Mol Genet Genomics (2007) 278:393–402
DOI 10.1007/s00438-007-0257-6

ORIGINAL PAPER

Dating the early evolution of plants: detection and molecular clock analyses of orthologs

Andreas Zimmer · Daniel Lang · Sandra Richardt · Wolfgang Frank · Ralf Reski · Stefan A. Rensing

Received: 10 April 2007 / Accepted: 24 May 2007 / Published online: 26 June 2007

Abstract Orthologs generally are under selective pressure against loss of function, while paralogs usually accumulate mutations and finally die or deviate in terms of function or regulation. Most ortholog detection methods contaminate the resulting datasets with a substantial amount of paralogs. Therefore we aimed to implement a straightforward method that allows the detection of ortholog clusters with a reduced amount of paralogs from completely sequenced genomes. The described cross-species expansion of the reciprocal best BLAST hit method is a time-effective method for ortholog detection, which results in 68% truly orthologous clusters and the procedure specifically enriches single-copy orthologs. The detection of true orthologs can provide a phylogenetic toolkit to better understand evolutionary processes. In a study across six photosynthetic eukaryotes, nuclear genes of putative mitochondrial origin were shown to be over-represented among single copy orthologs. These orthologs are involved in fundamental biological processes like amino acid metabolism or translation. Molecular clock analyses based on this dataset yielded divergence time estimates for the red/green algae (1,142 MYA), green algae/land plant (725 MYA), mosses/seed plant (496 MYA), gymno-/angiosperm (385 MYA) and monocotyledons/core eudicotyledons (301 MYA) divergence times.

Keywords Ortholog detection · Molecular clock · Phylogeny · Plant evolution

Introduction

The availability of several complete or nearly complete genome sequences of plant and algal model species provides the opportunity to investigate their evolution on a genomic scale. However, the extension of conclusions drawn from such analyses to, e.g., plants, as a whole, is difficult due to possible sampling bias. The reduction of the analysis to orthologous "benchmark" genes or marker regions, which can be feasibly compared between multiple species in parallel is the means to overcome taxon-sampling issues. Yet, if the selection of these benchmark genes is arbitrary or based on the criterion of availability, unequal evolutionary rates might falsify the analysis. Paralogous "contaminations" in a benchmark set can diffuse the signal, because paralogs often evolve at a different rate than their primordial ortholog (Blanc and Wolfe 2004; Fares et al. 2006). Hence, a generally applicable, fast and feasible method to detect orthologous relationships, even among species where only expressed sequence tag (EST) data are available, is needed. A classification based on strict orthologous relationships between gene products constitutes a benchmark for phylogenetic studies, enables calibration of the molecular clock and should eventually result in a deeper understanding of plants from genomic and evolutionary perspectives.

This study uses the predicted proteins of the core eudicotyledon *Arabidopsis thaliana*, the monocotyledon *Oryza*

Communicated by Y. Van de Peer.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-007-0257-6) contains supplementary material, which is available to authorized users.

A. Zimmer · D. Lang · S. Richardt · W. Frank ·
R. Reski · S. A. Rensing (✉)
Plant Biotechnology, Faculty of Biology,
University of Freiburg, Schaenzlestr. 1,
79104 Freiburg, Germany
e-mail: stefan.rensing@biologie.uni-freiburg.de
URL: www.plant-biotech.net; www.cosmoss.org

sativa, the gymnosperm *Pinus taeda*, the moss *Physcomitrella patens*, the green and red alga *Chlamydomonas reinhardtii* and *Cyanidioschyzon merolae*, thus covering major phyla of photosynthetic eukaryotes, i.e., plants *sensu lato*. Based on molecular clock analyses, approximately 1.6 BYA (Hedges et al. 2004; Yoon et al. 2004) plants split off from the animal and fungal lineage by a primary endosymbiotic event, the engulfment of a cyanobacteria-like prokaryotic cell, eventually leading to the formation of plastids. About 200 MY later (Hedges et al. 2004; Yoon et al. 2004) the ancestors of the red and green lineage (which would later give rise to the *Chlorophyta* and *Embryophyta*, i.e., green algae, water and land plants) diverged. From the ancestor of the red lineage the *Rhodophyta* (red algae) emerged as well as several algal divisions which acquired their plastids by secondary endosymbiosis. The land plants diverged from the green algae at least 1 BYA, while mosses and seed plants shared their last common ancestor at least 450 MYA, as deduced from molecular data (Theissen et al. 2001; Hedges et al. 2004). We did not include data from diatoms (Armbrust et al. 2004) or other secondary endosymbionts because their plastids were established several hundred MY later.

Orthologs, as defined by Fitch (Fitch 1970) often perform the same function in different organisms, while paralogs usually deviate in function or regulation from their archetypical gene (Li et al. 2005; Dutilh et al. 2006). Recently, the terminology of gene relationships has been expanded (Fitch 2000; Sonnhammer and Koonin 2002) in order to clarify the definition of paralogs, which is dependent on the frame of reference. While determination of true orthologs (i.e., a single gene in each organism that can be traced back to the common ancestor) is most confidently achieved by phylogenetic approaches, reciprocal homology searches have been established as a straightforward method to determine probable orthologs (Mushegian et al. 1998). While common orthologs of animals and yeast have already been analyzed (Mushegian et al. 1998) and orthologs of species-pairs have been made available (O'Brien et al. 2005), plants have not yet been the focus of such research, except for phylogenetic profiling approaches of *Arabidopsis* in comparison with fungi, animals, eubacteria and archaea (Gutierrez et al. 2004) as well as with a large group of clustered EST data from plants (Vandepoele and Van de Peer 2005). Although clusters of orthologous groups (COGs) have been established for eukaryotes (KOGs) (Tatusov et al. 2003), these homology-based clusters also contain paralogs. The same is true for so-called in-paralog approaches (Remm et al. 2001; O'Brien et al. 2005; Alexeyenko et al. 2006) as well as approaches yielding clusters of closely related sequences (Lee et al. 2002; Li et al. 2003). In order to reduce the amount of potentially disturbing paralogs we used reciprocal BLAST searches and analyzed all six organisms in parallel, creating clusters of orthologous genes, which contain a single gene from each studied organism. Hence, the aim of our study was the detection of orthologs *sensu strictu*, excluding paralogs, from six photosynthetic eukaryotes. We evaluated our method by annotation and phylogenetic analysis of the ortholog clusters. Subsequently, molecular clock analyses of the orthologs were used to estimate the divergence times of red and green algae, mosses, gymnosperms as well as monocotyledons and core eudicotyledons.

Results

Detection of cross-species ortholog clusters among photosynthetic eukaryotes

In order to avoid false positives, only clearly defined homologs were selected, using filtering criteria designed to avoid the twilight zone of protein sequence alignments in which homology cannot be unambiguously determined (Rost 1999). To enable the detection of cross-species orthologs, reciprocal BLAST searches of all against all species were performed. Subsequently, we used non-ambiguous keys to select from the resulting database in order to determine putatively orthologous genes that were shared by all, or any chosen subset of species (see Supplementary data for a graphical representation of the method). Using these restrictive criteria, we identified 9,497 such shared genes between the angiosperms rice (*Liliopsida*) and *Arabidopsis* (core eudicotyledons). These plants have 2,105 genes in common with the gymnosperm *Pinus*, while all land plants (*Embryophyta*) share a group of 540 genes. Land plants and green algae (= green lineage) share 203 genes, whereas the number drops to 93 clusters when the red alga is included into this analysis. We determined the global conservation of the gene clusters using multiple sequence alignments. Based on global alignments, the minimum average identity within an ortholog cluster (6-tuple) was around 38%, the maximum around 85% (see Supplementary data). On average, the genes are ~60% identical over the whole sequence length.

Phylogenetic analyses of the ortholog clusters

Phylogenetic analyses of orthologs have been carried out as a test for consistency (Mushegian et al. 1998; Raymond et al. 2002). For example, in an analysis of human, worm, fly and yeast orthologs, the majority of ortholog clusters supported the expected phylogeny. However, unequal evolutionary rates may disturb the phylogenetic reconstruction in some cases (Felsenstein 1978; Mushegian et al. 1998). In a similar approach the individual maximum likelihood

(ML) topologies for the 93 ortholog clusters of the photosynthetic eukaryotes were determined and a consensus tree was calculated subsequently. In this tree, the generally accepted species topology is recovered, yet with better support for the lower branches than for the seed plants (data not shown). In order to determine the reasons for this observation, the phylogenetic trees were analyzed in more in detail. Scatter plots of quartet puzzling support values versus effective alignment length and total branch length of the expected trees, respectively, did not reveal a clear correlation (data not shown). Therefore, the weak support for the upper branches does not seem to be due to either short sequence length or varying gene-specific rates. The ML differences between best and expected trees were found to be in a narrow range. For only four out of 93 trees, the likelihood differences were found to be significant based on the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999).

Evaluation of the ortholog detection method

The aim of this study was the identification and analysis of true orthologs common to photosynthetic eukaryotes. Such genes, common to six different species of major phyla of algae and plants, would be expected to represent orthologs *sensu strictu*, i.e., not containing in- or out-paralogs (with respect to species A, an in-paralog has arisen within species B after separation of A and B from the last common ancestor, an out-paralog before that event). By inferring phylogenies of the homologs surrounding and including the detected 6-tuples, we assessed the quality of our detection method. All gene clusters were associated with the Boolean qualifiers *true ortholog*, *species tree ok* and *single copy* based on manual inspection of the phylogenetic trees. The qualifier *true ortholog* was set if all six selected genes were correctly selected, i.e., there were no in-paralogs present or the in-paralog had a larger evolutionary distance to the other sequences than the selected ortholog. In total, 63 out of 93 clusters contained exclusively true orthologs (Fig. 1), i.e., 68% of the 6-tuples were accurately selected by the cross-species reciprocal BLAST search method. The expected species phylogeny was considered resolved if the cluster was not invaded by proteins from wrong taxonomic groups and the topology of the tree did not contradict the expected taxonomy, which was true for a total of 55 phylogenies (59%). We assigned the *single copy* flag if neither in- nor out-paralogs were present in the tree. However, we allowed for the exception that a single species might contain independently acquired paralogs. The single copy status was assigned to 47 (51%) of the clusters (Fig. 1), of those, 42 were true orthologs as well. Thus, single copy genes are significantly enriched ($P = 0.033$, Fisher's exact test) among the true orthologs. While the BLAST-based method selects ~32% of false-positive (paralog) containing clusters, this error rate drops to 11% for single copy genes. In the strict sense (not allowing for an aberrant species) 26 ortholog clusters contained single copy genes, i.e., no detectable paralog was present in any of the six species. All of these 26 clusters contained exclusively true orthologs.

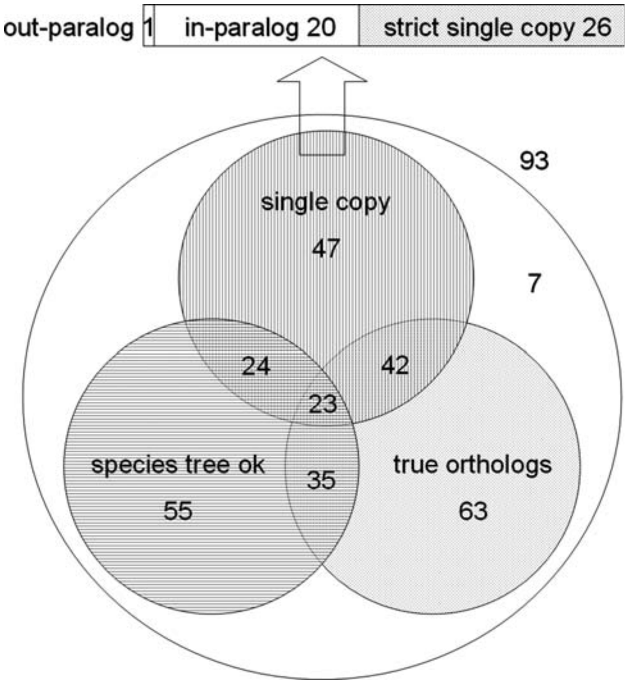

Fig. 1 Venn diagram of the ortholog clusters. Properties of the 93 ortholog clusters of photosynthetic eukaryotes. The intersections are not additive, i.e., they are already contained in the main sections. All ortholog clusters were associated with the boolean qualifiers "true ortholog", "single copy" and "species tree ok" based on manual inspection of phylogenetic trees. The qualifier true orthology was set if all six selected genes were correctly selected by the detection method. The expected species phylogeny was considered resolved if the ortholog cluster was not invaded by sequences from wrong taxonomic groups and the topology of the tree did not contradict the expected taxonomy. We assigned the single copy flag if neither in- nor out-paralogs were present in the tree. However, we allowed for the exception that a single species might contain paralogs. In the strict sense (not allowing for an aberrant species) 26 orthologs were single copy genes, i.e., no detectable paralog was present in any of the six species

In order to evaluate the selectivity and sensitivity of our procedure, it was compared to the results of InParanoid (Remm et al. 2001; O'Brien et al. 2005). InParanoid detects 1,878 clusters of orthologs among the green alga *Chlamydomonas* and the red alga *Cyanidioschyzon*, whereas our method identifies 1,439 orthologs; 1,208 orthologs are identical. Five hundred and seventy-nine orthologs are unique to InParanoid because these genes were dropped in our approach due to the 30% identity/alignment length cutoffs. One hundred and forty orthologs are unique to our result set because they miss InParanoid's criterion of 50% coverage but exceed the alignment length cutoff of 100.

A small set of 91 genes have different partners, either because we consider the best hit only or they are members of an InParanoid ortholog cluster ($n > 2$). Additionally, the results for all six species were compared to the MultiParanoid (Alexeyenko et al. 2006) method based on the pairwise results of InParanoid. Whereas our approach is equivalent to a complete linkage clustering, MultiParanoid generates single linkage clusters. Because the MultiParanoid method is based on pairwise InParanoid results the clusters can contain more than one protein per organism (orthologs, in-paralogs). Hence, the results of our method should be a subset of the MultiParanoid results. Indeed, with the exception of one cluster, the clusters derived from our method are a subset of the MultiParanoid clusters. This aberrant cluster (CO #22) differs only in one member, the *P. patens* protein PPP_312_C1, which does not appear in any MultiParanoid cluster, although it generates best BLAST hits during the pairwise searches. All other common proteins are seed orthologs (main orthologs) in InParanoid, respectively, MultiParanoid clusters. In addition the MultiParanoid result contains 1,194 clusters. Moreover, our method outperforms the MultiParanoid approach in terms of runtime and disk space usage. For the given set of organisms our method, including the BLAST searches and their parsing, it is about an order of magnitude faster than generating the corresponding MultiParanoid results while it requires only about half the amount of disk space.

Ancient single copy genes and their genetic parentage

We compared the strict single copy genes with the remainder of the 6-tuples in terms of alignment length, average sequence identity, average quartet support and the fraction of resolved quartets. Out of these parameters, only the average quartet support values were significantly increased ($P = 0.048$) among the strict single copy clusters. Hence, the proteins do not differ in length or conservation grade from those gene families that contain paralogs, but reconstruction of phylogenies from gene families that do not contain paralogs seems to yield better-supported topologies. In terms of pathway annotation, among the strict single copy clusters those genes that encode amino acid metabolism are significantly ($P < 0.05$) enriched (KEGG KO categories: 00300 Lysine biosynthesis; 00330 Arginine and proline metabolism; 00220 Urea cycle and metabolism of amino groups). Proteins involved in translation, energy metabolism and other basic processes are present as well (see Supplementary data).

Plastids have arisen by engulfment of a free-living cyanobacterial-like prokaryote and subsequent establishment of endosymbiosis. During evolution, the majority of prokaryotic genes have been transferred to other compartments, mainly to the nucleus, while the gene products are targeted to the plastid as well as to other destinations (Martin et al. 2002). The same holds true for mitochondria, which share their last common ancestor with extant alpha-proteobacteria and were established prior to plastids (Gray et al. 2001). Assignment of putative origin of the orthologs using BLAST searches revealed that 26 (28%) out of 93 genes were inherited from the mitochondrion, 19 (20%) from the plastid and 48 (52%) from the ancestral eukaryotic nucleus, i.e., genes of eukaryotic origin are enriched ($P < 0.002$). Among the 26 strict single copy genes, those of mitochondrial origin (46%) are over-represented ($P = 0.016$) in comparison with the other two fractions (nt: 23%; pt: 16%). Also, gene families of mitochondrial origin contain significantly less in-paralogs ($P < 0.027$) than those derived from the ancestral eukaryotic nucleus or the plastid.

Taxonomic profile

The taxonomic profile (i.e., the contribution of distinct taxonomic groups to the distribution of homologs) of the true ortholog clusters correlates well with their genetic origin (Fig. 2). Most genes of eukaryotic origin are clustered in the upper part of the heat map. Those clusters contain homologs from protists, plants, fungi and animals, whereas eubacterial genes are clearly under-represented. In the lower part, most of the prokaryotic genes are clustered. These genes generally prevail among several eubacterial groups while they are under-represented among most eukaryotic groups. Yet, there are some genes for which taxonomic distribution and heritage deviate. There are eukaryotic genes that cluster with those of prokaryotic origin because homologs are present to a large extent among eubacteria, like cluster 50, 65 and 45. Another trend is the prevalence of eukaryotic genes among the non-photosynthetic (other) protists, while this taxonomic group is clearly under-represented among the prokaryotic genes.

Molecular clock analyses date the early evolution of plants

We used the orthologs described here in order to determine the basic splits of early plant evolution. The 6-tuples were subjected to likelihood ratio tests in which 14 of the clusters supported the molecular clock hypothesis. Based on those, divergence times were calculated by applying the LF method (Langley-Fitch, assumes a molecular clock) and the PL method (penalized likelihood, a semiparametric approach) as implemented in r8s (Sanderson 2003), using constraints collected from the literature (Table 1). For each of the clusters the calculations were iterated in order to constrain four of the nodes and calculate a divergence time for the remaining node. The average calculated divergence times for the LF and PL method were not significantly different ($P = 0.97$), which confirms that the sequences

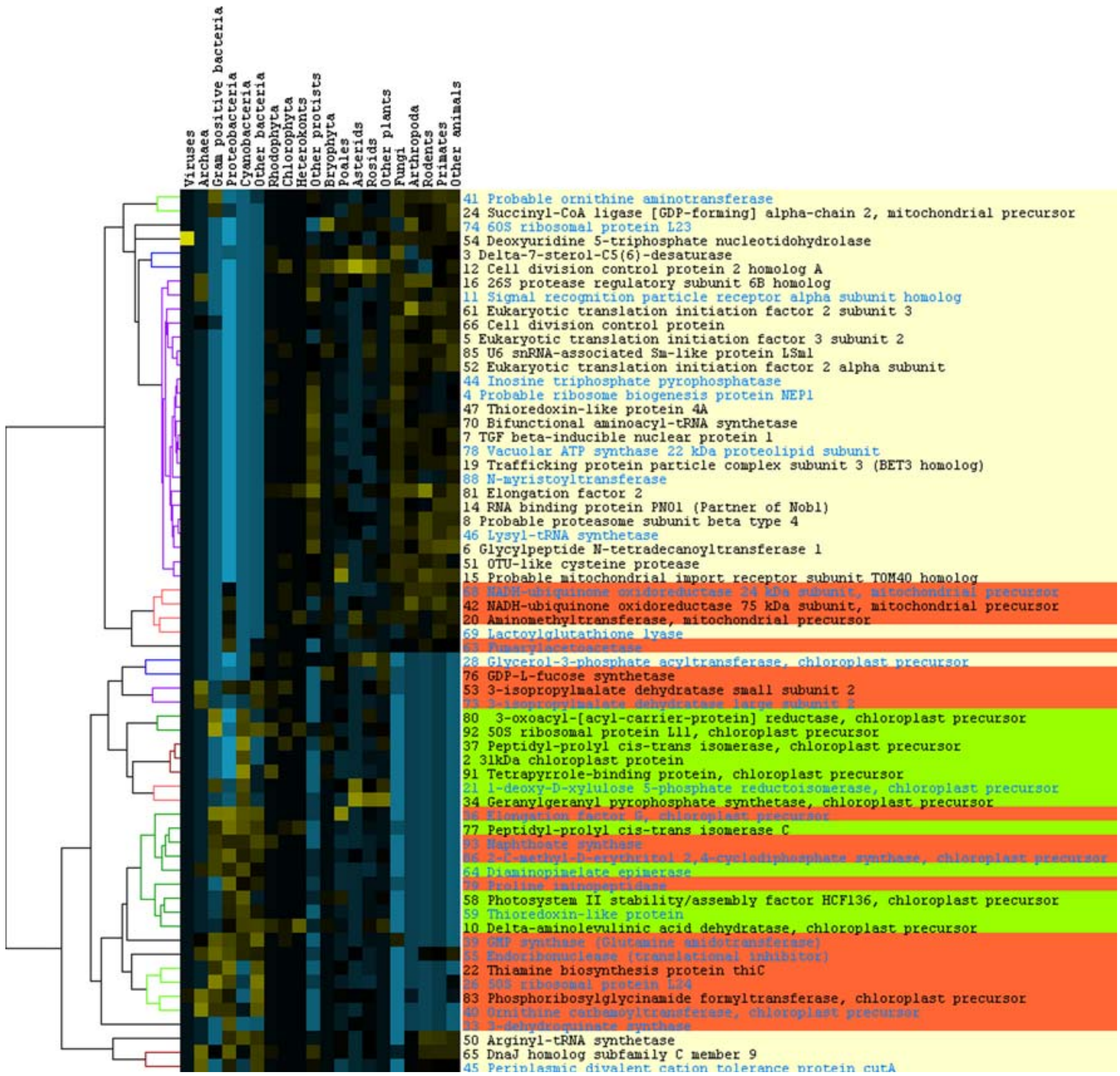

Fig. 2 Taxonomic profile of the ortholog clusters. The NCBI taxonomy information for all true orthologs was parsed per cluster. After normalization of the columns, the rows were clustered and visualized as a heat map (*yellow* indicates over-represented, *blue* indicates under-represented). The *columns* represent those taxonomic groups, which contributed significantly to the distribution; the remainder of the eubacteria, protists, plants and animals was pooled as "other", respectively. All clusters with a Pearson correlation coefficient $R \geq 0.7$ are displayed in *color to the left of the heat map*. The ortholog id and annotation are shown to the *right*. Color code: *yellow*—gene of ancestral eukaryotic origin; *green*—gene of cyanobacterial origin; *red*—gene of alpha-proteobacterial origin; *blue font*—strict single copy genes

show approximate clocklike behavior and that the estimates are not dependent on the method. Because only ten out of the 14 clusters were assigned true orthology, we also tested whether the estimates were different between these two populations, which was not the case ($P = 0.97$). In addition, the divergence times using a concatenated alignment instead of the individual alignments were calculated, which also led to insignificant differences ($P > 0.94$). The estimates generated using the LF method exhibit the smallest average standard deviation. For the full set of individual alignments LF dates the split between red and green algae to 1,142 ± 167 MY, the green algae/land plant divergence to 725 ± 138, mosses/seed plants 496 ± 84, gymno-/angiosperms 385 ± 72 and monocotyledons/core eudicotyledons to 301 ± 93 (Fig. 3, Table 1).

Discussion

Not surprisingly, the set of orthologs detected among the broad phylogenetic range of phototrophic eukaryotic organisms studied here exhibit a high degree of sequence conservation, being the result of strong purifying selection against functional divergence or loss of function of these genes. The genes studied here were already established before the split of the red and the green lineage at least 1 BYA

Table 1 Constraints used and results of the molecular clock analyses

	Monocot/core eudicot	STDEV/ref.	Angio-/gymnosperm	STDEV/ref.	Moss/seed plant	STDEV/ref.	Green algae/land plants	STDEV/ref.	Red/green algae	STDEV/ref.
Minimum constraint	90.00	Crane et al. 1995	290.00	Yoon et al. 2004	396.00	Taylor et al. 2005	786.00	Hedges et al. 2004	1174.00	Butterfield 2001
Maximum constraint	240.00	Wolfe et al. 1989	360.00	Troitsky et al. 1991	899.00	Hedges et al. 2004	1150.00	Hedges et al. 2004	1579.00	Hedges et al. 2004
PL average	304.72	111.25	372.33	76.56	476.65	93.76	725.63	134.97	1212.89	215.18
PL concatenated	256.46		365.46		454.90		756.32		1167.95	
LF average	300.99	93.37	385.36	71.51	495.82	83.71	725.43	138.03	1141.63	166.55
LF concatenated	264.60		393.77		472.33		778.28		1111.16	

The upper two rows list the minimum and maximum constraints that have been used for the molecular clock analyses together with the associated literature. The corresponding nodes are arranged as columns and describe the divergence of the respective last common ancestor. The lower four rows contain the results of the molecular clock dating using the PL and LF method, respectively, both for individual (averaged) and concatenated alignments together with the standard deviation of the average values. All numbers are in million years

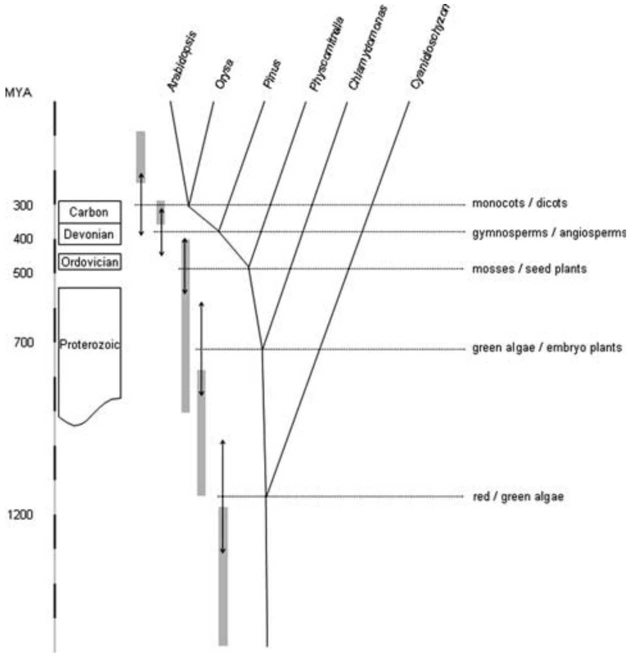

Fig. 3 Early evolution of plants. Schematic cladrogram of plant evolution summarizing the data of the molecular clock analyses. The *gray boxes* depict the literature-derived range used as constraints (see Table 1) while the *dotted lines* represent the estimated divergence times (LF average, Table 1) for the respective nodes (*arrows* indicate standard deviation)

(Hedges et al. 2004; Yoon et al. 2004) and subject to strong selective pressure, which, as expected, is reflected in the high conservation grade. It is also to be expected that there is some variation in gene-specific substitution rates, as has been demonstrated for orthologs of fungi and mammals (Mushegian et al. 1998). Indeed, a plot of ML gene distances of the clusters revealed that they follow a normal distribution (data not shown). The fact that the best tree cannot be distinguished from the expected tree in most of the cases, together with the low support of the higher branches, indicates that the seed plant sequences are too well conserved to unambiguously resolve their phylogeny.

By phylogenetic profiling of 32 plant transcriptomes, a total of 397 gene families have been detected previously that are shared among the green lineage (Vandepoele and Van de Peer 2005), while we detected 203 gene clusters among these phyla. The finding that more gene families are detected using a clustering approach than by cross-species reciprocal BLAST searches indicates that our method selects a lower amount of clusters because it does not allow paralogs. Apparently, our method preferentially selects single copy genes, most probably because paralogs will not always yield unambiguous best BLAST hits. Not surprisingly, all of these 26 single copy clusters contained exclusively true orthologs. In general, paralog retention of genes that encode essential biological functions can be under negative selection because of the potentially harmful dosage imbalance that might be associated with such a change (Birchler et al. 2005).

We compared our method to the pairwise InParanoid (O'Brien et al. 2005) results of the two algae. While the results for these two organisms differ in some aspects, we do not seem to exclude a large part of orthologs using our strict criteria as deduced from the fact that the majority of detected orthologs are identical. Also the comparison to the multi-species ortholog detection approach MultiParanoid (Alexeyenko et al. 2006) reveals a high amount of common ortholog clusters. Yet, our approach detects fewer orthologous clusters between the six photosynthetic organisms. This leads to the conclusion that our method is more restrictive through exclusion of potential false positives, while probably losing true positive hits. As our intent was not to select all orthologous clusters, but to detect and analyze orthologs; for, e.g., dating purposes, our method is adequately complete. Moreover, our method outperforms the MultiParanoid approach in terms of runtime and disk space usage.

Although small and large scale duplication events constantly occur among seed plants (De Bodt et al. 2005) we detected 26 genes among photosynthetic eukaryotes which did not retain a paralog over ~1.4 BY of evolution. While genes that were inherited from the ancestral eukaryotic nucleus dominate the total set of ortholog clusters, genes of mitochondrial origin are over-represented among the single copy gene families. Apparently, genes that were transferred to the nucleus between the primary endosymbiotic event giving rise to mitochondria and the division of the red and green lineage are less likely to be retained after duplication. Those genes encode proteins that are involved in amino acid metabolism and other primary metabolic processes. Apparently, ancient genes, which encode essential biological functions are under negative selection against paralog retention, particularly if they are of mitochondrial origin. These dominance of eukaryotic parentage among the whole set has also been found in a study of KOG's among seven eukaryotic genomes (Koonin et al. 2004), where 56% of the genes did not exhibit prokaryotic ancestry.

In terms of the taxonomic profile, there are some genes for which taxonomic distribution and heritage deviate (Fig. 2). E.g., the three mitochondrial ortholog clusters 68, 42 and 20 exhibit a typical eukaryotic profile except for the alpha-proteobacterial group from which they are derived. Apparently, other eubacteria use different genes to fulfill the functions associated with these orthologs. Among plants, there are several orthologous gene families with deviating distribution, e.g., between monocotyledons and core eudicotyledons, such as cluster 15, 34 and 36.

The detection of recently described panorthologs (Blair et al. 2005) relies on a pairwise clustering approach and subsequent deletion of paralogs. In our ortholog detection approach we only permit a single sequence to enter the comparison in the first hand. As both methods detect orthologs from more than two organisms, the outcome is expected to be approximately the same, i.e., detection of true orthologs among a group of more than two organisms. Panorthologs have been successfully applied to calculate divergence times based on a molecular clock (Blair et al. 2005): however, the analysis focused on animals. We used the orthologs described here to determine the basic splits of early plant evolution, namely red/green algae, green algae/land plants, mosses/vascular plants, gymnosperms/angiosperms and monocotyledons/core eudicotyledons. It has been shown that divergence times estimates might be influenced by, e.g., chosen method or selected dataset (Bell et al. 2005). Based on the data described above, however, our dataset yields robust estimates that are not significantly influenced by either gene sampling, methodology or multi-versus super-gene approach. All five estimates (Fig. 3, Table 1) are not significantly different (in terms of standard deviation) from the literature-based constraints. The fossil record can only provide us with a minimum estimation, i.e., with the date that a certain specimen appeared at the latest, it does not allow to date back to the last common ancestor and may suffer from biased sampling. The estimates from the literature for the nodes analyzed here display large variation; with the fossil-based dates usually defining the lower (more recent) and the molecular clock analyses the upper boundaries (Table 1). As an example, based on the fossil record the ancestors of mosses and seed plants were certainly separated by the time Rhynie chert was deposited in the Early Devonian, ~400 MYA (Taylor et al. 2005) or by the time spores associated with early land plants were deposited in the Middle Ordovician, ~460 MYA (Kenrick and Crane 1997), whereas the oldest known bryophyte fossils have been dated to the Late Devonian, ~360 MYA (Crum 2001). Based on molecular clock analyses, the estimates for the divergence of mosses and seed plants range from 450 MYA (Theissen et al. 2001) up to ~900 MYA (Hedges et al. 2004). In our analysis, this particular node was dated to 496 ± 84 (412–580) MYA, which is a narrower range. The same is true for the divergence time estimate of embryo plants/green algae and green/red algae (Fig. 3).

Taken together, our ortholog detection is a quick and easy method to detect cross-species single copy ortholog clusters. These can easily be selected (based on taxon strings) from the complete pool of clusters without carrying out a phylogenetic analysis. The cross-species orthologs presented in this work constitute a powerful set of markers, which can be used to increase the taxonomic resolution of phylogenetic analyses through inclusion of "pre-genome" organisms for which EST data is present. The ortholog clusters have been successfully applied to molecular clock analyses, dating the major separations of early plant evolution in concordance with the literature. Hence, our cross-species ortholog detection and subsequent rate analysis is a convenient method to update our knowledge about early evolution as the number of completely sequenced genomes increases.

Methods

Datasets

We used the following datasets to cover as thoroughly as possible the transcriptomes of all major lineages of photosynthetic eukaryotes.

Arabidopsis thaliana: 28,952 predicted proteins, The Institute for Genome Research (TIGR) ATH1.pep, (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/a_thaliana/annotation_dbs/ATH1.pep).

Oryza sativa: 88,149 predicted proteins, TIGR OSA1.pep, (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/BAC_PAC_clones/OSA1.pep).

Pinus taeda: TIGR PGI release 5 (35,053 tentative contigs, TCs) (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=pinus). All non-redundant Genbank full-length coding sequences from the genus *Pinus* (txid3337) (224 sequences) were used to calculate a codon usage table for ESTScan 2.0 (Iseli et al. 1999). The ESTScan prediction yielded 26,925 open reading frames (ORF) with an average length of 149 amino acids.

Physcomitrella patens: 26,123 clustered and assembled public EST (Lang et al. 2005), (http://www.cosmoss.org) were used for the prediction of ORF for *P. patens* using ESTcan 2.0 and a *Physcomitrella* specific model (Rensing et al. 2005), yielding 22,491 ORF.

Chlamydomonas reinhardtii: 19,832 predicted proteins from release 2.0, Joint Genome Institute (http://genome.jgi-psf.org/chlre2/chlre2.download.ftp.htm).

Cyanidioschyzon merolae: 5,013 translated mRNAs, (http://merolae.biol.s.u-tokyo.ac.jp/download/cds_nt.fasta).

Protein sequences of the cyanobacteria, the alpha-proteobacteria and non-photosynthetic eukaryotes were retrieved from Genbank (http://www.ncbi.nlm.nih.gov/entrez). For the collection of additional homologs throughout the available protein space using PSI-BLAST, the UniProt Knowledgebase, http://www.ebi.uniprot.org/database/download.shtml, was used.

BLAST searches and determination of orthologs

Parsing of the BLAST output, processing, and detection of orthologous proteins was done using Perl scripts and bioperl (Stajich et al. 2002) modules. A MySQL (http://www.mysql.com/) database was used to enable retrieval of the orthologs. In order to find orthologous protein pairs among a pair of genomes, we performed all-against-all searches between the two sequence spaces using BLASTP (Altschul et al. 1997). Significant hits were filtered according to these criteria: E-value <1E-4, >30% identity, alignment length >50% (with respect to the longer sequence) or >100aa, which places the hits above the twilight zone. A pair of proteins was labeled as orthologs if both sequences are each other's bi-directional best hit. To detect orthologs among more than two genomes, the accessions of orthologous pairs from the reciprocal all-against-all BLAST approach were concatenated. These strings were used as non-ambiguous keys for database queries. Genes were defined as orthologs if the exact string turns up in every BLAST result of the given subset of species (see Supplementary data). To determine the genetic origin of the orthologs, BLAST searches against all known proteins of cyanobacteria, alpha-proteobacteria and non-photosynthetic eukaryotes (cutoff 1E-4, 30% identity, hits at least 50 amino acids long) were carried out. From the respective best hits those with the highest percentage identity were selected. In addition, the gene family trees (see below) were manually inspected to determine as to which organisms were present in the sister branches with respect to the six orthologs.

Comparison with InParanoid and MultiParanoid

The pairwise InParanoid run for *Cyanidioschyzon merolae* and *Chlamydomonas reinhardtii* and also the runs for MultiParanoid were performed using the algorithm default values. Except for the confidence_score cut-off, which was set to 1.0, the MultiParanoid default values were used to retrieve only those proteins, which are main orthologs (seed orthologs) in InParanoid clusters.

Alignment, distances, phylogeny and rates

The individual ortholog clusters (6-tuples) were aligned using probcons 1.09 (Do et al. 2005). Pairwise identity values were calculated as inversed, uncorrected distances based on the multiple global alignments with the EMBOSS (http://emboss.sourceforge.net/) program Distmat. Based upon the multiple sequence alignment, pairwise distances within the 93 6-tuples of orthologs were calculated using PAML 3.14 (Yang 1997) with the Jones model and the cleandata option. ML trees were calculated from the 6-tuple alignments with TREE-PUZZLE 5.2 (Schmidt et al. 2002), applying the WAG (Whelan and Goldman 2001) substitution model with dataset-derived frequencies, using eight gamma-distributed rates. The WAG model was applied because it has been developed for globular proteins, which is the case for most proteins in our set. In addition, WAG was determined as the best model for 80% of the clusters of a random sample by applying ProtTest (Abascal et al. 2005). Calculation of likelihood values for the expected topology, likelihood ratio tests and statistical tests between best and expected trees were also carried out using TREE-PUZZLE. An extended majority rule consensus tree was calculated using the PHYLIP 3.6 (http://evolution.genetics.washington.edu/phylip.html) program Consense. Evolutionary rates were calculated based on the ML likelihood branch lengths using r8s 1.7 (Sanderson 2003). For the PL method the value of the smoothing parameter (10,000) was established using cross validation. For both the PL and the LF method cross validation and the TN algorithm were applied. Constraints for the nodes were set as lower and upper boundaries based on values from the literature (see Table 1). For each tree the calculations were iterated in order to constrain four of the five nodes and calculate a divergence time for the remaining node. In order to calculate phylogenies for the gene families surrounding and including each of the ortholog 6-tuples, homologs were detected using PSI-BLAST and the resulting hits were filtered, clustered and aligned as previously described (Richardt et al. 2007). Near identical sequences (containing up to 1% substitutions) were reduced to a single representative and phylogenetic trees were calculated using a combination of neighbor-joining (Saitou and Nei 1987) and ML (Schmidt et al. 2002) as previously described (Richardt et al. 2007).

Annotation and statistics

From the description lines of the ortholog 6-tuples a representative annotation was selected by majority rule, i.e., by best BLAST hit consensus annotation. In cases where no clear result could be achieved (11 out of 93), manual annotation was carried out with HMM profile searches using HMMer (http://hmmer.wustl.edu/) and the above-mentioned multiple global alignments as well as Interpro (Mulder et al. 2003) database cross references. GO terms were assigned to the sequences by performing InterPro searches and parsing of the results using the bioperl module (Bio::Tools::IPrScan) as described (Lang et al. 2005). The *A. thaliana* accessions were used to retrieve the GOA from TAIR (http://www.arabidopsis.org/tools/bulk/go/index.jsp). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were assigned to the orthologs using KAAS (KEGG Automatic Annotation Server 1.10; http://www.genome.jp/kegg/kaas/). Searches were performed against the "representative set" with a bit score threshold of 60 and the single-directional best-hit method (SBH). The resulting KO (KEGG Orthology) assignments were associated with the orthologs based upon the term assigned to the *Arabidopsis* sequence. In total, 68 (73%) of the orthologs could be assigned to a pathway. To check for significant deviations, *t*-tests and Fisher's exact tests were carried out. The resulting Fisher's exact tests *P* values were adjusted to control for multiple testing by calculating the false discovery rate (Benjamini and Hochberg 1995). Statistics were performed with R 2.1.0 (http://www.r-project.org/). For the visualization of the taxonomic profiles, the NCBI taxonomy information for all clustered sequences was retrieved, normalized (column-wise log ratio) and subsequently underwent average linkage clustering and heat map visualization, as described previously (Richardt et al. 2007).

Acknowledgments We would like to thank Wolfgang R. Hess, Clemens Kreutz and Klaas Vandepoele for helpful discussions. Financial support by the DFG (Re 837/10-1) is gratefully acknowledged.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9–e15

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86

Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: a molecular timescale without a clock. Evolution: Int J Org Evolution 59:1245–1258

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 57:289–300

Birchler JA, Riddle NC, Auger DL, Veitia RA (2005) Dosage balance in gene regulation: biological implications. Trends Genet 21:219–226

Blair JE, Shah P, Hedges SB (2005) Evolutionary sequence analysis of complete eukaryote genomes. BMC Bioinformatics 6:53

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16:1679–1691

Butterfield NJ (2001) Paleobiology of the late Mesoproterozoic (ca. 1200 Ma) hunting formation, Somerset Island, Arctic Canada. Precam Res 111:235–256

Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature 374:27–33

Crum HA (2001) Structural diversity of bryophytes. The University of Michigan Herbarium, Bloomfield Hills

De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol 523:1–7

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340

Dutilh BE, Huynen MA, Snel B (2006) A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. BMC Genomics 7:10

Fares MA, Byrne KP, Wolfe KH (2006) Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. Mol Biol Evol 23:245–253

Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool 27:401–410

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Fitch WM (2000) Homology a personal view on some of the problems. Trends Genet 16:227–231

Gray MW, Burger G, Lang BF (2001) The origin and early evolution of mitochondria. Genome Biol 2:REVIEWS1018

Gutierrez RA, Green PJ, Keegstra K, Ohlrogge JB (2004) Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? Genome Biol 5:15

Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Int Conf Intell Syst Mol Biol, 138–148

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature 389:33–39

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:15

Lang D, Eisinger J, Reski R, Rensing SA (2005) Representation and high-quality annotation of the *Physcomitrella patens* transcriptome demonstrates a high proportion of proteins involved in metabolism among mosses. Plant Biol 7:228–237

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 12:493–502

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189

Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends Genet 21:602–607

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99:12246–12251

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318

Mushegian AR, Garey JR, Martin J, Liu LX (1998) Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res 8:590–598

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476–D480

Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE (2002) Whole-genome analysis of photosynthetic prokaryotes. Science 298:1616–1620

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052

Rensing SA, Fritzowsky D, Lang D, Reski R (2005) Protein encoding genes in an ancient plant: analysis of codon usage, retained genes and splice sites in a moss, *Physcomitrella patens*. BMC Genomics 6:43

Richardt S, Lang D, Frank W, Reski R, Rensing SA (2007) PlanTAPDB: a phylogeny-based resource of plant transcription associated proteins. Plant Physiol 143:1452–1466

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Taylor TN, Kerp H, Hass H (2005) Life history biology of early land plants: deciphering the gametophyte phase. Proc Natl Acad Sci USA 102:5892–5897

Theissen G, Münster T, Henschel K (2001) Why don't mosses flower? New Phytol 150:1–8

Troitsky AV, Melekhovets Yu F, Rakhimova GM, Bobrova VK, Valiejo-Roman KM, Antonov AS (1991) Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. J Mol Evol 32:253–261

Vandepoele K, Van de Peer Y (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol 137:31–42

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809–818