PART OF A SPECIAL ISSUE ON EVOLUTION AND DEVELOPMENT

High time for a roll call: gene duplication and phylogenetic relationships of TCP-like genes in monocots

Mariana Mondragón-Palomino1,2,* and Charlotte Trontin2,†

1 Department of Cell Biology and Plant Biochemistry, University of Regensburg, Regensburg, Germany and 2 Department of Genetics, Friedrich Schiller University, Jena, Germany

† Present address: Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, F-78000 Versailles, France.

* For correspondence. E-mail mariana.mondragon@biologie.uni-regensburg.de

INTRODUCTION

The TCP transcription factors form an ancient, plant-specific family that originated in the Streptophyta about 650–800 million years ago (MYA) (Navaud et al., 2007). Its name is an acronym after the first three characterized members: teosinte-branched 1 (tb1) from Zea mays (Doebly et al., 1997), CYCLOIDEA (CYC) from Antirrhinum majus (Luo et al., 1996) and PROLIFERATION CELL FACTOR 1 and 2 (PCF1 and PCF2) from Oryza sativa (Kosugi and Ohashi, 1997). The members of this gene family encode a conserved TCP domain that is predicted to adopt a helix-loop-helix structure (Cubas et al., 1999a). This domain binds to motifs with the consensus GGNNCCAC and GTGGNCC, mediates homo- and heterodimeric protein–protein interactions needed for DNA binding (Kosugi and Ohashi, 2002) and might also be important for activation and/or repression of transcription (Martin-Trillo and Cubas, 2010).

Phylogenetic analyses based on amino acid sequence similarity of the TCP domain showed that the TCP family is divided in two major clades: Class I (Kosugi and Ohashi, 2002) or PCF (Cubas, 2002) and Class II also known as CYC/tb1 (Cubas, 2002; Howarth and Donoghue, 2006). The latter is subdivided in two clades: the CYC/tb1 clade or angiosperm-specific ECE clade and the more ancient CINNINATA (CIN) clade (Martin-Trillo and Cubas, 2010). While most of the members of the ECE clade contain a conserved R domain that might be involved in protein–protein interaction (Cubas et al., 1999a), some members of the CIN clade independently acquired an R domain (Cubas, 2002).

The phylogenetic relationships within the ECE clade have been well characterized in eudicots. In this group the ECE clade underwent two duplication events leading to three subgroups: CYC1, CYC2 and CYC3 (Howarth and Donoghue, 2006; Chapman et al., 2008). Within those subgroups, evolution of the TCP-like genes is quite complex with series of taxon-specific gene duplications and losses. In the case of DICHOTOMA (DIC) and CYC genes one of these duplications led to subfunctionalization (Hileman and Baum, 2003).

Regardless of the class where they belong, TCP transcription factors from Class I and Class II influence differential growth by activating or inhibiting cell division during development. For example, Class I TCPs PCF1 and PCF2 from Oryza sativa activate the expression of the gene PCNA, which

© The Author 2011. Published by Oxford University Press on behalf of the Annals of Botany Company. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
encodes a factor essential for many meristematic cellular functions (Kosugi and Ohashi, 1997), whereas AtTCP20 positively influences the expression of the cyclin CYCB1;1 (Li et al., 2005) and in certain contexts can also inhibit the transcription of other genes (Hervé et al., 2009). The protein CHE is also an inhibitor of transcription that binds to the promoter of CCA1, a key component of the circadian clock in Arabidopsis (Pruneda-Paz et al., 2009).

Several Class II TCP factors repress cell division in specific structures, for instance tb1 and tb1-like proteins from Zea mays, Oryza sativa, Sorghum bicolor and Arabidopsis thaliana repress the development of axillary branches (Hubbard et al., 2002; Takeda et al., 2003; Kebrom et al., 2006, Aguilar-Martínez et al., 2007; Finlayson, 2007). In Antirrhinum, CINCINNATA (CIN) promotes growth arrest of leaf margins and affects cell differentiation in leaves and petals (Nath et al., 2003; Crawford et al., 2004). The role of CIN-like transcription factors in the control of leaf morphogenesis has been analysed in other eudicots such as Arabidopsis thaliana (Koyama et al., 2007) and Solanum lycopersicum (Ori et al., 2007) but not in monocots. Class II TCP proteins also activate the transcription of other genes. For instance, AtTCP13 (PTF1) promotes the transcription of PSBD, the D2 component of photosystem II (Baba et al., 2001).

Class II proteins like CYC and DICH are involved in the establishment of flower and petal asymmetry by promoting stamen abortion and reducing cell division in dorsal petals. The molecular mechanisms behind the origin of zygomorphy are well understood in the model species Antirrhinum majus (snapdragon). In flowers of A. majus, dorsiventral asymmetry is observed in the second whorl, which is formed by two asymmetric dorsal petals, two asymmetric lateral petals and one symmetric ventral petal, while in the third whorl the dorsal stamen is reduced to a staminode.

During the early stage of flower development, CYC retards growth rate and reduces the number and growth of organ primordia in dorsal regions of the flower meristem (Luo et al., 1996). Later in development, the expression of CYC and DICH is restricted to the dorsal petal and stamen primordia and is different in each whorl (Luo et al., 1996, 1999). While in whorl 2 CYC promotes petal lobe growth, in whorl 3 it arrests the development of the dorsal stamen, which eventually becomes a staminode (Luo et al., 1996). The late effects of CYC are mediated by activation of the MYB-like transcription factor RADIALIS (RAD) in the dorsal and lateral domains of the flower meristem, where RAD inhibits the activity of the MYB-like transcription factor DIVARICATA (DIV) (Corley et al., 2005). It has been proposed that DIV establishes ventral petal identity by interacting with class B MADS domain transcription factors DEF and GLO (Perez-Rodriguez et al., 2005).

Within eudicots, the role of TCP proteins in the establishment of bilateral symmetry has been observed in species-rich taxa with zygomorphic flowers such as Fabales, Brassicales, Lamiales and Asterales (Cubas et al., 1999b; Busch and Zachgo, 2007; Wang et al., 2008; Chapman et al., 2008). These studies show that different evolutionary and developmental mechanisms are responsible for the transition from polysymmetry to monosymmetry (reviewed in Busch and Zachgo, 2009), e.g. change in the pattern of expression and the targets of TCPs proteins in Iberis amara (Busch and Zachgo, 2007), neofunctionalization of PsCYC3 a CYC2 gene from Psim sativum (Wang et al., 2008), subfunctionalization after duplication of DICH and CYC in Antirrhinum (Luo et al., 1996; Gübitz et al., 2003) and introgression of the RAY locus in Senecio vulgaris (Kim et al., 2008).

In comparison to the vast body of work developed in eudicots, the role of TCP proteins in the establishment of bilateral symmetry in monocot flowers has not been widely investigated. However, in monocots as in eudicots, zygomorphy evolved at least 23 times independently (Citerne et al., 2010) and it is considered a key innovation that played a role in the radiation and diversification of several monocots families such as Orchidaceae and Poaceae (Sargent, 2004). In monocots, Class II TCP-like genes could be feasibly involved in the evolution of monosymmetry because this pattern is due to reduction or suppression of organs (structural zygomorphy) as it is observed in Commelinids and Orchidaceae and to a lesser extent also in the differential colour patterning or organ displacement in some Liliales (Rudall and Bateman, 2004). In this context, Yuan et al. (2009) showed recently that the TCP protein REP1 encoded by a CYClb1 gene partially determined floral zygomorphy along the lemma-palea axis in O. sativa. Although rice, in comparison to many eudicot flowers, develops a highly reduced floret structure, this recent study suggests that members of the TCP family have been recruited several times independently during the evolution of flower monosymmetry.

Because of the recent availability of the genomes from Sorghum bicolor, Zea mays and Brachypodium distachyon, most published phylogenies of the TCP family in monocots deal with the evolutionary history of the transcription factor tb1 (Lukens and Doebley, 2001) or exclusively involve genes of O. sativa (Navaud et al., 2007; Yao et al., 2007). To follow up the evolution and function of TCP-like genes in a wider group of monocot species, a necessary first step is to perform a detailed phylogenetic analysis of all available sequence information. In this work, the results of such an analysis are presented, as well as specific information on the relationships of orthology, history of duplication and specific sequence motifs of 20 distinct groups of putative TCP-like genes from monocots. These findings can be readily applied to the systematic identification and comparison of new members of the TCP family in monocots. The proposed system of classification based on a molecular phylogeny facilitates building connections between what is known today about this family and what will be learnt in the future from functional and genomic studies.

MATERIALS AND METHODS

Database query and multiple sequence alignment

Between August and September 2010 a dataset of TCP-like sequences was assembled by querying with tblastn the databases of ESTs and non-redundant nucleotides of NCBI's GenBank. In this search, the amino acid sequences of known TCP-like proteins from Oryza sativa (Yao et al., 2007) and tb1 of Zea mays were employed as queries. Further sequences were retrieved from from EBI's BioMart (www.biomart.org) by downloading
all monocot entries with Interpro domains IPR017887 (subgroup of TCP transcription factors), IPR017888 (CYC/IB1, R domain) and IPR005333 (transcription factor TCP). Similar searches where performed in databases of the completely sequenced genomes of *O. sativa japonica*, *S. bicolor*, and *B. distachyon* in Gramene (www.gramene.org) as well as in the databases of Zea mays B73 (www.maizegdb.org and www.maizesequence.org). The output obtained from all databases was parsed with Perl scripts, aligned with MUSCLE v. 3.8.31 (Edgar, 2004) and those sequences that were redundant or with an incomplete TCP domain were eliminated. The searches retrieved from GenBank 27 *ibl*-like sequences from the tribe Andropogoneae. Because these sequences are highly similar to each other and their phylogenetic relationships have already been described (Lukens and Doebley, 2001), only 14 of them were employed in the present phylogenetic analyses. To distinguish the major clades of TCP-like genes, 22 sequences obtained from eudicot genes that have been experimentally characterized were included (Table S1 in Supplementary Data). To ensure a reliable and informative alignment of eudicot reference sequences, only those that are at least 40% identical at the amino acid level with the members of one of the groups of monocot sequences distinguished were employed. The sequences from *Oryza sativa japonica* and *Arabidopsis thaliana* employed here follow the nomenclature of Martin-Trillo and Cubas (2010).

A total of 153 (131 monocot, 22 eudicot) amino acid sequences spanning the TCP and carboxyl domains were aligned with the program MUSCLE (Text file in Supplementary Data). The carboxyl domain comprises all residues after the TCP domain up to the last amino acid encoded (the one before the stop codon) or the last residue reported for the sequences analysed. The resulting alignment was subsequently optimized with several rounds of visual inspection in SeaView v. 4.2.6 (Gouy et al., 2010) and re-alignment of specific sequences and sites with MUSCLE. The final amino acid alignment was employed to align the corresponding DNA sequences.

Isolation and sequencing of TCP-like genes from *Phalaenopsis*

Plants of *Phalaenopsis* hybrid ‘Athens’ were obtained from ‘Valerius Orchideas’ Berlin, Germany. Total RNA was isolated from young terminal floral buds with the standard Trizol protocol. To isolate members of the TCP family in *Phalaenopsis* hyb. ‘Athens’, total RNA from young floral buds was used to synthesize cDNA with the primer AB05 that binds the poly-A tail of mRNAs (5′-GAC TCG AGT CGA CAT CTG TTT TTT TTT TT-3′). The region between the TCP domain and the poly-A tail was amplified using the primer AB07 (5′-GAC TCG AGT CGA CAT CTG-3′) and two degenerated primers binding a conserved region at the beginning of the TCP domain TCPD1 (5′-AAR GAC CGN CAY AGY AAR RTK-3′) and TCPD2 (5′-AAR GAC CGR CAY AGY AAG RTK-3′). TCPD1 was designed based on the alignment of the closest TCP-like sequences obtained from GenBank when the only TCP sequence from an orchid was that of *Dendrobium* hyb. (ABF61889). The primer TCPD2 was designed based on the alignment of the first TCP-like genes that were isolated from the first round of cloning and sequencing: *PhalTCP1*, *PhalTCP2*, *PhalTCP3* and *PhalTCP4*. Polymerase chain amplifications were performed in 25-μL reactions as indicated by the manufacturer of *Taq* polymerase (Fermentas, St Leon-Rot, Germany). The PCR products were separated in a 1% agarose gel and purified with the QIAquick gel extraction kit (Qiagen Ltd, Dorking, Surrey, UK), ligated to the cloning vector pJet1 with the CloneJET PCR Cloning Kit (Fermentas) and transformed in thermo-competent *Escherichia coli* (DH5α). Cells were plated on selective medium with ampicillin. From each transformation, 96 clones were screened via PCR using vector-specific primers (Fermentas). The clones containing an insert of >700 bp were sequenced in both directions on an ABI 3730xl DNA analyser using big dye terminator chemistry. The resulting sequences were assembled and managed with the program SEQUECHER (v. 4.5; Gene Codes Corporation, Ann Arbor, MI, USA). The five sequences of TCP-like genes obtained are deposited in GenBank.

Phylogenetic analysis

Two final alignments of 153 nucleotide sequences were employed to reconstruct the molecular phylogeny of TCP-like genes from monocots (Text file in Supplementary Data). The first alignment includes all residues encoding the TCP domain and the carboxyl domain as defined above. The second alignment includes only the residues encoding the TCP domain. Each of these two alignments was analysed with the following three phylogenetic approaches.

1) Maximum likelihood. The program PhyML 3.0 (Guindon and Gascuel, 2003) implemented in SeaView v. 4.2.6 for MacOS was employed with the GTR model and four categories of rate substitution. The corresponding a parameter of the Γ distribution as well as the proportion of invariable sites were estimated from the data. Branch lengths and model parameters were optimized and the tree topology was obtained by using both strategies of nearest-neighbour interchange and subtree pruning and regrafting. The approximate likelihood test was computed to perform a Shimodaira–Hasegawa-like statistic to support every bifurcation.

2) Bayesian inference. The Linux version of jModelTest 0.1-1 (Posada, 2008) implemented in the computer cluster of the University of Regensburg determined that according to the corrected Akaike information criterion, the model of nucleotide substitution TVM + I + G (transversional model with estimated proportion of invariable sites and Γ distribution) is the best one fitting the alignment of TCP-like genes that spans the residues encoding the TCP domain to the end of the carboxyl domain. For the second alignment analysed, which comprised exclusively the residues encoding the TCP domain, the best fitting model TVMeF + G (transversional model with equal frequencies and Γ distribution) was employed. The parameters of these models were employed in the parallelized version 3.1-2 of MrBayes (Husonbeck and Ronquist, 2001; Altekar et al., 2004) to obtain the corresponding phylogeny. In these analyses, the parameters of the models previously mentioned were used with either 2 million generations and burn in of 3000 for the full alignment.
or 4 million generations and burn in of 22 000 generations for the alignment encoding only the TCP domain.

(3) Analysis of phylogenetic distances. The models of nucleotide substitution for each of the datasets analysed were employed to calculate the pairwise distances between OTUs (operational taxonomic units) and reconstruct the phylogenetic relationships with the method of neighbor-joining implemented in PAUP v. 4b10 (Swofford, 2002) for Linux. The statistic support of these analyses was tested with 10 000 bootstrap replicates.

Searches in database Phytozone

To compare the genomic context of the TCP-like genes isolated from B. distachyon, O. sativa, S. bicolor and Z. mays, information on their patterns of microsynteny (www.phytozone.net) was retrieved from the database Phytozone v. 5.0. To this end, the grass pre-duplication database via blastx v. 2.2.22 was queried using default parameters. Further analyses were performed on the matching group that had the smallest E-value and where the query sequence matched >70 % of the length of the hit sequences in the database.

RESULTS

There are at least 20 groups of highly similar putative TCP-like genes in monocots

Search and alignment of TCP-like sequences from public databases resulted in a total of 131 entries from monocots. About 75 % of these sequences belong to species in the family Poaceae, the rest is formed by representatives from Zingiberaceae, Amaryllidaceae, Musaceae and Orchidaceae (Fig. 1; Table S1 and Text file in Supplementary Data). In Orchidaceae, five of the seven sequences representing this family were isolated from Phalaenopsis hyb. ‘Athens’.

**Fig. 1.** Phylogeny of the species represented in the analysis of monocot TCP-like genes. The eudicot species illustrated are represented in the subset of the reference genes employed. The numbers between brackets indicate the number of sequences per family or subfamily. The phylogenetic relationships are based on Chase et al. (2006) and information from the following on-line databases: Germplasm Resources Information Network (www.ars-grin.gov) and Angiosperm Phylogeny Website v.10 (www.mobot.org).
Searches in the complete genomes of *O. sativa*, *S. bicolor*, *B. distachyon* and *Z. mays* identified 24, 24, 20 and 31 putative TCP-like genes, respectively. The phylogenetic relationships of most candidate TCP-like genes from *Z. mays*, *S. bicolor*, *B. distachyon* and the species in the Musaceae, Zingiberaceae, Orchidaceae and Amaryllidaceae have not been previously characterized, with the exception of *tb1* and *tb2* from *Zea mays*, 12 *tb1*-like sequences from the tribe Andropogoneae (Lukens and Doebley, 2001) and the sequences from *Oryza sativa* (Navaud et al., 2007; Yao et al., 2007).

Search and alignment show that the monocot representatives of the TCP family can be organized in at least 20 groups of sequences. Each of these groups is formed by sequences from different species that in average are \( \geq 64\% \) identical (except group 15) and that share amino acid motifs all along the residues spanning the TCP and carboxyl domains (Table 1; Text file in Supplementary Data). Although within the members of each group there is a high level of identity, between groups the proportion of identical residues beyond the TCP domain is significantly lower. While this may limit the comparison of members from different groups, it greatly increases the resolution of relationships within each group. However, several members of groups 12, 13, 14, 15, 17 and 18 share an R domain and/or target sites for miR319 in their corresponding DNA sequences (Table 1). The R domain forms a coiled coil that may be involved in protein–protein interactions (Lupas et al., 1991), while the family of microRNAs miR319 is present in a wide group of plant species (Sunkar and Jagadeeswaran, 2008) and predominantly regulates the post-transcriptional activity of TCP mRNAs during leaf and petal development (Ori et al., 2007; Palatnik et al., 2007; Nag et al., 2009).

While representatives from each plant family available populate most groups, so far four of them are formed exclusively by sequences from *S. bicolor* (group 4), *O. sativa* (group 11) or from species in the Zingiberaceae or Asparagales (groups 10 and 15, respectively).

**Most groups identified form well-supported clades**

The alignment of residues encoding the TCP and carboxyl domains of 131 monocot and 22 eudicot sequences had a length of 1695 positions. The molecular phylogeny based on this matrix shows that 13 of the 20 groups previously identified are supported with SH-like indexes higher than 0.90 in the correspondence between groups of amino acid identity \( \geq 64\% \) and well-supported clades observed in the phylogenetic analysis with maximum likelihood (Fig. 2). This phylogenetic analysis also supports the previously identified major divisions of Class I (PCF-like) as well as the *CINCNATA* (CIN-like) and *CYC/tb1*-like clades of Class II genes. The limits of these major clades are clearly stated by the phylogenetic locations of Class I genes PCF1 (*O. sativa*), AtTCP6, AtTCP7, AtTCP9, AtTCP14, AtTCP20 and AtTCP21 (A. thaliana) and TIC (*A. majus*); Class II genes CINCNATA (*A. majus*), LA (Solaranum lycopersicum), AtTCP2, AtTCP3 and AtTCP4 (*A. thaliana*) and StTCP1 (Solanum tuberosum) as well as the *CYC/tb1*-like genes CYCLOIDEA and DICHTOMA (both from *A. majus*), PICYC (Piantago lanceolata), GhCyc2 (Gherba hybrida), AtTCP1, AtTCP12 and AtTCP18 (A. thaliana) and *tb1* (*Z. mays*). Also in agreement with these major clades is the presence of residues encoding the R domain in Class II candidates from groups 13, 15 (both CIN-like) and 18 (tb1-like) and residues targeted by miR319 exclusively found in candidates from Class II CIN-like genes in groups 12, 13, 14, 15 and 17.

The putative genes of Class I (PCF-like) retrieved from databases comprise almost half of TCP-like sequences identified in the genomes of *Z. mays* (17 out of 31), *S. bicolor* (14 out of 24) and *B. distachyon* (11 out of 20). The clades of Class II in these genomes are represented by nine CIN- and six *tb1*-like putative genes in *Z. mays*, while in *B. distachyon* and *S. bicolor* these CIN-like genes are represented by six and seven, respectively, and *tb1*-like genes by three sequences in both species.

The 17 TCP-like genes from Musaceae, Zingiberaceae, Orchidaceae and Amaryllidaceae that have been isolated so far belong either to PCF-like genes (groups 1, 3, 5 or 7) or to CIN-like genes (groups 10, 15 and 17). These sequences are distinctly grouped outside the clades formed by sequences from Poaceae and their phylogenetic relationships to these have variable levels of statistical support. This suggest that the small number of TCP-like sequences in public databases from species outside Poaceae, still does not provide enough information to establish accurate relationships of orthology between them and the rest of the OTUs in the alignment. Similarly, the small number of sequences available from non-grass monocots may explain why so far no representatives from these species are classified in the *CYC/tb1* group. The existence of this clade in non-grass monocots will be clarified after the identification of all members of the TCP family in the complete genome of a species from this group.

In the present phylogenetic analysis, groups 1, 3, 5, 7, 9, 15, 17 and 18 include both sequences from eudicots and monocots (Fig. 2). This suggests that nearly 50 % of the groups here defined were already present in the common ancestor of both groups. This is a conservative estimate based on the clades formed by those eudicot TCP-like genes for which there is experimental information and that are at least 40 % identical to monocot sequences in the dataset. In the future, a better prediction of the common groups of TCP-like genes in the common ancestor of eudicots and monocots will be enabled by the isolation and phylogenetic analysis of more members of this family from species of basal eudicots and non-grass monocots.

The correspondence between groups of amino acid identity \( \geq 64\% \) and well-supported clades observed in the phylogenetic analysis with maximum likelihood partially holds in the analysis performed with Bayesian inference (Fig. S1 in Supplementary Data), but completely breaks down in the distance-based analysis where clades of Class I and Class II TCP-like genes are not defined and only a few clades associated with groups of identity are well supported (Fig. S2 in Supplementary Data).

The phylogenetic analysis of the dataset encoding the TCP domain with the methods of maximum likelihood and Bayesian inference associated the sequences in groups of high identity, but gave relatively lower statistical support to the major subdivisions and clades from each class (Figs S3 and S4 in Supplementary Data). In contrast with the results of the distance analysis of the longer alignment, the corresponding analysis of the matrix encoding the TCP domain.
| Table 1. Main sequence features of the groups of monocot TCP-like genes with high average amino acid identity identified in this study |
|---|---|---|---|
| Group, TCP class Members* | Average pairwise identity† | Specific sequence motifs‡,§ |
| 1, I 5 0.74 ± 0.10 | KDRHSKVDGRGRRIRMPIICAARVFQLTRELGHKSDGQTIEWLLRQAEPSIIAATGTGTTPASFSTS; WSFAAAPEMMV; SDRHAKVAGRGRRVRIPAMVAARVFQLTRELGHRTDGETIEWLLRQAEPSIIAATGTGVTPEEAPPAAVPIGS; PYYTALL; PPxADEPP; TAAEENNN |
| 2, I 5 0.81 ± 0.05 | KDRHSKVDGRGRRIRMPIICAARVFQLTRELGHKSDGQTIEWLLRQAEPSIIAATGTGTTPASFSTS; WSFAAAPEMMV; SDRHAKVAGRGRRVRIPAMVAARVFQLTRELGHRTDGETIEWLLRQAEPSIIAATGTGVTPEEAPPAAVPIGS; PYYTALL; PPxADEPP; TAAEENNN |
| 3, I 5 0.76 ± 0.14 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 4, I 5 0.82 ± 0.06 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 5, I 5 0.75 ± 0.10 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 6, I 5 0.64 ± 0.09 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 7, I 5 0.64 ± 0.08 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 8, I 5 0.75 ± 0.06 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 9, II 9 0.66 ± 0.08 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 10, II 3 0.63 ± 0.25 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 11, II 3 0.70 ± 0.11 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 12, II 16 0.70 ± 0.10 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 13, II 4 0.79 ± 0.07 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 14, II 5 0.73 ± 0.12 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 15, II 5 0.54 ± 0.11 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 16, II 3 0.67 ± 0.06 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 17, II 7 0.70 ± 0.10 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |
| 18, II 16 0.82 ± 0.01 | KDRHTKVDGRGRRIRMPALCAARIFQLTRELGHKSDGETVQWLLQQAEPAIVAATGTGT; PAPATSTASPVSFSPPKSx; ITSSVLv3xFvPQxLRLv3xWIDCAA |

* Number of sequences in the monocot species here analysed.
† Average number of identical amino acid residues among pairs of sequences from monocot species.
‡ Motifs larger than five continuous residues identified in the 60 % consensus of each group.
§ The TCP domain is underlined, the R domain is in bold italic type and the residues corresponding to the sites targeted by miR319 are in bold type underlined.
Fig. 2. Phylogeny of TCP-like genes from monocots based on the analysis with PhyML of the nucleotide residues encoding the TCP and carboxyl domains. To delimit the clades of Class I (PCF-like sequences) and Class II (CIN-like and tb1-like) 22 sequences of experimentally and phylogenetically characterized TCP-like genes from dicots were included. The name of each sequence analysed includes the group of high identity where it belongs (Table 1), its name in the database where it was retrieved, the gene name if available and the full name of the species where it was isolated. The sequences from Oryza sativa japonica and Arabidopsis thaliana follow the nomenclature of Martín-Trillo and Cubas (2010). The numbers on each node are the Shimodaira–Hasegawa (SH)-like test indexes of statistic support provided by PhyML. Indicated with an arrow and bold type are the indexes ≥ 0.90 that support each clade containing the sequences of a group of high identity. The black bars parallel to the phylogeny show the limits of each of these groups and the inset shows in detail the relationships between the monocot members of group 18.
yields a phylogeny where clades grouping Class I and Class II genes, as well as some clades comprising highly similar sequences, are supported with bootstrap indices $\geq 70$ (Fig. S5 in Supplementary Data).

The expansion of TCP-like genes in the Poaceae was driven by ancient genome duplications

It was observed that the sequences of *O. sativa*, *B. distachyon*, *Z. mays* and *S. bicolor* in groups 2, 5, 6, 7 (PCF-like genes) and 9 (CIN-like genes) correspond to well-supported clades and form several subclades. Interestingly, two to four TCP-like genes from *Z. mays* are present in these clades as well as in those, comprising sequences from groups 3, 14, 17, 18, 19 and 20.

To determine whether these clade-specific expansions resulted from gene or genome duplication, the genomic context of all TCP-like genes from *B. distachyon*, *O. sativa*, *Z. mays* and *S. bicolor* was investigated in Phytozome, a comparative database for plant genomics. Nearly 68% (68 out of 99) of the TCP-like genes from these species are located in regions of highly similar gene composition (Table S2 in Supplementary Data). Specifically, genes within groups 1, 2, 3, 5, 6, 7, 8, 9, 13, 14, 16, 17, 18, 19 and 20 are located in one or two specific microsyntenic regions in the genomes of the four species analysed (Table S2 in Supplementary Data). Moreover, the genes located in a specific microsyntenic region correspond to the subclades in which some of these groups are divided (Figs 2 and 3; Table S2 in Supplementary Data).

Systematic analysis of the orthologous clades of TCP-like genes from *B. distachyon*, *O. sativa*, *Z. mays* and *S. bicolor* suggest that in their common ancestor the TCP family was already formed by at least 21 genes. Specifically ten of these genes are duplicated in groups 2, 5, 6, 7 and 9, and the remaining 11 genes are present in single copies (Fig. 3). The second major expansion of TCP-like genes in these species took place on the lineage of *Z. mays*, where additional paralogues increased the size of groups 2, 3, 5, 6, 9, 14, 18, 19 and 20 (Fig. 3). This analysis and the fact 67% of the TCP-like genes from these four species are located in microsyntenic regions suggest that two episodes of whole genome duplication (WGD) are behind the expansion of TCP-like genes in these species. The present analysis of the family of TCP-like genes is in line with recent findings on the occurrence of WGDs before the radiation of the grasses and, more recently, after the divergence of *Sorghum* and *Zeae* (Fig. 3; Table S2 in Supplementary Data). In this context, the first WGD may have taken place before the radiation of the grass family between 50 and 70 MYA as has been suggested by several studies (Kellogg, 2001; Paterson *et al.*, 2004; Tian *et al.*, 2005). The second episode was exclusive of the maize lineage and may have taken place between 5 and 12 MYA after the divergence of *Sorghum* (Gaut and Doebley, 1997; Blanc and Wolfe, 2004; Swigonová *et al.*, 2004; Paterson *et al.*, 2009). As a result of this second WGD the number of orthologous genes in *Sorghum*, *Oryza* and *Brachypodium* keeps a 1 : 2 relationship with those in the genome of *Z. mays* (Schnable *et al.*, 2009), as is observed in the TCP-like genes from these species.

The number of synonymous substitutions per synonymous site ($K_s$) was employed to date the events of complete genome duplication that gave rise to paralogous TCP-like genes in the Poaceae. Although the values of $K_s$ for the case of *O. sativa* (data not shown) generally agree with previous estimates (Navaud *et al.*, 2007), recent analysis of the genomes of *O. sativa* and *Sorghum bicolor* showed that in these species about 42% of all genes have an average GC content in the third codon position ($G*C_3$) higher than 75% (Tang *et al.*, 2010). This bias results in serious under- or over-estimation of $K_s$ values and sequences with these characteristics are not used for dating gene or genome duplications (Tang *et al.*, 2010). According to the present analysis with YN00 from the PAML package, all TCP-like genes from *Zea*, *Sorghum*, *Brachypodium* and *Oryza* have an average $G*C_3$ over 80% and therefore $K_s$ values obtained from them are inappropriate for dating the evolution of this family.

**DISCUSSION**

Evolution of TCP-like genes in monocots: challenges and opportunities

Several studies on the evolution of the TCP gene family in land plants (Navaud *et al.*, 2007; Martín-Trillo and Cubas, 2010) have firmly established that the family is subdivided into Class I (PCF-like genes) and Class II genes, which are subdivided into the ancient clade of CINCINNATA (CIN-like genes) and the angiosperm-specific ECE clade which contains CYC/tb1-like genes. However, the high degree of sequence diversity between eudicot and monocot TCP-like genes requires of a group-specific analysis that facilitates establishing relationships of orthology and developing a systematic framework to organize and compare the impending deluge of information from genomic-level studies. A step towards this goal is to characterize the sequence diversity of TCP-like genes in monocots and infer their phylogenetic relationships. This is the first overall alignment and phylogenetic analysis of TCP-like genes in monocots, with emphasis on the members of completely sequenced genomes of *O. sativa*, *S. bicolor*, *Z. mays* and *B. distachyon*.

To distinguish and characterize the phylogenetic structure of this gene family within the major clades of PCF-, CIN- and CYC/tb1-like genes, phylogenetic analyses were performed on the alignment of the nucleotide residues encoding both TCP and carboxyl domains. Sequence information corresponding to the region before the TCP domain was not included because this is not available for all the sequences retrieved.

The maximum likelihood approach here employed yielded a phylogeny that lends statistical support to the relationships within 20 groups of highly similar sequences and to several of the clades that associate them. The fact that many of the putative genes in these groups are found in microsyntenic regions of four genomes strengthens the conclusion that the putative genes in each group are orthologues.

The associations found on the data at the level of sequences (group-specific motifs), phylogeny (clades) and genomic level (microsynteny) are the basis of a system to identify, classify and compare new sequences as well as functional information. In this system, putative or recognized monocot TCP-like genes are assigned to one of eight clades of PCF-like genes or eight of CIN-like genes or three of CYC/tb1-like genes. The fact that most sequences from
species in Musaceae, Zingiberaceae, Orchidaceae and Amaryllidaceae already form distinct subclades or even individual groups suggests that, in the future, novel groups within the PCF-, CIN- and CYC/tb1-like genes might be identified. The system of classification proposed accommodates any new clades that may not belong in the groups initially defined. The use of a system based on an all-encompassing phylogeny and corresponding genomic information sets up a comparative context for investigating the evolution of TCP-like genes that will become broader as new monocot genomes are completed.

The first comparisons between the genomes of O. sativa and Musa acuminata (Zingiberaceae) support the existence of microsynteny (Lescot et al., 2008). Although the TCP-like genes from O. sativa and M. acuminata analysed here are not in microsyntenic regions, the first reports of microsynteny between these species suggest that, in the future, information about the genomic context of TCP-like genes in Poaceae may be comparable to a certain extent between distantly related species. However, the discovery and use of microsynteny to understand the evolution and function of this gene family in the monocots cannot be solely based on comparisons with members of the Poaceae. For instance, although TCP-like genes from Allium cepa (Asparagales) could be reliably aligned to those of species in the Poaceae, recently several groups reported absence of synteny between the genome of rice and preliminary information on the genomes of A. cepa and Asparagus officinalis (Asparagales) (Martin et al., 2005;
Retention and expansion of ancient duplicates is relevant for the development of new morphologies

It is a well-known fact that after WGD the genes encoding duplicated transcription factors tend to be retained more often than those with other roles (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004). This has already been documented in Z. mays (Schnable et al., 2009) and the analysis presented here indicates that several events of WGD drove the expansion of TCP-like genes in the four grass species investigated, but it is particularly relevant in Z. mays where the increase was of 48% over the average. The preservation of these duplicates during the process of allotetraploidization that gave rise to the maize genome may be explained by the theory of balanced gene drive. This theory predicts that after an event of WGD purifying selection preserves duplicate genes encoding transcription factors to maintain their interactions with other molecules or their concentration, which is a limiting factor for the expression of other genes (Papp et al., 2003; Birchler et al., 2001, 2005; Veitia, 2002; Freeling and Thomas, 2006). According to the proponents of balanced gene drive, the divergence of duplicated regulatory gene networks is essential for the evolution of novel developmental pathways and morphological complexity. In the context of the TCP family two corollaries from this theory can be tested experimentally: (1) those TCP-like genes that resulted from the last maize WGD have been preserved because they are dose-sensitive and functionally associated in the developmental regulatory network; (2) the recently duplicated TCP-like genes are part of a regulatory network whose divergence may be associated with the morphological divergence of Z. mays from its common ancestors. The theory of balanced gene drive might not only be helpful to interpret and test the evolutionary and functional association between duplicated TCP-like genes, but may also offer a different explanation of why and how other TCP paralogues persist (e.g. analysis of CYC and DICH by Hileman and Baum, 2003) and subfunctionalize in the context of strong purifying selection.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following. Fig. S1: phylogeny of putative TCP-like genes from monocots based on the analysis with MrBayes of the nucleotide residues encoding the TCP and carboxyl domains. Fig. S2: phylogeny of putative TCP-like genes from monocots based on a neighbor-joining analysis of the nucleotide residues encoding the TCP and carboxyl domains. Fig. S3: phylogeny of putative TCP-like genes from monocots based on analysis with PhyML of the nucleotide residues encoding the TCP domain. Fig. S4: phylogeny of putative TCP-like genes from monocots based on analysis with MrBayes of the nucleotide residues encoding the TCP domain. Fig. S5: phylogeny of putative TCP-like genes from monocots based on a neighbor-joining analysis of the nucleotide residues encoding the TCP domain. Text file 1: amino acid alignment of monocot TCP-like sequences. Table S1: list of sequences analysed. Table S2: list of sequences in each group of similarity ≥ 64% and the corresponding patterns of microsynteny as documented in the database Phytozome (w.phytozome.net).


