Table of contents

1. Supplementary methods

1.1. *Ectocarpus* strain

Sequencing of genomic DNA and cDNA used material from *Ectocarpus siliculosus* strain Ec 32, which is a meiotic offspring of a field sporophyte collected in 1988 in San Juan de Marcona, Peru[1]. The strain is maintained in the *Ectocarpus* culture collection at Roscoff and in the Culture Collection of Algae and Protozoa at Oban, Scotland (accession CCAP1310/4).

1.2. DNA isolation and library construction

The DNA extraction protocol described by Apt et al.[2] was optimised for the Ec 32 strain. About one gram of tissue was ground to a powder under liquid nitrogen in a mortar and pestle using sand. After addition of 2 ml of extraction buffer (100 mM Tris-HCl pH 7.5, 1.5 M NaCl, 2% CTAB, 50 mM EDTA pH 4.5, 50 mM DTT) the tissue was extracted further in a Wheaton glass grinding tube. Extraction buffer was then added to 15 ml final volume, the sample shaken vigorously for 10 min and then incubated at 55°C for 2 hours in the presence of 25 units of proteinase K. After extraction with 1 volume of chloroform:isoamyl alcohol (24:1), polysaccharides were precipitated with 0.3 volumes of 100% ethanol and the chloroform-isoamyl alcohol extraction repeated. RNA was then removed by incubation overnight at 20°C in 4M LiCl and 1% β-mercaptoethanol and centrifugation at 13,000 rpm for 30 min. The DNA was then precipitated from the supernatant by addition of 0.8 volumes of isopropanol, redissolved in 500 µl of Tris EDTA buffer, extracted with phenol:chloroform:isoamyl alcohol (25:24:1) and then with chloroform:isoamyl alcohol (24:1) and precipitated in 0.3 M sodium acetate and 71% ethanol. The DNA was then purified on a CsCl gradient. Two plasmid libraries were constructed, one with 3 kb inserts in the vector pcDNA2.1 and the second with 10 kb inserts in vector pCNS. A BAC library was constructed in the pBELOBAC11 vector.

1.3. Analysis of DNA methylation

Enzymatic hydrolysis of DNA and HPLC analysis of nucleotide methylation was carried out as described in Monteuuis et al.[3], as modified from Jaligot et al.[4]. The *Ectocarpus* DNA was 50 µg of caesium chloride purified genomic DNA isolated from axenically grown strain Es32 sporophytes.

1.4. Genome sequencing

Paired, end-sequences were obtained from plasmid libraries with 3 kbp (2,233,253 reads) and 10 kbp (903,939 reads) inserts. Attempts to construct a large-insert bacterial artificial chromosome (BAC) library were unsuccessful but an additional 58,155 paired, end-sequence reads were obtained from a small-insert BAC library (average insert size 15 kbp). Assembly of all of these sequences (equivalent to more than 10-fold coverage) with Arachne[5] resulted in 14,043 contigs longer than 2 kbp that assembled into 1,902 scaffolds. Assembly of the mitochondrial and chloroplastic genomes was completed manually and sequences corresponding to these genomes were removed from the nuclear genome dataset. Several approaches were then used to remove contaminating bacterial sequence. Firstly, the assembly contained a near complete bacterial genome sequence. Additional sequencing was carried out to span gaps in this sequence and all the sequences corresponding to this complete bacterial genome were removed from the nuclear genome assembly. Additional scaffolds corresponding to contaminating bacterial DNA were identified based on the presence of intron-less genes that matched strongly to bacterial sequences in the public databases. Other information taken into account was the size of the scaffold (the smallest scaffolds often represented bacterial contamination) and percentage GC. Many of the bacterial contaminant scaffolds were also identified during the manual annotation. The final nuclear genome assembly, after removal of the organellar and bacterial sequences, consisted of 1561 scaffolds. Half of this assembled sequence was contained in scaffolds that were longer than 522 kbp (the N50 size).

1.5. Genetic map and analysis of pseudochromosomes

The construction of a genetic map for *Ectocarpus* will be described in detail elsewhere (S.H. *et al*., in preparation). Briefly, 408 polymorphic microsatellite markers were developed and used to genotype a population of 60 individuals derived from a cross between the sequenced strain Ec 32 and the Ec 568 strain from northern Chile. Mapping of 406 markers allowed 325 supercontigs to be associated with 34 linkage groups. Pseudochromosomes, generated by concatenating the mapped supercontigs for each linkage group, were analysed for evidence of large-scale heterogeneity in the distribution of features such as genes or transposable elements by calculating the percentage of bases assigned to these features within a 10 kbp sliding window. Percent GC content was also calculated using the same sliding window approach.

1.6. cDNA sequencing

In order to identify the transcribed regions of the genome, 91,041 sequencing reads were carried out on six cDNA libraries corresponding to different developmental stages and growth conditions. These included a normalised library corresponding to immature (non-fertile) sporophytes (57,520 ESTs) plus the following non-normalised libraries: immature (non-fertile) sporophytes (8,868 ESTs), immature gametophytes (17,200 ESTs), mature sporophyte with plurilocular and unilocular sporangia (3,695 ESTs), mature gametophytes with plurilocular gametangia (2,889 ESTs), and immature sporophytes stressed either in high salt medium or by addition of hydrogen peroxide (oxidative stress) (869 ESTs). Of the 66,388 immature sporophyte sequences, 44,000 corresponded to paired 5' and 3' end sequences.

Total RNA for cDNA libraries was extracted as described by Apt et al.[2] and polyA+ RNA was isolated using the PolyA-tract kit (Promega, Madison, USA). Both the normalised and non-normalised immature sporophyte libraries were constructed using the Cloneminer cDNA library construction kit (Invitrogen, Cergy Pontoise, France).

The cDNA libraries derived from immature gametophytes, mature sporophytes, mature gametophytes and immature sporophytes stressed either in high salt medium or by addition of hydrogen peroxide were constructed using the SMART method. Double stranded cDNA was prepared using the Super SMART PCR cDNA synthesis kit (Clontech, Mountain View, USA), amplified with 25 cycles of long distance PCR with the Advantage 2 polymerase mix and purified on a Chroma Spin column. Taq polymerase was then used to add

an adenine residue to the 3' ends of the cDNA molecules before cloning in pCR2.1-TOPO (Promega, Madison, USA). Following cDNA sequencing, low quality base calls and vector sequences were removed using Phred[6] and Seqclean (http://www.tigr.org/tdb/tgi/software/), respectively. Clustering and assembly of cDNA contigs was carried out using the TGICL tool, Megablast and the CAP3 assembler[7].

The EST sequences were used to determine the completeness of the genome sequence. To do this, paired reads from the opposite ends of the same clone were concatenated to produce 69,938 raw clone sequences from the 91,041 sequence reads. Poor quality or short clone sequences or clone sequences containing low complexity sequence were then removed. The remaining 64486 clone sequences were aligned with the genome using a two-step strategy. First, Blat[8] was used to generate alignments between the microsatellite repeat-masked EST sequences and the genomic sequence. The best match was retained (ID > 90%). Once the location of the transcript sequence was determined, the corresponding genomic region was extended by 1 kb on each side. Transcript sequences were then realigned to the extended region using EST_GENOME[9] (mismatch 2, gap penalty 3) to define transcript exons[10]. This approach allowed 62834 of the 64486 clone sequences to be matched with the draft genome assembly, indicating that the latter is 97.4% complete.

1.7. Whole genome tiling array analysis

A whole genome tiling array approach was used to generate a comprehensive transcriptome map of the *Ectocarpus* genome. This experiment involved hybridising eight microarrays bearing a total of 3,065,615 50-mer probes. These probes uniformly covered the entire *Ectocarpus* genome, excluding highly repetitive regions, separated by 10 bp gaps. In order to test the gene structure predictions, the slides carried 123,642 oligonucleotides that spanned predicted exon junctions (half the probe corresponding to the end of an exon and the other half to the beginning of the next exon downstream). These probes only produce a hybridisation signal if splicing occurs as predicted by the gene model. The oligonucleotide probes were synthesised on eight glass slides by Roche NimbleGen (Reykjavik, Iceland). See Stolc et al.[11] for details of the array design. Each slide carried the following control probes: 2,000 random sequence oligonucleotides and 929 oligonucleotides designed from the *Ectocarpus* genome. The slides were hybridised with two, labelled samples: 1) a mixture of labelled cDNA corresponding to RNA samples from mature sporophytes and gametophytes

and from immature sporophytes stressed either in high salt medium or by addition of hydrogen peroxide and 2) genomic DNA as a control. The genomic DNA control was used to identify probes that hybridised abnormally strongly to their target sequence. To do this, a cutoff was first determined based on the signal obtained with the random oligonucleotide control probes. Then, any probes that gave a signal above this cutoff following hybridisation with the genomic DNA were removed from the analysis, unless the signal obtained with the cDNA sample was at least twice as strong as for that detected with the genomic DNA. The expression data from the eight arrays were normalised using the quantile normalisation method and mapped onto the latest genome assembly. These data have been made available to the manual annotators via a graphical browser (http://ectocarp.us) during the genome annotation process. To test whether a gene had an unusually large number of high signal probes, the binomial distribution function was used to count the numbers of probes above and below the median value of the array, providing a statistical evaluation of whether the gene was expressed at a significant level[12].

1.8. Detection of transposable elements.

The *Ectocarpus* genome was analysed using the "REPET" pipeline (http://urgi.versailles.inra.fr/development/repet/), which uses methods for *de novo* TE identification that are adapted for the detection of nested and fragmented TEs. TE consensuses were generated "ab initio" by first searching for repeats using BLASTER, which implements an all-by-all BlastN[13] genome comparison, and then grouping the results using three clustering methods: GROUPER[14], RECON[15] and PILER[16] with default parameters. A consensus was then built for each group using the MAP[17] multiple sequence alignment program and each consensus was classified based on (i) BLASTER matches using TBlastX and BlastX[13] with the entire Repbase Update databank[18] and (ii) according to the presence of structural features such as terminal repeats (TIR, LTR, and polyA or SSR tails). For example, a consensus was defined as a MITE (i) if it contained TIRs; (ii) if it did not match via tBlastx or BlastX[13] with known TEs; (iii) and if its length, without its TIRs, was less than 500bp. The set of consensuses was then analyzed by an all-by-all BLASTER procedure to remove redundancies, *i.e.* when one consensus sequence was included within another at a 95% identity threshold and 98% length threshold. This step produced TE consensus sequences that represented ancestral copies of TEs subfamilies.

To identify TE families, the TE consensus sequences (often a structural variant of a family) were clustered into 197 groups, again using the GROUPER clustering method. Each group (representing a putative TE family) was then curated manually as follows: (i) the longest consensus was compared to the genome using BlastN, several genomic copies were then extracted along with 5 kbp of flanking sequences and a multiple alignment of these sequences was carried out to identify the full length element on the basis of sequence homology; (ii) a search for structural features in each full length sequence was carried out by comparing it to itself using BlastN to identify putative LTRs, TIRs, and by comparing it to an *Ectocarpus* tRNA library using BlastN; (iii) a search for similarity to known genes or TEs was carried out by translating the full length sequence into the six reading frames (with a minimum open reading frame length of 150 amino acids) and comparing it with the Genbank non-redundant protein sequences database using BlastP (putative genes identified in this way were re-integrated into the gene model repertoire) and with the Repbase library using TBlastX; (iv) the results of the previous steps were integrated and a sequence was confirmed as a TE if it showed characteristic structural features or sequence similarity to known TEs, as a tandem repeat if it had a tandem repeat structure, or as an unclassified repeat sequence if all the above steps failed to return conclusive results.

The *Ectocarpus* genome was then annotated with all the TE family reference sequences (which include the ~335 kb *Ectocarpus* double-stranded DNA virus EsV-1) using the "REPET" pipeline annotation step (TEannot). This pipeline is composed of the TE detection programs BLASTER[14], RepeatMasker[19] and Censor[20], and the satellite detection programs RepeatMasker, TRF[21] and Mreps[22].

To save computer time and reduce software memory requirements, we analysed the genomic sequences in segments of 200 kbp, which overlapped by 10 kbp. Each segment was independently analysed by the different programs. Simple repeats were used to filter out any spurious hits. TE or repeat copies of less than 20 bp, after removing simple repeat regions, were discarded.

To take into account the fact that TEs often insert into other TEs, resulting in fragmentation of the original TE, a specific "long join" annotation procedure was carried out, using an age estimate proxy to date repeat fragments. This employed the fact that the percentage identity between a fragment and its reference TE/repeat sequence can be used to estimate the age of the fragment. Consecutive fragments that occurred both in the genome and in the relevant reference repeat sequence were automatically joined (i) if the percentage

identity with the reference sequence differed by less than 2% (indicating that the two fragments have approximately the same age) and (ii) if they were separated by a gap of less than 5000 bp and/or by a mismatch region of less than 500 bp, or (iii) if there were nested repeats (i.e. the fragments were separated by a sequence, 95% of which that consisted of other younger repeat insertions all having a higher identity compared to their respective reference sequence). Fragments separated by more than 100 kbp were not joined. Finally, nested repeats were split if inner repeat fragments were longer than the outer fragments that had been joined.

The secondary structure of the Sower element RNA was predicted using a multiple alignment of 10 randomly selected full length copies on the RNAalifold web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi) with default settings.

Copy age estimations were calculated for categories and families of TEs by aligning each copy with a consensus of the family, obtaining the sequence identity, and converting it with the Jukes and Cantor formula to obtain an evolutionary distance.

For the expression analysis, we compared the *Ectocarpus* TE library to the *Ectocarpus* "non-stressed" cDNA library by BlastN. Significant hits (e-value $<e^{-20}$ and >95% identity) were selected and each hit was attributed to only one TE family (that with the highest identity). Frequency was then calculated for each TE family by dividing the number of matching ESTs by the total number of ESTs in the library. The loci corresponding to the top 100 positive peaks identified from the analysis of the *Ectocarpus* tiling expression array were extracted and compared to the *Ectocarpus* TE library using BlastN. Significant hits (evalue $<e^{-20}$ and >95% identity) were attributed to only one TE family (that with the highest identity).

Expression levels of the 98 fully annotated and putatively autonomous TE families with coding capacity was carried out by comparing these sequences with the *Ectocarpus* EST data set corresponding to unstressed growth conditions (82872 sequences). The comparison was carried out using BlastN with the filter for low complexity regions. Results were filtered for expectation values of less than e^{-25}, and identities of greater than 95%. Redundancy in the hits due to ESTs matching more than one element and/or the same element more than once was removed by retaining only the hit with highest identity to the query. We then counted the number of ESTs corresponding to each of the sequences queried. The EST frequency was calculated for each TE family as the number of ESTs found for a TE family divided by the size of the EST dataset (n=82872).

Searches were carried out for telomere sequences using a library of telomeric sequences from a broad range of eukaryotes. Details are available on request.

1.9. Analysis of methylation of transposable elements.

Ectocarpus or *Phaeodactylum tricornutum* genomic DNA (500 ng) was incubated with 50 units of endonuclease McrBC (which cleaves DNA containing methylcytosine; New England Biolabs, Hitchin, UK) in the presence or absence of 1 mM guanosine triphosphate GTP for 4 hours at 37°C. The reaction also contained 10 mM Tris-HCl (pH 7.9), 100 μg/ml bovine serum albumin, 50 mM NaCl, 10 mM $MgCl_2$ and 1 mM dithiothreitol. The enzyme was inactivated by incubation at 65°C for 10 minutes. Digestion of transposable elements was then assayed by semi-quantitative PCR using oligonucleotides specific to selected TE families.

1.10. Gene prediction and annotation of gene models

Gene prediction was carried out using the EuGène program[23], which incorporated the signal prediction program SpliceMachine[24]. A dataset of 1305 gene models (representing 9135 splice sites) was assembled manually using either cDNA sequence information or high quality alignments to homologous sequences from other genomes. This dataset was used to rebuild the interpolated Markov Models of the EuGène program and to optimise splice site prediction by SpliceMachine. A subset of 397 genes from the training dataset was used to optimise the programs to recognise genes in the *Ectocarpus* genome. EuGène was also optimised to incorporate several kinds of extrinsic data sources. These included protein sequences from the SwissProt and UniProt databases[25] and predicted proteins from four other stramenopiles: *Phytophthora sojae* and *Phytophthora ramorum*[26], *P. tricornutum*[27] and *Thalassiosira pseudonana*[28]. EuGène was also supplied with spliced alignments with the *Ectocarpus* cDNA sequences, generated using GenomeThreader[29]. Paired cDNA sequence reads were provided to EuGène as a separate data source so that the coupled information could be exploited. Prior to gene prediction, repeat sequences were masked, using RepeatMasker (http://www.repeatmasker.org), with the library of *Ectocarpus* repeat sequences described above. The predicted genes were analysed to generate genome wide statistics such as average

gene size or average number of introns per gene. All such analyses were also carried out on the subset of models that were fully supported by cDNA sequences to investigate the effect of annotation errors on the gene model statistics (data not shown). These analyses found no evidence that biases had been introduced by the gene prediction.

Functional annotation of the predicted gene models was carried out based on the identification of protein domains using the InterPro database and on BlastP matches against the SwissProt and UniProt databases[25]. Subcellular localisation was predicted using the Hectar program, which uses support vector machines to integrate data from several predictors[30]. A database and annotation interface (BOGAS: http://bioinformatics.psb.ugent.be/webtools/bogas/) was created for the manual annotation of the *Ectocarpus* genome and additional annotations were entered manually for 5439 (33.2 %) of the predicted genes (note that only 64.2% of the *Ectocarpus* proteins match a sequence in the public databases, with a Blast e-value of $< e^{-5}$).

To compare the 16,256 proteins encoded by the *Ectocarpus* genome with those encoded by previously sequenced genomes, each sequence was compared to the nr_prot database (NCBI) using Blast and all matches with an e-value of less than e^{-5} were retained. The taxononic information associated with each Blast match was then used to determine whether the *Ectocarpus* protein shared significant similarity with a sequence from one or more of the following groups: diatoms, oomycetes, plants, metazoans and fungi.

1.11. Searches for duplicated regions of the genome

A search for duplicated regions of the genome was carried out using the i-ADHoRe 2.0 program[31]. Gene pairs were regarded as homologues if the aligned region was longer than 150 amino acids and if the sequences shared more than 30% similarity (LiRost criterion). This list of gene pairs was then provided, as input, to the iADHoRe program (parameter settings: gap_size = 40, cluster_gap = 50, q_value = 0.9, prob_cutoff = 0.001, anchor_points = 3, level_2_only = false, tandem_gap = 2). The results of the i-ADHoRe analysis were also used to identify tandem duplicated genes. A tandem duplication was defined as two homologous genes separated by less than 2 non-homologous genes on the chromosome. The two tandem duplicated genes could be in any orientation with respect to each other.

An analysis of the level of synonymous substitutions (Ks) between gene pairs was carried out to look for evidence of genome duplication events using the method described by Rensing *et al*.[32].

The online version of the promoter prediction tool EP3[33] (available at http://bioinformatics.psb.ugent.be/software/details/EP3) was run on the 20 largest scaffolds of the *Ectocarpus* genome to investigate the feasibility of predicting *in silico* promoter prediction. The following parameter settings were used: Genome size < 2Gb, no precomputed model, no repeats masked.

1.12. Analysis of introns, untranslated regions and intergenic regions

Logos representing the sequence composition of splice sites (Supplementary Fig. 1) were generated by analysing 20 bases up- and downstream of all the donor and acceptor sites supported by an EST sequence. Sequence logos were constructed from these sets using the Weblogo program[34].

To analyse polyadenylation sites, the trimming of the cDNA sequences was repeated without removal of polyA regions. The set of cDNA sequences which terminated with a polyA tail was then established. These sequences were mapped onto the genome and sequences that mapped to multiple loci were removed. For each polyA site, the genomic region surrounding the position corresponding to the polyadenylation site was extracted from the genome sequence (300 bases upstream, 100 downstream). After removal of redundant cDNA sequences that corresponded to identical polyadenylation sites and sequences that contained undetermined bases (N), this resulted in a dataset of 4422 sequences. Of this set of sequences, 1260 sequences could be uniquely assigned to 511 gene models. All analyses were performed both on all the complete set of 4422 polyA regions and on the 1260 polyA regions that corresponded to gene models, unless noted otherwise.

Polyadenylation signals were identified using the UTRscan service[35] (http://utrdb.ba.itb.cnr.it/tool/utrscan). An additional filtering step was applied to the output based on the location of the signal (less than 40 bp from the polyadenylation site). Tools available on the RSAT website[36] (http://rsat.ulb.ac.be/rsat/) were used to search for new sequence motifs. The background frequency files were created by applying the criteria used for the polyA sites to the 5' ends of the genes.

1.13. Experimental and *in silico* identification of small RNAs

Sporophyte and gametophyte small RNAs were isolated and prepared for sequencing by Fasteris Life Sciences (Plan-les-Ouates, Switzerland). Ten micrograms of total RNA of each life cycle generation, extracted as described by Apt et al.[2], was separated on a polyacrylamide gel and the 15 to 30 nucleotide fraction isolated by excision. Addition of single-stranded adapters and PCR amplification was carried out using the DGE-Small RNA kit (Illumina, San Diego,USA) and a small number of inserts were sequenced to check the quality of the libraries. Flow-cell preparation on the Solexa Cluster Station and high throughput sequencing was carried out by Genoscope using a Solexa Genome Analyser (Illumina).

Read filtering and mapping onto the genome. In order to keep only full-length small RNAs and to eliminate truncated sequences, only the reads that had a signature corresponding to the first 6 nt of the 3' adapter sequence in their 3' end were retained. These reads were mapped to the *Ectocarpus* genome using BlastN (filter off, e-value cutoff 10) using all the scaffold sequences (199 Mbp) plus the set of smaller fragments (< 2kb) that could not be assembled in the main genome (representing an extra 15 Mbp). For each read, the longest alignment to the genome of between 18 and 30 nt without any gaps and mismatches was kept.

Small RNA expression normalization. The read counts were transformed to reads per million (RPM) to take into account the variable number of reads carried out for each of the two libraries. If several reads mapped to exactly the same location on the genome, therefore representing the same small RNA sequence, the counts of each read was added to have a unique count value for each unique small RNA sequence. Furthermore, each unique small RNA count was normalized by the total number of loci found for this sequence in the *Ectocarpus* genome. The final normalized expression values were obtained by log2 transformation.

Annotation of rRNAs, tRNAs and snoRNAs. The program t-RNAscan SE detected 579 tRNAs in the *Ectocarpus* genome[37]. For rRNAs, the RNAmmer program was used to predict 78 rRNA sequences, using a cutoff score value linked to the rRNA type[38]. The 245 snoRNAs were annotated using the snoSCAN software[39].

Transposable element-associated small RNA randomization tests. This was carried out using a similar procedure as the one described by Kasschau et al.[40]. The coordinates of transposable elements were transformed into segments of 250 nt, overlapping by 125 nt. A total of 10,000 bins were selected at random from this pool. This set was compared with a

genome wide set of 10,000 sets, each containing 10,000 randomly selected genome bins. The transposon set contained 2,321 small RNA loci, compared with 677 ± 118 for the ransdom genome sets. The corresponding z-score has a value of 13.9, with an associated p-value < 0.0001.

Annotation of miRNAs. Small RNAs that did not map to any rRNA, tRNA, snoRNA, exon sequence and that had between 1 and 30 loci in the genome were clustered using a window that included 100 nt upstream and downstream. In order to exclude dense regions with a very high number of small RNAs, clusters longer than 40 nt but shorter than 350 nt were retained for further analysis. For each cluster sequence, all possible stable secondary structures were predicted using RNALfold[41]. A set of recently published guidelines were followed to set the rules defining a *bona fide* miRNA sequence[42]. The secondary structure of the miRNA precursor had to encode a stem-loop structure with a free energy of ≤ -30 Kcal/mol. The miRNA/miRNA* duplex had to be supported by sequence reads for both the miRNA and the miRNA*, it had to be located in a stem region with no more than 5 unpaired bases, and bulges could not exceed 3 bases. There also had to be a differential expression between the miRNA and the miRNA*. The miRNA was defined as the sequence having the highest expression level. The miRNA/miRNA* duplex had to display the 2 nt (± 2 nt) overhangs that are characteristic of Dicer processing of the stem-loop structure (Supplementary Fig. 2).

Computational prediction of miRNA targets. We used a computational search similar to the one described by Allen and collaborators[43]. A FASTA search was done for all miRNAs against gene predictions, for both CDS and UTR sequences. The results were parsed and a score was calculated for each miRNA/target pair. Each mismatch was counted as 1 while GU pairs were counted as 0.5. Each mismatch or GU pair within the seed region (between positions 2 and 12 of the miRNA sequence) was counted as twice its normal score. Only one bulge (1 nt) was allowed in the duplex. Sequences with scores inferior or equal to 3.5 were selected as candidate miRNA targets.

Statistical analyses. All statistical analyzes and graphs were performed using the R statistical package[44].

1.14. Protein domain analysis

To identify protein domains, InterProScan was run with default settings on the complete sets of predicted proteins from the *Ectocarpus* genome and from 16 additional genomes from a broad range of species. The results of this analysis were then used to score for the presence and abundance of each domain in each genome. To test whether *Ectocarpus* possessed a significantly different number of proteins with a particular domain than another species, a Fisher Exact test was carried out. The p-values obtained were then corrected for multiple testing using the Bonferroni correction.

1.15. Dollo analysis of gene family loss and gain during evolution

The dataset for the Dollo analysis consisted of a broad range of complete genome sequences from unicellular and multicellular organisms (Supplementary Table 1). To delineate gene families, a similarity search was performed (all-against-all BlastP; e-value cutoff e^{-10}) with all proteins of the dataset (17 species). Gene families were constructed with MCLBLASTLINE[45] (http://micans.org/mcl/; inflation factor of 2.0) based on the BlastP analyses. Using the MCL-families, phylogenetic profiles were constructed reflecting the absence or presence of a certain gene family in every species. Based on these profiles, the gene families that were single copy in all organisms of the Dollo-dataset were extracted. For every single copy core gene family, a multiple alignment was created using MUSCLE[46]. The different multiple alignments were concatenated into one large alignment that was further edited using BioEdit[47], yielding a concatenated alignment of about 10.000 amino acids. A distance matrix was calculated for this edited concatenated alignment based on Poisson correction. The phylogenetic tree was constructed with the neighbour-joining algorithm, using the software package TREECON[48]. Bootstrap analysis with 500 replicates was performed to test the significance of the nodes. All nodes received a bootstrap-value of >70%. The DOLLOP program of the PHYLIP package[49] was then used to construct a parsimonious evolutionary scenario in which loss and acquisition of gene families was mapped onto the branches of a phylogenetic tree. The DOLLOP program is based on the Dollo parsimony principle, which assumes irreversibility of character loss[50].

Putative functions were assigned to genes and gene families using the Gene Ontology (GO) database[51]. Proteins were assigned to GO categories using Blast2GO[52] and

Interpro2GO. Note that proteins mapped to a particular GO category were also explicitly included into all parental categories. Gene families were annotated by listing the GO labels for all of the genes in each family. A weight, equal to the percentage of genes with each GO annotation from within the same subcategory (molecular function, cellular component, biological process), was attached to all of the GO labels. Only GO labels with a weight greater than 40% were considered as representative for the family. The statistical significance of functional GO enrichment was evaluated by using the hypergeometric distribution, whereas multiple hypotheses testing was carried out using FDR[53]. GO labels occurring in fewer than 10 gene families across the whole dataset were discarded before the statistical analysis. To identify more general trends, the GO annotation was converted into GO Slim annotation, using the script map2slim.pl provided by the GO-perl package of Gene Ontology (http://www.geneontology.org). The percentage of gene families annotated with each GO Slim category was calculated for all sets of gained and lost gene families at each timepoint of the phylogenetic tree (see Fig. 3). The number of families with a certain label was divided by the number of lost (or gained) families with GO annotation in that tree at the corresponding timepoint. The means and standard deviations of these percentages was then calculated and the matrix of percentages was converted to a matrix of z-scores. The z-scores were hierarchically clustered (complete linkage clustering) using Pearson correlation as a distance measure. Clustering and visualization was carried out using Genesis[54].

1.16. Gene family expansions

For all gene families present in *Ectocarpus* and in at least two other species, the mean gene family size and standard deviation of the phylogenetic profiles were calculated. The matrix of these profiles was transformed into a matrix of *z*-scores to centre and normalize the data. The profiles of interest were then hierarchically clustered (complete linkage clustering) using Pearson correlation as a distance measure. The clustering and visualization was carried out using Genesis[54]. A description was added to each family based on the most frequently occurring gene description in that family.

1.17. Clusters of genes with similar functions

To detect genes with similar functions that are clustered in the genome, a Gene-Ontology-based (molecular function or biological process) search was carried out using the C-Hunter program[55] (http://fcg.tamu.edu/C_Hunter/). The program was provided with the *Ectocarpus* genes and their corresponding GO labels, gene coordinates and orientation. An e-value cut-off of e^{-5} was used. Clusters may simply correspond to local gene duplications or may be clusters of genes of independent evolutionary origin involved in the same biological process.

1.18. Identification of endosymbiosis-derived genes

The predicted protein sequences from the V2 version gene predictions for *Ectocarpus* and V3 for *T. pseudonana* (for comparative purposes) were compared against a local database of predicted protein sequences derived from complete genome sequences. For the red algae, the database also included EST-derived data to improve the rather low genome sequence coverage within this taxon (only one complete red algal genome sequence was available, and this for the extremeophile *Cyanidioschyzon merolae*, which possesses only 5331 protein coding sequences). Gene prediction for the ESTs was carried out using Augustus[56]. Phylogenomic analyses were performed with an improved version of Phylogena referred to as Phylogena2[57] (B.B. *et al*. in preparation; the source code is available at http://aforge.awi.de/gf/project/phylogena/) using the entire set of *Ectocarpus* V2 gene predictions. Up to 50 best hits with an e-value below 10^{-5} were selected for phylogenetic analyses. Multiple alignments were calculated with muscle[46] using default parameters, and neighbour joining phylogenetic trees with 100 bootstrap replicates were generated with quicktree[58]. Phylogenies in which the query sequence grouped with particular taxa were identified using an interface to PhyloSort[59] implemented in Phylogena2. Phylogenetic profiles (i.e., the distribution of a sequence across the taxa examined) were also extracted from the phylogenies.

Genes putatively derived from endosymbiosis with a red alga (defined as "red" genes) were identified based on their forming a clade with red algal homologues. Homologues from other chromalveolates were allowed to group within such a clade. Similarly, genes putatively derived from endosymbiosis with a green alga (defined as "green" genes) were identified based on their forming a clade with homologues from the green lineage (archeaplastida) and

chromalveolate homologues were allowed to group within such clades. Genes were only classed as "red" or "green" if no non-stramenopile sequences branched within such clades and if the clade had a bootstrap support above 70 %. This criterion is more stringent than the MCL clustering used for the Dollo analysis.

To provide additional support for the candidate "red" and "green" genes, maximum likelihood trees were also generated using PhyML[60] with the WAG substitution model and SH-like support values (using a cutoff at 0.7 in support values).

1.19. Annotation of transcription associated proteins

Three sets of TAP classification rules for plants [61-63] were combined and expanded to yield a set of classification rules for 111 families. The initial set of rules was adopted from three previous publications, PlantTFDB [61], PlnTFDB [62] and PlanTAPDB [63]. Potential conflicts between those three sources were manually evaluated and eliminated based on an analysis of the scientific literature. This set was then expanded by adding recently defined families or subfamilies from published sources. The rule set for each family consists of at least one entry defining a "should" rule, i.e. a mandatory domain for that particular family. Additional entries may define further "should" or "should not" (forbidden) domains (Supplementary Table 2). All domains relevant for classifying the TAPs were represented by an “ls” HMM. If available, the HMMs were retrieved directly from the ‘PFAM_ls’ database [64]. For the remaining domains, HMMs were custom-made using multiple sequence alignments (MSAs) to identify the conserved domain(s) of interest. The MSAs used for creating the custom HMMs were downloaded from PlnTFDB [62]. For domains not represented in this database, MSAs were created as follows. Blast searches with a protein query containing the respective domain yielded homologous hits defined by having at least 30% sequence identity with the query over a minimum length of 80 amino acids [65]. Those hits were aligned using MAFFT linsi [66] and manually curated using Jalview [67]. The conserved domain of interest was extracted and the HMM calculated with HMMER 2.0 (http://hmmer.janelia.org/) using ‘hmmbuild’ with the default parameters to generate "ls" HMMs and subsequently ‘hmmcalibrate’ with the option ‘--seed 0’ to obtain reproducible results. Gathering cutoff (GA) values were defined for each custom HMM. The GA was set as the lowest score of a domain-containing protein (true positive) after a ‘hmmpfam’ search (using an E-value cutoff of 1e-5) against the full proteome sets of several different species and considering the alignments of all hits. In order to avoid

sampling bias, only fully sequenced genomes were used in this study. For each organism, the complete set of proteins derived by conceptual translation of the nuclear gene models (using the filtered/selected model per locus) was combined with the proteins encoded by the respective mitochondrial genome, if available. All proteins can be unambiguously identified via their fasta id. We used a unique five letter code for each organism (Supplementary Table 1) followed by "mt" (mitochondrial) or "pt" (plastid), if applicable, and the accession number of the gene model. Using all proteins of the investigated organisms as query, 'hmmpfam' searches were performed against an HMM library containing all 129 domains necessary for the TAP classification (Supplementary Table 3). The GA was used during this procedure to minimize the number of false positive hits. GA values were either provided with the 'PFAM' HMMs or defined as described above. The classification rules (Supplementary Table 2) were subsequently applied to all proteins for which at least one significant domain hit was found. In cases where the domain composition of a protein matched more than one classification rule, the 'should' rule with the highest score determined the family into which the protein was categorized. Highly similar domains which are often found in the same or overlapping regions of a protein were treated in similar fashion, i.e., the domain with the lowest E-value/higher score was used for the subsequent classification. This procedure was necessary in four cases, namely i) Myb_DNA-binding and G2-like_Domain, ii) NF-YB, NF-YC and CCAAT-Dr1_Domain, iii) PHD and Alfin-like and iv) GATA and zf-Dof. In addition, a Boolean "OR" rule was applied to three families. In these cases one out of two domains was found to be necessary and sufficient for a protein to be classified into the corresponding family. This rules was applied to the bZIP, HD-Zip and GARP_ARR-B families. Whenever the presence of a combination of domains lead to more than one possible family classification, TF was favoured over TR or PT. This situation encountered in 14 cases, resulting in 14 rules. Taken together, the ruleset (Supplementary Table 2) defined 111 TAP families using 223 rules of which 134 represented mandatory and 89 forbidden rules. In total, 129 domain HMMs were used, of which 16 were custom-made and 113 obtained from the PFAM database (Supplementary Table 3). The application of this ruleset identified a total of 401 putative TAP genes in the *Ectocarpus* genome (see Supplementary Table 4). Several of these gene families (AN1/A20 Zinc finger, ARID, bHLH, bZIP, CCAAT-HAP5, C2H2, MED7, RWP-RK) were manually annotated and curated using the genome browser. bZIP TFs were identified using sensitive, custom-made Hidden Markov Models (HMMs). The HMMs were trained on bZIP sequences from the animal lineage (vertebrates, other non-vertebrate chordates, insects, diploblastic animals), choanoflagellates, ascomycetes, the green lineage (dicots, monocots,

mosses, chlorophytes) and from other unicellular eukaryotic organisms, including diatoms and stramenopiles. Both lineage-specific and general eukaryotic HMMs were trained. Proteins retrieved by the HMMs were manually checked for the presence of a basic region with an adjacent coiled-coil domain and knowledge of the key biochemical characteristics of the bZIP domain was used to evaluate and curate the genes identified. The bZIP proteins were classed as either typical (the leucine zipper region was a predicted coiled-coil of three or more heptads) or atypical (in which case only the presence of the basic region was required). Significant expansion of individual families between different groups of organisms was analyzed using T-tests with subsequent false discovery rate correction [68] or 2-fold ratio of means (if less than two samples per group). Tests and visualization was performed using Analyst (Genedata, Basel, CH). Multiple alignment of the bZIP region was performed with the HMMalign program of the HMMER software and was manually curated. Phylogenetic analysis was performed with the Phylip software using the protdist and neighbor programs.

1.20. Annotation of P-loop GTPases

Representatives of all known P-loop GTPase lineages of the TRAFAC class were used to compile an inventory of homologs in *Ectocarpus* and other selected eukaryotes, including additional stramenopiles with sequenced genomes and representatives of other major eukaryotic lineages. Putative orthologous relationships of the *Ectocarpus* GTPases to proteins in the other eukaryotes analysed were established based on reciprocal BlastP comparisons, in some cases complemented by visual inspection of multiple alignments and/or maximum-likelihood phylogenetic analyses. The orthology assignment is tentative in several cases and needs be confirmed/refined by a broader sampling of homologs and more rigorous analyses. Functional annotation is based on surveys of the literature dealing with putative orthologs in "model" species (mainly human, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*). Domain architecture was investigated using SMART[69] and Pfam[70]. Possible organelle-targeting transit leader sequences were predicted using TargetP and SignalP[71].

1.21. Phylogenetic analyses of receptor kinases

A total of 50 protein sequences corresponding to representative eukaryotic receptor kinases and related cytosolic kinases were retrieved from public databases (see Accession numbers in Supplementary Table 5). Using the Phylogeny.fr web site[72], their cytosolic serine/threonine and tyrosine kinase domains were aligned with MUSCLE, together with the kinase domains of 5 putative membrane-spanning receptor protein kinases from *Ectocarpus* (see section 2.3.2.) and of 5 newly-identified paralogous proteins from *Phytophtora* (see Supplementary Table 5). After removal of gaps, phylogenetic analyses were carried out on 163 amino acid positions corresponding to the kinase domains of the 60 aligned sequences. Maximum Likelihood and neighbour-joining approaches were performed using PHYML and BioNJ methods, respectively, with the default parameters proposed by the web interface. For both methods, bootstrap analyses of 100 replicates were used to provide confidence estimates for the phylogenetic tree topologies.

2. Supplementary notes

2.1. Genome structure and organisation

2.1.1. Genome composition

A genetic linkage map was constructed for *Ectocarpus* using microsatellite markers derived from the genome sequence (see supplementary methods). This map allowed 325 supercontigs, representing 137 Mbp (70.1% of the genome), to be anchored onto 34 linkage groups (26 major linkage groups and eight minor linkage groups). It is likely that the eight minor linkage groups will coalesce with the larger linkage groups as more markers are added to the map, in which case the data from the genetic map agrees well with the estimated number of 25 chromosomes reported for *Ectocarpus*[73,74]. Pseudochromosomes were generated by concatenating the mapped supercontigs for each linkage group (Supplementary Fig. 3) and these structures were analysed for evidence of heterogeneity in terms of gene and TE distribution and GC content. The overall gene and TE densities on each linkage group were quite variable (Supplementary Fig. 4) but there was no clear evidence for regions that were particularly gene-poor or TE-rich within linkage groups (as might be expected for centromere regions or heterochromatic knobs, for example). Note however that such regions would have been expected to present greater problems at the assembly stage and therefore may be missing from the assembled genome sequence.

Overall, therefore the genome did not exhibit any obvious large-scale structures. However, we did note that linkage groups 29 and 30 were particularly rich in TEs and that linkage groups 30 and 32 exhibited a particularly low density of genes (Supplementary Fig. 4). Gene density for linkage group 30, for example, was only 60% of the average for the genome as a whole (0.58 genes per kbp compared with 0.89 on average), whilst TE density was almost twice as high as the average (8.4 per kbp compared with 4.6 on average). Moreover, a high percentage (74%) of the genes on linkage group 30 were annotated as hypothetical or conserved hypothetical indicating that the gene density may even be overestimated (assuming, as indicated by the tiling array analysis, that a proportion of the hypothetical genes are overpredicted). Analysis of the function of the genes on linkage group

30 indicated a statistically significant enrichment in GO-labels associated with metal ion transport. Linkage group 32 exhibited similar trends to linkage group 30, although they were less marked. In terms of gene function, there was evidence of enrichment in genes associated with protein localisation on linkage group 32. It is not known at present how variations in structure such as those observed for linkage groups 30 and 32 are related to genome function.

None of the pseudochromosomes had an unusual GC% content, they were all close to the genome average of 53%.

2.1.2. Genome methylation

Ectocarpus genomic DNA was analysed for deoxycytosine methylation (5mdC) by HPLC analysis of hydrolysed DNA. No 5mdC was detected in the *Ectocarpus* DNA sample but methylation was detected in a control sample from oil palm (Supplementary Fig. 5). Calibration with purified 5mdC to determine the sensitivity of the HPLC detection method indicated that the percentage of 5mdC in *Ectocarpus* is below 0.035%.

2.1.3. Duplication events and gene organisation

Two types of analysis were carried out to search for evidence of genome duplication events. These were based on an analysis of the number of synomymous substitutions per synomynous site in duplicated gene pairs (Ks analysis) and on a search for duplicated blocks of genes based on synteny at the gene level using the program iADHoRe. Neither of these methods detected any evidence that a large-scale duplication of the genome had occurred during evolution (although it is possible that an ancient duplication would not have been detected by this analysis). The histogram plot of the Ks analysis did not indicate any peaks (which would have indicated large scale duplication events) but showed only a continuous mode of small scale duplication (Supplementary Fig. 6), and an analysis with the iADHoRe program identified only one duplicated block of genes, a group of histone genes that was found at four locations in the genome (Supplementary Fig. 7). The *Ectocarpus* genome also has an unusually low number of tandem duplicated genes in proportion to its size (823; Supplementary Fig. 8). Most of the tandem duplications consisted of just two genes but a small number of more extensive arrays, including up to 20 duplicated genes, were also identified. One or two non-homologous genes were intercalated between the tandem

duplicated genes in 19.8% of the pairs identified. Many of the tandem duplicated gene pairs were divergently (16.1%) or convergently (15.5%) orientated.

About 61.5% of the genes in the *Ectocarpus* genome are arranged in an alternating manner, with adjacent genes on opposite strands of the DNA. Analysis of alternating genes in a range of other genomes indicated that small, compact genomes tend to have more alternating genes than expected (50% would be expected if the genes were organised randomly), whereas large genomes have less than expected (Supplementary Fig. 9). Compared to these other genomes, the *Ectocarpus* genome is unusual in that it has a high number of alternating genes relative to its size. The probable absence of genome duplication events and the scarcity of small-scale gene duplication events in *Ectocarpus* may have played an important role in the retention or maintenance of this structured, alternating pattern of gene organisation. A higher frequency of tandem gene duplication would probably have disrupted the alternating organisation of the genes. Another unusual feature is that the intergenic regions between divergently transcribed genes are often very short (less than 400 bp in 29% of cases; Supplementary Fig. 10). Short intergenic regions, as observed for many divergently expressed gene pairs, have been shown to stabilise genome structure by reducing the probability of recombination events occurring between genes[75].

In yeast, there is a tendancy for highly co-expressed genes to be adjacent in the genome[76] and divergently expressed genes exhibit a higher degree of co-expression than convergently expressed or co-oriented gene pairs[77,78]. In *Ectocarpus*, large numbers of genes have been shown to exhibit significant alterations in transcript abundance following stress treatments[79]. When adjacent pairs of genes that exhibited significant changes in expression level following these stress treatments were analysed, they consistently showed a greater degree of co-regulation (i.e. both genes up-regulated or both genes down-regulated) compared with a randomised sample of non-adjacent genes (Supplementary Fig. 11). Approximately 50% of non-adjacent pairs of genes exhibited co-regulation, as expected for a random sample. The effect was even more marked when the analysis was restricted to pairs of genes that had been detected as being differentially expressed following two or three of the stress treatments. However, in contrast to the situation in yeast, there was no evidence that divergently transcribed genes showed a great level of co-regulation than gene pairs with other orientations (data not shown). This is consistent with the fact that divergent gene pairs are found with almost the same frequency as convergent gene pairs in the genome (4078 and 4076 pairs,

respectively) suggesting that it is the alternating arrangement that is conserved rather than any particular orientation of gene pairs.

2.1.4. Gene structure

The *Ectocarpus* genome is predicted to contain a total of 16,256 genes. In general these genes are rich in introns, with an average of 6.98 per gene. Very few *Ectocarpus* genes lack introns completely; the fraction of genes without introns (5.3%) is the smallest for any genome reported to date (Supplementary Fig. 12). Typical genes consist of short exons interspersed with introns that are exceptionally large for a genome of this size (704 bp on average; Supplementary Fig. 13). In consequence, the fraction of the genome represented by the introns in *Ectocarpus* (40.86%) exceeds that of any of the other genomes we analysed (Supplementary Fig. 14). Consensus intron donor and acceptor sequences are shown in Supplementary Fig. 1. GT donor sites are most abundant but GC donor sites are predicted to occur in 4.6% of the introns.

To verify that the unusual features observed for the *Ectocarpus* genes were not due biases introduced during the gene annotation process, we calculated gene statistics for three well-supported gene sets: the set of manually annotated genes, the set of genes with full EST support (complete, overlapping EST coverage) and the set of single copy genes with single copy orthologues in four other stramenopile genomes (Supplementary Table 6, Supplementary Fig. 15). On the whole, this analysis supported the results obtained from analysing the complete gene set, except that the genes with full EST support tended to be shorter on average and, consequently, to possess fewer introns and exons. However, we believe that this difference was due principally to the tendency for short genes to have full EST coverage (because it is easier to produce full-length cDNAs from short transcripts) and hence did not indicate a bias in the complete gene set. This is supported by the fact that the two other sub-sets of genes showed similar statistics to the complete gene set. All subsequent analyses were carried out with the complete gene set.

Analyses carried out with the Easy Promoter Prediction Program[33] (EP3), which uses GC content and large-scale structural features of DNA, failed to detect any candidate promoter regions. Similarly, no evidence was found for a conserved consensus sequence surrounding the ATG initiation codons of *Ectocarpus* genes.

With a mean length of 848 bp, the 3'UTRs of *Ectocarpus* genes are exceptionally long for an organism with a relatively small and compact genome (Supplementary Table 7). This mean 3'UTR length is comparable to that of the mouse, which has a genome that is more than 13 times larger (Supplementary Fig. 16). In contrast, the lengths of the *Ectocarpus* 5' UTRs are not exceptional and, in consequence, the ratio of the lengths of the *Ectocarpus* 3'UTR/5'UTR was higher than that of any other organism we analysed.

One possible explanation for the existence of long 3'UTR regions of *Ectocarpus* genes may be that they contain elements that are important for gene regulation. We investigated the degree to which *Ectocarpus* genes use alternative polyadenylation sites, as alternative transcripts generated by this process could potentially bear different complements of 3'UTR regulatory elements. Analysis of the cDNA sequence data allowed the identification of 4422 polyadenylation sites, 1260 of which could be assigned uniquely to one of 511 gene models. Functional annotation of the genes with multiple polyadenylation sites indicated that 11% corresponded to genes involved in photosynthesis and 10% to ribosomal proteins.

The number of alternative polyadenylation sites identified per gene was highly variable. Multiple polyadenylation sites were detected in only 157 (31%) of the 511 genes analysed (representing 906 polyA sites). In 60% of these cases (546 polyadenylation sites) the multiple polyadenylation sites for each gene were located within a window of 30 bp. These sites most likely represent 'microheterogeneity'[80], in which case they can be treated as single polyadenylation sites. Based on this assumption, the true number of genes with multiple polyadenylation sites would be 104 (containing 360 polyadenylation sites), with the number of sites per gene ranging between two and 19. This suggests that there are 1.41 true alternative polyadenylation sites per *Ectocarpus* gene on average (at least for the genes that are expressed at a high enough level to be detected by the cDNA sequencing approach used). This value is not exceptionally high compared to other species that have been analysed (for example 1.46 and 2.10 for mouse and human, respectively[81]). This analysis did not, therefore, provide any support for the hypothesis that *Ectocarpus* uses a combination of long 3'UTRs and a high level of alterantive polyadenylation to create mRNAs bearing different 3' regulatory signals. However, it should be noted that this analysis does not rule out the possibility that the long 3'UTRs contain regulatory signals in a general sense.

To complete the analysis of the polyadenylation sites, searches were carried out for potential polyadenylation signals. The near upstream element (NUE) AAUAAA is highly conserved in mammals but is less conserved in non-mammals[82,83]. This is also the case in

Ectocarpus, where the canonical NUE was present in only 12.7% of the 511 regions upstream of polyadenylation sites (12.2% for the total set of 4422 sequences). This is only slightly higher than the percentage reported for flowering plants[80]. The GU-rich far upstream element (FUE) was not found in the *Ectocarpus* polyadenylation regions. Tools available at the RSAT website were used to search for novel conserved motifs. The sequence CACACG was detected in 35% of the 4422 polyadenylation regions. However its location relative to the polyadenylation site was too variable to directly link it to the polyadenylation process. The canonical AAUAAA element was also detected by this analysis (ranked 146th).

2.1.5. Whole genome tiling array analysis of gene expression

Identification and structural annotation of genes in the *Ectocarpus* genome was complicated by the elevated numbers of introns and by the absence of closely related genomes for comparative analyses. A whole genome tiling array approach was therefore used to complement cDNA sequence data and to provide a comprehensive transcriptome map. Supplementary Fig. 17a,b shows how the tiling array data was used to confirm exon structure, providing support for predicted gene structures. Within each gene the hybridisation signal decreased from the 3' to the 5' end of the gene because the cDNAs were labelled from 3' end. This phenomenon could be exploited to deduce strand information for a transcript, even though the hybridisations themselves were not strand-specific.

A total of 6,474 of the predicted genes exhibited a significant level of expression (2% confidence level) in the tiling array experiment (39.8% of the 16,256 predicted genes). For comparison, cDNA sequencing provided support for 6,710 genes (41.2%), including 4,374 of the genes detected as expressed using the tiling array. The overlap between these two sets of genes provides strong support for the expression of a large number of genes. Moreover, genes for which there were many EST sequences were usually found to have a high level of expression based on the tiling array data. However, it is also important to note that there were significant numbers of genes whose expression was detected by only one or the other of the two methods, illustrating the complimentarity of the two methods. Among the 100 genes with highest expression level according to the tiling array data, there were 16 that had no EST support. This shows that even relatively large-scale EST sequencing may not provide an exhaustive view of a transcriptome, and it may even fail to detect some highly expressed genes.

The tiling array data provided empirical support for 3,969 hypothetical genes and 1,832 conserved hypothetical genes. Of the hypothetical genes, 1,908 were not supported by any EST evidence and the tiling array data therefore represents the first empirical support that these are expressed genes. The tiling array provided evidence of expression for a greater proportion of the genes with homologues in other genomes (45%) than for the predicted genes with no such match (31%). This suggests that a proportion of the genes (about 31%) in the latter group may be false predictions. The exon-spanning probes (see supplementary methods section) were used by annotators to validate predicted gene structures.

In addition to providing information about the expression of the predicted genes, the tiling array also identified 8,741 expressed regions longer than 200 nucleotides that lie outside the predicted genes (Supplementary Fig. 17c). These regions represent potential novel protein-coding genes or non-coding RNA genes. Many of these expressed regions (2,315) are supported by EST data. One particularly interesting group of these novel expressed regions was found on sctg_0085, where they occur in a region containing a number of chlorophyll A-B binding protein genes. These expressed regions, which are also strongly supported by EST data, may therefore have a function related to that of the neighbouring chlorophyll A-B binding protein genes. When the 8,741 expressed regions longer than 200 nucleotides that lie outside the predicted genes were compared with the *T. pseudonana* genome only 34 regions exhibited significant similarity. This indicates that the vast majority of these transcribed regions have originated since divergence from the diatom lineage. To investigate further the origin of these additional expressed sequences, we determined whether they corresponded to repeated elements. Nearly half of the expressed regions (4224) overlapped with TEs and more than a quarter (2500) overlapped for more than a third of their length. This suggests that TEs have played an important role in the creation of the additional expressed regions.

An important influence of TEs was also observed when the 1000 most highly expressed regions (according to the tiling array data) were analysed. Eight hundred and fifty eight of these regions corresponded to predicted genes but 325 of these regions also included TE sequences and an additional 92 only matched TEs (the remaining 50 matched neither). TEs therefore appeared to make a significant contribution to the most highly expressed regions of the genome. Interestingly, when the fraction of the 1000 most highly expressed regions that corresponded to predicted genes were analysed, 139 of the 858 genes were members of a pair of convergently expressed loci on opposite strands with overlapping or nearly overlapping 3'UTRs. No functional information was available for the majority of these

genes but, in some cases, the two members of the pair had similar predicted functions (e.g. nitrate transport, light harvesting complex, transcription factor).

For some organisms, when genes have been divided into functional groups, marked differences in the percentage of genes that are expressed in each group have been observed (in the sea urchin embryo, for example, where a large proportion of the transcription factor and signalling categories are expressed compared to other functional groups[84]). No significant bias of this type was observed in *Ectocarpus*, although this may have been related to the complex mixture of cDNA samples that was used.

2.1.6. Alternative splicing

A total of 69.6% of the current *Ectocarpus* EST/cDNA collection unambiguously aligned to the genome using the stringent GENESEQER spliced alignment program (note that this percentage of matches differs from that obtained for the genome completeness calculation carried out in supplementary section 1.5. This difference is due to the fact that the matching carried out here used the raw cDNA sequences, without removal of poor quality reads, and because the matching method used by the GENESEQER program is more stringent). GENESEQER provides high quality alignments with overall similarity and coverage scores of at least 0.8. Local similarity scores 50 bp from both ends of each individual EST/cDNAs must also exceed 0.8. Hence, EST/cDNAs with overall similarity scores <0.8, and/or with local similarity scores 50 bp from either the 5' or 3' end that were <0.8, were excluded from the analysis. In addition for ESTs/cDNAs matching multiple genomic loci only the best match was used. Sequences that did not map to the genome were either derived from organellar (mitochondrial or chloroplast) genomes or were short, low-quality sequences. In total, 58,154 ESTs/cDNAs were mapped to the genome generating 86,105 cognate alignments; 48% of the ESTs had multiple cognate alignments. A total of 17,386 transcriptional units were identified, 91% correspond to annotated gene regions and 9% to unannotated regions. On average, 4.97 ESTs/cDNAs were found per *Ectocarpus* transcriptional unit.

Only a small proportion of the expressed genes in *Ectocarpus* are alternatively spliced. A total of 17,386 genes including 1,130 uncharacterized genes that were not included as annotated genes, were defined as "expressed genes" by GENESEQER. Of these, 472 protein coding sequences or 2.88% of the genes, exhibit a total of 772 alternative splicing (AS) events. The vast majority of the alternatively spliced genes in *Ectocarpus* (403 or 85%),

however, only exhibited a single AS event. Although, the incidence of alternatively spliced genes in *Ectocarpus* appears to be rare, this may have more to do with the depth of the *Ectocarpus* EST/cDNA set than the actual frequency or importance of AS in *Ectocarpus*. As larger and more comprehensively sampled tissue-specific *Ectocarpus* EST/cDNA collections become available the number of alternatively spliced genes identified is likely to increase.

As shown in Supplementary Table 8, *Ectocarpus,* unlike plants and animals, does not appear to have a preferred AS type. Of the five different alternative splicing types, while alternative acceptor is the most prevalent form, it is followed closely by intron retention and exon skipping in *Ectocarpus*. In land plants intron retention is the dominant type of AS while exon skipping is the most type prevalent in humans. Intron retention is probably less common in *Ectocarpus* than it is in higher plants because of the size of the former's introns, whose mean length is 674 nt compared to 171 and 438 for *Arabidopsis* and rice, respectively[85]. Intron retention is typically associated with short introns. The absence of a preferred AS type suggests that *Ectocarpus* may employ different of several modes of splice site recognition.

Approximately one-fifth of the observed AS events alter the reading frame generating an early stop codon and marking alternative mRNA isoforms as candidates for nonsense mediated decay. Alternative acceptor, alternative donor, and exon skipping events all display a similar 25-30% frequency of nonsense-mediated decay while a 10% frequency was observed for intron skipping and for alternative position. The overall incidence of nonsense-mediated decay observed in *Ectocarpus* is considerably lower than that observed in land plants, where more than one third of AS events may be coupled to nonsense mediated decay. This suggests that AS is not used to regulate message stability to the extent that it is in land plants, but rather may be more important in terms of affecting translation efficiency and/or increasing protein diversity in *Ectocarpus*.

While AS is predominant in some gene families in *Ectocarpus* and absent in others, with more than 75% of the alternatively spliced genes classified as unknown or conserved hypothetical proteins, it is difficult to determine whether AS might play a more significant role in some biological processes than others. Results from the limited number of functional analysis that have been performed suggest roles for AS in nucleotide binding, nucleic acid binding, hydrolase activity, transferase activity, protein binding, oxidoreductase activity, and ion binding, with more alternatively spliced genes (28) linked to nucleotide binding and hydrolase acivity (25) than to other categories (Supplementary Fig. 18). Alternatively spliced transcripts were distinctly absent for genes involved in processes related to sexual

reproduction, cell proliferation, gamete generation, reproductive development, cell cycle processes and cell death.

2.1.7. Subcellular localisation of *Ectocarpus* proteins

The Hectar algorithm[30] was used to predict the subcellular localisation of *Ectocarpus* proteins. This analysis predicted that 2097 proteins enter the secretory pathway (1643 bearing signal peptides and the rest possessing type II signal anchors) and that 445 and 615 proteins are targeted to the plastid and to mitochondria, respectively. Similar numbers of proteins were targeted to the plastid in *Ectocarpus* as in the two diatoms *T. pseudonana* and *P. tricornutum* (549 and 379 respectively) despite the fact that the *Ectocarpus* genome encodes about twice as many proteins.

Multicellular organisms might be expected to require larger numbers of extracellular proteins to mediate communication between cells[86]. We calculated the percentage of *Ectocarpus* proteins that enter the secretory pathway (proteins with a signal peptide or a type II signal anchor) based on an analysis carried out using the Hectar program and compared this with predictions for genomes from a broad range of unicellular and multicellular organisms. The results are summarised in Supplementary Fig. 19. This analysis indicated that, overall, multicellular organisms have a significantly higher percentage of proteins that are predicted to enter the secretory pathway than unicellular organisms (18.6% on average compared to 14.5% on average, respectively; P=0.038 in a student's t test). However, the values for individual genomes were very variable and several multicellular organisms, including *Ectocarpus* and *Physcomitrella patens*, were actually predicted to direct a smaller proportion of their proteins to the secretory pathway than the majority of the unicellular organisms analysed.

2.1.8. Domain analysis

To investigate the repertoire of biological and metabolic processes in *Ectocarpus*, we compared the abundance of individual protein domains in the genome with the corresponding abundances for a broad range of other eukaryotic species (Supplementary Table 9). The overall number of different domains represented in the *Ectocarpus* genome was comparable to the numbers obtained for the other species analysed but marked differences were observed in terms of the nature of the domains present and their abundances. Six protein domains were

significantly more abundant in *Ectocarpus* than in any of the other genomes analysed (indicated with an asterisk in Supplementary Table 9). These included domains that are predicted to be involved in carbohydrate binding (WSC domain IPR002889, see section 2.2.1.), photosynthesis (Chlorophyll A-B binding protein IPR001344, see section 2.2.2.) and G protein signalling (Regulator of G protein signalling superfamily IPR016137, see section 2.3.5.). Notch (IPR000800) and ankyrin (IPR002110 and PTHR18958) domain proteins were also significantly overrepresented. The abundance of the notch domain proteins is particularly interesting given the presence of this domain in proteins such as the notch receptor that are involved in intercellular communication in animals. Many of the notch domain containing proteins were predicted to be secreted or anchored in the membrane, consistent with possible roles in intercellular communication.

Interestingly, given the tendency for the proteins of multicellular organisms to possess a large proportion of small protein folds[87], many small domains are particularly abundant in the *Ectocarpus* genome, including the WW (IPR001202), FNIP (IPR008615), tetratricopeptide repeat (IPR011717) and NZF (SSF90213) domains, in addition to the notch and ankyrin domains mentioned above.

Comparison of the multicellular and unicellular organisms used for the protein domain analysis (see Supplementary Table 9) allowed the identification of eight domains that were consistently more abundant in the genomes of the former compared with those of the latter. These were the endonuclease/exonuclease/phosphatase (IPR005135), AMP-dependent synthetase and ligase (IPR000873), acetyl-CoA synthetase-like (SSF56801), carbohydrate kinase, FGGY (IPR000577), eukaryotic translation initiation factor 4 gamma (PTHR23253), lipoxygenase, C-terminal (IPR013819), UDP-glucuronosyl/UDP-glucosyltransferase (IPR002213) and peptidase T2, asparaginase 2 (IPR000246) domains. Moreover, none of the unicellular organisms possessed more NADPH oxidase domain (PTHR11972) genes than the multicellular organisms (see section 2.2.11.).

Several domains that are quite abundant (12 to 32 proteins) in the *Ectocarpus* genome were not found in any of the other stramenopile genomes analysed. These included the kinesin light chain (IPR002151), FNIP and transient receptor potential cation channel (PTHR13800) domains. Compared to other stramenopiles, the *Ectocarpus* genome also contains a large number of proteins with the tyrosine protein kinase domain (IPR001245).

2.1.9. Comparison of the complete set of *Ectocarpus* proteins with those of other genomes

Comparison of the 16,256 *Ectocarpus* proteins against the complete proteomes of two diatoms (*T. pseudonana* and *P. tricornutum*) and two *Phytophthora* species (*P. sojae* and *P. ramorum*) indicated that 6390 shared similarity (Blast, e-value < e^{-5}) with a diatom protein and 6317 shared similarity with an oomycete protein. Eight thousand and ninety five proteins did not match any protein from these stramenopiles.

To broaden the taxonomic distribution we also used Blast (e-value < e^{-5}) to compare the *Ectocarpus* proteome against the complete nr_prot database (NCBI) and grouped the hits based on the species from which they were derived. This analysis identified 4418, 7491, 7676 and 8282 proteins that shared similarity with a bacterial protein, an opisthokont protein, a green plant protein and a stramenopile protein, respectively. Overall, 6157 *Ectocarpus* proteins shared no similarity with proteins from any of these groups and 5822 proteins (35.8%) matched no proteins in the complete database (i.e. all species) at this stringency.

The highest number of matches was with the stramenopile genomes, as expected because *Ectocarpus* belongs to this group. The numbers of matches to opisthokont and green plant proteins was similar, reflecting the large phylogenetic distance that separates the stramenopiles from both of these groups. Supplementary Fig. 20 shows the overlaps between the groups of protein matches with bacteria, opistokonts, green plants and stramenopiles. The largest category in this diagram is the set of proteins shared with opistokonts, green plants and stramenopiles (3444 proteins). This group of proteins corresponds to a core set of proteins present in a wide range of eukaryotes.

2.1.10. Dollo analysis of gene family loss and gain during evolution

The Dollo parsimony principle[88] was used to map patterns of gain and loss of gene families onto a phylogenetic tree consisting of 17 organisms from a broad range of eukaryotic groups (see Fig. 3). The Gene Ontology (GO) vocabulary was used to link the loss and gain events with information about the molecular function or biological process the corresponding gene families were associated with[51]. Detailed information about this analysis can be accessed at http://bioinformatics.psb.ugent.be/dollo_analysis.

We were particularly interested in instances of gene family gain or loss that were predicted to have occurred since the divergence of the brown algal and diatom lineages (timepoint 4 in Fig. 3) because the transition to multicellularity presumably occurred during this time period (intermediate taxa are unicellular[89]). A marked enrichment for families with GO labels related to protein kinase activity was observed on this branch of the tree indicating that the evolution of novel kinase families has been one of the major evolutionary events since the separation of the two lineages (Supplementary Table 10, see also http://bioinformatics.psb.ugent.be/dollo_analysis for further details). The kinase families that were mapped to this branch of the tree by the Dollo analysis included a family of membrane-spanning receptor kinases (see section 2.3.2.) and several families of LRR kinases that are predicted to be targets of miRNAs (see section 2.1.14.).

A GO Slim[90] analysis was carried out to compare the pattern of gain and loss of gene families along the branch leading to *Ectocarpus* with the events that were predicted to have occurred on the other branches of the phylogenetic tree. This involved grouping the GO labels associated with gained and lost families under their corresponding GO Slim categories and calculating z-scores to identify differences between the patterns of loss and gain of gene families on each branch of the tree (Supplementary Fig. 21). The main conclusion from this analysis was that there was very little correlation between the patterns of gene loss and gene gain at each branch of the tree. In particular, no clear common trends were shared by the various branches of the tree that represent transitions to multicellularity (timepoints 4, 8, 16, 24 and 27). This may not be surprising, considering the large evolutionary distances separating these branches, but it does underline the independent nature of these events. One feature that did distinguish multicellular and unicellular organisms, however, concerned the number of lost and gained gene families. Calculated over their evolutionary histories, the muticellular organisms were predicted to have lost fewer gene families (1518 compared with 2131) and to have gained more gene families (4069 compared with 2436) on average than the unicellular lineages.

2.1.11. Gene family expansions

A comparative analysis of gene family expansions was carried out using the same 17 genomes that were analysed for gene family loss and gain. The analysis was carried out in two ways; the first approach focused on the 100 biggest gene families in *Ectocarpus* that were also

present in at least two other species. These were analysed to determine whether these families had expanded specifically in *Ectocarpus* or whether there were also correspondingly large families in the genomes of the other organisms analysed. To compare distributions and relative sizes of gene families, gene family sizes were converted into *z*-scores, which are expressed as standard deviations from their means (i.e., the mean copy number in all the species that possessed a particular gene family). Positive *z*-scores indicate gene copy numbers that are greater than the corresponding mean gene family size[88]. These *z*-score profiles were hierarchically clustered to allow comparison of the expansion patterns (Supplementary Fig. 22).

Of the 100 largest *Ectocarpus* gene families, 27 exhibited a greater degree of expansion in *Ectocarpus* than in any of the other genomes analysed. These families included genes with a broad range of predicted functions but genes involved in cytoskeleton and flagella function, in protein degradation and in protein phosphorylation and dephosphorylation were particularly well represented. Families that have only expanded in *Ectocarpus* include a glycosyl hydrolase family and a fucoxanthin chlorophyll a/c binding protein subfamily. Two additional fucoxanthin chlorophyll a/c binding protein subfamilies have expanded only in *Ectocarpus* and in one or both diatoms. Some examples of families that have only expanded in *Ectocarpus* and the *Phytophthora* species are a ubiquinol cytochrome c oxidoreductase biogenesis factor family, a sulphate transporter family protein and a catalase/peroxidae family.

The pattern of gene family expansions in the lineage leading to *Ectocarpus* has been markedly different to those leading to other complex multicellular lineages, underlining the independence of the evolutionary processes in each multicellular lineage (Supplementary Fig. 22).

We also searched for families that were significantly more expanded in *Ectocarpus* than in any of the other organisms analysed (families with a z-score greater than two), irrespective of gene family size (Supplementary Fig. 23). Again, this analysis identified families with a broad range of predicted functions, including families involved in transcription, DNA replication, proteolysis and protein modification.

2.1.12. Clusters of genes with similar functions

A gene-ontology-based search for clusters of genes with similar functions was carried out using the C-Hunter program (Supplementary Table 11). The *Ectocarpus* genome does not exhibit unusual levels of clustering of genes of related function compared to previously sequenced genomes. However, it is interesting to note that many of the clusters identified involved genes predicted to be involved in chromatin-related functions (e.g. chromosome organization and biogenesis, nucleosome, chromatin assembly).

2.1.13. Transposons, repeat sequences and telomeres

Transposable elements (TEs) account for a large proportion of many eukaryotic genomes. We searched for TEs in the *Ectocarpus* genome using a previously described transposable element annotation pipeline[14] (and see Supplementary methods). This procedure, followed by manual curation, allowed the establishment of *Ectocarpus* reference sequences for 612 unclassified repeats and 135 known TE families. These included Ty1/copia, Ty3/gypsy, and DIRS/Ngaro-like LTR-retrotransposons, non-LTR retrotransposons, as well as subclass I DNA transposons such as Harbinger, JERKY, and POGO-like elements, and subclass II elements (Helitrons) (Supplementary Table 12). The TEs identified also included a new family, which is similar to large retrotransposon derivatives (LARDs) in that extremities consist of long terminal repeats flanking a large internal domain but which has an unusual internal sequence containing an open reading frame encoding a protein of ~400 amino acids. This protein is predicted to contain a zinc finger domain at the C-terminus and to share weak similarity with GAG proteins from plant Ty1/copia-like elements. It is unknown whether the protein enables these large GAG-related elements (LGAs) to transpose autonomously. The *Ectocarpus* repeat complement also includes tandem repeats (minisatellites).

Repeated sequences constitute about 45 Mbp (22.7%) of the *Ectocarpus* genome. This is comparable to previously sequenced genomes of a similar size such as *Arabidopsis* (17% repeated sequences) and *Drosophila* (20%). The repeated sequence fraction in *Ectocarpus* consists of 41.3% class I elements and 12.3% class II elements (Supplementary Fig. 24). Among the class I elements, LTR retrotransposons are the most abundant (26% of the repeated component of the genome). We failed to detect SINEs in the genome although this observation should be treated with caution because SINE elements have few conserved features and are therefore difficult to detect. However, the rarity of LINEs in the genome may

account for the absence of SINEs. The most abundant individual repeat sequence found in the *Ectocarpus* genome was an unclassified 676 nucleotide repeat (denoted as Sower) with a 5' end consisting of 1, 2 or 3 (full length) repeats of a ~67bp motif that begins with a stretch of cytosines and another stretch of cytosines at the 3' end of the element (Supplementary Fig. 25a). Analysis of several copies of Sower failed to identify any target site duplication or any conserved open reading frame in the element. RNA secondary structure prediction for Sower indicated a stable (free energy of -387 kcal/mol) rod-like shape branched with numerous hairpins (Supplementary Fig. 25b).

TEs were found in both the intergenic regions and the within genes. TE distribution within different gene features was as follows: 18.9% in exons, 41.3% in introns, 4.0% in 5'UTRs and 4.8% in 3'UTRs. This corresponded to a significant overabundance of TEs in exons, introns and 3'UTRs ($P<2.2e^{-16}$). At the supercontig scale, the distribution of TEs appeared to be quite homogeneous (Supplementary Fig. 26). Also, when different pseudochromosomes were compared, the TE density was very similar, although pseudochromosomes 30 and 32 appeared to be particularly TE-rich and gene-poor (Supplementary Figs. 3, 4). See section 2.1.1. for further details.

Supplementary Fig. 27 shows an estimation of the divergence times between copies of several categories of TEs. Interestingly, this analysis indicated two peaks of TE activity at 0.02 and 0.15, involving a broad range of TE categories. These peaks suggest the occurrence of two events in the evolutionary past that led to the activation of diverse families of TE elements.

The cDNA sequence data obtained using *Ectocarpus* material grown under normal laboratory conditions were used to assess the levels of expression of the 98 best characterised (fully classified) TE families under non-stress conditions. This data set contains high numbers of ESTs (between 97 and 250 sequences) for five TE families (EsCopia1, EsCopia5, EsCopia6, EsCopia10 and EsLGA6) and a calculation that took into account the genome coverage of each family indicated that the members of these five families were expressed at higher levels, on average, than the members of the other 93 TE families (see Supplementary Fig. 28, Supplementary Table 13). In addition, four of these five families were represented among the 100 loci that exhibited the highest expression levels according to the whole genome tiling array analysis (Supplementary section 2.1.5) and all five families are predicted to have been involved in the most recent peak of TE activity described in the previous paragraph. When information concerning the estimated age, level of expression and sequence

conservation was combined, these five families appeared to be the best candidates for containing actively transposing elements. Note that four of these families are Ty1/Copia-like elements.

The relatively high level of expression of these five TE families is interesting because TEs are normally maintained silent by epigenetic processes under non-stress conditions in other organisms. Trancriptional activation of TEs in response to specific stresses has been described in several systems[91] but trancriptional activity of TEs under non-stress conditions is rare, usually occurring only at a particular stage of the life cycle such as expression of the Ty1 element during the haploid phase in yeast[92] or germ line-specific expression of the intracisternal A-particle in mice[93].

Epigenetic silencing of TEs commonly involves DNA methylation but TE sequences in the *Ectocarpus* genome appear to be free of such marks (Supplementary Fig. 29). This data is consistent with the apparent lack of DNA methyltransferases in the *Ectocarpus* genome. It thus appears that, as has been observed in a small number of unikont species such as *Caenorhabditis elegans*, *Ectocarpus* lacks this component of the TE silencing machinery. Consequently, the study of the epigenetic mechanisms involved in regulating chromatin compaction and TE silencing in *Ectocarpus* could be of particular interest. Note also that the small RNAs encoded by the *Ectocarpus* genome were found to be preferentially derived from TEs, suggesting the presence of a small-RNA-mediated TE-silencing mechanism that does not involve DNA methylation (see section 2.1.14).

A putative telomere region with 22 repeats of the sequence TTAGGG was found at the start of scaffold 562. Four additional regions with between 12 and 26 TTAGGG repeats were found in four small contigs that were below the 2 kbp size limit set for inclusion in the final set of genome scaffolds. The telomere sequences in *Ectocarpus* therefore resemble those of most other chromalveolates and of more distant species in the fungi and metazoa. However, as the esimated number of chromosomes is approximately 25[73,74], most of the telomeric regions do not appear to have been included in the genome assembly.

2.1.14. Small RNAs and RNAi

The endogenous small RNA repertoire of *Ectocarpus* was analysed by deep sequencing of two libraries built, respectively, from sporophyte and gametophyte RNA. A total of 7,114,682 reads were obtained for the two libraries (Supplementary Table 14). These reads

corresponded to 2,699,843 unique sequences, of which only the 1,262,711 that corresponded to full-length small RNA sequences were retained. Based on this set of sequences, 875,490 reads were mapped onto the *Ectocarpus* genome. Further filtering was carried out to retain only sequences that had been sequenced ≥ 5 times. This produced a reduced set of 63,511 reads, which represented 24,132 unique small RNA sequences and which mapped to 1,031,522 loci in the genome.

The expressed small RNA sequences were annotated based on their overlap with various structural elements of the genome (introns, exons, rRNAs, tRNAs, snoRNAs, transposable elements, intergenic regions). The most abundant source of small RNAs was rRNA sequences (45%), followed by intergenic regions (26%), transposons (13%) and introns (9%) (Supplementary Table 15, Supplementary Fig. 30). Analysis of the sizes of the small RNAs that originated from intergenic regions, transposons, introns and exons showed that, in each case, the largest size group was 21 nt. No such bias was observed for sequences originating from rRNA and tRNA (Supplementary Fig. 31). The clear peak of 21 nt small RNAs is likely to reflect functional selection whilst the uniform size distribution for small RNAs associated with tRNAs and rRNAs indicates that they are more likely to be derived from the degradation of longer transcripts. The small RNAs associated with rRNAs were expressed at the lowest levels, followed by the TE-associated small RNAs, while the other small RNAs tended to be more strongly expressed, with similar distributions of expression levels (Supplementary Fig. 32). There was a significant difference between the normalized expression levels associated with the different genomic location categories (one-way ANOVA df=5, f-value=6375, p-value < 0.001). Only 9,919 (41 %) of the 24,132 unique small RNAs were detected in both the gametophyte and the sporophyte libraries, indicating that the small RNA coverage may not yet be complete.

Transposable elements were one of the major sources of small RNAs in *Ectocarpus* (Supplementary Figs. 30, 33). We found a statistically significant link (z-score 13.9, p-value < 0.0001) between small RNAs and transposable elements in *Ectocarpus*, compared to random genome segments. Small RNAs have been associated with transposable elements and repeat silencing in many eukaryotic organisms[94], and therefore the small RNAs associated with TEs in *Ectocarpus* are likely to fulfil the same role. Most small RNAs that mediate silencing of transposable elements trigger DNA methylation[94] but, as no evidence was found of methylation of *Ectocarpus* DNA, small RNA mediated silencing could be achieved by other means in this species, such as histone modifications.

MicroRNAs (miRNAs) are a distinct class of small RNAs present in most eukaryotic organisms. They arise from typical imperfect stem-loop structures, are processed by Dicer enzymes and the mature 21-22 nt long sequence is incorporated into a silencing complex where they are associated with argonaute proteins. miRNAs have crucial roles in several cellular and developmental pathways in plants and animals[95,96]. There is evidence that an RNA interference system exists in *Ectocarpus*. The genome contains one Dicer, one Argonaute and two RNA-dependent RNA polymerase homologues (Supplementary Table 16), although no matches were found with many other genes associated with RNAi in other species, including *SGS3, HEN1* and *PolV1*. The RNA interference system therefore probably differs significantly from those found in green plants and animals, as would be expected given the phylogenetic distance separating *Ectocarpus* from these groups.

Using a set of stringent rules combining small RNA expression and structural features inspired by recent progress in the definition of miRNAs[42], 26 miRNA sequences were identified in *Ectocarpus*, defining 21 families (Supplementary Fig. 2, Supplementary Table 17). Most of the mature miRNA sequences are 21 nt long and begin with a U (24 sequences), the latter being a known preference of the plant Argonaute-1 protein (AGO1)[97]. Seventeen miRNA sequences are located in introns, eight are in intergenic regions, and one is located antisense to a transposable element (Supplementary Table 17). Intronic miRNAs are frequently found in animal genomes, but not in plants. A cluster of seven miRNA sequences was found in a 5 kbp region of one intron (Supplementary Table 17). Using a computational screen designed for plant miRNAs and a conservative cutoff score, we found 71 candidate target sequences for 12 of the 26 miRNAs. Interestingly, 53 of these target sequences (75%) contain leucine rich repeat (LRR) domains (Supplementary Table 18). Nine of the LRR targets are members of the ROCO family (LRR-GTPase). Altogether, 21 LRR-ROCO genes have been identified in *Ectocarpus* (excluding pseudogenes), so 43% of the members of the family are targets of miRNAs. The binding sites of some of these miRNAs correspond to the LRR domain (Supplementary Fig. 33).

Ectocarpus belongs to one of the major eukaryotic lineages (stramenopiles) that have diverged from the other lineages leading to green plants and animals several hundreds of million years ago[98]. A diverse small RNA population, including a few miRNAs, has been described in the social amoeba *Dictyostelium*[99], which also belongs to another ancient eukaryotic lineage, distinct from the stramenopiles. Combined with our results this suggests

that small RNAs, and more particularly miRNAs, were probably present from an early stage in eukaryotic evolution.

2.1.15. Endosymbiosis

Chromalveolate plastids are believed to have originated from a unicellular red alga that was captured by an ancestral host cell, the process of endosymbiosis involving a large-scale transfer of genes from the endosymbiont to the host nucleus. This process would have resulted in the transfer of a broad range of functions that included, but were certainly not limited to, photosynthetic roles[27,100]. The ancestral host cell mentioned above was originally thought to be non-photosynthetic but a recent analysis indicates that diatom genomes contain a large number of genes that are phylogenetically related to the green lineage, suggesting that this ancestral cell may have been already photosynthetic, containing a green-alga-derived plastid that was subsequently replaced by the red-alga-derived plastid[101].

Two independent phylogenomic analyses were carried out, using the neighbour joining and maximum likelihood methods respectively, to identify *Ectocarpus* genes that were predicted to have been derived either from a red algal ("red" genes) or a green algal ("green" genes) endosymbiont. The two methods detected 611 and 674 "red" genes, respectively, with 386 genes being detected by both approaches (Supplementary Table 43). As far as the "green" genes were concerned, the neighbour joining and maximum likelihood methods detected 2669 and 3987 genes, respectively, 2176 genes being common to both data sets (Supplementary Table 44). The following analyses were based on the "red" and "green" genes identified by the neighbour joining method because the use of this data set allowed us to make comparisons with similar analyses that had been carried out for other chromalveolates[101 and this study]. In general, however, the conclusions still held when the analysis was restricted to the more stringent data sets of "red" and "green" genes identified by both tree building algorithms.

The neighbour-joining-based phylogenomic analysis identified more "green" than "red" genes in the *Ectocarpus* genome (2669 of the former compared with 611 of the latter; 17.92% and 3.65% of the genome, respectively). A similar result was obtained when diatom genomes were analysed[101]. Homologues of both the "red" and "green" genes were found in the genomes of other chromalveolates (Supplementary Fig. 35). The number of "red" genes shared with *Ectocarpus* decreased when comparisons were made with diatoms, pelagophytes and haptophytes, respectively. To determine whether this pattern reflected the phylogenetic

relationship between these organisms, we constructed a phylogenetic tree for these groups using 12 concatenated "red" gene protein sequences (see section 2.1.16.). The tree (Supplementary Fig. 36) indicated that diatoms were most closely related to *Ectocarpus*, followed by pelagophytes and then haptophytes, concurring with the relative numbers of shared "red" genes. Oomycetes share the lowest number of red genes with *Ectocarpus,* probably because the former are not photosynthetic and thus have lost photosynthesis-related "red" genes from their genomes. Surprisingly though, the oomycete genomes share more "green" genes with *Ectocarpus* than do the genomes of the other stramenopiles analysed. This may be due, in part, to the fact that *Ectocarpus* and the oomycetes have larger genomes and therefore possess more genes in general, but it is also possible that the retention of these genes in the two lineages is linked to common features such as multicellular bodyplans (see below). Note that a similar phenomenon was not observed for *T. pseudonana*, which shares similar numbers of "green" genes with the oomycetes and *E. huxleyi*.

In terms of predicted functions, two of the "red" genes, which encode a glutamate/ornithine acetyltransferase (ISS) and an acetylornithine aminotransferase, are of particular interest because they could be involved in the urea cycle. This suggests that the urea cycle in brown algae is partly derived from the red algal lineage. Interestingly, *Ectocarpus* shares more "green" genes with oomycetes (non-photosynthetic chromalveolates which exhibit a simple form of multicellularity) than with any photosynthetic chromalveolate (425 versus 232-328). Analysis of the predicted functions of the "green" genes shared between oomyctes and *Ectocarpus* indicated that they included many protein kinases (28 genes), transporters (15), and flagellar proteins (14). The retention of the kinases and transporters may have played a role in the evolution of multicellularity in these two groups, allowing more complex signalling and transport processes. It is also interesting that "green" genes are predicted to contribute towards the construction of the flagellar apparatus in brown algae and oomycetes, indicating that this may have a chimeric structure in terms of its evolutionary origins.

The inclusion of cyanobacterial sequences in clades corresponding to "red" or "green" genes was interpreted as an indication that these genes have a role in plastid function. The numbers of *Ectocarpus* "red" and "green" genes that grouped with cyanobacterial sequences were similar to those observed for the diatom *T. pseudonana* (71 "red" and 84 "green" for *Ectocarpus* compared with 62 "red" and 72 "green" for the diatom), with “green” genes slightly outnumbering “red” genes. This is in accordance with the situation in diatoms which

possess in their nuclear genomes more plastid genes of green than of red origin[101]. The *Ectocarpus* "green" genes with plastid functions were predicted to include enzymes with roles in nitrate assimilation, sulfate assimilation and pigment synthesis.

None of the "red" genes that were shared by *Ectocarpus* and oomycetes were predicted to be plastidial. This would seem logical, as the oomycetes are not photosynthetic. However, the 365 "green" genes that were shared between oomycetes and *Ectocarpus* included 11 (3%) putative plastid proteins (compared with 143 identified by manual annotation, or 5.4%, of the total number of 2669 green genes in *Ectocarpus)*. None of these proteins are predicted to carry out functions in photosynthesis, but are rather associated with other metabolic processes. The presence of these genes suggests that some plastid metabolic functions have been retained in oomycetes and it will be interesting to ascertain how these proteins function and in which compartment. The "red" and "green" genes also contribute significantly to the pool of mitochondrial proteins (34 of the 611 "red" genes and 93 of the 2669 "green" genes, of a total of 658 predicted mitochondrial proteins in the genome). The brown algal mitochondrion therefore appears to be a chimeric structure in terms of the evolutionary origins of the proteins that function in this compartment, and this may have had a significant impact on the efficiency of this organelle.

2.1.16. Phylogenetic relationships among photosynthetic stramenopiles

In the past it has been difficult to clearly resolve the phylogenetic relationships among photosynthetic stramenopiles. One explanation could be that, depending on the genes selected, phylogenies could have been based on a mixture of "red", "green" (see section 2.1.15.) or other genes from the nuclear genomes. This problem could be overcome by selecting genes with a common origin in stramenopiles, i.e. genes of red algal origin. These likely entered stramenopiles through a single secondary endosymbiosis and thus are monophyletic in this taxon. We selected 12 genes present in all photosynthetic stramenopiles sequenced to date and calculated phylogenetic trees based on a concatenated alignment of the 12 genes (12,000 positions). Protein sequences were aligned using ClustalW[102], and phylogenetic tree was constructed using PhyML[103] (WAG protein model, gamma distribution describing among-site rate heterogeneity and SH-like approximate likelihood ratio test branch supports). The tree separates all taxa well with the earliest emergence of *Emiliana huxleyi*

(haptophyta) followed be *Aureococcus* (Pelagophycea), *Ectocarpus* and diatoms (Supplementary Fig. 36).

2.1.17. EsV-1 virus

E. siliculosus virus-1 (EsV-1) is a phaeovirus of the Phycodnavirus family. Phycodnaviruses are characterised by icosahedral morphology, internal lipid membranes and large double-stranded DNA genomes[104]. Phaeoviruses have been found to be pandemic in several brown algal species[104]. Like other phaeoviruses, EsV-1 only infects free-swimming (wall-less) gametes or spores. The viral DNA integrates into the cellular genome following infection and is then transmitted via mitosis to all the cells of the developing host[105-107]. The virus remains latent in vegetative cells and viral particles are only produced in the reproductive organs, the sporangia and gametangia, following a stimulus such as a change in light, sea water composition and temperature[104,108]. Infected algae show no apparent growth or developmental defects other than partial or total inhibition of reproduction. Viral DNA has been detected in the genomes of several species of *Ectocarpus* and *Feldmannia*[109,110,111]. Analysis of the *Ectocarpus* genome sequence identified the presence of a viral genome closely related to EsV-1. The viral genome is present as a single copy and most of the viral sequence is present as a single fragment on supercontig 0052 (Fig. 2). This is consistent with previous observations indicating that EsV-1 integrates at a single locus[105-107]. The integrated viral genome extends for at least 310,438 bp (from Esi0052_0297 to Esi0052_0174), has a GC content of 51.% (EsV-1 is 313,838 bp in length and is 51% GC[112]; and contains orthologues of 173 of the 231 EsV-1 genes. Comparison with EsV-1 indicated that the virus probably integrated as a circle (the "terminal" repeats ITRA and ITRA’ are adjacent), with recombination occurring in the region corresponding to the putative integrase gene (i.e. EsV-1-213 in EsV-1). Restriction digestion and electron microscopy initially indicated that EsV-1 was circular but the sequence of EsV-1 indicated a linear molecule with terminal repeats[113]. Analysis of the genome sequence has therefore confirmed that the virus can exist in circular form. It was not possible to identify precisely the DNA sequences involved in integration due to sequence divergence in this region of the viral genome compared to EsV-1. Moreover, a short region corresponding to the putative integrase gene EsV-1-213 is missing from this region of supercontig 0052. A second fragment of viral DNA of about 3400 bp that includes a sequence homologous to EsV-1-213 was found on supercontig 0371, suggesting that this region may have been transposed, either during integration or in a subsequent rearrangement

event (Fig. 2). The genetic map located supercontig 0371 on a different linkage group to supercontig 0052. Surprisingly, on supercontig 0052, the original integrase gene has been replaced by another integrase sequence which shares only 70% identity at a DNA level with the corresponding EsV-1 gene (whereas the putative integrase sequence on supercontig 0371 shares 97 % identity). Additional remnants of the putative integrase gene were found in supercontigs 0108, 0150, 0540 and 0545 and other additional, short viral regions were detected in supercontigs 0007, 0108, 0150, 0490, 0540 and 0545 (Supplementary Table 19). Most of these regions contain fragments of genes with one or several frame shifts but some of the viral genes are potentially functional (e.g. Esi0490_0016, which codes for a small subunit of the Replication Factor C, see also Supplementary Table 19).

Overall, gene order in the integrated viral sequence is conserved compared to EsV-1 but there are several insertions and deletions of genes and these appear to have occurred preferentially in regions of lower gene density (Supplementary Fig. 37, Supplementary Table 19). The majority of the inserted or deleted ORFs have no homology with known genes although FirrV-1 ORFs and ankyrin coding sequences were identified among the inserted genes, and the deleted genes included histidine protein kinases (Supplementary Table 19). The presence of FirrV-1 coding sequences suggests that phaeoviral genomes may represent mosaics of genes that were either originally shared by a common ancestor[114] or have been shuffled between genomes since the divergence of the viruses. A few genes coding for transposases are not located in the same positions relative to EsV-1[113]. The genes in EsV-1 do not contain introns[113] but a single intron was found in one of the EsV-1 gene homologues in the inserted viral sequence(Supplementary Fig. 37, Supplementary Table 19). This gene shares low similarity with its EsV-1 counterpart (EsV-1-164) and is oriented in the opposite direction, suggesting that the original viral gene may have been replaced by an algal gene. Esi0052_0151 belongs to a multigene family that encodes proteins with a discoidin domain (FA5/8 type C domain) and a nosD copper-binding region. Despite these differences between the integrated virus and EsV-1, there are also indications that at least part of the genome of the former has been maintained in a functional state. Many of the genes that encode proteins that are important for the life cycle of EsV-1 and FirrV-1 are located in two regions of the viral genomes[114] (Fig. 2a). The orthologues of these genes in the integrated sequence have not been affected by deletions or mutation and are potential functional. Moreover, all the NCLDV-core genes found in EsV-1 are also present in the inserted virus (Supplementary Table 19). Remarkably, distances between genes also appear to have been conserved within

the two conserved regions (Fig. 2a). Also, the ITRA and ITRA'-like repeat sequences in the genome are similar in structure and sequence to those of EsV-1. Nonetheless, *Ectocarpus* Ec 32 has not been observed to produce virions indicating that the integrated virus is unable to complete its life cycle. This may be due to modification at the putative integrase locus or to the absence of one or more of the other key genes in the genome. For example, the integrated sequence lacks three EsV-1 genes that encode histidine kinases. None of the 83,502 EST sequences correspond to the viral genes on supercontig 0052, suggesting that these genes are not transcribed (although three ESTs were identified for the putative integrase gene on supercontig 371). Analysis of microarray data[79] confirmed that the viral genes are transcriptionally silent (Fig. 2b). Only one of the 149 genes assayed exhibited an expression level above the background baseline defined by multiple, random-sequence control probes, and even this gene exhibited only a very weak level of expression. No significant increase in transcript abundance was observed for any of these genes following the application of several stress treatments (hyperosmotic, hypoosmotic and oxidative stress), nor when fertile gametophytes (carrying gametangia where viral particles are normally produced in infected strains) were analysed (data not shown). The transcriptional silence of the viral genes may be due to the absence of factors necessary for the transcription of these genes or, alternatively, could be due to a silencing mechanism imposed either by the virus, as part of the lysogenic process, or by the host. We consider that the latter hypothesis is the least likely of the three because we did not observe a preferential association of small RNAs with the viral sequence compared to other regions of the genome, as was observed for transposon-rich domains of the genome which are probably subjected to host-mediated transcriptional repression. One hundred and thirty one small RNAs matched a set of 10,000 virus segments compared with an average of 248 ± 51 for sets of 10,000 random genomic segments (giving a z-score of -2.29, corresponding to a p-value < 0.05).

Three viral DNA-containing segments of the genome of an *Ectocarpus* strain that produces EsV-1 particles (NZVic14) have been sequenced[115]. These segments contained complete and mutated viral genes related not only to EsV-1 but also to mimivirus[116]. One segment included a large, viral Origin Binding Protein (OBP) gene, a sequence that is only found in mimiviral, asfarviral and herpesviral genomes. OBP plays an important role in DNA replication of herpes viruses[117]. The *Ectocarpus* Ec 32 genome contains no regions that are completely colinear with these three segments, although several regions that share sequence identity were detected in the genome. In particular, two regions on two independent

supercontigs (positions 72951 to 76275 on supercontig 0448 and positions 30951 to 32058 on supercontig 0524) share high identity with part of one of the NZVic14 segments (segment C, position 40997 to 46187), which codes for a large viral Origin binding protein (OBP). However, neither of these supercontigs contains a complete OBP coding sequence.

The presence of these OBP coding sequences suggests that *Ectocarpus* strain Ec 32 was infected by other giant dsDNA viruses, in addition to the virus found on supercontig 0052. However, we cannot also exclude the possibility that these sequences have been transposed from the viral sequence on supercontig 0052, particularly as these sequences have similar GC contents. Note that the ancestor of phaeoviruses and NCLDVs is thought to have possessed OBP coding sequences[115].

In general, apart from the inserted viral genome, the *Ectocarpus* genome contains very few genes that are predicted to be of viral origin. This is unexpected considering the pandemic levels of virus infection seen in the field[118,119] and suggests that there is an effective barrier to gene acquisition via this route.

2.1.18. Organellar genomes

Complete sequences of both the plastid and mitochondrial genomes were obtained. The circular plastid genome (139,954 bp; EMBL acc. # FP102296) and its use as a source of genes for phylogenetic analyses based on concatenated data sets has been described elsewhere[120]. The mitochondrial genome of *Ectocarpus* (accession number FP885846) is a circular molecule of 37,189 bp (33.5 % GC content), which is highly similar to other sequenced phaeophyte mitochondrial genomes[121,122] both in terms of gene content and organisation.

2.2. Metabolism

2.2.1. Carbon storage and cell wall metabolism

Carbohydrate metabolism is one of the main traits that distinguish brown algae from other phyla. In contrast to most living cells which contain intracellular storage α-glucans (glycogen or starch), brown algae store carbon in vacuoles as laminarin, a medium molecular weight β-

1,3-glucan with occasional β-1,6-linked branches[123]. This is a feature that brown algae share with diatoms and oomycetes, which store glucans as the related molecules chrysolaminarin and mycolaminarin, respectively. Another unusual feature of brown algae is that the photoassimilate D-fructose 6-phosphate (F6P) is not used to produce sucrose as in flowering plants but is mainly converted into the alcohol sugar D-mannitol. It has been proposed that laminarin and mannitol are interchangeable storage substances, fulfilling the same functions as sucrose and starch in flowering plants[124]. Brown algal cell walls are also composed of complex polysaccharides. As in land plants, some neutral polysaccharides such as cellulose are synthesised, but brown algae also produce unique anionic polysaccharides: alginates and sulphated fucans[125].

Mannitol is synthesized in two steps in brown algae: F6P is reduced by mannitol-1-phosphate 5-dehydrogenase (M1PDH, EC 1.1.1.17) to produce mannitol-1P, which is then converted into mannitol by mannitol-1-phosphatase (M1Pase, EC 3.1.2.22). These enzymatic activities have been isolated from several brown algae[126], but none of the corresponding genes have been cloned. Recycling of mannitol is thought to involve mannitol 2-dehydrogenase (M2DH, EC 1.1.1.67) and hexokinase (EC 1.1.1.67)[127]. Three *Ectocarpus* proteins (Esi0017_0062, Esi0020_0181 and Esi0080_0017) share significant sequence identity (~30%) with the M1PDH from the apicomplexa *Eimeria tenella*. The only known M1Pase gene was also cloned from this parasite[128], but no homologue has been identified in *Ectocarpus*. A mannitol recycling M2DH gene was identified (Esi0135_0010), but surprisingly we did not found any homologue of eukaryotic hexokinases. Instead, *Ectocarpus sus* possesses a specific kinase (Esi0139_0025), closely related to fructokinase from cyanobacteria (EC 2.7.1.4). Thus, brown algae seem to be an exception to the tendency for multicellular eukaryotes to possess broad specificity hexokinases whereas bacteria and unicellular eukaryotes possess distinctive sugar-specific kinases[129].

The key enzymes for the synthesis and remodelling of oligo- and polysaccharides are glycoside hydrolases (GH) and glycosyltransferases (GT), which are classified into more than 200 Carbohydrate Active enZYme (CAZY) families http://www.cazy.org/, [130]. We have identified 41 glycoside hydrolases and 88 glycosyltransferases in the *Ectocarpus* genome. These belong to 18 GH and 32 GT families, respectively (Supplementary Table 20). *Ectocarpus* has slightly more GH/GT genes than the green microalgae *Micromonas sp.* and *Ostreococcus tauri*[131], but about 6 times less than land plants[132]. However, the CAZY component of this brown algal genome more closely resembles those of land plants, there

being 34 GH and 40 GT families in *Arabidopsis thaliana* for example. *Arabidopsis* possesses some highly expanded gene families (e.g. family GH28: 69 genes, family GT1: 121 genes)[132] but *Ectocarpus* has less functional redundancy with only a few genes in each CAZY family. Interestingly, this brown alga contains some families that are absent from plant genomes but present in other phyla such as bacteria (family GH88), fungi and animals (family GH30, glucosylceramidase), only animals (families GT23, GT49 and GT54, protein N-glycosylation) and Amoebozoa (families GT60 and GT74, O-glycosylation of Skp1 subunits of E3 ubiquitin-protein ligase).

Sucrose metabolism appears to be completely absent from *Ectocarpus*, as deduced from the lack of sucrose synthase and sucrose phosphate synthase (family GT4) and of invertases (families GH32 and GH100). Similarly, the enzymes involved in starch biosynthesis (families GT5 and GH13) and degradation (families GT35, GH13, GH14 and GH77) are missing. Also, no ADP-glucose pyrophoroylase was found, confirming the absence of starch metabolism in brown algae. In contrast, *Ectocarpus* possesses a complete trehalose pathway. Trehalose is synthesised by a family of six bifunctional enzymes encompassing a trehalose-phosphate synthase (EC 2.4.1.15, family GT20) fused to a trehalose phosphatase (EC 3.1.3.12), while the recycling of trehalose is assured by a single trehalase (EC 3.2.1.28, family GH37). The exact role of trehalose in brown algae is unknown, but this non-reducing disaccharide is a central metabolic regulator in plants[133].

The laminarin biosynthetic pathway is essentially unknown but we identified several genes that are likely to be involved in this metabolism. *Ectocarpus* contains two cytosolic isoforms of UDP-glucose pyrophoroylase (UGP, Esi0144_0004 and Esi0430_0005), supporting the hypothesis that UDP-glucose is the activated sugar needed for the production of laminarin. Interestingly, Esi0430_0005 encodes a modular protein in which the UGP domain is fused to a phosphoglucomutase. This bifunctional enzyme is predicted to catalyse two consecutive reactions (Glc6P ↔ Glc1P; UTP + Glc1P ↔ PPi + UDP-glucose), suggesting that these metabolic steps may be more efficient in brown algae. Three β-1,3-glucan synthases from the GT48 family were identified in *Ectocarpus* (Esi0033_0138, Esi0338_0032 and Esi0193_0029). These integral membrane proteins, which are closely related to plant callose synthases, probably polymerise the laminarin backbone. Moreover, *Ectocarpus* has two proteins (Esi0100_0034 and Esi0243_0020) that are homologous to KRE6, a GH16 family transglycosylase involved in the biosynthesis of cell wall β-1,6-glucans in yeast [134]. Therefore, these two proteins are good candidates for the synthesis of β-

1,6-linked branches of laminarin. The degradation of laminarin is potentially catalysed by ten endo-1,3-beta-glucanases (EC 3.2.1.39) belonging to three different families (GH16: 4 genes; GH17: 1 gene; GH81: 5 genes) and two exo-1,3-beta-glucanases (EC 3.2.1.58, family GH5). These numerous laminarinases show similarity to bacterial (family GH16), fungal (families GH5 and GH81) and plant (family GH17) genes, underlining the complexity of laminarin metabolism in brown algae. Laminarin oligosaccharides are likely to be further hydrolysed by three β-glucosidases from the GH1 family (Esi0061_0010, Esi0176_0045 and Esi0212_0019). The released glucose would then be subsequently phosphorylated by a glucokinase (Esi0000_0270) to enter glycolysis.

Little is known about cell wall metabolism in brown algae. The genome of *Ectocarpus* thus provides the first broad data resource allowing the prediction of some aspects of this crucial metabolism. We have identified nine cellulose synthase-like proteins (family GT2). Like β-1,3-glucan synthases, these integral membrane glycolystransferases are predicted to use UDP-glucose as the activated sugar. However, we found neither cellulases (despite the existence of twelve CAZY families of cellulases in other species) nor expansins, which are cell-wall loosening proteins in plants. Therefore, cellulose biosynthesis in brown algae seems to be similar to that of land plants, but the remodelling of cellulose fibres probably involves novel, unidentified families of cellulases.

The alginate biosynthetic pathway has been characterised in some pathogenic bacteria which secrete alginate as an exopolysaccharide[135]. In brown algae, however, only the final step of alginate biosynthesis, the epimerisation of β-1,4-D-mannuronic acid to α-1,4-L-guluronic acid, has been characterized in brown algae. This structural change is catalysed by a multigene family of mannuronan C-5-epimerases in *Laminaria digitata*[136]. Very few of the 13 bacterial proteins involved in alginate biosynthesis[135] are conserved in *Ectocarpus*: the phosphomannomutase AlgC (Esi0149_0030, EC 5.4.2.8); the GDP-mannose 6-dehydrogenase AlgD (Esi0051_0092 and Esi0164_0053, EC 1.1.1.132); and the mannuronan C5-epimerase AlgG. Like *L. digitata*, *Ectocarpus* contains a large family of mannuronan C5-epimerases (28 genes). Surprisingly, we did not find any known alginate lyase, although six families of polysaccharide lyase (PL) have been described in other species. However, *Ectocarpus* does possess six glycoside hydrolases which are homologous to D-4,5 unsaturated β-glucuronyl hydrolase (family GH88), an enzyme that is specific for unsaturated oligosaccharides released by polysaccharide lyases. The presence of enzymes of the GH88 family in *Ectocarpus* strongly suggests that brown algae possess alginate lyases belonging to

novel PL family(ies). Taken together, these data indicate that brown algae have independently evolved a novel alginate metabolism, apart from the final C5-epimerization step, which may have been obtained secondarily from bacteria.

Interestingly, three mannuronan C5-epimerases (Esi0882_0001, Esi0069_0059 and Esi0010_0210) also possess WSC domains, which are potential carbohydrate binding modules (CBM) that were initially discovered in fungi[137,138]. Since CBMs generally have similar substrate specificities to their associated catalytic modules[139], these WSC domains are likely to bind alginates. Moreover, WSC domain proteins have undergone a spectacular expansion in *Ectocarpus* (Supplementary Table 9). Altogether, 115 genes contain at least one WSC module. Most of these proteins possess an ER signal peptide and they probably constitute an important class of cell wall proteins. Analysis of the WSC domains from 86 *Ectocarpus* proteins showed that they fell into three clades (clade A, B and C; see Supplementary Fig. 38). The proteins in clade A grouped with a β-1,3 exoglucanase from *Trichoderma harzianum*[137]. These included mannuronan C5 epimerases (Esi0069_0059; Esi0010_0210), a thaumatin-like protein (Esi0533_0007) and PR1-like metalloproteases (Esi0013_0157; Esi0013_0166; Esi0013_0168; Esi0026_0140 and the hypothetical pseudogene Esi0838_0002). The *Ectocarpus* WSC domain proteins in clade C clustered with the yeast proteins WSC1, WSC2 and WSC3. This clade included several putative PR1 proteins (Esi0277_0030; Esi0277_0035 and Esi0277_0044). An additional mannuronan C5 epimerase (Esi0882_0001) and a thaumatin-like protein (Esi0212_0047) fell into clade B. The presence of mannuronan C5 epimerases and thaumatin-like proteins in both clade A and clade B, as well as PR1-related proteins in clades A and C, suggests that the fusion of a WSC domain to these proteins has occurred repeatedly during evolution. No mannuronan C5 epimerases with a WSC domain have been identified so far in other brown algae, including *Laminaria digitata* and *Saccharina japonica.* Comparison of the *Ectocarpus* WSC domains allowed a consensus WSC signature to be deduced (Supplementary Fig. 39).

The metabolism of sulphated fucans in brown algae has not yet been characterised. The fucanase FcnA from the marine bacterium *Mariniflexile fucanivoran*s is the only protein known to be specific for algal sulphated fucans[140,141] but this family GH107 glycoside hydrolase is not conserved in *Ectocarpus*. More surprisingly, *Ectocarpus* does not possess a-L-fucosidases of the GH29 family, whereas these enzymes are well conserved in bacteria, animals and plants. We identified several fucosyltransferases from families GT10, GT23 and GT65, but these enzymes are involved in protein glycosylation in other eukaryotic phyla.

However, *Ectocarpus* contains ten sulphotransferases homologous to glycosaminoglycan (GAG) sulphotransferases from animals, as well as nine formylglycine-dependent sulphatases related to GAG sulphatases. These enzymes are thus likely to be involved in the sulphation / desulphation of sulphated fucans. It is noteworthy that formylglycine-dependent sulphatases are completely absent in plants, which have likely lost this protein family during their adaptation to terrestrial environment, concomitantly with the disappearance of sulphated polysaccharides from their cell wall[142].

In conclusion, brown algae have evolved a carbohydrate metabolism profoundly different from that of land plants. In particular, cell wall metabolism largely remains an uncharted territory and numerous CAZY families await discovery in the brown algae.

2.2.2. Photosynthesis genes

Most of the genes necessary to encode the enzymes involved in photosynthetic inorganic carbon fixation (the Calvin-Benson cycle) were found in the *Ectocarpus* genome and the proteins were predicted to be directed to the appropriate compartment (Supplementary Table 21). It has been suggested that brown algae may use C4 or CAM metabolism[143,144] and, consistent with this hypothesis, the enzymes necessary for C4 photosynthesis were identified in the genome. However, the predicted localisations of the proteins were not concordant with established C4-models in other organisms; for example, PEP-carboxylase was predicted to be directed to the mitochondria. This is potentially interesting since large numbers of mitochondria have been detected close to the cell wall in the Fucaceae[143] and inhibition of mitochondrial respiration reduces photosynthesis in brown algae[145]. This suggests that mitochondria might have an important role in inorganic carbon uptake in *Ectocarpus,* and that the initial steps of inorganic carbon fixation may be partly located in the mitochondria.

As far as the light reactions of photosynthesis are concerned, most of the genes normally associated with the antennae system could be identified (Supplementary Table 22). Exceptions include the gene for plastocyanin, which was expected to be absent because this protein is normally replaced by cytochrome c in chromophyte algae[116]. *Ectocarpus* also lacks PsbQ, in contrast to all the other photosynthetic organisms for which complete genome sequences are currently available, except *O. tauri* and the two diatoms. *Arabidopsis* mutants with reduced levels of PsbQ exhibited a wild type phenotype under normal growth conditions, but became yellow under low light conditions[146]. The *Ectocarpus* genome contains more light

harvesting complex (LHC) protein genes (also referred to as chlorophyll fucoxanthin binding genes within the stramenopiles) than any green plant genome studied to date (53, although some are probably pseudogenes). For comparison, poplar has 39 LHC genes, *Arabidopsis thaliana* has 30[147] and *Chlamydomonas reinhardtii* has 20[148]. *Ectocarpus* lacks a PsbS-type LHC, which is important for non-photochemical quenching in land plants, but this gene is also absent from the *T. pseudonana* genome[149]. However, *Ectocarpus* possesses a cluster of 11 LHCP genes on supercontig 18 (LHCP12-22, LHCP 15 probably being a pseudogene) that are most similar to the LI818- or LHCSR-family of light-stress-related LHC proteins. LI818 proteins have been implicated in the photoprotective xanthophyll cycle in the green alga *C. reinhardtii*[150] and the diatom *Cyclotella meneghiniana*[151], fulfilling a similar function to PsbS in vascular plants. This abundance of LHC genes in *Ectocarpus* might reflect the fact that the intertidal zone is an exceptionally dynamic environment where effective control of photosynthetic efficiency is particularly important.

Comparison of the light reaction and electron transport system gene complements of a range of genomes of photosynthetic organisms indicated that the gene complement of *Ectocarpus* resembles most closely that of *Arabidopsis*, these two species having the highest number of genes in common (Supplementary Table 22). *Ectocarpus* has a more complete set of genes than many other sequenced eukaryotic algae, with the exception of *Micromonas*.

2.2.3. Biosynthesis of tetrapyrroles, carotenoids and sterols

The major pigments in brown algae are chlorophyll *a*, chlorophylls c_1 and c_2, fucoxanthin, violaxanthin and β-carotene[152,153]. Chlorophyll *a*, β-carotene and violaxanthin are also present in vascular plants and green algae, but the *c*-type chlorophylls and fucoxanthin are specific to chromalveolate algae such as haptophytes and many stramenopiles (including the phaeophytes). The enzymes that are involved in the formation of these particular pigments are unknown, but the *Ectocarpus* genome contains homologs of all genes that are known to be involved in the biosynthesis of chlorophyll *a* and the respective carotenoids in vascular plants and green algae (Supplementary Fig. 40, Supplementary Table 23).

Chlorophylls. Most reactions in chlorophyll biosynthesis appear to be catalyzed by unique gene products in *Ectocarpus*, but in three cases there is evidence for ancient paralogs: uroporphyrinogen III decarboxylase (UROD, three genes), coproporphyrinogen III oxidase (CPX, two genes), and the ChlH subunit of protoporphyrin IX magnesium chelatase (two

genes). Because the paralogs share on average only between 30% (ChlH) and 35% (UROD and CPX) identical amino acid positions it is likely that their regulation or catalytic function has been modified.

The interesting features of the chlorophyll pathway in *Ectocarpus* are the presence of plastid genes encoding the magnesium-protoporphyrin IX monomethyl ester cyclase (subunit CHL27; *acsF* plastid gene) and three subunits (*chlB*, *chlL* and *chlN* plastid genes) of the light-independent NADPH:protochlorophyllide oxidoreductase (DPOR)[154]. Chl27 is nucleus-encoded in seed plants and green algae, but found on the plastid genome in red algae. It is absent from the available nuclear and plastid genomes of diatoms, and from the plastid genomes of the cryptophyte *Guillardia theta* and the haptophyte *E. huxleyi*[155]. More recently, the *acsF* gene has also been found on the plastid genomes of the raphidophyte *Heterosigma akashiwo*, the xanthophyte *Vaucheria litorea* and the phaeophyte *Fucus vesiculosus*[154 and references therein], suggesting independent loss in the diatom lineage after their separation from other heterokont lineages. The DPOR genes show a patchy distribution among plastid genomes from plants and algae. Among phototrophs with a primary plastid, DPOR has been detected in gymnosperms, mosses, several green algae, the rhodophytes *Porphyra purpurea, P. yezoensis* and *Galdieria sulphuraria*, and the glaucophyte *Cyanophora paradoxa*, but the genes are absent from angiosperms, prasinophytes and the rhodophyte *C. merolae*. In algae containing secondary plastids, DPOR is lacking in diatoms, the raphidophyte *H. akashiwo*, the haptophyte *Emiliania* and the cryptophyte *Guillardia*. Recently, however, DPOR genes have been reported on the plastomes of two other cryptophytes[156] and in the plastomes of the xanthophyte *V. litorea* and the phaeophyte *F. vesiculosus*[154 and references therein]. The presence of DPOR supports previous observations indicating that species of the closely related Laminariales synthesize chlorophyll in the dark, allowing arctic species to grow during the winter[157 and references therein]. In cyanobacteria, DPOR has been shown to be important for chlorophyll biosynthesis not only in the dark but also under low light conditions[158]. Similarly, DPOR in *Ectocarpus* may be important for an effective colonization of light-limited habitats.

Siroheme and heme. Siroheme is the cofactor of the plastid-localized enzymes nitrite reductase and sulfite reductase and is synthesized from uroporphyrinogen III. The *Ectocarpus* genome contains single genes encoding the enzymes uroporphyrinogen-III C-methyltransferase and sirohydrochlorin ferrochelatase (SirB). As components of the cytochromes, hemes are needed in plastids, mitochondria and the cytosolic compartment (e.g., cytochrome P450 in ER membranes). The *Ectocarpus* genome does not contain an ortholog of

aminolevulinate synthase (ALAS), which catalyzes the formation of the tetrapyrrole-building block aminolevulinate in the mitochondria of animals, fungi and Euglenid algae. Seven of the nine enzymes shared by the pathways of chlorophyll and heme biosynthesis are encoded by single genes. All enzymes are predicted to contain an N-terminal bipartite targeting sequence typical for plastid proteins suggesting that plastids are the only site of heme biosynthesis in *Ectocarpus*, similar to the situation in green algae like *Chlamydomonas*[159].

Carotenoids. The *Ectocarpus* genome encodes all the proteins of the methyl-erythritol phosphate (MEP) pathway, which synthesises active isoprene in the plastid. Similarly, all the genes of the cytosolic mevalonate (MVA) pathway of isoprene biosynthesis are present, including two isopentenyl diphosphate isomerase (IDI) genes. One of the genes may encode a plastid-localized IDI, but the current gene models of both paralogs lack a canonical bipartite taregting signal. In the post-isoprene part of the carotenoid biosynthesis pathway, all steps are carried out by enzymes that are encoded by single genes. This is similar to the situation in diatoms with the exception of phytoene synthase (PSY), phytoene desaturase (PDS), and zeaxanthin epoxidase (ZEP), which are present as two or three copies in the latter. Like diatoms, *Ectocarpus* lacks a non-heme iron carotene hydroxylase (CHYB), but contains genes for two cytochrome P450 enzymes (CYP97E3 and CYP97F4) that may be involved in the formation of zeaxanthin from β-carotene (see also section 2.2.8.). Carotenoid biosynthesis is of particular interest in chromist algae because this pathway gives rise to fucoxanthin which allows these algae to efficiently harvest blue light for photosynthesis, and which is responsible for their brown coloration. Diatoms and haptophytes possess two xanthophyll-based systems for dissipating excess light energy in the plastid, the violaxanthin cycle and the diadinoxanthin cycle[160]. These cycles are part of the biosynthetic pathway that synthesises fucoxanthin and both are catalysed by the activity of two opposing enzymes, zeaxanthin epoxidase and violaxanthin de-epoxidase[155]. Brown algae only have the violaxanthin cycle and this was correlated with the presence of a solitary zeaxanthin epoxidase gene compared with the two or three copies that have been found in diatom and haptophyte genomes[161], suggesting that the additional zeaxanthin epoxidases in diatoms and haptophytes might be involved in the formation of diadinoxanthin.

Sterols. In brown algae, cholesterol, 24-methylene-cholesterol (chalinasterol), and fucosterol have been reported as the major sterols[162-164]. Sterols are synthesized in the cytosol at the ER from isoprene units that are supplied by the MVA pathway[165]. The *Ectocapus* genome contains homologs of most of the genes known to be involved in the biosynthesis of

ergosterol/cholesterol in yeast/animals and of isofucosterol in land plants. A major difference of sterol biosynthesis in yeast/animals and land plants occurs at the level of squalene epoxide which is cyclized to lanosterol in the former and to cycloartenol in the latter. Both reactions are catalyzed by closely related oxidosqualene cyclases, with cycloartenol synthases and lanosterol synthases differing in three catalytically important amino acid residues[166]. For the stramenopile oomycete *Aphanomyces euteiches*, biochemical evidence indicated that lanosterol is the precursor of both cholesterol and fucosterol[167], and in accordance with this observation the oxidosqualene cyclase of *A. euteiches* contains the three amino acid residues that are characteristic of lanosterol synthases. The ortholog from *Ectocarpus* (Esi0148_0074/Esi0901_0001; gene split between two supercontigs), however, displays the canonical cycloartenol synthase residues. The formation of cycloartenol as an intermediate of sterol biosynthesis in *Ectocarpus* is further corroborated by the presence of a gene (Esi0169_0047) encoding a putative cyclopropyl isomerase that acts specifically on the cyclopropyl ring of cycloartenol and its derivatives.

2.2.4. Nitrogen metabolism

Analysis of the *Ectocarpus* genome indicates that this organism obtains nitrogen in at least three different ways (see Supplementary Table 24). Firstly, the presence of not less than fifteen paralogous *AMT1* genes encoding ammonium transporters (see Supplementary Table 25) indicates that ammonium is very efficiently pumped from the surrounding seawater and is probably the privileged nitrogen source for *Ectocarpus*. It is not clear which of these transporters are involved in transporting ammonium from the sea into the cytosol and which ones would be involved in a second step, the transport from the cytosol into plastids.

Secondly, *Ectocarpus* can assimilate nitrate, possessing five paralogous *NRT2* genes encoding high affinity nitrate transporters (one of these being a pseudogene in this strain) suggesting that, when available, nitrate is an important nitrogen source for *Ectocarpus*. The *NAR2* genes encoding the accessory component of high affinity nitrate transport have not been found, which is probably due to the fact that NAR2 proteins are poorly conserved. The imported nitrate is converted into nitrite by a NAD(P)H:nitrate reductase encoded by a *NIA* gene. Nitrite is then transported from the cytosol into plastids and maybe also from the surrounding medium into the cytosol, as in *Chlamydomonas*[168], through the action of two nitrite transporters NAR1;1 and NAR1;2. Finally nitrite is converted into ammonium by nitrite

reductase. Interestingly, *Ectocarpus* possess two distinct proteins that could perform this task, using either ferredoxin as reducing substrate as in plants (NII) or NAD(P)H as in bacteria (NIR). The same co-occurrence of these two nitrite reductases has also been found in the diatoms *Thalassiosira* and *Phaeodactylum*, and this may provide these marine organisms with the capacity to reduce nitrite when reduced ferredoxin is limiting, e.g. at night and under low light.

Finally, *Ectocarpus* is also able to transport urea into its cells (*DUR3*) and convert it into ammonium using a urease (*UREABC*, *URED*, *UREF*, *UREG*). As observed in diatoms, the *Ectocarpus* genome also encodes a complete urea cycle including an arginosuccinate synthetase (Esi0112_0076), an ornithine transcarbamoylase (Esi0361_0020) an argininosuccinate lyase (Esi0081_0083) and an arginase (Esi0073_0060).

Contrary to what has been observed in prasinophytes[169-171], *Chlamydomonas*[168] and fungi, the nitrogen assimilation genes are dispersed all over the genome and are never clustered, even as gene pairs. On the other hand, some paralogous genes in the pathway are occasionally arranged in tandem copies, such as the five copies of *NTR2* on supercontig 278 and tandem copies of *AMT1,* which occur twice on supercontigs 464 and 62.

2.2.5. Amino acid biosynthesis

The *Ectocarpus* genome is predicted to encode all the enzymes necessary for the synthesis of Pro, Ala, Phe, Tyr, His, Cys, Lys, Gly, Trp, Thr, Met, Asp, Asn, Glu, Gln, Ser, Val, Leu, Ile, and Arg. Acetolactate synthase, which has a central role in valine, leucine and isoleucine synthesis is encoded by two genes (*ilvB* and *ilvH*) on the *Ectocarpus* plastid genome[154] but by two nuclear genes in the diatom *P. tricornutum*. No glutamate decarboxylase and GABA transaminase genes were found in *Ectocarpus*, indicating that gamma-aminobutyric acid (GABA) biosynthesis via the GABA shunt is absent. Both of these genes are present in *Phytophthora* but not in diatoms. GABA is an amino acid that is not found in proteins but which functions as a neurotransmitter in mammals and insects. It is produced by both green plants and animals.

2.2.6. Thiamine pyrophosphate (vitamin B1) biosynthesis

All the genes necessary for the biosynthesis of thiamine pyrophosphate (TPP) were found in the *Ectocarpus* genome (Supplementary Fig. 41) but none of those involved in thiamine transport, indicating that all the TPP required is synthesised within the cell. Surprisingly, the TPP pathway resembles that of bacteria in many ways, for example the thiazole precursor of thiamine is synthesised by ThiO and ThiG orthologs, rather than ThiM and Thi4 as described for other eukaryotes.

Single exon genes are rare in the *Ectocarpus* genome but *TPK1*, which converts thiamine to TPP, does consists of only one exon, as does the gene immediately upstream. TPK1 appears to be a fusion protein, the N-terminus being distantly related to the bacterial TMP phosphorylase ThiE (this particular fusion is unique, but TPK1 is also a fusion protein in *Ustilago*, in this case with a tRNA nuclease at the N-terminus). The *Ectocarpus* TPP pathway is also unusual in that it does not appear to be regulated by a TPP riboswitch (no riboswitches were found using the "Infernal" software[172]).

2.2.7. Lipid and fatty acid metabolism

Brown algae are able to produce both "plant-like" C18 and "animal-like" C20 PUFAs, molecules. These are likely to be important in these algae both as precursors of oxylipins involved in defence and stress responses and for the production of the sexual pheromone[173,174].

The *Ectocarpus* genome encodes the enzymes necessary for the biosynthesis of fatty acids from acetyl-CoA and malonyl-CoA in plastids. Fatty acids are then exported to the cytoplasm via the action of acyl-ACP thioesterases and acyl-CoA synthetases, where they can be used for the synthesis of polyunsaturated fatty acids (PUFAs), the most abundant being 18:2n-6, 18:3n-6 and 20:4n-6 for the omega 6 series, and 18:3n-3 and 20:5n-3 for the omega 3 series. No DHA has been found in *Ectocarpus*[175]. A complete set of desaturases and microsomal elongases potentially involved in PUFA biosynthesis were identified in the genome. These genes have homologues in diatoms and other microalgae with well-characterized functions.

The PUFAs can be oxygenated to produce oxylipins. Synthesis of these derivatives involves cleavage of PUFAs from phospholipids by phospholipases followed by the action of

lipoxygenases (LOXs) to produce hydroperoxides. There are four LOX genes in *Ectocarpus*, forming two groups (Esi0424_0005 and Esi0424_0006; Esi0010_0044 and Esi0010_0039) based on analysis of the primary sequences. Their specificity (C18 and/or C20 PUFA) could not be inferred from their sequences. However, the two groups are located on two different supercontigs, and the proteins encoded by the genes on sctg 0010 possess a bi-partite signal peptide and are predicted to be targeted to the chloroplast. In plants, C18 derivatives, such as jasmonic acid, are derived from alpha-linolenic acid via 13-hydroperoxide, which is converted to 12-OPDA by Allene Oxide Synthase (AOS), a member of the cytochrome P450 family (CYP5164A2 Esi0060_0078 and CYP5164B1 Esi0111_0095 are the best candidates for an *Ectocarpus* AOS or they could alternatively be a hydroperoxyde-lyase, an aldehyde-producing enzyme that is necessary for the production of brown algal pheromones; experimental verification is necessary.) and an Allene Oxide Cyclase (AOC, one candidate in the *Ectocarpus* genome). The 12-OPDA is then transferred to the peroxisome, either passively or by an ABC type transporter, where it is reduced by an NADH:flavin oxidoreductase. The *Ectocarpus* genome encodes several NADH:flavin oxidoreductases that could potentially catalyze this step. Reduction is followed by three cycles of beta-oxydation resulting in the production of jasmonic acid (JA). In land plants this molecule is then converted to methyl-jasmonate by a JA carboxyl methyltransferase (JMT), but no JMT homologues were found in the *Ectocarpus* genome. Note that both JA and 12-OPDA function as signalling molecules in land plants[176,177]. This may also be the case in brown algae as 12-OPDA was detected in *Laminaria digitata* following copper stress without production of JA[178]. Brown algae can also produce several types of C20 oxylipins such as prostaglandins[178]. The *Ectocarpus* genome encodes one potential alpha-dioxygenase/cyclooxygase (Esi0026_0091), a number of cytosolic Sigma class GSTs and three microsomal GSTs (Esi0122_0061, Esi0122_0054, Esi0122_0055; de Franco et al., 2009), which may be involved in the production of these oxylipins. Supplementary Fig. 42 presents an overview of fatty acid and lipid metabolism in *Ectocarpus*.

Another class of lipids, sphingolipids, which are structural components of membranes and have signalling roles in plants and mammals, has not been looked for in *Ectocarpus*, but genes encoding the enzymes necessary for sphingolipid synthesis are present in the genome (with the exception of a sphingolipid delta4-desaturase). Delta4- and delta8-unsaturated C18-LCB (long chain bases) sphingolipids have been reported in the diatom *T. pseudonana*[179].

As far as membrane lipids are concerned, the *Ectocarpus* genome encodes the glycerolipid enzymes MGDG synthase and DGDG synthase, and two enzymes involved in sulfolipid (SQDG) synthesis (UDP-sulfoquinovose synthase, sulfolipid synthase), but no sulfolipid 2'-O-acyltransferase, which catalyses the last step in sulfolipid synthesis. All the enzymes necessary for the synthesis of the glycerophospholipids PC, PE and PI are present. No betaine lipid synthase could be identified, explaining why *E. siliculosus* does not contain DGTA, in contrast to *Ectocarpus fasciculatus*[180].

In yeast and in plants, storage lipids (triacylglycerols) can be produced via either the acyl-CoA dependent pathway, called the Kennedy pathway, whose last step is catalysed by the diacylglycerol acyltransferase (DGAT), or the acyl-CoA independent pathway, which involves the transacylation of acyl groups from phospholipids to diacylglycerols. In *Ectocarpus*, genes potentially encoding both type 1 and type 2 DGATs were found but no phospholipid-diacylglycerol acyltransferase gene could be identified, suggesting that *Ectocarpus* does not possess the acyl-CoA independent pathway.

Finally, *Ectocarpus* encodes all the enzymes necessary for the degradation of fatty acids by beta-oxidation, some of them present as multigene families. As in diatoms[28], two beta-oxidation pathways are present, one localized in the mitochondria, the other in the peroxysomes.

2.2.8. P450 oxidoreductases

Eukaryotic cytochrome P450s are oxidoreductases that require an accessory source of electrons such as NADPH cytochrome P450 reductase to function[181]. These proteins are usually membrane bound, most frequently associated with the endoplasmic reticulum membrane, with the bulk of the protein being exposed to the cytosol. *Ectocarpus* has 12 cytochrome P450s and these have been assigned names using the standardised nomenclature, which is based on sequence relatedness[182]. Three of these sequences can be assigned to families that include enzymes of known function, whilst the remaining nine sequences belong to six new families of unknown function. The CYP97 family is involved in carotenoid hydroxylation in terrestrial plants[183-185] suggesting that the *Ectocarpus* proteins CYP97E3 and CYP97F4 act as beta-carotene and beta-cryptoxanthin hydroxylases, and are probably important for the biosynthesis of xanthophylls. Both *Phaeodactylum* and *Thalassiosira* possess orthologues of CYP97E3 and CYP97F4 and these genes are all most similar to the

CYP97B family in green plants. This suggests that the CYP97B family is ancestral compared to the green plant CYP97A and CYP97C families, which would have arisen later within the archaeplastida group. The CYP51 family consists of sterol 14 alpha demethylases involved in the synthesis of cholesterol in animals, ergosterol in fungi and related sterols in plants[186]. The *Ectocarpus* protein CYP51C1 therefore probably has a similar activity indicating that brown algae are able to synthesize sterols. All eukaryotes need some type of sterol-like molecule in their membranes (although ciliates use another tetrahymenol, which is synthesised without using oxygen, in its place[187]) and those that have lost the CYP51 family (protostomes, tunicates, *Giardia* and parasites such as *Plasmodium*) acquire sterols by ingestion or import. Conservation of P450 families is highly variable within the stramenopiles, for example *Ectocarpus* shares two and three P450 families with *Thalassiosira* and *Phaeodactylum*, respectively (including CYP51 in both cases), whereas it has none in common with the oomycetes *Phytophthora ramorum* and *Phytophthora sojae*.

2.2.9. Secondary metabolism

The shikimate pathway is fully conserved in *Ectocarpus* but some of the pathways that branch off the shikimate pathway in higher plants are absent (Supplementary Table 26), including pathways that produce important compounds such as phenylpropanoids and salicylic acid. Interestingly, the absence, in *Ectocarpus*, of a gene coding for Phenylalanine Ammonia Lyase (PAL), the enzyme which controls entry into the phenylpropanoid pathway, is consistent with a recent phylogenetic study of the evolution of PAL in bacteria and eukaryotes[188]. This study concluded that a horizontal gene transfer from bacteria was at the origin of phenylpropanoid metabolism in terrestrial plants.

In brown algae, a broad range of polyphenolic compounds involved in UV protection, cell wall strengthening, adhesion and defence are derived from the acetate-malonate pathway[189]. Biosynthesis of phloroglucinol, the precursor of brown algal tannins, probably involves three type III polyketide synthases (PKS), which are closely related to plant chalcone or stilbene synthases (Supplementary Table 26). One isoform of PKSIII is highly represented in the EST libraries, probably reflecting a key role in this pathway. Analysis of the locations of these genes in the genome did not find any evidence for their occurring in functional clusters dedicated to the synthesis of pigments or other secondary metabolites. No homologues of these PKS genes were found in oomycete or diatom genomes. Note also that

no multidomain type I PKS genes, as identified in Mamiellales genomes[131], were found in *Ectocarpus*. Flavonoid metabolism is also fully conserved in *Ectocarpus* compared to higher plants (Supplementary Table 26). It has been proposed that the complex phenylpropanoid pathway in higher plants with its branches through the flavonoid and lignin pathways evolved posteriorly to the horizontal acquisition of PAL from a bacterial genome[188]. However, the existance of the flavonoid pathway in a broad range of photosynthetic organisms, including stramenopiles, suggests rather that this pathway is ancestral and therefore that the PAL enzyme acquired by the first land plants would have been integrated into a pre-existing pathway.

One role of this complex pathway in *Ectocarpus* is probably to provide the repetitive units with aryl-ester or aryl-aryl bounds found in oligomers of phlorotannins. However, a recent report based on the analysis of two japanese kelp species has shown that brown algae also secrete three monomeric bromophenols (2,4-dibromophenol, 2,4,6-tribromophenol and dibromo-iodophenol, but not phloroglucinol or phlorotannins) into the surrounding seawater[190]. Consequently, there is currently considerable interest in reinvestigating the metabolome of *Ectocarpus*, including the exo-metabolome, as this may display a greater diversity in terms of phenolic metabolism than previously thought[189]. Metabolism of phenolics may also have interesting links with the unusual halide metabolism in macroalgae (see section 2.2.10.).

2.2.10. Halogen metabolism

The metabolism of halogenated compounds is widespread among both prokaryotes and eukaryotes, with the notable exception of terrestrial green plants. This metabolism has been well characterised in mammals, both with respect to the organification of iodine in the thyroid gland and in terms of dehalogenating processes[191]. Many macroalgae are able to accumulate high concentrations of halides (iodide and/or bromide) from seawater, and they can produce volatile halocarbons that are often viewed as microbiocidal metabolites by analogy with the halogenated compounds produced in mammalian phagocytes as potent oxidants[192]. A recent study described an unusual anti-oxidant system in kelps, based on the accumulation of iodide, which is then oxidized following oxidative stress through the catalytic activity of vanadium-dependent haloperoxidases (vHPO). This system could potentially have a significant impact on atmospheric chemistry[193]. vHPOs are likely to play a central role in halogen metabolism,

both in halide uptake and in the production of halogenated compounds, and they are also thought to be involved in the scavenging of activated oxygen species, produced in the cell in response to oxidative stress. vHPOs have been shown to catalyze oxidative cross-linking between cell wall polymers of brown algae, suggesting that they may be involved in spore and gamete adhesion and cell-wall strengthening[194]. These enzymes belong to a class of non-haem peroxidases and catalyze the oxidation of halides in the presence of hydrogen peroxide. The hypohalous acids formed by this reaction can then transfer halogen atoms onto a wide range of organic molecules. Whereas vanadium-dependent chloroperoxidases of terrestrial organisms catalyze the oxidation of chloride, bromide and iodide, vanadium-dependent bromoperoxidases (vBPOs), which are mainly present in brown and red seaweeds, react only with bromide and iodide. Iodoperoxidases (vIPOs), which specifically oxidise iodide, have been described only in kelps. Only one vanadium-dependent haloperoxidase gene was found in the genome of *Ectocarpus* (Supplementary Table 27), in striking contrast with *Laminaria digitata*, which possesses two large multigenic families of vBPOs and vIPOs[195,196]. The *Ectocarpus* vHPO is more closely related to *L. digitata* vBPOs (70% amino acid identity) than to vIPO (~ 30% amino acid identity), and the introns are in similar positions. The protein is predicted to possess a signal peptide indicating that it is an extracellular vBPO, targeted to the cell wall. No vIPO homologue was detected in the *Ectocarpus* genome. *Ectocarpus* accumulates halides (~ 0.08 mg iodine and 0.24 mg bromine / g dry weight) but, at least for iodine, accumulation is to a much lower level than has been observed in kelps (1000-fold accumulation compared to 30,000-fold for *L. digitata*). Consistent with this low level of halide accumulation, the *Ectocarpus* vBPO gene is only expressed at a low level in sporophytes (0.1% of the available ESTs, compared to 4% in *L. digitata* sporophyte ESTs). No transcripts were detected in the gametophyte EST collections. In *L. digitata* sporophytes, defence responses appear to involve tightly coordinated regulation of the two distinct haloperoxidase gene families, which are likely to have been derived from an ancestral gene duplication[196,197]. Hence there is a marked difference between the closely-related Ectocarpales and Laminariales, in that a highly developed, iodine-based defence metabolism has evolved in the macroscopic parachymentous sporophytes of kelps, but this system is not present, or at least not to the same degree, in the filamentous Ectocarpales.

Interestingly, other halogen-related enzymes have been identified on the *Ectocarpus* genome. These include at least three different families of haloacid dehalogenase (HAD; InterPro: IPR005834) (corresponding to 21 loci) and two haloalkane dehalogenases

(Supplementary Table 27). The HADs belong to a very large superfamily of hydrolases with diverse substrate specificity, including phosphatases and ATPases. The dehalogenase enzymes may serve to defend *Ectocarpus* against halogen-containing compounds produced as defence metabolites by kelps[193] allowing it to successfully grow as an epiphyte on the surfaces of the kelp thalli[198,199].

2.2.11. Mechanisms for alleviating oxidative and metal stress.

The production of reactive oxygen species (ROS) as part of an innate immunity primary response or in developmental responses is often mediated through the activation of plasma-membrane-bound NAPDH oxidases (NOX, also known as gp91 respiratory burst oxidase homologs or rboh). NADPH oxidases are also involved in various responses to mechanical stress and in developmental responses in both metazoans and plant cells[200]. Two NADPH oxidase genes (plus one pseudogene) have been found but these lack EF hand domains and are most like red algal NADPH oxidases (Supplementary Table 28). Interestingly, domain analysis (section 2.1.8.) indicated that the genomes of multicellular organisms, including *Ectocarpus*, consistently possessed more NADPH oxidase genes than those of unicellular organisms (6-20 genes for the former compared with 0-6 genes for the latter with the domain PTHR11972). This indicates a strong link between this gene family and the emergence of multicellularity.

Photosynthesis is also a source of ROS, because of the proximity of electron transport and oxygen production. In this case, the ROS are an unwanted by-product of the light reactions and photosynthetic organisms use a variety of mechanisms to scavenge these potentially destructive molecules. *Ectocarpus* has a number of other antioxidant proteins, in addition to the haloperoxidase mentioned above (see section 2.2.10.). These include six superoxide dismutase genes (four Fe/Mn and two Cu/Zn SOD) and a multigene family of 11 catalase genes, including three pseudogenes, which exhibits evidence of tandem duplications (Supplementary Table 28). Most of the *Ectocarpus* catalases are predicted to be targeted to the peroxisome. The two highly similar (90%) Cu/ZnSOD genes have a conserved dimerisation motif in their C-terminal regions. However, one of the two conserved histidine residues in the cofactor-binding motif has been replaced by a phenylalanine. As this histidine is required for the activity, it is possible that these enzymes are inactive. This would be consistent with the apparent absence of a CuZnSOD chaperone. No matches were found with

the recently described Ni SOD found in marine cyanobacteria[201]. The genome encodes at least nine glutathione peroxidases, eight GSTs and several peroxiredoxins and thioredoxin peroxidases. Glutathione peroxidases (GPX) catalyze the reduction of hydroperoxides by glutathione (GSH) oxidation. In plants they play an important antioxidant role in the GSH-ascorbic acid (AsA) and GPX cycles. The *Ectocarpus* genome encodes seven GPXs, all of which have a thioredoxin fold, a dimerisation site and Cys/Ura catalytic motifs. GPX6 is encoded by the mitochondrial genome. Interestingly this gene is highly similar to cyanobacterial GPXs. Glutathione reductases (GRs) are NADH oxidases that catalyze the reduction of oxidized gluthathione, using NAD(P)H as a cofactor[202]. They are important components of the ascorbate-glutathione and the GPX cycles. The *Ectocarpus* genome encodes two GRs (Supplementary Table 28) with conserved features such as an NAD(P)H binding domain (Pyr_redox) and a dimerisation domain (Pyr_redox-dim). Dehydroascorbate reductases (DHARs) are plant-specific monomeric enzymes that catalyse the reduction of DHA to produce AsA, using glutathione as the reductant. They allow plants to recycle oxidized AsA and are therefore important in the GSH-AsA cycle[202]. The single *Ectocarpus* dehydroascorbate reductase (DHAR) has typical DHAR features such as thioredoxin-like and DHAR-GST-C domains, together with a signal peptide, but its specific subcellular location is unclear. Monodeshydroascorbate reductase (MDAR) catalyses the reduction of Monodesydroascorbate into AsA, and is therefore important in AsA recycling via the GSH-AsA cycle. The single MDAR gene in the *Ectocarpus* genome possesses an NAD(P)H binding domain (Pyr_redox). One sequence coding for an ascorbate peroxidases (APX) was identified in the *Ectocarpus* chloroplast genome. These heme metalloenzymes catalyse the peroxidation of AsA to produce MDA, and are therefore important in the chloroplast water-water cycle and the GSH-AsA cycle. The heme domain and the iron cofactor histidine motif are conserved in the *Ectocarpus* protein.

ROS may also be produced during metal stress. *Ectocarpus* is a metal (copper)-tolerant alga, which is able to colonise ship hulls despite the presence of anti-fouling paints[203]. Interestingly, the metal-complexing proteins encoded by the genome include two metallothionein (MT) genes and one homolog of a phytochelatin synthase gene. Both metallothionein genes have conserved Cys residues, which are shared with plant, animal and algal sequences. EsMT1 shows high similarity with *Fucus* MT with over 60% of identity. This protein has a 16aa spacer and a total of six CXC motifs distributed equally within the two Cys domains, allowing it to be classed with the type I MTs[204]. However as with *Fucus*

MT, the spacer region is smaller than in other type I MTs and this feature could be characteristic of brown algal MT[205]. EsMT2 is smaller than EsMT1, and shows similarity to small plant type I MTs. This protein also has an extremely reduced spacer with CXC motifs within the Cys rich regions, a feature common with plant small MTs. Forty ESTs were detected for the EsMT2, indicating that it is highly expressed. Type I and II MTs appear to be important copper chelators in vascular plants[206]. In addition, the MT gene in *F. vesiculosus* is induced by copper excess[205]. A *Fucus* MT fusion protein showed a greater affinity for Cu than Cd and the pH required for dissociation of Cd from this protein was approximately 2 pH units higher than for a recombinant human MT.

In common with most plants and animals[207], *Ectocarpus* has a single phytochelatin synthase PCS gene. This protein shares over 60% identity with plant PCSs, whereas much lower identity (18%) was observed for a putative *T. pseudonana* PCS. Plant PCSs have a conserved N-terminal region and a variable C-terminal region containing multiple cysteine residues[204]. Interestingly, the conserved N-ter region of plant PCSs corresponds to the C-terminal region of EsPCS and the N-terminal region of the latter does not possess Cys residues. The gene model is supported by 11 ESTs. Additional copper tolerance genes include three multicopper oxidase genes and several transporters and protein chaperones (PCs). Following metal chelation PCs are transported into vacuoles by a multidrug ABC type transporter[208]. A putative PC ABC transporter with conserved sequence features such as 9 TMDs, a P-loop N-ATPase domain was identified in the *Ectocarpus* genome. However, this protein shares only 25% similarity with higher plant transporters, and further investigation is therefore required to test whether this gene or other ABC transporters in the *Ectocarpus* genome mediate transport of PC into the vacuole.

2.2.12. Iron uptake and storage

Iron is an essential element for all living organisms due to its ubiquitous role in redox enzymes, especially in the context of respiration and photosynthesis, and algae are no exception to this. However although it is the fourth most abundant element in the Earth's crust, it is present under aerobic conditions at neutral pH only in the form of extremely insoluble minerals like hematite, goethite, and pyrite or as polymeric oxide-hydrates, -carbonates, and -silicates that severely restrict the bioavailability of this metal. The iron level in open ocean waters is even lower than in most terrestrial environments[209-211], particularly so

since a large fraction of the limited iron available in the ocean is already tightly complexed[210,212].

In terrestrial plants, two basic strategies for iron uptake have been distinguished, with strategy I plants (mainly dicotyledons) using a mechanism involving soil acidification, lateral root formation, specific transfer cells in the rhizodermis as well as induction of an Fe(III)-chelate reductase (ferrireductase) and of transporter proteins, transferring Fe(II) into the cells[213,214]. In contrast, strategy II plants (in particular, monocotyledons / grasses) take up Fe as a phytosiderophore complex. Other iron uptake systems are known for eukaryotes. These include a reductive-oxidative pathway such as that utilized by yeast as well as an Fe(III) permease system. In addition heme uptake systems are common. Bacteria and fungi on the other hand have evolved sophisticated systems based on high-affinity iron specific binding compounds called siderophores to acquire, transport and process this essential, but biologically unavailable, metal ion. Their major role is the extracellular solubilization of iron from minerals and/or organic substrates and its specific transport into microbial cells.

The iron uptake system in *Ectocarpus* genome most closely resembles that of strategy I plants. *Ectocarpus* possesses homologues of FRO2, an iron chelate reductase, as well as NRAMP, a M^{2+}-H^{+} symporter with a preference for Fe(II). This is similar to what has been seen in the related pennate diatom *P. tricornutum* and distinct from the reductive-oxidative pathway found for the centric diatom *T. pseudonana*. *Ectocarpus* also has a suite of genes which could be part of a mugineic acid (phytosiderophore) biosynthesis pathway (i.e. NAAT, DMAS, IDS2 and IDS3) but it lacks a homologue of the first committed enzyme of the pathway, NAS (nicotianamine synthase), which converts S-adenosyl-L-methionine to nicotianamine or the phytosiderophore receptor YSL. This indicates that there is no active classic strategy II phytosiderophore although it is possible that there is a system based on a core structure different from nicotianamine. No receptors for common bacterial siderophores such as *fhu, fep or fec* were found. The presence in *Ectocarpus* of an excretable "siderophore-like" molecule was recently experimentally verified (data not shown) suggesting that at least some iron acquisition mechanisms in Ectocarpus may have been derived from an ancient horizontal gene transfer.

Although iron is an essential bioelement, excess iron is toxic due to the formation of ROS via Haber-Weiss-Fenton chemistry and thus its uptake must be tightly controlled and a means for safe storage is needed. Iron is typically stored as a complex with the ubiquitous proteins of the ferritin superfamily i.e. the ferritins ("maxi-ferritin"), bacterioferritins, or DPS

("mini-ferritin) widely found in animals, plants and bacteria. However, although the diatom *P. tricornutum* has ferritin genes, they have not been detected in any other stramenopile genome and *Ectocarpus* has no homologues of any of these proteins. Recently an alternate method for iron storage, involving vacuoles, has been elucidated in yeast and several other eukaryotes including the halotolerant alga *Dunaliella salina*. The *Ectocarpus* genome encodes NRAMPs, which may be important for iron release or mobilization as part of this storage strategy, but no homologues of the carrier CCC1p were found. Thus, at present, no iron storage system has been identified in *Ectocarpus*. Iron K-edge XAS analysis of *Ectocarpus* tissue showed that most of the Fe pool is present as Fe(III), with most of this being coordinated to sulphur and nitrogen in complexes such as Fe-S clusters and heme, respectively (FCK, Martin C. Feiters and Wolfram Meyer-Klaucke, unpublished data). Fe-O clusters were not detected, providing support for the absence of ferritin.

2.2.13. Selenoproteome

Usually only a small number of proteins contain selenocysteine (Sec) in any particular organism. Humans and some marine organisms, for example, have more than 20 selenoproteins whereas land plants do not have any. Sec most often replaces cysteine (Cys) and, because Sec is a stronger nucleophile than Cys, selenoproteins often function as redox partners in a diverse range of biological processes.

The biosynthesis of selenocysteine and of selenoproteins is a complex process. The main players in this process have only recently been identified and many details still remain to be elucidated[215]. A search of the *Ectocarpus* genome identified putative orthologues of all of the documented biosynthesis proteins (see Supplementary Table 29) apart from O-phosphoryl $tRNA^{Sec}$ kinase (PSTK), which is poorly conserved across species. Contrary to what has often been observed in other species, none of these proteins are themselves selenoproteins. Interestingly, the single gene encoding selenophosphate synthase is closely related to bacterial SELD and not to the eukaryotic SPS genes, which are often found in pairs in higher eukaryotes (SPS1, SPS2), one member of each pair encoding a selenoprotein. Prasinophytes also possess a single selenophosphate synthase gene that is closely related to bacterial genes.

As far as the selenoproteins themselves are concerned, initially a search was carried out by screening for homologues of known selenoproteins. Five selenoproteins (two

glutathione peroxidases and homologs of SELM, SELT and SELW) were identified by this approach (Supplementary Table 29). The functions of the latter three proteins are not known. To identify fast-evolving or novel selenoproteins a search was also carried out for SECIS elements, an approach that has previously been used successfully for prasinophytes[169,170]. SECIS elements are secondary structures located in the 3'UTRs of selenoprotein-encoding genes. When a SECIS element is present, all upstream UGA codons in the coding sequence (there is usually only one) are decoded as Sec instead of as stop codons. The screen for this element was carried out using the SECIS secondary structure from the RFAM database (RF00031) in conjunction with the Infernal software package[216]. However, no significantly predicted SECIS elements were found, not even in the vicinity of the five selenoprotein genes that had previously been identified by protein similarity. A local search for SECIS elements downstream of these five genes was therefore carried out using using SECISearch alone[217] or embedded in SECISaln[218]. Two SECIS elements were identified by this analysis, one downstream of the *SELT* gene and the other one downstream of the *SELW* (Supplementary Fig. 43), but the scores for both elements were low. No valid prediction can be made for the three other genes, whatever the parameters used. Taken together these analyses suggest that the SECIS element is not canonical in *Ectocarpus*. Known SECIS elements exhibit a significant level of variability[218] and it would not, therefore, be surprising for an organism such as *Ectocarpus*, from a poorly explored phylogenetic group, to possess non-canonical SECIS elements.

Although it is possible that additional selenoproteins will be identified in the future, the presence of only five selenoproteins in *Ectocarpus* suggests that it contains significantly less of these proteins than the prasinophytes, which consistently have more than 25 selenoproteins. Numbers of selenoproteins tend to be higher in marine organisms and this has been proposed to be an adaptation to the marine environment[219]. However, *Ectocarpus* is found in a coastal habitat and this may influence the need for selenoproteins in this organism.

2.3. Signalling and cell biology

2.3.1. Transcription associated proteins

Transcription associated proteins (TAPs) include transcription factors (TFs, proteins that bind to cis-regulatory elements enhancing or repressing gene transcription) and transcriptional regulators (TRs, proteins with indirect regulatory functions, such as the assembly of the RNA polymerase II complex, functioning as scaffold proteins in enhancer/repressor complexes or controlling chromatin structure by modifying histones or the DNA methylation). The expansion of TAP families has often been linked with the evolution of complexity in animals[220-223] and it has been proposed that there is a correlation between the complexity of an organism and the proportion of its protein coding genes that encode TAPs[221]. Similarly, a recent comparison revealed a significant increase in the average size of TF and TR gene families in land plants compared with unicellular algae[63].

A total of 401 TAP genes were found in the *Ectocarpus* genome (Supplementary Table 4). These genes were from 54 of the 111 TAP families analysed in this study (Supplementary Table 4). To search for evolutionary trends, the *Ectocarpus* TAP genes were compared with the TAP gene sets of a broad range of completely sequenced land plants, algae, protists, fungi and animals (Supplementary Table 4), the genomes analysed are listed in Supplementary Table 1). A comparison of the absolute numbers of TAPs per organism suggested a significant expansion of TAPs in land plants compared to algae (Supplementary Fig. 44a), which is congruent with a previous analysis based on fewer organisms[63]. The haptophyte *Emiliania huxleyi* appeared to be an exception to this trend, but *E. huxleyi* more closely resembled the other algae (Supplementary Fig. 44b) when the TAP gene component of each genome was presented as a proportion of the total gene content of the genome (reducing any bias due to large-scale duplication events). A principal component analysis on the TAP family sizes resulted in separation according to taxonomic group (Supplementary Fig. 45). The first two Eigen items contribute 42.8 and 16.4% of variance, respectively, and are not sufficient to separate the red/green lineage from the stramenopiles (Supplementary Fig. 45a,b). The third Eigen item contributes another 8.3% and more clearly separates according to taxa (Supplementary Fig. 45c). Interestingly, we observed no apparent correlation between multicellularity and the proportion of the genome that consisted of TAP genes (Supplementary Figs. 44, 45b).

To further investigate the relationship between TAP genes and multicellularity, comparisons of the sizes of each TAP family in multicellular and unicellular organisms were performed. This analysis identified 29 families that were consistently larger in multicellular organisms, 17 of which encoded TFs. When land plants were compared with algae, 27 families were found to be consistently larger in the former (18 of which encoded TFs). We compared this latter set of families with the families that were larger in multicellular organisms and removed 20 families that occurred in both sets (to eliminate the influence of the transition to the terrestrial environment in the plant lineage). The remaining seven families are good candidates for TAP genes that played a role in the transition to multicellularity. They include two TR families (MADS and LIM) and five TF families (GNAT, SWI/SNF_SNF2, SWI/SNF_SNF3, TRAF and AN1/A20 type zinc finger. A comparison between stramenopiles and the red/green lineage only identified one TAP family, namely the TF HSF family, that exhibited a significant size difference between the two lineages.

Several other comparisons were carried out, for example between photosynthetic and non-photosynthetic organisms, between stramenopiles and the red/green lineage, between red and green algae and between organisms derived from primary vs. secondary endosymbioses, but no TAP families were identified as clearly correlating with these divisions.

Several TAP families are present in *Ectocarpus* and oomycete genomes but absent from unicellular diatoms (the TF families ARID and RWP-RK and the TR families MBF1, MED6 and AN1/A20 Zinc finger). In most cases these were small families with one or two members, with the exception of the NIN-like family, which includes eight members and was therefore significantly overrepresented in *Ectocarpus* and oomycetes compared to diatoms. All families showing at least a two-fold pairwise size bias between *Ectocarpus*, the two diatoms and the two oomycetes, respectively, were visualized in a heat map (Supplementary Fig. 46). A total of 16 families are biased between *Ectocarpus* and the diatoms, 10 between *Ectocarpus* and oomycetes and 16 between diatoms and oomycetes. Five families are biased between *Ectocarpus* and both, diatoms and oomycetes. These are the TRs Coactivator p15, HMG and SWI/SNF_SNF3 and the TFs C2H2 and mTERF. Overall, the two diatom genomes have lost more stramenopile TAP families (i.e. TAP families that occur in at least one stramenopile) than have *Ectocarpus* and the oomycetes.

RWP-RK proteins. The RWP-RK domain superfamily or NIN-like proteins *sensu lato* are putative transcription factors characterized by the PFAM domain RWP-RK. These proteins are often involved in nitrogen-controlled development[224,225]. The gene family is

divided into two major subfamilies (Supplementary Fig. 47). One subfamily contains the seed plant NIN-like proteins *sensu strictu,* required for establishing symbioses between legumes and *Rhizobia*, the other subfamily contains minus dominance (MID) proteins from volvocine algae[226-228]. Land plants and most green algae possess quite a large number (8-19) of NIN-like proteins whereas red algae and prasinophytes have less (1-5). The eight genes in the *Ectocarpus* genome (Supplementary Fig. 47) cluster basal to the MID, RKD and NIN-like gene families, which could indicated possible functions in either nitrate signalling[229] and/or gametogenesis[230]. In volvocine algae gametes are produced in response to nitrogen starvation and the MID proteins determine the minus mating type[226]. The MID gene is also one of the determinants of oogamy in this group[228,231]. *Ectocarpus* gametes are isogamous, but the NIN-like family might have a role in mating type determination in this species or a related role during the life cycle. It is therefore interesting to note that the single member of this family (indicated with an asterisk in Supplementary Fig. 47) for which transcriptome-based microarray data is available is slightly, but significantly ($p \leq 0.05$), down-regulated in two mutant lines that carry the *immediate upright* mutation[232], which partially converts the sporophyte into a gametophyte (unpublished results).

C2H2 zinc finger proteins. The *Ectocarpus* genome encodes 52 C2H2 zinc finger proteins. An analysis of the domains present in these proteins indicated that they are not all transcription factors but are involved in a broad range of cellular processes including transcription regulation, splicing, polyubiquitination, deubiquitinylation, RNA binding, DNA polymerisation, proteolysis and protein dephosphorylation. The genes harbouring these domains show a low level of primary sequence conservation and thus do not usually group within gene families, although three families of nine, five and two genes were identified. The family of five genes encodes identical proteins and similarities between the upstream regions of these genes indicate that they were generated by segmental duplications.

bZIP transcription factors. The *Ectocarpus* genome contains 16 typical bZIP proteins and five atypical bZIP proteins (Supplementary Fig. 48). In a phylogenetic analysis, the majority of these bZIPs clusters together with the bZIPs of unicellular origin. Searches with custom-made HMMs allowed the identification of 6 bZIPs that were not identified by the global screens for TAPs. This study highlighted the potential benefits of developing custom-made search methodologies for the characterisation of gene families in lineages such as the stramenopiles, which are very distant phylogenetically from the majority of species currently represented in the databases.

2.3.2. Protein kinases

Many intracellular processes are regulated, at least in part, through the reversible addition of molecular switches, such as ubiquitin and SUMO proteins, to particular enzymes. Protein phosphorylation/dephosphorylation is one of the better characterised of these switch systems; the phosphoryl group, PO_3^-, may be added to a variety of target proteins by specific protein kinases and may be removed by a smaller number of less specific protein phosphatases. In eukaryotes, much of this protein phosphorylation is carried out by members of the eukaryotic protein kinase (ePK) superfamily, which share a conserved catalytic core[233]. ePK homologs are expected to be present in all eukaryotes and there is much experimental evidence for their operation in brown seaweeds[234-236].

We identified ePKs in the *Ectocarpus* genome in two ways. First, by PFAM, IPR and keyword searches of the gene models predicted by Eugène[237]. Second, and independently, by comparing the raw genome sequence against an ePK-specific multi-level profile Hidden Markov Model library[238]; this second method also provided an initial division of *Ectocarpus* ePKs into families. The resulting candidate ePKs were manually aligned against human[239] and *Arabidopsis*[240] datasets using Jalview 2.4[241] and errors were rectified if corrections were supported by the *Ectocarpus* EST library. The evolutionary histories of genes of interest were further inferred from phylogenetic analyses in *MEGA*4[242] using either Neighbor-Joining[243] or Maximum Parsimony[244], both with bootstrap support[245], and with all positions containing gaps having been eliminated from the dataset.

Together, these approaches identified 258 ePKs (Supplementary Table 30), corresponding to approximately 1.6 % of the proteins encoded by the *Ectocarpus* genome, a relatively high figure for a gene superfamily, but one which is consistent with genome counts in closely related diatoms[246] and with the 1.5-2.5 % predicted in other eukaryotes[247]. The ePK superfamily is divided into a number of families, according to the classification scheme first proposed by Hanks and Hunter in 1988[233] and most significantly emended by Manning and coworkers in 2002[239]. Not all of these families are to be expected in every eukaryote – the RGC family, for example, is found only in animals[239] – but all the expected ePK families are present in the *Ectocarpus* genome in the numbers given in Supplementary Table 30.

The resolution of these ePK families into sub-families will be presented in a later paper (D.M.-S., J.H.B., in preparation). Here we will focus on features of the kinase

superfamily that may be linked to the evolution of multicellularity in the brown algae. The Dollo analysis[88] of loss and gain of gene families (Section 2.1.10) identified a number of kinase gene families that are predicted to have evolved since divergence from the diatom lineage. Of the 700 gene families predicted to have been gained since divergence from the diatom lineage, 33 families (4.7% of the total gain) had gene ontology labels indicating protein kinase activity. These 33 families contained 100 genes, of which – after manual inspection to remove split genes and incomplete loci - 27 families were selected as potential ePKs.

Two of these 27 kinase families are of particular interest. First, *Ectocarpus* possesses a number of 'basal' ePK gene loci relative to the diatoms. That is to say, the evolutionary histories of several gene loci identified by the Dollo analysis place them near to the root of the ePK tree. It should be noted that some of these loci lack the conserved residues commonly associated with active ePKs[248]; while this makes their identification as ePKs uncertain, it does not rule out possible ePK activity because the traditional ePK conserved residues were identified primarily in animal ePKs[248]. We advance no adaptive explanations for these basal ePKs.

Second, *Ectocarpus* has a small group of 11 ePKs which strongly resemble the membrane-spanning receptor protein kinases seen in animals and plants. These consist of four domains: a) an extracellular signal peptide, identified by the HECTAR programme[249], b) several extracellular Leucine Rich Repeat (LRR) domains, of the sort traditionally associated with protein-protein interactions[250,251], c) a transmembrane domain, identified by TOPCONS[252], and d) a intracellular serine/threonine ePK domain. These LRR-PKs are members of a larger clade of 17 ePKs, some of which possess neither LRR nor transmembrane domains. The distribution of LRR-PKs within this clade suggests that the last common ancestor of the clade was an LRR-PK and that subsequent branches have lost the LRR domains.

Animal tyrosine and green plant serine/threonine receptor kinases form two separate monophyletic clades, indicating that these two families evolved independently[240], and the emergence of both of these families may be associated with the evolution of multicellularity[253,254]. In this respect, it is interesting that the *Ectocarpus* receptor kinases also form a monophyletic clade - separate from animal and green plant receptor kinases - indicating that the brown algal family of LRR-PKs also evolved independently. The evolution of membrane-spanning receptor kinases may, therefore, have been a key step in the evolution

of complex multicellularity in at least three of the five groups that have attained this level of developmental sophistication. No homologues of the *Ectocarpus* receptor kinases were found in the unicellular diatoms[246], but a detailed analysis of two complete oomycete genome sequences[255] identified another family of LRR-PK-like kinases. Bearing in mind the evolutionary proximity of the brown algal and oomycete lineages, it is possible that the receptor kinases found in the two groups families are derived from a common ancestral gene. However, phylogenetic analyses of the kinase domains of these proteins did not provide any support for monophyly and we are unable to rule out the independent evolution of *Ectocarpus* and oomycete gene families from cytosolic kinases. In the latter case, of course, independent families of receptor kinases could, again, represent two independent transitions to multicellularity, in the Phaeophyta and oomycetes.

2.3.3. Cell cycle genes

Genes controlling cell division in the *Ectocarpus* genome were annotated using a core cell cycle gene dataset from various organisms including metazoans, green plants and green and red algae. Overall *Ectocarpus* possesses a similar set of core cell cycle genes to the two diatoms *T. pseudonana* and *P. tricornutum*, including two cyclin-dependant kinase B (CDKB) homologues, a gene that has been described only in the green lineage (including both Streptophyta and Chlorophyta) and is absent in both Metazoans and Fungi. Both *Ectocarpus* and the two diatoms possess nearly all the cyclins found in flowering plants (cyclins A, B, D, H and T), with only cyclinD being absent. The kinase WEE1 and the phosphatase CDC25 are key, antagonistic enzymes that regulate M phase entry. *Ectocarpus, T. pseudonana* and *P. tricornutum* lack CDC25, a feature they share with all the members of the green lineage analysed so far, with the exception of Prasinophytes. Presumably, another phosphatase carries out this key function in these species. Overall, therefore, *Ectocarpus* has a set of cell cycle regulatory genes that is more similar to that of green plants than that of animals.

2.3.4. TOR kinase pathway

The conserved eukaryotic target of rapamycin (TOR) kinase is an important regulator of cell growth in both animals and green plants. *Ectocarpus* has two TOR kinases plus associated proteins such as Raptor and FKBP12. However, it lacks the Ras-related GTPase Rheb, which

is a crucial component of one pathway that activates TOR in animals, and this is all the more surprising because *Ectocarpus* has homologues of two RAG GTPases (Esi0006_0074 and Esi0351_0017) that promote the intracellular localisation of TOR to a compartment that also contains its activator Rheb in animals[256]. The Rheb GTPase is not present in the green lineage (land plants, *Chlamydomonas reinhardtii*, *Volvox carteri*, *Ostreococcus* spp., *Micromonas* spp.) but has been found in all stramenopile genomes sequenced so far apart from *Ectocarpus* (see Supplementary Table 31).

2.3.5. Putative membrane-localised receptors

The *Ectocarpus* genome encodes three putative G-protein coupled receptors (Esi0028_0043, Esi0192_0045, Esi0104_0004) plus three additional proteins that show similarity to GPCRs but are predicted to have only between three and six transmembrane domains (Esi0044_0105, Esi0123_0056, Esi0165_0058). GPCRs are also found in diatoms and oomycetes. Accordingly, these organisms also display subunits of the associated heterotrimeric G-proteins. The presence in *Ectocarpus* of six distinct paralogs of the Gα subunit (see Supplementary Table 31) compared to only one gene in *Phytophthora* and one (*Phaeodactylum*) or two (*Thalassiosira*) in diatoms may indicate an increased complexity of G-protein signalling associated with multicellularity in the former species. Interestingly, proteins with the domain "Regulator of G protein signalling superfamily" (IPR016137) were significantly more abundant in the *Ectocarpus* genome than in a broad range of other genomes (see section 2.1.8.).

The genome also contains three histidine kinases with predicted transmembrane domains and N-terminal Mase or Chase sensor domains (Esi0034_0049, Esi0197_0009, Esi0008_0194). Interestingly, one of these proteins (Esi0008_0194) is predicted to have seven transmembrane domains and seems to be a GPCR-histidine kinase fusion protein. No homologues of any of these three putative histidine kinase receptors were found in other chromalveolate genomes. The *Ectocarpus* genome also contains two response receiver genes (Esi0005_0242, Esi0038_0039) and an Hpt domain protein (Esi0172_0062).

2.3.6. Photoreceptors

Ectocarpus has a similar array of photoreceptors to diatoms, although in general the former has a greater number of genes encoding each type of receptor (Supplementary Table 32). The *Ectocarpus* genome encodes three cryptochromes, including a (6-4) family ("animal type") cryptochrome and two Cry-DASH genes, five aureochromes and three phytochromes. No phototropin genes were found, consistent with the current view that aureochromes are the stramenopile equivalents of these blue light receptors[257].

2.3.7. P-loop GTPases

Essentially, all important cellular processes (splicing, ribosome assembly, translation, protein and membrane transport, cytoskeletal dynamics, assembly of the flagellum, mitosis and cytokinesis, various signalling cascades) include as crucial regulatory factors representatives of TRAFAC GTPases, a vast, distinct class of so-called P-loop proteins (characterised by "Walker A" motif implicated in binding nucleotides, mostly ATP or GTP[258]). To get an overview of the complexity and diversity of cellular processes in *Ectocarpus*, we compiled an inventory of its TRAFAC GTPases and compared it to four other stramenopiles (the pelagophyte *Aureococcus anophagefferens*, the diatoms *T. pseudonana* and *P. tricornutum* and the oomycete *P. sojae*) and other reference species (the rhodophyte *Cyanidioschyzon merolae*, the flowering plant *A. thaliana*, the metazoan *Homo sapiens*, the yeast *S. cerevisiae*, and the amoeba *Dictyostelium discoideum*) (Supplementary Table 31).

The most conserved core of the TRAFAC class comprises proteins involved in ribosome assembly and export from the nucleus and in translation (initiation, elongation, termination and mRNA decay). *Ectocarpus* codes for all expected GTPases serving these roles in the nucleus/cytoplasm, in addition to equivalent cohorts of GTPases mediating these functions in the mitochondrion and the plastid. The cytosolic translation factors notably include SelB, a factor specific for co-translational incorporation of selenocystein in nascent proteins (see section 2.2.13.). The set of plastid-targeted GTPases in *Ectocarpus* has two interesting features. First, in common with other plastid-bearing stramenopiles (ochrophytes) but unlike land plants or the red alga *Cyanidioschyzon*, it includes the eubacterial-like translation release factor 3 (PRF3). Second, *Ectocarpus* and other stramenopiles encode a homolog of the *Arabidopsis* plastid-targeted GTPase PDE318 (pigment-defective 318), which in ochrophytes appears to be also plastid-targeted, whereas in *Phytophthora* it is presumably

targeted to the mitochondrion. At least one GTPase (a homologue of bacterial TrmE, which is involved in an essential tRNA modification) may exhibit dual organellar targeting; depending on the initiation codon used, the protein appears to bear either a bi-partite plastid targeting leader or a mitochondrial transit peptide, but these predictions need to be verified experimentally. *Ectocarpus* also encodes a homolog of Lrc5 (also known as YqeH), which was reported to be an NO synthase in *Arabidopsis*. Later investigations ruled this out, however, and indicated instead that it was an organellar GTPase involved in ribosome maturation[259]. Interestingly, an YqeH homolog in *Phaeodactylum* was reported to be plastid-targeted and to be implicated in regulation of NO signalling[260], but in light of more recent findings, the seeming implication of YqeH in NO signalling may be an indirect consequence of perturbation of the organelle. The *Ectocarpus* Lrc5/YqeH appears to have a mitochondrial transit peptide and may thus be a mitochondrial protein, in contrast to the diatom homologue.

The superfamily of dynamin-related proteins comprises GTPases generally implicated in modelling and constriction of membrane structures. The superfamily is represented by a number of proteins in *Ectocarpus*, including homologues of Dnm1 and ARC5, which are implicated in the division of mitochondria and plastids, respectively, but are not true dynamins. Another member, an orthologue of RME-1/EHD (actually a GTPase-like ATPase), which acts in the endosomal pathway, has an EH domain at the C-terminus as in Metazoa, rather than at the N-terminus as in plants. Four additional RME-1/EHD-like proteins are also encoded by *Ectocarpus*, but they lack an EH domain and may exhibit novel domain architectures, including a C-terminal extension with a trans-membrane segment. Further members of the dynamin superfamily in *Ectocarpus* include three homologues of the metazoan Mx proteins and five homologues of GBP (guanylate-binding) proteins. In Metazoa these families are involved in cellular defence against some viruses and bacteria[261]. Finally, *Ectocarpus* possesses the dynamin-related GTPases Atlastins, recently demonstrated to be responsible for homotypic fusion of ER membranes and the generation of the tubular ER network[262,263]. Interestingly, Atlastins have been characterised so far known only in Metazoa and the Atlastin function has been thought to be replaced in other eukaryotes by the distantly related GTPase RHD3/Sey1, which also has homologues in *Ectocarpus*. This species is therefore the first organism known to possess both GTPases, raising the question as to what extent their function is actually redundant.

The *Ectocarpus* genome contains a very interesting family of septin GTPase genes. Septins were originally believed to be restricted to metazoa and fungi, where they play a

critical role in cytokinesis and exocytosis, but recently septin homologs have been found in green algae and ciliates[264]. The *Ectocarpus* genome has nine septin-related loci, but interestingly, at least five of them are apparent pseudogenes with open-reading frames disrupted by frame-shifts, stop-codons, and/or deletions. The four intact *Ectocarpus* septins resemble the green algal and ciliate septins in that they are predicted to possess a C-terminal membrane-spanning segment (in contrast to septins from opisthokonts). The *Tetrahymena* septins localise to the outer mitochondrial membrane, raising the possibility that the *Ectocarpus* septins also have mitochondrial rather than cytokinetic or exocytic function.

The last major subgroup of TRAFAC GTPases is the Ras superfamily (sometimes called "small" G-proteins). An important group here is the RAB family, comprising GTPases regulating vesicle transport in the secretory and endocytic pathways. The *Ectocarpus* RAB family is only half as big as that of flowering plants, but is more diversified in terms of distinct RAB types. Our analysis also uncovered a novel large protein (Esi0102_0013) with a RAB-like domain, a coiled-coil region, and an SH2 domain, which apparently belongs to the ancestral stramenopile toolkit (it is shared with oomycetes) but has been lost from diatoms and *Aureococcus*. A notable portion of the Ras superfamily in *Ectocarpus* is represented by homologues of GTPases with flagellum-associated functions, including IFT27, FAP9/RABL5, ARL3, ARL6, ARL13, D2LIC and RJL, presumably active in zoospores and gametes. The RHO family forms in fungi, land plants, and especially in amoebae and Metazoa an extended set of GTPases with functions generally centred around the regulation of actin dynamics, cell polarity, and morphogenesis. In contrast, the RHO-based signalling in *Ectocarpus* is rather simple (Supplementary Table 33), with only a single RHO protein most closely resembling the opisthokont RAC and plant ROP (the prototypical eukaryotic types within the family). *Ectocarpus* also encodes four RhoGAP proteins (negative regulators of RHO) and four RhoGEF proteins (positive regulators), the latter being of the type present in Metazoa but absent from plants. While diatoms completely lack these RHO signalling devices, *Aureococcus* and *Phytophthora* exhibit this system at a complexity comparable to that in *Ectocarpus*, indicating that the origin of multicellularity in the brown algal lineage was not accompanied by expansion of RHO signalling (as it was in the metazoan lineage[265]). Metazoans are also characterised by an expanded set of RAS family paralogues with highly differentiated roles in the development and functioning of the multicellular body[266]. However, *Ectocarpus*, like flowering plants, lacks both the RAS family and associated regulators such

as proteins with the RasGEF, RasGAP and RapGAP domains, indicating that the RAS pathway is not essential for the evolution of complex multicellularity.

The most interesting feature of the *Ectocarpus* GTPase set is a greatly expanded ROCO family, which is generally defined by a conserved core consisting of the ROC GTPase domain fused to the family-specific COR domain. In different ROCO proteins this core is decorated by additional domains added to the N- and/or C-terminus[267]. The *Ectocarpus* genome contains at least 39 ROCO-related loci (Supplementary Table 31), but almost half of them are fragmentary and/or disrupted by in-frame stop-codons and frame-shifts. The intact ROCO genes encode proteins with a conserved structure with a variable number of leucine-rich repeats (LRRs) N-terminal to the GTPase domain. The specific cellular function of ROCO proteins in general is poorly characterized and may vary between species[267], so the role of the ROCO family in *Ectocarpus* cannot be predicted *per analogiam*. However, the apparently high evolutionary turnover of the ROCO genes and the presence of LRRs known to serve as flexible devices for interactions with other proteins suggest that the *Ectocarpus* ROCO proteins may be a part of anti-pathogen defence system (see also section 2.3.8).

Overall, the repertoire of TRAFAC GTPases in *Ectocarpus* is similar to those of other stramenopiles, but there are several notable differences (Supplementary Table 31). One trivial difference is that *Phytophthora* lacks GTPases that are predicted to function within the plastid in ochrophytes (with the exception of PDE318, see above). Compared to other ochrophytes, (especially diatoms), *Ectocarpus* has retained a less reduced set of putatively ancestral GTPases (most of which have also persisted in *Phytophthora*). Nevertheless, *Ectocarpus* appears to lack secondarily at least nine GTPase ancestrally present in other stramenopiles (Supplementary Table 34). Three of them may have been lost early in the ochrophyte lineage, whereas the remaining six have been retained by at least some ochrophytes. Among these, the most striking is the absence of Rheb, a widely conserved Ras-like GTPase (present in all the other stramenopiles analysed), which is known to regulate the TOR kinase signalling pathway in Metazoa (see also section 2.3.4).

2.3.8. Defence signaling and apoptosis

Our knowledge of immune systems is mostly restricted to plants and animals, which diverged from the stramenopile lineage, and from each other, during the eukaryotic crown radiation[268]. Despite the large phylogenetic distance separating plants and animals, their immune systems

share striking similarities at the molecular level, in particular in terms of pathogen receptors and signalling components[269]) and in the fact that innate immune responses commonly involve cell death programs such as the hypersensitive response in plants and apoptosis in animals.

E. siliculosus is the host for several pathogens including viruses, the oomycete *Eurychasma dicksonii*, the chytrid *Chytridium polysiphoniae*, the hyphochytrid *Anisolpidium ectocarpii* and the plasmodiophorid *Maullinia ectocarpii*[270] and is expected to possess defences against these organisms. A search was carried out for potential components of such defence systems based both on the likelihood that they share similarities with the systems that have been found in plants and animals and on known features of brown algal defence responses, such as their unusual halogen metabolism[271] (see Section 2.2.10.).

Candidate pathogen receptors. Defence reactions are triggered by the recognition of invading pathogens. In plants and animals, this is performed by a wealth of receptors that share some similarities. Structurally, they are often modular proteins containing a hypervariable, ligand-binding domain (such as Leucine Rich Repeats, LRRs), coupled to more conserved domains involved in signal transduction (CARD, DAPIN, TIR, etc.)[272]. In taxa as diverse as cnidarians, insects, jawless fishes and plants, LRR-containing proteins play key roles in defence responses, by interacting directly with antigens[273-275], contributing to signal transduction[276], or by monitoring the integrity of the host transduction pathways[277,278].

No CARD, DAPIN or TIR domain-containing proteins were found in *Ectocarpus*. However, *Ectocarpus* possesses more than 250 LRR-domain-containing genes, often grouped in small clusters of closely related genes and probable pseudogenes (Supplementary Table 35). This organization in clusters, and the high proportion of pseudogenes, is reminiscent of the fast evolution patterns described for plant disease resistance genes[279], making them good candidate defence genes (A. Zambounis, M.E., C.M.M.G., in preparation). NB-ARC domains were identified at 15 *Ectocarpus* loci (and 5 probable pseudogenes). This motive is shared by plant and animal proteins involved in disease resistance and apoptosis, such as plant resistance genes and the mammal Apaf-1[280]. All of the *Ectocarpus* NB-ARC-domain-containing proteins also have a C-terminal tetratricopeptide region, which is potentially involved in protein-protein interactions.

Programmed cell death and autophagy. The *Ectocarpus* genome encodes four candidate metacaspases (MCPs; Supplementary Fig. 49a,b), which are potentially involved in programmed cell death. These genes form a distinct clade to the MCPs of *T. pseudonana* and

P. tricornutum[281] but nonetheless retain the histidine/cysteine catalytic diad and the aspartate residues putatively involved in substrate binding. Interestingly, *Ectocarpus* also contains a legumain / vacuolar processing enzyme homologue, suggesting the potential existence of a vacuolar-mediated cell death program, as has been described in plants[282] (Supplementary Fig. 49c). Ubiquitous stress regulators were found and annotated, in particular one homologue each of RaR1 and Sgt1[283,284]. We also annotated heat shock proteins, which are known to interact with the RAR1/SGT1 complex and which are strongly induced following stress in *Laminaria*[285] and during infection in oomycetes (e.g. PPAT5[286]). In total, four Hsp90 and 26 Hsp70 genes were identified, including one probable pseudogene.

The presence of the MCPs and the absence of other key apoptosis-related genes that are found in animals, such as members of the BCL2 family or of the NFκB transduction pathway, indicates that programmed cell death pathways in *Ectocarpus* are more similar to those of plants than to those found in animals.

Autophagy, which is the capacity of a cell to digest its own components, is a cellular process that is highly conserved among eukaryotes[287]. It is triggered by stresses such as starvation, and is associated with programmed cell death pathways in both plant and animal models[288,289]. In yeast, genetic screens have identified 27 key autophagy genes[290], a core subset of which appears to be well conserved among eukaryotes. The *Ectocarpus* repertoire of autophagy-related genes is highly similar to those found in other stramenopile genomes (Supplementary Table 36).

Pathogenesis-related (PR) proteins. Since the assignment of a role for PR proteins in plant defence almost three decades ago[291], 17 PR protein families (PR1-PR17) have been recognised[292]. By definition PR proteins are induced upon infection or attack and are not constitutively expressed. However, several studies have shown that PR proteins can also be induced during plant development, for example during seed germination, flowering or senescence (see review by [293]). Moreover, hormones such as salicylic acid, ethylene and jasmonate can mimic pathogen attacks leading to the induction of PR proteins. PR proteins are structurally diverse, and in many cases, the biochemical function remains to be elucidated.

The *Ectocarpus* genome encodes three proteins that are similar to plant PR1 genes (Esi0277_0030; Esi0277_0035 and Esi0277_0044), except that the *Ectocarpus* proteins are longer due to an N-terminal extension which contains a WSC domain (Supplementary Fig 50a). This combination of a WSC domain (see section 2.2.1.) and a PR1 / SCP domain has not been found in any other organism to date, including the stramenopiles for which complete

genome data is available. The N-terminal WSC domain might bind to carbohydrates, mediating adhesion and / or positioning the C-terminal PR1 domain relative to a target. Key structural amino acids are conserved between the *Ectocarpus* proteins (Supplementary Fig. 50b) and the Tomato PR1 protein P14a, for which secondary structure information is available[294]. This suggests that the four alpha-helix, four beta-sheet structure of P14a might be conserved in the *Ectocarpus* proteins. A second class of *Ectocarpus* PR1-related proteins has EGF and metalloprotease domains downstream of the WSC and PR1/SCP domain (Supplementary Fig. 50a). The metalloprotease domain of these *Ectocarpus* proteins corresponds to the MEROPS peptidase family M8, exhibiting the typical zinc-binding HEXXH motif described by Rawlings and Barrett[295]. Moreover, four of these proteins (Esil0013_0157, Esi0013_0159, Esi0013_0166, Esi0013_0168; all on supercontig 0013) also possess the conserved histidine and methionine residues, with a characteristic 11 residue interval, downstream of the HEXXH motif as described for the metalloprotease leishmanolysin[296]. The three histidines are involved in zinc binding and the glutamate has a catalytic role. One *Ectocarpus* PR1-like metalloprotease gene (Esi0013_0168) is down-regulated during hypoosmotic stress[297], indicating a potential biological function for this gene family.

PR2 proteins are β-1.3 glucanases and *Ectocarpus* possesses several proteins of this family (see section S2.2.1.). One protein (Esi 0128_0015) showed homology to the plant type member of PR2 protein (tobacco) and has been assigned to family GH17. Plant glucanases involved in the response to pathogen attack are members of the GH 17 family; however glucanases of this family are also involved in various other processes and therefore the exact function of the *Ectocarpus* proteins remains to be investigated. Searches for the chitinase and chitin-binding proteins of the PR3, PR4, PR8 and PR11 families (including tobacco chitinases class I, III, IV, V, P, Q and potato and cucumber genes, glycosylhydrolases of the families GH18 and GH19 and PR4 chitin-binding proteins) did not identify homologues, despite the fact that related sequences have been found in other stramenopiles (*Phytophthora* species and *T. pseudonana*). Only one *Ectocarpus* protein (Esi0153_0024) was found to contain a chitinase II / glycosyl hydrolase domain. This protein is most similar to a *P. sojae* protein (Physo1:130467) and to animal chitinase-domain containing proteins. The apparent absence of chitinases in the *Ectocarpus* genome is surprising given the abundance of marine chitin-containing fungi and chytrid pathogens that infect this alga.

Plant PR5 proteins have antifungal properties[298]. They are 16-26 kD in size and have been called thaumatin-like proteins based on their similarity to thaumatin, a sweet tasting protein of the West African shrub *Thaumatococcus daniellii*. The antifungal action of plant PR5 proteins is hypothesized to result from the permeabilisation of the membranes of microbial organisms. Four thaumatin-like proteins were identified in *Ectocarpus* including one viral protein (Esi0052_0204) homologous to EsV-1-169[112]. Two of these proteins (Esi0212_0047: Esi0533_0007) possess an N-terminal WSC domain (discussed in section 2.2.1.) whereas the third protein (Esi010_0021) only has the thaumatin domain. The protein encoded by Esi 0212_0047 also contains a galactose-binding like domain upstream of the WSC domain. Thaumatin-domain-containing proteins are also present in the genomes of *P. sojae* and *P. ramorum* but, as with the PR1 proteins described above, the association of WSC and thaumatin domains is unique to *Ectocarpus* and appears to have been an innovation of the brown algae. All three of the *Ectocarpus* proteins are most similar to the EsV-1 protein EsV-1-169 and a *Feldmannia* virus sequence but, again, neither of the viral proteins contains a N-terminal WSC domain.

In higher plants it has been shown that some of the proteinase inhibitors of the PR6 family are induced upon wounding and herbivore attack[299]. One putative protease inhibitor with similarity to coffee PR6 and barley chymotrypsin-subtilisin inhibitor was found in the *Ectocarpus* genome (Esi0079_0059). It was assigned to the MEROPS I13 family of protease inhibitors[300] which groups proteins that inhibit the serine proteases (e.g. chymotrypsin and subtilisin).

Blast searches with PR15 and PR16 sequences (the barley germin or oxalate oxidase sequence[301] and barley germin-like protein[302]) identified six matches in *Ectocarpus* (Esi0118_0001, Esi0118_0007, Esi0128_0002, Esi0128_0009, Esi0128_0014, Esi0731_0004). These six genes were most similar to *Physarum* spherulin 1B which plays an important role during encystment following abiotic stress[303] and which shares similarity with germins[304]. Manganese ion-binding domains $HXHX_4EXxH$ are present in the *Ectocarpus* homologues as first described for barley germin[305] which exhibits oxalate oxidase / superoxidase activity. However, germins, germin-like proteins and spherulins are grouped within the cupin superfamily which includes proteins with diverse functions[306-308] and it is difficult to assign an exact function to these *Ectocarpus* proteins based on sequence similarity alone.

Searches with representatives of the PR9 (tobacco lignin-forming peroxidase), PR10 (parsley 'ribonuclease-like'), PR12 (radish defensin), PR13 (*Arabidopsis* thionin), PR14 (barley lipid-transfer protein) and PR17 (tobacco PRp27) families did not identify any homologues in *Ectocarpus*.

2.3.9. Ion channels and Ca signalling

Ectocarpus possesses a large family of transient receptor potential (TRP) channels (Supplementary Table 37). These are six transmembrane pass proteins with varying degrees of sequence homology that are generally permeable to cations, including Ca^{2+}. TRP channels can be activated by various stimuli including chemicals, temperature, mechanical stress and light, allowing cells to respond directly to the environment[309]. Animal TRP channels are classified into 2 main groups containing seven subfamilies dependant on their sequence, structure, function and mode of action. An eighth subfamily (TRPY) consists of yeast TRPs, which are distantly related to mammalian group 1 and 2 proteins. As with the yeast channels, the divergence of the *Ectocarpus* TRP sequences relative to the mammalian TRP sequences precludes their being assigned to specific groups. However, twelve of the *Ectocarpus* sequences have N-terminal ankyrin repeat domains, a feature generally associated with group 1 mammalian TRP channels (TRPC, TRPV, TRPA and TRPN). One of these sequences (Esi0124_0018) also possesses a carboxy-terminally located calmodulin binding domain. The absence of an N-terminal ankyrin region in the six other *Ectocarpus* TRP channel sequences is a feature that is generally associated with the group 2 channels TRPP and TRPML and three of these genes (Esi0255_0038, Esi0255_0034 and Esi0049_0030) are most similar to the mammalian group 2 TRPP subfamily protein polycystin kidney disease protein 2 (PKD2). Twelve of the *Ectocarpus* TRP genes (eight with an ankyrin domain and four without) contained a PTHR13800 domain, which is found in the non-ankyrin repeat containing mammalian group 2 TRP subfamily M. The PTHR13800 domain was absent from all of the TRP channels encoded by the published diatom and oomycete genomes, indicating an expansion of these TRP channel genes since divergence from the diatom lineage and loss from the sequenced diatom and oomycete genomes. The increased TRP channel complement in *Ectocarpus* could indicate a role for these proteins during the evolution of muticellularity or as an adaptation to the intertidal environment. The known functions of TRP channels in other species as environmental sensors is consistent with these hypotheses.

Three genes encoding bacterial-type, small conductance mechanosensitive (MscS) channel are present in the *Ectocarpus* genome. These channels are involved in osmoregulation in *E. coli*[310], chloroplast development in *Chlamydomonas* and *Arabidopsis*[311,312], and root hair mechanosensation in *Arabidopsis*[313], and could be involved in environmental sensing in brown algae, particularly in relation to osmosensing in the intertidal environment[314]. The tandemly repeated genes Esi0032_0081 and Esi0032_00822 are most similar to an *Arabidopsis* MscS domain containing protein, Y5208_ARATH. The third *Ectocarpus* MscS channel is most similar to the yeast MscS yna1 channel. A survey of the sequenced stramenopile genomes indicates that the diatom genomes contain more MscS channels than *Ectocarpus* (Supplementary Table 38), arguing against a role for these channels in the evolution of multicellularity. No sequences representing the large conductance mechanosensitive channels were identified in the *Ectocarpus* genome. *Ectocarpus* has four four-domain voltage gated calcium channels (VGCs). The VGC genes (Esi0029_0192, Esi0162_0059, Esi0206_0044 and Esi0234_0035) are not obviously related and reside at distinct loci. Analysis of the selectivity filter in the pore region of the protein[315] encoded by these loci suggests that they are calcium selective (Data not shown). Sequences similar to two-pore voltage gated calcium channels (TPCs) were identified at loci Esi0009_0041 and Esi0015_0142. Sequences corresponding to both the the two-pore and the four-pore VGC channels are more abundant in the *Ectocarpus* genome that the genomes of the other sequenced stramenopiles and could this have been important for the development of multicellularity in the brown algae. VGCs and TPCs are absent from plant genomes.

Esi0171_0053 encodes an inositol triphosphate (InsP3)/ryanodine type receptor (IP3R/RyR). Sequence analysis of this protein did not reveal to which of the two classes of receptor this sequence belongs and it may represent an ancestral form with features of both the IP3 and RYR receptors. However inositol triphosphate, the ligand for the animal IP3R, has been demonstrated to effect calcium release in the embryo of the brown alga *Fucus serratus*[315,316]. The absence of an homologous IP3R/RyR type receptor in the diatom or oomycete genomes suggests an important role for this class of protein in the brown algae.

Ionotropic Glutamate Receptors (iGR) function in animals as glutamate gated non-selective cation channels involved in cell-cell communication in the nervous system. Land plant genomes encode many Ionotropic Glutamate Like receptors (iGLR), although information concerning their function and gating is limited[317]. Two iGLR sequences with 80% amino acid identity were found at adjacent loci in the *Ectocarpus* genome

(Esi1051_0056, Esi 0151_0062). Analysis of the other stramenopile genomes identified only one partial iGLR-like sequence in the diatom *T. pseudonana* (Supplementary Table 38). This indicates that, unlike plants and animals, there has not been an expansion of this signalling family in *Ectocarpus*, although the presence of two iGLR receptors in this species and the absence of homologues in other stramenopiles hints at an important function in this algae. No cyclic nucleotide gated ion channels were found in the *Ectocarpus* genome.

2.3.10. mRNA maturation

Ectocarpus genes are intron-rich, raising the possibility that intron splicing may be an important regulatory step during gene expression. The genes encoding components of the splicing machinery were therefore annotated in detail.

Over recent years, the Sm and Sm-like (Lsm) proteins have been found to be widespread in the eukaryotic, archaeal and bacterial kingdoms, and have emerged as important players in many aspects of RNA metabolism, including splicing, nuclear RNA processing and messenger RNA decay[318-323]. The diversification of the Lsm domain family has been studied[324] but specific studies of their evolution in the algal lineage have not been carried out.

Ectocarpus, like the other heterokont species for which the full genome has been sequenced, possesses a single gene for each component of the Sm ring (the SmD3BD1D2FEG). This structure binds to the major U1, U2, U4 and U5 snRNAs, and assists the splicing of GT-AG introns. Single genes genes encoding each of Lsm1 to Lsm8 were also identified in *Ectocarpus*. The heteroheptameric ring proteins Lsm1 to Lsm7 are involved in messenger RNA (mRNA) degradation in the cytoplasm, whereas the Lsm2 to Lsm8 form a ring that is associated with U6 snRNA and functions during general RNA maturation in the nucleus.

We also searched for members of the recently identified group of Lsm proteins with long C- terminal tails[325]. Orthologues of Lsm12 and Lsm14 were found in most of the heterokonts including *Ectocarpus* (but with the exception of *A. anophagefferens*), in apicomplexans, in rhodophytes, in the green lineage (chlorophytes, prasinophytes and a streptophyte) and in metazoans. The conservation of these Lsm proteins suggests that these interaction partners may play important roles in mRNA degradation and in the control of the mitotic G2/M phase.

Another combination of Sm proteins has been shown to be a component of the U7 snRNP, and to play an essential role in the maturation of the 3' ends of metazoan replication-dependent histone mRNAs[326,327]. The Lsm10 and Lsm11 proteins replace SmD1 and SmD2 in the Sm ring that specifically binds to the U7 snRNA[327]. The histone cleavage site is flanked by evolutionarily conserved sequences that interact with trans-acting processing factors. Upstream of the cleavage site is a highly conserved sequence encompassing a hairpin structure which is recognized by the hairpin binding protein (HBP)[328-330]. The human HBP contains a central RNA binding domain (RBD) that does not resemble other known RNA binding proteins[331,332]. The second conserved sequence in the histone 3' UTR, the purine-rich histone downstream element (HDE), lies several nucleotides downstream of the cleavage site and interacts by base pairing with U7 snRNA. These conserved sequences, which were identified in the 3' UTR of metazoans, have also been shown to be present in some green algae[333]. This suggests that, contrary to the situation in land plants where the histone mRNAs are polyadenylated[334], the maturation of histone mRNAs in algae share feature with those of metazoans.

To obtain a comprehensive view of histone mRNA maturation in algae, we specifically looked for the conserved palindromic sequence, U7 specific Lsm proteins and for proteins containing a RBD characteristic of the hairpin binding protein. *Lsm10, Lsm11* and *HBP* genes were found in several heterokonts including *Ectocarpus*, in chlorophytes, in prasinophytes and in one streptophyte (*Physcomitrella patens* ssp *patens*) (Supplementary Fig. 51). However, the Lsm10 protein seemed to be less conserved and no clear orthologues could be found in the two diatom genomes. In addition, we found that, like animal type-1 histone genes, the 3' regions of the majority of heterokont, green algal and *Physcomitrella patens* histone genes contain a highly conserved motif with 3' palindrome and a putative spacer element (Supplementary Fig. 52). Even if important questions remain (such as the replacement of LSm10 by another Sm protein, possibly SmD1), the presence of both trans-acting elements and associated factors in the U7 snRNP strongly suggests that in most algae the maturation of most histone mRNAs involves U7snRNP rather than polyadenylation. Because the mature 3' end is involved in controlling later steps of the RNA's life cycle, such as translation and mRNA degradation, we believe that aspects of histone RNA metabolism are important features for gene regulation and might have important consequences in cell cycle regulation and development in *Ectocarpus*.

A search for genes encoding the U11-U12 associated proteins (U11/U12-65, U11/U12-48, U11/U12-35, U11/U12-31, U11/U12-25, U11/U12-20) of the minor spliceosome, using the human orthologues as bait, failed to detect any matches. This suggests that splicing of U12-dependent introns does not occur in *Ectocarpus,* unlike *Phytophthora*[335].

2.3.11. mRNA translation

Translation is a complex and sophisticated process involving a large number of factors including about 20 initiation, elongation and termination factors, many of which are complexes of several polypeptides. Regulation of translation is mostly exerted at the level of initiation (reviewed in[336,337]). During this step, the cap structure of the mRNA interacts with the cap-binding protein eIF4E, which forms a complex with initiation factors eIF4A and eIF4G. This interaction of the cap-binding complex is stabilized through the interaction of eIF4G with the poly(A)-binding PABP bound to the poly(A) tail. A 43S complex containing eIF3, eIF2, methionyl-tRNA and the 40S ribosomal subunit is then recruited to the mRNA via interaction of the multi-subunit eIF3 with eIF4G and the complex then scans the mRNA to arrive at the initiation codon. Recognition of the initiation codon triggers eIF2-bound GTP hydrolysis, release of initiation factors and binding of the large ribosomal subunit (Supplementary Fig. 53). In *Ectocarpus*, most of the initiation factors, with the exception of eIF4E and eIF4G, are encoded by single genes. There genome encodes four eIF4E proteins including the canonical translation factor eIF4E1 (Esi0116_0065), a protein (Esi0010_0026) that is most similar to the plant-specific nCBP (novel cap binding protein, of unknown function) and two proteins which are most similar to animal-type eIF4E2 and eIF4E3 (Esi0009_0153 and Esi0120_0052). Two eIF4G proteins are encoded (Esi0304_0009 and Esi0711_0004). These multiple eIF4E and eIF4G genes suggest that there may be novel processes regulating translation initiation in *Ectocarpus*. As in Angiosperms, 4E-BP, the eIF4E-binding protein that plays an important role in the regulation of mammalian protein synthesis by sequestering eIF4E, is absent from the *Ectocarpus* genome. In metazoans, eIF2a phosphorylation is critical for the regulation of eIF2 activity. The phosphorylatable serine on eIF2a is conserved in *Ectocarpus* (Esi0216_0037) but, as in Angiosperms, there are no strong homologues of the kinases that regulate eIF2a in animals, although Esi0000_0166 exhibits about 30% identity in its kinase domain with the eIF2kinase GCN2.

Binding of the large subunit marks the transition to the elongation step, which is dependent on two elongation factors: eEF1, comprising two components (eEF1A and eEF1B), and eEF2. eEF1A is encoded by two tandem duplicated genes in reverse orientation (Esi 0387_0021 and 00387_0022, encoding identical proteins). The elongation factor specific for the incorporation of selenocystein into proteins (eEF-Sec/SelB) is encoded by Esi0151_0054. The termination step, during which the protein is released from the ribosome, involves the release factors eRF1 and eRF3. Genes encoding all of these elongation and release factors were identified in the *Ectocarpus* genome. Supplementary Table 39 lists all the translation factors involved in the cytoplasmic translation found in the *Ectocarpus* genome. Overall, the set of translation regulators in *Ectocarpus* resembles that of green plants more than that of animals.

2.3.12. Meiosis

As expected, given the sexual life cycle of *Ectocarpus*, most of the core meiotic genes conserved between plants, animals, and fungi were also identifiable in the *Ectocarpus* genome. Particularly noteworthy was the higher conservation of several gene families in *Ectocarpus* compared to the unicellular Heterokonts for which genomes are available, such as the diatoms *T. pseudonana* and *P. tricornutum*. *Ectocarpus* possesses homologues of six of the seven known members of the Rad51 family[338]. The *Ectocarpus* Rad51 homologues included a likely orthologue of DMC1, which is highly specific to meiotic recombination in both plants and animals[339,340]. Esi0034_0089 is considered a probable DMC1 ortholog because it is the closest reciprocal blast hit of well-studied yeast, animal, and plant DMC1 proteins. This assignment was also revealed in trees using alignments including only the *Ectocarpus* Rad51 homologues with the most confident gene models. However, the tight phylogenetic association with DMC1 was not evident in some trees made including all *Ectocarpus* and diatom Rad51 homologues because fewer sites could be included in the alignments. The poorly conserved XRCC2 gene was the only member of the Rad51 family that did not have a clear homolog in the genome of *Ectocarpus*. The diatom genomes each contain only three clear homologs of the Rad51 family and none of these genes appear to be homologues of DMC1. Although sex has not yet been documented in these two diatom species[341], sex is well described in other members of this group, including close relatives of both species. Similarly, the oomycetes are known to be sexual, yet in the genomes of these organisms only four putative Rad51 homologues could be identified. The similarity of these

genes to known Rad51 family members was weak and, consequently, confident assignment of orthology was not possible. Similarly, yeasts, which are unicellular Opisthokonts, retain only four Rad51-like proteins compared to the seven distinct members commonly found in multicellular animals (also Opisthokonts). This suggests that unicellular lineages may be more prone to loss or greater divergence of Rad51 family members than lineages that have adopted multicellularity, independently of whether meiosis has been retained.

In addition to the Rad51 family of nuclear recombination genes, both stramenopiles and land plants also contain homologues of the prokaryotic recA family of recombination proteins, which are ancestrally related to the more diverse Rad51 proteins of eukaryotes and to the RadA and RadB proteins of Archaea[338]. Eukaryotic recA proteins, which are not present in animals or fungi, are thought to have been acquired during the primary endosymbiosis events that produced chloroplasts and mitochondria. The genomes of both diatoms and non-photosynthetic *Phytophthora* species encode only a single recA, whereas the *Ectocarpus* genome contains two divergent recA genes (Supplementary Fig. 54). *recA1* is most closely related to the diatom recAs and groups with the chloroplast *recA* of the red algae *Cyanidioschyzon merolae*, the cyanobacteria, and chloroplast recAs of green lineage organisms. This is as predicted because stramenopile plastids originate from a red algal endosymbiont. The Hectar program predicts that recA1 is targeted to the plastid. The second recA gene, *recA2*, is highly similar to the *Phytophthora* recAs, which are orthologous to the mitochondrial recAs of land plants and the ameobozoid *D. discoideum*. Curiously, no recA homologues could be identified in the genome of another unicellular photosynthetic stramenopile, *A. anophageferens*. Together these data suggest that the ancestral stramenopile genome encoded a recA of mitochondrial origin and that a recA of chloroplast origin was subsequently acquired during secondary endosymbiosis. Single-celled lineages then differentially lost one or both of these recA genes. The dramatic reduction of the mitochondrial genome size in animals, fungi and prasinophytes compared to terrestrial plants has been hypothesized to result from loss of mitochondrial-targeted recA in the former lineages. However, the generally small mitochondrial genome sizes in stramenopiles, irrespective of whether a mitochondrial recA was retained, indicates that retention of this gene was not the factor that determined mitochondrial genome size in this group.

Ectocarpus contains two homologues of Spo11, a protein that creates double strand DNA breaks, initiating meiotic recombination[340,342,343]. A single Spo11 homologue is identifiable in fungi and animals but three homologues of Spo11 are identifiable in vascular

plants, including Spo11-1 and Spo11-2, which are specifically required for meiosis, and Spo11-3/Top6A, which functions with Top6B as a topoisomerase and is required for non-meiotic functions[344-346]. As with diatoms, the two *Ectocarpus* Spo11/Top6A homologues grouped respectively with Spo11-2 and the non-meiotic Top6A of *Arabidopsis* in a phylogenetic analysis. Both *Ectocarpus* and diatom genomes also contain a Top6B homologue but this gene was not identified in the *Phytophthora* genomes. Although neither Top6A nor Top6B are found in fungi or animals, these genes have been found in the *C. merolae* genome. This suggests that perhaps the ancestral eukaryote, or at least the ancestral bikont, contained topoisomerase 6 function. Top6A and Top6B are required for endoreduplication in vascular plants, in which the ploidy of somatic cells is increased leading to higher cell volumes. However, certain animal cells (e.g., megalokaryocytes) are capable of endoreduplication without Top6[347,348]. Top6A and Top6B presumably mediate an ancestrally shared mechanism for packaging and regulation of chromatin that has been lost by certain lineages.

In yeast and plants, chromosomal crossovers involve two basic pathways: Class I crossovers are facilitated by the Mer3 DNA helicase and are sensitive to interference whereas class II crossovers are facilitated by Mus81 and Mms4 and do not depend on Msh4 and Msh5[349,350]. Both the diatom and *Ectocarpus* genomes contain homologues of *Mer3* but neither contains an identifiable homologue of *Mms4*. Also, *Ectocarpus* and diatoms both contain genes encoding ERCC4-like proteins with limited similarity to the XPF/Mus81 family of nucleases but none were clear *Mus81* homologues. It is possible that *Mms4* and *Mus81* are too highly divergent in stramenopiles to permit their identification. Alternatively, stramenopiles may have only interference-sensitive, Class I-type crossovers.

2.3.13. Integrins

Integrins are a family of dimeric cell adhesion receptors composed of one α and one β subunit that are involved in the transmission of mechanical signals perceived at the cell surface to the cytoskeleton[351]. At least 18 α and eight β subunits have been identified in Human. Binding of integrin to extracellular matrix (ECM) proteins, such as collagen, fibronectin or vitronectin promotes cell spreading and activates signalling cascades critical for adhesion-dependent growth and survival. In opisthokont cells, adhesion results in the aggregation of integrin within structures called focal adhesion plaques, which form the link between the extracellular

space and the intracellular cytoskeleton[352]. Precursors of integrins were identified in the ascomycete *Candida albicans* and in the amoebae *D. discoideum* and were shown respectively to be necessary and sufficient for the filamentous growth of the fungi[353], and to be involved in cell adhesion and phagocytosis[354].

Three integrin alpha sub-units were identified in the *Ectocarpus* genome. The sequence similarity is limited to the N-terminal domains, with the presence of a 7-bladed beta-propellor domain and a specific integrin alpha domain[351], whereas the C-terminal ends are divergent. In animals, the N-terminus is extracellular and interacts with proteins such as collagen, fibronectin and vitronectin, whereas the C-terminus corresponds to the intracellular region of the protein and mediates the transmission of mechanical signals from the external medium to intracellular actors, allowing the cell responding to a mechanical stimulus. No collagen, fibronectin and vitronectin homologues were found in *Ectocarpus* but the genome does encode homologues of the intracellular integrin partners talin and α-actinin (but not vinculin) (Supplementary Table 40). In animals, talin, α-actinin and vinculin interact with actin microfilaments[355] and contribute, in conjunction with several kinase cascades (FAK, Rho GTPases et ERK[356]), to the transmission of the mechanical signal to the proteins WASP and WAVE/Scar (Wiskott-Aldrich syndrome protein (WASP) WASP and WASP-family verprolin-homologous protein (WAVE)[357,358]. These components of the cytoskeleton are discussed below.

2.3.14. Cytoskeleton

Cytokinesis in brown algae shares features with both green plant and animal cells; centrosomes function as microtubule organising centres (as in animal cells) but cytokinesis involves the formation of a structure resembling a cell plate, which extends out to the plasma membrane[359]. However, in contrast to the situation in green plants, no specialised phragmoplast is formed and this is consistent with the absence of genes encoding dynamin-related phragmoplastins in the *Ectocarpus* genome. On the other hand, *Ectocarpus* does have the centrosome-associated tubulins δ and ε in addition to α-, β- and γ-tubulin. No η and ζ tubulin genes were found. Note that δ and ε tubulins are ubiquitous among organisms with triplet microtubules (found in centrioles and flagellar basal bodies[360-363]). They are not present in diatoms, which have flagella that lack a central MT. Tubulins are related to FtsZ proteins, prokaryote cytoskeleton GTPases that are also found in eukaryotes where they are involved in

organelle division. There are also several *FtsZ* genes in *Ectocarpus*, putatively involved in both chloroplast and mitochondrial division. The chloroplast form of FtsZ is widespread in photosynthetic organisms, while the mitochondrial form has been found in *D. discoideum*, a glaucocystophyte *Cyanophora pardoxa*, a red alga *C. merolae*, an oomycete *Phytophthora infestans*, diatoms, haptophyte algae, *Mallomonas* and *Ectocarpus*, but not in groups of the green algae and plants, the Opisthokonts and the Apicomplexa[364].

Cytokinesis in brown algae is different to the equivalent process in land plants, involving a cortical actin network which plays an important role in determining the polarity of apical and axillary cell division and in cell wall morphogenesis[365]. The genome contains a single actin gene, plus the nucleation promoting factors formin[366] (1 formin gene plus 5 formin-like) and profilin. Most of the proteins involved in the depolymerisation or severing of the actin filaments are also conserved (two homologues of villin and three of fimbrin), and the actin molecular motor myosin is also present. *Ectocarpus,* like diatoms and oomycetes, does not possess WASP or WAVE/Scar proteins. In metazoan and land plants, WAVE/scar activates the reticulation of actin microfilaments through activation of the ARP2/3 complex[367]. ARP2/3-complex-mediated actin polymerization is crucial for the reorganization of the actin cytoskeleton at the cell cortex during processes such as cell movement, vesicular trafficking and pathogen infection. The ARP2/3 complex is composed of 7 sub-units, all of which are encoded in *Ectocarpus* genome. Evidence from studies in *Fucus* supports a role for the ARP2/3 complex in polarised growth of brown algae[368].

2.3.15. Vesicle trafficking

With some exceptions, the cellular trafficking machinery tends to be more complex in multicellular than in unicellular organisms[369]. Ultra-structure studies performed on vegetative *Ectocarpus* cells indicate that they have a very active intracellular trafficking system[370,371]. A chloroplastic endoplasmic reticulum surrounds the chloroplast envelope and is fused to the nuclear envelope, the whole system being in close vicinity to the Golgi apparatus. This compact network of vesicles could allow the direct transit of photosynthetates from the chloroplast to the Golgi apparatus. Moreover, numerous osmiophilic bodies have been observed, fusing with the cell membrane (reviewed in[270]). The complexity of the gene families predicted to be involved in vesicle trafficking in *Ectocarpus* is consistent with the presence of this very active trafficking system (Supplementary Table 41). For example, the genome encodes 20 SNARE proteins, which are involved in membrane fusion processes within the secretory pathway[372,373]. *Ectocarpus* also possesses significant numbers of coat

protein complex proteins[374], which are involved in vesicle formation and fission during trafficking between the endoplasmic reticulum and the Golgi apparatus or exo/endocytosis (23 COPI and COPII proteins and 5 clathrins).

2.3.16. Flagella

Ectocarpus gametes and spores bear the two heteromorphic flagella typical of stramenopiles, the anterior flagella bearing mastigonemes. Comparison with the known flagellar proteins in *Chlamydomonas*, indicated that the *Ectocarpus* genome contains a well-conserved set of flagellar-related genes, with some exceptions (Supplementary Fig. 55). Although *T. pseudonana* is also a stramenopile, its flagella seem to be highly different from those of *Ectocarpus.* This may reflect the unusual configuration of the flagellar apparatus in the diatoms, where there is a 9+0 microtubular configuration in the flagellar axoneme and a ring of nine doublets of microtubules in the basal body. With regard to the radial spoke protein (RSP) that regulates the pattern of flagella bending, *Ectocarpus* has only 6 of the 17 radial spoke proteins that have been found in *Chlamydomonas* (Supplementary Table 42). Moreover, most of the radial spoke proteins were not conserved between *T. pseudonana*, *M. pusilla* and *H. sapiens*, indicating that RSPs vary considerably between organisms that have different patterns of flagella bending. The function of intraflagellar transport (IFT) proteins in flagellar assembly and maintenance has been extensively investigated[375-377]. However, the molecular mechanism of IFT regulation is not clear. Recently, it has been postulated that a phosphoprotein component of the IFT particle complex B, IFT25, for which there is not a homologue in *Ectocarpus*, directly interacts with IFT27, and the association and disassociation between the sub-complex of IFT25 and IFT27 and complex B might be involved in the regulation of IFT[378]. Bardet-Biedl syndrome (BBS) genes encode proteins of unknown function that are located in the basal body and cilia in animal cells. The *Ectocarpus* genome encodes all the known BBS proteins that have been detected in *Chlamydomonas*. The *T. pseudonana* genome encodes only BBS13 and BBS14. The strong similarity between *Ectocarpus*, green algal and animal BBS sequences, suggests that these proteins may perform similar functions in organisms with cilia and flagella. With regard to the BBS proteins which were not found in *Chlamydomonas* (BBS6, 10, 11 and 12), the *Ectocarpus* genome possesses only BBS6. The lack of many BBS proteins in *T. pseudonana* may reflect the unusual configuration of the basal body. Orthologues of SF-assemblin, a component of system I fibers, which run parallel to flagellar root microtubules from the basal bodies[379], were also

found in the *Ectocarpus* genome. Recently, it was reported that SF-assemblin is widely distributed among the green algae, the alveolates and the stramenopiles[380]. The *Ectocarpus* genome also encodes orthologues of several proteins that have been found to be components of the stramenopile-specific tripartite tubular mastigonemes covering the anterior flagellum in the chrysophyte *Ochromonas danica*: Ocm1 (which is homologous to the sexually-induced SIG1 protein from the diatom *Thalassiosira weissflogii*), Ocm2, Ocm3 and Ocm4[381-383].

References

1 Peters, A. *et al.* Life-cycle-generation-specific developmental processes are modified in the immediate upright mutant of the brown alga *Ectocarpus siliculosus*. *Development* **135**, 1503-1512 (2008).

2 Apt, K., Clendennen, S., Powers, D. & Grossman, A. The gene family encoding the fucoxanthin chlorophyll proteins from the brown alga *Macrocystis pyrifera*. *Mol. Gen. Genet.* **246**, 455-464 (1995).

3 Monteuuis, O., Doulbeau, S. & Verdeil, J. L. DNA methylation in different origin clonal offspring from a mature *Sequoiadendron giganteum* genotype. *Trees - Structure and Function* **22**, 779-784 (2008).

4 Jaligot, E., Rival, A., Beulé, T., Dussert, S. & Verdeil, J.-L. Somaclonal variation in oil palm (*Elaeis guineensis* Jacq.): the DNA methylation hypothesis. *Plant Cell Reports* **7**, 684-690 (2000).

5 Jaffe, D. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-96 (2003).

6 Ewing, B., Hillier, L., Wendl, M. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175-185 (1998).

7 Pertea, G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652 (2003).

8 Kent, W. BLAT - the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

9 Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477-478 (1997).

10 Castelli, V. *et al.* Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res.* **14**, 406-413 (2004).

11 Stolc, V. *et al.* Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci. USA* **102**, 4453-4458 (2005).

12 Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).

13 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

14 Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, 166-175 (2005).

15 Bao, Z. & Eddy, S. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269-1276 (2002).

16 Edgar, R. & Myers, E. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152-158 (2005).
17 Huang, X. On global sequence alignment. *Comput. Appl. Biosci.* **10**, 227-235 (1994).
18 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467 (2005).
19 Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker.* <http://repeatmasker.org>
20 Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119-121 (1996).
21 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
22 Kolpakov, R., Bana, G. & Kucherov, G. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **31**, 3672-3678 (2003).
23 Foissac, S. *et al.* Genome Annotation in Plants and Fungi: EuGene as a model platform. *Curr. Bioinfor.* **3**, 87-97 (2008).
24 Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332-1338 (2005).
25 Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154-159 (2005).
26 Tyler, B. *et al. Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-1266 (2006).
27 Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244 (2008).
28 Armbrust, E. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86 (2004).
29 Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inform. Software Tech.* **47**, 965-978 (2005).
30 Gschloessl, B., Guermeur, Y. & Cock, J. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* **9**, 393 (2008).
31 Simillion, C., Janssens, K., Sterck, L. & Van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24**, 127-128 (2008).
32 Rensing, S. *et al.* An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* **7**, 130 (2007).
33 Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P. & Van de Peer, Y. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.* **18**, 310-323 (2008).
34 Crooks, G., Hon, G., Chandonia, J. & Brenner, S. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188-1190 (2004).
35 Mignone, F. *et al.* UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **33**, D141-146 (2005).
36 Thomas-Chollier, M. *et al.* RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* **36**, W119-127 (2008).
37 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955-964 (1997).
38 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100-3108 (2007).

39 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686-689 (2005).

40 Kasschau, K. D. *et al.* Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol* **5**, e57 (2007).

41 Hofacker, I. *et al.* Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.* **125**, 167-188 (1994).

42 Meyers, B. C. *et al.* Criteria for annotation of plant MicroRNAs. *Plant Cell* **20**, 3186-3190 (2008).

43 Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121**, 207-221 (2005).

44 R, Development, Core & Team. (2009).

45 van Dongen, S. *Graph Clustering by Flow Simulation.* University of Utrecht, (2000).

46 Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

47 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98 (1999).

48 Van de Peer, Y. & De Wachter, R. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.* **10**, 569-570 (1994).

49 Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* **266**, 418-427 (1996).

50 Farris, J. Phylogenetic analysis under Dollo's law. *Syst. Zool.* **26**, 77–88 (1977).

51 Harris, M. Developing an ontology. *Methods Mol Biol* **452**, 111-124 (2008).

52 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005).

53 Storey, J. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445 (2003).

54 Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-208 (2002).

55 Yi, G., Sze, S. & Thon, M. Identifying clusters of functionally related genes in genomes. *Bioinformatics* **23**, 1053-1060 (2007).

56 Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).

57 Hanekamp, K., Bohnebeck, U., Beszteri, B. & Valentin, K. PhyloGena - a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics* **23**, 793-801 (2007).

58 Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18**, 1546-1547 (2002).

59 Moustafa, A. & Bhattacharya, D. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. *BMC Evol. Biol.* **8**, 6 (2008).

60 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).

61 Guo, A. Y. *et al.* PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.* **36**, D966-969 (2008).

62 Riano-Pachon, D. M., Ruzicic, S., Dreyer, I. & Mueller-Roeber, B. PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* **8**, 42 (2007).

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

64 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281-288 (2008).

65 Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85-94 (1999).

66 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518 (2005).

67 Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427 (2004).

68 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Soc. Series B-Method.* **57**, 289-300 (1995).

69 Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, D229-232 (2009).

70 Finn, R. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281-288 (2008).

71 Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* **2**, 953-971 (2007).

72 Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36**, W465-469 (2008).

73 Müller, D. G. Untersuchungen zur Entwicklungsgeschichte der Braunalge *Ectocarpus siliculosus* aus Neapel. *Planta* **68** (1966).

74 Müller, D. G. Generationswechsel, Kernphasenwechsel und Sexualität der Braunalge *Ectocarpus siliculosus* im Kulturversuch. *Planta* **75**, 39-54 (1967).

75 Hurst, L., Williams, E. & Pál, C. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet*. **18**, 604-606 (2002).

76 Kruglyak, S. & Tang, H. Regulation of adjacent yeast genes. *Trends Genet.* **16**, 109-111 (2000).

77 Kensche, P., Oti, M., Dutilh, B. & Huynen, M. Conservation of divergent transcription in fungi. *Trends Genet.* **24**, 207-211 (2008).

78 Trinklein, N. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62-66 (2004).

79 Dittami, S. *et al.* Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biol.* **10**, R66 (2009).

80 Shen, Y. *et al.* Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res.* **36**, 3150-3161 (2008).

81 Tian, B., Hu, J., Zhang, H. & Lutz, C. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201-212 (2005).

82 Li, Q. & Hunt, A. The Polyadenylation of RNA in Plants. *Plant Physiol.* **115**, 321-325 (1997).

83 Gilmartin, G. Eukaryotic mRNA 3' processing: a common means to different ends. *Genes Dev.* **19**, 2517-2521 (2005).

84 Samanta, M. *et al.* The transcriptome of the sea urchin embryo. *Science* **314**, 960-962 (2006).

85 Wang, B., O'Toole, M., Brendel, V. & Young, N. Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol.* **8**, 17 (2008).

86 Hazkani-Covo, E., Levanon, E., Rotman, G., Graur, D. & Novik, A. Evolution of multicellularity in Metazoa: comparative analysis of the subcellular localization of proteins in *Saccharomyces, Drosophila* and *Caenorhabditis*. *Cell Biol. Int.* **28**, 171-178 (2004).

87 Gerstein, M. & Levitt, M. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* **94**, 11911-11916 (1997).

88 Martens, C., Vandepoele, K. & Van de Peer, Y. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. USA* **105**, 3427-3432 (2008).

89 Guillou, L. *et al. Bolidomonas*: a new genus with two species belonging to a new algal class, the Bolidophyceae (Heterokonta). *J. Phycol.* **35**, 368–381 (1999).

90 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

91 Capy, P. *et al.* Sexual isolation of genetically differentiated sympatric populations of *Drosophila melanogaster* in Brazzaville, Congo: the first step towards speciation? *Heredity* **84 (Pt 4)**, 468-475 (2000).

92 Elder, R., St John, T., Stinchcomb, D., Davis, R. & Scherer, S. Studies on the transposable element Ty1 of yeast. I. RNA homologous to Ty1. II. Recombination and expression of Ty1 and adjacent sequences. *Cold Spring Harb. Symp. Quant. Biol.* **45**, 581-591 (1981).

93 Dupressoir, A. & Heidmann, T. Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. *Mol. Cell. Biol.* **16**, 4495-4503 (1996).

94 Malone, C. D. & Hannon, G. J. Small RNAs as guardians of the genome. *Cell* **136**, 656-668 (2009).

95 Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell* **136**, 669-687 (2009).

96 Carthew, R. W. & Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**, 642-655 (2009).

97 Mi, S. *et al.* Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5' terminal nucleotide. *Cell* **133**, 116-127 (2008).

98 Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977 (2000).

99 Hinas, A. *et al.* The small RNA repertoire of *Dictyostelium discoideum* and its regulation by components of the RNAi pathway. *Nucleic Acids Res.* **35**, 6714-6726 (2007).

100 Keeling, P. J. Diversity and evolutionary history of plastids and their hosts. *Am J Bot* **91**, 1481–1493 (2004).

101 Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724-1726 (2009).

102 Thompson, J., Higgins, D. & Gibson, T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).

103 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).

104 Müller, D. G., Kapp, M. & Knippers, R. Viruses in marine brown algae. *Adv. Virus Res.* **50**, 49-67 (1998).

105 Brautigam, M., Klein, M., Knippers, R. & Müller, D. G. Inheritance and meiotic elimination of a virus genome in the host *Ectocarpus siliculosus* (phaeophyceae). *J. Phycol.* **31**, 823-827 (1995).
106 Delaroque, N., Maier, I., Knippers, R. & Müller, D. G. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol.* **80**, 1367-1370 (1999).
107 Müller, D. G. Mendelian segregation of a virus genome during host meiosis in the marine brown alga *Ectocarpus siliculosus*. *J. Plant Physiol.* **137**, 739-743 (1991).
108 Müller, D. G. K., H. Stache, B. Lanka, S. A virus infection in the marine brown alga Ectocarpus siliculosus (Phaeophyceae). *Botanica Acta* **103**, 72-82 (1990).
109 Lee, A. M., Ivey, R. G. & Meints, R. H. Repetitive DNA insertion in a protein kinase ORF of a latent FSV (*Feldmannia sp.* virus) genome. *Virology* **248**, 35-45 (1998).
110 Delaroque, N. *et al.* The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol.* **8**, 110. (2008).
111 Meints, R. H., Ivey, R. G., Lee, A. M. & Choi, T. J. Identification of two virus integration sites in the brown alga *Feldmannia* chromosome. *J Virol.* **82**, 1407-1413 (2008).
112 Delaroque, N. *et al.* The complete DNA sequence of the *Ectocarpus siliculosus* Virus EsV-1 genome. *Virology* **287**, 112-132 (2001).
113 Delaroque, N. *et al.* The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology* **287**, 112-132 (2001).
114 Delaroque, N., Boland, W., Muller, D. G. & Knippers, R. Comparisons of two large phaeoviral genomes and evolutionary implications. *J. Mol. Evol.* **57**, 613-622 (2003).
115 Delaroque, N. & Boland, W. The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol.* **8**, 110 (2008).
116 Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344-1350 (2004).
117 Lehman, I. & Boehmer, P. Replication of herpes simplex virus DNA. *J. Biol. Chem.* **274**, 28059-28062 (1999).
118 Dixon, N. M., Leadbeater, B. S. C. & Wood, K. R. Frequency of viral infection in a field population of *Ectocarpus fasciculatus* (Ectocarpales, Phaeophyceae). *Phycologia* **39**, 258-263 (2000).
119 Müller, D. G. *et al.* Massive prevalence of viral DNA in *Ectocarpus* (Phaeophyceae, Ectocarpales) from two habitats in the North Atlantic and South Pacific. *Bot. Mar.* **43**, 157-159 (2000).
120 Le Corguille, G. *et al.* Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids. *BMC Evol. Biol.* **9**, 253 (2009).
121 Oudot-Le Secq, M., Loiseaux-de Goër, S., Stam, W. & Olsen, J. Complete mitochondrial genomes of the three brown algae (Heterokonta: Phaeophyceae) *Dictyota dichotoma, Fucus vesiculosus* and *Desmarestia viridis*. *Curr. Genet.* **49**, 47-58 (2006).
122 Oudot-Le Secq, M.-P., Kloareg, B. & Loiseaux-De Goër, S. The mitochondrial genome of the brown alga *Laminaria digitata*: a comparative analysis. *Eur. J. Phycol.* **37**, 163-172 (2002).
123 Read, S. M., Currie, G. & Bacic, A. Analysis of the structural heterogeneity of laminarin by electrospray-ionisation-mass spectrometry. *Carbohydrate Res.* **281**, 187-201 (1996).

124 Yamaguchi, T., Ikawa, T. & Nisizawa, K. Incorporation of radioactive carbon from $H^{14}CO_3^-$ into sugar constituents by a brown alga, *Eisenia bicyclis*, during photosynthesis and its fate in the dark. *Plant Cell Physiol.* **7**, 217-229 (1966).

125 Kloareg, B. & Quatrano, R. Structure of the cell walls of marine algae and ecophysiological functions of the matrix polysaccharides. *Oceanogr. Mar. Biol. Ann. Rev.* **26**, 259-315 (1988).

126 Ikawa, T., Watanabe, T. & Nisizawa, K. Enzymes involved in the last steps of the biosynthesis of mannitol in brown algae. *Plant Cell Physiol.* **13**, 1017-1029 (1972).

127 Iwamoto, K. & Shiraiwa, Y. Salt-regulated mannitol metabolism in algae. *Marine Biotech.* **7**, 407-415 (2005).

128 Liberator, P. *et al.* Molecular cloning and functional expression of mannitol-1-phosphatase from the apicomplexan parasite *Eimeria tenella*. *J. Biol. Chem.* **273**, 4237-4244 (1998).

129 Cardenas, M. L., Cornish-Bowden, A. & Ureta, T. Evolution and regulatory role of the hexokinases. *Biochim. Biophys. Acta* **1401**, 242-264 (1998).

130 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-238 (2009).

131 Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268-272 (2009).

132 Henrissat, B., Coutinho, P. M. & Davies, G. J. A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Mol. Biol.* **47**, 55-72 (2001).

133 Paul, M. J., Primavesi, L. F., Jhurreea, D. & Zhang, Y. Trehalose metabolism and signaling. *Ann. Rev. Plant Biol.* **59**, 417-441 (2008).

134 Montijn, R. C. *et al.* Localization of synthesis of beta1,6-glucan in *Saccharomyces cerevisiae*. *J. Bacteriol.* **181**, 7414-7420 (1999).

135 Ramsey, D. M. & Wozniak, D. J. Understanding the control of *Pseudomonas aeruginosa* alginate synthesis and the prospects for management of chronic infections in cystic fibrosis. *Mol. Microbiol.* **56**, 309-322 (2005).

136 Nyvall, P. *et al.* Characterization of mannuronan C-5-epimerase genes from the brown alga *Laminaria digitata*. *Plant Physiol.* **133**, 726-735 (2003).

137 Cohen-Kupiec, R., Broglie, K. E., Friesem, D., Broglie, R. M. & Chet, I. Molecular characterization of a novel beta-1,3-exoglucanase related to mycoparasitism of *Trichoderma harzianum*. *Gene* **226**, 147-154 (1999).

138 Verna, J., Lodder, A., Lee, K., Vagts, A. & Ballester, R. A family of genes required for maintenance of cell wall integrity and for the stress response in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **94**, 13804-13809 (1997).

139 Michel, G., Barbeyron, T., Kloareg, B. & Czjzek, M. The family 6 carbohydrate-binding modules have coevolved with their appended catalytic modules toward similar substrate specificity. *Glycobiology* **19**, 615-623 (2009).

140 Barbeyron, T., L'Haridon, S., Michel, G. & Czjzek, M. *Mariniflexile fucanivorans* sp. nov., a marine member of the Flavobacteriaceae that degrades sulphated fucans from brown algae. *Int. J. Syst. Evol. Microbiol.* **58**, 2107-2113 (2008).

141 Colin, S. *et al.* Cloning and biochemical characterization of the fucanase FcnA: definition of a novel glycoside hydrolase family specific for sulfated fucans. *Glycobiology* **16**, 1021-1032 (2006).

142 Barbeyron, T. *et al.* Matching the sulfur diversity: classification of sulfatases and insights into their evolution. *Mol. Biol. Evol.* **Submitted** (2009).

143 Axelsson, L. Change in pH as a measure of photosynthesis by marine macroalgae. *Mar. Biol.* **97** (1988).

144 Kremer, B. P. & Küppers, U. Carboxylating enzymes and pathway of photosynthetic carbon assimilation in different marine algae- evidence for the C4-pathway? *Planta* **133**, 191-196 (1977).
145 Carr, H. *Energy balance during active carbon uptake and at excess irradiance in three marine macrophytes.* Stockholm University, (2005).
146 Yi, X., Hargett, S., Frankel, L. & Bricker, T. The PsbQ protein is required in *Arabidopsis* for photosystem II assembly/stability and photoautotrophy under low light conditions. *J. Biol. Chem.* **281**, 26260-26267 (2006).
147 Klimmek, F., Sjödin, A., Noutsos, C., Leister, D. & Jansson, S. Abundantly and rarely expressed Lhc protein genes exhibit distinct regulation patterns in plants. *Plant Physiol.* **140**, 793-804 (2006).
148 Elrad, D. & Grossman, A. A genome's-eye view of the light-harvesting polypeptides of Chlamydomonas reinhardtii. *Curr. Genet.* **45**, 61-75 (2004).
149 Niyogi, K., Li, X., Rosenberg, V. & Jung, H. Is PsbS the site of non-photochemical quenching in photosynthesis? *J. Exp. Bot.* **56**, 375-382 (2005).
150 Peers, G. *et al.* An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature* **462**, 518-521 (2009).
151 Gundermann, K. & Büchel, C. The fluorescence yield of the trimeric fucoxanthin-chlorophyll-protein FCPa in the diatom *Cyclotella meneghiniana* is dependent on the amount of bound diatoxanthin. *Photosynth. Res.* **95**, 229-235 (2007).
152 Jeffrey, S. W. The occurrence of chlorophyll c1 and c2 in algae. *J. Phycol.* **12**, 349-354 (1976).
153 Bjørnland, T. & Liaaen-Jensen, S. in *The Chromophyte Algae. Problems and Perspectives,* eds J. C. Green, B. S. C. Leadbeater, & W. L. Diver) 37–61 (Clarendon Press, 1989).
154 Le Corguillé, G. *et al.* Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids. *BMC Evol. Biol.* **9**, 253 (2009).
155 Wilhelm, C. *et al.* The regulation of carbon and nutrient assimilation in diatoms is significantly different from green algae. *Protist* **157**, 91-124 (2006).
156 Fong, A. & Archibald, J. Evolutionary dynamics of light-independent protochlorophyllide oxidoreductase genes in the secondary plastids of cryptophyte algae. *Eukaryot. Cell* **7**, 550-553 (2008).
157 Lüning, K. *Seaweeds: Their Environment, Biogeography, and Ecophysiology*. (John Wiley & Sons, Inc., 1990).
158 Shui, J. *et al.* Light-dependent and light-independent protochlorophyllide oxidoreductases in the chromatically adapting cyanobacterium *Fremyella diplosiphon* UTEX 481. *Plant Cell Physiol.* **50**, 1507-1521 (2009).
159 van Lis, R., Atteia, A., Nogaj, L. & Beale, S. Subcellular localization and light-regulated expression of protoporphyrinogen IX oxidase and ferrochelatase in *Chlamydomonas reinhardtii*. *Plant Physiol.* **139**, 1946-1958 (2005).
160 Lohr, M. & Wilhelm, C. Algae displaying the diadinoxanthin cycle also possess the violaxanthin cycle. *Proc. Natl. Acad. Sci. USA* **96**, 8784-8789 (1999).
161 Frommolt, R. *et al.* Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol. Biol. Evol.* **25**, 2653-2667 (2008).
162 Aknin, M. *et al.* Fatty acid and sterol compositions of eight brown algae from the Senegalese coast. *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.* **102**, 841-843 (1992).
163 Fleury, B. G. *et al.* Sterols from Brazilian marine brown algae. *Phytochemistry* **37**, 1447-1449 (1994).

164 Al Easa, H. S., Kornprobst, J. M. & Rizk, A. M. Major sterol composition of some algae from Qatar. *Phytochemistry* **39**, 373-374 (1995).
165 Benveniste, P. in *The* Arabidopsis *Book. The* Arabidopsis *Book* 1-31 (The American Society of Plant Biologists, 2002).
166 Lodeiro, S., Schulz-Gasch, T. & Matsuda, S. Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. *J. Am. Chem. Soc.* **127**, 14132-14133 (2005).
167 Madoui, M., Bertrand-Michel, J., Gaulin, E. & Dumas, B. Sterol metabolism in the oomycete *Aphanomyces euteiches*, a legume root pathogen. *New Phytol.* **183**, 291-300 (2009).
168 Fernandez, E. & Galvan, A. Inorganic nitrogen assimilation in *Chlamydomonas*. *J. Exp. Bot.* **58**, 2279-2287 (2007).
169 Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. USA* **103**, 11647-11652 (2006).
170 Palenik, B. *et al.* The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. USA* **104**, 7705-7710 (2007).
171 Worden, A. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268-272 (2009).
172 Nawrocki, E. & Eddy, S. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.* **3**, e56 (2007).
173 Muller, D. G., Jaenicke, L., Donike, M. & Akintobi, T. Sex Attractant in a Brown Alga: Chemical Structure. *Science* **171**, 815-817 (1971).
174 Pohnert, G. & Boland, W. The oxylipin chemistry of attraction and defense in brown algae and diatoms. *Nat. Prod. Rep.* **19**, 108-122 (2002).
175 Schmid, C. E., Müller, D. G. & Eichenberger, W. Isolation and characterization of a new phospholipid from brown algae. Intracellular localization and site of biosynthesis. *J. Plant Physiol.* **143**, 570-574 (1994).
176 Ribot, C., Zimmerli, C., Farmer, E., Reymond, P. & Poirier, Y. Induction of the *Arabidopsis PHO1;H10* gene by 12-oxo-phytodienoic acid but not jasmonic acid via a CORONATINE INSENSITIVE1-dependent pathway. *Plant Physiol.* **147**, 696-706 (2008).
177 Taki, N. *et al.* 12-oxo-phytodienoic acid triggers expression of a distinct set of genes and plays a role in wound-induced gene expression in *Arabidopsis*. *Plant Physiol.* **139**, 1268-1283 (2005).
178 Ritter, A. *et al.* Copper stress induces biosynthesis of octadecanoid and eicosanoid oxygenated derivatives in the brown algal kelp *Laminaria digitata*. *New Phytol.* **180**, 809-821 (2008).
179 Tonon, T. *et al.* Fatty acid desaturases from the microalga *Thalassiosira pseudonana*. *FEBS J.* **272**, 3401-3412 (2005).
180 Eichenberger, W., Araki, S. & Müller, D. G. Betaine lipids and phospholipids in brown algae. *Phytochemistry* **34**, 1323-1333 (1993).
181 Coon, M. Cytochrome P450: nature's most versatile biological catalyst. *Annu. Rev. Pharmacol. Toxicol.* **45**, 1-25 (2005).
182 Nelson, D. *et al.* P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6**, 1-42 (1996).
183 Tian, L., Musetti, V., Kim, J., Magallanes-Lundback, M. & DellaPenna, D. The *Arabidopsis LUT1* locus encodes a member of the cytochrome p450 family that is required for carotenoid epsilon-ring hydroxylation activity. *Proc. Natl. Acad. Sci. USA* **101**, 402-407 (2004).

184 Kim, J. & DellaPenna, D. Defining the primary route for lutein synthesis in plants: the role of *Arabidopsis* carotenoid beta-ring hydroxylase CYP97A3. *Proc. Natl. Acad. Sci. USA* **103**, 3474-3479 (2006).
185 Kim, J., Smith, J., Tian, L. & Dellapenna, D. The evolution and function of carotenoid hydroxylases in *Arabidopsis*. *Plant Cell Physiol.* **50**, 463-479 (2009).
186 Lepesheva, G. & Waterman, M. Sterol 14alpha-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. *Biochim. Biophys. Acta* **1770**, 467-477 (2007).
187 Beedle, A., Munday, K. & Wilton, D. Studies on the biosynthesis of tetrahymanol in *Tetrahymena pyriformis*. The mechanism of inhibition by cholesterol. *Biochem. J.* **142**, 57-64 (1974).
188 Emiliani, G., Fondi, M., Fani, R. & Gribaldo, S. A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. *Biol. Direct* **4**, 7 (2009).
189 Amsler, C. D. & Fairhead, V. A. Defensive and sensory chemical ecology of brown algae. *Adv. Bot. Res.* **43**, 2-91 (2006).
190 Shibata, T., Ham, Y., Miyasaki, T., Ito, M. & Nakamura, T. Extracellular secretion of phenolic substances from living brown algae. *J. Appl. Phycol.* **18**, 787–794 (2006).
191 Valverde, C. *et al.* Halometabolites and cellular dehalogenase systems: an evolutionary perspective. *Int. Rev. Cytol.* **234**, 143-199 (2004).
192 Halliwell, B. Phagocyte-derived reactive species: salvation or suicide? *Trends Biochem. Sci.* **31**, 509-515 (2006).
193 Kupper, F. C. *et al.* Iodide accumulation provides kelp with an inorganic antioxidant impacting atmospheric chemistry. *Proc. Natl. Acad. Sci. USA* **105**, 6954-6958 (2008).
194 Potin, P. & Leblanc, C. in *Biological adhesives* Vol. 105-124 eds A.M. Smith & J.A. Callow) (Springer-Verlag, 2006).
195 Colin, C. *et al.* The brown algal kelp *Laminaria digitata* features distinct bromoperoxidase and iodoperoxidase activities. *J. Biol. Chem.* **278**, 23545-23552 (2003).
196 Colin, C. *et al.* Vanadium-dependent iodoperoxidases in *Laminaria digitata*, a novel biochemical function diverging from brown algal bromoperoxidases. *J. Biol. Inorg. Chem.* **10**, 156-166 (2005).
197 Cosse, A., Potin, P. & Leblanc, C. Patterns of gene expression induced by oligoguluronates reveal conserved and environment-specific molecular defense responses in the brown alga *Laminaria digitata*. *New Phytol.* **182**, 239-250 (2009).
198 Russell, G. Formation of an Ectocarpoid Epiflora on Blades of *Laminaria digitata*. *Marine Ecology-Progress Series* **11**, 181-187 (1983).
199 Russell, G. Parallel Growth-Patterns in Algal Epiphytes and Laminaria Blades. *Marine Ecology-Progress Series* **13**, 303-304 (1983).
200 Kawahara, T., Quinn, M. & Lambeth, J. Molecular evolution of the reactive oxygen-generating NADPH oxidase (Nox/Duox) family of enzymes. *BMC Evol. Biol.* **7**, 109 (2007).
201 Eitinger, T. In vivo production of active nickel superoxide dismutase from *Prochlorococcus marinus* MIT9313 is dependent on its cognate peptidase. *J Bacteriol*. **186**, 7821-7825 (2004).
202 Mittler, R. Oxidative stress, antioxidants and stress tolerance. *Trends Plant Sci.* **7**, 405-410 (2002).
203 Russell, G. & Morris, O. P. Copper tolerance in the marine fouling alga *Ectocarpus siliculosus*. *Nature* **228**, 288-289 (1970).

204 Cobbett, C. & Goldsbrough, P. Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. *Annu. Rev. Plant Biol.* **53**, 159-182 (2002).

205 Morris, C., Nicolaus, B., Sampson, V., Harwood, J. & Kille, P. Identification and characterization of a recombinant metallothionein protein from a marine alga, *Fucus vesiculosus*. *Biochem J.* **338**, 553-560 (1999).

206 Murphy, A., Zhou, J., Goldsbrough, P. & Taiz, L. Purification and immunological identification of metallothioneins 1 and 2 from *Arabidopsis thaliana*. *Plant Physiol.* **113**, 1293-1301 (1997).

207 Ha, J. C. *et al.* beta-Agarase from *Pseudomonas* sp. W7: purification of the recombinant enzyme from *Escherichia coli* and the effects of salt on its activity. *Biotech. Appl. Biochem.* **26**, 1-6 (1997).

208 Ortiz, D., Ruscitti, T., McCue, K. & Ow, D. Transport of metal-binding peptides by HMT1, a fission yeast ABC-type vacuolar membrane protein. *J. Biol. Chem.* **270**, 4721-4728 (1995).

209 Martin, J. H. & Fitzwater, S. E. Iron-deficiency limits phytoplankton growth in the Northeast Subarctic Pacific. *Nature* **331**, 341-343 (1988).

210 Bruland, K. W., Donat, J. R. & Hutchins, D. A. Interactive influences of bioactive trace-metals on biological production in oceanic waters *Limnol. Oceanography* **36**, 1555-1577 (1991).

211 Wu, J. F. & Luther, G. W. Size-Fractionated Iron Concentrations in the Water Column of the Western North-Atlantic Ocean. *Limnol. Oceanography* **39**, 1119-1129 (1994).

212 Rue, E. L. & Bruland, K. W. The role of organic complexation on ambient iron chemistry in the equatorial Pacific Ocean and the response of a mesoscale iron addition experiment. *Limnol. Oceanography* **42**, 901-910 (1997).

213 Moog, P. R. & Bruggemann, W. Iron reductase systems on the plant plasma-membrane - a review. *Plant and Soil* **165**, 241-260 (1994).

214 Robinson, N., Procter, C., Connolly, E. & Guerinot, M. A ferric-chelate reductase for iron uptake from soils. *Nature* **397**, 694-697 (1999).

215 Allmang, C., Wurth, L. & Krol, A. The selenium to selenoprotein pathway in eukaryotes: More molecular partners than anticipated. *Biochim. Biophys. Acta* **1790**, 1415-1423 (2009).

216 Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439-441 (2003).

217 Kryukov, G. *et al.* Characterization of mammalian selenoproteomes. *Science* **300**, 1439-1443 (2003).

218 Chapple, C., Guigó, R. & Krol, A. SECISaln, a web-based tool for the creation of structure-based alignments of eukaryotic SECIS elements. *Bioinformatics* **25**, 674-675 (2009).

219 Lobanov, A. *et al.* Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol.* **8**, R198 (2007).

220 Hsia, C. C. & McGinnis, W. Evolution of transcription factor function. *Curr. Opin. Genet. Develop.* **13**, 199-206 (2003).

221 Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147-151 (2003).

222 Gutierrez, R. A., Green, P. J., Keegstra, K. & Ohlrogge, J. B. Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? *Genome Biol.* **5**, 15 (2004).

223 Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol* **3**, e245 (2005).

224 Schauser, L., Roussis, A., Stiller, J. & Stougaard, J. A plant regulator controlling development of symbiotic root nodules. *Nature* **402**, 191-195 (1999).
225 Schauser, L., Wieloch, W. & Stougaard, J. Evolution of NIN-like proteins in *Arabidopsis*, rice, and *Lotus japonicus*. *J Mol. Evol.* **60**, 229-237 (2005).
226 Ferris, P. J., Armbrust, E. V. & Goodenough, U. W. Genetic structure of the mating-type locus of *Chlamydomonas reinhardtii*. *Genetics* **160**, 181-200 (2002).
227 Lin, H. & Goodenough, U. W. Gametogenesis in the *Chlamydomonas reinhardtii* minus mating type is controlled by two genes, *MID* and *MTD1*. *Genetics* **176**, 913-925 (2007).
228 Nozaki, H., Mori, T., Misumi, O., Matsunaga, S. & Kuroiwa, T. Males evolved from the dominant isogametic mating type. *Curr. Biol.* **16**, R1018-1020 (2006).
229 Fernandez, E. & Galvan, A. Nitrate assimilation in *Chlamydomonas*. *Eukaryot. Cell* **7**, 555-559 (2008).
230 Lin, H. & Goodenough, U. Gametogenesis in the *Chlamydomonas reinhardtii* minus mating type is controlled by two genes, *MID* and *MTD1*. *Genetics* **176**, 913-925 (2007).
231 Kirk, D. L. Oogamy: inventing the sexes. *Curr. Biol.* **16**, R1028-1030 (2006).
232 Peters, A. F. *et al.* Life-cycle-generation-specific developmental processes are modified in the immediate upright mutant of the brown alga *Ectocarpus siliculosus*. *Development* **135**, 1503-1512 (2008).
233 Hanks, S. K., Quinn, A. M. & Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52 (1988).
234 John, P. C. L., Sek, F. J. & Lee, M. G. A Homolog of the Cell Cycle Control Protein p34*cdc2* Participates in the Division Cycle of *Chlamydomonas*, and a Similar Protein is Detectable in Higher Plants and Remote Taxa. *Plant Cell* **1**, 1185-1193 (1989).
235 Pu, R. & Robinson, K. R. The involvement of Ca^{2+} gradients, Ca^{2+} fluxes, and CaM kinase II in polarization and germination of *Silvetia compressa* zygotes. *Planta* **217**, 407-416 (2003).
236 Küpper, F. C., Kloareg, B., Guern, J. & Potin, P. Oligoguluronates Elicit an Oxidative Burst in the Brown Algal Kelp *Laminaria digitata*. *Plant Physiol.* **125**, 278-291 (2001).
237 Foissac, S. *et al.* Genome Annotation in Plants and Fungi: EuGene as a model platform. *Curr. Bioinfo.* **3**, 87-97 (2008).
238 Miranda-Saavedra, D. & Barton, G. J. Classification and Functional Annotation of Eukaryotic Protein Kinases. *Proteins* **68**, 893-914 (2007).
239 Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298**, 1912-1934 (2002).
240 Shiu, S. H. & Bleecker, A. B. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proc. Natl. Acad. Sci. USA* **98**, 10763-10768 (2001).
241 Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java Alignment Editor. *Bioinformatics* **20**, 426-427 (2004).
242 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599 (2007).
243 Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).
244 Eck, R. V. & Dayhoff, M. O. *Atlas of Protein Sequence and Structure*. (National Biomedical Research Foundation, 1966).

245 Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783-791 (1985).
246 Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86 (2004).
247 Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends. Biol. Sci.* **27**, 514-520 (2002).
248 Hanks, S. K. & Quinn, A. M. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **200**, 38-62 (1991).
249 Gschloessl, B., Guermeur, Y. & Cock, J. M. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* **9**, 393 (2008).
250 Lopez, J. A. *et al.* Cloning of the alpha chain of human platelet glycoprotein Ib: a transmembrane protein with homology to leucine-rich alpha 2-glycoprotein. *Proc Natl Acad Sci USA* **84**, 5615-5619 (1987).
251 Titani, K., Takio, K., Handa, M. & Ruggeri, Z. M. Amino acid sequence of the von Willebrand factor-binding domain of platelet membrane glycoprotein Ib. *Proc Natl Acad Sci USA* **84**, 5610-5614 (1987).
252 Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **3**, W465-468 (2009).
253 Manning, G., Young, S. L., Miller, W. T. & Zhai, Y. The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc. Natl. Acad. Sci. USA* **105**, 9674-9679 (2008).
254 Rokas, A. The molecular origins of multicellular transitions. *Curr. Opin.Genet. Develop.* **18**, 472–478 (2008).
255 Tyler, B. M. *et al. Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-1266 (2006).
256 Sancak, Y. *et al.* The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320**, 1496-1501 (2008).
257 Ishikawa, M. *et al.* Distribution and phylogeny of the blue light receptors aureochromes in eukaryotes. *Planta* **230**, 543-552 (2009).
258 Leipe, D., Wolf, Y., Koonin, E. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41-72 (2002).
259 Gas, E., Flores-Pérez, U., Sauret-Güeto, S. & Rodríguez-Concepción, M. Hunting for plant nitric oxide synthase provides new evidence of a central role for plastids in nitric oxide metabolism. *Plant Cell* **21**, 18-23 (2009).
260 Vardi, A. *et al.* A diatom gene regulating nitric-oxide signaling and susceptibility to diatom-derived aldehydes. *Curr. Biol.* **18**, 895-899 (2008).
261 Martens, S. & Howard, J. The interferon-inducible GTPases. *Annu. Rev. Cell Dev. Biol.* **22**, 559-589 (2006).
262 Hu, J. *et al.* A class of dynamin-like GTPases involved in the generation of the tubular ER network. *Cell* **138**, 549-561 (2009).
263 Orso, G. *et al.* Homotypic fusion of ER membranes requires the dynamin-like GTPase atlastin. *Nature* **460**, 978-983 (2009).
264 Wloga, D., Strzyzewska-Jówko, I., Gaertig, J. & Jerka-Dziadosz, M. Septins stabilize mitochondria in *Tetrahymena thermophila*. *Eukaryot. Cell* **7**, 1373-1386 (2008).
265 Boureux, A., Vignal, E., Faure, S. & Fort, P. Evolution of the Rho family of ras-like GTPases in eukaryotes. *Mol. Biol. Evol.* **24**, 203-216 (2007).
266 Reuther, G. & Der, C. The Ras branch of small GTPases: Ras family members don't fall far from the tree. *Curr. Opin. Cell Biol.* **12**, 157-165 (2000).

267 Marín, I., van Egmond, W. & van Haastert, P. The Roco protein family: a functional perspective. *FASEB J.* **22**, 3103-3110 (2008).

268 Baldauf, S. L. An overview of the phylogeny and diversity of eukaryotes. *J. Syst. Evol.* **46**, 263–273 (2008).

269 Nürnberger, T., Brunner, F., Kemmerling, B. & Piater, L. Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol. Rev.* **198**, 249-266 (2004).

270 Charrier, B. *et al.* Development and physiology of the brown alga *Ectocarpus siliculosus*: two centuries of research. *New Phytol.* **177**, 319-332 (2008).

271 Küpper, F. C. *et al.* Iodide accumulation provides kelp with an inorganic antioxidant impacting atmospheric chemistry. *Proc. Natl. Acad. Sci. USA* **105**, 6954-6958 (2008).

272 Palsson-McDermott, E. M. & O'Neill, L. A. J. Building an immune system from nine domains. *Biochem. Soc. Trans.* **35**, 1437-1444 (2007).

273 Pancer, Z. *et al.* Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**, 174-180 (2004).

274 Bosch, T. C. G. *et al.* Uncovering the evolutionary history of innate immunity: The simple metazoan Hydra uses epithelial cells for host defence. *Develop. Comp. Immunol.* **33**, 559-569 (2009).

275 Ausubel, F. M. Are innate immune signaling pathways in plants and animals conserved? *Nature Immunol.* **6**, 973-979 (2005).

276 Povelones, M., Waterhouse, R. M., Kafatos, F. C. & Christophides, G. K. Leucine-Rich Repeat Protein Complex Activates Mosquito Complement in Defense Against Plasmodium Parasites. *Science* **324**, 258-261 (2009).

277 Dangl, J. F. & Jones, J. D. G. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826-833 (2001).

278 Tameling, W. I. L. & Joosten, M. The diverse roles of NB-LRR proteins in plants. *Physiol. Mol. Plant Pathol.* **71**, 126-134 (2007).

279 Meyers, B. C., Kaushik, S. & Nandety, R. S. Evolving disease resistance genes. *Curr. Opin. Plant Biol.* **8**, 129-134 (2005).

280 Van der Biezen, E. A. & Jones, J. D. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8**, R226-227 (1998).

281 Bidle, K. D. & Bender, S. J. Iron starvation and culture age activate metacaspases and programmed cell death in the marine diatom *Thalassiosira pseudonana*. *Eukaryot. Cell* **7**, 223-236 (2008).

282 Hatsugai, N., Kuroyanagi, M., Nishimura, M. & Hara-Nishimura, I. A cellular suicide strategy of plants: vacuole-mediated cell death. *Apoptosis* **11**, 905-911 (2006).

283 Noël, L. *et al.* Interaction between SGT1 and cytosolic/nuclear HSC70 chaperones regulates *Arabidopsis* immune responses. *Plant Cell* **19**, 4061-4076 (2007).

284 da Silva Correia, J., Miranda, Y., Leonard, N. & Ulevitch, R. SGT1 is essential for Nod1 activation. *Proc. Natl. Acad. Sci. USA* **104**, 6764-6769 (2007).

285 Roeder, V. *et al.* Identification of stress gene transcripts in *Laminaria digitata* (Phaeophyceae) protoplast cultures by expressed sequence tag analysis. *J. Phycol.* **41**, 1227-1235 (2005).

286 Allen, R. L. *et al.* Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. *Science* **306**, 1957-1960 (2004).

287 Xie, Z. P. & Klionsky, D. J. Autophagosome formation: Core machinery and adaptations. *Nature Cell Biol.* **9**, 1102-1109 (2007).

288 Kourtis, N. & Tavernarakis, N. Autophagy and cell death in model organisms. *Cell Death Diff.* **16**, 21-30 (2009).

289 Hofius, D. *et al.* Autophagic Components Contribute to Hypersensitive Cell Death in *Arabidopsis*. *Cell* **137**, 773-783 (2009).
290 Klionsky, D. J. *et al.* A unified nomenclature for yeast autophagy-related genes. *Developmental Cell* **5**, 539-545 (2003).
291 Antoniw, J. F., Ritter, C. E., Pierpoint, W. S. & Van Loon, L. C. Comparison of three pathogenesis-related proteins from plants of two cultivars of tobacco infected with TMV. *J. Gen. Virol.***47**, 79-87 (1980).
292 Van Loon, L. C., Rep, M. & Pieterse, C. M. J. Significance of inducible defense-related proteins in infected plants. *Annu. Rev. Phytopath.* **44**, 135-162 (2006).
293 Edreva, A. Pathogenesis-related proteins: Research progress in the last 15 years. *Gen. Appl. Plant Physiol.* **31**, 105-124 (2005).
294 Férnandez, C. *et al.* NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576-593 (1997).
295 Rawlings, N. D. & Barrett, A. J. Evolutionary families of metallopeptidases. *Methods Enzymol.* **248**, 183-228 (1995).
296 Schlagenhauf, E., Etges, R. & Metcalf, P. The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). *Structure* **6**, 1035-1046 (1998).
297 Dittami, S. M. *et al.* Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biol.* **10** (2009).
298 Vigers, A. J. *et al.* Thaumatin-like pathogenesis-related proteins are antifungal. *Plant Sci.* **83**, 155-161 (1992).
299 Green, T. R. & Ryan, C. A. Wound-induced proteinase inhibitors in plant leaves: possible defense mechanism against insects. *Science* **175**, 776-777 (1972).
300 Rawlings, N. D., Morton, F. R., Kok, C. Y., Kong, J. & Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Res.* **36**, 320-325 (2008).
301 Zhang, Z., Collinge, D. B. & Thordal-Christensen, H. Germin-like oxalate oxidase, a H_2O_2-producing enzyme, accumulates in barley attacked by the powdery mildew fungus. *Plant J.* **8**, 139-145 (1995).
302 Wei, Y. *et al.* An epidermis/papilla-specific oxalate oxidase-like protein in the defence response of barley attacked by the powdery mildew fungus. *Plant Mol. Biol.* **36**, 101-112 (1998).
303 Bernier, F., Lemieux, G. & Pallotta, D. Gene families encode the meajor encystment-specific proteins of *Physarum polycephalum* plasmodia. *Gene* **59**, 265-277 (1987).
304 Lane, B. G. *et al.* Homologies between members of the germin gene family in hexaploid wheat and similarities between these wheat germins and certain *Physarum* spherulins. *J. Biol. Chem.* **266**, 10461-10469 (1991).
305 Woo, E., Dunwell, J., Goodenough, P., Marvier, A. & Pickersgill, R. Germin is a manganese containing homohexamer with oxalate oxidase and superoxide dismutase activities. *Nat Struct. Biol.* **7**, 1036-1040 (2000).
306 Khuri, S., Bakker, F. & Dunwell, J. Phylogeny, function, and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. *Mol. Biol. Evol.* **18**, 593-605 (2001).
307 Dunwell, J., Purvis, A. & Khuri, S. Cupins: the most functionally diverse protein superfamily? *Phytochemistry* **65**, 7-17 (2004).
308 Bernier, F. & Berna, A. Germins and germin-like proteins: Plant do-all proteins. But what do they do exactly? *Plant Physiol. Biochem.* **39**, 545-554 (2001).
309 Venkatachalam, K. & Montell, C. TRP channels. *Annu. Rev. Biochem.* **76**, 387-417 (2007).

310 Levina, N. *et al.* Protection of *Escherichia coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. *Embo J.* **18**, 1730-1737 (1999).

311 Haswell, E. S. & Meyerowitz, E. M. MscS-like proteins control plastid size and shape in *Arabidopsis thaliana*. *Curr. Biol.* **16**, 1-11 (2006).

312 Nakayama, Y., Fujiu, K., Sokabe, M. & Yoshimura, K. Molecular and electrophysiological characterization of a mechanosensitive channel expressed in the chloroplasts of *Chlamydomonas*. *Proc. Natl. Acad. Sci. USA* **104**, 5883-5888 (2007).

313 Haswell, E. S., Peyronnet, R., Barbier-Brygoo, H., Meyerowitz, E. M. & Frachisse, J. M. Two MscS homologs provide mechanosensitive channel activities in the *Arabidopsis* root. *Curr. Biol.* **18**, 730-734 (2008).

314 Taylor, A., Manison, N., Fernandez, C., Wood, J. & Brownlee, C. Spatial Organization of Calcium Signaling Involved in Cell Volume Control in the *Fucus* Rhizoid. *Plant Cell* **8**, 2015-2031 (1996).

315 Goddard, H., Manison, N., Tomos, D. & Brownlee, C. Elemental propagation of calcium signals in response-specific patterns determined by environmental stimulus strength. *Proc. Natl. Acad. Sci. USA* **97**, 1932-1937 (2000).

316 Coelho, S. M. *et al.* Spatiotemporal patterning of reactive oxygen production and Ca2+ wave propagation in fucus rhizoid cells. *Plant Cell* **14**, 2369-2381 (2002).

317 Ward, J. M., Maser, P. & Schroeder, J. I. Plant Ion Channels: Gene Families, Physiology, and Functional Genomics Analyses. *Annu. Rev. Physiol.* **71**, 59-82 (2009).

318 Fernandez, C., Pannone, B., Chen, X., Fuchs, G. & Wolin, S. An Lsm2-Lsm7 complex in *Saccharomyces cerevisiae* associates with the small nucleolar RNA snR5. *Mol. Biol. Cell* **15**, 2842-2852 (2004).

319 Khusial, P., Plaag, R. & Zieve, G. LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends Biochem. Sci.* **30**, 522-528 (2005).

320 Kufel, J., Allmang, C., Verdone, L., Beggs, J. & Tollervey, D. Lsm proteins are required for normal processing of pre-tRNAs and their efficient association with La-homologous protein Lhp1p. *Mol. Cell. Biol.* **22**, 5248-5256 (2002).

321 Wilusz, C. & Wilusz, J. Eukaryotic Lsm proteins: lessons from bacteria. *Nat. Struct. Mol. Biol.* **12**, 1031-1036 (2005).

322 Séraphin, B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* **14**, 2089-2098 (1995).

323 Møller, T. *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol. Cell* **9**, 23-30 (2002).

324 Scofield, D. & Lynch, M. Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol. Biol. Evol.* **25**, 2255-2267 (2008).

325 Albrecht, M. & Lengauer, T. Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett.* **569**, 18-26 (2004).

326 Pillai, R. *et al.* Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes Dev.* **17**, 2321-2333 (2003).

327 Schümperli, D. & Pillai, R. The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein. *Cell Mol. Life Sci.* **61**, 2560-2570 (2004).

328 Dominski, Z. & Marzluff, W. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene* **396**, 373-390 (2007).

329 Marzluff, W. & Duronio, R. Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences. *Curr. Opin. Cell Biol*. **14**, 692-699 (2002).
330 Mullen, T., Kaygun, H. & Marzluff, W. Chapter 2. Cell-cycle regulation of histone mRNA degradation in Mammalian cells: role of translation and oligouridylation. *Methods Enzymol.* **449**, 23-45 (2008).
331 Jaeger, S., Eriani, G. & Martin, F. Critical residues for RNA discrimination of the histone hairpin binding protein (HBP) investigated by the yeast three-hybrid system. *FEBS Lett.* **556**, 265-270 (2004).
332 Wang, Z., Whitfield, M., Ingledue, T. r., Dominski, Z. & Marzluff, W. The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing. *Genes Dev.* **10**, 3028-3040 (1996).
333 Fabry, S. *et al.* The organization structure and regulatory elements of *Chlamydomonas* histone genes reveal features linking plant and animal genes. *Curr. Genet.* **28**, 333-345 (1995).
334 Chabouté, M., Chaubet, N., Gigot, C. & Philipps, G. Histones and histone genes in higher plants: structure and genomic organization. *Biochimie* **75**, 523-531 (1993).
335 Russell, A., Charette, J., Spencer, D. & Gray, M. An early evolutionary origin for the minor spliceosome. *Nature* **443**, 863-866 (2006).
336 Hershey, J. W. & Merrick, W. C. in *Translational control of gene expression* eds N. Sonenberg, J.W. Hershey, & M.B. Mathews) 33-88 (Cold Spring Harbor Laboratory Press, 2000).
337 Browning, K. Plant translation initiation factors: it is not easy to be green. *Biochem. Soc. Trans.* **32**, 589-591 (2004).
338 Lin, Z., Kong, H., Nei, M. & Ma, H. Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc. Natl. Acad. Sci. USA* **103**, 10328-10333 (2006).
339 Bleuyard, J., Gallego, M. & White, C. Recent advances in understanding of the DNA double-strand break repair machinery of plants. *DNA Repair (Amst)* **5**, 1-12 (2006).
340 Krogh, B. & Symington, L. Recombination proteins in yeast. *Annu. Rev. Genet.* **38**, 233-271 (2004).
341 Chepurnov, V. *et al.* In search of new tractable diatoms for experimental biology. *Bioessays* **30**, 692-702 (2008).
342 Keeney, S., Giroux, C. & Kleckner, N. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375-384 (1997).
343 Lichten, M. Meiotic recombination: breaking the genome to save it. *Curr. Biol.* **11**, R253-256 (2001).
344 Hartung, F. *et al.* The catalytically active tyrosine residues of both SPO11-1 and SPO11-2 are required for meiotic double-strand break induction in *Arabidopsis*. *Plant Cell* **19**, 3090-3099 (2007).
345 Sugimoto-Shirasu, K., Stacey, N., Corsar, J., Roberts, K. & McCann, M. DNA topoisomerase VI is essential for endoreduplication in *Arabidopsis*. *Curr. Biol.* **12**, 1782-1786 (2002).
346 Yin, Y. *et al.* A crucial role for the putative *Arabidopsis* topoisomerase VI in plant growth and development. *Proc. Natl. Acad. Sci. USA* **99**, 10191-10196 (2002).
347 Ravid, K., Lu, J., Zimmet, J. & Jones, M. Roads to polyploidy: the megakaryocyte example. *J. Cell Physiol.* **190**, 7-20 (2002).
348 Zimmet, J. & Ravid, K. Polyploidy: occurrence in nature, mechanisms, and significance for the megakaryocyte-platelet system. *Exp. Hematol.* **28**, 3-16 (2000).

349 Chen, C., Zhang, W., Timofejeva, L., Gerardin, Y. & Ma, H. The *Arabidopsis ROCK-N-ROLLERS* gene encodes a homolog of the yeast ATP-dependent DNA helicase MER3 and is required for normal meiotic crossover formation. *Plant J.* **43**, 321-334 (2005).

350 Mercier, R. *et al.* Two meiotic crossover classes cohabit in *Arabidopsis*: one is dependent on MER3,whereas the other one is not. *Curr. Biol.* **15**, 692-701 (2005).

351 Arnaout, M., Goodman, S. & Xiong, J. Structure and mechanics of integrin-based cell adhesion. *Curr. Opin. Cell Biol.* **19**, 495-507 (2007).

352 Takagi, J. Structural basis for ligand recognition by integrins. *Curr. Opin. Cell Biol.* **19**, 557-564 (2007).

353 Gale, C. *et al.* Linkage of adhesion, filamentous growth, and virulence in *Candida albicans* to a single gene, INT1. *Science* **279**, 1355-1358 (1998).

354 Cornillon, S. *et al.* An adhesion molecule in free-living *Dictyostelium* amoebae with integrin beta features. *EMBO Rep.* **7**, 617-621 (2006).

355 Ziegler, W., Gingras, A., Critchley, D. & Emsley, J. Integrin connections to the cytoskeleton through talin and vinculin. *Biochem. Soc. Trans.* **36**, 235-239 (2008).

356 Assoian, R. & Klein, E. Growth control by intracellular tension and extracellular stiffness. *Trends Cell Biol.* **18**, 347-352 (2008).

357 Takenawa, T. & Suetsugu, S. The WASP-WAVE protein network: connecting the membrane to the cytoskeleton. *Nat. Rev. Mol. Cell. Biol.* **8**, 37-48 (2007).

358 Perroud, P. & Quatrano, R. *BRICK1* is required for apical cell growth in filaments of the moss *Physcomitrella patens* but not for gametophore morphology. *Plant Cell* **20**, 411-422 (2008).

359 Nagasato, C. & Motomura, T. Influence of the centrosome in cytokinesis of brown algae: polyspermic zygotes of *Scytosiphon lomentaria* (Scytosiphonales, Phaeophyceae). *J. Cell Sci.* **115**, 2541-2548 (2002).

360 Lange, B. & Gull, K. A molecular marker for centriole maturation in the mammalian cell cycle. *J. Cell Biol.* **130**, 919-927 (1995).

361 Chang, P. & Stearns, T. Delta-tubulin and epsilon-tubulin: two new human centrosomal tubulins reveal new aspects of centrosome structure and function. *Nat. Cell Biol.* **2**, 30-35 (2000).

362 Dutcher, S. The tubulin fraternity: alpha to eta. *Curr. Opin. Cell Biol.* **13**, 49-54 (2001).

363 Dutcher, S. Long-lost relatives reappear: identification of new members of the tubulin superfamily. *Curr. Opin. Microbiol.* **6**, 634-640 (2003).

364 Kiefel, B., Gilson, P. & Beech, P. Diverse eukaryotes have retained mitochondrial homologues of the bacterial division protein FtsZ. *Protist* **155**, 105-115 (2004).

365 Katsaros, C., Karyophyllis, D. & Galatis, B. Cytoskeleton and morphogenesis in brown algae. *Ann. Bot. (Lond)* **97**, 679-693 (2006).

366 Faix, J. & Grosse, R. Staying in shape with formins. *Dev Cell* **10**, 693-706 (2006).

367 Pollard, T. Regulation of actin filament assembly by Arp2/3 complex and formins. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 451-477 (2007).

368 Hable, W. & Kropf, D. The Arp2/3 complex nucleates actin arrays during zygote polarity establishment and growth. *Cell Motil. Cytoskeleton* **61**, 9-20 (2005).

369 Dacks, J. & Field, M. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *J. Cell Sci.* **120**, 2977-2985 (2007).

370 Bouck, G. Fine structure and organelle association in brown algae. *J. Cell Biol.* **26**, 523-537 (1965).

371 Oliveira, L. & Bisalputra, T. Studies in the brown alga *Ectocarpus* in culture. *J. Submicrob. Cytol.* **5**, 107-120 (1973).

372 Jahn, R. & Scheller, R. SNAREs - engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **7**, 631-643 (2006).
373 Lipka, V., Kwon, C. & Panstruga, R. SNARE-ware: the role of SNARE-domain proteins in plant biology. *Annu. Rev. Cell Dev. Biol.* **23**, 147-174 (2007).
374 Paul, M. & Frigerio, L. Coated vesicles in plant cells. *Semin. Cell Dev. Biol.* **18**, 471-478 (2007).
375 Rosenbaum, J. & Witman, G. Intraflagellar transport. *Nat. Rev. Mol. Cell Biol.* **3**, 813-825 (2002).
376 Scholey, J. Intraflagellar transport. *Annu. Rev. Cell Dev. Biol.* **19**, 423-443 (2003).
377 Pan, J. & Snell, W. *Chlamydomonas* shortens its flagella by activating axonemal disassembly, stimulating IFT particle trafficking, and blocking anterograde cargo loading. *Dev. Cell* **9**, 431-438 (2005).
378 Wang, Z., Fan, Z., Williamson, S. & Qin, H. Intraflagellar transport (IFT) protein IFT25 is a phosphoprotein component of IFT complex B and physically interacts with IFT27 in *Chlamydomonas*. *PLoS One* **4**, e5384 (2009).
379 Lechtreck, K. & Melkonian, M. Striated microtubule-associated fibers: identification of assemblin, a novel 34-kD protein that forms paracrystals of 2-nm filaments in vitro. *J. Cell Biol.* **115**, 705-716 (1991).
380 Harper, J., Thuet, J., Lechtreck, K. & Hardham, A. Proteins related to green algal striated fiber assemblin are present in stramenopiles and alveolates. *Protoplasma* **236**, 97-101 (2009).
381 Honda, D. *et al.* Homologs of the *sexually induced gene 1* (*sig1*) product constitute the stramenopile mastigonemes. *Protist* **158**, 77-88 (2007).
382 Yamagishi, T., Motomura, T., Nagasato, C., Kato, A. & Kawai, H. A tubular mastigoneme-related protein, Ocm1, Isolated from the flagellum of a chromophyte alga, *Ochromonas danica*. *J. Phycol.* **43**, 519-527 (2007).
383 Yamagishi, T., Motomura, T., Nagasato, C. & Kawai, H. Novel proteins comprising the stramenopile tripartite mastigoneme in *Ochromonas danica* (Chrysophyceae). *J. Phycol.* **45**, 1100-1105 (2009).
384 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).
385 Waterhouse, A., Procter, J., Martin, D., Clamp, M. & Barton, G. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
386 Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
387 Ronquist, F. & Huelsenbeck, J. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
388 Rambaut, A. *FigTree, a graphical viewer of phylogenetic trees, version 1.2.*, <http://tree.bio.ed.ac.uk/software/figtree/> (2009).
389 Schauser, L., Wieloch, W. & Stougaard, J. Evolution of NIN-like proteins in *Arabidopsis*, rice, and *Lotus japonicus*. *J. Mol. Evol.* **60**, 229-237 (2005).
390 Schlagenhauf, E., Etges, R. & Metcalf, P. The crystal structure of the *Leishmania* major surface proteinase leishmanolysin (gp63). *Structure* **6**, 1035-1046 (1998).
391 Fernández, C. *et al.* NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576-593 (1997).
392 Pazour, G., Agrin, N., Leszyk, J. & Witman, G. Proteomic analysis of a eukaryotic cilium. *J. Cell Biol.* **170**, 103-113 (2005).