeavor ...'.

With the hope of identifying numerous nuclear loci for general plant phylogenetic investigations that require little preliminary work to use, we took an integrative approach that combines the advantages of both approaches mentioned above and avoids the disadvantages of each. The general idea is to identify a large number of putative orthologous loci that are well characterized and information-rich. The availability of such online databases as POGs/PlantRBP (Walker *et al.*, 2007; http://plantrbp.uoregon.edu/) makes this strategy straightforward. The POGs/PlantRBP assigns proteins (with corresponding gene loci) in the rice (*Oryza sativa*) and *Arabidopsis thaliana* proteomes to putative orthologous groups (POGs) via a 'mutual-best-hits' strategy (Walker *et al.*, 2007; see also http://plantrbp.uoregon.edu/about-pogs.php for a schematic illustration of this strategy). Among the assigned proteins, predicted RNA-binding proteins (RBPs) are particularly well annotated. By mining this database, we found that the enormous pentatricopeptide repeat (PPR) gene family, coding for RNA-binding proteins, may have tremendous potential in plant phylogenetic applications.

The PPR gene family contains *c.* 450 members in *A. thaliana*, 477 in *O. sativa*, and 103 in *Physcomitrella patens*, whereas there are only a handful of loci in the genomes of green algae and nonplant eukaryotic organisms (Lurin *et al.*, 2004; O'Toole *et al.*, 2008), and virtually none in prokaryotic genomes (Lurin *et al.*, 2004; Pusnik *et al.*, 2007). The *A. thaliana* PPR genes are more or less evenly distributed throughout the 10 chromosome arms (Lurin *et al.*, 2004). An interesting observation is that *c.* 80% of the PPR genes in both *A. thaliana* and rice are intronless (Lurin *et al.*, 2004; O'Toole *et al.*, 2008). PPR proteins are characterized by 2–26 tandem repeats of a highly degenerate 35 amino acid motif, and divided into two subfamilies and four subclasses based on their conserved C-terminal domain structure (Lurin *et al.*, 2004). These proteins are targeted to organelles (i.e. mitochondria and plastids) and involved in many post-transcriptional processes undergone by organellar transcripts, including splicing, editing, processing, and translation (reviewed in Delannoy *et al.*, 2007). The presence of one of the four subclasses (i.e. the DYW subclass) is strictly correlated with the existence of RNA editing in land plants (Salone *et al.*, 2007; Rudinger *et al.*, 2008). Together with other evidence, this led Salone *et al.* (2007) to propose that the DYW domain found exclusively in PPR proteins is the catalytic domain conducting the enigmatic organelle RNA editing process.

It might be counterintuitive at first glance that such a huge gene family with most members being intronless can have any phylogenetic utility. (1) Won't the massive number of gene copies make orthology assessment extremely difficult? (2) Can these protein-coding sequences provide sufficient variation to address phylogenetic problems at lower taxonomic levels? A phylogenetic analysis including all of the rice and *A. thaliana* PPR genes revealed that an extraordinarily large proportion of these genes form well-supported pairs that are probably *A. thaliana* and rice orthologues (O'Toole *et al.*, 2008), which suggested that most of the PPR gene loci predate the divergence of eudicots and monocots with few duplications since then. This is consistent with the finding from the POGs/PlantRBP database (Walker *et al.*, 2007) that the majority of PPR genes have a single orthologue in both *A. thaliana* and rice genomes. These results suggest that evaluating the orthology of most PPR genes should not be difficult. In addition, considering that PPR proteins probably function as RNA-binding molecules in a sequence-specific manner (Delannoy *et al.*, 2007), they may have a rapid rate of evolution to adjust to changes in the targeted RNA species. This means that, within a putative orthologue group, sequences could be divergent enough to provide variation that could be used in resolving relationships at lower taxonomic levels, despite the lack of rapidly evolving introns. In fact, the absence of introns can be a great advantage for many phylogenetic applications such as resolving intergeneric relationships (see the Discussion). It is these three appealing characteristics – a huge number of loci, but an easily assessed orthology; an absence of introns; the likelihood of a rapid rate of evolution – that stimulate us to explore the potential of PPR genes in plant phylogenetic studies.

In this paper we have aimed: (1) to compile a comprehensive list of PPR genes that are intronless and have a single orthologue in both *A. thaliana* and rice; (2) to compute the pairwise distance at these compiled loci between two *Arabidopsis* species (*A. thaliana* and *Arabidopsis lyrata*) and among three Poaceae genera (*Oryza*, *Zea* and *Sorghum*) in order to obtain a cursory estimation of variation at each locus; and (3) to select a small proportion of these loci to evaluate their utility in resolving relationships among six Poaceae genera and nine diploid *Oryza* species. There is a large amount of genomic sequence data for 12 *Oryza* species (not including cultivated rice, *O. sativa*) in GenBank as trace archives from the *Oryza* Map Alignment Project (OMAP; http://www.omap.org/; Wing *et al.*, 2005). Eight of them are diploid species, representing all six diploid *Oryza* genome types (Nayar, 1973; Aggarwal *et al.*, 1997; Ge *et al.*, 1999). There are also quite abundant ESTs of three other Poaceae genera, *Triticum* (*Triticum aestivum*), *Hordeum* (*Hordeum vulgare*) and *Saccharum* (*Saccharum*

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

officinarum), available from The Institute for Genomic Research (TIGR) Plant Transcript Assemblies database (http://plantta.tigr.org/; Childs *et al.*, 2007). These publicly available genomic data provide a good opportunity to examine the phylogenetic utility of PPR gene loci.

Materials and Methods

Locus screening, sequence retrieval and annotation

We retrieved all the *Arabidopsis thaliana* (L.) Heynh. PPR gene family members with their putative orthologues in rice (*Oryza sativa* L.) from the POGs/PlantRBP database (Walker *et al.*, 2007; http://plantrbp.uoregon.edu/) by searching 'At*' by gene AND 'PPR' by domain. The *A. thaliana* PPR genes are assigned to 418 POGs, most of which contain a single locus in both rice and *A. thaliana*. The results were downloaded to an Excel file (see the Supporting Information, Table S1). We then screened loci for phylogenetic utility in a stepwise manner.

(1) If the POG contains a single locus in both rice and *A. thaliana*, continue to (2); otherwise, abandon.

(2) If the gene pair in the remaining POGs is marked as 'well supported' in POGs/PlantRBP, continue to (3); otherwise, abandon. When building the POGs/PlantRBP database, Walker *et al.* (2007) took a phylogenetic approach to evaluate POGs assigned by the 'mutual-best-hit' method. The top blast hits with > 50% coverage (either hit/query or query/hit) for each protein were retrieved to produce a multiple alignment and corresponding guide tree. Only those POG assignments supported by the tree topology were marked as 'well supported'.

(3) If the *A. thaliana* gene in the remaining POGs is intronless, continue to (4); otherwise, abandon. This was done by comparing the *A. thaliana* locus ID against the '*Arabidopsis* intronless gene list' from Jain *et al.* (2008).

(4) Follow the POGs/PlantRBP link for each rice locus to TIGR (http://www.tigr.org/) and for each *A. thaliana* locus to The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/). These linked pages have comprehensive gene descriptions for the corresponding loci. If the rice gene in the remaining POGs is intronless, continue to (5) with sequence retrieval and annotation; otherwise, abandon.

(5) For each remaining POG, download the rice coding sequence (CDS) from TIGR and the *A. thaliana* CDS from TAIR. In many cases the CDS is the same as the cDNA as well as the genomic DNA sequence, while in other cases the genomic DNA sequence is longer than the CDS by regulatory regions at the 5′ and/or 3′ ends. Blast the *A. thaliana* sequence against *Arabidopsis lyrata* (L.) O'Kane & Al-Shenbaz trace archives using the MEGABLAST program and default parameters (http://blast.ncbi.nlm.mih.gov/). Sequences with an *E*-value < e^{-100} were downloaded as SCF files, and were edited and assembled using SEQUENCHER 4.7 (Gene Codes Corporation, Ann Arbor, MI, USA). The positions of start and stop codons were determined by comparison with the *A. thaliana* sequence. Similarly, sequences of *Sorghum bicolor* (L.) Moench and *Zea mays* L. were retrieved by blasting rice sequences against the *S. bicolor* and *Z. mays* trace archives, but using the DISCONTIGUOUS MEGABLAST program (http://blast.ncbi.nlm.mih.gov/) as the divergence between rice and *S. bicolor* or *Z. mays* is much greater than that between the two *Arabidopsis* species. Downloaded sequences were subsequently annotated using the SEQUENCHER 4.7 software. In a small proportion of the retained POG loci, one or more of the three annotated sequences (*A. lyrata*, *S. bicolor* and *Z. mays*) had > 5% polymorphic sites (see Table S1). These loci were also abandoned as the last step to minimize the possibility that our data for each selected locus include paralogous sequences from recent gene duplications. A set of 127 loci was finally retained.

Estimation of variation for each selected locus

For each of the 127 loci, sequence alignments between the two *Arabidopsis* species (*A. thaliana* and *A. lyrata*) and among the three Poaceae genera (*Oryza*, *Sorghum* and *Zea*) were performed manually using SE-AL version 2.0a11 (Rambaut, 1996). The uncorrected p-distance was then calculated for the two sets of aligned sequences using the 'Pairwise Base Differences' function implemented in PAUP* version 4.0b10 (Swofford, 2002), as a cursory estimation of variation level. To compare the variation level of these selected PPR gene loci with that of loci extensively used in plant phylogenetic studies, we also retrieved, annotated, and aligned sequences of *rbc*L, *ndh*F, *mat*K, *trn*L-F and ITS, for the two *Arabidopsis* species and the three Poaceae genera, following the procedures described above. The uncorrected p-distance was also subsequently calculated for the five extensively used loci.

Evaluation of phylogenetic utility

We blasted the rice sequence of each of the 127 loci against those of the eight diploid *Oryza* species (*Oryza australiensis* Domin, *Oryza brachyantha* A.Chev. & Roehrich, *Oryza glaberrima* Steud., *Oryza granulata* Nees, *Oryza nivara* S.D.Sharma & Shastry, *Oryza officinalis* Wall., *Oryza punctata* Kotschy ex Steud. and *Oryza rufipogon* Griff.) that have trace archive sequences in GenBank (http://blast.ncbi.nlm.mih.gov/). Thirteen of the 127 loci, which have partial sequences available for six or more of the eight *Oryza* species, were selected for further analyses (see Table S2). Then we blasted the rice sequence of each of the 13 loci against the ESTs of *Triticum aestivum* L., *Hordeum vulgare* L. and *Saccharum officinarum* L. in the TIGR Plant Transcript Assemblies database (http://blast.jcvi.org/euk-blast/plantta_blast.cgi). Partial genomic sequences of the eight *Oryza* species and the ESTs of the additional three Poaceae genera, whenever available, were downloaded, annotated, and aligned with rice, *S. bicolor* and *Z. mays* sequences for each locus, following the procedures described above.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Parsimony analysis was performed on each data set of the 13 loci separately, and both parsimony and Bayesian analyses were performed on the combined data set of all 13 loci. Parsimony analyses were conducted using PAUP* version 4.0b10 (Swofford, 2002). Heuristic searches were performed with 200 random stepwise addition replicates and tree-bisection-reconnection (TBR) branch swapping with MULTREES on. Clade support was determined by bootstrap analyses (Felsenstein, 1985) of 500 replicates, each with 10 random stepwise addition replicates and TBR branch swapping with the MULTREES option effective. Bayesian analysis was conducted using MRBAYES version 3.1.2 (Ronquist & Huelsenbeck, 2003). MODELTEST 3.7 (Posada & Crandall, 1998) was employed to determine the sequence evolution model that best fits the data. The GTR+G model was selected by the Akaike information criterion (AIC; Akaike, 1974) for the combined 13 loci data set. We carried out two independent runs of 1 000 000 generations using the default priors and four Markov chains (one cold and three heated chains), sampling one tree every 100 generations. The first 2500 trees were discarded as burn-in.

Results

Selected loci and their variation level

A set of 127 PPR gene loci was finally obtained via the screening process for potential phylogenetic utility. They all are intronless and have a single orthologue in both rice and *A. thaliana* as well as the other three annotated taxa (*A. lyrata*, *S. bicolor* and *Z. mays*). Table 1 includes a comprehensive list of these loci with the uncorrected p-distance between *A. thaliana* and *A. lyrata*, and the distances among *Oryza*, *Zea* and *Sorghum*. Although these loci consist entirely of protein coding regions, they have a relatively rapid rate of evolution. For example, the pairwise distance between the two *Arabidopsis* species at the PPR loci ranged from 0.0244 (At1G10270) to 0.0985 (At3G25060), with an average of 0.0512, which is 6.7, 4.1, 2.7, 1.4 and 0.9 times the distance at *rbc*L, *ndh*F, *mat*K, *trn*L-F and ITS, respectively (Table 1). Thirty-seven loci had a larger distance than ITS. Similarly, the pairwise distance between rice (*O. sativa*) and maize (*Z. mays*) ranged from 0.1262 (At5G18390) to 0.2789 (At1G10330), with an average of 0.1890, which is 4.7, 3.0, 2.3 and 2.4 times the distance at *rbc*L, *ndh*F, *mat*K and *trn*L-F, respectively (Table 1). The ITS sequences could not be confidently aligned in the three Poaceae genera because of extensive length variation, and therefore no pairwise distances were available for the ITS region among these genera. Fig. 1 is a graphic view of the variation level of these 127 PPR gene loci as well as the five additional loci that have been used extensively in plant phylogenetics. The sizes of these 127 loci in *A. thaliana* are also listed in Table 1. They range from 909 to 3339 bp, with an average of 1977 bp.

Phylogenetic utility

Because of incomplete sequences of the eight diploid *Oryza* species and *S. officinarum*, *H. vulgare* and *T. aestivum*, there are substantial amounts of missing data at the 13 loci that we used to reconstruct the intergeneric relationships within Poaceae and interspecific relationships within *Oryza*. Taxa that have no sequences at all at a certain locus were excluded from phylogenetic analysis of that locus and were not taken into account when calculating the percentage of missing data. The loci AT4G38150 and AT1G11290 have the lowest (11%) and highest (53%) percentages of data missing, respectively (see Fig. 2 and Table S2). All 14 taxa were included in the analyses of the concatenated data, in which 48% of data are missing.

Despite the large amount of missing data, almost every individual locus generated well-resolved gene trees (Fig. 2). The intergeneric relationships within Poaceae were congruent across all 13 loci, and are consistent with the Grass Phylogeny Working Group subfamily classification (Barker *et al.*, 2001; see subfamily designations in Fig. 3). Within the genus *Oryza*, the monophyly of A-genome species and the phylogenetic position of E-genome species were very consistent among the 13 loci. However, the relationships among the A-, B- and C-genomes and the relationships among the F-genome, G-genome and all other genome types were incongruent among these loci (see genome type designation in Fig. 3). These results are consistent with a recent study of the relationships of the six *Oryza* diploid genome types using 142 nuclear genes (Zou *et al.*, 2008). Fig. 3 represents the single most parsimonious tree inferred from the concatenated data for all 13 loci. Intergeneric relationships are the same as shown in gene trees resulting from individual data sets. The *Oryza* genome type relationships are very similar to that reported in the aforementioned study (Zou *et al.*, 2008), except that the positions of the F- and G-genomes are switched.

Discussion

The PPR genes have three major characteristics that make them excellent candidates for plant phylogenetic studies. First of all, there are a large number of loci but orthology assessment is straightforward. Each of the 127 loci obtained from the screening process in the present study should have a single orthologue in the vast majority of diploid flowering plants, given the fact that a single orthologue was retained in both rice and *A. thaliana* after the deep split between monocots and eudicots. To test this intuitive assumption, we randomly drew 10 PPR loci from the list (Table 1), and blasted the *A. thaliana* nucleotide sequences against two other sequenced genomes, those of *Populus trichocarpa* (Tuskan *et al.*, 2006; http://www.ncbi.nlm.nih.gov/projects/genome/seq/BlastGen/BlastGen.cgi?pid=10770) and *Vitis vinifera* (Jaillon *et al.*, 2007; http://www.ncbi.nlm.nih.gov/projects/genome/

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Table 1 List of the 127 pentatricopeptide repeat loci as well as *rbc*L, *ndh*F, *mat*K, *trn*L-F and ITS, with their corresponding uncorrected p-distances and sequence lengths in *Arabidopsis thaliana*

	Locus	*Arabidopsis thaliana* vs *Arabidopsis lyrata*	*Oryza sativa* vs *Zea mays*	*Oryza sativa* vs *Sorghum bicolor*	*Sorghum bicolor* vs *Zea mays*	Length in *A. thaliana* (bp)
1	*rbc*L	0.0076	0.0399	0.0406	0.0077	1440
2	*ndh*F	0.0125	0.064	0.0649	0.0099	2241
3	*mat*K	0.0191	0.0814	0.0794	0.0169	1515
4	*trn*L-F	0.0367	0.0779	0.0792	0.0114	1384
5	ITS	0.055	NA	NA	NA	639
6	At1G02420	0.0373	0.1906	0.1698	0.0711	1476
7	At1G03510	0.0543	0.181	0.1787	0.0538	1290
8	At1G03560	0.0459	0.1565	0.1563	0.0577	1983
9	At1G05600	0.0397	0.2141	0.2032	0.0749	1515
10	At1G05750	0.0472	0.1646	0.1592	0.0604	1503
11	At1G09680	0.0582	0.2136	0.2016	0.0679	1824
12	At1G10270	0.0244	0.1718	0.1528	0.043	2742
13	At1G10330	0.0605	0.2789	0.2493	0.0811	1404
14	At1G11290	0.0547	0.1648	0.1513	0.0747	2430
15	At1G13040	0.0418	0.2109	0.2004	0.0618	1554
16	At1G15510	0.0346	0.1793	0.1719	0.046	2601
17	At1G19290	0.0531	0.2322	0.2227	0.06	2715
18	At1G20230	0.0442	0.1808	0.1727	0.0611	2283
19	At1G20300	0.0471	0.2058	0.1957	0.0728	1614
20	At1G22960	0.0456	0.2045	0.1932	0.0593	2157
21	At1G25360	0.0506	0.1594	0.1495	0.0497	2373
22	At1G26500	0.0645	0.2013	0.1869	0.0521	1518
23	At1G28690	0.0518	0.2078	0.2051	0.0883	1563
24	At1G31430	0.0671	0.1837	0.1711	0.0547	1713
25	At1G33350	0.0544	0.1868	0.174	0.0576	1617
26	At1G53330	0.0593	0.2405	0.2346	0.0865	1416
27	At1G59720	0.057	0.1843	0.1915	0.084	1860
28	At1G64310	0.0464	0.2025	0.1983	0.0701	1659
29	At1G66345	0.0981	0.235	0.233	0.0946	1617
30	At1G68930	0.0535	0.1807	0.174	0.0581	2232
31	At1G71060	0.062	0.2133	0.2179	0.0572	1533
32	At1G71210	0.047	0.2366	0.2004	0.107	2595
33	At1G73400	0.0516	0.1651	0.1591	0.0571	1401
34	At1G74600	0.0598	0.2071	0.1924	0.0497	2688
35	At1G77010	0.0508	0.2316	0.2214	0.0771	2028
36	At1G77360	0.0436	0.146	0.1281	0.0546	1446
37	At1G79490	0.038	0.1438	0.1332	0.0497	2512
38	At1G79540	0.0354	0.2326	0.2148	0.0668	2343
39	At1G80150	0.0419	0.1848	0.1772	0.067	1194
40	At1G80550	0.0472	0.1511	0.1397	0.0534	1347
41	At2G01860	0.0592	0.2156	0.2057	0.0623	1461
42	At2G02980	0.0629	0.1614	0.1471	0.0714	1812
43	At2G03380	0.0618	0.2146	0.203	0.0603	2070
44	At2G03880	0.0417	0.1845	0.168	0.0827	1149
45	At2G13600	0.0516	0.1528	0.1523	0.0493	2094
46	At2G15630	0.0706	0.1778	0.18	0.0476	1869
47	At2G15690	0.0607	0.1365	0.1333	0.0373	1470
48	At2G15980	0.0688	0.2146	0.1999	0.0669	1497
49	At2G16880	0.0412	0.1855	0.1758	0.0634	2232
50	At2G17670	0.0526	0.2136	0.2081	0.087	1392
51	At2G18940	0.0501	0.1837	0.1474	0.103	2469
52	At2G20540	0.053	0.1891	0.1875	0.0676	1605
53	At2G22070	0.0466	0.1917	0.1913	0.0715	2361
54	At2G22410	0.0594	0.2054	0.1978	0.0581	2046
55	At2G27610	0.047	0.2199	0.2089	0.0691	2607
56	At2G32630	0.0475	0.2101	0.1921	0.0936	1875
57	At2G33680	0.043	0.2055	0.2	0.0823	2184
58	At2G33760	0.0576	0.1801	0.165	0.0683	1752

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Table 1 *continued*

	Locus	*Arabidopsis thaliana* vs *Arabidopsis lyrata*	*Oryza sativa* vs *Zea mays*	*Oryza sativa* vs *Sorghum bicolor*	*Sorghum bicolor* vs *Zea mays*	Length in *A. thaliana* (bp)
59	At2G35030	0.0467	0.1892	0.1735	0.0591	1894
60	At2G36240	0.0469	0.2062	0.1943	0.0674	1494
61	At2G36730	0.052	0.2049	0.1962	0.0886	1506
62	At2G37230	0.0463	0.1557	0.1553	0.0616	2274
63	At2G37310	0.0476	0.1952	0.1887	0.0705	1974
64	At2G41080	0.0424	0.1606	0.1564	0.0692	1698
65	At2G42920	0.0602	0.1581	0.1409	0.0549	1680
66	At3G04130	0.0481	0.2359	0.2205	0.0722	1527
67	At3G04750	0.0481	0.1991	0.1898	0.0648	1974
68	At3G05340	0.0379	0.2229	0.1899	0.1014	1977
69	At3G09040	0.04	0.1989	0.1911	0.0587	3087
70	At3G09060	0.046	0.2057	0.2063	0.0553	2064
71	At3G11460	0.0429	0.1976	0.1918	0.068	1872
72	At3G14730	0.0572	0.2233	0.2151	0.0591	1962
73	At3G15130	0.0411	0.1734	0.1652	0.0604	2070
74	At3G16890	0.0592	0.2116	0.2056	0.0791	1980
75	At3G18020	0.0543	0.2281	0.2215	0.0732	2067
76	At3G21470	0.0549	0.2247	0.2154	0.0658	1329
77	At3G22150	0.0422	0.1735	0.1726	0.0501	2463
78	At3G22670	0.0558	0.2229	0.2236	0.0457	1689
79	At3G23020	0.0499	0.2018	0.1965	0.0538	2526
80	At3G25060	0.0985	0.2019	0.1944	0.0778	1782
81	At3G25970	0.0772	0.189	0.1753	0.0835	1941
82	At3G26540	0.0442	0.1896	0.1746	0.0727	2103
83	At3G29230	0.0577	0.1667	0.1585	0.0513	1803
84	At3G46790	0.038	0.1744	0.1649	0.0452	1974
85	At3G47530	0.0401	0.2007	0.1776	0.0691	1776
86	At3G47840	0.0523	0.1899	0.1842	0.061	2121
87	At3G48810	0.0398	0.2205	0.2182	0.0713	1980
88	At3G49240	0.0487	0.1508	0.1378	0.0607	1890
89	At3G53360	0.0458	0.1896	0.1817	0.0626	2307
90	At3G57430	0.0616	0.1931	0.1746	0.0512	2673
91	At4G01570	0.0548	0.1717	0.1627	0.0538	2418
92	At4G14170	0.0712	0.1793	0.1665	0.07	1377
93	At4G20740	0.0517	0.1893	0.1824	0.0643	2184
94	At4G20770	0.0778	0.1689	0.1615	0.0582	2223
95	At4G21300	0.0433	0.1876	0.1848	0.0584	2574
96	At4G30700	0.0319	0.1809	0.1792	0.0626	2379
97	At4G30825	0.0588	0.19	0.1909	0.0619	2715
98	At4G31850	0.0399	0.1888	0.1774	0.0484	3339
99	At4G32430	0.0401	0.1916	0.1822	0.0639	2292
100	At4G33990	0.051	0.1702	0.1635	0.0495	2472
101	At4G35130	0.0562	0.1807	0.1743	0.0649	2415
102	At4G37170	0.0612	0.1704	0.1657	0.0561	2076
103	At4G37380	0.0479	0.2073	0.1937	0.0679	1899
104	At4G38150	0.0497	0.1922	0.1689	0.0612	909
105	At4G39530	0.0423	0.2094	0.2036	0.0582	2505
106	At5G01110	0.0558	0.176	0.1549	0.061	2190
107	At5G02860	0.0378	0.1657	0.1625	0.0546	2460
108	At5G03800	0.0505	0.2013	0.1917	0.0761	2691
109	At5G04780	0.0478	0.1831	0.1762	0.0554	1884
110	At5G06400	0.0492	0.2633	0.2559	0.0595	3093
111	At5G08490	0.0349	0.2081	0.1997	0.0654	2550
112	At5G09950	0.0368	0.1785	0.1757	0.0491	2988
113	At5G12100	0.0522	0.222	0.2065	0.0577	2451
114	At5G13230	0.0494	0.1953	0.1853	0.0632	2469
115	At5G13770	0.0497	0.2024	0.1921	0.0607	1830
116	At5G15300	0.0431	0.1969	0.1808	0.0574	1647
117	At5G16420	0.0456	0.1859	0.1863	0.0698	1608

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Table 1 *continued*

	Locus	*Arabidopsis thaliana* vs *Arabidopsis lyrata*	*Oryza sativa* vs *Zea mays*	*Oryza sativa* vs *Sorghum bicolor*	*Sorghum bicolor* vs *Zea mays*	Length in *A. thaliana* (bp)
118	At5G18390	0.0507	0.1262	0.1207	0.0483	1380
119	At5G18475	0.0546	0.1909	0.1873	0.0542	1521
120	At5G37570	0.045	0.1775	0.167	0.0686	1653
121	At5G39350	0.0549	0.2237	0.2069	0.0794	2034
122	At5G39680	0.0553	0.1761	0.1741	0.0549	2133
123	At5G39980	0.0403	0.15	0.1458	0.0463	2037
124	At5G42450	0.0696	0.1853	0.1883	0.0741	1179
125	At5G44230	0.0481	0.1934	0.186	0.069	1974
126	At5G47360	0.0642	0.234	0.2303	0.0587	1434
127	At5G52630	0.056	0.1595	0.1378	0.0524	1767
128	At5G55740	0.0547	0.2241	0.1949	0.0809	2493
129	At5G56310	0.0621	0.1895	0.1799	0.0642	1593
130	At5G59600	0.0467	0.2015	0.1942	0.0648	1605
131	At5G60960	0.0473	0.1481	0.1419	0.0604	1566
132	At5G61400	0.0583	0.2234	0.2175	0.0693	1965

'NA' indicates that the uncorrected p-distance is not available because the ITS sequences cannot be unambiguously aligned between the corresponding pair of taxa.

seq/BlastGen/BlastGen.cgi?pid=12992), using the CROSS-SPECIES MEGABLAST program. The *A. thaliana* sequence hit a single locus in both genomes at nine of the 10 loci and did not produce any significant hit at the other locus (data not shown). We then blasted the amino acid sequence of this exceptional locus against the same databases using the TBLASTN program and this time it produced a single best hit (*E*-value < e^{-100}) in both genomes. In addition, the successful retrieval of unique orthologous sequences from *S. officinarum*, *H. vulgare* and *T. aestivum* using the rice sequences of the 13 loci used in our phylogenetic analyses corroborates this assumption.

Secondly, the majority of PPR genes are intronless (Lurin *et al.*, 2004; O'Toole *et al.*, 2008). In fact, the 127 loci listed in Table 1 are all intronless, as this was one of the selection criteria. There are two important practical advantages in choosing intronless loci. (1) Alignment is straightforward. Alignments of noncoding DNA sequences such as introns or intergenic spacers can be problematic because of extensive length variation among all but the most closely related species. This often necessitates the introduction of numerous (sometimes impracticably numerous) large or small gaps ('indels') into the alignment. Intronless genes, in contrast, tend to contain few fixed length mutations and are easy to align if the taxa of interest are not too distantly related (e.g. belong to different major clades of angiosperms). (2) Sequencing requires minimal effort. When a nuclear locus is heterozygous, direct sequencing of intron or intergenic spacer regions becomes almost impossible. A simple deletion or insertion that occurred in one allele but not the other(s) will affect all the sequence reads after this point, and cloning is necessary to generate good quality sequences in this situation. For intronless loci, although polymorphism will be observed if there is allelic variation, sequence reads after the polymorphic sites will probably not be affected, because allelic polymorphisms usually do not involve length mutations in protein coding regions. In addition, nuclear gene introns often contain polynucleotide (e.g. poly-A) or/and microsatellite regions that are extremely difficult to sequence through, whereas intronless genes tend not to contain such regions.

The difference between intronless loci and noncoding regions might be trivial if recovering allelic polymorphisms is the main focus of a study, as in some phylogeography or population genetics studies. In such studies, separating multiple alleles within an individual via cloning is desirable, no matter whether the targeted loci are intronless or noncoding. However, being intronless is an obvious advantage when the main question is phylogenetic relationships of organisms and allelic polymorphism is not an issue (i.e. incomplete lineage sorting is trivial). This advantage will be substantially inflated when resolving intergeneric relationships is the primary interest of a study. Nuclear gene intron sequences often diverge rapidly and may not be aligned at all between distantly related genera. Exon sequences are the only source of useful data. Unfortunately, one may need to sequence across several intron regions to generate sufficient exon sequences from a locus that contains both exons and introns. What is worse is that cloning is likely to be necessary to overcome the length mutation problem in introns for many organisms. The laborious cloning work and wasted effort in generating intron sequences that may be useless in resolving intergeneric relationships can be completely avoided by employing these intronless loci. Of course, one may argue that protein coding regions usually diverge much more slowly than intron regions. An intronless locus without sufficient variation to resolve the targeted phylogenetic problem, particularly at lower taxonomic levels, is not very helpful. The third characteristic of PPR genes suggests that this is not a problem.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Fig. 1 Graphic view of the variation levels of the 127 pentatricopeptide repeat (PPR) loci and *rbc*L, *ndh*F, *mat*K, *trn*L-F and ITS, indicated by the uncorrected p-distance. The order of loci follows Table 1. Uncorrected p-distances are shown (a) between *Arabidopsis thaliana* and *Arabidopsis lyrata*, (b) between *Sorghum bicolor* and *Zea mays*, (c) between *Oryza sativa* and *Z. mays*, and (d) between *O. sativa* and *S. bicolor*. The arrow in (b), (c) and (d) indicates that the pairwise distance at the ITS region is not available because the ITS sequences of these taxa cannot be unambiguously aligned because of extensive length variation.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

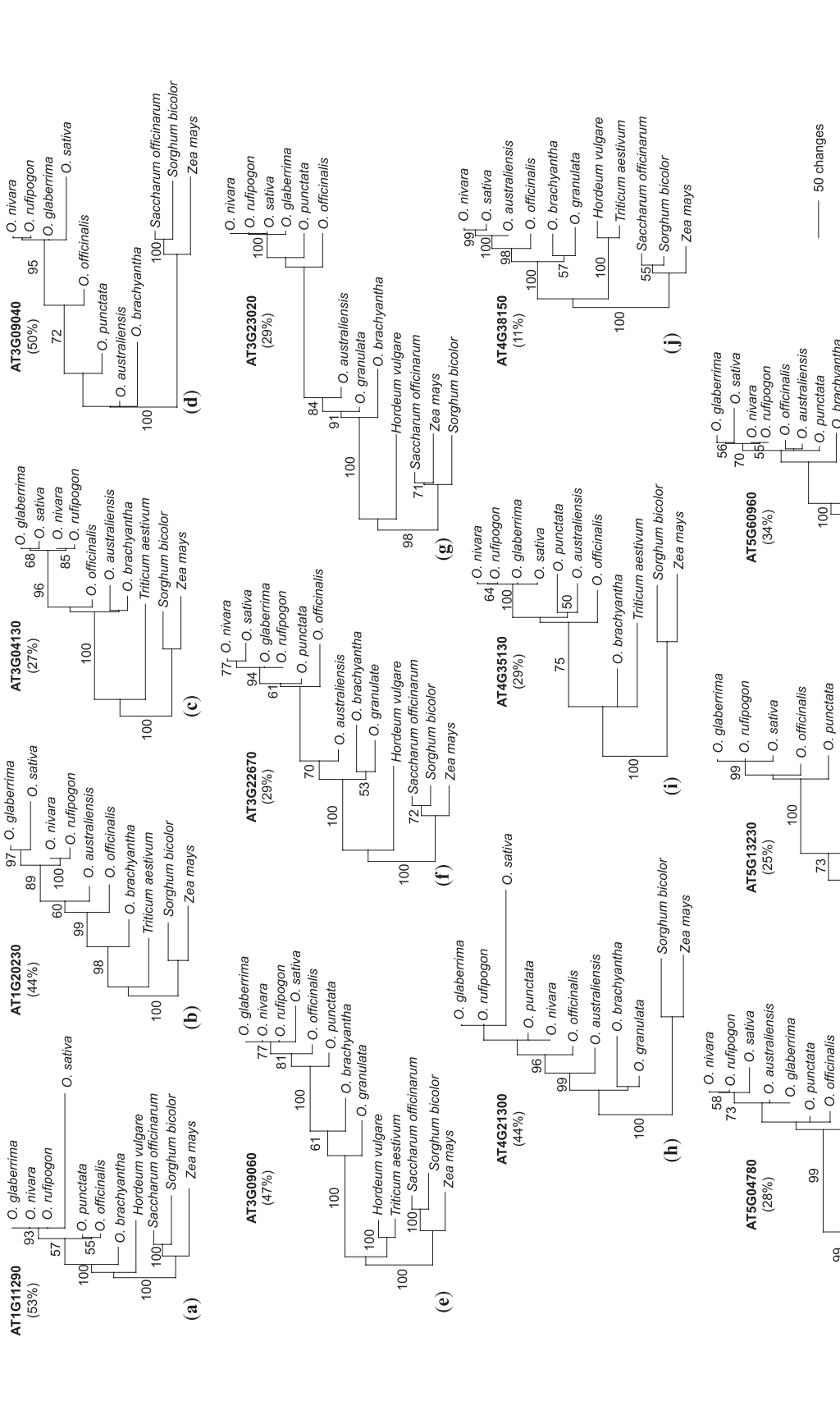

Fig. 2 Gene trees resulting from parsimony analyses of individual data sets for the 13 loci. All trees are drawn to the same scale. Bootstrap values are shown along the branches when > 50. The ID of each locus is in bold and indicated above the gene tree. The number below the locus name (in parentheses) indicates the percentage of missing data at that locus. (a), (b), (e), (i), (l) and (m) show the single maximum parsimony (MP) tree inferred from the corresponding locus. (c) and (k) show one of the six MP trees; (d) shows one of the nine MP trees; (f) shows one of the three MP trees; (g) and (j) show one of the two MP trees; (h) shows one of the 10 MP trees.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

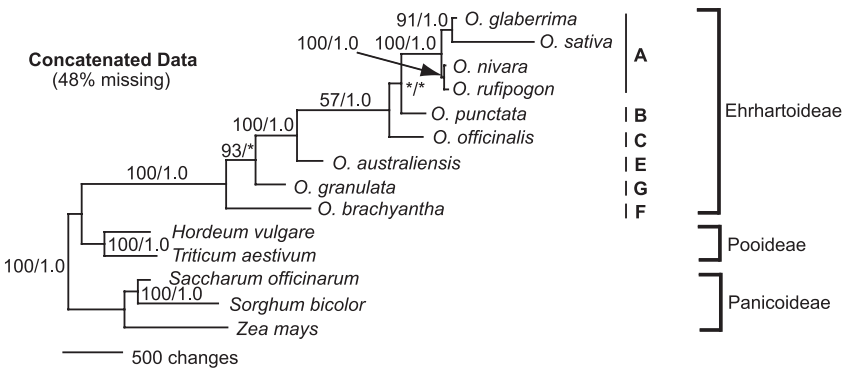

Fig. 3 The single maximum parsimony (MP) tree inferred from the concatenated data for all 13 loci, 48% of which are missing data. Bootstrap values (BS) and Bayesian posterior probabilities (PP) supporting the corresponding nodes are shown along the branches (BS/PP). The asterisks indicate BS < 50 or PP < 0.95. Subfamily designations of the family Poaceae and genome type designations of the genus *Oryza* are represented on the right.

The third property of PPR genes is that they have a rapid rate of evolution. Figure 2 shows the general pattern of variation across the 127 loci we selected, in comparison with that of the four chloroplast DNA regions and ITS region. The average pairwise distance for the selected PPR loci between *A. thaliana* and *A. lyrata* was 1.4 times that for *trn*L-F and 0.9 times that for ITS. The average distances for PPR loci among the three Poaceae genera were 2.3–5.6 times those for *trn*L-F. These data suggest that PPR loci can certainly be used at interspecies and intergeneric levels, considering that both *trn*L-F and ITS have been extensively used for resolving interspecific and intergeneric relationships (Alvarez & Wendel, 2003; Shaw *et al.*, 2005). Our phylogenetic analyses of partial sequences of 13 selected loci confirm this conclusion. Despite the substantial amount of missing data, individual data sets for the 13 loci generated well-resolved gene trees (Fig. 2). The intergeneric relationships were congruent across all 13 loci and consistent with the subfamily classification (Barker *et al.*, 2001). Within the genus *Oryza*, there were both congruent (e.g. the position of the E-genome, *O. australiensis*) and incongruent relationships (e.g. among A-, B- and C-genomes) from one locus to another. These results are consistent with a recent phylogenomic study of the *Oryza* diploid genome types (Zou *et al.*, 2008). Additionally, considering that these loci are intronless, we speculate that they might also be useful to resolve relationships between closely related families, but this possibility needs to be evaluated in future studies.

The unique combination of these three properties gives PPR gene loci many advantages over other nuclear gene loci as phylogenetic tools. They provide numerous loci with established orthology assessment to use. Generating sequence data of these loci requires only minimal effort and aligning these sequences is straightforward. They have a rapid rate of evolution despite being intronless, and versatile utility at various levels (interspecific, intergeneric, and potentially interfamiliar between closely related families). We believe that these loci will play a key role in resolving intergeneric relationships using nuclear gene data, given their extraordinary advantages in this respect, as discussed above. By the present report, we wish to bring the tremendous potential of these PPR gene loci as phylogenetic tools to the attention of plant systematists and to ameliorate the pessimistic view that 'identifying phylogenetically informative LCN markers remains a time-consuming endeavor' (Steele *et al.*, 2008).

There are two final issues that we consider to be worth mentioning from a practical point of view. The first concerns the selection of loci from among these 127 loci for a specific project. Variation level (Fig. 1) and locus size (i.e. sequence length; Table 1) are two informative factors that one can use as guidance to select appropriate loci. However, we should caution that variation level might be lineage specific – the locus with the most rapid rate of evolution in *Arabidopsis* does not necessarily evolve most rapidly in another group. In this sense, locus size may be a more consistent parameter to guide locus selection. The second issue concerns primer design. While universal primers that can be used to amplify a locus across a broad spectrum of organisms (e.g. all angiosperms) are ideal choices, it is more and more widely recognized that such universal primers may not exist for most nuclear loci (Sang, 2002; Steele *et al.*, 2008). For loci that have such a rapid rate of evolution as the PPR genes, primer design in a lineage-specific fashion is probably more fruitful than searching for universal primers. With the rapid development of whole genome sequence and EST databases (e.g. the National Center for Biotechnology Information (NCBI) plant genome project database: http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi?p3=11:Plants&taxgroup=11:Plants|12%3; the TIGR Plant Transcript Assemblies database: http://plantta.tigr.org/) and bioinformatics tools (e.g. BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi), it has become much easier to design lineage-specific primers. The general idea is to use these public databases to search for orthologous sequences of a selected locus from several other plant species, especially those most closely related to the study group. The alignment of these sequences can then provide a basis for the identification of conserved motifs and the design of working primers. As a matter of fact, we have employed this approach and designed Lamiales-specific primers for five more or less arbitrarily selected loci from Table 1. Using these primers we have successfully amplified the targeted loci as single bands in the family Verbenaceae, a typical non-model-system group that has

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

been poorly studied to date. While the details of these empirical data and phylogenetic results will be published elsewhere, we are assured that these loci are quite easy to use in practice.

Acknowledgements

The authors are grateful to Bruce Baldwin and an anonymous reviewer for comments on the manuscript. This research was supported by a Graduate Fellowship in Plant Molecular Systematics from the University of Washington Department of Biology, an NSF Doctoral Dissertation Improvement Grant (DDIG) (DEB-0710026) to RGO for the first author's dissertation research, and an NSF Grant (DEB-0542493) to RGO.

References

Aggarwal RK, Brar DS, Khush GS. 1997. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Molecular & General Genetics* **254**: 1–12.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

Alvarez I, Costa A, Feliner GN. 2008. Selecting single-copy nuclear genes for plant phylogenetics: a preliminary analysis for the Senecioneae (Asteraceae). *Journal of Molecular Evolution* **66**: 276–291.

Alvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**: 417–434.

Bailey CD, Doyle JJ. 1999. Potential phylogenetic utility of the low-copy nuclear gene pistillata in dicotyledonous plants: comparison to nrDNA ITS and trnL intron in Sphaerocardamum and other Brassicaceae. *Molecular Phylogenetics and Evolution* **13**: 20–30.

Barker NP, Clark LG, Davis JI, Duvall MR, Guala GF, Hsiao C, Kellogg EA, Linder HP, Mason-Gamer RJ, Mathews SY *et al.* 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden* **88**: 373–457.

Chapman MA, Chang J, Weisman D, Kesseli RV, Burke JM. 2007. Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). *Theoretical and Applied Genetics* **115**: 747–755.

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP. 2007. The TIGR plant transcript assemblies database. *Nucleic Acids Research* **35**: D846–D851.

Crawford DJ, Mort ME. 2004. Single-locus molecular markers for inferring relationships at lower taxonomic levels: observations and comments. *Taxon* **53**: 631–635.

Delannoy E, Stanley WA, Bond CS, Small ID. 2007. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochemical Society Transactions* **35**: 1643–1647.

Felsenstein J. 1985. Confidence limits on phylogenies – an approach using the bootstrap. *Evolution* **39**: 783–791.

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

Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences, USA* **96**: 14400–14405.

Howarth DG, Baum DA. 2002. Phylogenetic utility of a nuclear intron from nitrate reductase for the study of closely related plant species. *Molecular Phylogenetics and Evolution* **23**: 525–528.

Hughes CE, Eastwood RJ, Bailey CD. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society B – Biological Sciences* **361**: 211–225.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–U5.

Jain M, Khurana P, Tyagi AK, Khurana JP. 2008. Genome-wide analysis of intronless genes in rice and Arabidopsis. *Functional & Integrative Genomics* **8**: 69–78.

Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B *et al.* 2004. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**: 2089–2103.

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

Nayar NM. 1973. Origin and cytogenetics of rice. *Advances in Genetics Incorporating Molecular Genetic Medicine* **17**: 153–292.

O'Toole N, Hattori M, Andres C, Iida K, Lurin C, Schmitz-Linneweber C, Sugita M, Small I. 2008. On the expansion of the pentatricopeptide repeat gene family in plants. *Molecular Biology and Evolution* **25**: 1120–1128.

Olsen KM, Schaal BA. 1999. Evidence on the origin of cassava: phylogeography of manihot esculenta. *Proceedings of the National Academy of Sciences, USA* **96**: 5586–5591.

Padolina JM. 2006. *Phylogenetic reconstruction of Phalaenopsis using nuclear and chloroplast DNA sequence data and using Phalaenopsis as a natural system for assessing methods to reconstruct hybrid evolution in phylogenetic analyses*. PhD dissertation, The University of Texas at Austin, Austin, TX, USA.

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

Pusnik M, Small I, Read LK, Fabbro T, Schneider A. 2007. Pentatricopeptide repeat proteins in trypanosoma brucei function in mitochondrial ribosomes. *Molecular and Cellular Biology* **27**: 6876–6888.

Rambaut A. 1996. *Se-al: sequence alignment editor*. Oxford, UK: University of Oxford. http://evolve.zoo.ox.ac.uk/.

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

Rudinger M, Polsakiewicz M, Knoop V. 2008. Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat proteins in jungermanniid but not in marchantiid liverworts. *Molecular Biology and Evolution* **25**: 1405–1414.

Salone V, Rudinger M, Polsakiewicz M, Hoffmann B, Groth-Malonek M, Szurek B, Small I, Knoop V, Lurin C. 2007. A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Letters* **581**: 4132–4138.

Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* **37**: 121–147.

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92**: 142–166.

Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* **17**: 145–170.

Small RL, Wendel JF. 2000. Phylogeny, duplication, and intraspecific variation of Adh sequences in new world diploid cottons (*Gossypium* L., Malvaceae). *Molecular Phylogenetics and Evolution* **16**: 73–84.

Steele PR, Guisinger-Bellian M, Linder CR, Jansen RK. 2008. Phylogenetic utility of 141 low-copy nuclear regions in taxa at different taxonomic levels in two distantly related families of rosids. *Molecular Phylogenetics and Evolution* **48**: 1013–1026.

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)

Strand AE, LeebensMack J, Milligan BG. 1997. Nuclear DNA-based markers for plant evolutionary biology. *Molecular Ecology* **6**: 113–118.

Swofford DL. 2002. *PAUP*: phylogenetic analysis using parsimony (*and other methods)*, version 4b10. Sunderland, MA, USA: Sinauer Associates, Inc.

Tank DC, Sang T. 2001. Phylogenetic utility of the glycerol-3-phosphate acyltransferase gene: evolution and implications in Paeonia (Paeoniaceae). *Molecular Phylogenetics and Evolution* **19**: 421–429.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Walker NS, Stiffler N, Barkan A. 2007. POGs/PlantRBP: a resource for comparative genomics in plants. *Nucleic Acids Research* **35**: D852–D856.

Wing RA, Ammiraju JSS, Luo MZX, Kim H, Yu YS, Kudrna D, Goicoechea JL, Wang WM, Nelson W, Rao K *et al.* 2005. The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Molecular Biology* **59**: 53–62.

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

Yuan YW, Olmstead RG. 2008. Evolution and phylogenetic utility of the *PHOT* gene duplicates in the *Verbena* complex (Verbenaceae): dramatic intron size variation and footprint of ancestral recombination. *American Journal of Botany* **95**: 1166–1176.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biology* **9**: R49.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Putative orthologous groups of pentatricopeptide repeat (PPR) genes

Table S2 Missing data information for the 13 loci selected for phylogenetic analyses

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

About *New Phytologist*

- *New Phytologist* is owned by a non-profit-making **charitable trust** dedicated to the promotion of plant science, facilitating projects from symposia to open access for our Tansley reviews. Complete information is available at **www.newphytologist.org**.
- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as-ready' via *Early View* – our average submission to decision time is just 29 days. Online-only colour is **free**, and essential print colour costs will be met if necessary. We also provide 25 offprints as well as a PDF for each article.
- For online summaries and ToC alerts, go to the website and click on 'Journal online'. You can take out a **personal subscription** to the journal for a fraction of the institutional price. Rates start at £139 in Europe/$259 in the USA & Canada for the online edition (click on 'Subscribe' at the website).
- If you have any questions, do get in touch with Central Office (**newphytol@lancaster.ac.uk**; tel +44 1524 594691) or, for a local contact in North America, the US Office (**newphytol@ornl.gov**; tel +1 865 576 5261).

© The Authors (2009)
Journal compilation © *New Phytologist* (2009)