Expressed sequence tag analysis in Cycas, the most primitive living seed plant

Eric D Brenner*, Dennis W Stevenson*, Richard W McCombie†, Manpreet S Katari†, Stephen A Rudd‡, Klaus FX Mayer‡, Peter M Palenchar§, Suzan J Runko*, Richard W Twigg*, Guangwei Dai¶, Rob A Martienssen¥, Phillip N Benfey# and Gloria M Coruzzi§

Addresses: *The New York Botanical Garden, 200th Street and Kazimiroff, Bronx, NY 10458-5126, USA. †Genome Research Center, Cold Spring Harbor Laboratory, 500 Sunnyside Blvd, Woodbury, NY 11797, USA. ‡Institut für Bioinformatik (IBI), GSF National Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany. §§New York University, Department of Biology 1009 Main Building, New York, NY 10003, USA. ¶Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012-1185. ¥Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. #Biology Department, Duke University, Box 91000, Durham, NC 27708, USA.

Correspondence: Eric D Brenner. E-mail: ebrenner@nybg.org

© 2003 Brenner et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article’s original URL.

Abstract

Background: Cycads are ancient seed plants (living fossils) with origins in the Paleozoic. Cycads are sometimes considered a 'missing link' as they exhibit characteristics intermediate between vascular non-seed plants and the more derived seed plants. Cycads have also been implicated as the source of 'Guam's dementia', possibly due to the production of S(+)-beta-methyl-alpha, beta-diaminopropionic acid (BMAA), which is an agonist of animal glutamate receptors.

Results: A total of 4,200 expressed sequence tags (ESTs) were created from Cycas rumphii and clustered into 2,458 contigs, of which 1,764 had low-stringency BLAST similarity to other plant genes. Among those cycad contigs with similarity to plant genes, 1,718 cycad 'hits' are to angiosperms, 1,310 match genes in gymnosperms and 734 match lower (non-seed) plants. Forty-six contigs were found that matched only genes in lower plants and gymnosperms. Upon obtaining the complete sequence from the clones of 37/46 contigs, 14 still matched only gymnosperms. Among those cycad contigs common to higher plants, ESTs were discovered that correspond to those involved in development and signaling in present-day flowering plants. We purified a cycad EST for a glutamate receptor (GLR)-like gene, as well as ESTs potentially involved in the synthesis of the GLR agonist BMAA.

Conclusions: Analysis of cycad ESTs uncovered conserved and potentially novel genes. Furthermore, the presence of a glutamate receptor agonist, as well as a glutamate receptor-like gene in cycads, supports the hypothesis that such neuroactive plant products are not merely herbivore deterrents but may also serve a role in plant signaling.
Background

The Cycadales (cycads) are the most primitive living seed plants and have endured over 270-280 million years since their origins in the Lower Permian [1,2]. Cycads have a fern or palm-like appearance, largely due to their pinnately compound leaves (Figure 1a,b). Unlike ferns or palms, however, cycads belong to the gymnosperms, or non-flowering seed plants. Of the four orders that comprise the gymnosperms, the Cycadales are considered to be the most ancestral compared to Ginkgoales, Gnetales and Coniferales (Figure 2) [3,4]. Cycads (non-flowering seed plants) exhibit a number of characteristics that reflect their evolutionary position between ferns (non-seed plants) and angiosperms (flowering seed plants). Such characteristics include pollen tubes, which release motile sperm before fertilization; dichotomous branching (versus axillary branching in higher plants); and ovules, which contain a large, free-nuclear megagametophytic stage, that are borne on the margins of leaf-like megasporophylls [5-7]. These characteristics, among others, place cycads at a key node in plant evolution.

In addition to their evolutionary importance, cycads have also been studied in the field of medicine, because they produce neurotoxic compounds. In particular, cycads produce a secondary compound, BMAA (S(+)-beta-methyl-alpha, beta-diaminopropionic acid), which has been implicated as the possible cause of Guam’s dementia [8]. This disorder occurs among the indigenous Chomorro people, who ate cycads as food, and now suffer from Alzheimer’s and Parkinson’s dementia [9-11]. BMAA production is unique to cycads, where it has been used as a monophyletic character in plant classification [7]. It is present in both seeds and leaves of all genera of the Cycadaceae [12]. BMAA is neurotoxic in mammals [9,13] because of its excitotoxic action as an agonist of glutamate receptors (GLRs) [14]. The discovery of GLR-like genes in *Arabidopsis* suggests that plant-derived GLR agonists, as well as acting as potential deterrents to herbivores, might also operate in signaling during plant growth and development, by interacting with native plant GLRs [15]. In partial support of this hypothesis, BMAA was shown to affect the development of *Arabidopsis* and consequently was used in a pharmacologically-based genetic screen to isolate mutants in a putative GLR pathway in *Arabidopsis* [16].

Despite the importance of cycads in the study of plant evolution, and their role in neurological disorders in humans, nothing is known about the genes responsible for these traits - primarily because cycads are recalcitrant to genetic analysis. Unlike genetically tractable plants such as tomato, maize and *Arabidopsis*, cycads are dioecious (male and female organs on separate plants), produce a limited number of seeds and take up to 30 years to become reproductive. Furthermore, cycad genomes are large (20,000-30,000 million base-pairs (Mbp)) [17,18] compared to *Arabidopsis* (125 Mbp) [19]. Consequently, cycads have remained outside the realm of both traditional genetic studies and modern genome-sequencing initiatives. Fortunately, recent advances in plant genomics [20,21], provide new tools to study genetically complex species such as cycads. In particular, the availability of the
Cycads are the sister group to the seed plants. A phylogenetic tree shows that cycads (highlighted) are the least derived of the seed plants. Cycads are believed to be the oldest extant seed plants.

To begin a survey of expressed genes of cycads, the genus Cycas was chosen for expressed sequence tag (EST) analysis because Cycas is at the basal node - that is, the sister taxon to the rest of the Cycadales [25-27]. Furthermore, the species Cycas rumphii Miq. was selected for this analysis as it is suspected to be the dietary cause of Guam’s dementia. It has been established that in C. rumphii, from which the EST library was made, BMAA levels are nearly 0.1 mg/g tissue [28]. Because of its evolutionary position as a key node within the plant kingdom, as well as its medicinal significance to humans, Cycas is ideally suited for genomic prospecting [29].

Here, we describe the construction of a cycad EST database from RNA of young C. rumphii leaves. Using this database, our comparison revealed conserved genes, including those involved in development and signaling in present-day flowering plants. Our analysis defined a set of cycad clones that have no similarity to any known angiosperm genes, but possess similarity only to genes of other gymnosperms. Furthermore, as a first step to understanding the function of neurtotoxins produced in cycads, we defined a number of candidate genes that encode putative enzymes involved in the biosynthesis of BMAA, as well as a cycad GLR-like gene, the suspected target of BMAA action in animal brains. These cDNA tools will be useful to test whether BMAA, which has been postulated to serve as an herbivore deterrent [5], also acts to regulate GLR function in plants.

Results

Construction of a cDNA library from Cycas rumphii

At maturity, C. rumphii leaves can reach up to 3 meters in length (Figure 1a). The tissue used in this study consisted of 10 to 40 cm of the immature leaf terminus protruding from the crown collected shortly after emergence (Figure 1b). Immature leaves consist of a petiole, a central rachis and circinate leaflets composed of both expanding and meristemetic cells [30]. RNA extracted from this tissue was used to construct a cDNA library from C. rumphii. Size fractionation was used to enrich for full-length cDNAs during library construction. It was determined that 53% of the cDNA clones were over 500 bp long. From this cDNA library, 4,210 sequence reads (ESTs) were generated. The majority of these reads (3,917) were generated from the 5' end of the cDNA; however, a small subgroup (293) were sequenced from the 3' end. Cluster analysis performed at the Munich Information Center for Protein Sequences (MIPS) of the entire EST data-set produced a UniGene set of 2,458 contigs consisting of 1,917 singletons and 541 assemblies. Of the clustered ESTs, the longest contig was 1,836 bp. The entire UniGene set can be viewed on the MIPS Sputnik website [31], which features sequence annotations and peptide sequence predictions. At the MIPS Sputnik site there are links to download the complete cycad sequences as an EST fasta file, a cluster fasta file or as the derived peptide fasta file.

Classification of C. rumphii ESTs by functional categories

Each contig from the database was automatically assigned to a functional category on the basis of its top match against the complete genomic sequence of Saccharomyces cerevisiae and A. thaliana databases using BLASTP. A non-stringent expect value (E-value) of <1e-10 was chosen as the threshold. The pie chart in Figure 3 illustrates the relative fraction that each functional category comprises within the entire UniGene set. The four largest predominant categories of cycad ESTs according to this functional categorization are: 'cellular organization' (22%), 'metabolism' (10%), 'unclassified proteins' (10%), and 'cell growth, cell division/DNA synthesis' (9%).

Cycad contig matches to genes in angiosperms, gymnosperms and lower plants

Using TBLASTX, a comparison was made between the C. rumphii UniGene set versus all available ESTs from GenBank and predicted Arabidopsis genes from The Arabidopsis Information Resource (TAIR). Both EST and predicted genes were grouped into three subcategories: angiosperms, gymnosperms, and lower plants. The angiosperm database encompasses all annotated rice and Arabidopsis genes identified from their respective genomic sequences, as well as all higher plants.
plant ESTs. The gymnosperm database contains ESTs from all gymnosperms, the majority of which came from the Pinus taeda EST sequencing project [32,33]. The lower plant databases included genes from all remaining plant ESTs including ferns, fern allies, bryophytes and algae available in GenBank. The angiosperm subgroup consisted of 84.5%, the gymnosperms 6.5% and lower plants 9.0% of the total genes used in this analysis.

The Venn diagram shown in Figure 4 displays the total number of cycad contigs shared between one or more of the plant gene datasets at very low BLAST stringency values (expect < 1e-5). The majority of cycad contigs (1,764/2,458) have counterparts in other plants, leaving 694 with no match to other plant genes. As one would expect, most Cycas hits (1,718) are to angiosperms, because of the predominance of angiosperm accessions in GenBank. Many of the cycad matches to angiosperms also match gymnosperms and/or lower plants (1,416). There are 1,310 cycad contigs that match gymnosperm genes and 734 that match genes from lower plants.

Full-length sequencing of cycad clones that match only gymnosperm genes

As shown in Figure 4, 44 Cycas ESTs specifically match only genes in the gymnosperm subgroup. Two additional Cycas ESTs match genes from gymnosperms and lower plants, but not angiosperms. To further analyze these 46 contigs that match only gymnosperms and/or lower plants, we next sequenced these Cycas cDNAs in their entirety to determine whether this 'gymnosperm/lower plant' specific grouping held up when the remaining portions of the cDNA were sequenced. Because ESTs, even when clustered into contigs, usually represent only a portion of the actual gene (particularly for genes poorly represented in the library) 37 of the 46 Cycas cDNAs were sequenced in their entirety (the remaining nine clones were not successfully recovered for sequencing), and this sequence can be downloaded from the Internet [34]. Of these 37 fully sequenced cDNAs, 14 clones still showed no similarity to any known angiosperm genes, even at this low stringency cut-off. The insert size for each clone ranges from 586 bp to 1,899 bp, with predicted open reading frames (ORFs) varying from 69 to 527 residues (Table 1). None of these 14 Cycas cDNA clones is homologous to any known genes outside the plant kingdom, although Interpro analysis identified a small number of conserved motifs, which are listed in Table 1. To confirm that these genes were indeed derived from C. rumphii, gene-specific primers designed to each of the 14 genes were able to amplify a fragment from genomic DNA isolated from a different C. rumphii specimen and different tissue (sporophyll) from the source tissue of the cDNA library (data not shown). This distinct C. rumphii specimen was cultivated in a geographically separate location (Florida) from the cDNA source C. rumphii specimen used for cDNA library construction (New York).

Cycad genes similar to developmental regulators

A survey of the cycad EST dataset reveals a surprisingly large number of genes with highest similarity (BLASTP score < e-5) to genes with defined roles in growth and development in
A predicted pathway for BMAA synthesis in Cycas is supported by EST analysis

BMAA, an agonist of mammalian GLRs, is a suspect causative agent of neurological disorders [9,13]. However, nothing is known about the genes and enzymes involved in the biosynthesis of BMAA. Because the structure of BMAA is similar to other beta-substituted alamines [45,46], it is likely that BMAA biosynthesis utilizes phosphoserine, cysteine, o-acetylserine or cyanalanine as a beginning substrate. On this basis, a likely BMAA biosynthetic pathway is shown in Figure 5. This would require a two-step reaction initiated with the transfer of NH$_3$ at the beta-carbon of the substituted alanine (Figure 5a), followed by an addition of CH$_3$ (Figure 5b) to produce BMAA (Figure 5c). NH$_3$ transfer would require a nucleophilic reaction catalyzed by a cysteine synthase-like protein. A preliminary survey of genes in the cycad EST library identified candidate genes for both of these enzymatic steps (Table 2). The cycad leaf EST library contains two ESTs, which each encode a cysteine synthase. To catalyze the second step of BMAA synthesis, the EST library contains two potential methyltransferases (caffeic acid O-methyltransferase II and caffeoyl-CoA 3-O-methyltransferase). The second step would require a methyl donor, the most likely candidate being S-adenosylmethionine (SAdM). Consumption of SAdM would require the presence of enzymes to regenerate SAdM. A number of cycad ESTs can be implicated in SAdM recycling including: adenosylhomocysteinase, S-adenosylmethionine synthetase and homocysteine methyltransferase. Taken together, the cycad EST library contains candidate genes for all of the enzymes predicted to be present during the biosynthesis of BMAA.

Discussion

Cycads can be regarded as living fossils

Extant genera, such as Cycas, have changed little in morphology from their extinct relatives, such as Crossozamia, which existed during the Permian [1,2]. The study of cycads has proved to be useful in reconstructing plant evolution, in particular in understanding the rise of important plant structural innovations such as the evolution of seeds [47]. Cycads also produce a variety of neuroactive compounds, some of which are suspected to be the source of Guam’s dementia [11,48]. However, despite their scientific importance in plant biology and medicine, virtually nothing is known regarding gene expression, development and signaling in the Cycadales.

A cycad EST database: a foundation to study the evolution of early seed plants

One advantage of a genomics approach is that it provides rapid access to genes important for evolutionary studies. The more traditional homology-based gene-cloning approach is limited by tedious gene-by-gene purification. It is also limited

\[ \text{Total cycad contigs} = 2,458 \]

\[ \begin{align*}
\text{Gymnosperms} & \quad 44 \\
\text{Angiosperms} & \quad 302 \\
\text{Cycads} & \quad 684 \\
\text{Lower plants} & \quad 152 \\
\text{No match} & \quad 694
\end{align*} \]

**Figure 4**

A Venn diagram reveals shared gene sets between cycad contigs versus lower plants, gymnosperms and/or angiosperms. BLASTX (cut-off E value > 1e-5) was used to compare the cycad contigs against all angiosperm ESTs and annotated genes from the full Arabidopsis and rice genome sequence from GenBank. Genes that do not have a match to angiosperm genes were then compared to available ESTs from all gymnosperms or lower-plant ESTs available in GenBank. Genes that are common to cycads and more than one group are shown in the intersecting (shaded) regions.

angiosperms (Table 2). Some of these Cycas genes have similarity to Arabidopsis transcription factors, including CONSTANS [25,36], two distinct homeobox genes [37] and a YABBY gene [38,39]. Other cycad ESTs have similarity to other regulators of Arabidopsis development, including ARGONAUT [40] and COP9 [41,42].

**Cycas genes with similarity to Arabidopsis genes involved in signaling**

A number of genes in our cycad EST library showed similarity to components of signaling pathways found in higher plants (Table 2). These genes include a photolyase blue-light receptor, genes involved in secondary signaling (including those for calmodulin, kinases, and phosphatases), a 14-3-3 protein, and genes involved in phytosphoral responses, including auxin (IAA-9 and IAA-13) pathways as reviewed in Chory and Wu [43]. Surprisingly, a Cycas EST with high similarity to plant GLR-like genes was also found (Table 2) [15,44]. The presence of a GLR-like gene in cycads is of particular interest as it relates to BMAA, as described below.
in that it may miss related genes if the degeneracy is too great or if nonconserved regions of the protein are chosen during primer design. Finally, the targeted gene approach can never be used to discover new genes.

Sequence analysis of contigs with BLAST similarity to gymnosperms but not angiosperms
An EST project in *Pinus taeda* (loblolly pine) sampled 59,797 transcripts from wood-forming tissues [32]. In this analysis, 66 *P. taeda* contigs showed BLAST similarity at low stringency only to other gymnosperms. Similarly, in our analysis, we found 46 cycad contigs that only matched gymnosperm and lower plant genes, were fully sequenced, of which 14 clones (listed) still have no similarity to angiosperms. To confirm that these genes were of cycad origin, all 14 were successfully amplified from the DNA of a *C. rumphii* specimen other than the one used to construct the cDNA library. The transcript length, as well as the predicted translation product size, is shown. Interpro analysis identified conserved motifs in three of these cycad ESTs as shown.

| Contig   | GenBank accession number | Transcript length (bp) | Peptide (residues) | InterPro result          |
|----------|--------------------------|------------------------|--------------------|--------------------------|
| gy79c04_704 | CB090702                 | 586                    | 72                 | No matches found         |
| gy78g12_244 | CB090673                 | 627                    | 84                 | No matches found         |
| te82h02_741 | CB093328                 | 647                    | 107                | No matches found         |
| he95e08_721 | CB091708                 | 651                    | 114                | No matches found         |
| hf04g07_288 | CB092366                 | 684                    | 141                | ASP_RICH (unintegrated)  |
| hck42a07_743 | CB093061                | 790                    | 142                | No matches found         |
| gp23c01_369 | CB089407                 | 791                    | 189                | No matches found         |
| gp26f08_297 | CB089628                 | 827                    | 69                 | No matches found         |
| gy82g05_181 | CB090964                 | 840                    | 118                | No matches found         |
| he92f06_688 | CB091462                 | 935                    | 170                | No matches found         |
| gy81e11_544 | CB090877                 | 948                    | 211                | ASP_RICH (unintegrated)  |
| he97c12_740 | CB091858                 | 965                    | 140                | No matches found         |
| gp32b03_724 | CB089926                 | 1311                   | 335                | Peptidoglycan-binding LysM |
| te83a03_729 | CB093338                 | 1899                   | 527                | No matches found         |
| Average   |                         | 893                    | 173                |                          |

All available ESTs and annotated genes from GenBank were divided into three categories (angiosperms, gymnosperms and lower plants) and compared with the *C. rumphii* UniGene set. Forty-six cycad ESTs that had no similarity to angiosperm genes, but matched gymnosperm and lower plant genes, were fully sequenced, of which 14 clones (listed) still have no similarity to angiosperms. To confirm that these genes were of cycad origin, all 14 were successfully amplified from the DNA of a *C. rumphii* specimen other than the one used to construct the cDNA library. The transcript length, as well as the predicted translation product size, is shown. Interpro analysis identified conserved motifs in three of these cycad ESTs as shown.

**Genes with potential developmental roles in cycads**
As in higher plants, cycad leaves are derived from the shoot apical meristem (SAM) [30]. In *Cycas* leaflet primordia, meristematic growth ceases at the apex, while proceeding basipetally where it becomes localized to the leaflet margins [30]. The presence of these marginal meristems may explain why a surprising number of developmental genes were identified in a relatively small number of ESTs from young cycad leaves (Table 2).

A gene with identity to the *YABBY* gene family was among the cycad ESTs. *YABBY* genes encode transcription factors expressed on the abaxial side of all lateral organs that promote abaxial cell fate [38]. In *Arabidopsis*, mutations in the *YABBY* gene *INO (INNER-NO-OUTER)*, lead to the loss of the outer integument [49] reminiscent of gymnosperm (and cycad) unigamy (the presence of a single integument). Unigamy is considered to be the ancestral condition in seed plants [5,47]. An analysis of *YABBY* gene expression in cycads may help to explain the origin of the integument in gymnosperms, and/or possibly the second integument in angiosperms. One cycad EST from the library has highest similarity to *COP9*. *COP9* encodes a subunit of the COP9

**Table 1**

| Contig   | GenBank accession number | Transcript length (bp) | Peptide (residues) | InterPro result          |
|----------|--------------------------|------------------------|--------------------|--------------------------|
| gy79c04_704 | CB090702                 | 586                    | 72                 | No matches found         |
| gy78g12_244 | CB090673                 | 627                    | 84                 | No matches found         |
| te82h02_741 | CB093328                 | 647                    | 107                | No matches found         |
| he95e08_721 | CB091708                 | 651                    | 114                | No matches found         |
| hf04g07_288 | CB092366                 | 684                    | 141                | ASP_RICH (unintegrated)  |
| hck42a07_743 | CB093061                | 790                    | 142                | No matches found         |
| gp23c01_369 | CB089407                 | 791                    | 189                | No matches found         |
| gp26f08_297 | CB089628                 | 827                    | 69                 | No matches found         |
| gy82g05_181 | CB090964                 | 840                    | 118                | No matches found         |
| he92f06_688 | CB091462                 | 935                    | 170                | No matches found         |
| gy81e11_544 | CB090877                 | 948                    | 211                | ASP_RICH (unintegrated)  |
| he97c12_740 | CB091858                 | 965                    | 140                | No matches found         |
| gp32b03_724 | CB089926                 | 1311                   | 335                | Peptidoglycan-binding LysM |
| te83a03_729 | CB093338                 | 1899                   | 527                | No matches found         |
| Average   |                         | 893                    | 173                |                          |
Table 2

| Genes in Cycas rumphii with potential roles in signaling, development and biosynthesis of BMAA |
|---------------------------------------------------------------|
| **GenBank accession number** | **Subject description** | **E-value** | **% ID** | **Score** |
| **Development** |
| CB089278 | Argonaute-like protein 1 (AGO1) Arabidopsis thaliana | 8.00e-10 | 0.85 | 34 |
| CB089033 | YABBY2 A. thaliana | 2.00e-36 | 0.58 | 151 |
| CB089539 | Multisubunit regulator protein COP9 - spinach COP9 Spinacia oleracea | 2.00e-31 | 0.62 | 98 |
| CB092157 | CONSTANS B-box zinc finger family protein A. thaliana | 1.00e-47 | 0.48 | 221 |
| CB092462 | CRHB3 homeoprotein Ceratopteris richardii | 3.00e-44 | 0.70 | 131 |
| CB089344 | Homeodomain protein HB2 Picea abies | 3.00e-29 | 0.62 | 117 |
| **Signaling** |
| CB089945 | Photolyase/blue-light receptor PHR2 | 8.00e-76 | 0.69 | 197 |
| CB091652 | Putative glutamate receptor protein GLR3.4b | 2.00e-45 | 0.54 | 161 |
| CB093220 | Calmodulin-like protein; protein ids At5g4460.1 A. thaliana | 3.00e-07 | 0.58 | 45 |
| CB089469 | 14-3-3 protein Fritillaria cirrhosa | 8.00e-38 | 0.80 | 94 |
| CB091066 | Ser/Thr protein kinase isolog; protein ids, supported by cDNAs Arabidopsis | 1.00e-10 | 0.28 | 185 |
| CB090652 | Ser/Thr specific protein phosphatase 2A B regulatory subunit beta Medicago | 4.00e-86 | 0.94 | 162 |
| CB093099 | Auxin regulated protein (IAA13) A. thaliana | 1.00e-34 | 0.63 | 125 |
| CB089385 | Auxin-induced protein IAA9 A. thaliana | 8.00e-29 | 0.55 | 111 |
| **Biosynthetic enzymes of cycad specific phytochemicals (BMAA)** |
| Cysteine synthase |
| CB089577 | Cysteine synthase (O-acetylserine sulphydrylase) | 3.00e-50 | 0.75 | 128 |
| CB092214 | Plastid cysteine synthase 2 Solanum tuberosum | 5.00e-27 | 0.64 | 83 |
| Methyl transferases |
| CB091906 | Caffeic acid O-methyltransferase II Nicotiana tabacum | 3.00e-35 | 0.56 | 142 |
| CB090738 | Caffeoyl-CoA 3-O-methyltransferase Oryza sativa | 1.00e-37 | 0.47 | 188 |
| **SAdM metabolism** |
| Adenosylhomocysteinase (S-adenosyl-L-homocysteine hydrolase) |
| CB091477 | Adenosylhomocysteinase Phalaenopsis | 1.00e-87 | 0.84 | 185 |
| CB091821 | Adenosylhomocysteinase Triticum aestivum | 3.00e-78 | 0.90 | 156 |
| CB090818 | Adenosylhomocysteinase Medicago sativa | 2.00e-18 | 0.68 | 66 |
| S-adenosylmethionine synthase |
| CB091682 | S-adenosylmethionine synthetase Brassica juncea | 4.00e-90 | 0.94 | 167 |
| CB090997 | S-adenosylmethionine synthetase (methionine adenosyltransferase) Petunia | 1.00e-69 | 0.94 | 133 |
| CB090407 | S-adenosyl-L-methionine synthetase Elaeagnus umbellata | 1.00e-93 | 0.88 | 191 |
| CB092344 | Methionine synthase protein Sorghum bicolor | 4.00e-94 | 0.90 | 190 |
| CB091647 | S-methyltetrahydropteroylglutamate - homocysteine S-methyltransferase | 3.00e-79 | 0.76 | 205 |

C. rumphii ESTs were compared to GenBank with a BLASTP score < e-5. The top match produced from the BLAST search to the cycad EST is listed under subject description.

Signalosome complex, which controls multiple signaling pathways that regulate development in all eukaryotes [42,50]. In Arabidopsis, the cop9 mutant is constitutively photomorphogenic in dark-grown seedlings [51]. Some gymnosperms, (in particular the Coniferales) are constitutively photomorphogenic when grown in the dark [52,53]. As yet, the phenotype of dark-grown cycad seedlings has not been fully evaluated. The discovery of a gene encoding a putative subunit of the COP9 complex in cycads could be a first step to define the ancestral, developmental role of the signalosome in gymnosperms, particularly with regard to its role in photomorphogenesis.

Another gene potentially involved in cycad development has highest similarity to the CONSTANS gene family, which are regulators of flowering time that follow internal and external (environmental) inputs in Arabidopsis [35]. Because cycads predate the evolution of flowers, it would be of interest to...
determine if CONSTANS genes in cycads temporally regulate sporophyll and cone induction, which typically follows a yearly cycle [5,6].

A cycad GLR-like gene expressed in tissue producing the GLR agonist BMAA

An unexpected finding of the Arabidopsis EST genome project was the discovery of GLR-like genes, or 'neural' receptor genes, in plants [15]. In Arabidopsis, the GLR-like gene family comprises 20 members [54]. Pharmacological evidence has linked Arabidopsis GLRs to light and/or growth signaling pathways [15,16]. Supplying exogenous BMAA to growing Arabidopsis seedlings was shown to block light-induced hypocotyl shortening and cotyledon expansion [16]. Because BMAA has such profound effects on Arabidopsis development, we have previously proposed that BMAA, or glutamate, the natural agonist of GLRs in humans, plays a physiological role in Arabidopsis [15,16]. Continuing genetic studies in Arabidopsis aim to identify the endogenous components of the BMAA-targeted pathway in plants [16].

Cycads produce BMAA [8,9]. One EST uncovered in the C. rumphii leaf cDNA library has a high degree of similarity to plant GLR genes (Table 2). This discovery is intriguing, because it suggests that BMAA might be interacting with native GLR gene products in cycads. To further investigate the relationship between cycad GLR genes and BMAA, we sought to identify cycad genes potentially involved in BMAA synthesis.

From the structure of BMAA, we hypothesized that cycads produce BMAA in a simple two-step pathway, beginning with a β-substituted alanine. To enhance the probability of finding genes involved in BMAA synthesis, we made our cDNA library from tissues that produce relatively large quantities of BMAA (nearly 0.1 mg/g tissue) [28]. According to Ohrogge and Benning, there is a 95% chance of finding the gene for a specified enzyme when it is expressed at 0.1% mRNA/protein by sampling only 3,000 ESTs from an unnormalized library [55]. Considering the prevalence of BMAA in Cycas, it is not surprising that we discovered cognate genes for the predicted enzymes for this BMAA biosynthetic pathway in the cycad EST database (Figure 5, Table 2). Future biochemical and molecular studies will determine if these genes play a part in BMAA synthesis.

The discovery of GLR-like genes in C. rumphii raises the intriguing possibility that endogenous BMAA may interact with native cycad GLRs as a regulatory molecule. Future studies aim to understand the role of GLRs in plants, as well as the role of BMAA in herbivore defense versus endogenous signaling. The production of additional ESTs from cycads will increase the variety of genes available for study, so that a detailed expression profile can be evaluated during cycad development. Complementation studies of these genes in orthologous Arabidopsis mutations will help define their roles in cycads. This combined approach to studying cycad gene structure and function will help reveal molecular changes in genes involved in signaling, metabolic and developmental pathways that led to the rise of the seed plants.

Materials and methods

Tissue collection and library construction and DNA purification

Newly emerged immature leaves from the crown of a C. rumphii tree, accession 808/59 A, were collected from the New York Botanical Garden Conservatory. Leaves collected ranged from 5 to 30 cm in length. Tissue was frozen in liquid nitrogen. RNA was extracted from pulverized, frozen tissue in a mortar and pestle with the RNaseasy maxi kit (Qiagen, Valencia, CA) according to the manufacturer’s protocol. Purified
Cycas RNA was precipitated in 2 M LiCl, washed twice with 70% ethanol, and resuspended in 50 µl water. Poly(A) RNA was subsequently purified from total RNA with the Oligotex Maxi kit (Qiagen). A CDNA library was constructed using the Lambda ZAP-CMV cDNA synthesis kit (Stratagene, La Jolla, CA) using 10 µg poly(A) RNA. Before cloning, cDNA was size fractionated over a Sepharose CL-6b column. The first five fractions containing a total of around 100 ng cDNA were collected, pooled and precipitated in 70% ethanol/0.3 M sodium acetate and resuspended in 3.5 µl water. cDNA (0.5 µl) was then directionally subcloned into the vector at the EcoRI and XhoI sites.

DNA was collected from unemerged C. rumphii sporophylls using the DNeasy purification kit (Qiagen).

EST sequencing
Plasmid DNA was collected as described in the manual (Stratagene) catalog number 200450 in the in vivo mass excision section. Sequence analysis was performed at Cold Spring Harbor Laboratory using an ABI 3700 capillary sequencer (Applied Biosystems, Foster City, CA) for separation and nucleotide detection. Reactions were performed using a 1/16 Big Dye Terminator. Sequencing was performed with either the -21 M13 forward and/or reverse primer. Sequencing was performed with either the -21 M13 forward and/or reverse primer.

EST clustering and assignment into functional categories
The EST sequences were clustered and assembled using the HarvESTer application (Biomax informatics, Martinsried, Germany). The default HarvESTer settings were optimized to screen for vector against the UniVec nonredundant database of vector and polyn linker sequences [56]. The Hashed Position Tree (HPT) clustering used a similarity link threshold of 0.7 and a maximum distance of six steps was required to define a cluster from the similarity network, thus encouraging the separation of likely paralogs. Cluster consensus sequences and concomitant alignments were derived from the HPT clusters using the CAP3 application with default settings. The HarvESTer assemblies and coordinate alignments were imported into the Sputnik EST and cluster analysis application [57].

Peptide extraction
BLASTX [58] was performed against a nonredundant protein database for each of the cluster consensus sequences. Likely coding sequences were derived for each cluster consensus sequence by parsing the best BLASTX match and filtering the results using the arbitrary expect value <1e-10. Dicodon usage frequencies and probabilities were extracted using tools from the ESTate package [56]. A peptide sequence was predicted for each of the cluster consensus sequences using the Framefinder application from the ESTate package with the cycad-specific codon usage statistics. Framefinder was run using the default parameters. The derived peptide sequences were used as the basic scaffold for peptide-based annotation in Sputnik.

Sequence annotation
Sequence annotation on each of the cycad cluster consensus sequences and derived peptides were performed within the Sputnik application. Results were assessed for possible contamination by searching for homology to the Escherichia coli and human genomes and were scored for homology to a wide range of noncoding RNAs and plant chloroplast and mitochondrial genomes. Similarity searches were performed using the BLAST application [58] and results were filtered using the expectation value < 1e-10. Functional assignment was performed on both cluster consensus sequence and the peptide sequence. Assignments were made using BLASTX and BLASTP respectively against the MIPS catalog of functionally assigned proteins (funcat) [60,61]: tentative functional assignments were filtered using the expectation value < 1e-10.

Categorization of cycad contig
All cycad contigs sequences were aligned against the PlantEST database using TblastX [58] and BlastX against the NR(aa) database. The PlantEST database was created by downloading all plant ESTs in GenBank and assembling them using Phrap [60,61]. Todd Wood from Clemson University provided the PERL script that creates the PlantEST databases as described above. The NR(aa) database is a nonredundant database of protein sequences from GenBank.

Determination of gymnosperm-specific genes
All available plant ESTs were downloaded from GenBank and separated into three datasets consisting of angiosperms (monocots and dicots), gymnosperms, or lower plants (ferns, mosses and algae). Downloaded ESTs were assembled using Phrap [60,61]. All matches with an expect value < 1e-5 were considered significant.

Acknowledgements
We thank Francesco Coelho, Javier Francisco Ortega and the Montgomery Botanical Center, Florida for providing plant tissue; Dan Chamovitz and Trevor Stokes for reviewing the manuscript; Vivekanand Balija and Nelly Dethia for sequence generation and curation; Eduardo de la Torre and Eugene Mueller for helpful discussions; and Alex Clark and Ayelet Levy for technical help. Funding for this work comes from the Plant Genomics Consortium. The Plant Genomics Consortium is made possible by the generosity of the Altira Group, The Mary Flager Cary Charitable Trust, The Eppley Foundation for Research, The Leon Lowenstein Foundation, The Ambrose Monell Foundation, The Wallace Genetic Foundation and the National Institutes of Health, grant number GM-32877 to G.C. and an NIH postdoctoral fellowship to E.B.

References
1. May M: Cycads: fossil evidence of late paleozoic origin. Science 1969, 164:295-296.
2. Gao Z and Thomas BA: A review of fossil cycad megasporophylls, with new evidence of Crossozamia pomel and its associated leaves from the lower Permian of Taiyuan, China. Rev Paleobot Paleogeogr 1989, 60:205-223.
3. Nixon K, Crepet W, Stevenson DW and Friis E: A reevaluation of seed plant phylogeny. Ann Missouri Bot Garden 1994, 81:484-583.
4. Solits DE, Solits PS and Zannis MJ: Phylogeny of seed plants based on evidence from eight genes. Am J Bot 2002, 89:1670-1681.
5. Norstog KJ and Nicholls TJ: The Biology of the Cycads Ithaca, NY: Cornell University Press; 1997.

6. Chamberlin C: The Living Cycads Chicago: University of Chicago Press; 1919.

7. Loconte H and Stevenson DW: Cladistics of the Spermaphyta. Brittonia 1990, 42:197-211.

8. Vega A and Bell EA: Alpha-amino-beta-methylaminopropionic acid, a new amino acid from seeds of Cycas circinalis. Phytochemistry 1967, 6:759-762.

9. Spencer PS, Huns PB, Ross SM, Hunn PB, Nugon I, Ludolph AC, Liang QS, Nestor RI, and Stevenson DW: Glutamate receptor agonist. J Ethnopharmacol 2002, 82:159-167.

10. Whiting MG: Toxicity of cycads. Econ Bot 1963, 17:271-302.

11. Kurland LT: An appraisal of the neurotoxicity of cycad and the etiology of amotrophic lateral sclerosis on Guam. Fed Proc 1972, 31:1540-1543.

12. Brenner ED, Martinez-Barboza N, Clark AP, Liang QS, Stevenson DW and Coruzzi GM: A draft sequence of the flowering plant Arabidopsis thaliana. Nature 2000, 408:796-815.

13. Mayer K, Mewes HW, Robertson RS, Robertson RC, Aronsson G, Gasteiger E, Bairoch A, Herneberger B, and Gough J: Plant development and cell wall biosynthesis in loblolly pine. Plant Mol Biol 2001, 47:275-291.

14. Chory J and Wu D: The COP9 signalosome. Curr Opin Plant Biol 2001, 4:42-52.

15. Villanueva JM, Broadhurst J, Hausser BA, Meister RJ, Schneitz K, Pasqualotto AC, and Deng XW: COPS signalosome revisited: a novel mediator of protein degradation. Trends Cell Biol 2001, 11:420-426.

16. McLaren A, Gagnon JM, Durbin R, and Flicek P: The UCSC Genome Browser database. Nucleic Acids Res 2010, 38:D749-754.

17. Chory J and Wu D: Weaving the complex web of signal transduction. Plant Physiol 2002, 129:77-80.

18. Chory J, Brenner ED, DeSalle R, Nitabach MN, Holmes TC and Coruzzi GM: Phylogenetic and expression analysis of the glutamate-receptor-like gene family in Arabidopsis thaliana. Mol Biol Evol 2002, 19:1066-1082.

19. Warrilow AG and Hawkesford MJ: Cysteine synthase (O-acetyserylserine (thiol) lyase) substrate specificities classify the mitochondrial isomerase as a cycanoalanine synthase. J Exp Bot 2000, 51:985-993.

20. Chory J and Wu D: Weaving the complex web of signal transduction. Plant Physiol 2002, 129:77-80.

21. McLaren A, Gagnon JM, Durbin R, and Flicek P: The UCSC Genome Browser database. Nucleic Acids Res 2010, 38:D749-754.

22. Chory J and Wu D: Weaving the complex web of signal transduction. Plant Physiol 2002, 129:77-80.

23. Warrilow AG and Hawkesford MJ: Modulation of cyanoalanine synthase and O-acetyserylserine (thiol) lyases A and B activity by beta-substituted alanyl and anion inhibitors. J Exp Bot 2002, 53:439-445.

24. Foster AS and Gifford EM: Comparative Morphology of Vascular Plants 2nd edition. San Francisco: WH Freeman; 1974.

25. Khazabin I, Bains JS, Pasqualotto AC, Warrilow AG, and Deng XW: COPS signalosome revisited: a novel mediator of protein degradation. Trends Cell Biol 2001, 11:420-426.

26. Villanueva JM, Broadhurst J, Hausser BA, Meister RJ, Schneitz K and Gasser CS: INNER NO OUTER regulates abaxial-adaxial patterning in Arabidopsis. Dev Growth Differ 2002, 19:245-252.

27. GenBank. A database platform for comparative plant genomics. Nucleic Acids Res 2003, 31:128-131.

28. Sputnik: A database platform for comparative plant genomics. Nucleic Acids Res 2003, 31:128-131.

29. Sputnik: A database platform for comparative plant genomics. Nucleic Acids Res 2003, 31:128-131.
59. Slater GSC: *Algorithms for the Analysis of ESTs*. PhD thesis University of Cambridge; 2000.

60. Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998, 8:186-194.

61. Ewing B, Hillier L, Wendl MC and Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998, 8:175-185.