Historical profiling of maize duplicate genes sheds light on the evolution of C4 photosynthesis in grasses

Yao-Ming Chang, Chia-Lin Chang, Wen-Hsiung Li, Arthur Chun-Chieh Shih

Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA
Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan

Article history:
Available online 27 August 2012

Keywords:
Duplicate genes
C4 evolution
C4 photosynthesis

Abstract

C4 plants evolved from C3 plants through a series of complex evolutionary steps. On the basis of the evolution of key C4 enzyme genes, the evolution of C4 photosynthesis has been considered a story of gene/genome duplications and subsequent modifications of gene function. If whole-genome duplication has contributed to the evolution of C4 photosynthesis, other genes should have been duplicated together with these C4 genes. However, which genes were co-duplicated with C4 genes and whether they have also played a role in C4 evolution are largely unknown. In this study, we developed a simple method to characterize the historical profile of the paralogs of a gene by tracing back to the most recent common ancestor (MRCA) of the gene and its paralog(s) and then counting the number of paralogs at each MRCA. We clustered the genes into clusters with similar duplication profiles and inferred their functional enrichments. Applying our method to maize, a familiar C4 plant, we identified many genes that show similar duplication profiles with those of the key C4 enzyme genes and found that the functional preferences of the C4 gene clusters are not only similar to those identified by an experimental approach in a recent study but also highly consistent with the functions required for the C4 photosynthesis evolutionary model proposed by S.F. Sage. Some of these genes might have co-evolved with the key C4 enzyme genes to increase the strength of C4 photosynthesis. Moreover, our results suggested that most key C4 enzyme genes had different origins and have undergone a long evolutionary process before the emergence of C4 grasses (Andropogoneae), consistent with the conclusion proposed by previous authors.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

C4 plants, such as maize, have more efficient photosynthesis than C3 plants and can survive better in more extreme environments, such as high temperatures and arid lands. In the past several decades, the biochemical pathway of C4 photosynthesis has been extensively studied and a set of key C4 enzymes and transporters has been identified, including pyruvate Pi dikinase (PPDK), NADP-malic enzyme (NADP-ME), carbonic anhydrase (CA), dicarboxylate transporter (DCT2), PEP carboxylase (PEPC), rubisco small subunit (rbcS), and oxoglutarate/malate transporter (OMT1) (see Hibberd and Covshoff, 2010). With these key enzymes and transporters, C4 plants can effect a massive efflux of C4-related metabolites between mesophyll and bundle sheath cells and between the cytosol and organelles in the cells (see Weber and von Caemmerer, 2010).

It has been recognized that C4 plants evolved from C3 plants through a series of complex evolutionary steps (Dengler and Taylor, 2000; Gowik and Westhoff, 2011; Sage, 2004). From the phylogenies of C4 plant lineages, several studies have estimated that the evolution from C3 to C4 has occurred independently at least 31 (Kellogg, 1999), 45 (Sage, 2004), or 62 times (Sage et al., 2011). However, some authors have questioned the estimation procedure based solely on phylogenetic analysis (Christin et al., 2009) and proposed to integrate other data, such as ecosystem data (Edwards et al., 2010). On the other hand, using key C4 enzyme genes as a basis, some studies have considered the phylogenetic trees of homologous genes in different C3 and C4 plants and concluded that the evolution of C4 photosynthesis is a story of gene/genome duplications and subsequent modifications of gene function (Christin et al., 2009; Monson, 2003; Wang et al., 2009).

If whole-genome duplication has contributed to the evolution of C4 photosynthesis, many other genes should have been duplicated together with these C4 genes. A similar comment applies
to duplications that involved large DNA segments (i.e., segmental duplications). However, which genes were co-duplicated with key C4 enzyme genes and whether they have also played a role in C4 evolution are largely unknown.

Identifying genes that were co-duplicated is difficult because duplicate genes may have been derived from single-gene duplications, segmental duplications, or whole genome duplications (Blanc and Wolfe, 2004; Bowers et al., 2003; Hughes, 2002; Li, 1982; Roth et al., 2007; Yu et al., 2005). Instead, it is simpler to infer whether two groups of duplicate genes (paralogs) share similar evolutionary histories; that is, whether they can be traced back to the same common ancestors.

In this study, we propose a simple method to characterize the temporal or historical profile of the paralogs of a gene; that is, we trace back each paralog of the gene to their most recent common ancestor (MRCA), and then count the number of paralogs at each MRCA. We then cluster the genes in a genome into clusters with similar historical profiles and infer their functional preferences (enrichments). We focus on maize genes because maize is a familiar C4 plant and its genome has been sequenced and annotated. We are particularly interested in maize genes that show similar duplication profiles with those of the key C4 enzyme genes because this analysis may help us infer the genes that have contributed to the evolution of C4 photosynthesis. Interestingly, the functional preferences of our identified C4 gene clusters are similar to those identified by an experimental approach in a recent study (Gowik et al., 2011).

2. Materials and methods

2.1. Maize data and genome information for MRCA inference

The maize genes and their paralog data were downloaded from MaizeSequence (release 5a.59) (http://maizesequence.org/index.html) and Ensembl Plants (release 14) (http://plants.ensembl.org/index.html). The Ensembl group has inferred the maize duplicate paralogs and their most recent common ancestors (MRCA), using the method of (Vilella et al., 2009). Nineteen plant genomes were used to infer the homologs and the MRCA of paralogs in Ensembl Plants (release 14), including Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Glyceria max, Oryza glaberrima, Oryza sativa Nipponbare, Oryza indica 93-11, Physcomitrella patens, Populus trichocarpa, Sorghum bicolor BTx623, Vitis vinifera, Selaginella moellendorffii, Chlamydomonas reinhardtii, Cyanidioschyzon merolae, Brassica rapa, Oryza brachyantha, Setaria italica, Solanum lycopersicum, and Zea mays. These 19 species include one fern, one green alga and one red alga, and the inferred MRCA includes Zea mays (Zm), Andropogoneae (An), Panicicoideae (Pa), Poaceae (Po), Magnoliophyta (Ma), Tracheophyta (Tr), Embryophyta (Em), Viridiplantae (Vi) and Eukaryota (Eu). These data were all downloaded from http://plants.ensembl.org/biomart/martview/.

2.2. Paralog and MRCA inference by Ensembl

In Ensembl, the paralogs were inferred by a three-step pipeline (http://plants.ensembl.org/info/docs/compara/homology_method.html): (a) loading the longest translation of each gene from all plant species in Ensembl, (b) applying the WU-Blastp and the SmithWaterman algorithms to align every gene against every other in the genomes (both within and between species), and (c) building a sparse graph of gene relations based on the Blast scores and generating clusters using “hcluster_sg”, a hierarchical clustering software. In each cluster, the genes within the same species are called paralogs and designated as a paralog group.

The MRCA of paralogs were inferred by another three-step pipeline in Ensembl: (a) building a multiple alignment of the protein sequences using a combination of multiple aligners and then determining the consensus by MCoffee for each paralog cluster, (b) building a phylogenetic tree by TreeBeST using the coding DNA sequence (CDS) back-translation of the protein multiple alignment from the original DNA sequences for each aligned paralog cluster, and (c) inferring gene pairwise relations of orthology and paralogy types for each gene tree. In the last step, several criteria were used to determine each internal node as a duplication or speciation node. If two genes are orthologs in a gene tree, their most recent internal node is assigned as a speciation node. If two genes are paralogs, their most recent internal node is assigned as a duplication node. In a gene tree, each node represents one ancestral taxon. However, many speciation nodes have not yet been defined. In this case, each of the internal nodes was assigned to a nominated taxonomic node by looking at the commonly shared ancestors of children nodes. As a result, many internal nodes may share the same taxonomic node. For more details, refer to http://plants.ensembl.org/info/docs/compara/homology_method.html and (Vilella et al., 2009).

2.3. Historical profiling of paralogs

For each gene that has paralogs, we defined the duplication profile vector (DPV) as a vector in which each element represents the number of paralogs of the gene at a MRCA; in the present study the vector has nine dimensions because there are nine MRCA (see above).

The original idea and the meaning of DPV are described below. When a gene tree includes hundreds of homologous genes, it is difficult to make sense of a fully expanded tree with a large number of terminal nodes. In Ensembl, a gene tree shown at the website is first to collapse all its sub-trees except for that of a query gene as an initial tree for visualization and interaction. For example, GRMZM2G149481, GRMZM2G045057, and GRMZM2G127949 are three within-species paralogs of GRMZM2G129413 in the maize genome. A complete maximum likelihood phylogenetic tree has been constructed for their orthologs and other non-within-species paralogs; there are 138 genes in total and the gene tree ID is ENSGT0014000000699. If we only look at their initial gene trees, we can find that those of GRMZM2G129413 and GRMZM2G149481 are very similar but both are quite different from those of GRMZM2G045057 and GRMZM2G127949 (Supplementary Fig. S1). That is, the initial gene trees of GRMZM2G129413 and GRMZM2G149481 have the same paralog distributions at the internal nodes, in contrast to the different distributions of GRMZM2G045057 and GRMZM2G127949. Thus, comparing initial gene trees of different genes can help infer whether these genes have undergone similar evolutionary histories.

Next, we explain how to transform an initial tree into a linear vector. Supplementary Fig. S2a shows an almost fully expanded tree of ENSGT0014000000699 that includes GRMZM2G129413 and its four within-species paralogs located in three sub-trees. After collapsing these sub-trees, we obtain the initial gene tree of GRMZM2G129413 that includes three clades where Clades I, II, and III have two, one, and one paralog, respectively (Supplementary Fig. S2b). Then, the initial tree can be collapsed into a linear shape without losing any information (Supplementary Fig. S2c). The MRCA of the internal nodes of speciation have already been given at the Ensembl website. Thus, the DPV of GRMZM2G129413 is (0, 0, 0, 2, 0, 0, 0, 0, 0) and the corresponding MRCA are Zm, An, Pa, Po, Ma, Tr, Em, Vi and Eu, respectively.

2.4. Reliability assessment

For each DPV, we used two measures to assess their reliability: the average bootstrap value and the average duplication
consistency score. The average bootstrap value is the average of all bootstrap values of the internal nodes visited by each target gene tracing back to the root of the tree. The duplication consistency score is based on the assumption that duplication followed by reciprocal complementary gene losses on the left and right branches of a duplication node is unlikely and it is defined as the ratio of the intersections of species between left and right branches over the unions of species between left and right branches (Vilella et al., 2009). In this study, we defined the average duplication consistency score as the average of all duplication consistency scores of the internal nodes with maize duplication events that were visited by the target gene tracing back to the root of the tree.

Both the bootstrap values and the duplication consistency scores have been calculated already and can be downloaded from the Ensembl ftp site. Thus, the average bootstrap values and duplication consistency scores can be calculated directly. As an example (Supplementary Fig. S2), the average bootstrap value and duplication consistency score of the DPV of GRMZM2G129413 are 64% and 60%, respectively. All the values and scores for each DPV can be found in Supplementary Table S1.

2.5. Profile clustering

The clustering tool we used was gCLUTO (Graphical Clustering Toolkit) and was downloaded from http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview. This tool provides different kinds of hierarchical clustering methods; a repeated bisection method was selected in this study. We calculated the Pearson correlation between each pair of DPVs of the genes under study and used the correlation values to cluster the DPVs. To obtain a reliable number of clusters, a scalar quantity, called the Figure of Merit (FOM) (Yeung et al., 2001), was used to assess the quality of the clustering algorithm and to determine the optimal number of clusters. Let $k$, $R(x,e)$, and $\mu_C(e)$ be, respectively, the total number of clusters, the expression level of gene $x$ under condition $e$, and the average expression level in condition $e$ of genes in cluster $C$. FOM is defined as:

$$FOM(e, k) = \frac{1}{R} \sum_{i=1}^{k} \sum_{x \in C_i} (R(x,e) - \mu_C(e))^2$$

where $n$ is the total numbers of genes. The aggregate figure of merit for all conditions is

$$FOM(k) = \sum_{e=1}^{m} FOM(e,k),$$

where $m$ is the total numbers of experimental conditions of a typical gene expression data set. $FOM(k)$ is a measurement of the total variance among all within-cluster genes. If the value of $FOM(k)$ is small, the total similarity of all within-cluster genes is high (e.g. the quality of the clustering algorithm is high). FOM is useful for determining the best of input parameters for a clustering algorithm and has been widely applied to many microarray and deep sequencing transcriptomic data analyses (Li et al. 2010; Gowik et al., 2011; Pick et al. 2011).

In this study, we redefined $R(x,e)$ and $\mu_C(e)$ as the total paralog number of gene $x$ at MRCA $e$ and the average paralog number of genes in cluster $C_i$ respectively. In addition, $n$ in Eq. (1) and $m$ in Eq. (2) are the total numbers of genes and MRCA$s$, respectively. Then, we used a $k$-means algorithm in the Mev program (http://www.tm4.org/mev) by assigning various cluster numbers to calculate the FOM values. The program executed the $k$-means algorithm for each $k$ ($1 \leq k \leq 50$) with random initialization 20 times and then calculated the mean adjusted FOM values and standard deviations. The result shows that the value of the adjusted FOM run decreases steeply until the number of clusters reaches between 20 and 30 and after that the rate of change levels off (Supplementary Fig. S3). This suggests that the optimal cluster number of our data is around 25.

The $k$-means clustering algorithm is a good tool for the estimation of FOM values because using different initializations it can assess the variations of clustering results. However, the generated clustering results are usually not repeatable because different initializations may have somewhat different results. Thus, in the later step we used a repeated bisection approach to obtain the final clustering result. A repeated bisection approach is a top down hierarchical clustering algorithm that separates samples with the longest distance. The two major reasons for using this algorithm are: (a) the clustering result is deterministic because there is no initialization problem; and (b) any valid measure of distance can be used and the distribution of input data is not limited to a normal distribution.

2.6. Functional annotation

The annotation tool used was MapMan (Thimm et al., 2004) (http://mapman.gabipd.org/). The annotations with the functional category of “not assigned.unknown” were excluded in this study. We then used the Chi-square test to examine whether a function was significantly over-represented in a selected set of genes against the set of all genes with MapMan annotations ($p<0.01$).

3. Results

3.1. Duplication profile of maize genes

In the downloaded Ensembl maize database, there were 110,028 genes, including nucleus, chloroplast, and mitochondrial genes. Only 45,069 genes have been inferred to have paralogs, which can be divided into 8825 paralog groups. For those genes without paralogs (~60% of the total genes), only 3134 genes (i.e., 4.8%) have orthologs in the other 18 plant species (data not shown). In contrast, the majority of the genes with paralogs (~75%) have orthologs in the other 18 species. We considered the genes with at least one paralog and recorded the number of paralogs at each of the nine MRCA$s$. If a gene has many paralogs most of which can be traced back to a recent MRCA, then gene duplication was very active in recent times. In contrast, if most paralogs can be traced back to an old MRCA, then the majority of duplication events related to this gene were old. In the Ensembl Plants database, the MRCA$s$ used for maize genes are Zea mays (Zm), Andropogoneae (An), Panicoideae (Pa), Poaceae (Po), Magnoliophyta (Ma), Tracheophyta (Tr), Embryophyta (Em), Viridiplantae (Vi) and Eukaryota (Eu), from the latest to the oldest. The vector that records the paralog numbers at these MRCA$s$ is called the duplication profile vector (DPV).

Fig. 1 gives an example of gene duplication profiling. Fig. 1a shows the phylogeny of 10 maize PEPC duplicate genes, including the maize PEPC1 gene (NM_001111948; GRMZM2G083841). If we trace back from the terminal node of GRMZM2G083841 to the root of the phylogenetic tree, the paralog numbers of GRMZM2G083841 at the nine MRCA$s$, from the latest to the oldest, are 2, 0, 2, 2, 0, 1, 2 and 0 respectively. Fig. 1b demonstrates the sub-tree notation of GRMZM2G083841 with collapsed branches where the number of paralogs at a MRCA was marked beside the internal node (a MRCA taxon). Since there is no duplication event at the MRCA$s$, An, Pa, Tr, and Eu, we used “0” to represent each of these nodes. Therefore, the DPV of GRMZM2G083841 was (2, 0, 2, 2, 0, 1, 2, 0). Similarly, the DPV$s$ of the three paralogs GRMZM2G074122, GRMZM2G473001, and GRMZM2G384597 are (1, 0, 0, 3, 2, 0, 1, 2), (0, 0, 0, 1, 5, 0, 1, 2, 0), and (0, 0, 0, 0, 0,
These vectors clearly show that the duplication profiles of GRMZM2G083841, GRMZM2G074122, and GRMZM2G473001 are similar to each other, but very different from that of GRMZM2G384597 (Fig. 1f). Note that the paralog number distribution at the nine MRCA for a gene can be very different from those of other duplicate genes in the same paralog group, implying that genes in the same paralog group may have undergone different duplication events. However, genes in different paralog groups may have similar duplication profiles. For example, DCT2 (AB112938; GRMZM2G086258) and TIL1 (TILTED 1, GRMZM2G426847) have 7 and 8 paralogs, respectively, and their phylogenies provided by Ensembl are somewhat different (Fig. 2a and b). However, in terms of the paralog distributions at the MRCA, these two genes may have undergone similar patterns of duplication events (Fig. 2c and d) as can be seen from their DPVs.

The DPVs can serve the following purposes. First, from the DPV of a gene one can infer whether it has been derived from an ancestor that is also the ancestor of many paralogs. Second, by comparing the DPV of a gene with those of others one can infer whether the paralog group of this gene had a relatively high duplication rate. Third, from the DPVs, one can infer which MRCA has the largest number of descendent duplicate genes. Fourth, one can cluster the genes by their DPVs and examine whether the genes in a cluster have functional enrichments (see later).

3.2. Clustering DPVs of maize genes

We clustered the maize genes under study into 25 clusters and calculated the average vector of all DPVs of the genes in the same cluster, called the mean DPV of the cluster (Fig. 3). Since the average of all values in the mean DPVs is 1.55, we assigned a name to each cluster by using the name of the oldest MRCA with an element >1 and also the name of the second oldest MRCA with an element >1 if it exists to emphasize the paralog numbers above the average at the MRCA. For example, if the mean DPV of a cluster is (0, 0, 0, 2, 0, 0, 0, 0, 0), we call it “Zm” because only Zm had an element >1. As another example, if the mean DPV of a cluster is (0, 0, 0, 0, 0, 0, 0, 0, 0), we call it “Eu–Po” because the oldest MRCA that had an element >1 was Eu and the second MRCA that had an element >1 was Po. As another example, if the mean DPV of a cluster is (0, 0, 0, 0, 0, 0, 0, 0, 0), we call it “Zm” because only Zm had an element >1. In some cases, two clusters had the same assigned names. For example, if the mean DPV of two clusters were (0, 0, 0, 0, 0, 0, 0, 0, 0) and (0, 0, 0, 0, 0, 0, 0, 0, 0), their...
assigned names were both Ma–Po. To distinguish between them, we called the first cluster “Ma–Po1” and the second cluster “Ma–Po2.”

Using the earliest MRCA of a non-zero paralog number in the mean DPVs, we divided the clusters into seven groups (Fig. 3). The proportions of gene numbers among these clusters are non-uniformly distributed (Supplementary Fig. S4). Over 65% of the genes under study belong to five clusters, Zm (37.3%), Eu (8.6%), Po (7.8%), Ma (7.6%), and Em (6.5%). In cluster Zm, no paralogs of the genes could be traced back to Andropogoneae, so some of these paralogs may be related to the specific phenotypes of maize and its relatives. The genes in the Po cluster do not seem to have duplicate genes that arose after the emergence of Andropogoneae, but those in Po–Zm1 and Po–Zm2 still continued to duplicate new copies. A similar comment applies to the genes in the Ma cluster.

### 3.3. Functional analysis of the genes in each DPV cluster

We used MapMan to check the functional enrichments of the genes in each DPV cluster. In total, 20,666 genes have been annotated by MapMan. We first checked whether the number proportions of all paralogs and the paralogs with MapMan annotations in each cluster were similar. After the comparison, we found that the top five clusters with the highest proportions in these two sets are the same. However, in the Zm cluster the proportion of the genes with annotations is dramatically decreased from 37% to 11% (Supplementary Fig. S4). It indicates that many paralogs in this cluster have not been annotated yet. For each annotated functional category, we calculated the ratio of the number of paralogs in that category over the total number of paralogs and tested the statistical significance for each cluster (Supplementary Table S2).
Among the 25 clusters, only four clusters (Eu, Em, Ma, and Zm) have three or more over-represented functional categories (each paralog ratio >30%; \( p < 0.01 \)) (Fig. 4). In cluster Eu, fermentation, OPP (Oxidative Pentose Phosphate), protein folding, amino acid activation, TCA/ORG (tricarboxylic acid/organic acid) transformation, nitrogen metabolism, and RNA processing are over-represented. Cluster Em has the functional preference in protein glycosylation, biodegradation of xenobiotics, and auxin hormone metabolism. In cluster Ma, metal handling, cytokinin and salicylic acid hormone metabolism are over-represented. In cluster Zm, the over-represented functional categories include sulfur assimilation, C1-metabolism, mitochondrial electron transport/ATP synthesis, protein assembly and cofactor ligation, tetrapyrrole synthesis and glycolysis. These observations indicate that the genes in different clusters have their own specific functions.

We also checked whether several important functions related to C4 plants were significantly more abundant in some DPV clusters. Photosynthesis is an important function in plants. In the maize genome, there are 371 genes related to photosynthesis annotated by MapMan and 306 of them have paralogs. The numbers of photosynthesis-related genes in clusters Eu-Zm, Ma-Po1, Zm, and “Others” are significantly higher than expected from the background distribution of total genes with MapMan annotations (all \( p < 0.01 \); Fig. 5a).

Auxin biosynthesis in plants is complex and its de novo biosynthesis plays an essential role in virtually every aspect of plant development (Zhao, 2008, 2010). In the maize genome, there are 315 auxin-related genes and 300 of them have paralogs. We checked the distribution of the 300 genes in the DPV clusters. Relative to the background distribution, the auxin-related gene numbers are significantly higher in clusters Em–Ma1, Em and Ma–Po2 but lower in cluster Eu (all \( p < 0.01 \); Fig. 5b).

Some previous studies have indicated that C4 plants evolved to tolerate adverse environments such as windy, dry, high light or extreme temperatures (Sage et al., 2012). It implies that genes related to abiotic stress played an essential role during the evolution from C3 to C4. Moreover, the cell wall of bundle sheath cells in C4 plants became thicker to prevent CO2 leaking from bundle sheath cells to mesophyll cells but it also needs to be loosened and modified for cell enlargement. Therefore, we analyzed the genes related to abiotic stress and cell wall (Fig. 5c and d, respectively). Both gene numbers are significantly higher in clusters Em and the numbers of abiotic stress-related genes are also high in clusters Em–Zm and Po–Zm2. Moreover, the numbers of cell wall-related genes are high in clusters Eu–Em, Eu–Ma1, Em–Ma2, and Ma–Po1.

In summary, gene duplications at different MRCAs have different functional preferences.

### 3.4. Duplication profiles of key C4 enzyme genes

Gene duplication is regarded as the major source for functional innovation in C4 plant evolution (Monson, 2003), though it did not occur in all C4 genes (Wang et al., 2009). Several key C4 photosynthesis specific genes in maize have been identified in the literature,
including NADP-ME, DCT2, OMT1, PEPC, CA, PCK, rbcS, PPDK and PPDK-RP (see Table 1 for the full gene names). However, PPDK-RP (GRMZM2G131286) and rbcSs (GRMZM2G098520 and GRMZM2G113033) have only one identified paralog in the Ensembl Plants database. Thus, these genes were not considered in this study.

The first duplications for PPDK, NADP-ME, DCT2, PEPC, CA, and OMT1 all occurred before the MRCA Tracheophyta and, except for OMT1, subsequently more duplications occurred at younger MRCA. On the other hand, those for PCK, RRDK-RP, and rbcS occurred at the three latest MRCA. Poaceae, Andropogoneae, and
275, 2915 and 832 genes, respectively (Supplementary Fig. S5).

4. Discussion

In this study, we defined “duplication profile” in terms of evolutionary history and identified genes that shared similar duplication profiles; we used the inferred duplicate genes and their MRCA provided by Ensembl Plants. Interestingly, many duplicate genes have only a single MRCA and over 37% of these genes have their paralogs found only in the maize genome. These genes and their paralogs would be more likely to be related to maize-lineage specific phenotypes but less likely to be related to C4 evolution because C4 photosynthesis evolved before the emergence of maize. When studying the paralog profiles of eight key C4 photosynthesis enzymes or transporters, we found that most C4 genes were separated in different duplication profile clusters. Moreover, the MRCA of these genes could be traced back to Embryophyta (Viridiplantae and Eukaryota) and the oldest MRCA of PPDK and PPDK-RP were Poaceae and Andropogoneae, respectively. These results suggest that most key C4 genes originated at different evolutionary times and have undergone a long evolutionary process before the emergence of C4 grasses (Andropogoneae), a view that is consistent with the conclusion of Wang et al. (2009).

Sage (2004) has proposed a model of C4 photosynthesis evolution based on the comparative analyses of C3, C4 and C3–C4 species. In his model, to create and maintain large numbers of duplicated genes is the general prerequisite during C4 evolution. In the maize genome, there are about 20,000 paralogs whose MRCA are Poaceae or older, providing large numbers of duplicate genes for evolutionary modifications. In the present study, we identified many genes that share similar duplication profiles with known key C4 genes and we determined their over-represented functional categories with MapMan annotations. Interestingly, most of these categories were consistent with the functions...
required for the C4 photosynthesis evolutionary model (Sage, 2004). For example, the first step toward C4 evolution in this model was the development of the kranz anatomy that can efficiently concentrate CO2 from mesophyll to the adjacent bundle sheath cells. To achieve the particular structure of compartmentation between mesophyll and bundle sheath, the vein density had to be increased in leaves. Previous studies showed that the vein formation is triggered by polar auxin flow mediated by auxin efflux carriers (Scarpella et al., 2006). In our study, auxin metabolism is one of the over-represented functional categories of genes that share similar DPVs with C4 enzyme genes (Fig. 6). Moreover, the ARF (auxin response factor) transcription factor family is also over-represented (Fig. 6). Two paralogous genes of the ARF transcription factor MP (MONOPTERO) that can directly and positively regulate the procamium differentiation and vascular formation (Donner et al., 2010) were also in the same cluster of Em. In addition, many PIN family genes (PIN4, PIN5, and PIN8), the auxin efflux carriers, were also in the same cluster of Em (Supplementary Table S1).

Moreover, the 2nd phase in Sage’s model is that the enlargement of the bundle sheath cells and the increase in the number of organelles in this tissue might be a secondary effect of the higher vein density. When a cell was in progress of expansion or enlarge-ment, the cell wall needs to be loosened, modified, and then synthe-sized. In addition, the bundle sheath cells in C4 species became thicker to prevent CO2 escape. In our study, many functional categories related to cell wall (degradation, modification, synthesis, etc.) were also over-represented (Fig. 6). Finally, we found that the G2-like family transcription factor G2 belongs in the cluster of OMT1 (Supplementary Table S1) and it had been shown to regulate the development of chloroplast in maize bundle sheath cells (Hall et al., 1998; Rossini et al., 2001).

Phase 3 in the model is to establish a photosynthetic CO2 pump. This phase is an important intermediate step on the way toward C4 photosynthesis. In the present study, both the functional category of photosrespiration and C4 carbon concentrating mechanism were over-represented (Fig. 6). Since three key C4 genes NADP-ME, CA, and PPDK were annotated in the category of carbon concentrating mechanism, many genes in this functional category have similar duplication profiles, as we expected.

The categories of major CHO metabolism, photosynthesis light reaction, abiotic stress drought/salt, and signaling of light were also over-represented in the clusters including at least one or more C4 genes. Interestingly, the photosynthesis light reaction and light signaling have been identified as the significantly up-regulated functional categories in C4 or C3–C4 Flaveria species (Gowik et al., 2011). Moreover, C4 plants should have developed some particular mechanisms that can grow faster and better than C3 plants under abiotic stresses, such as a high temperature or an arid land. Indeed, we found that abiotic stress was over-represented in our data.

Most genes in the functional category of RNA regulation of transcription are transcription factors. In addition to the ARF and G2-like families, which may be involved in the C4 photosynthesis evolution mentioned above, the AP2/EREBP, CO-like, ARR, GRAS, SBP, C2H2, and zf-HD families were also over-represented. Some of these TF genes may also be related to the evolution of C4 photosynthesis.

Acknowledgments

Morris Goodman was one of the first scientists to use protein sequences to study the molecular evolution of duplicate genes. We have been inspired by his studies and we dedicate this paper to him. The study was supported by Academia Sinica and National Science Council, Taiwan (NSC 99-2621-B-001-003-MY3).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.ympev.2012.08.009.

References


Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J.,
Zhang, Y., Li, R., Xu, Z., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi,
J., Liu, J., Li, H., Li, L., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Li, D.,
Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J.,
Chen, F., Wu, S., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Xu, J., Zhou, Y., Yu, Y.,
Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X.,
Fang, L., Huang, X., Su, Z., Tong, W., Tong, Z., Ye, J., Wang, L., Lei, T., Chen, C.,
Chen, H., Huang, H., Zhang, F., Li, N., Zhao, C., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W.,
Zhao, Y., 2008. The role of local biosynthesis of auxin and cytokinin in plant
Plant Biol. 61, 49–64.