A Phylogenomic Investigation of \textit{CYCLOIDEA}-Like TCP Genes in the Leguminosae\textsuperscript{1}

Hélène L. Citerne*, Da Luo, R. Toby Pennington, Enrico Coen, and Quentin C.B. Cronk\textsuperscript{2}

Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, United Kingdom (H.C., R.T.P., Q.C.B.C.); Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, United Kingdom (H.C., Q.C.B.C.); Shanghai Institute of Plant Physiology, Chinese Academy of Sciences, 300 Fenglin Road, Shanghai 200032, China (D.L.); and Genetics Department, John Innes Centre, Colney Lane, Norwich NR4 7UH, United Kingdom (E.C.)

Numerous TCP genes (transcription factors with a TCP domain) occur in legumes. Genes of this class in Arabidopsis (TCP1) and snapdragon (Antirrhinum majus; CYCLOIDEA) have been shown to be asymmetrically expressed in developing floral primordia, and in snapdragon, they are required for floral zygomorphy (bilaterally symmetrical flowers). These genes are therefore particularly interesting in Leguminosae, a family that is thought to have evolved zygomorphy independently from other zygomorphic angiosperm lineages. Using a phylogenomic approach, we show that homologs of TCP1/CYCLOIDEA occur in legumes and may be divided into two main classes (LEGYC group I and II), apparently the result of an early duplication, and each class is characterized by a typical amino acid signature in the TCP domain. Furthermore, group I genes in legumes may be divided into two subclasses (LEGYC IA and IB), apparently the result of a duplication near the base of the papilionoid legumes or below. Most papilionoid legumes investigated have all three genes present (LEGYC IA, IB, and II), inviting further work to investigate possible functional difference between the three types. However, within these three major gene groups, the precise relationships of the paralogs between species are difficult to determine probably because of a complex history of duplication and loss with lineage sorting or heterotachy (within-site rate variation) due to functional differentiation. The results illustrate both the potential and the difficulties of orthology determination in variable gene families, on which the phylogenomic approach to formulating hypotheses of function depends.

The considerable advances in plant developmental genetics from a few model species have provided a starting point for studying plant morphological diversity and evolution at the molecular level. Genes that control development have been implicated in the evolution of novel phenotypes (for review, see Baum, 1998; Doebley and Lukens, 1998; McSteen and Hake, 1998; Cronk, 2001; Shepard and Purugganan, 2002). There is now a growing interest in expanding this knowledge to other species less amenable to genetic studies but displaying patterns of morphological variation that could be accounted for by changes in the expression of developmental genes.

Comparative expression studies rely on a phylogenetic framework to help identify candidate genes (Eisen, 1998). This approach has been used to find putative orthologs of MADS-box genes in non-model species of basal eudicots (Kramer and Irish, 1999). We present here a study of the evolution of putative homologs of the floral symmetry gene CYCLOIDEA (CYC) in legumes, with particular emphasis on the subfamily Papilionoideae. Using relatively wide sampling within Leguminosae is potentially a useful way of identifying the different subgroups within a gene family, as represented in legumes.

In snapdragon (Antirrhinum majus L. [Lamiales, Veronicaceae]), floral dorsal identity is controlled by two closely related nuclear genes CYC and DICHTOMA (DIC; Luo et al., 1996, 1999; Almeida et al., 1997). In floral meristems, CYC and DICH have overlapping expression patterns on the adaxial side, but they have diverged so that CYC is expressed slightly later in development than DICH but has a greater effect on phenotype. These two genes belong to a gene family of putative transcription factors characterized by a basic helix-loop-helix domain referred to as the TCP domain (Cubas et al., 1999a). In Arabidopsis, 24 members have been identified. A subclass of this gene family, to which CYC/DICH and the maize (Zea mays) architecture gene TEOSINTE BRANCHED 1 belong, also has a highly conserved Arg-rich R domain (Cubas, 2002). CYC-like genes have been implicated in the control of floral symmetry in other species in the Lamiales, such as Linaria vulgaris Miller (Veronicaceae; Cubas et al., 1999b). The homolog of CYC in Arabidopsis, TCP1, has recently been shown to be expressed transiently at the adaxial base of floral and axillary meristems (Cubas

\textsuperscript{1}This work was supported by The Carnegie Trust for the Universities of Scotland and by the Systematics Association.

\textsuperscript{2}Present address: Botanical Garden and Centre for Plant Research, University of British Columbia, 6804 Southwest Marine Drive, Vancouver, BC, Canada V6T 1Z4.

*Corresponding author; e-mail h.citerne@rbge.org.uk; fax 44–131–248–2901.

Article publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.102.016311.
et al., 2001). This suggests that asymmetric expression of CYC-like genes may predate the divergence of the Asteridae (e.g. snapdragon) and the Rosidae (e.g. Arabidopsis and Leguminosae). Such asymmetrically expressed genes may have been recruited repeatedly for the evolution of zygomorphy in separate lineages.

The Leguminosae is one such plant family where zygomorphy is believed to have evolved separately from the Lamiales (Stebbins, 1974; Donoghue et al., 1998). With approximately 18,000 species, it is one of the most species-rich angiosperm families, with the greatest number of species (approximately 12,000) found in the subfamily Papilionoideae. Papilionoids are characterized by highly zygomorphic flowers, with an enlarged dorsal (standard) petal, and lateral (wings) and ventral (keel) petals surrounding the reproductive organs. This highly specialized floral form, an adaptation to bee pollination, contrasts with that of the other two subfamilies Caesalpinioideae and Mimosoideae. Mimosoid flowers are typically actinomorphic, with reduced outer whorls, whereas Caesalpinioideae display more variation in floral morphology ranging from near radial symmetry to zygomorphy. Current molecular evidence suggests that mimosoids and papilionoids have evolved from different lineages of a paraphyletic caesalpinioid group (Doyle et al., 1997; Bruneau et al., 2001; Kajita et al., 2001; Fig. 1).

Within the Papilionoideae, a few taxa with atypical near radial symmetry have traditionally been considered basal members of this subfamily, even transitional between caesalpinoids and papilionoids (Polhill, 1981). However, recent molecular evidence suggests that these unusual taxa are derived from typical papilionoids (Pennington et al., 2000). These putative reversals from zygomorphy to actinomorphy provide a framework for studying the control of floral symmetry in legumes.

In the model legumes Lotus japonicus, soybean (Glycine max), and pea (Pisum sativum), CYC-like genes have been isolated, and in the case of L. japonicus, two genes have been found to be asymmetrically expressed in the developing flower (D. Luo, unpublished data). This study aims to expand these findings to other taxa from other major papilionoid groups such as the dalbergioid and genistoid clades as well as basal lineages (Pennington et al., 2001) where most of the morphological variation lies. This study comprises species with unusual flower morphology, such as Acosmium subelegans (Mohl.) Yakovlev and Cadia purpurea (Picc.) Aiton from the Genistoid clade, and Swartzia jorori Harms from the basal papilionoid group (Polhill, 1981; Pennington et al., 2001). C. purpurea, in particular, has open, near radial flowers, with equal free stamens arranged in a ring (Fig. 2a). This contrasts with typical papilionoids from the Genistoid group such as Lupinus (Lupinus nanus; Fig. 2b). Inclusion of legumes with unusual floral morphology is likely to be useful in studies of the origin of derived modifications in floral symmetry.

As functional gene studies expand from model organisms to related species, it becomes necessary to identify the functional counterparts of genes well-characterized in model species. The phylogenomic method proposes that orthology (i.e. common descent) is a likely predictor of functional equivalence (Eisen, 1998). Modern phylogenetic techniques now often permit robust determination of orthology relations of genes. We have thus taken a phylogenomic approach to investigate orthologs of CYC in legumes, with sampling that ensures coverage of all the main clades of papilionoid legumes (Fig. 3).
RESULTS

Legume CYC Sequence Characterization

Thirty-eight sequences with a TCP and R domain were amplified using primers LEGCYC/F1 and R1 in 16 different taxa. Sequence number per taxon ranged from one to four, with only one sequence isolated from non-papilionoid taxa. However, basal papilionoid taxa, such as *S. jorori* and *Dussia macrophyllata* Harms, had multiple copies comparable in number with more derived papilionoid species (see Table I for summary and GenBank accession no.). No evident sequence modifications (e.g. premature stop codons) were observed in papilionoids with unusual floral morphology.

Fragment length ranged from 274 bp (Pisum 1) to 427 bp (Clitoria 1), with a mean length of 333.81 (± 40.2) bp. These fragments were also highly variable in sequence (at the amino acid and nucleotide level), with numerous substitutions and indel events in the region between the TCP and R domain. As a result, unambiguous sequence alignment for all legume CYC-like sequences was only possible in the TCP and R domains.

Table I. List of sequences obtained with primers LEGCYC-F1 and R1, and corresponding GenBank accession number

<table>
<thead>
<tr>
<th>Sequence</th>
<th>GenBank Accession No.</th>
<th>Sequence</th>
<th>GenBank Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceratonia 1</td>
<td>AY225810</td>
<td>Lupinus sp. 1</td>
<td>AY225832</td>
</tr>
<tr>
<td>Dialium 1</td>
<td>AY225811</td>
<td>Lupinus sp. 2</td>
<td>AY225834</td>
</tr>
<tr>
<td>Zapoteca 1</td>
<td>AY225812</td>
<td>Lupinus sp. 3</td>
<td>AY225833</td>
</tr>
<tr>
<td>Pisum 1</td>
<td>AY225813</td>
<td>Lupinus sp. 4</td>
<td>AY225835</td>
</tr>
<tr>
<td>Anthyllis 1</td>
<td>AY225814</td>
<td>Lupinus nanus 1</td>
<td>AY225836</td>
</tr>
<tr>
<td>Anthyllis 2</td>
<td>AY225815</td>
<td>Lupinus nanus 2</td>
<td>AY225837</td>
</tr>
<tr>
<td>Anthyllis 3</td>
<td>AY225816</td>
<td>Lupinus nanus 3</td>
<td>AY225838</td>
</tr>
<tr>
<td>Lotus berthelotii 1</td>
<td>AY225817</td>
<td>Lupinus angustifolius 1</td>
<td>AY225839</td>
</tr>
<tr>
<td>Lotus berthelotii 2</td>
<td>AY225818</td>
<td>Lupinus angustifolius 2</td>
<td>AY225840</td>
</tr>
<tr>
<td>Indigofera 1</td>
<td>AY225819</td>
<td>Machaerium 1</td>
<td>AY225841</td>
</tr>
<tr>
<td>Indigofera 2</td>
<td>AY225820</td>
<td>Machaerium 2</td>
<td>AY225842</td>
</tr>
<tr>
<td>Indigofera 3</td>
<td>AY225821</td>
<td>Amicia 1</td>
<td>AY225843</td>
</tr>
<tr>
<td>Clitoria 1</td>
<td>AY225822</td>
<td>Amicia 2</td>
<td>AY225844</td>
</tr>
<tr>
<td>Clitoria 2</td>
<td>AY225823</td>
<td>Dussia 1</td>
<td>AY225845</td>
</tr>
<tr>
<td>Clitoria 3</td>
<td>AY225824</td>
<td>Dussia 2</td>
<td>AY225846</td>
</tr>
<tr>
<td>Cadia 1</td>
<td>AY225825</td>
<td>Dussia 3</td>
<td>AY225847</td>
</tr>
<tr>
<td>Cadia 2</td>
<td>AY225826</td>
<td>Swartzia 1</td>
<td>AY225848</td>
</tr>
<tr>
<td>Cadia 3</td>
<td>AY225827</td>
<td>Swartzia 2</td>
<td>AY225849</td>
</tr>
<tr>
<td>Cadia 4</td>
<td>AY225828</td>
<td>Swartzia 3</td>
<td>AY225850</td>
</tr>
<tr>
<td>Acosmium 1</td>
<td>AY225829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acosmium 2</td>
<td>AY225830</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acosmium 3</td>
<td>AY225831</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TCP domains of seven legume CYC-like protein sequences from two species, *C. purpurea* and *L. japonicus*, were analyzed in the context of the TCP gene family. Analysis of the TCP domain peptide matrix using protein distance, parsimony, maximum likelihood (ML), and Bayesian methods resulted in congruent trees with strong support values for the major groups. Figure 4 shows the protein ML unrooted phylogram, with support values obtained by Bayesian analysis of the data. The 50% majority rule (MR) protein distance and maximum parsimony trees are also shown for comparison (Figs. 5 and 6, respectively). All analyses strongly suggest that the TCP gene
family can be divided into two main groups: the PCF group (recovered in every analyses with 100% support values) and a second group containing CYC/TB1 and, among others, the five Arabidopsis genes (TCP1, TCP12, TCP18, TCP2, and TCP24) with an R domain. These results confirm the conclusions of Cубас (2002), but with greater sampling and more comprehensive phylogenetic analysis. Within the latter group, CYC/TB1 genes form a separate group from another well-supported clade (in all analyses) of yet uncharacterized proteins. Although unrooted trees are difficult to interpret evolutionarily, because the point of origin is uncertain, these trees strongly suggest that the legume sequences here are the best candidates for CYC/TCP1 orthologs.

All analyses suggest that the legume CYC (LEG-CYC) sequences from *C. purpurea* and *L. japonicus* (with the exception of Cadia 4) form a strongly supported group (found in 92% of Bayesian trees). This monophyletic group (here called LEG-CYC) is sister to the CYC-TCP1 clade in the ML, Bayesian (Fig. 4) and distance (Fig. 5) trees. LEG-CYC genes are therefore putative orthologs of CYC and TCP1. Cadia 4 is recovered in ML (Fig. 4) and distance (Fig. 5) analyses in the clade containing TB1, TCP12, and TCP18. The parsimony analysis is not informative because the relationship between the LEG-CYC clade, Cadia 4, the CYC/LCYC/DICH clade, TCP1, TCP12, TCP18, and TB1 collapses in a 50% MR consensus tree (Fig. 6).

**Evolution of LEG-CYC Genes: Partial TCP and R Nucleotide Analyses**

To recover major groups within the LEG-CYC genes, we analyzed a matrix of 29 legume nucleotide sequences, rooted using snapdragon CYC and DICH, chosen to represent the full range of papilionoid legume taxa and sequence variation. The legume sequences could only be aligned with the snapdragon sequences using the highly conserved TCP and R domains. Parsimony analysis of the 67 informative sites out of 145 in the partial TCP and R nucleotide sequences produced 168 trees with a minimal length of 278 steps (additional branch swapping did not recover any more maximum parsimony trees), a consistency index (CI) of 0.424 and a retention index (RI) of 0.636, indicating fairly high homoplasy (parallel evolution) in the data. A strict consensus tree (Fig. 7), rooted on snapdragon genes CYC and DICH, resolves only one large supported clade within the ingroup (corresponding to group II, see below). Otherwise, only the relationship between sequences from different species of the same genus (e.g. *Lupinus* spp.) or related genera (e.g. *Anthyllis* and *Lotus* spp.) were supported in this analysis.

Model-based methods, such as Bayesian inference, are less sensitive to long-branch attraction and may therefore be better alternatives for analyzing homoplastic data. Bayesian analysis (Fig. 8) recovered two groups of legume sequences with support values (called here group I and group II). Group I had very high (97%) Bayesian support, whereas group I had weak support of 52%. Both groups include species from basal as well as more derived papilionoids and would appear to represent an early duplication event. However, relationships between sequences other than from closely related species or genera (e.g. *Lupinus* spp.) were difficult to interpret.

Therefore although parsimony analysis of this small data set did not resolve relationships between
LEGYC genes well, Bayesian analysis gave a more fully resolved tree. The poor performance of parsimony analysis was probably due to high homoplasy in the data set coupled with the low number of informative characters with consequent low phylogenetic signal.

**Evolution of LEGYC Genes: Inclusion of Sequence Data between the TCP and R Domains**

The region between the TCP and R domains was then added to the initial data set, together with additional legume sequences. Due to the high length and sequence variability of this region, it could not be aligned with nonlegume sequences, and so all analyses are unrooted. Furthermore, because of length variability, alignment was difficult even within legumes. For this reason some of the positions in which the alignment was ambiguous were excluded from the analysis (300 aligned positions). Eight LEGYC sequences were excluded altogether from this analysis for the same reason. The remaining 38 sequences covered 292 unambiguously aligned characters, which required the insertion of 34 gaps of 1- to 6-bp triplets for alignment.

Parsimony analysis of the resulting 153 parsimony informative characters from the extended data set resulted in a single most parsimonious tree of 748 steps, with CI = 0.452 and RI = 0.601. The tree recovered two clades (groups I and II from the previous analyses) with a bootstrap value of 65%, although sequence relationship within these groups had little bootstrap support with the exception of sequences from closely related taxa (Fig. 9). The topology of the ML tree and the 50% MR consensus tree from the Bayesian analysis was identical, with only three nodes collapsing in the Bayesian consensus tree. The topology of those trees was also similar to the tree from the parsimony analysis, but the level of support for the nodes (estimated by Bayesian inference) was much higher in the model-based analysis. For instance, group I and II were recovered in the Bayesian analysis with high support (Fig. 10). Comparison of the partial TCP domains of amino acid sequences from group I and II showed that there were five synapomorphies, which suggests these clades are genuine (Fig. 11). These groupings were also supported by considerable differences in the variable region, such as presence or absence of motifs, which could not be included in the analysis.

**Figure 5.** Fifty percent MR consensus tree of the protein distance analysis using the PAM-Dayhoff model of protein substitution (PROTDIST; Felsenstein, 1993) of the TCP domain. Values >50% of the 100 jackknife replicates are given at branch nodes. Taxa as in legend to Figure 4.
Within group I, two sequences from most taxa were found. These segregated into two clades (A and B, see Fig. 9), which for the most part contained one sequence per taxon, with a few exceptions (for example Machaerium 1 and 2). Clade A contained one LEGCYC sequence from representatives from both the genistoid (Lupinus spp., Cadiia sp., and Acosemium spp.) and robiniod (Lotus spp. and Anthyllis sp.) clades, whereas clade B contained another LEGCYC sequence from these taxa. Although these clades have no bootstrap support in the parsimony analysis, they were found the ML tree and in most Bayesian trees. This suggests a putative orthology relationship between sequences within these clades (IA and IB) and a further conserved duplication in LEGCYC sequences (LEGCYC IA and IB) of possible functional significance.

**DISCUSSION**

**Presence of TCP1/CYC Orthologs in Leguminosae**

In the TCP gene family analyses, evidence from sequence similarity (PROTDIST) and evolution (ML and Bayesian analyses) strongly suggest that the legume CYC-like sequences examined here are homologous to the floral symmetry genes in snapdragon, CYC and DICH, and to the adaxially expressed floral gene TCP1 in Arabidopsis. Within this legume clade, a lower estimate of three CYC-like copies were found within the Papilionoideae, in species ranging from the basal-most clade (S. jorori) to higher papilionoids (e.g. the robinoid A. hermannia). Because of their apparent orthology with snapdragon CYC, these genes are candidates for floral developmental genes in the Leguminosae. However, these analyses, many of which lead to poorly resolved trees, highlight some of the difficulties in making detailed orthology statements within gene families and CYC-like genes in particular.

**Complex Evolution of CYC-Like Genes in the Leguminosae**

No simple pattern of gene evolution tracking organismal phylogeny within the legume CYC family was recovered in the phylogenetic analyses. Possible confounding factors such as intermediate levels of concerted evolution, variation in the rate of sequence
evolution, and independent gene loss and duplication events, which render the interpretation of gene trees difficult (Doyle, 1994), cannot be ruled out here. Different levels of variation in different parts of the sequences also made analysis difficult. The highly conserved TCP and R boxes were alignable but contained little phylogenetically informative information, whereas the variable region contained much variation but was difficult to align. Furthermore, the variation in the TCP and R domains was mainly at the synonymous third codon position and had a high degree of homoplastic variation (accounting for two-thirds of the steps required). High levels of homoplasy, possibly resulting in long-branch attraction and therefore artificial groupings, is suggested by the low support values of the trees from this analysis and the collapse of many nodes in the maximum parsimony strict consensus trees. Also, because the analysis includes clades between which functional differentiation may exist, particular amino acid positions may be subject to different selection pressure in different parts of the tree. This within-site rate variation, or heterotachy (Lopez et al., 2002), is also likely to make phylogenetic reconstruction more difficult.

Two Major Subgroups (I and II) of Legume CYC-Like Genes Represent a Probable Early Duplication

Despite the problematic nature of the data, certain patterns do emerge from the analyses. Results of the rooted Bayesian analysis suggests that LEGCYC genes can be divided into two main groups (referred to as I and II), which are characterized by different amino acid signatures. The results of the unrooted legume analyses of the extended dataset are also consistent with the two-group hypothesis, and these groups, although only moderately supported by maximum parsimony, are strongly supported by Bayesian inference. Taxa ranging from the basal-most papilionoids to highly derived species (from the “inverse repeat loss” clade, e.g. pea) have both groups of genes suggesting that these genes probably diverged after a duplication event before the evolution of the Papilionoideae. In addition to the putative amino acid synapomorphies in the TCP domain (Fig. 11),
these groups are also distinguished by specific motifs in the otherwise variable region between the TCP and R domains.

Evidence for Two Subgroups (IA and IB) of Group I LEGYC Sequences

Within group I, one other major duplication event appears to have occurred, giving rise to two subgroups IA and IB. We recovered genes belonging to both clades in a wide range of the species sampled here, implying that this duplication occurred at least early in the diversification of the papilionoids.

However, the relationships between sequences within these groups appear complex and require further investigation. Even though our sampling is fairly extensive compared with many studies of developmental gene phylogeny, further sampling may help resolve relationships within and between gene copies. However, these results are in agreement with a trend of independent duplications, and possible losses, with rapid gene evolution outside of the conserved TCP and R domains, previously documented in CYC-like genes families from other plant groups (e.g. Gesneriaceae; Citerne et al., 2000).

The Limitations and Potential of Phylogenomics

The lack of resolution resulting from problematic analyses (particularly using parsimony) highlights the limitations of phylogenomics, at least in rapidly evolving genes with high levels of homoplasy and in gene families where functional differentiation may lead to high levels of heterotachy (within-site rate variation). These problems may lead to difficulties in robust orthology estimation and hence functional prediction. In this study, Bayesian inference gives better resolution than parsimony; with the large
amount of homoplasy in these data it is likely that model-based methods such as Bayesian inference will outperform parsimony.

The recognition of a major legume CYC-like (LEG-CYC) group in this study does however suggest likely candidate genes for functional equivalents of CYC/TCP1. Furthermore, within this group of legume CYC candidates, further subgroups are recognized in this study (LEG-CYC IA, IB, and II), inviting investigation of possible functional differences between these. Thus even where phylogenetic analyses are difficult, partial resolution may still enable hypotheses to be generated. Although we recognize the limitation of phylogenomics, we still regard this approach as extremely promising even with relatively intractable gene families.

MATERIALS AND METHODS

Molecular Methods: DNA Extraction, PCR, Cloning, and Sequencing

For each species, genomic DNA was extracted from either fresh or silica dried leaf material following a modification of the cetyl-trimethyl-ammonium bromide procedure of Doyle and Doyle (1987). Previously extracted DNA was available for *Dialium guinanense* (R.T. Pennington, Royal Botanic Garden Edinburgh), pea (line 399; J. Hofer, John Innes Centre), and *Lupinus angustifolius* cv Merrit (S. Barker, University of Western Australia, Perth).

The region delimited by the conserved TCP and R domains was amplified using primers LEGCYC/F1, 5'-TCA GGG SYT GAG GGA CCG-3' and LEG-CYC/R1, 5'-TCC CTT GCT CTT GCT CTT GC-3'. These primers were designed based on available sequences of CYC-like genes from *Lotus japonicus* and soybean (Glycine max; D. Luo, unpublished data), compared with nucleotide sequences of the TCP and R domains from snapdragon (Antirrhinum majus; Y16313; and DICH, AF199465), Arabidopsis (TCP1, AC002130; TCP12, AC011914; and TCP18, AP001303) and maize (Zea mays subsp. mays; TB1, AF340199). PCR amplifications were carried out using Taq and reagents...
Phylogenetic Analysis: Taxon and Sequence Selection

CYC-like genes from legumes were placed in the context of the TCP gene family, represented by certain key sequences from *L. japonicus* and *C. purpurea* (Lotus japonicus 1 and 2, Cadia 1–4; Table I). To simplify the analysis, certain Arabidopsis TCP genes belonging to the PCF group (Cubas, 2002) were excluded (TCP7, TCP8, TCP14, TCP5, TCP20, TCP21, and TCP22 following the nomenclature of Cubas [2002]), whereas other sequences of particular interest were added: *Gossypium hirsutum* ALIXN (AF165924), *Lupinus albus* ‘TCP1’ (AJ426419), *Linaria vulgaris* LCYC (AF161252), and snapdragon DICH (AF199465). The 58 amino acids of the TCP domain were aligned manually. The matrix of 31 sequences was analyzed using not only protein distance methods similar to that of Cubas (2002), but also maximum parsimony, ML, and Bayesian methods (see below).

Results from these analyses guided the choice of sequences sampled to investigate the evolution of CYC-like genes in the legume family, using nucleotides of the TCP and R domains, with CYC, DICH, and TCPJ as outgroups. Twenty-nine taxa were sampled to represent the phylogenetic range of the papilionoids.

For the detailed analysis within the legumes including the nucleotide region between the TCP and R domains, a larger number of species was used, with representatives from the three subfamilies Caesalpinioideae, Mimosoideae, and Papilionoideae (Table II). Particular emphasis was placed on sampling representatives from all major papilionoid groups defined by current molecular phylogenetic evidence (Doyle et al., 1997; Hu et al., 2000; Kajita et al., 2001; Lavin et al., 2001; Pennington et al., 2001; M. Wojciechowski, M. Lavin, and M. Sanderson, unpublished data; Fig. 3, names of groups follow [Pennington et al., 2001]). All legume sequences obtained with primers LEGCYC/F1-R1, with the exception of Cadia 4, were selected as the ingroup. Additional legume sequences from separate studies were included in this analysis: *L. japonicus* (Lotus japonicus 1, Lotus japonicus 2),

### Table II. Species used in survey of CYC-like genes using primers LEGCYC-F1 and R1

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Clade</th>
<th>Taxon</th>
<th>Source*</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caesalpinioideae</td>
<td></td>
<td><em>Coratonia oreoetha</em> (Hill) Lewis &amp; Verdc.</td>
<td>1996 0942A</td>
<td>Oman</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dialium guianense</em> (Aubl.) Sandw.</td>
<td>R.T. Pennington 639</td>
<td>Napo, Ecuador</td>
</tr>
<tr>
<td>Mimosoideae</td>
<td></td>
<td><em>Zapotoca tetragona</em> (Willd.) H.M. Hernandez</td>
<td>1999 1149</td>
<td>Guatemala</td>
</tr>
<tr>
<td>Papilionoideae</td>
<td>Inverse Repeat Loss clade</td>
<td><em>Pea</em> (Pisum sativum) line 399</td>
<td>–</td>
<td>cultivated, John Innes Centre, Norwich, UK</td>
</tr>
<tr>
<td></td>
<td>Robinioid clade</td>
<td><em>Anthyllis hermanniae</em> L.</td>
<td>1975 1501</td>
<td>Mediterranean</td>
</tr>
<tr>
<td></td>
<td>Old World Tropical clade</td>
<td><em>Lotus bertheloti</em> Masf.</td>
<td>1978 0702B</td>
<td>Canary Islands</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Indigotiera pendula</em> Franch.</td>
<td>1991 0547A</td>
<td>China</td>
</tr>
<tr>
<td>Genistoid clade</td>
<td></td>
<td><em>Clitoria sp.</em></td>
<td>R.T. Pennington 990</td>
<td>San Martín, Peru</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Cadia purpurea</em> (Picc.) Aiton</td>
<td>1994 2001A</td>
<td>Yemen</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Acosmium subegens</em> (Mohl.) Yakovlev</td>
<td>Bridgewater 358</td>
<td>Mato Grosso do Sul, Brazil</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Lupinus sp.</em></td>
<td>R.T. Pennington 815</td>
<td>Piura, Peru</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>L. nanus</em> Doug. ex Benth.</td>
<td>–</td>
<td>commercial seed (Sutton Seeds, Paignton, Devon, UK), cultivated, University of Western Australia, Perth</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Lupinus angustifolius</em> L. cv Merrit</td>
<td>–</td>
<td>Brazil</td>
</tr>
<tr>
<td>Dalbergioid clade</td>
<td></td>
<td><em>Maechaerium scleroyxylon</em> Tul.</td>
<td>1999 0888A</td>
<td>Brazil</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Amicia glandulosa</em> Kanth</td>
<td>R.T. Pennington 654</td>
<td>Loja, Ecuador</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dussia macropropylillata</em> Harms</td>
<td>1995 1539A</td>
<td>Heredia, Costa Rica</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Swartzia joroni</em> Harms</td>
<td>R.T. Pennington 938</td>
<td>Santa Cruz, Bolivia</td>
</tr>
</tbody>
</table>

*Source number refers to either RBGE living collection number (e.g. 1996 0942A) or collector’s voucher number from wild collections (e.g. R.T. Pennington 639). All herbarium vouchers at RGBE.
soybean (Soya 1), pea (Pisum CYC1, Pism CYC2; D. Luo, personal communication), and Medicago truncatula (Medicago 1, BG455908). Snaphdragon CYC and DICH and Arabidopsis TCP1 were chosen as outgroups in the partial TCP and R domains nucleotide sequence analysis.

DNA Sequence Alignment

Unambiguous alignment of all 54 legume CYC-like DNA sequences from 25 taxa was only possible in the TCP and R domains and reduced the matrix to 145 nucleotide characters. However, by excluding certain problematic sequences, it was possible to align certain parts of the variable region between these two conserved domains as protein sequences that were then analyzed as nucleotide sequences. Protein sequences were aligned using ClustalX (Thompson et al., 1997), followed by manual adjustments taking both amino acids and nucleotides into consideration.

Phylogenetic Analysis

Protein Methods

Protein distance analysis was carried out using program from the PHYLIP package (Felsenstein, 1993). One hundred half-deletion jackknife data sets were obtained with SEQBOOT, distance matrices were calculated with PROTDIST using the PAM-Dayhoff model of amino acid substitution, neighbor-joining trees were obtained with NEIGHBOR, and a consensus tree was produced by CONSENSE. Branches with <50% support were collapsed. Protein ML analysis was also carried out using PHYLIP. The most parsimonious trees were calculated with PROTPARS (Felsenstein, 1993), with support values obtained by 100 half-deletion jackknife replicates as described above. A 50% MR consensus tree was obtained with CONSENSE, collapsing branches with <50% jackknife support. Protein ML analysis was carried out using TREEPUZZLE v5 (Schmidt et al., 2005) with the BLOSUM 62 model of substitution (Henikoff and Henikoff, 1992) allowing for two rates of heterogeneity (1 invariable + 1 variable). To provide support values, Bayesian analysis was carried out using MrBayes v2.0 (Huelsenbeck and Ronquist, 2001), using the PAM-Dayhoff amino acid substitution model with one million generations sampled every 100 generations with a burn-in of 100,000 generations.

DNA Methods

Maximum parsimony analysis was carried out using PAUP* 4.0b7 (Phylogenetic Analysis Using Parsimony, version 4.0b7, Sinaur Associates, Sunderland, MA). Heuristic searches with 1,000 random addition replicates (to avoid local optima) and tree bisection reconnection (TBR) branch swapping were conducted with steepest descent and multrees options selected. A maximum of 10 minimal length trees was retained per replicate, and a further heuristic search by TBR was carried out on the shortest trees. Branch support values were calculated by 1,000 bootstrap replicates with simple sequence addition and a maximum of 10 minimal length trees retained per replicate. This search method was carried out both for the TCP and R nucleotide matrices, as well as the matrix incorporating certain variable regions. Bayesian phylogenetic analysis of the TCP plus R data set was carried out using MrBayes v2.0 (Huelsenbeck and Ronquist, 2001), using a general time reversible (GTR) model and site-specific rates partitioned by codon. Chains were run for 600,000 generations (burn-in of 100,000 generations) each. Resultant trees were used to generate a 50% MR consensus tree in PAUP* v4.0b7.

ML analyses were carried out for the matrix incorporating the more variable regions. The best-fit model was GTR + I + G (GTR model estimating the proportion of invariable sites and γ-distribution; Rodriguez et al., 1990), selected as the best-fit model of nucleotide substitution by the Akaike Information Criterion using Modeltest v3.06 (Posada and Crandall, 1998). A heuristic ML analysis with TBR branch swapping was carried out using PAUP* v4.0b7 with the parameters defined from above.

Distribution of Materials

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining permissions will be the responsibility of the requestor.

ACKNOWLEDMENTS

We thank the Royal Botanic Garden Edinburgh for use of laboratory and glasshouse facilities, the horticultural staff, and the laboratory staff (particularly Michelle Hollingsworth and Alex Ponge) for assistance. Julie Hofer (John Innes Centre) and Susan Barker (University of Western Australia) kindly made available DNA samples. We thank Debbie White (RGBE) for the photographs.

Received October 16, 2002; returned for revision November 20, 2002; accepted December 29, 2002.

LITERATURE CITED

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Seattle, Washington

Copyright © 2003 American Society of Plant Biologists. All rights reserved.
Phylogenomic Investigation of CYCLOIDEA-Like Genes in Leguminosae


