The *Chondrus crispus* genome – Supplementary material

1 Genome sequencing, assembly and annotation

1.1 Sequencing

The *Chondrus crispus* genome was sequenced using a whole genome shotgun strategy. The data were generated by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers. The main raw data for the genome project are 14-fold coverage sequencing, produced from genomic libraries with various insert sizes (Table S1.1).

Table S1.1. Libraries used for sequencing.

Insert size (kb)	Coverage (fold)	Number of reads
3	10.8	1,884,732
10	3.2	541,850
20	0.3	47,645

1.2 Genome assembly

The sequence reads were assembled with Arachne (1); generating 5,415 contigs linked into 1,266 scaffolds covering 122.2 Mb. The contig N50 was 64 kb, and the scaffolds N50 was 240 kb. Post-processing of the assembly was applied to remove bacterial contaminations. Each scaffold was cut into 0.1-kb overlapping windows of 1 kb, and each window was aligned against the nr databases with blastx. The windows were categorized by their best hit in one of the following categories: bacterial, archaea, eukaryote or unknown. Scaffolds composed of at least 60% of bacterial windows were considered as bacterial contaminations and were removed. The final assembly is composed of 921 scaffolds (> 2kb) and covered 104.7 Mb.

1.3 Annotation pipeline

Most of the genome comparisons were performed with repeat masked sequences. For this purpose, we sequentially searched and masked several kinds of repeats: known repeats and transposons available in Repbase with the RepeatMasker program (2); tandem repeats with the TRF program (3); and *ab initio* repeat detection with RepeatScout (4). From this pipeline, 3% of assembled bases were masked.

The Uniprot (5) database was used to detect conserved genes between *C. crispus* and other species. As Genewise (6) is processor intensive, the Uniprot database was first aligned with the genome assembly using BLAT (7). Subsequently, we extracted the genomic regions where no proteic hit had been found by BLAT and realigned Uniprot proteins with more permissive parameters. Each significant match was then refined using Genewise in order to identify exon and intron boundaries. Geneid (8) and SNAP (9) *ab inito* gene prediction software were trained on *C. crispus* cDNAs.

Chondrus crispus full-length cDNA libraries were constructed from the sequenced strain using pooled RNA from a diurnal sampling and after desiccation stress. Sequences were generated using 454

GS-FLX technology, which produced 333,836 useful reads. After a cleaning procedure, these reads were aligned to the genome assembly with the following pipeline: the sequences were aligned with BLAT on the assembly and only the best match (with identity greater than 90%) for each read was selected. Then each match was extended by 1 kb on each end and realigned with the read using the Est2genome software (10). Also, the EST reads were assembled with Newbler (v2.0.00.20) into 8,212 contigs and by Phrap (v1.080812) into 15,776 contigs. These two pools of contigs were mapped on the genome assembly following the same pipeline used for reads mapping.

A collection of 41,255 public mRNA sequences from red algae (downloaded from the EMBL database) was first aligned with the genome assembly by Blat (7), using default parameters between translated genomic and translated ESTs. To refine the Blat alignment, we used Est2Genome (10). Only the best match (from Blat alignments) was selected for each mRNA sequences.

All the resources described above were later used to automatically build gene models using GAZE (11). Individual predictions from each of the programs (geneid, SNAP, Genewise, Est2genome) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop). Exons predicted by *ab initio* softwares, Genewise and Est2genome, were used as coding segments. Introns predicted by Genewise and Est2genome were used as intron segments. When an mRNA was aligned to a genome locus, an intergenic segment with a negative score was created from the span (coercing GAZE not to split genes). Predicted repeats were used as intron and intergenic segments to avoid prediction of genes coding proteins in such regions. Additionally, transcript stop signals were extracted from the ends of mRNAs (polyA tail positions). Each segment coming from software predicting exon boundaries (like Genewise, Est2genome or *ab initio* predictors), was used by GAZE only if GAZE chose the same boundaries. A value was assigned to each segment or signal from a given program, reflecting confidence in the data, and these values were used as scores for the arcs of the GAZE automaton. All signals were given a fixed score, but segment scores were context sensitive: coding segment scores were linked to the percentage of identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. A weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplier for the score of each information source, before processing by GAZE. Finally, gene predictions created by GAZE were filtered following their scores and their lengths. When applied to the entire assembled sequence, GAZE predicted 9,606 gene models (Table S1.2).

Table S1.2. General gene characteristics of the *C. crispus* genome.

	Number	Average size (kb)	Genomic coverage (%)
Genes	9,606	1.242	11.3
monoexonic	8,059	0.983	
Exons	12,633	0.789	10.8 (9.3 CDS, 1.3 UTR)
Introns	3,066	0.182	0.5
Intergenic regions	-	-	88.7

The vast majority (88%) of the genes are monoexonic, resulting in an average number of exons per gene of 1.32. The proportion of intron-containing genes is not as low as in the very compact red algal genome of *Cyanidioschyzon merolae*, but it is lower than in those of unicellular green algae (Table S1.3), suggesting that a low number of introns is a characteristic of red algal genomes. Among the 18 *C. merolae* intron-containing genes that have orthologs (defined as best reciprocal hits, see below) in *C. crispus*, only eleven contain introns in *C. crispus* among which only five share identical intron positions between both species. Thus, the majority of introns that were maintained in *C. merolae* are different from those that were maintained in *C. crispus*.

Table S1.3. General genome characteristics between red and green algae**.**

	Red algae		Green algae	
	Chondrus crispus	*Cyanidioschyzon merolae*	*Micromonas pusilla*	*Ostreococcus lucimarinus*
Genome size	104.8 Mb	16.5 Mb	21 Mb	13.2 Mb
Number of genes	9,606	4,789	10,108	7,805
Exons per gene	1.32	1.01	1.55	1.31
Intron-containing genes	12%	0.6%	23.2%	21.2%

1.4 Expert annotation

Following the automatic annotation of the genome there was also an expert structural and functional annotation. This involved the annotation of groups of genes to provide knowledge on the gene content, but also to correct if needed the structural annotation. The identification of the genes included generally verification using BLAST searches verified with for instance Interpro domains, Pfam, HMMs as well as manual verification of EST support. The data is thus on inferences, this is, however, something that all genome projects outside well-studied organism have in common. This was also well known for the expert annotators that took great care of creating a relevant annotation. Even though many genes are lacking introns we are confident in their eukaryotic origins since only contigs larger

than 2 kb were used and all of the contigs show a eukaryotic signature with either genes with introns or non-coding regions

1.5 Intron conservation profiles in orthologous proteins

We identified orthologous genes within nine species: *Arabidopsis thaliana, Nematostella vectensis, Thalassiosira pseudonana, Cyanidioschyzon merolae, Ostreococcus lucimarinus, Micromonas pusilla, Phaeodactylum tricornutum, Guillardia theta* and *Vitis vinifera*. Each pair of predicted gene sets was aligned with the Smith-Waterman algorithm, and alignments with a score higher than 300 (BLOSUM62, gapo=10, gape=1) were retained. Orthologs were defined as best reciprocal hits (BRH).

We compared intron positions in *C. crispus* genes and their orthologs in five species of algae and land plants: *A. thaliana* (At), *V. vinifera* (Vv), *Physcomitrella patens* (Pp), *O. lucimarinus* (Ol) and *M. pusilla* (Mp). When available, the orthologous gene in the sea anemone *Nematostella vectensis* (Nv) was added since this species was shown to contain ancestral intron positions*(12)*. It was found that 1,488 *C. crispus* genes with orthologs (BRH) in one angiosperm (*A. thaliana* or *V. vinifera*), one green alga (*O. lucimarinus* or *M. pusilla*) and the moss *P. patens* were identified, among which 355 (23.8%) contain introns. This proportion is higher than for all genes (12%) and reflects the tendency of intron-containing genes to be more conserved than monoexonic genes. The protein sequences from the 355 sets of orthologs were aligned using MUSCLE*(13)* and the highly conserved blocks were identified using Gblocks*(14)* (parameters: -p=t -s=n –b5=a –b2=5 –b1=5 –b3=6). Intron positions were then mapped in the alignments so that a conservation profile was assigned to each intron, listing in which species it is found: profile "Cc" means the intron was only found in *C. crispus*, profile "Cc, At, Vv, Mp, Ol, Pp, Nv" means the intron is conserved among all seven species compared. We retained introns that were either in Gblocks-selected blocks (well conserved parts of the alignment) or found in at least four species among the five species we compared, and filtered out intron positions that were distant of less than five amino acids, since we could not rule out alignment issues such as gaps around the introns. The 355 proteins contain 746 introns in *C. crispus*, 251 of which are validated by cDNA sequences. Their conservation profiles are shown in Table S1.4. Most introns are specific to *C. crispus*: they were either lost in other lineages or gained in the *C. crispus* lineage. Among the 81 introns that are conserved in other species, 79% are shared with sea anemone, 69% with angiosperms, 63% with moss, and 27% with green algae.

We also investigated the characteristics of monoexonic and intron-containing genes; intron-containing genes are less compact than monoexonic genes, with longer CDS and UTR (Fig. S1.1A). Intron-containing genes also tend to be more highly expressed than monoexonic genes (Fig. S1.1B), as observed in *O. lucimarinus(15)*. Finally, intron-containing genes are generally more conserved with their orthologs in other species than monoexonic genes (Fig. S1.1C) and have more orthologs (defined as BRH) than monoexonic genes (40% and 24% respectively).

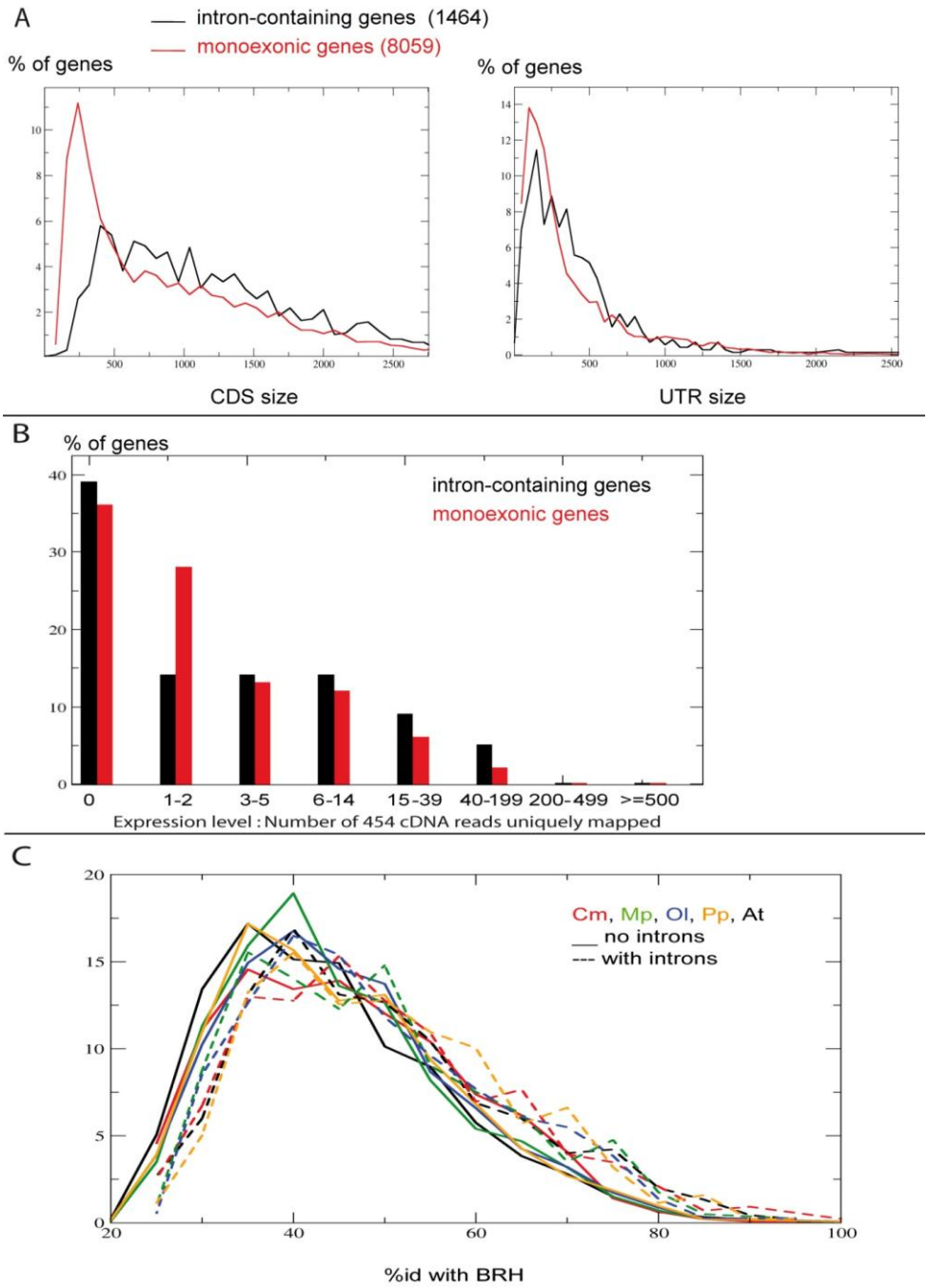

Fig. S1.1. Characteristics of intron-containing genes in *C. crispus***.** A) Frequency of genes as a function of UTR and CDS size. B) Relative expression of intron-containing and monoexonic genes. C) Conservation between orthologs and number of orthologs (defined as BRH). *Arabidopsis thaliana* (At), *Vitis vinifera* (Vv), *Physcomitrella patens* (Pp), *Ostreococcus lucimarinus* (Ol) and *Micromonas pusilla* (Mp).

Table S1.4. Conservation profiles of introns in *C. crispus. Arabidopsis thaliana* (At), *Vitis vinifera* (Vv), *Physcomitrella patens* (Pp), *Ostreococcus lucimarinus* (Ol) and *Micromonas pusilla* (Mp).

Conservation profile	Number of introns	Conservation profile	Number of introns
Cc	170	Cc,Vv	1
Cc,Nv	22	Cc,Pp,Nv	1
Cc,At,Vv,Pp,Nv	21	Cc,Ol,Nv	1
Cc,At,Vv,Pp	11	Cc,Ol	1
Cc,At,Vv,Mp,Pp,Nv	6	Cc,At,Vv,Mp,Pp	1
Cc,At,Vv,Mp,Ol,Pp,Nv	5	Cc,At,Vv,Mp,Nv	1
Cc,Vv,Mp,Pp,Nv	2	Cc,At,Vv	1
Cc,At,Vv,Nv	2	Cc,At,Ol,Pp,Nv	1
Cc,At,Vv,Mp,Ol,Pp	2	Cc,At,Mp,Ol,Pp,Nv	1
Cc,Vv,Mp,Nv	1		

In order to elucidate in more detail the pattern of intron loss in *C. crispus* and other lineages, we used the cnidarian *N. vectensis* (Nv) as an outgroup, and tagged introns present in *N. vectensis* and in at least one other species (Cc, At, Vv, Mp, Ol or Pp) as ancestral. 583 such ancestral introns were identified, among which 87% have been lost in *C. crispus* (Table S1.5).

Table S1.5 Characteristics of intron loss.

583 ancestral introns	Number of conserved introns	% of lost introns
Chondrus crispus	74	87%
Micromonas pusilla	64	89%
Ostreococcus lucimarinus	31	95%
Physcomitrella patens	436	25%
Arabidopsis thaliana	489	16%
Vitis vinifera	507	13%

Intron loss is even more pronounced in the green algal lineage (*M. pusilla*, *O. lucimarinus*) but lower in the land plant lineage (*P. patens*, *A. thaliana*, *V. vinifera*). We investigated the intron position bias on the genes from the 355 groups of orthologs that were present in the seven species*(16)* (Fig. S1.2B). We used an intragene analysis where the intron positions were mapped into a (0-1) interval from 5' to 3', and counted the number of introns at positions < 0.5 ("n5") and at positions > 0.5 ("n3"). A gene was qualified as "5'-biased" when n5 >= n3+2 and as "3'biased" when n3 >= n5+2. In *A. thaliana*, *V. Vinifera*, and *P. patens*, no significant difference between the number of 5'-biased genes and the number of 3'-biased genes was observed. However, the number of 5'-biased genes is significantly higher than the number of 3'-biased genes in *C. crispus*, *M. pusilla* and *O. lucimarinus*. Interestingly,

those lineages (red and green algae) are the lineages where the intron loss rate is the highest, suggesting that the positional bias is likely due to preferential intron loss in 3' of the genes.

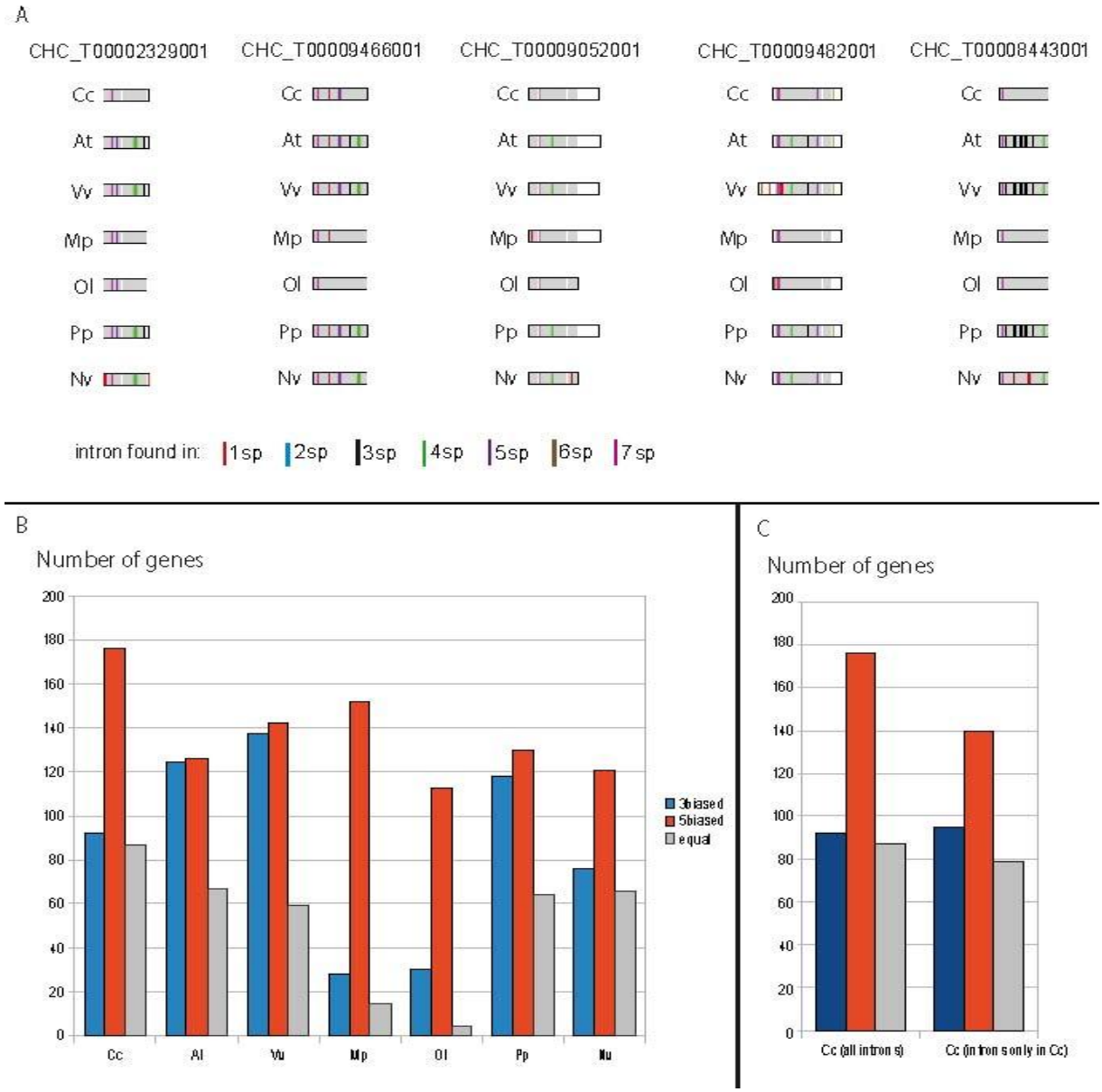

Fig. S1.2. Characteristics of intron containing genes in *C. crispus*. A) shows the five groups of orthologous proteins that contain introns that are conserved among seven species. B) Intron position bias. C) Intron position bias in genes with unique introns compared to all genes with introns. *Arabidopsis thaliana* (At), *Vitis vinifera* (Vv), *Physcomitrella patens* (Pp), *Ostreococcus lucimarinus* (Ol) and *Micromonas pusilla* (Mp).

Considering the low rate of intron loss in plants and sea anemone, it is unlikely that the *C. crispus* introns not found in any other species were lost in all the other species. They were more likely gained in *C. crispus*. Interestingly, these introns show a weaker 5' positional bias (Fig. S1.2C), suggesting a mechanism of intron gain preferentially acting in 3' of the genes like the mechanism of intron loss that may imply a reverse transcription step (17). This mechanism is thus possibly also mediated by reverse

transcription, as was proposed for reverse splicing (18) (19) (20) (21) (22). However, although this analysis was restricted to introns validated by cDNA reads, one cannot exclude that some of them are artefacts from the automatic annotation pipeline.

1.6 Contamination verification

In order to check for contaminations in the *C. crispus* genome, a two step analysis was conducted. First, the GC content of the coding sequences was calculated using the geecee program as implemented in the EMBOSS suite. The histogram of the GC content (Fig. S1.3) reveals only one local maximum, which is an indication that a contamination is not probable. The second step of the analysis was done by using MEGAN*(23)*. All CDS were blasted against the NCBI nr database with an E-value cut-off of 0.1. The results of the BLAST search were then analyzed by MEGAN for the creation of a taxonomic classification (Fig. S1.4). The MEGAN output did not hint at contaminations either.

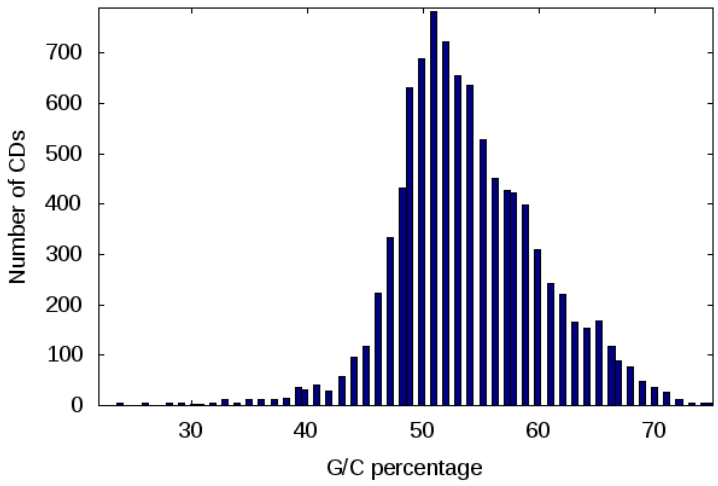

Fig. S1.3. Histogram of the G/C content of *C. crispus* coding sequences.

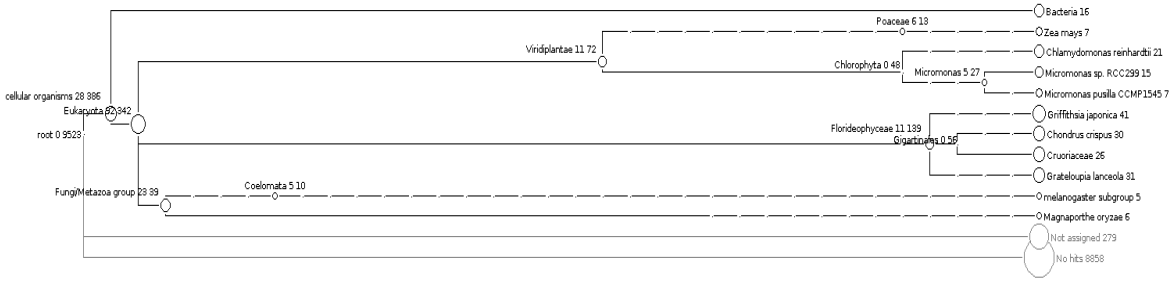

Fig. S1.4. MEGAN analysis of the *C. crispus* genome. After being subjected to Blast against the NCBI nr database, MEGAN was used to map the Blast hits against the NCBI taxonomy and summarize/order the results (bit score threshold 70.0).

2 Genome and gene structure

2.1 Plastid genome

During genome assembly four scaffolds were identified as part of the plastid genome. To obtain the full length sequence of the plastid genome, these four scaffolds were assembled with *C. crispus* EST data presenting sequence similarities with the plastid genome of *Gracilaria tenuistipitata* (24), *Porphyra purpurea* (25) and *Pyropia (Porphyra) yezoensis* (GenBank accession NC_007932). The resulting sequence (average coverage 87-fold) was further verified by direct sequencing of PCR-amplified fragments from enriched-organellar DNA when necessary, and manually annotated.

The complete plastid genome is a circular molecule of 180,086 bp (28.7 % GC), encoding three ribosomal RNA, 30 tRNA, one miscellaneous RNA, and 200 protein-coding genes.

2.2 Mitochondrial genome

The 26-kb mitochondrial genome (mtDNA) of *C. crispus* previously characterized (27) was used to identify mitochondrial sequences in the ESTs. And among the scaffolds wwo large contigs of 13 kb were obtained, corresponding to the two main transcripts of the mtDNA (28).

2.3 Organellar sequences in the nuclear genome

Both complete plastid and mitochondrial sequences were compared with the nuclear genome of *C. crispus* using blastn. For the plastid sequence, 13 nuclear scaffolds showed significant hits (e-value < 0.0001) and six hits featured an alignment with plastid sequence over 75 bp (Table S2.1).

Table S2.1. Plastid sequence in the nuclear genome. Similarities between the complete plastid sequence and the nuclear genome of *C. crispus* was investigated using blastn (e-value<0.0001).

Scaffold	**Blastn**			**cpDNA region**
	e-value	Alignment (bp)	Identities	
102	E-124	233	99%	*preA* gene
179	4,00E-40	88	100%	*tRNA-Ala*
20	1,00E-34	95	95%	*apcF* gene
248	9,00E-29	93	93%	*apcE* gene
223	1,00E-12	86	87%	*rpoA* gene
86	3,00E-10	86	86%	*rpoA* gene

In contrast, the comparison with the mitochondrial sequence revealed longer hits on the nuclear genome (Table S2.2). Among them, a ~2 kb fragment of the mitochondrial genome was located on the

nuclear scaffold 8. The alignment of both sequences showed that the mitochondrial-derived nuclear sequence is first inverted and split by a ~20 kb long insertion that contains putative LTR-retrotransposon sequence features. This portion of the nuclear genome illustrates a mitochondrial incorporation through both recombination and LTR-retrotransposon action.

Table S2.2. Mitochondrial sequences in the nuclear genome. Similarities between the complete mitochondrial sequence and the nuclear genome of *C. crispus* using blastn (e-value<0.0001).

Scaffold	Blastn			mtDNA region
	e-value	Alignment (bp)	Identities	
8	0	656	95%	cob-sdh3
8	0	586	93%	cob-sdh3
8	E-156	338	95%	cob-sdh3
8	E-123	254	97%	cob-sdh3
8	2,00E-56	122	98%	cob-sdh3
8	5,00E-29	80	96%	cob-sdh3
288	E-126	260	96%	rps12-trnfM
155	1,00E-82	214	94%	cox2-cox3
121	2,00E-77	157	98%	cox1
20	1,00E-54	111	100%	orf172
243	6,00E-41	121	91%	nad6
103	3,00E-33	75	100%	nad6
13	5,00E-20	93	98%	stem-loop

2.4 Codon usage

Codon usage was investigated using INteractive Codon usage Analysis (INCA) (29) All codon combinations were used for protein synthesis by *C. crispus* and in most cases with relatively low bias (Table S2.3). The codon usage was similar to *Gracilaria tenuistipitata* (30), while a more even usage has been reported for *Pyropia yezoensis* (31).

Table S2.3. Codon usage in *C. crispus*.

AA	Codon	%	AA	Codon	%	AA	Codon	%	AA	Codon	%
Lys	AAA	35	Gln	CAA	46	Glu	GAA	44	Stop	TAA	26
Lys	AAG	65	Gln	GAG	54	Glu	GAG	56	Stop	TAG	34
Asn	AAC	57	His	CAC	53	Asp	GAC	54	Tyr	TAC	60
Asn	AAT	43	His	CAT	47	Asp	GAT	46	Tyr	TAT	40
Arg	AGA	14	Arg	CGA	16	Gly	GGA	24	Stop	TGA	40
Arg	AGG	15	Arg	CGG	14	Gly	GGG	24	Trp	TGG	100
Ser	AGC	18	Arg	CGC	27	Gly	GGC	32	Cys	TGC	62
Ser	AGT	12	Arg	CGT	14	Gly	GGT	20	Cys	TGT	38
Thr	ACA	24	Pro	CCA	23	Ala	GCA	23	Ser	TCA	15
Thr	ACG	28	Pro	CCG	32	Ala	GCG	26	Ser	TCG	19
Thr	ACC	26	Pro	CCC	25	Ala	GCC	30	Ser	TCC	19
Thr	ACT	22	Pro	CCT	20	Ala	GCT	21	Ser	TCT	17
Ile	ATA	17	Leu	CTA	9	Val	GTA	13	Leu	TTA	7
Ile	ATC	43	Leu	CTG	22	Val	GTG	31	Leu	TTG	20
Ile	ATT	40	Leu	CTC	24	Val	GTC	33	Phe	TTC	51
Met	ATG	100	Leu	CTT	18	Val	GTT	23	Phe	TTT	49

2.5 Transposable elements

The transposable elements in the genome of *C. crispus* were identified by combining a series of existing programs and filtering the output with in-house scripts. The outcome was used to identify reference transposable elements that represent distinct families. We used the reverse transcriptase to perform a phylogenetic study of the different families (Fig. S2.1). The results suggest that *C. crispus* is in the middle of a very active transposition event. This analysis also shows that while elements diverge, they generally retain their capacity to transpose, resulting in complicated family structures.

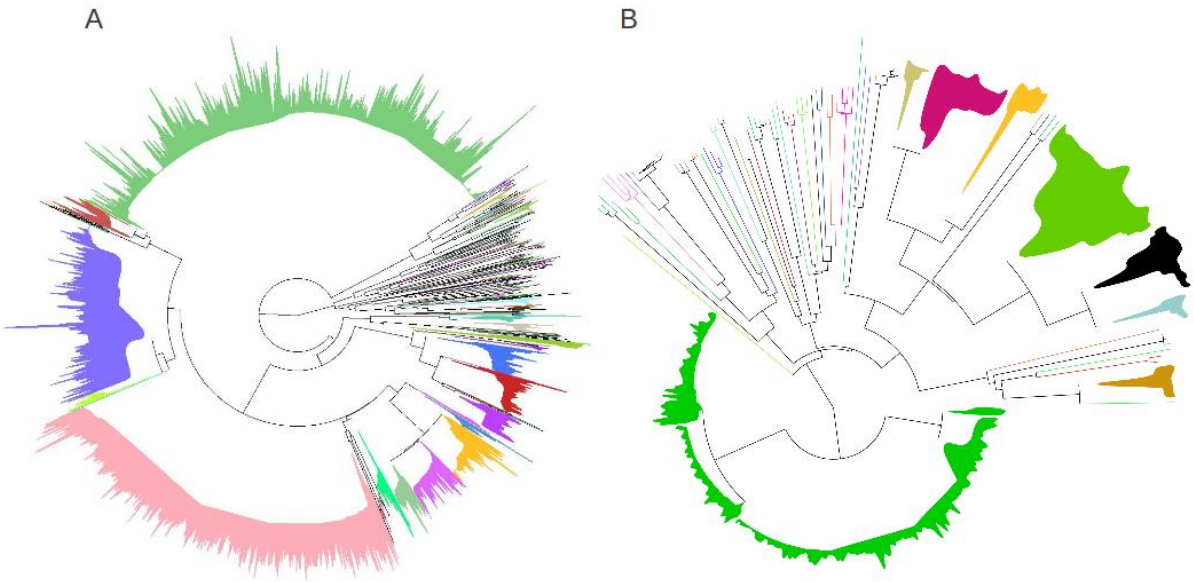

Fig. S2.1 Phylogenetic analysis of the reverse transcriptase in *C. crispus*. A) Gypsy tree showing in darker green the gyp2/gyp19/gyp35/gyp58 family and in pink the gyp4/gyp14 family. B) Copia tree showing the ccr-rlc-cop22 family in darker green.

The most abundant elements are class I LTR retrotransposons comprising ~ 58 Mb of the genome (Table S2.4). This class shows a very recent transpositional burst that is still active and is responsible for increasing the size of the genome by ~18 Mb. Retrotransposons are an extremely complex group, not only due to the large amount of recently transposed elements, but also because each family has several active copies that have diverged and form what we call "cousins". To estimate the time of occurrence of the last burst, we calculated the similarity of LTRs of the complete elements that were identified while constructing the protogenome. The histogram of the similarity between LTRs shows a unimodal curve (Fig. S2.2) suggesting that the transposition of all the elements (copias, gypsies and LARDs confounded) has occurred concomitantly and is very recent. The average similarity is 98% and more than 100 elements have identical LTRs. This allows us to date the transposition to the last 330,000 years and clearly shows that the transposition burst is still active.

Table S2.4. Retro LTR coverage in *C. crispus*.

Superfamily	Coverage (kb)	Count	Families	Whole elements
Copia	8,312	3,633	74	353
Gypsy	25,517	8,435	97	916
LARD	23,638	16,503	222	418

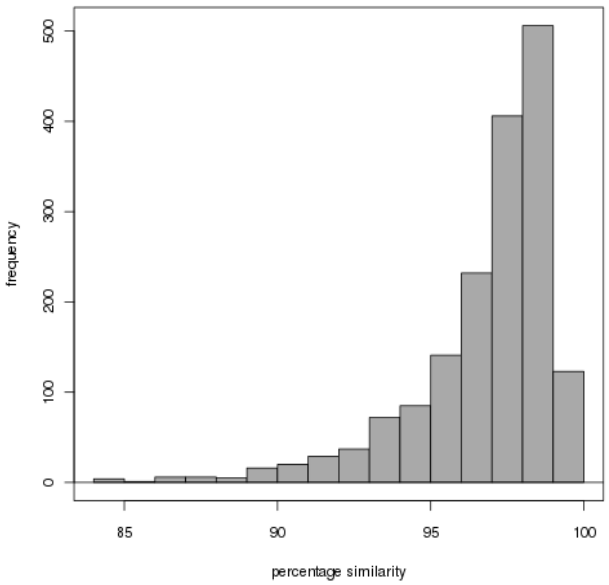

Fig. S2.2. LTR similarity in *C. crispus*. Histogram represents the similarity between LTRs of all complete elements identified in the genome.

We manually selected reference elements to represent each family, producing 74 elements for the copia superfamily and 97 for the gypsy superfamily. As stated above, several of these families are considered as "cousins", nevertheless they show enough divergence to be considered as separate families even if they have a common ancestor. The inclusion of "cousins" as reference elements greatly improved the annotation process as it allows identifying more precisely the complete elements and their identity. These cousins not only show an overall divergence along the elements but also marked differences in LTR sequences with some (active) elements containing only portions of the original LTRs (truncated LTRs).

We observe very few insertions/deletions (indels) when comparing elements from the same family. This could be due to the recent insertion of the elements, hindering the identification of older members of the family which might show a more common pattern of indels. Another explanation might be that the haploid phase of the life cycle reduces the recombination events. It is worth noting that the sizes of the different families of Copia elements are remarkably similar compared to those of the Gypsy families (Figs. S2.3 & S2.4).

The phylogenetic analysis of the Copia superfamily showed an ancient transpositional history that has created diversity in the family as well as evidence of a burst of transposition. Within the Copia superfamily, we distinguish eight major families that are actively transposing, and especially ccr-rlc-cop22 which contributes the most to the recent increase in size of the genome. The Gypsy families are more numerous than the Copia families, as it is the case for most other plants, and 18 major Gypsy families have been important in the last transpositional burst. The gyp4/gyp14 and

gyp2/gyp19/gyp35/gyp58 families have contributed the most to the last increase in size of the genome. No LINE or SINE elements pertaining to the non-LTR retrotransposons were found.

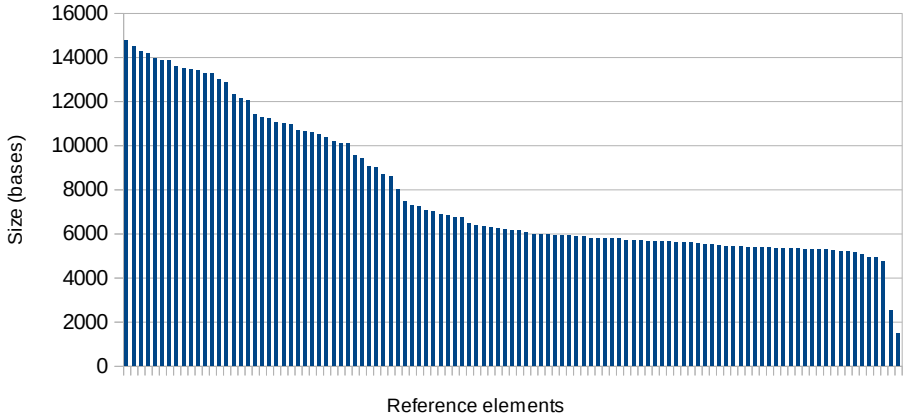

Fig. S2.3 Gypsy size distribution**.** Sorted sizes of reference elements for each identified family of Gypsy elements. About half of the elements have an average size between 5-6 kb.

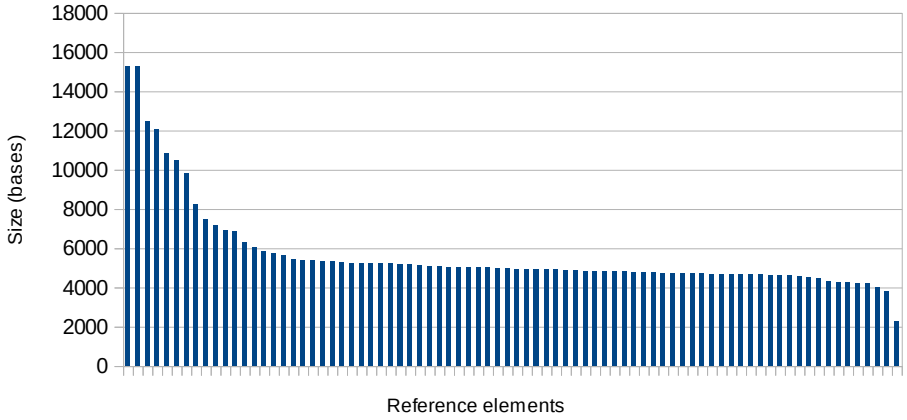

Fig. S2.4. Copia size distribution. Sorted sizes of reference elements for each identified family of Copia elements. About three quarters of the elements have an average size of ~5kb.

We found 21 families of class II-terminal inverted repeat transposons representing ~13 Mb of the genome, and ~10 Mb are contained in LTR retrotransposons. The most diversified superfamilies were Mariner, PiggyBac and Harbinger. The SMAR2 like element was the most abundant, representing ~3 Mb of the genome. We found one family of helitron elements which has actively transposed in recent times with twelve near identical copies, and representing at least 1 Mb of the genome.

2.6 Selenoproteins

No selenoproteins were identified, although some homologues with cysteine in place of selenocysteine were found for some families: glutathione peroxidases (GPx), methionine sulfoxide reductase A/B (MsrA, SelR), selenoprotein families O and W (SelO, SelW). The majority of the components of the

selenoprotein machinery (factors necessary in a species for selenocysteine production and insertion) are not present. We conclude that *C. crispus* does not synthesize selenoproteins.

2.7 Large scale duplication analysis

A large scale duplication analysis was performed using a paralog Ks plot (32). The histogram of the distribution of Ks values (Fig. S2.5) revealed three secondary peaks that might represent segmental/chromosomal duplication events. Cluster members from the three peaks were analyzed for their number of BLAST hit occurrences in the scaffolds. For each peak, members of the Ks clusters with the highest number of occurrences were selected, blasted against the genome and the hits subjected to restrictive filtering (80% identity, alignment covering 80% of the query length). After filtering, the queries were annotated by blast analyses, revealing the presence of many P-loop NTPase domains (Table S2.5). This suggests that the three secondary peaks arose from paralog acquisition and retention through segmental duplications or transposon activity.

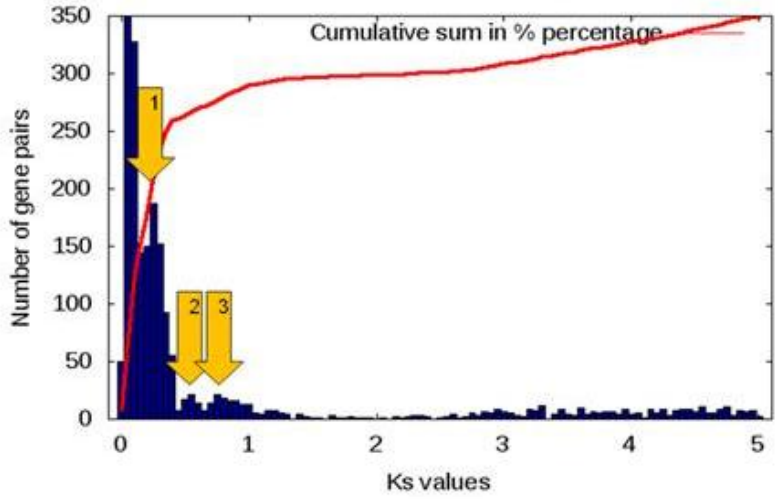

Fig. S2.5. Histogram of the distribution of Ks values in *C. crispus* gene pairs. The height of the bars corresponds to the number of gene pairs in the respective bin. The red line is the cumulative sum and the y-axis for that is from 0 to 100%. Arrows mark secondary peaks 1, 2 and 3.

Table S2.6 Results of the BLAST-annotated queries or scaffolds contributing to the secondary Ks histogram peaks.

Query	Ks_cluster_size	#hits*	annotation		
CHC_T00008974001	38	1	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00009077001	38	3	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00008853001	45	13	ATP-dependent DNA helicase PIF1	3	P-loop NTPase
CHC_T00008272001	38	10	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00008293001	38	1	WD repeat-containing protein alr3466	3	WD40 repeats
CHC_T00008306001	45	4	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00008492001	45	8	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00005954001	28	9	Retrovirus-related Pol polyprotein	1	Integrase, reverse transcriptase, Rnase H
CHC_T00008901001	1	4	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00008636001	38	5	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00008426001	38	3	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00008351001	38	3	Apoptotic protease-activating factor 1	1	P-loop NTPase, Caspase activation
CHC_T00005001001	45	14	Uncharacterized protein 030L	1	P-loop NTPase
CHC_T00009313001	45	14	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00009097001	45	14	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00008991001	45	6	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00008751001	45	4	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00009023001	45	18	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00009432001	45	6	ATP-dependent DNA helicase PIF1	1	P-loop NTPase
CHC_T00009042001	45	1	Uncharacterized protein 030L	1	P-loop NTPase
CHC_T00008278001	11	3	Vegetative incompatibility protein HET-E-1	2	P-loop NTPase, WD40 repeats
CHC_T00009514001	38	8	Vegetative incompatibility protein HET-E-1	2	P-loop NTPase, WD40 repeats
CHC_T00009419001	38	4	Vegetative incompatibility protein HET-E-1	2	P-loop NTPase, WD40 repeats
CHC_T00008903001	38	9	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats
CHC_T00008769001	38	9	Vegetative incompatibility protein HET-E-1	3	P-loop NTPase, WD40 repeats

2.8 Transcription factors and transcriptional regulators

We used a set of rules as described previously (33) to annotate all transcription associated proteins (TAPs), comprising transcription factors (TF), and transcriptional regulators (TR), in the genome. The total complement of TAPs consists of 193 proteins, which is surprisingly small. Correlation between TAP complement and morphological complexity (33) shows that the morphological complexity of *C. crispus* is not in accordance with its small TAP complement. However, cross species partial least squares (PLS) analysis (performed using Genedata Analyst 2.1) of TAP family sizes using numbers of different cell types as the activity vector predicts six cell types for *C. crispus*, a value close to what is expected based on morphology of this red alga (cortical, medullar, several types of reproductive cells).

By comparing the TAP families that are present in *C. crispus* with the ones present in the genome of *C. merolae* and in other red algae, we observe that 13 families deviate in *C. crispus*: eight are absent, two are bigger, and three are uniquely present in *C. crispus* (Table S2.6), mirroring the compact genome situation. The C2H2 zinc finger TF family is twice more abundant in *C. crispus* genome compared to that of *C. merolae*, and contributes significantly to the loadings of the 1st component (explaining 87.6% of covariance) in the above-mentioned PLS, suggesting the involvement of this family in the evolution of multicellularity in *C. crispus*.

Table S2.6 TAP families that deviate in *C. crispus* as compared to other red algae.

TAP family	Deviation
Argonaute	present
C2C2_CO-like	present
C2H2	bigger
Dicer	present
Sin3	bigger
ARID	absent
CCAAT_HAP3	absent
Coactivator p15	absent
HSF	absent
MADS	absent
MED6	absent
Zinc finger, AN1 and A20 type	absent
Zinc finger, ZPR1	absent

3 Evolutionary studies

3.1 Comparison with other red algae.

The predicted proteins from *C. crispus* genomic, mitochondrial and chloroplastic DNA, were compared with the 5,064 proteins from *Cyanidioschyzon merolae* (http://merolae.biol.s.u-tokyo.ac.jp/), the 23,961 predicted proteins of *Calliarthron tuberculosum* (34), and the 839 proteins of *Pyropia* (*Porphyra) yezoensis* present in GenBank. This set of red algal proteins was completed with amino acid sequences deduced from 22,431 ESTs of *P. yezoensis* and from 36,167 ESTs of *Porphyridium cruentum(34).*

The orthology relationship among these different sets were searched using blastp (35) and orthomcl (36) (37). The ESTs were filtered with ESTscan (38) and the resulting proteins used with orthomcl. Since the data for *C. tuberculosum*, *P. yezoensis* and *P. cruentum* are heterogeneous and partial, all predicted proteins have been considered as a single set.

Based on this orthology relationship, research of synteny has been conducted between the only two species with complete known genome sequence (*C. crispus* and *C. merolae)*. Position of each gene with the surrounding genes up to the 5th flanking position has been checked on both genomes to identify potential physical co-localization of genetic loci; 227 genes were found to be co-localised. The vast majority of the co-localised genes correspond to plastidial and mitochondrial genes. Of the 197 plastidial genes, 134 have an identified ortholog and 124 are syntenic. Of the 29 mitochondrial genes, 14 have an identified orthologs with *C. merolae* and 12 are syntenic. On the nuclear DNA, 44 pairs of genes have been identified to be syntenic with their orthologs, and only one group of three genes seems to be conserved.

3.2 The phylome

A phylome has been defined as the complete collection of phylogenies for all genes encoded in a given genome*(39),* and informs on the evolution of a genome from the perspective of all of its genes. It thus constitutes a valuable resource for inquiring about evolutionary processes. We reconstructed the *C. crispus* phylome in the context of 21 fully-sequenced plant genomes, including the genome of *Cyanidioschyzon merolae*, and five fungi/metazoan genomes (used as out-groups). Details are available at http://phylomedb.org/phylome_96. All trees and alignments have been deposited in PhylomeDB*(40)* (http://phylomedb.org). The pipeline used to reconstruct the phylogenetic tree is described below.

First, a Smith-Waterman (41) search was used to retrieve homologs using an e-value cut-off of 10E-5, and considering only sequences that aligned with a continuous region representing more than 50% of the query sequence. Once the sets of homologous sequences were defined, the phylogenetic trees were reconstructed (40). In brief, selected sequences were aligned using three different programs: MUSCLE v3.7 (42), MAFFT v6.721b (43), and DIALIGN-TX (44). Alignments were performed in

forward and reverse direction, i.e. using the Head or Tail approach (45), and the six resulting alignments were combined using M-COFFEE (46). The resulting combined alignment was subsequently trimmed with trimAl v1.3 (47), using a consistency score cut-off of 0.1667 and a gap score cut-off of 0.1, to remove poorly aligned regions.

Phylogenetic trees based on Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model. The selection of the evolutionary model best fitting each protein family was performed as follow. A phylogenetic tree was reconstructed using a Neighbor Joining (NJ) approach as implemented in BioNJ (48); the likelihood of this topology was computed, allowing branch-length optimization, using eight different models (JTT, LG, WAG, Blosum62, MtREV, VT, CpREV and DCMut), as implemented in PhyML v3.0(49). The two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion (50). Then, ML trees were derived using the two best-fitting models with the default tree topology search method NNI (Nearest Neighbor Interchange), and the one with best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been shown previously to be highly accurate (40). Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four categories plus invariant sites was used, estimating the gamma parameter and the fraction of invariant positions from the data. The resulting phylome comprises 4,451 gene trees, which represent 48% of the predicted C. crispus genes.

3.3 Phylogeny-based orthology prediction

We ran phylogeny-based orthology predictions to find orthologs in other species (51). To facilitate genome annotation, we searched for genes that had one-to-one orthology relationships with the model species *Arabidopsis thaliana*. Of the 1,023 one-to-one orthologs (around 11% of the genome), 918 mapped to an *A. thaliana* gene with some GO annotation. These orthology relationships and annotations are provided in the Supplementary Tables, Table S1.

3.4 Species tree

Evolutionary relationships among the species included in our analyses were inferred using two complementary approaches. Firstly, a super-tree was inferred from all the trees in the phylome by using a gene tree parsimony approach, as implemented in the dup-tree algorithm (52) that finds the species topology that minimizes the number of total duplications implied by a collection of gene family trees, *i.e.* the phylome. Secondly, 19 gene families, with a clear phylogeny based one-to-one orthology in at least 20 of the 27 species included in the analyses, were used to perform a multi-gene phylogenetic analyses. Protein sequence alignments were performed as described above and then

concatenated into a single alignment. Species relationships were inferred from this alignment using a ML approach as implemented in PhyML, using LG as evolutionary model, since for 16 out of 19 gene families this model was the best-fitting. Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution. Both trees (Fig. S3.1) are largely congruent and place the red algae genomes outside Viridiplantae, supporting the monophyly of Rhodophyta and Viridiplantae.

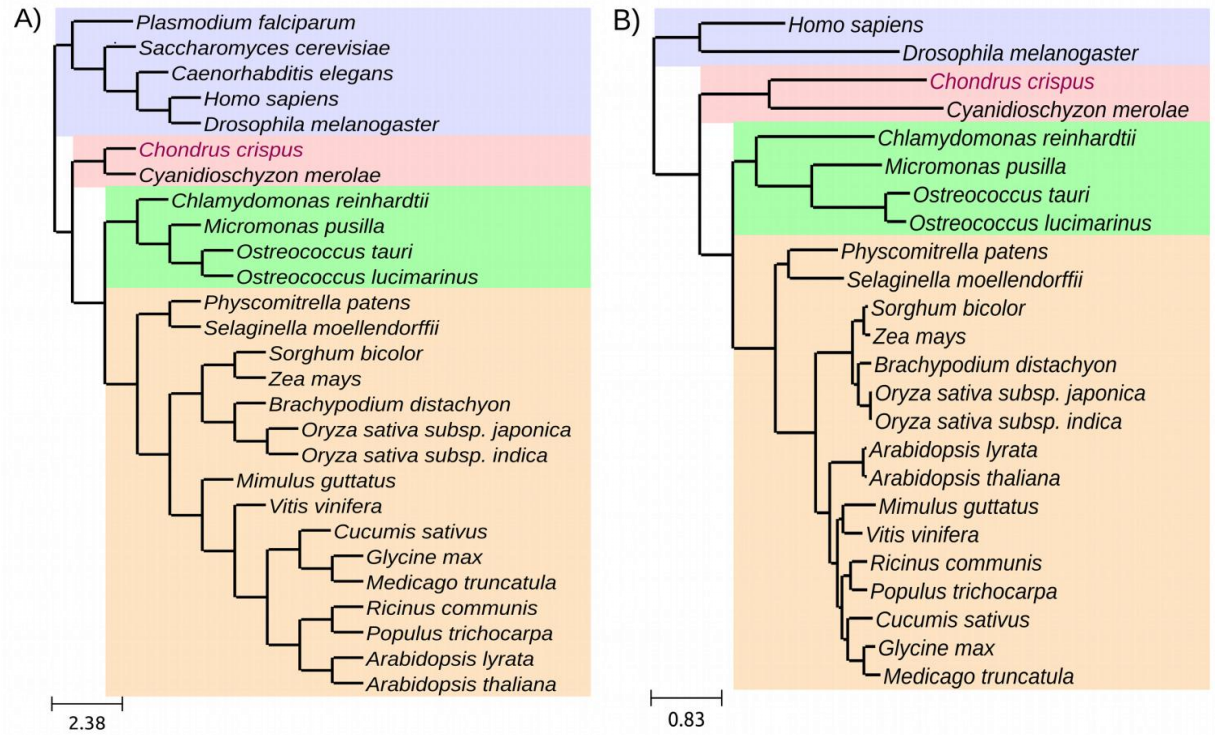

Fig. S3.1. Phylogeny of *C. crispus* and other genome-sequenced plants species. Colours correspond to the main clades: blue (out-groups), green (Chlorophyta), pink (Rhodophyta) and brown (Streptophyta). A) Super-tree, including the 27 species used in the analysis, inferred from all single trees (4,451) available in the *C. crispus* phylome using a gene tree parsimony approach as implemented in dup-tree algorithm. B) Concatenate tree (some out-groups species were removed) reconstructed using 19 concatenated genes with a clear phylogenetic one-to-one orthology relationships in at least 20 of 27 species from the *C. crispus* phylome. Species (*Plasmodium falciparum*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*) in less than 10 genes were removed from final tree.

3.5 Lineage specific duplications

We scanned all trees in the *C. crispus* phylome to detect duplication nodes at the Rhodophyta-specific and *C. crispus*-specific lineages. 2,332 genes (~ 25%) of the *C. crispus* genome result from either Rhodophyta-specific (663 genes; see green bars graph, Fig. S3.2A), or species-specific expansions (1,669 genes; see blue bars graph, Fig. S3.2B) Among these gene families of various sizes, we focused on the eleven biggest expansions, which include 20 or more paralogs, and predicted their putative

function based on the domain composition of the proteins (Supplementary tables, table S2). At least five cases correspond to putative transposases that were not masked in the gene prediction phase. This reinforces the observation that *C. crispus* has active retrotransposons. One of them corresponds to the second biggest expansion (81 copies), and seems to be a Tc3-type or Tc1-type transposase; another one has some similarity with the Harbinger transposase family. For others, no similarity can be detected to known families but retrotranscriptase/integrase domains are apparent.

Additional expansions are clearly non-transposase related and correspond to real expanded gene families with a potential important role in *C. crispus*. Interestingly, the biggest expansion (88 copies) has similarity to disease resistance proteins, while another one seems to encode a histone H3-like protein. For the rest, the function is unclear but protein kinase or TOLL-receptor like domains can be recognized.

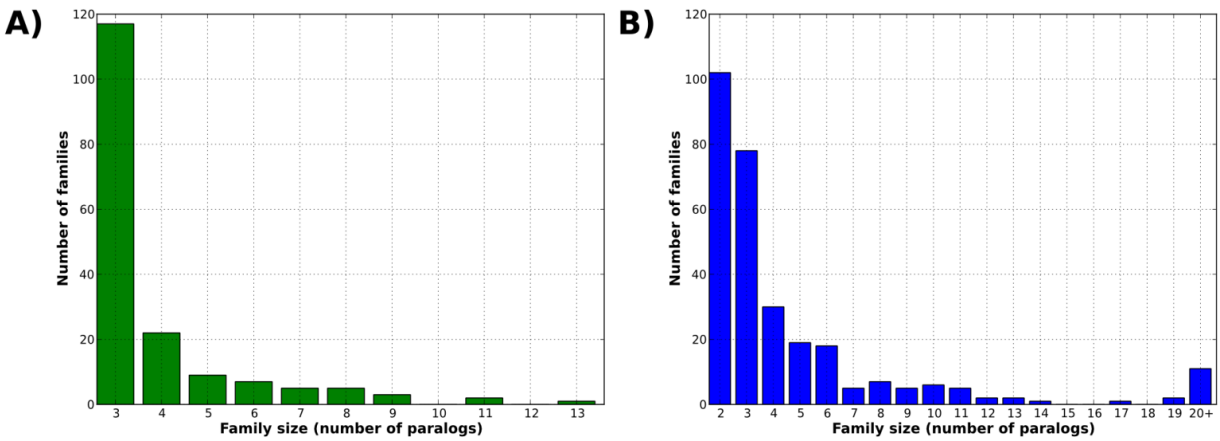

Fig. S3.2. Distribution of sizes for protein families that expanded specifically within the *C. merolae* genome (green bars, panel A) and in that of *C. crispus* (blue bars, panel B). Height of the bar indicates number of in-paralogs resulting from the each kind of expansion. Expansions at the red algal level were removed from the *C. crispus* level, if necessary, to avoid over-estimation of expanded protein families.

3.6 Comparative genome analysis

We compared the *C. crispus* genome with a set of other fully-sequenced genomes of eukaryotes (219) and prokaryotes (1,200) downloaded from the KEGG database (53) (Fig. S3.3). Given the large evolutionary distance among species, we used an e-value cut-off of 1E-3 for the Blast Search (Smith-Waterman algorithm), and only considered sequences that aligned with a continuous region representing more than 30% of the query sequence. Remarkably, for 4,301 genes (~ 45 %) of the genome, we could not detect any homology in other sequenced genomes. This seems to arise from a large degree of sequence divergence, as the plot of e-values from *C. crispus* best blast-hits against sequenced plant genomes is similar to that of the distant species *Homo sapiens* (Fig. S3.4), suggesting

that the level of substitutions in *C. crispus* sequence may have placed many real homologs below the threshold level. Interestingly, similar results were obtained for the *C. merolae* best blast-hits.

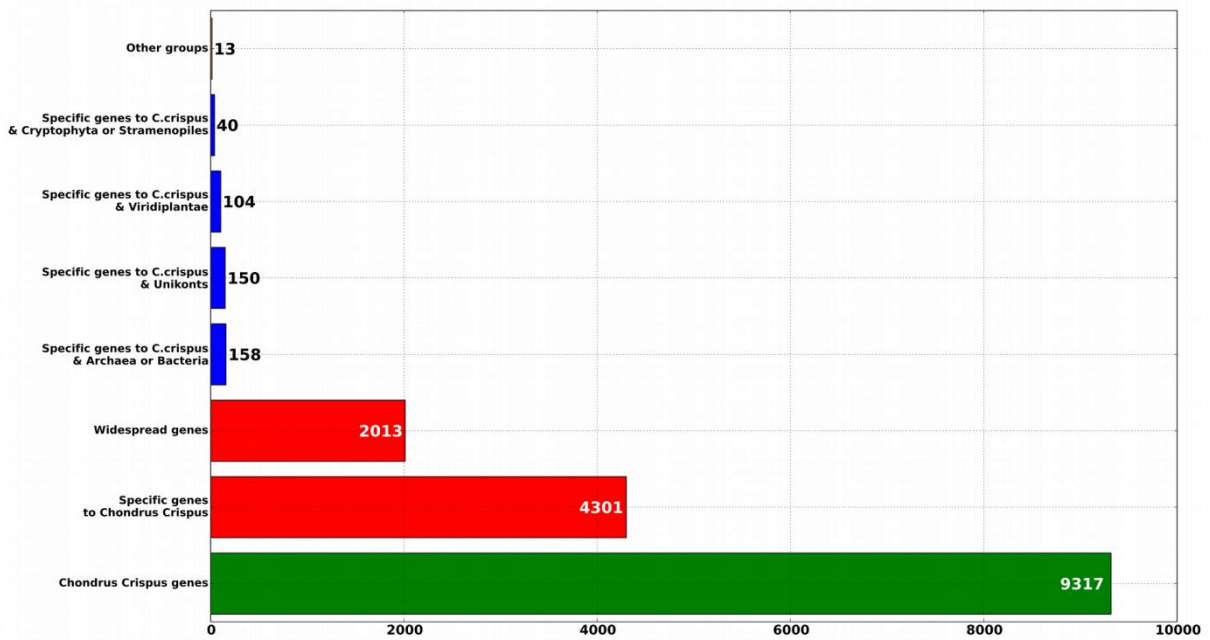

Fig. S3.3. Comparative analysis of the *C. crispus* genome against other fully sequenced eukaryotes (219), including *C. merolae*, and prokaryotes (1,200) species. Green bar shows total number of *C. crispus* genes; genes with the same sequence (206) were collapsed into a unique sequence. Red bars show genes specific to *C. crispus*, *i.e.* genes with no detectable homologs, and wide-spread genes, with at least a hit (e-value < 10^{-3}) within each taxonomic class. Blue bars show number of *C. crispus* genes with homologs only with specific taxonomic group. Groups with less than 10 specific hits were collapsed into a single group.

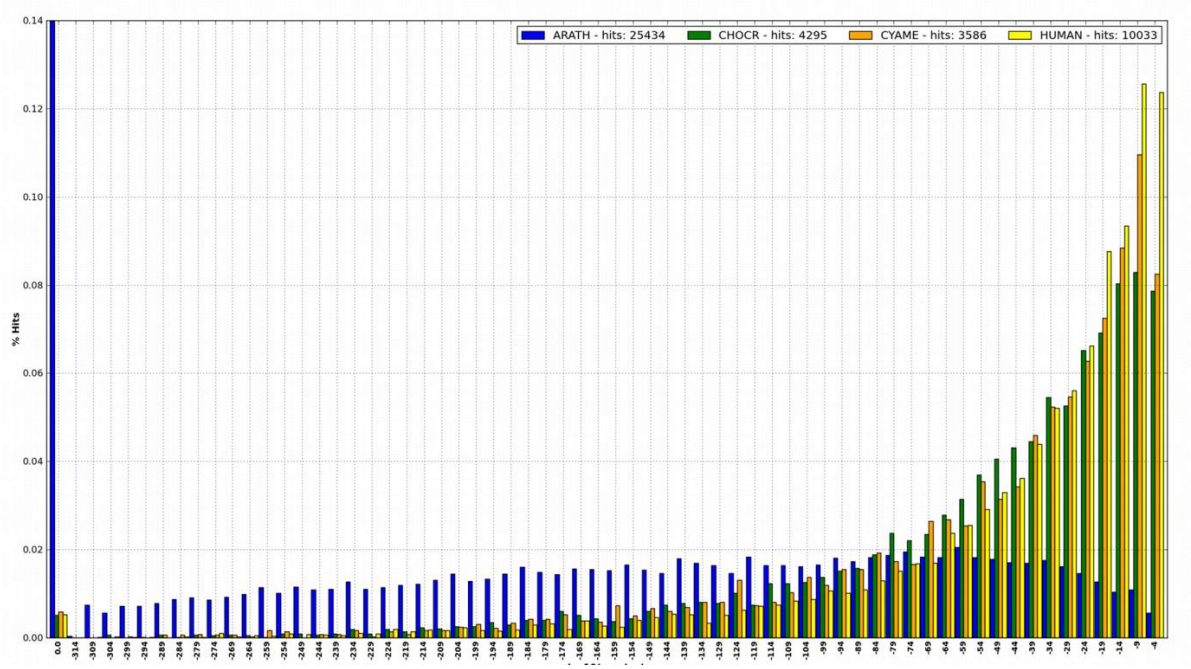

Fig. S3.4. Distribution of best blast-hits e-value for *A. thaliana* (ARATH), *H. sapiens* (HUMAN), *C. crispus* (CHOCR) and *C. merolae* (CYAME) proteins against the 22 plants species from the *C. crispus* phylome. Best blast-hits within the same species were not taken into account for *A. thaliana*, *C. crispus* and *C. merolae*.

Some interesting patterns can be recognized from the distribution of blast-hits (Fig. S3.5). For instance, links between eukaryotic species that underwent secondary endosymbiosis with ancestral red algae (stramenopiles and alveolates) is clearly apparent from the amount of shared genes. In addition, genes with homologs in cyanobacteria and α-proteobacteria may represent descendants of the plastid and mitochondrial symbioses. Several small sets of genes are shared between *C. crispus* and prokaryotes, while being absent in the rest of eukaryotes, or shared between *C. crispus* and specific group of eukaryotes. These are potential candidates for horizontal gene transfers.

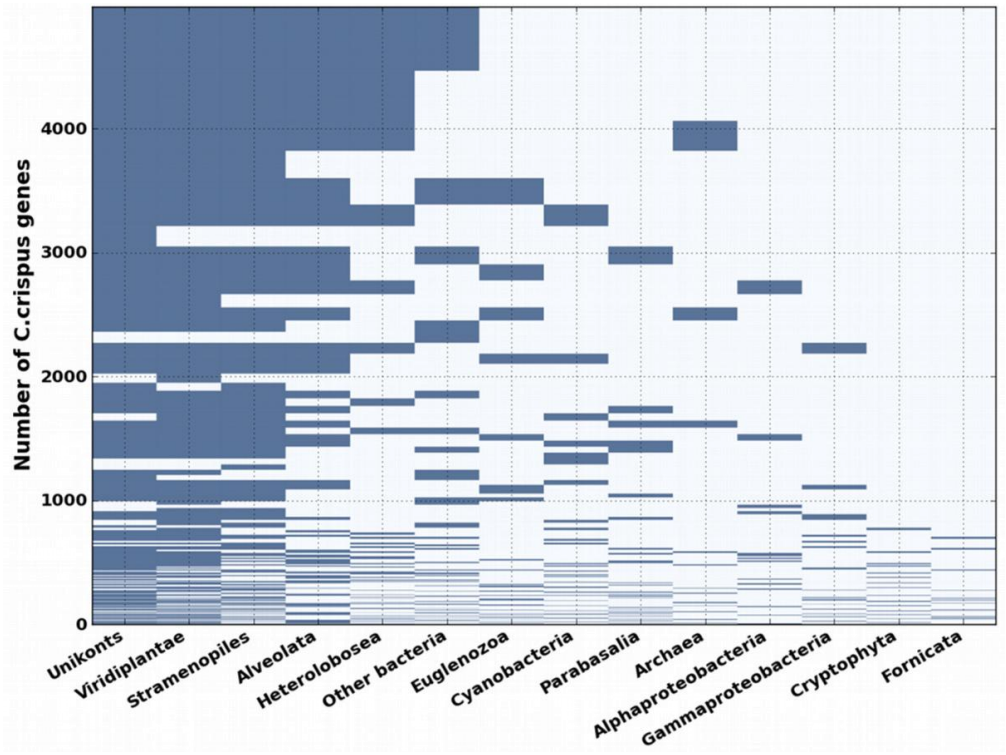

Fig. S3.5. Comparative genomics analysis across eukaryotes and prokaryotes species. Plots show phylogenetic profiles summarized with different combinations of taxonomic groups for each *C. crispus* gene with at least one homolog in the considered group.

Finally, we explored the fact that *C. crispus* seems to have fewer genes than other organisms (*i.e.* unikonts, Viridiplantae and stramenopiles) with similar complexity. This could be partly explained by a larger degree of diversification (by gene duplication) in the lineages leading to Viridiplantae and unikonts. To test this, we compared the ratio of "orthologs in species X" vs. "orthologs in *Chondrus*" for all the families. To avoid the effect of outliers, we removed families with more than ten members

in any of the given species pair. As shown in the Supplementary tables, table S3, most ratios are less that 1, indicating higher levels of diversification for Viridiplantae.

3.7 Horizontal transfer

We focused on *C. crispus* genes featuring homology with prokaryotic sequences, *i.e.* 3,288 genes, putting special attention to those that a) have their best hits in cyanobacteria (putative descendants of the proto-plastid endosymbiont), *i.e.* 612 genes, and b) have their best hits in α-proteobacteria (putative descendants of the proto-mitochondrial endosymbiont), *i.e.* 283 genes. For all genes with homologous sequences in prokaryotes we performed additional phylogenetic analysis including relevant prokaryotes and eukaryotes homologs.

Of the 612 genes with a bacterial best blast-hit in a cyanobacterial genome, 329 showed monophyly of eukaryotic and cyanobacterial genes in a Maximum Likelihood phylogenetic reconstruction (see phylome reconstruction for details). For 78 genes of this latter group, a putative plastid targeting signal was predicted by Target-P (54) (confidence class 1 to 3). Similarly, of the 283 *C. crispus* genes with a bacterial best blast-hit in an α-proteobacterial genome, 103 show a phylogeny consistent with a direct descent from the α-proteobacterial proto-mitochondrion, and 23 of them are predicted to be located to mitochondria. Additionally, 760 *C. crispus* genes, including 63 plastidial, 74 mitochondrial, 46 secreted, show a specific phylogenetic affiliation to prokaryotic genomes other than cyanobacteria or α-proteobacteria, suggesting recent horizontal gene transfers (HGT). Yet other 394 *C. crispus* genes, including 60 plastidial, 39 mitochondrial, 13 secreted, show close relationships with prokaryotic genomes and few eukaryotic ones, potentially representing cases of ancient or multiple HGT events.

4 Replication, transcription and translation

4.1 The exosome complex

The exosome (or PM/Scl) complex (55) degrades mRNA, rRNA and many species of small RNA. It is also involved in various aspects of RNA maturation, including degradation, processing, and quality control. Furthermore, it has recently been shown that it is involved in gene silencing through its role in the regulation of the expression of many types of non-coding RNAs. The typical exosome complex consists of a core of six RNase proteins, plus three S1 RNA binding domain, and two K-homology domain proteins, consisting of nine different proteins in all. In addition, it includes two associated proteins, the hydrolytic exoribonucleases Rrp44 and Rrp6. All the known components of the exosome, including the two associated proteins, appear to be encoded by the *C. crispus* genome. There is a single gene for each component plus three additional Rrp41-like genes (Table S4.1).

Table S4.1. The exosome complex of *C. crispus*.

Exosome component (alternative name)	Locus
Csl4 (EXOSC1, Ski4)	CHC_T00009354001
Rrp4 (EXOSC2)	CHC_T00008051001
Rrp40 (EXOSC3)	CHC_T00005365001
Rrp41 (EXOSC4)	CHC_T00008657001
Rrp41-like (EXOSC4-like)	CHC_T00009211001
"	CHC_T00008824001
"	CHC_T00008945001
Rrp46 (EXOSC5)	CACHCT135000004001
Mtr3 (EXOSC6)	CACHCT641000001001
Rrp42 (EXOSC7)	CHC_T00003458001
Rrp43 (EXOSC8)	CACHCT4000002001
Rrp45 (EXOSC9)	CHC_T00006256001
Rrp6 (EXOSC10)	CHC_T00009099001
Rrp44 (DIS3)	CHC_T00008438001

4.2 Spliceosome

To get insight in the spliceosome of *C. crispus*, we performed comparative analyses based on the described human 170 spliceosome-associated protein factors*(56)*. Blast analyses and protein sequence comparisons allowed to identify about 140 genes that encode for putative orthologs of the human spliceosomal proteins, including: core snRNP proteins, proteins associated to the U RNAs, hnRNPs, SR-proteins (with some proteins more closely related to plants and other to metazoans), a large number of proteins belonging or related to the Prp19 complex, proteins involved in the exon junction complex proteins, and members of the THO complex (required for the generation of functional mRNA-protein complexes).

4.3 mRNA maturation

Here, we searched the *C. crispus* genome for genes encoding proteins involved in cleavage and polyadenylation of precursor mRNA (Table S4.2). We found homologs of the Cleavage and polyadenylation specificity factor (CPSF) complex proteins: CPSF160, CPSF100, CPSF73-I, CPSF73-II and a putative CPSF30, components of the CstF complex (CstF77 and CstF64), an ortholog of Fip1, a putative FY protein, a gene encoding a symplekin-like protein, a putative Clp1, CFIm25 cleavage factor, a canonical PAP, but we did not found a homolog of PCF-protein that is known to interact with CstF. We also identified a factor that presents homology (e-value = 8E-12) to the

Schizosaccharomyces pombe polynucleotide 5'-hydroxyl-kinase GRC3, and that might contribute to the processing of pre-rRNAs.

Table S4.2. Candidate genes coding for proteins of the CFSF and CstF complex**.**

Protein	Gene ID
CFSF complex protein	
CPSF160	CHC_T00007588001
CPSF100	CHC_T00003473001
CPSF73-I	CHC_T00002052001
CPSF73-II	CHC_T00000489001
CPSF30	CACHCG67000007001
CstF complex	
CstF77	CACHCG73000001001
CstF64	CACHCG19000010001[f]
FY protein	CHC_T00004189001
symplekin-like protein	CHC_T00007172001
Clp1	CHC_T00005930001
CFIm25 cleavage factor	CHC_T00003836001
canonical poly(A) polymerase	CHC_T00004050001
PCF-protein	Not found
possible 5'-hydroxyl-kinase GRC3	CHC_T00007178001

Altogether, these bioinformatic analyses revealed that the machinery involved in pre-mRNA 3'-end processing of is closely related to the plant machinery (57), but with a single copy of each gene. CPSF73-II and FY have been associated with embryo development and flowering time controls in plants (58).

A recent analysis of eight species in the plant lineage has revealed a striking evolutionary history of canonical PAPs (59); these proteins are members of the large group of DNA polymerase β-like nucleotidyltransferases (60). It was proposed that the plant PAP gene family expanded via a series of duplications, followed by change in function of gene copies. This analysis also revealed that *Chlamydomonas reinhardtii* possesses a single PAP gene, while the mosses *Physcomitrella patens* and *Selaginella moellendorffii* possess two possible PAP genes (59). It was also shown that various angiosperms possess between four and six putative PAP genes (61). Since canonical PAPs could play an important role in plant growth and development, or mammals in spermatogenesis, we analysed the canonical PAPs in algae. We found that all unicellular chlorophytes (*C. reinhardtii*, *Volvox carteri*, *Micromona*s sp. RCC299, *M. pusilla*, *Ostreococcus lucimarinus*, *O. tauri*, *O.* RCC809, *Chlorella*

variabilis NC64A and *Coccomyxa* sp. C-169), the haptophyte *Emiliania huxleyi* CCMP1516, and the freshwater amoeboflagellate *Naegleria gruberi* present a single gene encoding for a canonical PAP. We also found that duplication of PAPs has occurred in Stramenopiles with two copies on the genome of diatoms (*Fragilariopsis cylindrus*, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*) of *Aureococcus anophagefferens* (Pelagophyceae), and of the brown alga *Ectocarpus siliculosus*, and four copies on that of the oomycetes (*Phytophthora capsici*, *P. infestans*, *P. ramorum* and *P. sojae*). However, on the *C. crispus* genome (GSCHC2T00008735001), as well as in the two other rhodophyte genomes available (*Cyanidioschyzon merolae* and *Galdieria sulphuraria*), we found a single gene encoding a PAP.

Canonical RNA polymerases are not the only enzymes catalyzing the addition of oligo(A) tails in eukaryotes. It was shown that PAP-related members can add poly(A) or poly(U) tails to a variety of RNAs (cytoplasmic RNAs, rRNAs, siRNAs, miR, histone, U6 snRNA, etc), and therefore carry out multiple cellular functions (62) (63) (64) (65) (66). Non canonical PAPs (ncPAPs) catalyze the template independent incorporation of ribonucleotide monophosphate via a two-metal ion catalytic mechanism (67). Their catalytic domains present the helix-turn-helix signature of the polymerase β superfamily, but in contrast to canonical PAPs, ncPAPs usually do not contain canonical RNA-binding motif (RRM), except for U6 snRNA terminal uridyl transferases (U6TUTases). Genome surveys have revealed nine ncPAPs in *C. reinhardtii(68)*, six in *S. pombe* (69), twelve in *Caenorhabditis elegans* (70), seven in *Homo sapiens* (69), eight in *A. thaliana* (68), and only two in *Saccharomyces cerevisiae* (71). By performing blast search using *S. pombe* ncPAPs, we identified only two putative ncPAPs (GSCHC2T00003048001 and GSCHC2T00016675001) in *C. crispus* genome, and only one in those of *C. merolae* (CMT581C) and *G. sulphuraria* (for which only partial sequence is available). The evolutionary history of these genes, inferred using Neighbour-joining or Maximum Likeliood methods from more than 350 sequences of ncPAPs aligned by Muscle revealed that one of the closest homologs to the rhodophyte genes (GSCHC2T00003048001 and CMT581C) is the *S. pombe* gene (SpCid1) that has been shown to be a cytoplasmic poly(A) polymerase involved in checkpoint regulation (72) (73). The second *C. crispus* gene (GSCHC2T00016675001) encodes a highly divergent ncPAP more related to the *SpCid16* gene, that has not been fully characterized (74).

4.4 DNA Replication

The *C. crispus* genome encodes representatives of all known eukaryotic replicative proteins. In eukaryotic cells, the initiation of the replication starts from multiple origins. The first step is the recognition of origins by the ORC complex which is composed of six subunits (ORC 1-6). After this event, a new complex MCM helicase binds to the origin. This complex is composed of six subunits (MCM 1-6). The ORC and MCM proteins together constitute the replisome progression complex. The GINS genes (Sld5, Psf1, Psf2 and Psf3) are also identified in the genome. These genes are essential for

the initiation and elongation steps in the replication process (75) (76) (77) (78) (79). Genes encoding for PCNA, RPA, Dna2, DNA ligase 1, and the DNA polymerases α, δ and ε are also present. These polymerases are involved in the replication of the DNA in the nucleus (80). We have also found two plant organellar DNA polymerases (POPs), involved in the replication of plastidial DNA. Both POP proteins have a plastid transit peptide in the N-terminal region. One of the proteins has some homology only with marine bacteria Polymerase I as it is the case for a POP gene coding by the nuclear genome of the red alga *C. merolae*[98 99].

4.5 DNA Repair and recombination

Algae are affected by environmental abiotic stressors like desiccation or UV-light which can induce DNA damages. Several DNA repair mechanisms have been reported (photoactivation, nucleotide excision repair, base excision repair, and mismatch repair (82), and all corresponding proteins have been identified in the *C. crispus* genome.

Another mechanism to repair damaged DNA is DNA recombination, which is also involved in replication and double strain break repair. It operates by homologous recombination or non homologous end joining, and all the genes encoding related proteins were found in the genome.

4.6 Meiosis

Meiosis appears to be an ancestral trait within eukaryotes probably present in the common ancestor of animals, plants and fungi (83). Our objective was to determine whether genes coding for proteins related to meiotic recombination were present in *C. crispus*. We focused our analysis on comparison with *E. siliculosus*, *A. thaliana*, *C. merolae* and *C. reinhardtii* (Table S4.3). The *C. crispus* genome contains homologs for most of the core meiotic recombination machine members; however, no clear candidate could be found for Hop1 and MLH3 that are included in the extended core meiosis machinery list (84). Hop1 is a meiosis-specific DNA binding protein required for homologous chromosome synapsis and chiasma formation and is present in all eukaryotes analyzed so far, except for a couple of fungi (84) and *C. merolae*. MLH3 is a DNA mismatch repair protein which is required for normal levels of meiotic crossovers in plants (85).

Table S4.3. Comparative list of genes involved in meiosis and key related-functions. Species: *A. thaliana* (At), *Ectocarpus siliculosus* (Es), *C. merolae* (Cm), and *Chlamydomonas reinhardtii* (Cr). +, presence; -, absence.

Gene	Gene ID	Function	At	Cr	Cm	Es
Spo11-1	CHC_T00006880001	Initiates double strand break formation	+	+	+	+
Mre11	CHC_T00003794001	Pairs with Rad50 to repair double strand breaks	+	+	+	+
Rad50	CHC_T00001019001	Pairs with Mre11 to repair double strand breaks	+	+	+	+
Rad51A	CHC_T00003437001	Involved in homology searching and strand exchange in homologous recombination	+	+	+	+
Rad51C	CHC_T00003437001	"	+	+	+	+
DMC1	CHC_T00003794001	"	+	+	+	+
MSH4	CHC_T00008573001	Promotes cross-over formation.	+	+	+	+
MSH5	CHC_T00006612001	Partner of MSH4 (MutS homolog 4) and promotes cross-over formation.	+	+	+	+
MLH1	CHC_T00003921001	DNA mismatch repair protein	+	+	+	+
MutL	CHC_T00005808001	Involved in DNA mismatch repair	+	+	+	+
MLH3	Not found	DNA mismatch repair protein	+	+	+	+
MSH1	CHC_T00006346001	Mismatch repair and recombination	+	+	+	+
MSH2	CHC_T00003068001	"	+	+	+	+
MSH6	CHC_T00006922001	DNA mismatch repair protein	+	+	+	+
Hop2/meu13/TBPIP	CHC_T00008571001	The Hop2-Mnd1 complex facilitates loading of Rad51 and DMC1	+	+	+-	+-
Hop1/Asy1	Not found	Required for homologous chromosome synapsis and chiasma formation	+	+	-	+
MND1	CHC_T00000509001	Usually interact with Hop2/AHP2.	+	+	+	+

4.7 Chromatin modifying genes

The chromatin modifications are performed by a wide array of genes classified based on their specific function. Genes such as argonaute and dicers which are components of RNA silencing machinery are widely present in eukaryotes. Such genes work in a close coordination with another multimeric protein group called polycomb group of proteins (PcG). Components of the PcG complex are proteins containing SET domain, which methylates the histone and leads to the chromatin modification. Table S4.4 provides a list of genes in the genome of *C. crispus* coding for this type of proteins.

Among several components which contribute to the RNA silencing machinery, the argonaute and dicer proteins are fundamental (86). A comparative analysis of such proteins in the genome of *Arabidopsis* revealed a five-fold increase in the number of argonaute proteins compared to *C. crispus*. This increase in *Arabidopsis* could be attributed to the diversification of RNA silencing genes during the course of evolution to provide new and enhanced gene control mechanisms (87).

The evolutionary conserved SET domain proteins regulate the epigenetic control of genes by either promoting or inhibiting gene expression. There is biochemical evidence to suggest that SET domain proteins can methylate histones (88) (89).

Table S4.4. Chromatin-modifying genes families.

Gene Category	*Arabidopsis thaliana*	*Chondrus crispus*	*Ectocarpus siliculosus*
SET Domain	41	5	18
RNA Interference			
Dicer	4	1	1
Argonaute	10	2	2
Histone and histone linker proteins	59	11	28

Polycomb group (PcG) proteins act as determinants of body patterning during development (90) and are required to maintain cell identity through the repression of many genes involved in alternative genetic programs (91). PcG proteins are conserved factors that regulate hundreds of different genomic loci (92). Among the three different classes of PcG proteins, polycomb repressive complex 2 (PRC 2) has been widely studied owing to its conservation across different taxa. PRC2 consists of four core proteins which are well conserved in metazoans and plants: 1) the histone methyltransferase enhancer of zeste, E(z), which serves as the catalytic subunit methylating H3K27 via its SET domain; 2) the WD40 domain-containing polypeptide extra sex comb (ESC); 3) the C2H2-type zinc finger protein suppressor of zeste 12 (Su(z)12); and 4) the nucleosome remodelling factor 55-kDa subunit (Nurf55) (93) (94) (91). The ubiquitous presence of E(z), Nurf 55, and ESC homologs in organisms belonging to independent lineages such as opisthokontes, stramenopiles, haptophytes and archaeplastides (C. crispus), supports an early apparition of PRC2 in eukaryote evolution (Table S4.5). The absence of Su(Z) in *C. crispus* indicates that this protein might have been lost during the course of evolution as in other organisms (91). PRC2 complex has also been shown to be involved in regulating the transition between gametophyte to sporophyte (95), and its presence in *C. crispus* may imply that functional regulation of alternation of generation might be partly conserved across different lineages within Archaeplastida.

Table S4.5. Polycomb related genes (+, presence; -, absence).

	Species	E(z)	ESC	Su(Z)	Nurf 55
Excavata	*Trypanosoma cruzi*	-	-	-	+
	Leishmania major	-	-	-	+
	Giardia lamblia	-	-	-	+
Stramenopiles	*Phytophthora sojae*	+	+	+	+
	Ectocarpus siliculosus	-	-	-	+
Haptophyta	*Emiliania huxleyi*	+	+	-	+
Archaeplastida	*Cyanidioschyzon merolae*	+	+	+	+
	Arabidopsis thaliana	+	+	+	+
	Chondrus crispus	+	+	-	+
	Volvox carteri	+	+	-	+
Opisthokonta	*Drosophila melanogaster*	+	+	+	+
	Caenorhabditis elegans	+	+	+	+
	Saccharomyces cerevisiae	-	-	-	+

4.8 RecQ helicase

The RecQ helicase family is actively involved in genome surveillance (96). Single RecQ genes are found in bacteria and yeast, but five to eight copies are found in plants (97). These helicases are essential for DNA metabolism; to escape effects of mutations and some organisms have duplicated one of these genes. For instance, in *A. thaliana*, two copies of RecQ4 exhibit 70% of identity (98). These proteins can dissolve telomere interactions between non homologous chromosomes in meiotic prophase I *in vivo* (99). In *C. crispus*, we found 19 genes with similarity to RecQ helicases (Table S4.6). All proteins contain the minimum domain to be active (100); twelve genes are 100 % identical in nucleotides. For most of them, the genes were found on a scaffold devoid of other genes. When present with other genes, four genes were associated with DNA adenine methylation and two with a ribonuclease II.

Table S4.6. Families of RecQ helicases in *C. crispus.*

Gene ID	Conserved domains	Homology/comments
CHC_T00007861001	Helicase//RQC	RecQ4
CHC_T00003907001	Helicase	RecQ1
CHC_T00009451001	Helicase	RecQl3
CHC_T00003907001	Helicase	RecQ family
CHC_T00008621001	Helicase	Telomere linked helicase
CHC_T00009335001	Helicase	RecQ family
CHC_T00008412001	Helicase	RecQ family
SNAPCcT00020422001	Helicase	RecQ/12 identical copies

4.9 Initiation factors

Translation of mRNA into proteins by the ribosome is universally conserved in all cellular life and can be divided into four stages: initiation, elongation, termination, and recycling*(101)*. The translation factors (Table S4.7) were identified using translation factors from *C. merolae*.

Table S4.7 Initiation factors in *C. crispus.*

Gene	Gene ID
Chloroplast translation elongation factor G (EF-G)	CHC_T00009455001
Eukaryotic elongation factor 1A EF-Tu	CHC_T00009375001
Eukaryotic elongation factor 1A EF-Tu bacteria predicted	CHC_T00008671001
Eukaryotic elongation factor 1A EF-Tu mitochondrial	CHC_T00009351001
Eukaryotic elongation factor 1B	Not found
Eukaryotic elongation factor 2 EF-G	CHC_T00009339001
Eukaryotic initiation factor 1	GSCHC2T00012203001
Eukaryotic initiation factor 1A	CHC_T00008281001
Eukaryotic initiation factor 2 α subunit	CHC_T00009249001
Eukaryotic initiation factor 2 β subunit	CHC_T00009562001
Eukaryotic initiation factor 2 γ subunit	CHC_T00008676001
Eukaryotic initiation factor 2B α subunit	Not found
Eukaryotic initiation factor 2B β subunit	CACHCG100001028102
Eukaryotic initiation factor 2B Δ subunit	GSCHC2T00013388001
Eukaryotic initiation factor 2B ε subunit	CHC_T00008286001
Eukaryotic initiation factor 2B γ subunit	CHC_T00009357001
Eukaryotic initiation factor 2C	Not found
Eukaryotic initiation factor 3a	CHC_T00008890001

Eukaryotic initiation factor 3b	CHC_T00009520001
Eukaryotic initiation factor 3c	CHC_T00009117001
Eukaryotic initiation factor 3d	CHC_T00008997001
Eukaryotic initiation factor 3e	CACHCG19000024001
Eukaryotic initiation factor 3f	CHC_T00008632001
Eukaryotic initiation factor 3g	CHC_T00009064001
Eukaryotic initiation factor 3h	CHC_T00008730001
Eukaryotic initiation factor 3i	CHC_T00008431001
Eukaryotic initiation factor 3j	GSCHC2T00003639001
Eukaryotic initiation factor 3k	CHC_T00008737001
Eukaryotic initiation factor 3l	CHC_T00008880001
Eukaryotic initiation factor 3m	CHC_T00009012001
Eukaryotic initiation factor 4A	CHC_T00008748001
Eukaryotic initiation factor 4B	Not found
Eukaryotic initiation factor 4E	CHC_T00008457001
Eukaryotic initiation factor 4G	CHC_T00008675001
Eukaryotic initiation factor 5	CHC_T00009078001
Eukaryotic initiation factor 5A	CHC_T00009569001
Eukaryotic initiation factor 5B	CHC_T00008845001
Eukaryotic initiation factor 6	CHC_T00009074001
Eukaryotic initiation factor iso 4E	CHC_T00008457001
Eukaryotic initiation factor iso 4G	Not found
eukaryotic peptide chain release factor eRF1	CHC_T00008673001
eukaryotic peptide chain release factor eRF3	CHC_T00009564001
nCBP	Not found
PABP	CHC_T00009470001
ribosome recycling factor	CHC_T00008444001

4.10 Ribosomal proteins

Ribosomal proteins together with rRNA make up the subunits of the ribosomes. The eukaryotic ribosome contains about 80 ribosomal proteins (102). *C. crispus* ribosomal proteins were identified using ribosomal proteins from *A. thaliana* (www.arabidopsis.org) and classified in three groups: small, acidic and large. Homologs for most proteins were found, but in most cases only one of each type was found. There are around 82 ribosomal proteins in *C. crispus* compared to 349 found in *A. thaliana* genome (Tables S4.8 to S4.10).

Table S4.8. Summary of ribosomal proteins in *C. crispus.*

Type	*Chondrus*	*Arabidopsis*
Small	31	101
Acidic	4	13
Large	47	135
Total	82	349

Table S4.9. Small ribosomal proteins in *C. crispus.*

40S ribosomal protein	Gene ID	40S ribosomal protein	Gene ID
SA	CHC_G00007710001	S15a	CHC_T00007037001
S2	CHC_T00001700001	S16	CHC_T00005582001
S3	CHC_T00005966001	S17	CHC_T00004275001
S3A	CHC_T00007130001	S18	CHC_T00002940001
S4	CHC_T00001977001	S19	CHC_T00005237001
S5	CHC_T00008165001	S20	CHC_T00002262001
S6	CHC_T00001323001	S21	CHC_T00008141001
S7	CHC_T00008131001	S23	CHC_T00005645001
S8	CHC_T00005146001	S24	CHC_T00002329001
S9 family	GSCHCT00004170001 GSCHC2T00004947001	S25	CHC_T00007712001
S10	CHC_T00002497001	S26	CHC_T00008627001
S11	CHC_T00008112001	S27	CHC_T00007530001
S12	CHC_T00003271001	S28	CHC_T00006345001
S13	CHC_T00002855001	S29	CHC_T00008776001
S14	CHC_T00008111001	S30	CHC_T00004409001
S15	CHC_T00006223001		

Table S4.10. Large ribosomal proteins in *C. crispus.* P, partial

60S ribosomal protein	Gene	60S ribosomal protein	Gene
Acidic			
P0	CHC_T00006921001	L17	CHC_T00002308001
	CHC_T00005100001	L18	CHC_T00004290001
P1	CHC_T00004192001	L18A	CHC_T00003914001
P2	CHC_T00003758001	L19	CHC_T00006661001
P3	not found	L21	CHC_T00007031001
		L22	CHC_T00006993001
Large subunit		L23	CHC_T00001050001
L2	CHC_T00000319001	L23a	CHC_T00000211001
L3	CHC_T00007885001	L24	CHC_T00006137001
L4	CHC_T00004115001	L26	CHC_T00006453001
L5	CHC_T00005732001	L27	CHC_T00002279001
L6	CHC_T00005545001	L27A	CHC_T00005567001
		L28	not found
L7	CHC_T00004046001	L29	CHC_T00002555001
L7a	CHC_T00005213001	L30	CHC_T00004447001
L8	CHC_T00002156001	L31	CC81:229271..229630
L9	CHC_T00001204001	L32	CHC_T00006843001
L10	CHC_T00000922001^{P}	L34	CHC_T00002925001
L10a	CHC_T00008715001	L35a	CHC_T00008110001
L11	CHC_T00006638001	L36	not found
L12	CHC_T00003759001	L44	CHC_T00002645001
L13	CHC_T00002756001	L37	CHC_T00001954001
L13A	CHC_T00003724001	L37a	CHC_T00005101001
L14	CHC_T00003038001	L38	CHC_T00003573001
L15	CHC_T00000015001	L39	CHC_T00000218001

5 Cell structure and regulation

5.1 Cytoskeleton

The *C. crispus* genome displays a good conservation for proteins involved in the structure, organisation and dynamics of the cytoskeleton (Table S5.1). Homologues of tubulin α, β, and γ monomers were found. Five genes coding for actin are also present, and the polymerization of actin

filaments by nucleation promoting factors seems conceivable, as homologues of formin and profilin were found in the genome. In metazoans and plants, actin reticulation is crucial for processes such as cell movement, vesicular trafficking and pathogen infection. It is carried out by the actin-related protein 2/3 multiprotein complex (ARP2/3) (103), the activation of which is mediated by the WASP and WASP-family verprolin-homologous protein (WAVE) (104). Surprisingly, this complex seems to be only partially present in *C. crispus*. The two core proteins of the complex, ARP2 and ARP3 proteins, as well as the sub-units ARPC1B (p41-Arc) and ARPC2 (p34-Arc), are missing. On the other hand, homologues of the sub-units ARPC3 (p21-Arc), ARPC4 (p20-Arc) and ARPC5 (p16-Arc) are present. The lack of most of the complex subunits is puzzling, especially as protein-protein interactions between some of the missing proteins were shown to be required for the activity of the complex in other organisms (103). The *C. crispus* genome also lacks WASP proteins, which have been found only in opisthokonts so far. Besides, WAVE/SCAR proteins are absent, while they have been described in the green lineage (105). In contrast, the genome codes for at least one homologue of severin, which is involved in severing of actin filaments.

Several molecular motors, such as myosins and kinesins, which are proteins using the scaffold of actin filaments and microtubules to promote motion of organelles or vesicles through the cell (106) (107) (108), are present.

Table S5.1 Cytoskeleton-related proteins in *C. crispus.*

Cytoskeleton related proteins	Number of genes	Gene ID
Actin	6	GSHC2T00011459001
		GSCHCT00005597001
		CHC_T00008840001
		CHC_T00008795001
		GSHC2T00011459001
		CHC_T00008795001
ARPC5	1	CHC_T00009298001
ARPC3	1	CHC_T00001790001
ARPC4	1	CHC_T00008614001
Actin-related protein	3	CHC_T00008614001
		CHC_T00008511001
		CHC_T00009298001
Formin	1	CHC_T00009469001
Kinesin	9	CHC_T00008304001
		CHC_T00009138001
		CHC_T00006378001
		CHC_T00009381001
		CHC_T00009381001
		CHC_T00008641001
		CHC_T00009001001
		CHC_T00008720001
		CHC_T00009265001
Myosin	1	CHC_T00008462001
Profilin	1	CHC_T00006281001
Severin	1	CHC_T00008685001
Tubulin α	2	CHC_T00008699001
		CHC_T00008870001
Tubulin β	2	CHC_T00008909001
		CHC_T00004301001
Tubulin γ	1	CHC_T00008366001
Dynein light chain	1	CHC_T00008555001
Dynein intermediate chain	1	CHC_T00009309001

5.2 Vesicle trafficking

The maintenance and functioning of the complex endomembrane system of the eukaryotic cell relies on vesicle trafficking mediating secretory and endocytic transport of membranes and proteins. To assess the "completeness" of this conserved core in *C. crispus*, we used a blast-based search strategy to explore four crucial parts of membrane-trafficking protein machinery: (1) membrane coats responsible for the formation of vesicles at donor compartments or the plasma membrane (Table S5.2); (2) complexes responsible for tethering of the vesicles to the target membranes (Table S5.3); (3) SNARE and SM proteins responsible for membrane fusion in a specific and regulated manner (Table S5.4); and (4) GTPases and their regulators orchestrating most of the processes mediated by the aforementioned elements (see the paragraph 5.3).

Orthologs of all conserved subunits of five coat complexes could be found in the *C. crispus* genome (Table S5.2): the seven subunits of the COPI (coatomer) complex, the four subunits of the COPII complex (with the Sec24 subunits being encoded by three different loci), the four subunits of the retromer complex, and the four subunits of each of the adaptor protein complexes AP-1 and AP-3 plus the associated clathrin (both the heavy and the light chain). On the other hand, neither subunit of the adaptor complexes AP-2, AP-4, and AP-5 could be identified, suggesting that these coat complexes are altogether missing. Whereas AP-4 implicated in trans-Golgi network (TGN) to cell surface transport and AP-5 involved in endosomal sorting have been secondarily lost several times during the evolution of the eukaryotes, the absence of AP-2 involved in clathrin-dependent endocytosis, is rather rare among the eukaryotic lineages investigated so far (109) (110).

Concerning the tethering complexes, five of them have orthologs of all or most subunits identifiable in the *C. crispus* genome, namely the Dsl1 complex, TRAPP, COG, GARP, and HOPS (Table S5.3). The CORVET complex shares four subunits with the HOPS complex (functioning at late endosomes and the vacuolar/lysosomal compartment), but has two specific subunits Vps3 and Vps8, the latter of which mediates recruitment of the CORVET complex to endosomes by a RAB GTPase of the RAB5 subfamily (111). No Vps3 and Vps8 orthologs could be detected, suggesting the absence of the CORVET complex and correlating with the absence of a RAB5 ortholog and hence a generally simplified endocytic mechanisms. However, simplification of the secretory pathway is also apparent due to the absence of orthologs of most subunits of the exocyst complex, which tethers exocytic vesicles to the plasma membrane in polarised exocytosis*(112)*. An ortholog of only the Sec6 subunits could be identified in *C. crispus*, echoing the situation reported for *Cyanidioschyzon merolae* (113). It is possible that without the remaining seven subunits, the "orphan" Sec6 protein functions in a novel way in red algal cells.

SNAREs and the associated SM (Sec1/Munc18) proteins are involved in the actual membrane fusion within the eukaryotic endomembrane system (114). There are four major types of the SM proteins (113), and orthologs of all four are encoded by the genome, with Sly1 present as two duplicated genes (Table S5.4). On the other hand, whereas most of the generally conserved types of

SNARE proteins (115) are retained in *C. crispus*, some are apparently missing, suggesting simplification in SNARE-mediated membrane fusion events (e.g. in the endosomal/vacuolar pathway due to the absence of any candidate Qc-SNARE, which would be represented by Syntaxin 8/Vam7 or SYP7 orthologs). Interestingly, a similar pattern of the presence/absence of the major SNARE types have been reported for *C. merolae* (115), suggesting an extensive simplification in the red algal stem lineage. However, it is possible that some of the SNAREs in *C. crispus* have diverged beyond recognition by simple homology detection methods, calling for more sophisticated searches and also for experimental characterization of the different SNARE complexes in rhodophytes.

In summary, while most of the elements of the general eukaryotic core of the membrane-trafficking machinery are conserved in *C. crispus*, notable losses of some of the components, particularly in the endocytic pathaway (AP-2, CORVET, endocytic Qc-SNARE; see also the absence of RAB5 and Vps9 described in the paragraph 5.3) points to less complex membrane trafficking processes in red algae than in most other eykaryotic groups.

Table S5.2. Membrane coat complexes in *C. crispus.*

Coat complex	Subunit	Gene ID
COP I (Coatomer)	Alpha	CHC_T00008629001
	Beta	CHC_T00003146001
	Beta’	GSCHC2T00004460001
	Gamma	GSCHC2T00006120001
	Delta	CHC_T00009496001
	Epsilon	CHC_T00003371001
	Zeta	CHC_T00009109001
COP II	Sec13	GSCHC2T00010051001
	Sec23	GSCHC2T00012335001
	Sec24	GSCHC2T00012808001
		GSCHC2T00000372001
		GSCHC2T00005550001
	Sec31	GSCHC2T00009517001
Clathrin	Clathrin heavy chain	CHC_T00008543001
		CHC_T00009384001
	Clathrin light chain	GSCHCT00013606001

AP-1	Gamma	CHC_T00009316001
	Beta1	GSCHC2T00013924001
	Mu1	GSCHC2T00017657001
	Sigma1	GSCHC2T00011114001
AP-2	All subunits missing	-
AP-3	Delta	GSCHC2T00017491001
	Beta3	GSCHC2T00015355001
	Mu3	GSCHC2T00008913001
	Sigma3	GSCHC2T00011214001
AP-4	All subunits missing	-
AP-5	All subunits missing	-
Retromer	SNX (sorting nexin)	GSCHC2T00010143001
	Vps26	GSCHC2T00009419001
	Vps29	GSCHC2T00009242001
	Vps35	GSCHC2T00001932001

Table S5.3. Vesicle tethering complexes in *C. crispus.*

Tethering complex	Subunit	Gene ID
Dsl1 complex	Dsl1/ZW10	GSCHC2T00004211001
	Tip20/RINT1	GSCHC2T00015283001
	Sec39	Not found
TRAPP complexes	Bet5/ TRAPPC1	CACHCT100000014001
	Trs20/ TRAPPC2	GSCHC2T00008722001
	Bet3/ TRAPPC3	GSCHC2T00004000001
	Trs23/ TRAPPC4	GSCHC2T00016992001
	Trs31/ TRAPPC5	CACHCT96000007001
	Trs33/ TRAPPC6	CACHCT76000009001
	Trs65	GSCHC2T00004952001

	Trs120/ TRAPPC9	GSCHC2T00008721001
	Trs130/ TRAPPC10	GSCHC2T00009013001
	Tca17/ TRAPPC2L	SNAPCcT00037762001
	Trs85/ TRAPPC8	Not found
	TRAPPC11	Not found
	TRAPPC12	Not found
Exocyst	Sec3	Not found
	Sec5	Not found
	Sec6	GSCHC2T00009920001
	Sec8	Not found
	Sec10	Not found
	Sec15	Not found
	Exo70	Not found
	Exo84	Not found
COG complex	COG1	GSCHC2T00011023001
	COG2	GSCHC2T00008537001
	COG3	GSCHC2T00005252001
	COG4	GSCHC2T00003291001
	COG5	GSCHC2T00014493001
	COG6	GSCHC2T00007709001
	COG7	GSCHC2T00013218001
	COG8	GSCHC2T00015815001
GARP complex	Vps51	Not found
	Vps52	GSCHC2T00007695001
		GSCHC2T00012400001
	Vps53	GSCHC2T00016643001
	Vps54	GSCHC2T00005665001
CORVET and HOPS	Vps3	Not found

complexes	Vps8	Not found
	Vps11	GSCHC2T00006874001
	Vps16	SNAPCcT00006356001
	Vps18	GSCHC2T00011924001
	Vps33	Incomplete sequence
	Vps39	GSCHC2T00004756001
	Vps41	GSCHC2T00009339001

Table S5.4. SNARE and SM proteins in *C. crispus.*

Major group	gene or subfamily name	Gene ID
Qa-SNARE	Syn5/Sed5	GSCHC2T00003005001
	Syn16/Tlg2	GSCHC2T00017223001
	Syn18/Ufe1	CACHCT2300000050714202
	Syn7/12/Pep12/Vam3	Incomplete sequence
	Plasma membrane syntaxins/Sso1/2	GSCHC2T00011521001
		GSCHC2T00000823001
		CACHCT574000002001
Qb-SNARE	Gos1/GS28	GSCHC2T00005460001
	Vti1	CACHCT42000008001
	Bos1/Membrin	Not found
	Sec20	GSCHC2T00004654001
	NPSN1	Not found
Qbc-SNARE	Sec9-like	SNAPCcT00003116001
Qc-SNARE	Use1	GSCHC2T00013468001
		SNAPCcT00001330001
	Bet1/Stf1/GS15	CACHCT78000008001
		GIDCcT00001643001
	Syn6/10/Tlg1	CACHCT80000011001

	Syn8/Vam7	Not found
	SYP7	Not found
R-SNARE	Sec22	GSCHC2T00009458001
	Ykt6	GSCHC2T00011456001
	VAMP7/Nyv1	CACHCT100001039102
		CACHCT400002104102
	R-brevins/Snc1/2	Not found
	Tomosyn/Sro7	Not found
SM proteins	Sec1	GSCHC2T00006255001
	Sly1	GSCHC2T00010622001
		GSCHC2T00012003001
	Vps45	GSCHC2T00008191001
	Vps33	Incomplete sequence

5.3 Kinases

The eukaryotic protein kinases (ePKs) in the genome were identified by two independent means: 1) by PFAM, IPR and keyword searches of the gene models, 2) by comparing the raw genome sequence against an ePK-specific multi-level profile Hidden Markov Model library. This second method also provided an initial classification of *C. crispus* ePKs into families. All together, these approaches identified 209 probable ePKs (Table S5.5), a relatively high number for a gene superfamily, but consistent with the 1.5-2.5 % predicted in other eukaryotes. The ePK superfamily is divided into a number of families (116) (117). Not all of these families are to be expected in every eukaryote – the RGC group, for example, is found only in animals – but all the expected ePK groups are present in the *C. crispus* genome.

This ePK distribution is different compared to land plants, largely because of the enormous expansion of the TKL family in land plants; see Table S5.5 for distribution into families within kinomes of *Phytophthora* (118), *Thalassiosira* (118), yeast, slime mold (119) and rice (120). Atypical kinases, including for instance the APHs, ABC1, and the bromodomain representatives, are present in large numbers. A relatively high number of TKs are present in *C. crispus*, corroborating the link between multicellular organisms and this group of extracellular signal transduction proteins (121). Overall, *C. crispus* has a fairly standard eukaryotic kinase complement without the massive expansions seen in some organisms such as plants and choanoflagellates.

Table S5.5. Protein kinases distribution. *C. crispus* (Cc), *Oryza sativa* (*Os*), *Phytophthora infestans* (*Pi*), *Thalassiosira pseudonana* (*Tp*), *Saccharomyces cerevisiae* (*Sc*) and *Dictyostelium discoideum* (*Dd*).

	Number of protein kinase genes						
Group	**Os**	**Cc**	**Pi**	**Tp**	**Sc**	**Dd**	**Total**
AGC	59	12	47	29	17	21	185
CAMK	163	39	70	52	22	21	367
RGC	0	0	1	0	0	0	1
CK1	33	4	3	4	4	3	51
CMGC	151	19	46	25	23	30	294
STE	73	10	15	5	14	45	162
TK	0	12	2	0	0	0	14
TKL	1387	21	137	14	0	67	1626
Other	31	20	34	23	38	68	214
Atypical	Not given	72	Not given	Not given	13	39	124
Total - atypical	1897		355	152			
Total		209			131	294	

5.4 The GTPase regulation

GTPases (or GTP-binding proteins) are a broad class of proteins mediating many crucial signalling and regulatory steps in the cell (122). To probe the range of GTPase-based processes in *C. crispus*, blast searches were performed to identify and annotate GTPases of the TRAFAC class, the major GTPase subgroup including a majority of the "classical" GTPase families (122). Searches using HMMER program (http://hmmer.janelia.org) were further employed to probe for the presence of some of the know regulators of GTPase signalling.

Most major families of TRAFAC GTPases are represented in the genome (Table S5.6). A fairly standard set of translation factor superfamily is present, including versions targeted to the plastid and the mitochondrion to assist the translation mediated by the organellar ribosomes; a notable missing gene is *SelB*, a translation factor specific for the incorporation of selenocysteine during translation, which is consistent with the predicted absence of selenoproteins in *C. crispus* (see section 2.6). The OBG-HflX-like and TrmE-Era-EngA-YihA-Septin-like superfamilies, and the YlqF/YawG (also called YRG) family all comprise GTPases primarily involved in various steps of the biogenesis of the small or the large subunit of the ribosome in both the nucleus/cytoplasm and the organelles (123) (124) (125). Most of the expected GTPases of this category are present in *C. crispus*, with a notable exception of a plastid-targeted form of the Era GTPase; while this GTPase is present in most other plastid-bearing eukaryotes. It is also missing in *C. merolae* genome, indicating a possible loss in the red algal lineage. The septin family is represented in *C. crispus* by only two putative receptor

subunits of the TOC complex in the outer plastid membrane (Toc34/159-related proteins (126)), whereas the true septins are absent (as in many other eukaryotic groups).

Only two dynamin family genes, involved in division of either the mitochondrion (CmDnm1) or the plastid (CmDnm2), have been identified in the *C. merolae* genome (127). Strikingly, an ortholog of only the former can be found in the *C. crispus* genome. Since CmDnm2 is conserved in other plastid-bearing eukaryotes and its embryophyte orthologs (called ARC5/DRP5B) are essential for proper plastid division*(128)*, the apparent lack of a corresponding dynamin-related protein in *C. crispus* suggests that the mechanism of plastid division may differ in this organism. On the other hand, *C. crispus* has four extra dynamin family members not seen in *C. merolae*. Two of them are most similar to the metazoan Mx proteins involved in anti-viral defence (129). Related proteins have been identified in many other eukaryotes (121) and may represent hitherto unrecognized general element of the eukaryotic innate immunity system. Another *C. crispus* dynamin-related protein group represents an apparently novel subgroup with only one additional known member in the brown alga *Ectocarpus siliculosus*; no functional prediction can be made for this protein at this stage. The last dynamin-related protein family in *C. crispus* is related to the EHD/RME-1 subfamily implicated in endocytic transport and characterized by a GTPase-related ATPase domain associated with the EH domain (130). The domain architecture of the *C. crispus* protein is unusual in that it contains an extra FYVE domain at the N-terminus, indicating some lineage-specific embellishments in the processes mediated by EHD/RME-1.

The GB1/RHD3 family comprises several conserved subgroups, of which *C. crispus* notably lacks the so-called gyanylate-binding proteins that have been shown in some species to be implicated in defence against intracellular pathogens (131). On the other hand, the *C. crispus* genome encodes an unusually extended family (seven paralogs) of RHD3/Sey1-like proteins that are known to be involved in regulating the ER morphology and ER-to-Golgi transport (132); the significance of this expansion remains unclear.

A real hallmark for eukaryotes is an expanded set of genes of the Ras GTPase superfamily implicated in diverse cellular processes such as membrane trafficking, flagellum biogenesis, nucleo-cytoplasmic transport, actin dynamics, or various signalling pathways. The generally conserved core of the superfamily may be expanded to tens (or even hundreds) of paralogs in a single genome. However, *C. crispus* and potentially red algae in general turned out to have one of the smallest Ras superfamily gene sets among all eukaryotes investigated so far (Table 5.6). The superfamily comprises several major families (ARF/ARL/SAR1, RAB, RHO, RAS, RAG/GTR, ROCO, Gα subunit of heterotrimeric G-proteins, dynein light intermediate chain) and an array of more or less unclassified members (122) (133).

The ARF/ARL/SAR1 family includes GTPases with diverse role in membrane and protein trafficking, and some of them (ARL3, ARL6, ARL13) are specifically associated with transport steps towards and within the flagellum (134). Unsurprisingly, both *C. crispus* and *C. merolae* lack all the

flagellum-associated ARL paralogs. Otherwise, the composition of the ARF/ARL/SAR1 family in *C. crispus* appears quite standard, except an interesting case of two paralogous genes representing a divergent ARF-like group shared with at least another red alga *Pyropia*, indicating a novel red algal-specific feature.

The RAB family, comprising master regulators of compartment specificity in vesicle trafficking (135), is very small in *C. crispus*, with only seven genes representing six ubiquitous paralogs RAB1, RAB2, RAB6, RAB7, RAB11 (duplicated in *C. crispus*), and RAB18. The RAB family in Cyanidiales is even smaller, as *C. merolae* has only one RAB11 paralog, and *Galdieria* lacks RAB18. Since a complex RAB family (around 20 paralogs) was inferred to exist in early eukaryotes and preserved in most extant lineages including the related Chloroplastida lineage (136) (137), it seems that many RAB genes have been lost in the ancestor of red algae. It is especially striking not to see in red algae some highly conserved and widespread paralogs, such as RAB8 involved in exocytosis, or RAB5 involved in endocytic transport. In agreement with this, RAB5-specific positive regulators (GEFs) characterised by the VPS9 domain (138) are also missing in the *C. crispus* genome, indicating a secondarily simplified endocytic machinery. Less surprising is the lack in the *C. crispus* genome of orthologs of all RABs and RAB-like proteins known or predicted to be involved in flagellum-associated processes: RAB23, IFT27/RABL4, RJL and IFT22/RABL5*(139) (140)*. These paralogs are generally conserved in species with the flagellar apparatus, and are missing in those unable to form flagella, including red algae.

GTPases of the RHO family are most typically associated with regulation of the actin dynamics and cell polarity (141) (142). Similarly to the situation in Cyanidiales but in contrast to Metazoa, Fungi, Amoebozoa or flowering plants, the RHO GTPase-based signalling in *C. crispus* seems to be very simple – just one RHO GTPase regulated by a single RhoGAP protein (a negative regulator; CHC_T00000147001) and by a single member of the Archaeplastida-specific PRONE class of GEFs (positive regulator; CHC_T00003971001). The canonical eukaryotic RhoGEF (DH) family or the Dock180/Zizimin family are missing. In addition, there is no RhoGDI (another general RHO GTPase regulator) in the *C. crispus* genome (as well as in *C. merolae*), in contrast to the vast majority of eukaryotes (143).

The RAS family comprises the most typical "signalling" GTPases that serve in pathways regulating cell growth and division in response to external clues*(144)*. Similarly to Chloroplastida and some other eukaryotic lineages, true Ras or Rap proteins are missing in *C. crispus* (and in Cyanidiales), together with dedicated positive and negative regulators containing the RasGEF, RasGAP, and RapGAP domains. Thus, the only member of the RAS family is Rheb, which is conserved in all red algal genomes investigated (even duplicated in *Galdieria*). Rheb has been established as a regulator of the TOR signalling pathway related to cell growth and division (145). Unexpectedly, *C. crispus* seems to lack another conserved type of Ras superfamily GTPases involved in TOR regulation, the GTR/RAG family always occurring as a pair of ancestral paralogs Gtr1

(RagA/RagB in vertebrates) and Gtr2 (RagC/RagD in vertebrates (145)). Both Gtr1 and Gtr2 are present in Cyanidiales, suggesting a simplification of the TOR pathway in *C. crispus*.

Chondrus crispus also harbours a series of conserved Ras superfamily GTPases not easily classified into major families, including RAN (the master regulator of nucleo-cytoplasmic transport (146)), MIRO (implicated in mitochondrial dynamics (147)), SPG1/TEM1 (involved in regulation of the cell cycle (148)), LIP1/RABL3 (possibly involved in circadian rhythm regulation (149)), and the β subunit of the signal recognition particle receptor (SRβ). However, some other widespread components of the Ras superfamily are missing in the genome (and *C. merolae*), including: 1) heterotrimeric G proteins, at least the alpha and beta subunits (the gamma subunits are small poorly conserved proteins that are difficult to identify by *in silico* approaches); this is really remarkable, as heterotrimeric G proteins are almost omnipresent element of the eukaryotic signalling toolkit; 2) two paralogous light intermediate chains of the cytoplasmic dynein 1 (D1LIC) and cytoplasmic dynein 2 (D2LIC); their absence is in accordance with the absence of the actual dynein motors; and 3) two unique GTPases with poorly defined or unknown cellular function but yet broad phylogenetic distribution, namely RBEL1/PARF (150) and RTW/RABL2.

The ROCO family is characterized by two conserved domains: a GTPase domain (Roc) belonging to the Ras superfamily, coupled with a unique dimerisation domain (COR). The cellular roles of different ROCO proteins may be diverse and depend on the other domains attached at the N- or C-terminus of the conserved Roc-COR core (151) (152). Recently, an expanded ROCO family has been characterized in *E. siliculosus* and suggested to be involved in recognition/transduction events linked to immunity (153). *Chondrus crispus* has eight ROCO loci (one of them probably being a pseudogene) coding for proteins featuring the Roc-COR core with a TIR domain at the C-terminus and either another TIR domain, a NB-ARC domain, or Sel1 repeats at the N-terminus. Since at least the TIR and the NB-ARC domains are typically found in proteins involved in anti-pathogen defence mechanisms*(154)*, it is possible that the ROCO proteins are also somehow involved in these processes.

In conclusion, the analysis of GTPases in the *C. crispus* genome indicates few group-specific innovations but considerable secondary simplifications in GTPase-utilising cellular processes, specifically in diverse trafficking and signalling pathways.

Table S5.6. A comparative analysis of P-loop GTPases of the TRAFAC class in *C. crispus* and other selected eukaryotes. Names applied to the GTPase genes generally follow names of well-characterised orthologs, mostly from Metazoa, *S. cerevisiae*, or *E. coli*. "P-" or "M-" before a gene name is applied for prokaryotic-like GTPase localised to the plastid or mitochondrion respectively, where no specific name has been adopted for the eukaryotic form. Presence or absence of putative orthologs in other species is indicated by "+" or "-". Species abbreviations: *Aureococcus anophagefferens* (Aa), *Thalassiosira pseudonana* (Tp), *Phaeodactylum tricornutum* (Pt) *Phytophthora sojae* (Ps), *Cyanidioschyzon merolae* (Cm), *Arabidopsis thaliana* (At), *Homo sapiens* (Hs); Sc – *Saccharomyces*

cerevisiae (Sc), *Dictyostelium discoideum* (Dd). p, plastidial genome; f, fragment; i, incomplete, EST, EST support without genome sequence .

Gene name	Gene ID	Cm	Es	Aa	Tp	Pt	Ps	At	Hs	Sc	Dd
Translation factor superfamily											
EFL	CHC_T00009511001	-	-	-	-	-	-	-	-	-	-
EF-1α	-	+	+	+	+	+	+	+	+	+	+
Tuf1	CHC_T00009351001	+	+	+	+	+	+	+	+	+	+
TufA	p	+[p]	+[p]	+[p]	+[p]	+[p]	-	+	-	-	-
EF2	CHC_T00009339001	+	+	+	+	+	+	+	+	+	+
Mef1	CHC_T00009147001	+	+	+	+	+	+	+	+	+	+
Mef2	CHC_T00008408001i	+	+	+	+	+	+	-	+	+	+
P-EFG	CHC_T00009455001	+	+	+	+	+	-	+	-	-	-
Ria1	CHC_T00005107001	+	+	+	+	+	+	+	+	+	+
Snu114	CHC_T00009253001	+	+	+	+	+	+	+	+	+	+
Sup35	CHC_T00009564001	+	+	+	+	+	+	+	+	+	+
HBS1	CHC_T00009014001	+	+	-	+	+	+	+	+	+	+
SelB	-	+	+	+	+	+	+	-	+	-	+
EIF5B	CHC_T00008845001	+	+	+	+	+	+	+	+	+	+
M-TIF2	CHC_T00009026001	+	+	+	+	+	+	+	+	+	+
InfB	p	+[p]	+	+	+	+	-	+	-	-	-
EIF2S3	CHC_T00008676001	+	+	+	+	+	+	+	+	+	+
GTPBP1	CHC_T00008939001	+	+	+	+	+	+	-	+	-	+
Guf1	CHC_T00008692001	-	+	+	+	+	+	+	+	+	+
P-EF4	CHC_T00008445001	+	+	+	+	+	-	+	-	-	-
PRF3	CHC_T00008434001	-	+	+	+	+	-	-	-	-	-
M-BipA	CHC_T00004182001	-	+	-	+	+	+	+	-	-	+
P-BipA	CHC_T00009437001	+	+	+	+	+	-	+	-	-	-
Bms1	CHC_T00002933001	+	+	+	+	+	+	+	+	+	+
OBG-HflX-like superfamily											
M-HflX	-	-	+	-	-	-	-	+	+	-	-
P-HflX	CHC_T00009489001	+	+	+	+	+	-	+	-	-	-
Nog1	CHC_T00001903001	+	+	+	+	+	+	+	+	+	+
PDE318	-	-	+	+	+	+	+	+	-	-	-
DRG1	CHC_T00008770001	+	+	+	+	+	+	+	+	+	+
DRG2	CHC_T00008792001	+	+	+	+	+	+	+	+	+	+
Mtg2	SNAPCcT00040968001[i]	+	+	+	+	+	+	+	+	+	+

P-Obg	CHC_T00004853001	+	+	+	+	+	-	+	-	-	-
OLA1	CHC_T00009137001	+	+	+	+	+	+	+	+	+	+
P-YchF	CHC_T00000245001	+	+	+	+	+	-	+	-	-	-
Ygr210c	-	-	-	+	+	+	+	-	-	+	+
TrmE-Era-EngA-YihA-Septin-like superfamily											
M-EngA	CHC_T00005918001	+	+	+	+	+	+	+	-	-	+
P-EngA	CHC_T00002792001	+	+	+	+	+	-	+	-	-	-
TrmE	CHC_T00006415001	+	+	+	+	+	+	+	+	+	+
YihA	CHC_T00003059001	+	+	+	+	+	+	+	+	+	+
M-Era-1	CHC_T00003305001	+	+	-	+	+	+	+	+	-	+
M-Era-2	CHC_T00007501001										
P-Era	-	-	+	+	+	+	-	+	-	-	-
Ngp1	CHC_T00003975001	+	+	+	+	+	+	+	+	+	+
Nug1	CHC_T00002347001	+	+	+	+	+	+	+	+	+	+
Lsg1	CHC_T00002034001	+	+	+	+	+	+	+	+	+	+
GNL1	CHC_T00005403001	-	-	-	-	-	+	-	+	-	+
Mtg1	-	-	+	-	-	-	+	+	+	+	+
P-YlqF	CHC_T00000923001	+	+	+	+	+	-	+	-	-	-
NOA1	CHC_T00002510001	+	+	+	+	+	+	+	+	+	+
Toc34a	CHC_T00004674001	+	-	-	-	-	-	+	-	-	-
Toc34b	SNAPCcT00008033001[i]										
Septin	-	-	+	-	-	-	-	-	+	+	-
IMAP/AIG1	-	-	-	+	-	-	-	+	+	-	-
Dynamin/Fzo/YdjA family											
Dnm1	CHC_T00003610001	+	+	+	+	+	+	+	+	+	+
ARC5	-	+	+	+	+	+	-	+	-	-	-
Mx1	CHC_T00001332001	-	+	+	-	+	+		+	-	-
Mx2	CHC_T00004690001										
DRP	CHC_T00005209001	-	+	-	-	-	-	-	-	-	-
RME1	CHC_T00005843001	-	+	-	+	+	+	-	+	-	+
GB1/RHD3 family											
GBP1	-	-	+	+	-	-	+	+	+	-	+
Atlastin	-	-	+	+	+	+	+	-	+	-	-
RHD3a	CHC_T00004993001	+	+	-	-	-	-	+	-	+	+
RHD3b	CHC_T00007280001										
RHD3c	CHC_T00007883001										

RHD3d	CHC_T00007777001										
RHD3e	SNAPCcT00025553001										
RHD3f	CHC_T00002113001										
RHD3g	CHC_T00006439001										
Ras-like superfamily											
Rab1	CHC_T00009150001	+	+	+	+	+	+	+	+	+	+
Rab2	CHC_T00000605001	+	+	+	+	+	+	+	+	-	+
Rab4	-	-	-	-	-	-	-	-	+	-	+
Rab5	-	-	+	+	+	+	+	+	+	+	+
Rab6	CHC_T00008465001	+	+	+	+	+	+	+	+	+	+
Rab7	CHC_T00009482001	+	+	+	+	+	+	+	+	+	+
Rab8	-	-	+	+	+	+	+	+	+	+	+
Rab11a	CHC_T00009363001i	+	+	+	+	+	+	+	+	+	+
Rab11b	EST										
Rab14	-	-	-	-	-	-	-	-	+	-	-
Rab18	CHC_T00009081001	+	+	+	+	+	+	+	+	-	+
Rab20	-	-	-	-	-	-	-	-	+	-	-
Rab21	-	-		+	+	-	+	-	+	-	+
Rab22	-	-	+	+	+	+	+	-	+	-	-
Rab23	-	-	-	-	-	-	+	-	+	-	-
Rab24	-	-	-	-	-	-	-	-	+	-	-
Rab28	-	-	+	+	-	-	+	-	+	-	-
Rab32A	-	-	+	+	-	-	+	-	+	-	+
Rab32B	-	-	-	-	-	-	-	-	-	-	+
Rab34	-	-	-	-	-	-	-	-	+	-	-
Rab50	-	-	+	+	+	+	+	-	-	-	+
RabTitan	-	-	+	-	-	-	+	-	-	-	-
RJL	-	-	+	+	+	-	+	-	+	-	-
RTW	-	-	+	+	+	-	+	-	+	-	-
IFT27	-	-	+	+	+	-	+	-	+	-	-
SPG1	EST	+	+	+	-	-	+	+	-	+	+
RAN	CHC_T00009201001	+	+	+	+	+	+	+	+	+	+
RAC	CHC_T00000146001	+	+	+	-	-	+	+	+	+	+
Ras	-	-	-	-	-	-	+	-	+	+	+
Rap	-	-	-	-	-	-	-	-	+	+	+
Rheb	CACHCT59000012001	+	-	+	+	+	+	-	+	+	+

MIRO	CHC_T00006657001	+	+	-	+	+	+	+	+	+	+
LIP1	CHC_T00005607001	-	+	-	-	-	+	+	+	-	+
IFT22	-	-	+	+	-	-	+	-	+	-	-
Arf1a	CHC_T00008825001	+	+	+	+	+	+	+	+	+	+
Arf1b	EST										
ARL1	CHC_T00008499001	+	+	+	+	+	+	+	+	+	+
ARL2	CHC_T00008923001	+	+	+	+	+	+	+	+	+	+
ARL3	-	-	+	+	+	-	+	-	+	-	-
ARL5	CHC_T00009553001	-	-	-	-	-	-	+	+	-	+
ARL6	-	-	+	+	-	-	+	-	+	-	-
ARL8	CACHCT42000009001	-	+	+	+	+	+	+	+	-	+
ARL13	-	-	+	-	-	-	+	-	+	-	-
ARL16	-	-	-	+	-	-	+	-	+	-	-
ARFRP1	CACHCT67000010001	+	+	+	+	+	+	+	+	+	+
ArlX1	CACHCT155000015001	-	-	-	-	-	-	-	-	-	
ArlX2	CACHCT259000005001										
Sar1	CHC_T00009285001	+	+	+	+	+	+	+	+	+	+
SRPRB	CHC_T00001378001	+	+	+	+	+	+	+	+	+	+
GPA	-	-	+	+	+	+	+	+	+	+	+
RBEL1	-	-	+	+	-	-	+	-	+	+	+
GTR1	-	+	+	+	+	+	+	-	+	-	+
GTR2	-	+	+	+	+	+	+	-	+	-	+
D1LIC	-	-	+	-	-	-	+	-	+	+	+
D2LIC	-	-	+	+	-	-	+	-	+	-	-
Roco	8	-	+	+	+	+	+	+	+	-	+
	CACHCT580000116102										
	GIDCcT00018537001										
	CHC_T00008269001										
	GIDCcT00005227001										
	CHC_T00008933001										
	CACHCT245000004001										
	CHC_T00009069001										
	CHC_T00008460001f										

6 Photosynthesis and respiration

6.1 Light-harvesting complexes

Red algae possess sophisticated light-harvesting systems, which represent an evolutionary intermediate between those found in cyanobacteria and in green algae or plants. They share with cyanobacteria the presence of large extrinsic antennae (or phycobilisomes) coupled to photosystem II, and with other eukaryotic phototrophs a membrane-intrinsic LhcA-type antenna surrounding photosystem I (155).

Like in most red algae, *C. crispus* phycobilisomes are constituted of a core made of allophycocyanin and rods made of R-phycocyanin and R-phycoerythrin (156). As in other red algal species, genes encoding the α- and β-subunits of all three phycobiliproteins, as well as those coding for the allophycocyanin-associated linker polypeptides, ApcC (Lc) and ApcE (Lcm), and for the sole rod-core linker (CpcG), are all located in the chloroplast genome. In contrast, all putative rod linker polypeptides are nuclear-encoded. Besides possessing three genes encoding chromophorylated γ-type linkers, the *C. crispus* genome also contains six genes encoding colourless rod linkers. One of these possesses a C-terminal CpcD-like domain, suggesting that it may be associated with phycocyanin, whereas others are probably linked to phycoerythrin. All of the *C. crispus* linker genes are expressed, as suggested by the occurrence of many ESTs for each sequence. A large variety of phycobilisome architectures exists in red algae, since they can be hemidiscoidal like in *Rhodella violacea*, hemi-ellipsoidal like in *Porphyridium cruentum*, ellipsoidal like in *Antithamnion glanduliferum*, or even a mixture of types, *e.g.* ellipsoidal and hemidiscoidal in *Porphyra umbilicalis* (156) (157). Although the type of phycobilisome arrangement found in *C. crispus* has not been described yet, the presence of such a large set of rod linkers strongly suggests that this species possesses sophisticated phycobilisomes, possibly combining different configurations. This contrasts with the small unicellular red alga *Cyanidioschyzon merolae*, the nuclear genome of which encodes only two candidate colourless phycobilisome rod linkers and no γ-type linker (127) (158), pointing out that this primitive phycoerythrin-lacking organism possesses a much simpler phycobilisome architecture.

The light-harvesting antenna of photosystem I (LHCI) of *C. crispus* is constituted of seven subunits. Genes coding for these chlorophyll-binding proteins have strong EST support. Yet, when aligned with other LCHI proteins, the quite low similarity observed suggests that these antenna subunits are only distantly related to antennae of other non-red photosynthetic eukaryotes. *C. merolae* possesses only three LHCI proteins, named LHCR. Their weak identity scores, only between 30 and 40% with *C. crispus* LHCR proteins, strongly suggest a divergent evolutionary history.

Chondrus crispus also possesses several other chlorophyll binding (CAB) proteins. Seven of them have one transmembrane helix, and are therefore homologous to cyanobacterial high light inducible polypeptides (HLIPs), which are short, chlorophyll-binding proteins known for their role in photoprotection. Although HLIPs were previously thought to be encoded in the plastidial

genome*(159)*, only one sequence has been found in *C. crispus* chloroplast, whereas the six other HLIPs were found in the nuclear genome. Although green plants possess a wide variety of CAB proteins, including two-helix stress enhanced proteins (SEPs), three-helix early light induced proteins (ELIPs) and four-helix PsbS (160), only one ELIP sequence was found in the nuclear genome of *C. crispus*. No ELIP have been observed so far in cyanobacteria, and thus this photoprotective complex must have been acquired by red algae and green plants after the primary endosymbiosis.

6.2 Pigment biosynthesis

The *C. crispus* genome contains all genes necessary for the biosynthesis of chlorophyll, with the exception of the *chlB*, *chlL* and *chlN* genes, which encode the three subunits of the light-independent protochlorophyllide reductase. In contrast, *C. crispus* possesses one *LPOR* gene encoding a light-dependent protochlorophyllide oxido-reductase. This suggests that, like *Ostreococcus tauri* (161), it can biosynthesize chlorophyll during light periods but not in the dark. In addition, all carotenoid biosynthesis related genes existing in plants and *C. merolae* have been found in the *C. crispus* genome, except the violanxanthin de-epoxidase and xanthophyll epoxidase, both involved in the xanthophyll cycle known to be absent in red algae and cyanobacteria.

6.3 Photoreceptors and circadian clock players

Circadian clocks are essential for fitness and survival of living organisms since they orchestrate biological processes along the day/night cycle an allow them to anticipate predictable environmental changes that are linked to diurnal cycle. We searched for circadian clock homologues based on genes of land plants (*Arabidopsis*), green algae (including *Ostreococcus* and *Chlamydomonas*), animals (mammals and invertebrates), and fungi (*Neurospora*). No homologue of animal and fungi specific clock genes was found.

Circadian clock genes such as circadian clock associated 1 gene, *CCA1*, have been shown to be conserved in angiosperms (162), however, we were not able to identify putative homologues in *C. crispus* using plant clock genes sequences and Blast-based program. Therefore conserved domains of plant clock genes such as MYB, CCT or REC were searched. A putative circadian clock associated 1 (CCA1) like protein was found. This MYB transcription factor (SANT subfamily) exhibits a specific signature (I/LPPPRPKRKPXXPYPQ/RK) near the MYB domain and conserved in *Chlamydomonas*, *Ostreococcus* and *Cyanidioschyzon* CCA1-like proteins (163) (164). Both *Ostreococcus* and *Chlamydomonas* proteins were shown to possess circadian rhythms. Both *Arabidopsis PCL1/LUX* transcription factor and B type response regulators (RRB), identified as master clock components, contain a GARP domain found also in the ROC15 and ROC75 clock proteins in *Chlamydomonas* (165) (166). A GARP domain protein was identified in *C. crispus*. TOC1 and Constans-like are clock proteins which contain a CCT domain ([CO], CONSTANS-like [COL], TIMING OF CAB

EXPRESSION) that is also found in the *Ostreococcus* TOC1 homologue and in the *Chlamydomonas* ROC66 clock components (163) (167) (164). Five *C. crispus* proteins contain such CCT-like domains.

It is noteworthy that several of these clock components have a regulator-like structure (*e.g.* TOC1 and RRB) including a receiver-like domain and a transactivator domain (*e.g.* CCT or GARP domain), in plants and green algae. However, two-component systems modules (receiver, histidine kinases and histidine phosphor-transfer) are absent in *C. crispus* and *C. merolae*. Taken together, these sequence analyses, combined with bibliography data on functional analysis of relevant proteins in green algae and land plants, suggest that *C. crispus* contains proteins with plant-specific domains, such as SANT, GARP and CCT, which could potentially have a circadian clock function.

Photoreceptors mediate light input to the circadian clock and more generally many physiological responses to light quality. An extensive search did not yield any candidate proteins for phytochrome (red light photoreceptor), phototropin and LOV domain containing proteins (blue light photoreceptors), nor rhodopsins (green to orange light photoreceptors). Two or potentially three putative homologues of cryptochromes were identified.

6.4 Carbon uptake

The *C. crispus* genome encodes five carbonic anhydrases (CAs, Table S6.1): three α-CAs (two with a signal peptide; one without signal peptide), one β-CA predicted to be targeted to the chloroplast, and one γ-CA without signal peptide. No δ-CA, a class of CAs typical for diatoms (168), was found. α-CAs are classically found in animals, but also in *E. siliculosus*, β-CAs are best known from prokaryotes and plant chloroplasts, and γ-CAs are usually found in methane producing bacteria, but were also present in several copies in the *E. siliculosus* genome (169). From a physiological point of view, the fact that the otherwise reduced genome of *C. crispus* encodes at least five CAs from three families, and with predicted localizations outside the cell, in the cytoplasm, and in the chloroplast, indicates that these enzymes are important for carbon uptake in *C. crispus*. This is further supported by the presence of two putative sodium bicarbonate cotransporters in the genome (Table S6.1). From an evolutionary point of view, it is interesting to note that *E. siliculosus* and *C. crispus* possess the same CA families.

With respect to genes potentially involved in organic carbon concentrating mechanisms (CCMs), *C. crispus* possesses two NADP-malic enzymes, three malate dehydrogenases (MDHs), one phosphoenolpyruvate carboxylase (PEPc), but no pyruvate carboxylase (PC), phosphoenolpyruvate carboxykinase (PepCK), or pyruvate phosphate dikinase (PPDK) were found. The absence of PepCK or PPDK is surprising, as this enzyme is present in all plant organisms included in the Phytozome database (http://www.phytozome.net), as well as in the red alga *C. merolae*. The absence of PepCK would exclude the possibility of a classical aspartate-based C4-like organic CCM. Malate-based

CCMs usually also rely on PPDK to furnish PEP. In *C. crispus*, PEP may be synthesized from pyruvate via the activity of a pyruvate kinase.

Table S6.1 Carbonic anhydrases and genes potentially involved in organic carbon concentrating mechanisms in *C. crispus*.

Enzyme	EC	Gene model
α-carbonic anhydrase (signal P)	4.2.1.1	SNAPCcT00041376001
α-carbonic anhydrase (no signal)	"	CHC_T00004895001
α-carbonic anhydrase (signal P)	"	SNAPCcT00020012001
putative α-carbonic anhydrase (no signal)	"	CHC_T00002086001
β-carbonic anhydrase (chloroplastic)	"	CHC_T00007098001
γ-carbonic anhydrase (no signal)	"	CHC_T00002095001
δ-carbonic anhydrase	"	Not found
putative sodium bicarbonate cotransporter		CHC_T00008463001
"		CHC_T00008507001
succinic semialdehyde dehydrogenase	1.2.1.24	CHC_T00008575001
phosphoenolpyruvate carboxylase (PepC)	4.1.1.31	CHC_T00008437001
phosphoenolpyruvate carboxykinase (PepCK)	4.1.1.-	Not found
pyruvate carboxylase (PC)	6.4.1.1	Not found
malate dehydrogenase (MDH)	1.1.1.37	CHC_T00008600001
"		CHC_T00008733001
"		CHC_T00009350001
NAD-malic enzyme	1.1.1.39	Not found
NADP-malic enzyme (ME)	1.1.1.40	CHC_T00009563001
"		CHC_T00008878001
pyruvate phosphate dikinase (PPDK)	2.7.9.1	Not found
pyruvate, water dikinase	2.7.9.2	Not found
pyruvate kinase	2.7.1.40	CHC_T00008482001

6.5 Oxidative phosphorylation

The genes identified in *C. crispus* using literature data (170) (171) are presented in Tables S6.2-S6.6. All the bacterial NADH dehydrogenase orthologues for the complex I (rotenone-sensitive NADH:ubiquinone oxidoreductase) have been identified, as well as the large majority of the eukaryote specific subunits. The missing genes are the ones coding for NUOA1, NUOB12, and NNUOB10 subunits. Only two of the plant specific subunits were identified, while none of the mammal or fungus specific subunits were found.

Concerning the complex II (succinate:ubiquinone oxidoreductase), the genes *sdh2, 3* and *4* are encoded in the mitochondrial genome of red algae and a gene homologous to SDH1 have been identified in the nuclear genome. Only six out of the ten classical subunits of complex III (ubiquinol-cytochrome c oxidoreductase) have been identified. Among those not found are QCR8 and QCR10. Because putative genes for these proteins in *Chlamydomonas reinhardtii* also show very low identity with other eukaryote orthologs, it is thus possible that these genes are present in *C. crispus* but were not identified due to low sequence homology. The other two genes not found are for the core1 protein and QCR6. As in most other systems, the complex IV (cytochrome c oxidase) cox1, cox2 and cox3 subunits that form the catalytic core of the enzyme are encoded on the mitochondrial genome. Only one nuclear putative cytochrome c oxidase gene has been found, corresponding to subunit 6b. Several putative genes encoding assembly factors were identified, and it is possible that other cytochrome c oxidase genes exist but were too divergent to be identified. Four putative genes of the F0 subcomplex (complex V F_0F_1-ATP synthase) have been identified, three of which were also found in *C. reinhardtii* in addition with a putative gene for the ATP synthase:F0 component subunit 5H/D. The situation is similar for the F1 subcomplex where putative genes of all five subunits were found in *C. crispus* while only four where identified in *Chlamydomonas*.

Table S6.2. Genomic analysis of mitochondrial respiration Complex I components in C. *crispus*. Gene models in italics refer to genes found in the mitochondrial genome.

Protein	Gene ID/name
Bacterial NADH dehydrogenase orthologues	
NADH dehydrogenase subunit 1	*nad1*
NADH dehydrogenase subunit 2	*nad2*
NADH dehydrogenase subunit 3	*nad3*
NADH dehydrogenase subunit 4	*nad4*
NADH dehydrogenase subunit 4L	*nad4L*
NADH dehydrogenase subunit 5	*nad5*
NADH dehydrogenase subunit 6	*nad6*
NADH dehydrogenase subunit B or 10	CHC_T00009457001
NADH dehydrogenase subunit I	CHC_T00003883001
NADH dehydrogenase subunit E	CHC_T00009333001
NADH dehydrogenase subunit C	CHC_T00006423001
NADH dehydrogenase subunit D	CHC_T00008797001
NADH dehydrogenase subunit F	CHC_T00008807001
NADH dehydrogenase subunit G (C-ter)	CHC_T00000309001
NADH dehydrogenase subunit G (N-ter)	CHC_T00003765001

Eukaryote-specific subunits	
NDUFA1 subunit	Not found
Acyl carrier protein 1	CHC_T00009135001
NDUFA2 subunit	CHC_T00006825001
″	CHC_T00009060001
NDUFB3 subunit	Not found
NDUFA5 subunit	scaffold_509:4853..5384
NDUFS6 subunit	CHC_T00009546001
NDUFA6 subunit	CHC_T00007090001
NP17.3 subunit	CHC_T00003268001
NDUFS5 subunit	CHC_T00004832001
GRIM-19 protein	CHC_T00008785001
NDUFA12 subunit	CHC_T00000607001
NDUFB7 subunit	CHC_T00007379001
NDUFA4 subunit	CHC_T00006420001
NDUFA8 subunit	CHC_T00008582001
NDUFB9 subunit	CHC_T00007577001
NDUFB10 subunit	Not found
NDUFA9 subunit	CHC_T00009463001
Plant-specific subunits	
NADH:ubiquinone oxidoreductase 10 kDa subunit	Not found
NADH:ubiquinone oxidoreductase 13 kDa subunit	CHC_T00002428001
NADH:ubiquinone oxidoreductase 19 kDa subunit	Not found
γ carbonic anhydrase like	CHC_T00002095001
NDH11	Not found
NDH16	Not found
NADH:ubiquinone oxidoreductase 9 kDa subunit	Not found
NADH:ubiquinone oxidoreductase 16 kDa subunit	Not found
NADH:ubiquinone oxidoreductase 19 kDa subunit	Not found
subunits identified in several lineages	
TIM17/22 protein	scaffold_259:57551..58174
NDUFB4 subunit	Not found
NADH:ubiquinone oxidoreductase 20,9 kDa-like subunit	Not found
Other NADH dehydrogenase matches	
internal NADH dehydrogenase	CHC_T00001734001

NADH dehydrogenase/disulfide oxidoreductase	"
NAD(P)H dehydrogenase	"
Complex I intermediate associated protein	CHC_T00004991001
Type II NADH dehydrogenase	CHC_T00004265001
NAD(FAD)-dependent dehydrogenase like	CHC_T00007960001
Small nuclear ribonucleoprotein-associated protein E	CHC_T00009457001

Table S6.3. Genomic analysis of mitochondrial respiration Complex II components in *C. crispus*. Gene models in italics refer to genes found in the mitochondrial genome.

Protein	Gene ID/name
Succinate dehydrogenase subunit 1	CHC_T00008943001
Succinate dehydrogenase subunit 2	*sdh2*
Succinate dehydrogenase subunit 3	*sdh3*
Succinate dehydrogenase subunit 4	*sdh4*

Table S6.4 Genomic analysis of mitochondrial respiration Complex III components in *C. crispus*. Gene models in italics refer to genes found in the mitochondrial genome.

Protein	Gene ID/name
Core1	CHC_T00009305001
Core2	Not found
cytochrome c1	CHC_T00009218001
cytochrome b	*cob*
iron ferredoxin subunit1 (RIP1)	CHC_T00008759001
cytochrome c reductase subunit 6 (cytochrome b-566)	Not found
cytochrome c reductase subunit 7	CHC_T00006225001
cytochrome c reductase subunit 8	Not found
cytochrome c reductase subunit 9	CHC_T00007973001
cytochrome c reductase subunit 10	Not found

Table S6.5. Genomic analysis of mitochondrial respiration Complex IV components in *C. crispus*. Gene models in italics refer to genes found in the mitochondrial genome.

Protein	Gene ID/name
cytochrome c oxidase subunit 1	*cox1*
cytochrome c oxidase subunit 2	*cox2*
cytochrome c oxidase subunit 3	*cox3*
cytochrome c oxidase subunit 4	Not found
cytochrome c oxidase subunit 5a	Not found
cytochrome c oxidase subunit 5b	Not found
cytochrome c oxidase subunit 5c	Not found
cytochrome c oxidase subunit 6	Not found
cytochrome c oxidase subunit 6a or 13	Not found
cytochrome c oxidase subunit 6b or 12	CHC_T00009022001
cytochrome c oxidase subunit 7	Not found
cytochrome c oxidase subunit 7a	Not found
cytochrome c oxidase subunit 8	Not found
Assembly factors	
cytochrome c oxidase assembly protein COX10	CHC_T00005639001
cytochrome c oxidase assembly protein COX11	CHC_T00004335001
cytochrome c oxidase assembly protein COX14	Not found
cytochrome c oxidase assembly protein COX15	Not found
cytochrome c oxidase assembly protein COX16	Not found
cytochrome c oxidase assembly protein COX17	Not found
cytochrome c oxidase assembly protein COX18	CHC_T00001318001
cytochrome c oxidase assembly protein COX19	scaffold_2:361756..362127
cytochrome c oxidase assembly protein COX23	Not found
SCO1/SenC family protein	CHC_T00004179001
SCO2	scaffold_209:161298..161717
cytochrome c oxidase assembly SURF1/SHY1	CHC_T00001990001

Table S6.6. Genomic analysis of mitochondrial respiration Complex V components in *C. crispus*. Gene models in italics refer to genes found in the mitochondrial genome.

Protein	Gene ID/name
F1 subcomplex	
ATP synthase: F1 component, subunit α	CHC_T00009194001
ATP synthase: F1 component, subunit β	CHC_T00009073001
ATP synthase: F1 component, subunit γ	CHC_T00008404001
"	CHC_T00008719001
ATP synthase: F1 component, subunit δ	CHC_T00005902001
ATP synthase: F1 component, subunit ε	scaffold_323:90163..90339
Assembly factors	
ATP synthase: assembly factor for F1 component (ATP10)	CHC_T00002011001
ATP synthase: assembly factor for F1 component (ATP11)	CHC_T00004910001
ATP synthase: assembly factor for F1 component (ATP12)	CHC_T00007556001
F0 subcomplex	
ATP synthase: F0 component, subunit 6	atp6
ATP synthase: F0 component subunit 5H/D	CHC_T00006891001
ATP synthase: subunit 0 or OSCP	CHC_T00006782001
ATP synthase: F0 component, subunit 8	Not found
ATP synthase: F0 component, subunit 9	atp9
Assembly factors	
ATP synthase: assembly factor for F0 component (ATP23)	CHC_T00005910001

7 Metabolism

7.1 Amino acid metabolism

With the exception of genes related to nitrogen uptake, the urea cycle, GABA metabolism, and proline degradation, the *C. crispus* genome appears to contain a set of amino acid metabolism-related genes typical of plants. Except for the glutamine synthase, aspartate transaminase, and urea transporter genes, annotated sequences were found only as single copies (Table S7.1).

Table S7.1. Primary amino-acid metabolism related genes annotated in *C. crispus*. NI, genome support, but not included in genome (scaffold < 2kb)

Enzyme	EC	Gene ID
Nitrogen assimilation		
glutamate dehydrogenase	1.4.1.3	CHC_T00004408001

nitrate reductase	1.7.1.1	CHC_T00003369001
nitrite reductase	1.7.7.1	CHC_T00003048001
glutamine synthetase	6.3.1.2	CHC_T00005451001
″		CHC_T00005454001
″		CHC_T00001505001
glutamate synthase	1.4.7.1	Not found
″	1.4.1.13/14	Not found
asparagine synthetase	6.3.5.4	CHC_T00004458001
aspartate transaminase	2.6.1.1	CHC_T00005085001
″		CHC_T00001302001
″		CHC_T00000848001
Urea metabolism/cycle		
ornithine carbamoyltransferase	2.1.3.3	CHC_T00005123001
argininosuccinate synthase	6.3.4.5	CHC_T00000625001
argininosuccinate lyase	4.3.2.1	CHC_T00004007001
arginine-tRNA ligase	6.1.1.19	CHC_T00002717001
arginase	3.5.3.1	Not found
urease	3.1.5.1	Not found
ornithine transaminase	2.6.1.13	CHC_T00002893001
ornithine decarboxylase	4.1.1.17	SNAPCcT00025013001
urea amidolyase/allophanate hydrolase	3.5.1.54	CHC_T00009426001
urea amidolyase/urea carboxylase	6.3.4.6	CHC_T00009426001
urea transporter		CHC_T00009208001
″		CHC_T00008887001
γ-aminobutyric acid (GABA)		
glutamate decarboxylase	4.1.1.15	Not found
GABA aminotransferase	2.6.1.22	Not found
diamine oxidase	1.4.3.22	Not found
Proline		
pyrroline-5-carboxylate reductase K	1.5.1.2	CHC_T00001376001
proline dehydrogenase	1.5.99.8	Not found
1-pyrroline-5-carboxylate dehydrogenase	1.5.1.12	NI
1-pyrroline-5-carboxylate synthase		CHC_T00008349001

7.2 Nitrogen uptake

One nitrite reductase, one nitrate reductase, three glutamine synthases (GS), and one glutamate dehydrogenase (GDH), but no glutamate synthase (GOGAT) were found in the genome (Fig. S7.1). GOGAT is part of the GS/GOGAT cycle, which serves nitrogen assimilation in land plants (172) (173). In this cycle, one molecule of glutamate, 2-oxagluarate, and ammonium serve as substrates for glutamine synthesis through the activity of GS. GOGAT then cleaves glutamine producing two molecules of glutamate. A homolog of a GOGAT also seems to be missing in the *C. merolae* genome, so that in both red algae, glutamate would need to be synthesized directly via GDH. This enzyme is used for nitrogen assimilation in bacteria (174), and is considered to play a regulatory role in carbon- and nitrogen metabolism in land plants (175).

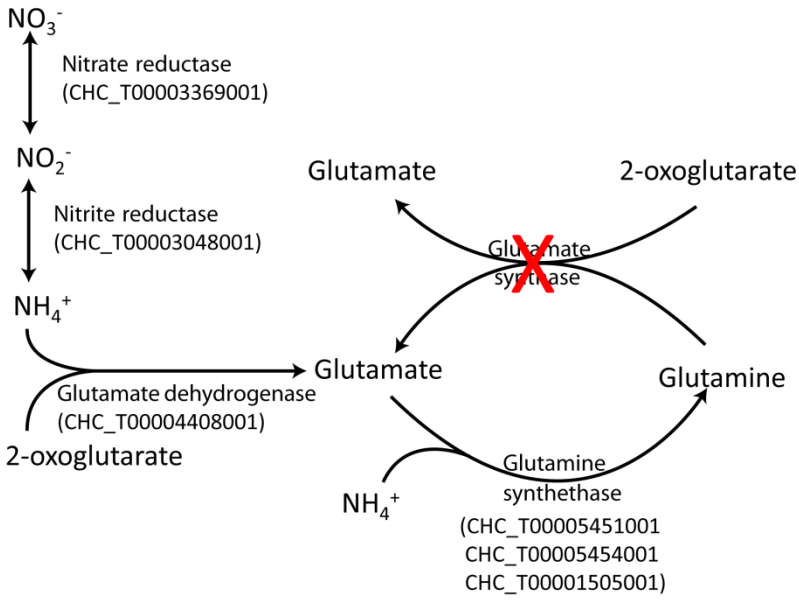

Fig. S7.1. Nitrogen assimilation in *C. crispus*.

7.3 Urea metabolism and urea cycle

No arginases or ureases were found in the *C. crispus* genome, and only a fragment of an argininosuccinate synthase was present (Table S7.1). The *C. merolae* genome contains a complete argininosuccinate synthase (gnl|CMER|CMT305C), but also lacks arginases and ureases. These enzymes are part of or close to the urea cycle, and homologs of characterized proteins are present in *Arabidopsis* and all other plant genomes included in the Phytozome database (www.phytozome.net), except in those of the green algae *Volvox carteri* and *C. reinhardtii*.

Unlike *C. merolae*, *C. crispus* possesses a putative urea carboxylase/allophanate hydrolase. This enzyme (or pair of enzymes, together referred to as urea amidolyase) has been described in yeast (176), but also detected in several green algae (177). Urea amidolyases degrade urea, producing

ammonium and CO_2 at the expense of ATP, and may thus substitute ureases. In *C. crispus*, both enzymatic activities of urea amidolyases are predicted to be contained in one fusion enzyme, the carboxylase part of the enzyme is in the N-terminal part of the protein, and the hydrolase part in the C-terminal region. In yeast, the opposite order is found. In *C. reinhardtii*, both functional units are predicted as separate proteins but in the same order as in *C. crispus*. Furthermore, the *C. crispus* genome, just as that of *C. reinhardtii* and yeast, encodes putative urea transporters. Together these data indicate that *C. crispus* can probably metabolize urea.

7.4 Gamma-aminobutyric acid

Gamma-aminobutyric acid (GABA) is a non-protein amino acid which serves as neurotransmitter in animals (178). In land plants, this compound is known to act as a phytohormone, but also to furnish intermediates for the TCA cycle (179). Just as the brown alga *E. siliculosus* (169), *C. crispus* lacks genes encoding glutamate decarboxylases (GADs) and GABA aminotransferases (GABA-Ts), both genes involved in the synthesis and degradation of GABA. *Cyanidioschyzon merolae* possesses a GAD (gnl|CMER|CMF072C), but probably no GABA-T. *Ectocarpus siliculosus* has been suggested to use an alternative pathway to synthesize low quantities of GABA from polyamines, especially in response to salt stress (180). However, *C. crispus* also lacks the putative key gene of this alternative pathway, a diamine oxidase. It is therefore probably not able to produce GABA.

7.5 Proline metabolism

The genome codes for a pyrroline-5-carboxylate reductase as well as a 1-pyrroline-5-carboxylate synthase. *Chondrus crispus* can thus be expected to be able to synthesize proline from glutamate. However, due to the afore-mentioned lack of arginases, proline synthesis can most likely not take place from arginine. With respect to proline degradation, a proline dehydrogenase was not found in the *C. crispus* genome, but a 1-pyrroline-5-carboxylate dehydrogenase sequence required to oxidize 1-pyrroline-5-carboxylate to glutamate was identified and grouped with plant sequences in a phylogenetic analysis. This is in contrast to our analysis of the *C. merolae* genome, where a proline dehydrogenase gene was found (gnl|CMER|CMG213C), but not a 1-pyrroline-5-carboxylate dehydrogenase gene.

Both proline dehydrogenase and 1-pyrroline-5-carboxylate dehydrogenase are present in all 22 plant genomes in the Phytozome database (www.phytozome.net). In this sense, the comparison of *C. crispus* and *C. merolae* is of particular interest: both genomes lack different genes involved in the degradation of proline, indicating that this pathway may currently be in the process of being reduced independently in both species.

7.6 Fatty acid metabolism

Production of plastidial fatty acids, polyunsaturated fatty acids and sphingolipids

Several reports have been published on the composition and partitioning of fatty acids (FAs) in red algae, in particular in membrane lipids in *C. cripus* (181) (182) (183) (184). Genes encoding enzymes catalysing the different steps of plastidial FA synthesis from acetyl-CoA and malonyl-CoA have been identified in the genome, as well as proteins involved in the transport of these fatty acids to the cytosol. *C. crispus*, like *C. merolae*, does not contain any ACP desaturase, suggesting that no FA desaturation occurs in the plastid. However, an ortholog of CMM045C, corresponding to a stearyol-CoA desaturase (to produce 18:1 from 18:0 in the endoplasmic reticulum), has been identified. The oleic acid can then be desaturated by a Δ12- and a Δ15-desaturases since one sequence for each enzyme has been identified. This is different when compared to the situation in *E. siliculosus*, where two forms (one microsomal and one plastidic) were identified for both types of desaturases. To complete the set of desaturases necessary for the production up to 18:4n3, 20:4n6 and 20:5n3, two other genes potentially encoding polyunsaturated fatty acids (PUFAs) Δ6- and Δ5-desaturases have been annotated. However, specificity of these desaturases cannot be inferred solely from the sequence.

Beside desaturases, proteins involved in the fatty acid elongase complex have been searched for. It involves four enzymes, the first one conferring the specificity to the complex. This first protein, a 3-ketoacyl-CoA synthase or "elongase", can be of two types, in relation with the specificity of the complex: elongation of saturated/monounsaturated fatty acid or elongation of PUFAs. Other activities are 3-keto-acyl-CoA reductase (step 2), 3-hydroxyacyl-CoA dehydratase (step 3), and enoyl-CoA reductase (step 4). At least one gene has been identified for each enzyme, indicating that *C. crispus* contains a functional elongase complex, and considering the low level of redundancy in its genome, it is worth to mention that the PUFA elongases may be represented by four genes. PUFAs are transported in the cytoplasm after binding to an acyl-CoA molecule to increase their solubility, and several candidates have been annotated as putative acyl-CoA synthetases. In parallel, one gene coding for an acyl-CoA-binding protein was found, that binds acyl-CoA esters to protect acyl-CoAs from degradation by microsomal acyl-hydrolases.

Sphingolipids represent another class of FA and lipids, which are structural components of membranes, and have also signalling roles in plants and mammals. Sphingolipids have not been profiled in details in red algae, even if two publications have reported the presence of sphingolipids containing inositol (185) (186) in some rhodophytes. In the *C. crispus* genome, genes have been searched using literature data (187). Orthologs of the genes encoding enzymes necessary for the production of 4-hydroxysphinganine, sphingosine-1-phosphate and ceramides from serine and palmitoyl-CoA have been identified. It is interesting to note that the red algal genome contains two potential sphingolipid Δ4-desaturases (which is different from *E. siliculosus* where no candidates for this activity were identified), and one potential sphingolipid Δ8-desaturase.

One gene potentially coding for a ceramide glucosyltransferase (GCS, GT2 family, found in land plants) has been identified. In the same vein, several candidates for GT enzymes were found that could encode potential ceramide galactosyltransferases. No gene for a ceramide kinase (normally found in plants) was identified. In contrast, one candidate for an inositolphosphorylceramide synthase, otherwise found only in fungi and bacteria, was identified and may represent an example of lateral gene transfer.

Table S7.2. Fatty acid biosynthesis and plastidial transport in *C. crispus.*

Enzyme	EC number	Gene ID
acetyl-CoA carboxylase/biotin carboxylase	6.4.1.2 /6.3.4.14	CHC_T00009359001
″		CHC_T00007145001
acetyl-CoA carboxylase	6.4.1.2	CHC_T00008359001
acyl carrier protein	1.6.5.3; 1.6.99.3	CHC_T00009135001
″		CHC_T00009356001
[acyl-carrier-protein] malonyl-CoA:ACP transacylase	2.3.1.39	CHC_T00008765001
β-ketoacyl synthase (FabB)	2.3.1.41	CHC_T00009465001
3-oxoacyl-[acyl-carrier-protein] synthase II (FabF)	2.3.1.179	CHC_T00008601001
3-oxoacyl-[acyl-carrier-protein] reductase (FabG)	1.1.1.100	CHC_T00008517001
″		CHC_T00008477001
″		CHC_T00008496001
β-hydroxyacyl-ACP dehydratase (FabZ)	4.2.1.17	CHC_T00009190001
enoyl-ACP reductase (FabI)	1.3.1.9	CHC_T00009325001
fatty acyl-ACP thioesterase A (FatA) or B (FatB)	3.1.2.-	Not found

Table S7.3. FA elongation in *C. crispus* mitochondria**.**

Enzyme	EC number	Gene ID
mitochondrial substrate carrier family protein		CHC_T00009209001
″		CHC_T00009479001
″		CHC_T00008818001
acetyl-CoA acyltransferase	2.3.1.16	CHC_T00009349001
3-hydroxyacyl-CoA dehydrogenase	4.2.1.17	CHC_T00009110001
″	4.2.1.17	CHC_T00009422001
″	4.2.1.17	CHC_T00008794001
″	4.2.1.17/1.1.1.35	CHC_T00009349001
″	4.2.1.17/4.2.1.55	CHC_T00008882001
mitochondrial trans-2-enoyl-CoA reductase	1.3.1.38	CHC_T00008594001

Table S7.4. Polyunsaturated fatty acid biosynthesis: microsomal elongase complex in *C. crispus*.

Enzyme	Comments	Gene ID
PUFA elongase	first step	CHC_T00007134001
FA elongase	first step	CHC_T00009084001
FA elongase	first step	CHC_T00009032001
FA elongase	first step	CHC_T00008855001
fatty acid elongase 3-ketoacyl-CoA synthase	first step	CHC_T00009104001
3-keto-acyl-CoA reductase	second step	CHC_T00008557001
″		scaffold_80:161845..162537
3-hydroxyacyl-CoA dehydratase	third step	CHC_T00008580001
enoyl-CoA reductase	fourth step	CHC_T00008438001

Table S7.5. Fatty acids related desaturases in *C. crispus*. NI, genome support, but not included in genome (scaffold < 2kb)

Enzyme	Gene ID
stearoyl-CoA desaturase	CHC_T00009557001
Δ12-desaturase	CHC_T00008296001
Δ15-desaturase	CHC_T00009454001
PUFA desaturase	CHC_T00008301001
″	2 NI
PUFA or sphingolipid desaturase	CHC_T00007134001
fatty acid hydroxylase	scaffold_43:204464..205681
sphingolipid Δ-4 desaturase	CHC_T00009401001
	CHC_T00009094001

Table S7.6. Acyl-CoA synthetases in *C. crispus*.

Enzyme	EC number	Gene ID
fatty acyl-CoA synthetase	6.2.1.3	CHC_T00009428001
″		CHC_T00008811001
″		CHC_T00008500001
″		CHC_T00008160001
″		CHC_T00008472001
″		CHC_T00008703001

Table S7.7. Kennedy pathway: Acyl-CoA dependant TAG synthesis in *C. crispus*. NI, genome support, but not included in genome (scaffold < 2kb)

Enzyme	EC number	Gene ID
glycerol kinase	2.7.1.30	CHC_T00008744001
glycerol-3-phosphate O-acyltransferase	2.3.1.15	CHC_T00008773001
gysocardiolipin acyltransferase	2.3.1.51	CHC_T00008713001
or lysophosphatidyl acyl acyltransferase		CHC_T00009396001
or 1-acylglycerol-3-phosphate O-acyltransferase		CHC_T00009195001
lysophospholipid acyltransferase		CHC_T00008526001
phosphatidic acid phosphatase	3.1.3.4	CHC_T00009530001
diacylglycerol kinase	2.7.1.107	CHC_T00008540001
acyl-CoA DAG acyltransferase	2.3.1.20	CHC_T00009240001
DGAT type 1		scaffold_63:334465..334962
″		scaffold_63:333734..333997
″		CHC_T00007755001
″		scaffold_83:190481..190744
DGAT type 2 and/or MGAT		NI
lecithin-cholesterol acyltransferase or phospholipid-cholesterol acyltransferase	2.3.1.-	CHC_T00009171001

Synthesis of membrane (plasmic and chloroplastic) lipids

The genome contains a full set of genes encoding enzymes involved in the synthesis of the main plasmic membrane phospholipids. In addition, different genes corresponding to enzymes involved in the synthesis of thylakoid membranes were identified: UDP-sulfoquinovose synthase (SQD1, 2 genes), sulfolipid synthase (SQD2, 1 gene), monogalactosyldiacylglycerol synthase (MGDGS, 1 gene), and digalactosyldiacylglycerol synthase (DGDGS, 1 gene). These represent fewer copies than for the *Arabidopsis* genome which contains three and two genes for MGDGS and DGDGS respectively.

Table S7.8. Acyl-CoA independent TAG synthesis and glycerophospholipid metabolism in *C. crispus*. NI, genome support, but not included in genome (scaffold < 2kb)

Enzyme	EC number	Gene ID
phospholipid:diacylglycerol acyltransferase		NI
glycerol-3-phosphate dehydrogenase cytosolic	1.1.1.8	CHC_T00009494001
″		CHC_T00008461001

ethanolamine-phosphate cytidylyltransferase	2.7.7.14	CHC_T00008907001
metallophosphoesterase)	3.6.1.16	CHC_T00008291001
″		CHC_T00009409001
phospholipase D	3.1.4.4	CHC_T00005313001
″		scaffold_175:36876..39733
glycerol-3-phosphate acyltransferase	2.3.1.15	CHC_T00004435001
″		CHC_T00008773001
1-acyl-sn-glycerol-3-phosphate acyltransferase	2.3.1.51	CHC_T00009396001
″		CHC_T00009195001
″		CHC_T00008713001
″		CHC_T00008325001
″		CHC_T00008526001
″		CHC_T00008844001
phosphatidate cytidylyltransferase	2.7.7.41	CHC_T00006028001
″		CHC_T00009577001
CDP-diacylglycerol-inositol 3-phosphatidyltransferase	2.7.8.11	CHC_T00008593001
″		CHC_T00009507001
phosphatidylcholine synthase	2.7.8.24	CHC_T00008854001
cardiolipin synthase	2.7.8.5	CHC_T00009507001
″		CHC_T00008428001
phosphatidylserine decarboxylase	4.1.1.65	CHC_T00008892001
phosphatidic acid phosphatase type 2	3.1.3.4	CHC_T00009530001
″		CHC_T00008446001
diacylglycerol kinase	2.7.1.107	CHC_T00008540001
phosphatidylcholine synthase	2.7.8.24	CHC_T00008854001
phosphatidylethanolamine N-methyltransferase	2.1.1.17	CHC_T00008258001
″		CHC_T00009046001
″		CHC_T00009414001
lysophospholipase I	3.1.1.5	CHC_T00009450001
″		CHC_T00009177001
glycerophosphodiester phosphodiesterase	3.1.4.46	CHC_T00006160001
″		CHC_T00009254001
choline-phosphate cytidylyltransferase	2.7.7.14	CHC_T00008907001
choline/ethanolamine kinase	2.7.1.82	CHC_T00008383001

Table S7.9. Sulfolipid biosynthesis in *C. crispus.*

Enzyme	EC number	Gene ID
UDP-sulfoquinovose synthase SQD1	3.13.1.1	CHC_T00009188001
″		CHC_T00008950001
″		CHC_T00009498001
sulfolipid synthase	2.4.1.-	CHC_T00003861001
monogalactosyldiacylglycerol synthase (MGDGS)	2.4.1.46	CHC_T00008619001
digalactosyldiacylglycerol synthase (DGDGS)	2.4.1.241	CHC_T00008344001
glycerol dehydrogenase	1.1.1.6	CHC_T00001993001
dihydroxyacetone kinase=glycerone kinase	2.7.1.29	CHC_T00009570001

Synthesis of neutral/storage lipids: triacylglycerols (TAGs)

Genes coding for enzymes involved in the different steps of the Kennedy pathway (acyl-CoA dependent pathway for TAG synthesis) have been identified; in particular, five sequences very similar have been annotated for the acyl-CoA diacylglycerol acyltransferase of type 1, even if none of them is supported by ESTs. Moreover, one candidate for a DGAT of type 2 was identified. One candidate for a potential lecithin: cholesterol acyltransferase or a phospholipid: diacylglycerol acyltransferase (PDAT) was found. If this gene really corresponds to a PDAT, this implies that *C. crispus* can use both the acyl-CoA dependent and independent pathways to store fatty acids in TAGs, in contrast to the situation observed in *E. siliculosus*. No gene encoding oleosin was found.

Oxylipin synthesis and derivatives

Profiles for oxylipin synthesis and derivatives compounds have been reported for *C. crispus* (188) (189), and show the production of C18 (plant like) and C20 (animal like) oxylipins from free FAs cleaved by phospholipases (several candidates in the red algal genome). Interestingly, only two genes encoding lipoxygenase (LOX) have been identified, which is surprising considering the diversity of oxylipins observed in this alga. They only exhibited 20% of identity after pairwise comparison. One of these LOXs features a high identity level with a LOX previously identified in *Porphyra purpurea* (190).

The presence of methyl jasmonate (MeJA), a well known plant hormone involved in stress signalling, has been detected *in vitro* using cell-free homogenates of *C. crispus* after incubation with linolenic acid (188). However, no candidates for an allene oxide synthase (AOS), belonging to the CYP74 family of proteins was found. No candidate was found for an allene oxide cyclase (AOC), while two genes were identified as 12-oxo-PDA (OPDA) reductase. OPDA is further decarboxylated by three steps of β-oxidation, and a full set of genes encoding enzymes involved in this process was found in the genome. No candidate for jasmonic acid carboxyl methyltransferase (JMT) was

identified, indicating that it may be due in *C. crispus* to an enzyme different from the ones characterized so far in land plants.

Other enzymes involved in oxylipin synthesis and belonging to the CYP74 family, such as hydroperoxide lyases (HPL) and divinyl ether synthases (DES), have been searched for but not found. In addition, a blast analysis with two polyenoic fatty acid isomerases from the red alga *Ptilota filicina* did not provide any hits in the *C. crispus* genome.

Table S7.10. Oxylipins and derivates synthesis in *C. crispus.* NI, genome support, but not included in genome (scaffold < 2kb)

Enzyme	EC number	Gene ID
LOX	1.13.11.-	CHC_T00008739001
fatty acid isomerase		NI
AOS = CYP74A	4.2.1.92	Not found
HPL =CYP74B and C		Not found
12-oxo-PDA reductase	1.3.1.42	CHC_T00000698001
or 12-oxophytodienoate reductase		CHC_T00009377001
dioxygenase (DIOX or DOX, C18)		CHC_T00008635001
″		CHC_T00009516001
cyclooxygenase (COX, C20)	1.14.99.-	CHC_T00009490001
″		CHC_T00001140001
″		CHC_T00008635001
divinyl ether synthase (DES)=CYP74D		Not found

Table S7.11. Sphingolipids and ceramides synthesis. EST, EST support without genome sequence support.

Enzyme	EC number	Gene ID
serine palmitoyltransferase	2.3.1.50	
″	LCB2	CHC_T00008697001
″	LCB1	CHC_T00008326001
3-ketodihydrosphingosine reductase	1.1.1.102	CHC_T00008374001
acyl-CoA binding protein, fatty acyl-CoA carrier		CHC_T00008546001
(acyl-CoA dependent) ceramide synthase	2.3.1.24	CHC_T00009025001
sphinganine-1-phosphate aldolase/lyase	4.1.2.27	CHC_T00009510001
sphingosine kinase	2.7.1.91	scaffold_105:73236..73910
glycosyl-ceramide synthase (GCS)	2.4.1.80	EST
UDP-galactose-ceramide galactosyltransferase.	2.4.1.45	P_UA0AAA63YO16FM1
inositolphosphoryl ceramide synthase (IPCS)		CHC_T00008908001

Table S7.12. Fatty acid metabolism in *C. crispus.*

Enzyme	EC number	Gene ID
AMP-dependent synthetase and ligase, long chain acyl-CoA synthetase	6.2.1.3	CHC_T00009428001
		CHC_T00008811001
acyl-CoA oxidase	1.3.3.6	CHC_T00008400001
″		CHC_T00008478001
″		CHC_T00009531001
″		CHC_T00007910001
″		CHC_T00008935001
″		CHC_T00009142001
aminoglycoside phosphotransferase + acyl-CoA dehydrogenase	1.3.99.	CHC_T00009166001
alcohol dehydrogenase	1.1.1.1/1.1.1.184	CHC_T00007717001
aldehyde dehydrogenase, putative	1.2.1.3	CHC_T00008341001
″		CHC_T00009129001
acetyl-CoA C-acetyltransferase	2.3.1.9	CHC_T00008981001

7.7 Carbon storage and cell wall metabolism

Carbohydrate metabolism is one of the traits which render the red algae a unique group of eukaryotes. In contrast to plants and green algae which produce starch in their chloroplasts, red algae store carbon in their cytosol as insoluble starch granules (floridean starch). Red algae also produce complex

polysaccharides which constitute their cell wall. They synthesize some neutral polysaccharides in common with land plants, such as crystalline cellulose, mannan, and xylans. But above all, red algae produce unique sulphated polysaccharides: sulphated hemicelluloses (*e.g.* sulphated 1,3-1,4-β-D-glucans), and sulphated galactans, which constitute the essential fraction of the cell wall matrix (191).

The key enzymes for the synthesis and remodelling of oligo- and polysaccharides are glycoside hydrolases (GH) and glycosyltransferases (GT), which are classified in more than 200 carbohydrate active enzyme (CAZY) families (192). We have identified 31 glycoside hydrolases and 65 glycosyltransferases, belonging to 16 GH and 27 GT families, respectively (Table S7.13). *Chondrus crispus* has slightly less GH/GT genes than the brown alga *E. siliculosus* (193) (194), and about six times less than found in land plants (195). As observed for brown algae, the main difference between red macroalgae and plants is the presence of large multgenic families in plants. With the exception of families GT14, GT39, and GT77, the other GH and GT families of *C. crispus* contain few paralogous genes. Interestingly, this red alga contains some families which are absent in plants but conserved with other phyla: bacteria (family GT45 and GT78), fungi (GH6, GH45, and GT39) and animals (GT7).

The CAZymes involved in starch biosynthesis (families GT5 and GH13) and degradation (families GT35, GH13, GH14 and GH77) are all found the genome of *C. crispus*. The starch metabolism in this red alga is described elsewhere (Section 7.12). This red alga also possesses a complete trehalose pathway. Trehalose is synthesized by a family of four 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). In Plants, this non-reducing disaccharide is a central metabolic regulator, coordinating the starch and sucrose pathways (196). However, sucrose metabolism is completely absent from *C. crispus*, as deduced from the lack of sucrose synthase and sucrose phosphate synthase (family GT4), and of invertases (families GH32 and GH100). Therefore, the role of trehalose in red algae is unclear and remains to be determined.

The most surprising finding concerning the synthesis of low-molecular weight carbohydrates is the presence in the *C. crispus* genome of a gene highly similar to the GT78 mannosylglycerate synthase from the marine bacterium *Rhodothermus marinus* (48% identity). This enzyme catalyzes the transfer of a mannose residue from GDP-mannose to a glycerate moiety (197). Mannosylglycerate is almost restricted to thermophilic and hyperthermophilic bacteria and Archaea and is required for osmotic and thermal adaptation (198). This compound is nonetheless found in red algae as photo-assimilate. Mannosylglycerate, also referred to as digeneaside, accumulates during photosynthesis in some red algae and is commonly used for chemical taxonomy (199). Very few GT78 homologues are found in Genbank NR. These proteins are only found in some Archae, in the red alga *Griffithsia japonica*, and in two early land plants, the moss *Physcomitrella pantens* and the lycophyte *Selaginella moellendorffii*. A phylogenetic analysis of the GT78 family reveals that the sequences from red algae

and from the early land plants constitute two distinct clades which are both rooted by the mannosylglycerate synthase from *R. marinus* (Fig. 7.2). A tblastn search in the EST Genbank database identified several additional organisms expressing putative GT78: four red algae (*Kappaphycus alvarezii*, *Gracilaria changii*, *Porphyridium cruentum* and *Eucheuma denticulatum*), and four green algae (*Penium margaritaceum*, *Klebsormidium flaccidum*, *Spirogyra pratensis*, *Chaetosphaeridium globosum*). Despite the scarcity of genomic and transcriptomic data on red algae, GT78 genes seem present in most red algal species. In contrast, such genes are absent in all available genomes of higher plants and of volvocale and prasinophyte green algae. In the green lineage, mannosylglycerate synthase genes are only present in few charaphytic algae and in two early land plants. It is highly unlikey that GT78 genes would be present in the common ancestor of red algae and green algae / plants, because this scenario would require multiple gene losses to explain the paucity of this gene family in the green lineage. Altogether, the most parcimonious scenario is that red algae horizontally acquired mannosylglycerate synthase from an ancestral thermophilic marine bacterium and likely retain digeneaside as a specific osmolyte. Mannosylglycerate has not been identified in *C. crispus* yet, but this species is known to accumulate other heterosides: floridoside (2-O-D-glycerol-α-D-galactopyranoside), and isofloridoside (1-O-D-glycerol-α-D-galactopyranoside) (199). Digeneaside may be only produced in *C. crispus* in response to an osmotic stress. An alternative hypothesis is a change in the enzymatic reaction of the GT78 in *C. crispus* compared to the role of this enzyme in *R. marinus*, and its potential involvement in the synthesis of floridoside or isofloridoside. Similarly, the presence of mannosylglycerate synthase genes in a limited number of green algae and early land plants is best explained by independent horizontal gene transfer. The existence of mannosylglycerate in these plants remained to be demonstrated.

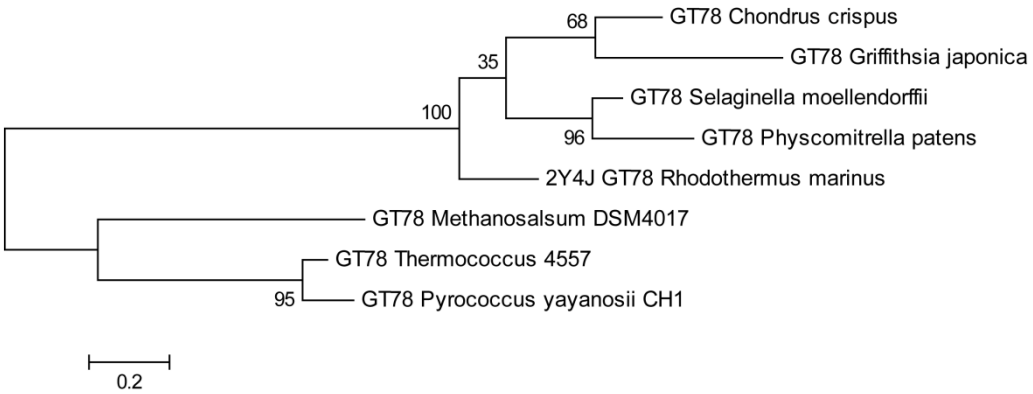

Fig. S7.2. Unrooted phylogenetic tree of glycosyltransferases of the family **78 (GT78)**.

The current knowledge on cell wall metabolism in red seaweed is very limited. The genome of *C. crispus* is thus the first broad data resource allowing the prediction of some aspects of this crucial metabolism. We have identified three cellulose synthase-like proteins (family GT2). These integral membrane glycosyltransferases used UDP-glucose as activated sugar. The first two proteins, which are

highly similar to the cellulose synthases (CESA) from the red algae *P. yezoensis* (200) and *Griffithsia monilis* (201), are related to the cyanobacterial CESA. Surprisingly, the third GT2 emerges in a phylogenetic clade comprising the CESA from firmicutes and *E. coli*, but not the plant and cyanobacterial CESA. Thus, cellulose biosynthesis in red algae may have a dual origin: the cyanobacterial primary endosymbiosis, and a second horizontal gene transfer (HGT) event involving a distinct ancestral bacterium.

The genome features numerous genes in GT families involved in the synthesis of glycosaminoglycans in animals and of hemicelluloses and pectins in land plants (families GT8, GT14, GT47 and GT64). These GT families are likely involved in hemicellulose biosynthesis in red algae. *Chondrus crispus* also possesses genes homologous to animal glycosaminoglycan sulfotransferases, and chondroitin β-1,4 N-acetylgalactosaminyltransferase (GT7). These enzymes are involved in the biosynthesis of sulphated glycosaminoglycans which constitute the extracellular matrix of animals. Therefore, these homologous genes in red algae likely play a role in the biosynthesis of carrageenans. Carbohydrate sulfotransferases are also conserved in brown algae, but are absent in land plants, confirming that the synthesis of sulphated polysaccharides is an ancient eukaryotic capacity, which has been lost by plants during their land conquest (193).

The three GH16 enzymes of *C. crispus* are closely related to κ-carrageenases from marine bacteria (202) and are likely involved in cell wall remodelling and expansion. The genome contains three families of cellulases (GH5, GH6, GH45), which are absent in plants, but surprisingly conserved with fungi and amoebozoa. These cellulases are thus ancestral enzymes, predating the cyanobacterial primary endosymbiosis, and initially involved in the degradation of bacterial cellulose. After the acquisition of the cellulose biosynthetic pathway, these red algal enzymes likely evolved to participate to the cell wall remodelling.

All together, the common ancestor of red algae was probably a cellulolytic protist feeding on bacteria (such as the extant *Dictyostelium*), and thus repeatedly exposed to prokaryotic genes, explaining the HGTs of various bacterial origins (primary endosymbiosis, but also GT2, GT45 and GT78 for instance). The GT78 HGT also indicates that the red algal common ancestor was exposed to a marine thermophilic environment at some point of its evolutionary history. Finally, red algae have evolved a carbohydrate metabolism profoundly different from land plants. In particular, cell wall metabolism largely remains an uncharted territory and numerous CAZY families wait to be discovered in red algae.

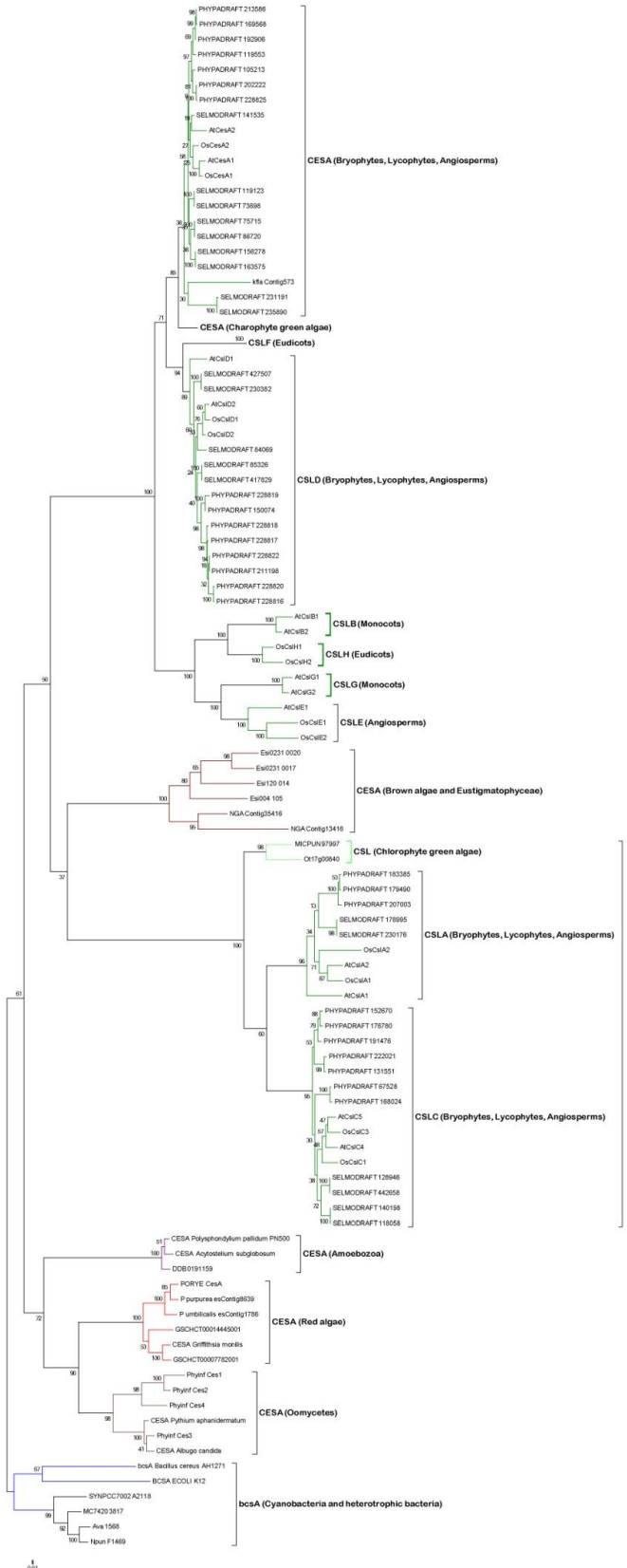

Fig. S7.3. Phylogenetic tree of the cellulose synthases (CESA) and cellulose synthase-like (CSL) proteins (family GT2). All of the phylogenetic trees presented here were constructed using the maximum likelihood (ML) approach with the program MEGA5. Numbers indicate the bootstrap values in the ML analysis. The tree is rooted by the bacterial CESA.

Table S7.13. Number of genes in each family of glycoside hydrolases (GH) and glycosyltransferases (GT) identified in the genome of *C. crispus*.

GH1	**GH5**	**GH6**	**GH13**	**GH14**	**GH16**	**GH31**	**GH35**	**GH36**
2	1	2	4	1	3	2	2	2
GH37	**GH38**	**GH45**	**GH47**	**GH63**	**GH77**	**GH85**		
1	1	3	4	1	1	1		
GT2	**GT4**	**GT5**	**GT7**	**GT8**	**GT13**	**GT14**	**GT20**	**GT22**
3	3	1	4	2	2	9	4	1
GT24	**GT27**	**GT28**	**GT33**	**GT34**	**GT35**	**GT39**	**GT41**	**GT45**
1	2	4	1	1	1	6	1	1
GT47	**GT57**	**GT58**	**GT59**	**GT64**	**GT66**	**GT77**	**GT78**	**GT90**
3	2	1	1	2	1	5	1	2

7.8 Polysaccharide sulphate active enzymes

Chondrus crispus is a carrageenophyte, in the gametophyte the cell wall is composed mainly of κ- and ι-carrageenan, whereas λ-carrageenan is found in the sporophytic phase of the life cycle. Carrageenans are linear galactans classified according to the number and position of sulphate groups, and by the occurrence of the 3,6-anhydro bridges in the α-linked residues (203) (204) (205). The anabolic pathways of carrageenans are not well known. However, it is generally accepted that the carrageenans are synthesized in the Golgi apparatus as a linear and neutral backbone of galactan by β-galactosyl transferases that catalyze the polymerization of galactose residues. Then, sulphotransferases and galactose-6-sulphurylases decorate the neutral galactan by incorporation of sulphate groups, probably using 3'-phosphoadenosine 5'-phosphosulphate (PAPS) as sulphate donor, and remove the C6-sulphate from precursors to form 3,6-anhydro rings (206). Finally, it is possible that carrageenans undergo many modifications of their sulphation level by sulphatases in order to change the plasticity of cell wall as a response to environmental conditions. Occurrence of sulphated galactans probably suggests the existence of an extensive sulphur metabolism.

7.9 Sulphotransferases

Sulphotransferases catalyze the transfer of a sulphate group from a donor to an acceptor molecule. In the *C. crispus* genome, twelve genes encoding sulphotransferases have been found. On the basis of

sequence similarity, sulphotransferases are classified in several families and members of family 1, family 2 and galactose-3-O-sulphotransferase family were identified.

Family 1

This family is divided into two subfamilies. Subfamily 1 is essentially composed of non-carbohydrate sulphotransferases such as arylsulphotransferases that catalyze the sulphate conjugation of catecholamines, phenolic drugs, and neurotransmitters (207), of alcohol and hydroxysteroid sulpho-transferases that catalyze sulphation of hydroxysteroids and xenobiotics (208), of amine sulphotransferases that catalyze N-sulphation of amines (209), of estrogen sulphotransferases that catalyze the sulphation of estradiol and estrone in animals (210), of thyroxin sulphotransferase (211), of flavonol 3-sulpho-transferases that may play a role in auxin transport, and of desulpho-glucosinolate sulphotransferases involved in the biosynthesis of the glucosinolate (212). Subfamily 1 also comprises, some carbohydrate sulphotransferases such as N-acetylgalactosamine (GalNAc) 4-sulphate 6-O-sulphotransferases that transfer sulphate to the C-6 of GalNAc 4-sulphate residue present in chondroitin sulphate and dermatan sulphate (213), heparan sulphate glucosamine 3-O-sulphotransferases involved in the biosynthesis of anticoagulant heparan sulphate (214), [heparan sulphate]-glucosamine N-sulphotransferases that are bifunctional enzymes catalyzing both the N-deacetylation and the N-sulphation of glucosamine (GlcNAc) in heparan sulphate, and glycolipid sulpho-transferases involved in the of bacterial cell wall sulpholipids biosynthesis (215).

The subfamily 2 is composed of carbohydrate sulphotransferases specifically involved in the transfer of sulphate group to position 6 of Gal, GlcNAc and GalNAc residues. It corresponds to chondroitin 6-O-sulphotransferases that catalyze the transfer of sulphate to N-acetylgalactosamine residues of chondroitin (216), to N-acetylglucosamine 6-O-sulphotransferases that catalyzes the transfer of sulphate to of non-reducing N-acetylglucosamine residues within keratan-like structures on N-linked glycans and within mucin-associated glycans (217), and to galactose 6-O-sulphotransferases catalysing the transfer of sulphate group on galactose residues in keratan (218).

Four sulphotransferase genes belonging to the family 1 of sulphotransferases were found. The phylogenetic analysis of the family 1 shows that three sequences form a clade that is related to the glycolipid sulphotransferases (Fig. S7.4). This result seems to indicate that these proteins could be involved in sulphation of lipids in membranes. CHC_T00008796001 is distantly related to the clade E1 of *E. siliculosus*. As it was postulated that these sequences could be involved in the fucoidan biosynthesis in *E. siliculosus* (193), it is possible that the sequence CHC_T00008796001 participates to the carrageenan biosynthesis. It is interesting to note that, for this family, 14 proteins putatively involved in the biosynthesis of fucans were identified in *E. siliculosus*, while only one sequence was found in *C. crispus*. Moreover, four sequences from *E. siliculosus* are related to animals arylsulphotransferases (193), putatively involved in sulphation of hormones, drugs, or xenobiotic

compounds, while no sequences from *C. crispus* are related to arylsulphotransferases, suggesting the absence of arylsulphate compounds in this marine red alga.

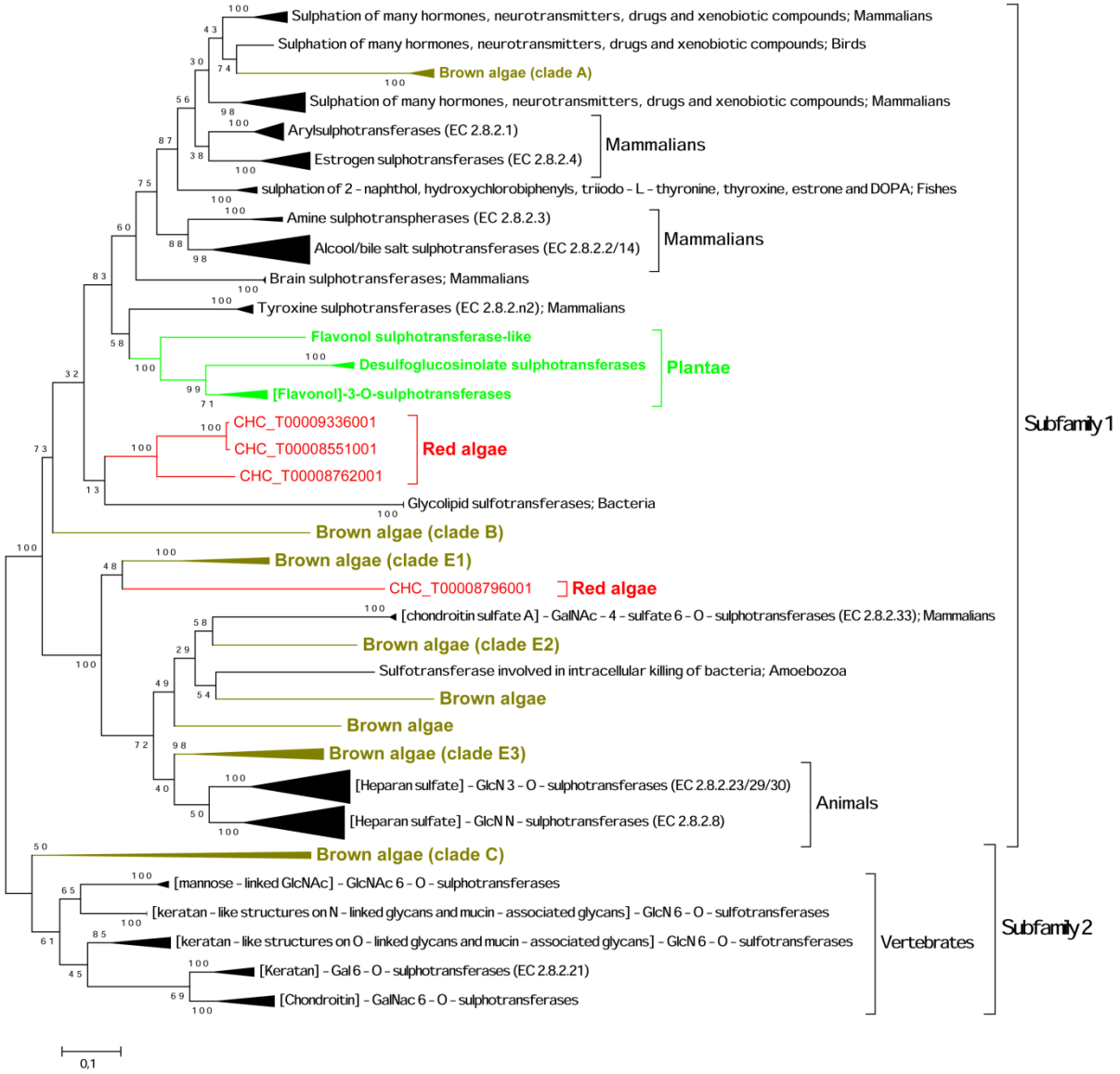

Fig. S7.4. Unrooted phylogenetic tree of family 1 of sulphotransferases. Clades of sequences of brown algae refer to *E. siliculosus* sulphotransferases*(193)*.

Family 2

Seven other genes from *C. crispus* encode sulphotransferases. They are related to the family 2 of sulphotransferases, composed so far only of carbohydrate sulphotransferases from vertebrates. These enzymes are involved in sulphation to position 3 of terminal glucuronic acid of both protein- and lipid-linked oligosaccharides (219), and in sulphation to position 4 of non-reducing GalNAc residues found in both N-glycans and O-glycans (220), of glycoprotein carbonic anhydrase VI (221), of chondroitin, of desulphated dermatan (222), and of dermatan sulphate (223). These seven proteins are good candidates to function as genuine carrageenan sulphotransferases (Fig. S7.5).

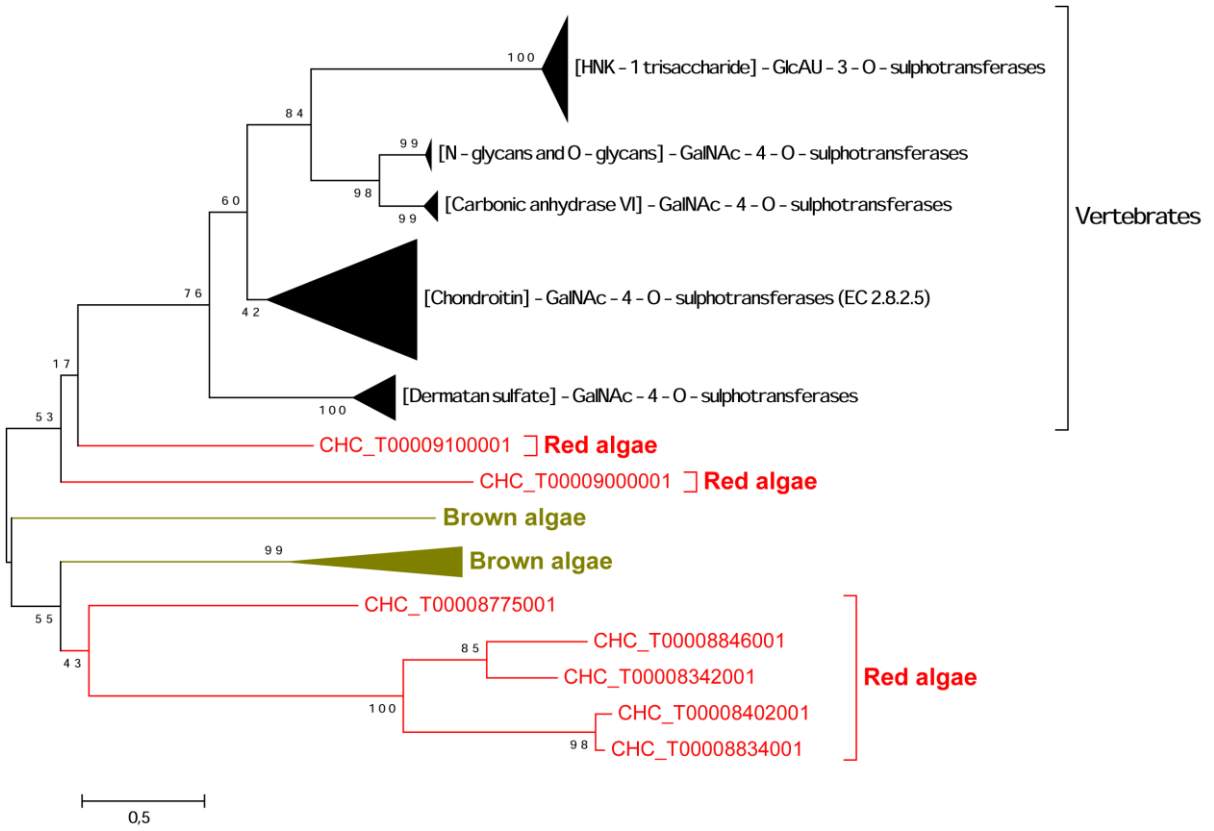

Fig. S7.5. Unrooted phylogenetic tree of family 2 of sulphotransferases. Except the *C. crispus* sequences CHC_T00009100001 and CHC_T00009000001, all sequences from brown and red algae are outgroups.

Galactose-3-O-sulphotransferase family

Finally a last gene encodes a putative sulphotransferase distantly related to the galactose-3-O-sulphotransferase family. This small family (eleven sequences in UniProt_SwissProt) contains carbohydrate sulphotransferases from mammalians transferring a sulphate group to the hydroxyl group at C3 of non-reducing β-galactosyl residues (224), and galactosylceramide sulfotransferases (that catalyze the sulphation of β-glycosides at the non-reducing termini of sugar chains attached to a lipid moiety in membrane glycolipids (225). Whereas seven sequences from *E. siliculosus* belong to this family (193), only one gene encoding a protein related to this family was found in *C. crispus*. Interestingly, as shown by phylogenetic analysis, all sequences from algae probably represent some new activities (Fig. S7.6).

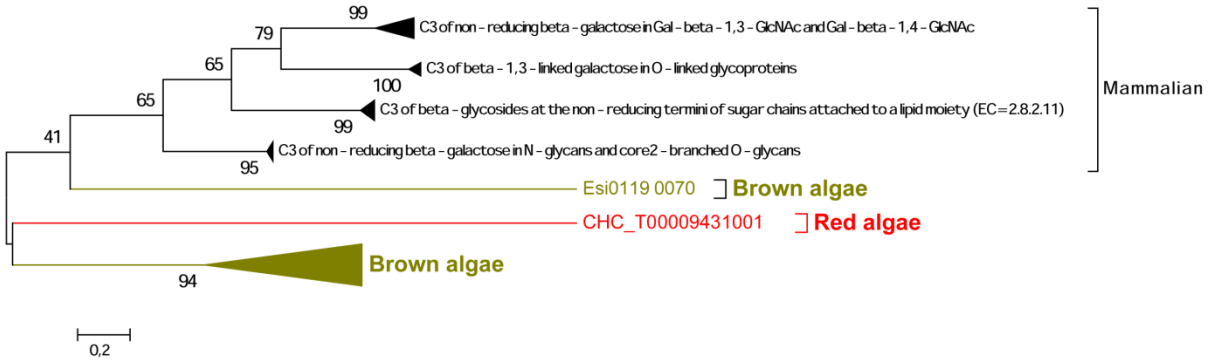

Fig. S7.6. Unrooted phylogenetic tree of the galactose-3-O-sulphotransferase family.

APS transporter

All sulphotransferases function with the PAPS as sulphate donor. As the sulphation reactions occur in Golgi apparatus, it is necessary to import the adenosine 5'-phosphosulphate (APS) in this organelle. The transport of the APS in the Golgi apparatus is carried out *via* the integral membrane protein adenosine 3'-phospho 5'-phosphosulfate transporter. Only one gene encoding a well conserved APS transporter was found (CHC_T00008958001).

7.10 Sulphurylases

The final step of biosynthesis of carrageenans is the formation of a 3,6-anhydro bridge of residue D-galactose, conferring the gelling properties of carrageenans. Two proteins, sulphurylase I and sulphurylase II, were purified from *C. crispus* (206). Both enzymes, called D-galactose-2,6-sulphurylases, convert the ν-carrageenan to ι-carrageenan by desulphation of carbon 6 in D-galactose and cyclisation of 3,6 anhydro-bridge in this residue. The kinetics experiments of conversion of ν-carrageenan to ι-carrageenan suggest a different mode of action of both enzymes on the galactan backbone.

The D-galactose-2,6-sulfurylase I (CHC_T00008516001) probably uses flavin as prosthetic group. Eleven genes coding for D-galactose-2,6-sulphurylase II were found. These genes represent a rare case of multigenic family found in the genome. At the sequence similarities level, seven proteins display between 70% and 99% of identity with the protein encoded by the gene CHC_T00009416001 taken as reference. Three proteins show only 67, 57 and 55% of identity with the reference protein. As expected by the sequences similarity levels, the phylogenetic analysis reveals a clade of three genes well separated from other sequences (Fig. S7.7). These three proteins could be good candidates for D-galactose-6-sulfurylases involved in the conversion of precursor μ-carrageenan to its mature polysaccharide κ-carrageenan.

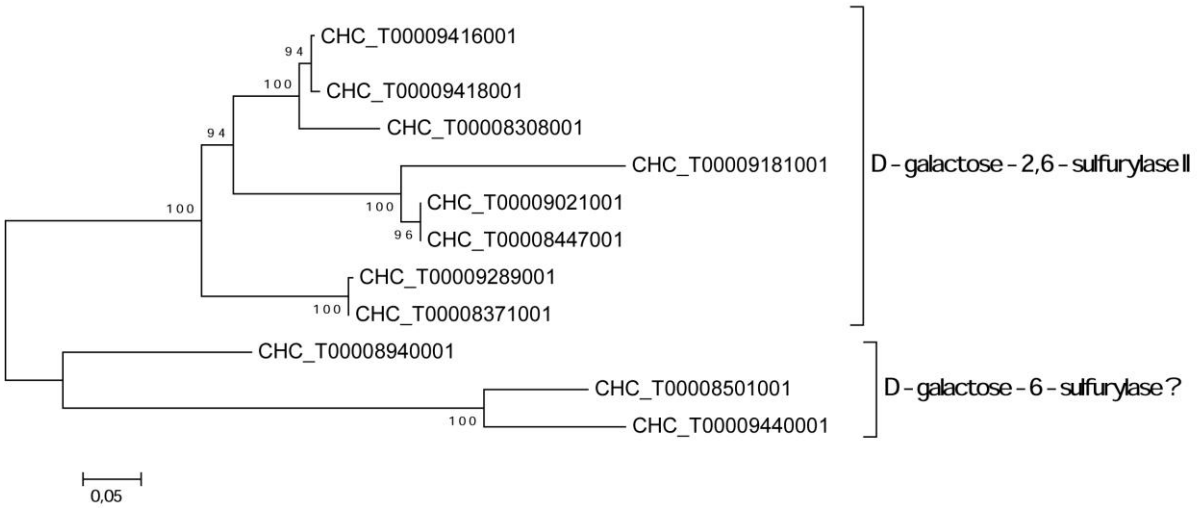

Fig. S7.7. Unrooted phylogenetic tree of type II-sulphurylases and the related group of potential sulphurylases in *C. crispus*.

7.11 Sulphatases

During the developmental cycle of *C. crispus* and under changes of environmental conditions, it is reasonable to assume that modifications of cell wall carrageenans, in particular their level of sulphation, are necessary and are potentially performed by sulphatases. There are four families of sulphatases. Family 1, groups the formylglycine-dependent sulphatases (226). Family 2 groups the Fe(II) α-ketoglutarate-dependent sulphatases (227). The families 3 and 4, that belong to the zinc-dependent β-lactamases superfamily, group alkylsulphatases (228) and aryl-sulphatases (229) respectively. Unexpectedly, no sulphatases belonging to any family were found in the *C. crispus* genome. Moreover, no gene encoding sulphatase-modifying factor 1 and 2 (C-α-formylglycine-generating enzyme 1), responsible of the post-translational modification of the sulphatases from the family 1, were found. This result suggests that no modification occur in carrageenan after its biosynthesis, or that sulphatases belonging to new families exist in the genome. In comparison, nine genes encoding for sulphatases belonging to the family 1 are present in *E. siliculosus* (193).

7.12 Starch

The *C. crispus* genome content with respect to starch metabolism is summarized in Table S7.14, and compared to the enzyme networks of *C. merolae, G. sulphuraria*, *O. tauri* and *A. thaliana*. Detailed explanations concerning the function of this enzyme network can be found in recent reviews (230) (231) (232) (233).

Starch accumulation in red algae and glaucophytes occurs solely in the cytosol while in the Chloroplastida storage granules are always plastidial; this location is thus taken as a distinctive feature of this lineage. This is illustrated by the occurrence, in both *O. tauri* and land plants, of enzymes related to the ones associated with storage polysaccharide synthesis in bacteria, such as ADP-glucose

pyro-phosphorylases and ADP-glucose utilizing starch synthases. The Chloroplastida starch synthases belong to the GT5 CAZY (carbohydrate active enzyme) family which typify bacterial glycogen metabolism, while other eukaryotes synthesize glycogen from UDP-glucose using either a GT3 (fungi and animals) or a GT5 (alveolates, parabasalids, amoebozoa) type of transferase. The *C. crispus* enzyme (CHC_G00009277001) groups, with strong support, to the GT5 UDP-glucose-specific enzyme of heterotrophic eukaryotes. This correlates with the absence of enzymes responsible for ADP-glucose synthesis (Table S7.14), and the reported biochemical purification of a UDP-glucose specific starch synthase in other Florideophycideae. The finding of a single elongation enzyme in *C. crispus* comes in stark contrast to all other documented studies of starch metabolism. Effectively, in green plants and algae, a minimum of four distinct soluble starch synthases are required to prime starch synthesis and to elongate chains of different length within amylopectin. Observation in *C. crispus* suggests that a single enzyme in red algae is sufficient to take over all of these distinctive functions. It is not known if such an enzyme would require a glycogenin type of priming mechanism. The single GT8-glycogenin-like sequence in Rhodophyceae seems fused to a GT64 domain which is related to enzymes transferring N-acetylglucosamine, and may not be related to starch or glycogen metabolism. The Florideophycideae subclass was initially reported to accumulate cytosolic starch with no amylose fraction, and the name floridean starch was coined to describe this situation (234) (232). *Chondrus crispus* tetrasporophyte and gametophyte starch clearly lack both the amylose fraction and the enzyme of amylose synthesis GBSS (granule-bound starch synthase). However, this is not a universal feature among Rhodophyceae since other Porphyridiales red algae are known to contain both of these. *Chondrus crispus* contains one copy of all other starch metabolism enzymes reported in Chloroplastida, including starch branching enzyme (CHC_G00008662001), starch phosphorylase (CHC_G00008988001), transglucosidase (dpe2-type of α-1,4 glucanotransferase, CHC_G00009120001), glucan-water dikinase (CHC_G00008789001), phosphoglucan-water dikinase (CHC_G00008992001), laforin-type of glucan-phosphatase (CHC_G00008421001), β-amylase (CHC_G00008493001), and pullulanase (CHC_G00008915001). In addition to these classical enzymes of starch metabolism, *C. crispus* seems to contain a novel type of debranching enzyme (CHC_G00008393001) found only in archaea and cyanobacteria, and that is related to the α-1,6 glucosidase domain of the bifunctional indirect debranching enzyme of eukaryotic glycogen metabolism (235). Nevertheless, as was documented for *C. merolae*, it contains two types of isoamylase-GlgX-type of direct debranching enzymes (CHC_G00009540001, CHC_G00008313001). Such enzymes are known to be involved in both starch biosynthesis and degradation in Chloroplastida, where 3 distinct isoforms are reported. Two of these, isa1 and isa2, function during biosynthesis, to splice out those misplaced branches introduced by branching enzyme, and that slow down polysaccharide aggregation into semi-crystalline insoluble granules. The third isoamylase plays an important role in debranching those α-1,6 branches left over after degradation mediated by β-amylase (232). Our phylogenetic trees are consistent with the

maintenance of this important dual function in starch metabolism, although the precise role taken by each isoform needs to be ascertained experimentally. *Chondrus crispus* displays the full suite of enzymes involved in phospho-starch metabolism (233). These include two distinct types of glucan water dikinases that are phylogenetically related to the enzymes responsible for introducing phosphates on C6 (GWD) and C3 (PWD) of crystalline amylopectin. This local phosphorylation of amylopectin, which loosens its crystalline packing, is required to facilitate attack by starch catabolism enzymes which are otherwise unable to degrade solid starch*(233)*. Phospho-starch metabolism also includes a phosphatase thought to be required to hydrolyse the phosphates left over after glucan degradation. This laforin-like phosphatase is related both to laforin and to the *A. thaliana* sex4-protein which have been demonstrated to be active in starch catabolism. The GWD-PWD enzymes are distinctive of the starch pathway in Archaeplastida and some of their secondary endosymbiosis derivatives. It is striking to note that *Galdieria*, which is known to accumulate glycogen rather than starch, does not display such sequences.

Chondrus crispus displays a very much reduced diversity of α-1,4 glucanotransferases, and only one dpe2 transglucosidase sequence has been recovered through our detailed bioinformatic analysis. Such transglucosidases have been reported to be highly selective for the β-maltose isomer as a donor, and somewhat less selective for the nature of the acceptor. Transglucosidase is coupled tọ β-amylase-mediated degradation in the cytosol since it generates its substrate. A growing body of evidence points to a cytosolic heteroglycan as a major acceptor of glucose in the monosaccharide transfer reaction catalyzed by transglucosidase. Malto-oligosaccharides of DP3 (degree of polymerization 3), or longer, are certainly not donors in such transfer reactions and may also define poor acceptors. The *C. crispus* cytosol lacks enzymes able to metabolize maltotriose and maltotetraose, while maltopentaose and longer malto-oligosacccharide series may be broken down to maltotetraose with concomitant release of glucose-1-P by the cytosolic starch phosphorylase. In Chloroplastida, metabolism of maltotriose and maltotetraose occurs in the plastid through disproportionating enzyme also known as dpe1, another α-1,4 glucanotransferase that can use maltotriose and longer oligo-saccharides both as donors and as acceptors, and leading to maltopentaose and longer maltooligosaccharides that define suitable substrates for the phosphorylases. Comparable α-1,4-glucanotransferases are clearly lacking in Rhodophyceae. The presence of two transglucosidases in the *C. merolae* cytosol may reflect the requirement for at least one isoform to allow for a more efficient use of maltotriose and maltotetraose as potential acceptors of the transglucosylation reaction catalyzed from β-maltose. However, the single enzyme in *C. crispus* may also display such properties. Thus the ability to process maltotriose and maltotetraose needs to be ascertained experimentally on the corresponding recombinant transglucosidase enzymes.

To summarize, *C. crispus* is likely to both prime starch synthesis and elongate chains from UDP-glucose through the action of a single GT5 type of soluble starch synthase. The chains would be

branched by the single branching enzyme and the misplaced branches would be trimmed out by one of the two isoamylase isoforms leading to polysaccharide aggregation into huge insoluble floridean starch granules. The maltooligosaccharides released through action of the trimming isoamylase would be recessed to maltotetraose through the action of starch phosphorylase with concomitant release of glucose-1-P. Maltotetraose and maltotriose could be metabolized either as acceptors during a transfer reaction involving maltose as donor through the action of transglucosidase, or through a yet to be defined mechanism.

Starch degradation would be initiated through C6 phosphorylation of selective glucose residues by GWD that would thereafter be subjected to C3 phosphorylation by PWD. The crystalline amylopectin disrupted through phosphorylation would then be attacked by β-amylases and debranching enzymes such as a second isoamylase isoform, pullulanase, or the novel archaeal type of debranching enzyme found in *C. crispus*. The phosphorylated glucans would be recycled through the action of the laforin-like phosphatase. This hydrolytic attack of the starch granule would yield both maltose and longer maltooligosaccharides. Maltose would be selectively metabolized by the trans-glucosidase, while the longer maltooligosaccharides would be metabolized through the action of both cytosolic phosphorylase and transglucosidase, or through a yet to be defined enzyme. In all cases, the suspected activities should be checked experimentally on recombinant enzymes.

Chondrus crispus presently defines the first genome of a red alga that accumulates *bona fide* starch granules. Cyanidiales that thrive in hot environments are expected to live at temperatures where some starch structures may swell and melt. Not surprisingly, *G. sulphuraria* has been reported to accumulate glycogen-like structures, while *C. merolae* was described as accumulating starch-like polysaccharides somewhat similar to the semi-amylopectin type described in unicellular diazotrophic cyanobacteria. Interestingly, phosphoglucan metabolism is entirely absent from *G. sulphuraria*, and is restricted to the sole presence of GWD and laforin in *C. merolae* in which starch-like structure should therefore lack phosphates on the C3 position.

The finding in *C. crispus* of mere twelve genes required to produce and mobilize starch revolutionizes our understanding of the building of this important polymer. Indeed, it was assumed, until very recently, that the complexity of starch metabolism in the green lineage reflected the complexity of the structure of starch. The *C. crispus* genome clearly invalidates this opinion.

Table S7.14. Starch metabolism in *C. crispus*.

Enzyme	Viridiplantae		Rhodophyta		
	A. thaliana	*O. tauri*	*C. merolae*	*G. sulphuraria*	*C. crispus*
ADP-glucose pyrophosphorylase	6	3	-	-	-
starch synthase (ADPG)	5	5	-	-	-
starch synthase (UDPG)	-	-	1	1	1
GBSS I	1	1	-	-	-
branching enzymes	3	2	1	1	1
isoamylases	3	3	2	2	2
pullulanase	1	1	1	0	1
archaeal DBE	-	-	-	-	1
phosphorylases	2	3	1	2	1
glucanotransferase (dpe1)	1	1	-	-	-
trans-glucosidase (dpe2)	1	1	2	1	1
β-amylase	3	2	1	3	1
α-amylase	3	3	1	3	-
glucan water dikinase	2	3	1	0	1
phosphoglucan water dikinase	1	2	-	-	1
laforin (sex4)	1	1	1	-	1
Total	33	31	12	13	12

7.13 Mannitol metabolism

Red algae feature a high diversity in terms of production of low molecular weight compounds, including heterosides (floridoside, D-isofloridoside, L-isofloridoside, digeneaside, digalactosyl-glycerol), and sugar alcohols (mannitol, sorbitol, dulcitol). Low molecular weight compounds have important biological roles, such as primary photosynthetic products and involvement in osmo-acclimation. As an example, mannitol metabolism has been described in mangrove macroalgae of the genus *Caloglossa* (Florideophyceae) (236) (237) (238), and in the unicellular *Dixoniella grisea* (Rhodellophyceae) (239) (240).

Genes potentially involved in the mannitol cycle were searched for using the genes identified in the genome of the brown alga *E. siliculosus* (121) (193) (241). Mannitol-1-phosphate dehydrogenases (M1PDHs) were not found. In contrast, several sequences similar to mannitol-1-

phosphatases (M1Pases) were identified, and belong to the haloacid dehalogenase-like hydrolase family whose members can carry out reactions on various types of substrates. Therefore, considering the absence of candidates for the first enzyme of the mannitol cycle (M1PDH), the sequences of M1Pases identified in the *C. crispus* genome may not encode enzymes involved in this cycle, but in other metabolic pathways. Concerning mannitol-2-dehydrogenase (M2DH), one good candidate was found. For the hexokinase (HK), two possible orthologs were identified, but it is difficult to infer specificity of kinases merely based on analysis of amino acid sequences.

The conclusion is that *C. crispus* do not feature the full set of enzymes related to the mannitol cycle. This observation may be interesting in a more general discussion on evolution within the red algae, with most of the species not featuring this cycle, while a few of them contain the corresponding enzymes and use mannitol for osmo-regulation (*Caloglossa*, *Dixoniella*).

7.14 Lignin biosynthesis

Although lignin is a complex phenolic polymer typically associated with vascular plants and land colonisation, its presence was recently unequivocally reported in cells of the calcareous red alga *Calliarthron cheilosporioides* (242). This finding significantly questioned the origin and early evolutionary history of the lignin biosynthetic pathway. By a systematic comparative genomic survey of green plants and algae it has been shown that the complete lignin pathway is likely to have first appeared in moss, followed by a further expansion of gene families after the divergence of monocotyledons and dicotyledons (243). A more ancient origin is not excluded, but in such scenario the pathway would have been subsequently lost at least in green algae and others but not all plant lineages (191) (243) (244). For the red alga *C. cheilosporioides*, the independent occurrence of lignin by convergent evolution was favoured (242), which would imply the concomitant innovation of all the pathways involved in this complex metabolic system (245). Without a whole red algal genome sequence available, both assumptions stayed highly speculative. Although true lignin has only been identified in *C. cheilosporioides* within the red algae so far, the genomic analyses of *C. crispus* can shed light on the lignin evolutionary history between rhodophytes and green plants. The monolignol biosynthesis pathway is highly conserved in vascular plants, with ten enzymes involved in the conversion of phenylalanine into different monolignol molecules (Table S7.15). For most enzymes, no homologues are found in the *C. crispus* genome, and when present, they are only distantly related to their plant counterparts, suggesting distinct functional assignments. The phenylpropanoid pathway constitutes the entry point of the lignin route in land plants and, the PAL, C4H and 4CL enzymes are absent in *C. crispus*. Although the *C. crispus* genome encodes 20 cytochrome P450 monooxygenases, none of them are related to C3H and C4H (respectively from CYP98 and CYP73 families) which are believed to be the most conserved gene families involved in the monolignol synthesis in plants (243). Among the medium-length dehydrogenase/reductase (MDR) superfamily (246), only one gene in *C. crispus* is related to the CAD (Table S7.15), exhibiting 35% identity with its *A. thaliana* counterparts.

One gene in *C. crispus* can account for a potential CCR, but this sequence shows greater similarity to dihydroflavonol reductases within this family (247), an enzyme leading to flavonoid end-products in plants. CCoAOMT and COMT are both responsible for the methylation of the monolignol precursors. Seven genes in *C. crispus* are more closely related to CCoAOMT than to any other O-methyltransferases, but again sequence identities with the plant members is low, *i.e.* around 30% at best. No ferulate-5-hydroxylase (F5H) homologue is found in *C. crispus*, which is not surprising as the enzyme involved in lignin biosynthesis is thought to be an independent innovation of lycophytes and angiosperms (243). Lignification is the final key step in the lignin deposition in plants, and represents the reaction by which the monomers are cross-coupled to the growing polymer. The class III peroxidase enzymes, along with the laccases, are the dominant form of lignin polymerization enzymes in plants. No homologues are found in the *C. crispus* genome.

Table S7.15. Core enzymes involved in monolignol biosynthesis and lignifications.

Enzyme	Enzyme name	Genes
PAL	phenylalanine ammonia-lyase	0
C4H	cinnamate-4-hydroxylase	0
4CL	4-coumarate-CoA ligase	0
HCT	hydroxycinnamoyl transferase	0
C3H	4-coumarate 3-hydroxylase	0
CCR	cinnamoyl-CoA reductase	1
CAD	cinnamyl-alcohol dehydrogenase	1
CCoAOMT	cinnamoyl CoA O-methyltransferase	7
COMT	caffeic acid O-methyltransferase	0
F5H	ferulate 5-hydroxylase	0
-	class III peroxidase	0
-	laccase	0

In conclusion, these results show that *C. crispus* lack most of conserved genes involved in lignin biosynthesis. This clearly impairs a putative orthology of the lignin biosynthetic pathway between land plants and red algae. A screening of the partial sequence genome of *C. cheilosporioides* also failed to retrieve any closely related orthologues of the pathway (http://dbdata.rutgers.edu/data/plantae/).

7.15 General secondary metabolism

Several classes of secondary metabolites have been described in red algae, such as isoprenoids, terpenoids, acetogenins, and in majority from the genus *Laurencia* (248) but, so far, phenylpropanoid

and flavonoid compounds have not been reported. In *C. crispus*, the knowledge on the composition of secondary metabolites is limited and mainly restricted to oxylipins (188) (189) (see paragraph 7.6). In plants and microorganisms, the chorismate pathway can lead to aromatic amino acids, folic acid, phylloquinone, and possibly indole alkaloids. The analysis of the *C. crispus* genome revealed that most enzymes of the shikimate and chorismate utilizing pathways are present (Tables S7.15 & S7.16), but none of the genes for the production of benzenoids, salicylate, or dihydroxybenzoic acid found in land plants were identified in the *C. crispus* genome. In addition, other key genes known in land plants and encoding phenylalanine ammonia lyase (PAL) and type III polyketide synthase (PKS), respectively involved in the phenylpropanoid and flavonoid pathways, are absent (Tables S7.17 & S7.18). This result is consistent with a phylogenetic analysis of PAL showing that a recent horizontal gene transfer from bacteria allowed the emergence of the phenylpropanoid metabolism in land plants (249).

Most of the typical genes of the phenylpropanoid and flavonoid pathways are absent in *C. crispus* apart some exceptions. For example, homolog to cinnamoyl-alcohol dehydrogenase or chalcone isomerase-like are found in the *C. crispus* genome, and an explanation would be that these enzymes have been recruited in phenylpropanoid and flavonoid pathways in land plants whereas they were initially involved in another metabolic function in the common ancestor with red algae.

While PKS genes have been found in brown algae*(121)*, there is no evidence of corresponding enzymes in red algae, even in Corallinales like *C. cheilosporioides* (34). Alternatively, type I and type II multi-domain PKS enzymes have not been detected, suggesting that acetogenins, if they exist in *C. crispus,* would be preferably produced by the action of fatty acid synthases (248). Phenolic substances are rare in red algae and less diverse and less complex than in brown algae. However, the existence of brominated phenols (250) in the former organisms suggest that an alternate pathway exists. UV-absorbing molecules are specifically mobilized in response to pathogenic attack of *C. crispus*, and it was suggested that these compounds are phenolics (188). The molecular bases for the production of phenols in red algae remains to be discovered (Table S7.19) but it could be related to an alternative pathway starting from tyrosine, as proposed for the green alga *Ulva lactuca* (251). In addition, other types of aromatic compounds have been documented in red algae, such as indole alkaloid derivatives (252), but their biosynthetic pathway is unknown. The last class of secondary metabolites investigated in this study was the mycosporine-like amino acids (MAAs). MAAs are small molecules known to be implicated in UV protection in aquatic organisms (253). The MAA shinorine has been purified from *C. crispus*, and probably other MAAs can be synthesized by this red alga (254). Recently, biosynthesis pathway of MAAs was elucidated in cyanobacteria (255), and all the corresponding genes have putative homologues in *C. crispus* (Table S7.20).

Table S7.15. Shikimate pathway genes.

EC number	Enzyme name	Gene ID
2.5.1.54	3-deoxy-D-arabino-heptulosonate 7-phosphate synthase	CHC_T00012752001
		CHC_T00002804001
		CHC_T00007516001
4.2.3.4	3-dehydroquinate synthase	CHC_T00010757001
1.1.99.25	pyrroloquinoline-quinone dehydrogenase	Not found
1.1.1.24	quinate dehydrogenase	Not found
1.1.1.282	quinate/shikimate dehydrogenase	Not found
4.2.1.10	bifunctional 3-dehydroquinate dehydratase/	CHC_T00013277001
1.1.1.25	shikimate dehydrogenase	
2.7.1.71	shikimate kinase	EST support only
2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase	CHC_T00014266001
4.2.3.5	chorismate synthase	CHC_T00014540001

Table S7.16. Chorismate utilizing pathway genes.

EC number	Enzyme name	Genes
4.1.3.27	anthranilate synthase (Component I)	CHC_T00005725001
	anthranilate synthase (Component II)	(scaffold_613)
5.4.99.5	chorismate mutase	CHC_T00007046001
5.4.4.2/	bifunctional isochorismate synthase /	CHC_T00004414001
2.2.1.9	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase	
	isochorismate pyruvate lyase	Not found
4.3.1.-	salicylate synthase	Not found
2.6.1.85	para-aminobenzoate synthase	CHC_T00012525001
4.1.3.40	chorismate pyruvate lyase	Not found
	aminodeoxychorismate lyase	Not found
1.14.13.-	benzoic acid 2-hydroxylase	Not found

Table S7.17. The phenylpropanoid pathway genes.

EC number	Enzyme name	Gene ID
4.3.1.24	phenylalanine ammonia lyase	Not found
4.3.1.23	tyrosine ammonia lyase	Not found
4.3.1.25	phenylalanine/tyrosine ammonia-lyase	Not found
1.14.13.11	cinnamate 4-hydroxylase	Not found

6.2.1.12	4-coumarate-CoA ligase	Not found
1.14.13.14	cinnamate 2-hydroxylase	Not found
1.2.1.44	cinnamoyl-CoA reductase	CHC_T00004103001
1.1.1.195	cinnamoyl-alcohol dehydrogenase	CHC_T00005929001
1.1.1.194	coniferyl-alcohol dehydrogenase	Not found
2.3.1.91	sinapine synthase	Not found
2.3.1.92	sinapoylglucose-malate O-sinapoyltransferase	Not found
2.3.1.99	quinate O-hydroxycinnamoyltransferase	Not found
2.3.1.115	isoflavone-7-O-β-glucoside 6''-O-malonyltransferase	Not found
1.13.11.22	caffeate 3,4-dioxygenase	Not found
2.4.1.128	scopoletin glucosyltransferase	Not found
2.4.1.120	sinapate 1-glucosyltransferase	Not found
2.4.1.114	2-coumarate O-β-glucosyltransferase	Not found
2.4.1.111	coniferyl-alcohol glucosyltransferase	Not found
3.1.1.49	sinapine esterase	Not found
3.2.1.126	coniferin β-glucosidase	Not found

Table S7.18. The flavonoid pathway genes.

EC number	Enzyme name	Gene ID
2.3.1.-.	Type III polyketide synthase	Not found
5.5.1.6	chalcone isomerase	CHC_T00005167001
2.3.1.170	chalcone reductase	Not found
1.14.13.-	isoflavone synthase	Not found
1.3.1.45	isoflavone reductase	Not found
1.3.1.77	anthocyanidin reductase	Not found
1.14.11.22	flavone synthase I	Not found
1.14.11.23	flavonol synthase	Not found
1.14.11.9	flavanone 3-dioxygenase	Not found
1.14.11.19	leucocyanidin dioxygenase	Not found
1.14.11.-	2OG-Fe(II) oxygenase superfamily	GIDCcT00001700001
		GIDCcT00017193001
1.1.1.219	dihydroflavanol 4-reductase	Not found
1.1.1.234	flavanone 4-reductase	CHC_T00004103001
1.17.1.3	leucoanthocyanidin reductase	Not found
2.8.2.25	flavonol 3-sulfotransferase	Not found
2.8.2.27	flavonol 4'-sulfotransferase	Not found

2.1.1.6	catechol O-methyltransferase	Not found
2.1.1.68	caffeate O-methyltransferase	GIDCcT00011942001
2.1.1.75	flavonoid O-methyltransferase	CHC_T00014019001
2.1.1.76	flavonol 3-O-methyltransferase	Not found
2.1.1.104	caffeoyl-CoA O-methyltransferase (putative bifunctional sulfurtransferase/O-methyltransferase family 3)	CHC_T00008286001
2.4.1.91	flavonol 3-O-glucosyltransferase	Not found
2.4.1.115	anthocyanidin 3-O-glucosyltransferase	Not found
2.4.1.159	flavonol-3-O-glucoside L-rhamnosyltransferase	Not found
2.4.1.81	flavone 7-O-β-glucosyltransferase	Not found
2.3.1.133	shikimate O-hydroxycinnamoyltransferase	Not found
1.21.3.6	sureusidin synthase	Not found

Table S7.19. Phenols and other aromatic derivatives related enzymes.

EC number	Enzyme name	Gene ID
2.6.1.5	L-tyrosine aminotransferase	CHC_T00004847001
1.1.1.26	glyoxylate/hydroxyphenylpyruvate reductase	CHC_T00012832001
1.2.3.13	4-Hydroxyphenylpyruvate oxidase	CHC_T00008423001
1.11.1.-	vanadium haloperoxidase	See Sup. Table 8.1.
	2,6-dihydroxypyridine-3-hydroxylase	CHC2_T00011355001
1.14.13.1	salicylate 1-hydroxylase	Not found
1.14.13.2	4-hydroxybenzoate 3-monooxygenase	Not found
1.14.13.3	4-hydroxyphenylacetate 3-monooxygenase	Not found
1.14.13.6	orcinol hydroxylase	Not found
1.14.13.7	phenol 2-monooxygenase	Not found
1.14.13.9	kynurenine 3-hydroxylase	CHC_T00011483001
1.14.13.12	cytochrome P450 family protein (similar to benzoate 4-monooxygenase)	CHC_T00004253001
1.14.13.16	cyclopentanone monooxygenase	Not found
1.14.13.18	4-hydroxyphenylacetate 1-monooxygenase	Not found
1.14.13.20	4-hydroxyphenylacetate 1-monooxygenase	Not found
1.14.13.22	cyclohexanone monooxygenase	Not found
1.14.18.1/	polyphenol oxidase family	SNAPCcT00025313001
1.10.3.1		CHC_T00009477001
1.11.1.11	ascorbate peroxidase	See Sup. Table 8.3.
1.11.1.14	lignin peroxidase	Not found

1.11.1.7	guaiacol peroxidase	Not found
	horseradish peroxidase	Not found
	extensin peroxidase	Not found
	catalase-peroxidase	Not found

Table S7.20. Mycosporine-like amino acids pathway genes.

EC number	Enzyme	Gene ID
4.2.3.-/2.1.1.-	bifunctional 3-dehydroquinate synthase-like / O-methyltransferase	CHC_T00003360001
4.2.3.-	3-dehydroquinate synthase-like	CHC_T00004575001
2.1.1.-	O-methyltransferase	CHC_T00007036001
		CHC_T00001993001
2.8.2.-/2.1.1.-	sulfurtransferase/O-methyltransferase	CHC_T00008286001
6.3.2.4	ATP-dependent carboxylate-amine ligase/ D-Ala-D-Ala ligase	GIDCcT00003179001
	non-ribosomal peptide synthetase	CHC2_T00003711001
		CHC_T00003145001
		CHC2_T00014030001
		SNAPCcT00038662001

Table S7.21. The phylloquinone (vitamin K1) pathway genes.

EC number	Enzyme name	Gene ID
5.4.4.2/ 2.2.1.9	bifunctional isochorismate synthase/ 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase	CHC_T00004414001
6.2.1.26/ 4.2.1.113	bifunctional 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylic acid synthase/ o-succinylbenzoic acid synthase	CHC_T00004398001
2.1.1.163	ubiquinone/menaquinone biosynthesis methyltransferase	CHC_T00009844001
		CHC_T00000396001
6.2.1.26	o-succinylbenzoate-CoA ligase	CHC_T00014931001
4.1.3.36	1,4-dihydroxy-2-naphthoyl-CoA synthase	CHC_T00004408001
2.5.1.74	1,4-dihydroxy-2-naphthoate polyprenyltransferase	CHC_T00005613001

7.16 Glutathione S-transferases

The GSTs (glutathione S-transferases) have emerged independently several times throughout evolutionary history, producing distinct GST families*(256)*. These enzymes have been found in

virtually all types of organisms and are mainly involved in the detoxification of hydrophobic substrates. They separate into clades: cytosolic, mitochondrial (the kappa-class), and microsomal (MAPEG). Beyond detoxification and involvement in various defence responses, there is increasing evidence indicating that GSTs may be implicated in many other physiological processes, including isomerisation of specific metabolites (257), transportation of endogenous substrates (258), cell growth, and development (259).

Table S7.22. Genes encoding cytosolic and microsomal (MAPEG) glutathione S-transferases (GSTs) in the *C. crispus* genome**.**

Name	Short name	Gene ID	Tentative class
Cytosolic GST	CcGST1	CHC_T00009444001	Sigma class
Cytosolic GST	CcGST2	CHC_T00008810001	Sigma class
Cytosolic GST		CHC_T00008274001	Sigma class
Cytosolic GST		CHC_T00008425001	Zeta class
Cytosolic GST		CHC_T00008766001	Beta class
Cytosolic GST	CcGST3	CHC_T00008724001	Beta class
Cytosolic GST		CHC_T00008791001	New and algal specific
MAPEG		CHC_T00009312001	-
MAPEG		CHC_T00008331001	-

Cytosolic GST

The cytosolic GSTs have been assigned to eight widely studied classes (Mu, Pi, Alpha, Sigma, Phi, Theta, Zeta and Omega) but others have been more recently recognized (256). Previous phylogenetic analyses on cytosolic GST genes from *C. crispus* showed that some members were describing a new class closely related to the Sigma class found in animals (260). Seven genes encoding cytosolic GSTs can actually be identified in the genome (Table S7.22): one is closely related to the Zeta class (occurring in a large variety of organisms), three are closely related to the Sigma class (specifically found in animals), two to the Beta class (specifically found in bacteria), and one belongs to a new algal-specific class (Table S7.22) sharing a putative common ancestor with both Omega, Lambda and Tau classes (the Lambda and Tau classes is plant-specific; the Omega class occurs preferentially in animals, although some members have also been found in plants (261)). An additional sequence (CHC_T00009287001) is found within a bacterial cluster closed to the Beta class, but its assignation to *C. crispus* will have to be fully confirmed as it is found isolated in the genome sequence, on a scaffold-terminus, with no EST to support the gene, and with a strong protein identity to bacterial sequences (~55%).

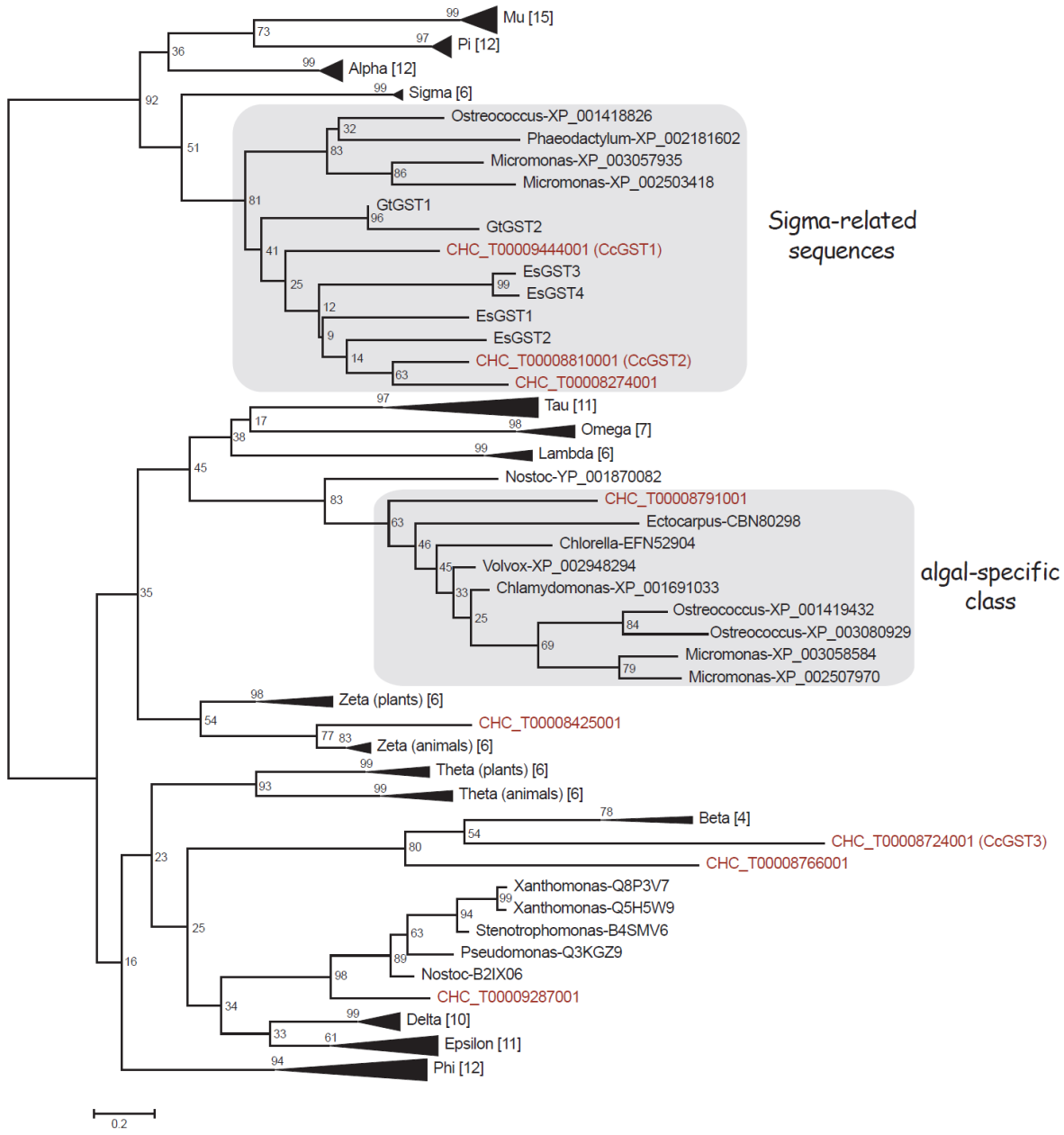

Fig. S7.8. Phylogenetic tree showing the relationships between GSTs from *C. crispus* and the major GST classes. The multiple-protein alignment was performed using the MAFFT program and analysed using the neighbour-joining method, based on distances derived from the PAM matrix, with the MEGA5.05 program. Clustering of proteins into GST classes is indicated. Numbers in square brackets refer to the number of sequences included for each class that have been collapsed in the Figure. The numbers on the branches indicate the percentages of node support from 1,000 bootstrap replicates. Es, *E. siliculosus*; Gt, *Gracilaria tenuistipitata*.

These new sequences provide clues to the evolution of GST classes. A total of 162 GSTs, including the *C. crispus* sequences, were integrated into a multiple-protein alignment to generate a phylogenetic tree using a distance method (Fig. S7.8). The results obtained confirm that the Zeta class have an

ancestral origin that predates the animal and plant split. With the identification of Sigma-related GSTs in red algae (260) and heterokonts (262), it was already suggested that the Sigma-class diverged before the Alpha, Mu, and Pi group, an hypothesis previously brought forward(263), and in contrast with the past dogma which proposed that the Alpha class enzymes were the oldest members of this clade (256). The present results strengthen the hypothesis of the Sigma-class being ancestral, as novel sequences have also been identified through database screening in other eukaryotic lineages (i.e. green algae), whereas the Alpha, Pi and Mu classes show a more restricted occurrence (animals only). In addition, the short branch lengths of these latter classes of GSTs support their recent evolutionary diversification. The two GST genes in *C. crispus* close to the Beta class have probably arisen and diverged from a horizontal gene transfer with a bacterium, together with the gene for which *C. crispus* inference will have to be confirmed (CHC_T00009287001). Finally, the last gene belongs to a newly identified cluster which contained only algal sequences rooted by a cyanobacterial isoform (Fig. S7.7). This group contains mainly green algal GSTs (with 40-50% identity between them), but also one gene from the brown alga *E. siliculosus* not previously assigned to a class. The red and green algal isoforms probably have a cyanobacterial origin, and developed after the primary endosymbiotic event which gave rise to the green and red lineages. The brown alga sequence would have been gained during the secondary endosymbiotic event with the red alga. The relationships of this algal-specific class with the Lambda, Omega and Tau classes are not clear yet. There is a consensus suggesting that Lambda and Omega GSTs are ancient cytosolic GST members (263). They both feature a cysteine residue as an essential active site residue, while the Tau-class, like many other classes, has a critical serine residue instead of the cysteine. Multiple protein-alignments also identified such a conserved serine residue in all algal-specific sequences. If all these GST classes may share a common cytosolic GST ancestor, their exact order of appearance cannot be inferred from these analyses only. A more dedicated phylogenetic investigation, with strongly supported nodes, will have to be pursued in order to draw solid conclusions on this specific aspect of GST evolution. All together, these results show a dual origin of the cytosolic GST sequences in *C. crispus*, with some isoforms inherited from the ancestral eukaryote and symbiont, and others acquired by horizontal gene transfer from bacteria.

Mitochondrial GSTs

The mitochondrial enzymes (kappa-class) share strong similarities with a family of prokaryotic 2-hydroxychromene-2-carboxylic acid isomerases (HCCI) involved in the catabolism of dibenzyl compounds (263) (256). The complete lack of sequence overlap between mitochondrial glutathione transferase/HCCI and the cytosolic GSTs strongly suggests an independent evolutionary origin. The evolutionary history of this mitochondrial-kappa family is considerably less complex than that of the cytosolic GSTs, although the mitochondrial family is also supposed to be very ancient, with homologs in bacteria and eukaryotes. Current annotations suggest that most vertebrate genomes contain only one copy of the GST kappa gene (264). In *E. siliculosus*, no mitochondrial GST isoform was found, while

one kappa member has been identified (262). Some counterparts have also been identified in nematodes, insects and more scarcely in land plants and fungi (264). Considering that kappa-enzyme homologues occur in multiple independent phylogenetic lineages, even if not in all, the most parsimonious evolutionary scenario would suggest an ancient origin for this class of enzymes, with numerous subsequent losses of the gene in distinct lineages, including red algae.

Microsomal GSTs

The microsomal GSTs (MAPEG) differ in size and structure from both the cytosolic and the mitochondrial GST families. Based on multiple sequence alignments, the MAPEG family can be further divided into four subgroups containing various proteins (263) (265). The first group includes 5-lipoxygenase-activating proteins (FLAP), leukotriene C4 (LTC4) synthases, and microsomal glutathione S-transferases 2 (MGST2). The plant and fungal members form a second subgroup with MGST3 sequences, while the bacterial proteins form a group of their own, and the MGST1 and prostaglandin E synthase 1 (PGSE1) proteins constitute a fourth subgroup. Most MAPEG proteins are involved in the synthesis of eicosanoids, leukotrienes and prostaglandins, others having different functions and catalytic activities more typical of cytosolic GSTs (263). Animal and bacterial species often possess more than one MAPEG member, either of the same or different subgroups. In contrast, plants and fungi express only MAPEG proteins of the MGST3-type. The genome possesses two MAPEG isoforms that both belong to the MGST3 subgroup. One of these proteins (mGST1) is closely related to the three brown algal isoforms from *E. siliculosus* (262). Phylogenetic analyses show that the subgroups containing the MGST2 and MGST3 homologues share a common ancestor (265). Considering that MGST3 isoforms are present in various phylogenetic distant taxa, it is very likely that the ancestral isoform was of an MGST3-type form from which the MGST2 members emerged in animals. In *E. siliculosus*, the ancestral gene encoding the microsomal GSTs was probably gained from the red algal endosymbiont.

8 Defence-related and stress genes

8.1 Candidate pathogen receptors or resistance genes.

In all species investigated, the onset of immune reactions relies on successful pathogen recognition, followed by signal transduction, and the induction of defence effectors. Specific pathogen recognition processes have previously been reported in *C. crispus* (266), but their molecular mechanisms have yet to be identified. On the other hand, pathogen recognition has been extensively studied in land plants, to which Rhodophyta are phylogenetically related. In plants, pathogen recognition is mediated by resistance genes and pattern recognition receptors, some if which directly recognize microbe-derived elicitors (267). Others monitor the integrity of endogenous proteins targeted by pathogens (268). These genes belong to families of up to several hundred members, which contain a leucine rich repeat

(LRR) domain, coupled to various domains, notably NB-ARC, coiled-coil (CC), and toll-interleukin 1R (TIR), the latter being shared with the pattern recognition receptors of vertebrates. TIR-NBS-LRR genes are also present in the moss *P. patens*, although their potential role in defence remains to be investigated (269). LRR domains are also prominent in animal proteins related to immunity, such as the well-characterised TOLL-like receptors (270) and the CATERPILLER family (271). A direct role for LRR proteins in antigen recognition has also been recently uncovered or hypothesized in jawless fishes and mosquito (272) (273).

Because of their involvement in pathogen recognition in both animal and plant systems, we set out to annotate the genes encoding TIR, NB-ARC, and LRR-containing proteins in the *C. crispus* genome. In stark contrast to land plants, *C. crispus* only contains eleven LRR genes (one of them containing an F-box, Fig. S8.1). Importantly, neither TIR-NBS-LRR nor CC-NB-ARC-LRR homologue of plant resistance genes was identified. On the other hand, 45 Sel1-repeats containing genes (one of them fused with TIR domain), 26 TPR -containing genes (two of them fused with the NB-ARC domain), as well as 85 WD40-repeats containing genes (45 of them fused to TIR and/or NB-ARC domains, eight fused to other functional domains) were found. At least twelve of the 45 NB-ARC WD40-repeat fusions contain an additional TIR domain. Considering that WD40 and LRR repeats are both typically involved in specific and highly variable ligand-binding*(274)*, we hypothesize that *C. crispus* (TIR-) NB-ARC-WD40 genes are the best candidates to be functional equivalent of plant resistance genes. To the best of our knowledge, the TIR-Sel1 domain fusion is as yet unreported in any other organisms. We also searched for proteins that may contain the LysM domain [Pfam01476], a peptidoglycan recognition module involved in fungal and bacterial perception in plants*(275) (276)*. Although LysM motives are widely distributed both in prokaryotes and eukaryotes, no homologue was identified in *C. crispus*.

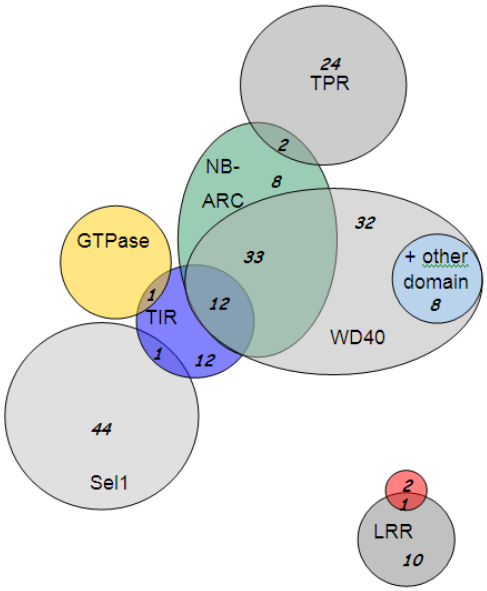

Fig. S8.1. Overview of multigenic families potentially involved in pathogen recognition, defense reactions, and signalling. Grey shades: repeats involved in interactions with specific ligands (WD40, LRR, TPR, Sel1). Colours: functional domains involved in signalling, including defence (NB-ARC, TIR, GTPase, F-box).

8.2 Candidate defence effectors

Thallus bleaching associated to the presence of epiphytic bacteria is common in several red algal species, and elicitor-triggered hypersensitive cell death has been described in the agarophyte *Gracilaria conferta* (277). Orthologues of apoptosis- and programmed cell death-related genes were annotated, leading to an overall picture similar of the one found in plants. In addition, a clear homologue of Bax inhibitor (CHC_T00004990001), also present in *C. merolae*, was found. However, neither corresponding Bax interactor nor indeed any member of the Bcl2 family was identified. The latter contains animal antiapoptotic and proapoptotic regulators that are notoriously absent from plant genomes. More surprisingly, a clear Sgt1 orthologue was found, but its interactant Rar1 doesn't seem to be present in *C. crispus*. The genome also contains three metacaspases (proteases involved in various steps of programmed cell death, senescence, and other developmental processes, Fig. S8.2): one type I metacaspase of fungal origin (CHC_T00000832001), one type II metacaspase (CHC_T00009024001), and one truncated sequence lacking the conserved H-C catalytic dyad of caspases, suggesting that this last gene is a pseudogene. No homologue of plant vacuolar processing enzymes (another group of proteases involved in plant cell death programs) was identified.

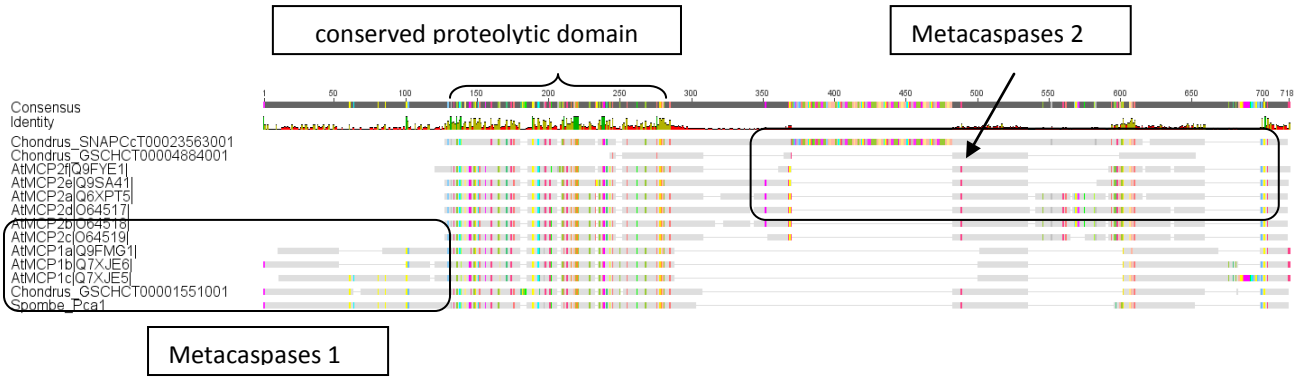

Fig. S8.2. Domain organisation and assignation of *C. crispus* metacaspases. GENEID changes

Over 15 additional homologues of plant, yeast or animal genes involved in cell death were identified and annotated (apoptosis inducing factor, anamorsin, exportin, rpp8, pheophorbide oxigenase, etc...). However, they are mostly involved in fundamental cellular processes, the disruption of which induces cell death. In the absence of specific functional information in red algae, it would be highly speculative to infer whether they are indeed involved in controlling cell death in *C. crispus*.

Additionally, autophagy is a universal process critical in plant, animal, and other eukaryote immune responses. However, the search for autophagy-related genes (Atg1-26 genes in yeast) did not unveil any unambiguous orthologues. Most candidate hits are so divergent that orthology with plant or yeast autophagy genes is difficult to infer.

8.3 Halogen metabolism

Marine red macroalgae are known to contain a large diversity of organo-halogenated compounds, especially brominated ones. Firstly considered as by-products of anti-oxidative processes, there is increasing evidence suggesting a role of these metabolites in the active chemical defence in macroalgae, by acting as antibiotics, antifouling, or anti-settlement compounds (278). The annotation of the *C. crispus* nuclear genome has revealed several large gene families related to such a metabolism, and the transcriptomic data indicate that some of these enzymes are also highly expressed (Table S8.1), suggesting an active halogen metabolism in the red alga.

Table S8.1. Halogen-related gene families identified in *C. crispus* and number of associated ESTs.

Heme-peroxidase	# ESTs	**PAP2/Vanadium haloperoxidase-like**	# ESTs	**Haloalkane dehalogenase**	# ESTs
CHC_T00009226001	12	CHC_T00009459001	21	CHC_T00009353001	35
CHC_T00006012001	17	CHC_T00000904001	22	CHC_T00008448001	34
CHC_T00004260001	12	CHC_T00009216001	20	CHC_T00009155001	>50
CHC_T00006807001	23	CHC_T00000065001	65	CHC_T00001736001	>50
CHC_T00009008001	4	CHC_T00008835001	0	CHC_T00001738001	>25
CHC_T00009515001	3	CHC_T00000016001	11	CHC_T00000781001	18
CHC_T00008590001	4	CHC_T00008578001	1		
CHC_T00009006001	4	CHC_T00008466001	4		
CHC_T00008836001	6	scaffold_3:109456..111063	2		
CHC_T00008271001	5	CHC_T00008343001	16		
CHC_T00004197001	22	CHC_T00008360001	14		
CHC_T00001140001	8	CHC_T00007429001	>100		
CHC_T00009490001	0	scaffold_41:29677..31194	>100		
CHC_T00008635001	13	CHC_T00009121001	>100		
CHC_T00007571001	>300	CHC_T00003497001	>100		
CHC_T00009516001	0				
Scaffold_67:222618..220849 *	10				
CHC_T00000317001	50				
scaffold_546:22417..20666 *	10				
scaffold_519:12794..10896 *	10				
CHC_T00000268001	8				
CHC_T00007624001	2				
CHC_T00008635001	0				

*pseudogene

A first and unexpected large family is formed by 20 genes (plus three highly conserved pseudogenes) closely related to the animal heme-peroxidase homologues. The red algal encoded proteins share ~35% of identity with the full-length animal isoforms, but some motifs are particularly well conserved. All the residues known from vertebrate myeloperoxidases to be necessary for heme or heme iron binding, and for catalytic activity, are conserved (Fig. S8.3).

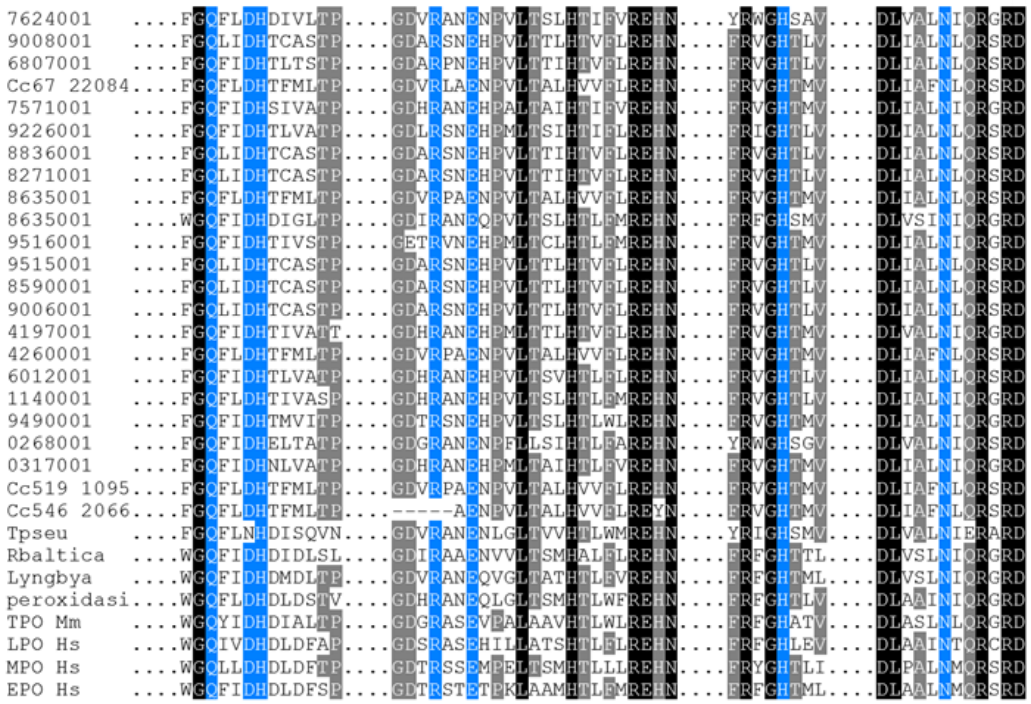

Fig. S8.3. Portions of multiple amino acid sequence alignment of *C. crispus* heme peroxidase with homologues from diverse origins, showing the conservation of catalytic and heme covalently linked residues (coloured in blue). Abbreviations: number N, gene ID CHC_T0000N; CcXX_YY, *C. crispus* scaffold XX, gene starting at YY position for pseudogenes; Tpseu, *T. pseudonana* heme peroxidase (XP_002286634); Rbaltica, *Rhodopirellula baltica* peroxidase (NP_864006); Lyngbya, *Lyngbya sp.* peroxidase (ZP_01620446); peroxidasi, *Homo sapiens* peroxidasin (NP_036425); TPO_Mm, *Mus musculus* thyroid peroxidase (NP_033443); LPO_Hs, *H. sapiens* lactoperoxidase (NP_006142); MPO_Hs, *H. sapiens* myeloperoxidase (P05164); EPO_Hs, *H. sapiens* eosinophil peroxidase (P11678). The conserved residues in all sequences are highlighted in black, and the residues that occur in at least 80% of the sequences are boxed in grey.

In mammals, these proteins also constitute large multigenic gene families which play a major role during pathogen ingress, releasing hypohalous acids through halide oxidation with the hydrogen peroxide during the respiratory burst (279). A similar role of these enzymes can be suspected in C. crispus as four NADPH oxidase genes are also found in the genome, and one of these members has previously been shown to be involved in defence responses in the red alga (280). Almost all of these heme peroxidase sequences in *C. crispus* are supported by ESTs (Table S8.1). Both the family size and the transcriptomic data point to an important role played by these proteins in *C. crispus*. Cyanobacterial homologues have recently been described (281), and current public database screenings retrieve additional prokaryotic homologues from aquatic origin. No isoforms were found in the genome sequence of the unicellular red algae *G. sulphuraria* and *C. merolae*, but at least one homologous partial sequence is present in the red algal *P. yezoensis* ESTs. Five genes in the brown

alga *E. siliculosus* show high similarities with the *C. crispus* genes, but do not feature the conserved residues involved in activity and heme prothetic group fixation, whereas an homologous gene seems to be present in the *T. pseudonana* genome (Fig. S8.4).

Phylogenetic analyses showed that *C. crispus* heme-peroxidases form an independent well supported clade, distinct from those of animals and bacteria (Fig. S8.4). This provides additional evidence for an ancient bacterial origin of proteins belonging to the peroxidase-cyclooxygenase superfamily (like animal-like heme peroxidases) (281). The primordial peroxidases probably played a protective role against oxidative damage during the early stages of the oxygen-containing biosphere era in the ancestral bacteria.

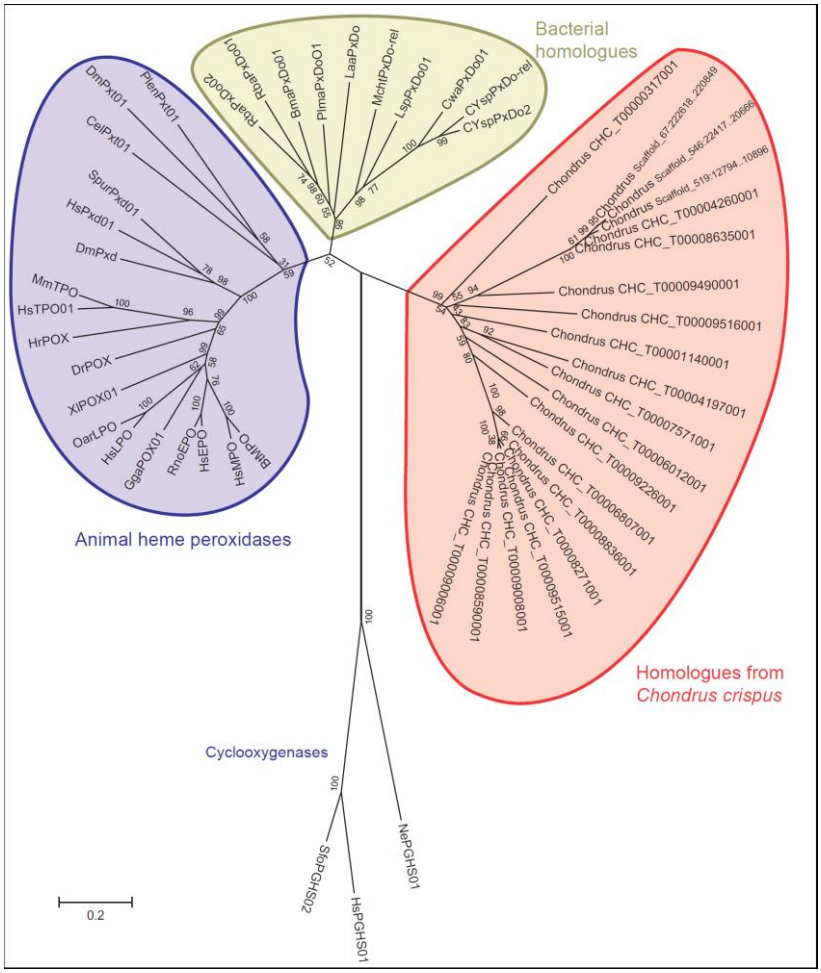

Fig. S8.4. Phylogeny of animal-heme peroxidases and homologues. The Neighbour-Joining method was applied using a multiple alignment of peroxidase domains with 1,000 bootstrap replications and the Dayhoff model for amino acid substitutions. The tree was obtained and presented using the MEGA5 package*(282)*. All sequences except the *C. crispus* ones are from PeroxiBase (http://peroxibase.toulouse.inra.fr/index.php).

In addition, the genome encodes at least 15 proteins belonging to the PAP2 (phosphatidic acid phosphatase type 2)-haloperoxidase family (Table S8.1). Like the heme-peroxidases, these enzymes

catalyze, in presence of hydrogen peroxide, the oxidation of halides which can then be used to transfer halogen on a wide range of organic molecules. Finally, there is also a group of six genes showing strong similarities with haloalkane dehalogenase (HLD) enzymes, which catalyze the cleavage of carbon-halogen bonds by a hydrolytic mechanism. Up to now, HLDs have been mainly characterized in bacteria, (*e.g.* in Actinobacteria and Proteobacteria), where they are involved in bioremediation and detoxification of organo-halogenated pollutants (283) (284).

The large size of these three gene families, putatively involved in both halogenation and dehalogenation enzymatic processes, might reflect a need to finely modulate halogen metabolism in *C. crispus*, and supports its major role in algal physiology in marine environment.

8.4 Cytochrome P450

Cytochrome P450s (CYPs) are extremely versatile membrane bound heme-thiolate oxidoreductases, which are present in domains of life (285) (286). Analysis of the *C. crispus* genome identified 20 CYPs (17 genes and three pseudogenes) which have been named following the above mentioned nomenclature system and updated CYP family order (http://drnelson.uthsc.edu/cytochromep450.html). These CYPs belong to seven families and 16 subfamilies, five of these families being new ones (from CYP805 to CYP809), and the two other being established families (CYP51 and CYP97). Functional information is only available for these two latter families. CYP51 enzymes catalyse the removal of the 14α-methyl group from sterol molecules, which have been highly conserved during the course of evolution*(287)*. *C. crispus* CYP51G1 belongs to the common CYP51G obtusifoliol 14α-demethylase gene subfamily of plants, usually present as a single copy per species, and likely has similar activity. The genome contains one member of plant CYP97 family, which is involved in carotenoid hydroxylation*(288)*. CYP97G1 is most similar to members from the CYP97E and CYP97B subfamilies, and is the first CYP97 so far identified in red algae. Considering functional characterization of other members of the CYP97 family, we can hypothesize that CYP97G1 is involved in either carotenoid ε- and/or β-ring hydroxylation as carotenoids with both rings are present in *C. crispus* (289). Among the five new *C. crispus* CYP families, CYP808 is the largest with 12 members, nine genes distributed in eight subfamilies and three pseudogenes, followed by CYP805 family, which count three genes distributed in three subfamilies. The other three new families, *i.e.* CYP805, CYP806, and CYP809 families, contain a single gene. Finally, it is worth to be noted that CYP710 family members (sterol C22-desaturases) seem to be absent from the genome, despite the apparent requirement for this activity as several Δ22-sterols have been found in *C. crispus* (181). This is quite intriguing as CYP710 family, which has an equivalent in fungi (CYP61), is present in *C. merolae* and *G. sulphuraria*, as well as in all green plants studied (290). Sterol C22-desaturase acting in *C. crispus* must then be highly divergent or different from currently known C22-desaturases as none of the *C. crispus* CYPs contain the conserved characteristic substrate recognition site (SRS4) of C22-desaturases (291).

Table S8.2. Cytochrome P-450 related genes in *C. crispus.*

Name	EC number	Gene ID
CYP51G1 obtusifoliol 14α- demethylase	1.14.13.70	CHC_T00009303001
CYP97G1 carotenoid hydroxylase	1.14.13.-	CHC_T00009187001
CYP805A1	1.14.-.-	CHC_T00008302001
CYP805B1	1.14.-.-	CHC_T00009144001
CYP805C1	1.14.-.-	CHC_T00008799001
CYP806A1	1.14.-.-	CHC_T00008886001
CYP807A1	1.14.-.-	CHC_T00008350001
CYP808A1	1.14.-.-	CHC_T00009499001
CYP808A2	1.14.-.-	CHC_T00008663001
CYP808A3p		(GIDCcT00019422001)*
CYP808A4p		(GIDCcT00012554001)*
CYP808B1	1.14.-.-	(GIDCcT00009562001)
CYP808C1	1.14.-.-	CHC_T00002715001
CYP808D1	1.14.-.-	CHC_T00009131001
CYP808E1	1.14.-.-	CHC_T00008672001
CYP808F1	1.14.-.-	CHC_T00009062001
CYP808G1	1.14.-.-	CHC_T00008616001
CYP808H1	1.14.-.-	CHC_T00008813001
CYP808I1p		(SNAPCcT00026414001)*
CYP809A1	1.14.-.-	CHC_T00009041001

*pseudogene

8.5 Stress genes

The intertidal is a dynamic environment with rapid changes in environmental parameters forcing active responses from its inhabitants *(*292*)*. It is therefore important to have effective methods to reduce effects of environmental stressors. The genome codes for numerous chaperones, chaperone related proteins and enzymes involved in reactive oxygen scavenging (Table S8.3). All eight subunits of the T complex were found, as well as four out of five subunits of the tubulin binding complex (Table S8.5); the missing subunit is a small peptide which could have been missed by Blast algorithms.

Table S8.3. Reactive oxygen scavenging enzymes in *C. crispus.*

ROS scavenging enzyme	*Chondrus crispus*	*Arabidopsis thaliana*	*Ectocarpus siliculosus*	Gene ID
ascorbate peroxidase	2	9	3	CHC_T00005198001
″				CHC_T00008640001
Cu/Zn superoxide dismutase	2	4	6	CHC_T00008473001
″				SNAPCcT00042582001
Fe/Mn superoxide dismutase	2	5	6	CHC_T00003067001
″				CHC_T00005214001
Catalase	2	3	12	CHC_T00007915001
″				CHC_T00005477001
glutathione reductase	1	2	3	CHC_T00004712001
monodehydroascorbate reductase	1	5	2	CHC_T00006866001
dehydroascorbate reductase	2	4	1	CHC_T00005252001
glutathione peroxidase	2	8	7	CHC_T00003751001
″				CHC_T00005150001
peroxiredoxin	4	6	3	CHC_T00001518001
″				CHC_T00009369001
″				CHC_T00008233001
″				CHC_T00004545001
glutaredoxin	4	±30	8	CHC_T00004545001
″				CHC_T00007132001
″				CHC_T00008414001
″				CHC_T00003957001
thioredoxin reductase	1	6	4	CHC_T00001349001
Total	25	83	62	

Table S8.4. Heat shock proteins and related proteins in *C. crispus*.

Proteins	Gene ID
HSP90	CHC_T00002029001
″	CHC_T00003878001
″	CHC_T00004017001
″	CHC_T00004336001*
HSP100	CHC_T00007938001

″	CHC_T00001090001[f]
HSP20	CHC_T00009442001
″	CHC_T00007118001
″	CHC_T00001909001
″	CHC_T00008399001
″	CHC_T00007114001
″	CHC_T00008937001
DnaK chaperone	CHC_T00001524001
″	CHC_T00006407001
″	CHC_T00002766001
″	CHC_T00008003001
″	CHC_T00000474001
″	CHC_T00005189001
″	CHC_T00005757001
″	CHC_T00002781001*
HSP40/DnaJ	CHC_T00003603001
″	CHC_T00002087001
″	CHC_T00007355001
″	CHC_T00004401001
″	CHC_T00001918001
″	CHC_T00000787001
″	CHC_T00003939001
HSP70 binding protein	CHC_T00002243001
HSP GrpE	CHC_T00009434001
″	CHC_T00008372001
CPN60/GroEL	CHC_T00006165001
″	CHC_T00008986001
″	CHC_T00009543001*
HSP10/GroES	CHC_T00009161001
″	CHC_T00005425001
heat shock transcription factor	CHC_T00002794001

*pseudogene; [f] fragment

Table S8.5 Stress related genes in *C. crispus.*

Protein	Gene ID
T-Complex protein	
alpha	CHC_T00006799001
beta	CHC_T00003679001
gamma	CHC_T00001659001
delta	CHC_T00005927001
epsilon	CHC_T00007860001
zeta	CHC_T00003824001
eta	CHC_T00005762001
theta	CHC_T00007622001
tubulin binding complex	
A	Not found
B	CHC_T00000407001
C	CHC_T00000351001
D	CHC_T00004881001
E	CHC_T00000070001
protein disulfide isomerase	CHC_T00009389001
″	CHC_T00008626001
peptidyl-prolyl cis-trans isomerase/calreticulin	CHC_T00009552001
″	CHC_T00008963001
″	CHC_T00002560001
″	CHC_T00008968001
″	CHC_T00008420001
″	CHC_T00006996001
″	CHC_T00009539001
″	CHC_T00001971001
″	CHC_T00009116001
″	CHC_T00008726001
calnexin	CHC_T00007100001

9 References

1. Batzoglou S, Jaffe D, Stanley K, Butler J (2002) ARACHNE: a whole-genome shotgun assembler. *Genome*:177–189.

2. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Prot Bioinform* Chapter 4:Unit 4.10.

3. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl Acids Res* 27:573–580.

4. Price AL, Jones NC, Pevzner P a (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1:i351–8.

5. Bairoch A et al. (2005) The universal protein resource (UniProt). *Nucleic Acids Res* 33:D154–9.

6. Birney E, Durbin R (2000) Using GeneWise in the Drosophila annotation experiment. *Genome Res* 10:547–548.

7. Kent WJ (2002) BLAT---The BLAST-like alignment tool. *Genome Res* 12:656–664.

8. Parra GN (2000) GeneID in Drosophila. *Genome Res* 10:511–515.

9. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

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

11. Howe KL, Chothia T, Durbin R (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res* 12:1418–1427.

12. Putnam NH et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317:86–94.

13. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

14. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.

15. Lanier W, Moustafa A, Bhattacharya D, Comeron JM (2008) EST analysis of Ostreococcus lucimarinus, the most compact eukaryotic genome, shows an excess of introns in highly expressed genes. *PloS one* 3:e2171.

16. Lin K, Zhang D-Y (2005) The excess of 5’ introns in eukaryotic genomes. *Nucleic Acids Res* 33:6522–6527.

17. Roy SW, Gilbert W (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci USA* 102:5773–5778.

18. Cavalier-Smith T (1985) Selfish DNA and the origin of introns. *Nature* 315:283–284.

19. Coghlan A, Wolfe KH (2004) Origins of recently gained introns in Caenorhabditis. *Proc Natl Acad Sci USA* 101:11362–11367.

20. Denoeud F et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330:1381–1385.

21. Sharp PA, others (1985) On the origin of RNA splicing and introns. *Cell* 42:397–400.

22. Roy SW, Irimia M (2009) Mystery of intron gain: new data and new models. *Trends Genet* 25:67–73.

23. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.

24. Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, De Oliveira MC (2004) Comparative analysis of the complete plastid genome sequence of the red alga Gracilaria tenuistipitata var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J Mol Evol* 59:464–77.

25. Reith M, Munholland J (1995) Complete nucleotide sequence of the Porphyra purpurea chloroplast genome. *Plant Mol Biol Rep* 13:333–335.

26. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403.

27. Leblanc C, Boyen C, Loiseaux-de Goër S (1995) Organisation of the plastid genome from the rhodophyte Chondrus crispus (Gigartinales); sequence and phylogeny of the 16S rRNA gene. *Eur J Phycol* 30:133–140.

28. Richard O, Bonnard G, Grienenberger JM, Kloareg B, Boyen C (1998) Transcription initiation and RNA processing in the mitochondria of the red alga Chondrus crispus: convergence in the evolution of transcription mechanisms in mitochondria. *J Mol Biol* 283:549–557.

29. Supek F, Vlahovicek K (2004) INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics* 20:2329–2330.

30. Nyvall Collén P et al. (2011) Analysis of expressed sequence tags from the agarophyte Gracilaria tenuistipitata (Rhodophyta). *J Appl Phycol*.

31. Nikaido I et al. (2000) Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, Porphyra yezoensis. *Dna Res* 7:223–227.

32. Rensing S a et al. (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol Biol* 7:130.

33. Lang D et al. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Gen Biol Evol* 2:488–503.

34. Chan CX et al. (2011) Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. *Curr Biol* 21:328–333.

35. Altschul S, Gish W, Miller W, Myers E (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

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

37. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One* 2:e383.

38. Iseli C, Jongeneel C V, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *International Conference on Intelligent Systems for Molecular Biology*:138–148.

39. Sicheritz-Pontén T, Andersson SG (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* 29:545–552.

40. Huerta-Cepas J et al. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 39:D556–60.

41. Smith T, Waterman M (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197.

42. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

43. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9:286–298.

44. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithm Mol Biol* 3:6.

45. Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol* 24:1380–1383.

46. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692–1699.

47. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

48. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695.

49. Guindon S et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321.

50. Akaike H (1973) in *Second international symposium on information theory* (Springer Verlag), pp 267–281.

51. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 9:235.

52. Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24:1540–1541.

53. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38:D355–360.

54. Emanuelsson O, Brunak S, Von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.

55. Belostotsky D (2009) Exosome complex and pervasive transcription in eukaryotic genomes. *Curr Opin Cell Biol* 21:352–358.

56. Wahl MC, Will CL, Lührmann R (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* 136:701–718.

57. Hunt AG et al. (2008) Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics* 9:220.

58. Zhao H, Xing D, Li QQ (2009) Unique features of plant cleavage and polyadenylation specificity factor revealed by proteomic studies. *Plant Physiol* 151:1546–1556.

59. Meeks LR, Addepalli B, Hunt AG (2009) Characterization of genes encoding poly(A) polymerases in plants: evidence for duplication and functional specialization. *PloS One* 4:e8082.

60. Aravind L, Koonin E V (1999) DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acids Res* 27:1609–1618.

61. Zhuang T, Kashiwabara S, Nogushi J (2004) Transgenic expression of testis-specific poly (A) polymerase TPAP in wild-type and TPAP-deficient mice. *J Reprod Dev* 50:207–213.

62. Aphasizhev R, Aphasizheva I (2008) Terminal RNA uridylyltransferases of trypanosomes. *Biochim Biophys Acta* 1779:270–280.

63. Rissland OS, Norbury CJ (2008) The Cid1 poly(U) polymerase. *Biochim Biophys Acta* 1779:286–294.

64. Norbury CJ (2010) 3 uridylation and the regulation of RNA function in the cytoplasm. *Biochem Soc Trans* 38:1150–1153.

65. Schmidt M-J, Norbury CJ (2010) Polyadenylation and beyond: emerging roles for noncanonical poly(A) polymerases. *Wiley Interdiscip Rev*:142–151.

66. Wang S-W, Stevenson AL, Kearsey SE, Watt S, Bähler J (2008) Global role for polyadenylation-assisted nuclear RNA degradation in posttranscriptional gene silencing. *Mol Cell Biol* 28:656–665.

67. Aphasizhev R (2005) RNA uridylyltransferases. *Cell Mol Life Sci* 62:2194–203.

68. Zimmer SL, Fei Z, Stern DB (2008) Genome-based analysis of Chlamydomonas reinhardtii exoribonucleases and poly(A) polymerases predicts unexpected organellar and exosomal features. *Genetics* 179:125–136.

69. Stevenson AL, Norbury CJ (2006) The Cid1 family of non-canonical poly (A) polymerases. *Yeast* 23:991–1000.

70. Martin G, Keller W (2007) RNA-specific ribonucleotidyl transferases. *RNA* 13:1834–1849.

71. San Paolo S et al. (2009) Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. *PLoS Genetics* 5:e1000555.

72. Wang SW, Norbury C, Harris a L, Toda T (1999) Caffeine can override the S-M checkpoint in fission yeast. *J Cell Sci* 112:927–937.

73. Wang SW, Toda T, MacCallum R, Harris AL, Norbury C (2000) Cid1, a fission yeast protein required for SM checkpoint control when DNA polymerase delta or epsilon is inactivated. *Mol Cell Biol* 20:3234–3244.

74. Matsuyama A et al. (2006) ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat Biotech* 24:841–847.

75. Labib K, Gambus A (2007) A key role for the GINS complex at DNA replication forks. *Trend Cell Biol* 17:271–278.

76. MacNeill S a (2010) Structure and function of the GINS complex, a key component of the eukaryotic replisome. *Biochem J* 425:489–500.

77. Remus D et al. (2009) Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* 139:719–730.

78. Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. *Ann Rev Biochem* 71:333–374.

79. Remus D, Diffley JFX (2009) Eukaryotic DNA replication control: lock and load, then fire. *Curr Opin Cell Biol* 21:771–777.

80. Hubscher U, Maga G, Spadari S (2002) Eukaryotic DNA polymerases. *Ann Rev Biochem* 71:133–163.

81. Moriyama T, Terasawa K, Fujiwara M, Sato N (2008) Purification and characterization of organellar DNA polymerases in the red alga Cyanidioschyzon merolae. *FEBS J* 275:2899–2918.

82. Kimura S, Sakaguchi K (2006) DNA repair in plants. *Chem Rev* 106:753–766.

83. Villeneuve AM, Hillers KJ (2001) Whence meiosis? *Cell* 106:647–650.

84. Ramesh MA, Malik SB, Logsdon JM (2005) A phylogenomic inventory of meiotic genes: evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Curr Biol* 15:185–191.

85. Jackson N et al. (2006) Reduced meiotic crossovers and delayed prophase I progression in AtMLH3-deficient Arabidopsis. *EMBO J* 25:1315–1323.

86. Tomari Y, Zamore PD (2005) Perspective: machines for RNAi. *Gen Dev* 19:517–529.

87. Nakayashiki H, Kadotani N, Mayama S (2006) Evolution and diversification of RNA silencing proteins in fungi. *J Mol Evol* 63:127–135.

88. Nakayama J, Rice JC, Strahl BD, Allis CD, Grewal SI (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* 292:110–113.

89. Tachibana M, Sugimoto K, Fukushima T, Shinkai Y (2001) Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *J Biol Chem* 276:25309–25317.

90. Struhl G (1981) A gene product required for correct initiation of segmental determination in Drosophila. *Nature* 293:36–41.

91. Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H (2010) Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics* 5:301–312.

92. Grimaud C et al. (2006) RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* 124:957–971.

93. Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Gene Dev* 16:2893–2905.

94. Birve a et al. (2001) Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. *Development* 128:3371–3379.

95. Okano Y et al. (2009) A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. *Proc Natl Acad Sci USA* 106:16321–16326.

96. Singh DK, Ahn B, Bohr VA (2009) Roles of RECQ helicases in recombination based DNA repair, genomic stability and aging. *Biogerontology* 10:235–252.

97. Hartung F, Puchta H (2006) The RecQ gene family in plants. *J Plant Physiol* 163:287–96.

98. Hartung F, Suer S, Puchta H (2007) Two closely related RecQ helicases have antagonistic roles in homologous recombination and DNA repair in Arabidopsis thaliana. *Proc Natl Acad Sci USA* 104:18836–18841.

99. Kodner RB, Summons RE, Pearson A, King N, Knoll AH (2008) Sterols in a unicellular relative of the metazoans. *Proc Natl Acad Sci USA* 105:9897–9902.

100. Vindigni A, Marino F, Gileadi O (2010) Probing the structural basis of RecQ helicase function. *Biophys Chem* 149:67–77.

101. Kapp LD, Lorsch JR (2004) The molecular mechanics of eukaryotic translation. *Ann Rev Biochem* 73:657–704.

102. Barakat A et al. (2001) The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. *Plant Physiol* 127:398–415.

103. Robinson RC et al. (2001) Crystal structure of Arp2/3 complex. *Science* 294:1679–1684.

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

105. Perroud P-F, Quatrano RS (2008) BRICK1 is required for apical cell growth in filaments of the moss Physcomitrella patens but not for gametophore morphology. *Plant Cell* 20:411–22.

106. Shimmen T, Ridge R, Lambiris I, Plazinski J (2000) Plant myosins. *Protoplasma*:1–10.

107. King SM (2002) Dyneins motor on in plants. *Traffic* 3:930–1.

108. Petrásek J, Schwarzerová K (2009) Actin and microtubule cytoskeleton interactions. *Curr Opin Plant Biol* 12:728–734.

109. Dacks JB, Poon PP, Field MC (2008) Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. *Proc Natl Acad Sci USA* 105:588–593.

110. Hirst J et al. (2011) The fifth adaptor protein complex. *PLoS Biol* 9:e1001170.

111. Nickerson D, Brett C (2009) Vps-C complexes: gatekeepers of endolysosomal traffic. *Curr Opin Cell Biol* 21:543–551.

112. He B, Guo W (2009) The exocyst complex in polarized exocytosis. *Curr Opin Cell Biol* 21:537–542.

113. Koumandou VL, Dacks JB, Coulson RMR, Field MC (2007) Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. *BMC Evol Biol* 7:29.

114. Carr C (2010) At the junction of SNARE and SM protein function. *Curr Opin Cell Biol* 22:488–495.

115. Sanderfoot A (2007) Increases in the number of SNARE genes parallels the rise of multicellularity among the green plants. *Plant Physiol* 144:16–17.

116. Hanks SK, Hunter T (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J* 9:576–596.

117. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298:1912–1934.

118. Judelson HS, Ah-Fong AM V (2010) The kinome of Phytophthora infestans reveals oomycete-specific innovations and links to other taxonomic groups. *BMC Genomics* 11:700.

119. Goldberg JM et al. (2006) The dictyostelium kinome--analysis of the protein kinases from a simple model organism. *PLoS Genetics* 2:e38.

120. Dardick C, Chen J, Richter T, Ouyang S, Ronald P (2007) The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiol* 143:579–586.

121. Cock JM et al. (2010) The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–21.

122. Leipe DD, Wolf YI, Koonin E V, Aravind L (2002) Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol* 317:41–72.

123. Reynaud EG et al. (2005) Human Lsg1 defines a family of essential GTPases that correlates with the evolution of compartmentalization. *BMC Biol* 3:21.

124. Karbstein K (2007) Role of GTPases in ribosome assembly. *Biopolymers* 87:1–11.

125. Verstraeten N, Fauvart M, Versées W, Michiels J (2011) The universally conserved prokaryotic GTPases. *Microbiol Mol Biol Rev* 75:507–542.

126. Andrès C, Agne B, Kessler F (2010) The TOC complex: preprotein gateway to the chloroplast. *Biochim Biophys Acta* 1803:715–723.

127. Matsuzaki M et al. (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* 428:653–657.

128. Miyagishima S (2011) Mechanism of plastid division: from a bacterium to an organelle. *Plant Physiol* 155:1533–1544.

129. Haller O, Stertz S, Kochs G (2007) The Mx GTPase family of interferon-induced antiviral proteins. *Microbes Infect* 9:1636–1643.

130. Grant BD, Caplan S (2008) Mechanisms of EHD/RME-1 protein function in endocytic transport. *Traffic* 9:2043–2052.

131. Vestal DJ, Jeyaratnam JA (2011) The guanylate-binding proteins: Emerging insights into the biochemical properties and functions of this family of large interferon-induced guanosine triphosphatase. *J Interferon Cytokine Res* 31:89–97.

132. Hu J, Prinz WA, Rapoport TA (2011) Weaving the web of ER tubules. *Cell* 147:1226–1231.

133. Rojas a. M, Fuentes G, Rausell A, Valencia A (2012) The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *J Cell Biol* 196:189–201.

134. Gillingham AK, Munro S (2007) The small G proteins of the Arf family and their regulators. *Ann Rev Cell Dev Biol* 23:579–611.

135. Brighouse A, Dacks JB, Field MC (2010) Rab protein evolution and the history of the eukaryotic endomembrane system. *Cell Mol Life Sci* 67:3449–3465.

136. Diekmann Y et al. (2011) Thousands of Rab GTPases for the Cell Biologist. *PLoS Comput Biol* 7:e1002217.

137. Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M.C., & Dacks J. (2012) Sculpting the endomembrane system in deep time: High resolution phylogenetics of Rab GTPases. *J Cell Sci*:In press.

138. Carney DS, Davies B a, Horazdovsky BF (2006) Vps9 domain-containing proteins: activators of Rab5 GTPases from yeast to neurons. *Trend Cell Biol* 16:27–35.

139. Elias M, Archibald JM (2009) The RJL family of small GTPases is an ancient eukaryotic invention probably functionally associated with the flagellar apparatus. *Gene* 442:63–72.

140. Li Y, Hu J (2011) Small GTPases and cilia. *Prot Cell* 2:13–25.

141. Wennerberg K, Der CJ (2004) Rho-family GTPases: it's not only Rac and Rho (and I like it). *J Cell Sci* 117:1301–1312.

142. Jaffe AB, Hall A (2005) Rho GTPases: biochemistry and biology. *Ann Rev Cell Dev Biol* 21:247–269.

143. Brembu T, Winge P, Bones AM, Yang Z (2006) A RHOse by any other name: a comparative analysis of animal and plant Rho GTPases. *Cell Res* 16:435–445.

144. Van Dam TJP, Bos JL, Snel B (2011) Evolution of the Ras-like small GTPases and their regulators. *Small GTPases* 2:4–16.

145. Durán R V, Hall MN (2012) Regulation of TOR by small GTPases. *EMBO Rep* 13:121–128.

146. Rensen W, Mangiacasale R, Ciciarello M, Lavia P (2008) The GTPase Ran: regulation of cell life and potential roles in cell transformation. *Front Biosci* 13:4097.

147. Vlahou G, Eliáš M, Von Kleist-Retzow J-C, Wiesner RJ, Rivero F (2011) The Ras related GTPase Miro is not required for mitochondrial transport in Dictyostelium discoideum. *Eur J Cell Biol* 90:342–355.

148. McCollum D, Gould KL (2001) Timing is everything: regulation of mitotic exit and cytokinesis by the MEN and SIN. *Trend Cell Biol* 11:89–95.

149. Kevei E et al. (2007) Arabidopsis thaliana circadian clock is regulated by the small GTPase LIP1. *Curr Biol* 17:1456–1464.

150. Montalbano J, Lui K, Sheikh MS, Huang Y (2009) Identification and characterization of RBEL1 subfamily of GTPases in the Ras superfamily involved in cell growth regulation. *J Biol Chem* 284:18129–18142.

151. Gotthardt K, Weyand M, Kortholt A, Van Haastert PJM, Wittinghofer A (2008) Structure of the Roc-COR domain tandem of C. tepidum, a prokaryotic homologue of the human LRRK2 Parkinson kinase. *EMBO J* 27:2239–2249.

152. Marín I, Van Egmond WN, Van Haastert PJM (2008) The Roco protein family: a functional perspective. *FASEB J* 22:3103–3110.

153. Zambounis A, Elias M, Sterck L, Maumus F, Gachon CMM (2012) Highly dynamic exon shuffling in candidate pathogen receptors ... What if brown algae were capable of adaptive immunity? *Mol Biol Evol* 29:1263–76.

154. Staal J, Dixelius C (2007) Tracing the ancient origins of plant innate immunity. *Trend Plant Sci* 12:334–42.

155. Arteni A a et al. (2008) Structure and organization of phycobilisomes on membranes of the red alga Porphyridium cruentum. *Photosynth Res* 95:169–174.

156. Algarra P, Thomas JC, Mousseau A (1990) Phycobilisome heterogeneity in the red alga Porphyra umbilicalis. *Plant Physiol* 92:570–576.

157. Redecker D, Wehrmeyer W, Reuter W (1993) Core substructure of the hemiellipsoidal phycobilisome from the red alga Porphyridium cruentum. *Eur J Cell Biol* 62:442–450.

158. Busch A, Nield J, Hippler M (2010) The composition and structure of photosystem I-associated antenna from Cyanidioschyzon merolae. *Plant J* 62:886–897.

159. Heddad M, Adamska I (2002) The evolution of light stress proteins in photosynthetic organisms. *Comp Funct Genom* 3:504–10.

160. Neilson J a D, Durnford DG (2010) Structural and functional diversification of the light-harvesting complexes in photosynthetic eukaryotes. *Photosynth Res* 106:57–71.

161. Derelle E et al. (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci USA* 103:11647–11652.

162. Murakami M, Ashikari M, Miura K, Yamashino T, Mizuno T (2003) The evolutionarily conserved OsPRR quintet: rice pseudo-response regulators implicated in circadian rhythm. *Plant Cell Physiol* 44:1229–1236.

163. Corellou F et al. (2009) Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *Plant Cell* 21:3436–3449.

164. Matsuo T et al. (2008) A systematic forward genetic analysis identified components of the Chlamydomonas circadian system. *Gen Dev* 22:918–930.

165. Onai K, Ishiura M (2005) PHYTOCLOCK 1 encoding a novel GARP protein essential for the Arabidopsis circadian clock. *Gen Cell* 10:963–972.

166. Hazen SP et al. (2005) LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms. *Proc Natl Acad Sci USA* 102:10387–10392.

167. Robson F et al. (2001) Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. *Plant J* 28:619–631.

168. Roberts SB, Lane TW, Morel FMM (1997) Carbonic anhydrase in the marine diatom Thalassiosira weissflogii (Bacillariophyceae). *J Phycol* 33:845–850.

169. Gravot A et al. (2010) Diurnal oscillations of metabolite abundances and gene analysis provide new insights into central metabolic processes of the brown alga Ectocarpus siliculosus. *New Phytol* 188:98–110.

170. Cardol P (2005) The mitochondrial oxidative phosphorylation proteome of Chlamydomonas reinhardtii deduced from the Genome Sequencing Project. *Plant Physiol* 137:447–459.

171. Cardol P et al. (2004) Higher plant-like subunit composition of mitochondrial complex I from Chlamydomonas reinhardtii: 31 conserved components among eukaryotes. *Biochim Biophys Acta* 1658:212–224.

172. Miflin B, Lea P (1977) Amino acid metabolism. *Ann Rev Plant Physiol* 28:299–329.

173. Lea P, Miflin B (1974) Alternative route for nitrogen assimilation in higher plants. *Nature* 251:614–616.

174. Lightfoot D, Baron A, Wootton J (1988) Expression of the Escherichia coli glutamate dehydrogenase gene in the cyanobacterium Synechococcus PCC6301 causes ammonium tolerance. *Plant mol biol* 11:335–344.

175. Robinson S et al. (1991) The role of glutamate dehydrogenase in plant nitrogen metabolism. *Plant Physiol* 95:509–516.

176. Genbauffe FS, Cooper TG (1991) The urea amidolyase (DUR1, 2) gene of Saccharomyces cerevisiae. *DNA Seq* 2:19–32.

177. Leftley J, Syrett P (1973) Urease and ATP: urea amidolyase activity in unicellular algae. *J Gen Microbiol* 77:109–115.

178. Li K, Xu E (2008) The role and the mechanism of γ-aminobutyric acid during central nervous system development. *Neurosci Bull* 24:195–200.

179. Allan WLAWL, Shelp BJSBJ (2006) Fluctuations of gamma-aminobutyrate, gamma-hydroxybutyrate, and related amino acids in Arabidopsis leaves as a function of the light-dark cycle, leaf age, and N stress. *Can J Bot* 84:1339–1346.

180. Dittami SM et al. (2011) Integrative analysis of metabolite and transcript abundance during the short-term response to saline and oxidative stress in the brown alga Ectocarpus siliculosus. *Plant Cell Environ* 34:629–642.

181. Tasende M (2000) Fatty acid and sterol composition of gametophytes and sporophytes of Chondrus crispus (Gigartinaceae, Rhodophyta). *Sci Mar* 64:421–426.

182. Pettitt TR, Harwood JL (1989) Alterations in lipid metabolism caused by illumination of the marine red algae Chondrus crispus and Polysiphonia lanosa. *Phytochemistry* 28:3295–3300.

183. Pettitt TR, Jones AL, Harwood JL (1989) Lipid metabolism in the red marine algae Chondrus crispus and Polysiphonia lanosa as modified by temperature. *Phytochemistry* 28:2053–2058.

184. Trevor R, Pettitt A, Jones L, Harwood JL (1989) Lipids of the marine red algae, Chondrus crispus and Polysiphonia lanosa. *Phytochemistry* 28:399–405.

185. Khotimchenko S, Vas' kovsky V (2004) An inositol-containing sphingolipid from the red alga Gracilaria verrucosa. *Bioorg Khim* 30:168–171.

186. Khotimchenko SV, Klochkova N, Vaskovsky V (1990) Polar lipids of marine macrophytic algae as chemotaxonomic markers. *Biochem Syst Ecol* 18:93–101.

187. Dunn TM, Lynch D V, Michaelson L V, Napier J (2004) A post-genomic approach to understanding sphingolipid metabolism in Arabidopsis thaliana. *Ann Bot* 93:483–497.

188. Bouarab K et al. (2004) The innate immunity of a marine red alga involves oxylipins from both the eicosanoid and octadecanoid pathways. *Plant Physiol* 135:1838.

189. Gaquerel E et al. (2007) Evidence for oxylipin synthesis and induction of a new polyunsaturated fatty acid hydroxylase activity in Chondrus crispus in response to methyljasmonate. *Biochim Biophys Acta* 1771:565–75.

190. Liu Q, Reith M (1994) Isolation of a gametophyte-specific cDNA encoding a lipoxygenase from the red alga Porphyra purpurea. *Mol Mar Biol Biotech* 3:206–209.

191. Popper ZA et al. (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu Rev Plant Biol* 62:567–590.

192. Cantarel BL et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* 37:D233–238.

193. Michel G, Tonon T, Scornet D, Cock JM, Kloareg B (2010) The cell wall polysaccharide metabolism of the brown alga Ectocarpus siliculosus. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *New Phytol* 188:82–97.

194. Michel G, Tonon T, Scornet D, Cock JM, Kloareg B (2010) Central and storage carbon metabolism of the brown alga Ectocarpus siliculosus: insights into the origin and evolution of storage carbohydrates in Eukaryotes. *New Phytol* 188:67–81.

195. Henrissat B, Coutinho PM, Davies GJ (2001) A census of carbohydrate-active enzymes in the genome of Arabidopsis thaliana. *Plant Mol Biol* 47:55–72.

196. Paul MJ, Primavesi LF, Jhurreea D, Zhang Y (2008) Trehalose metabolism and signaling. *Ann Rev Plant Biol* 59:417–441.

197. Pakker H et al. (2000) Effects of temperature on the photoreactivation of ultraviolet-B-induced DNA damage in Palmaria palmata (Rhodophyta). *J Phycol* 36:334–341.

198. Empadinhas N, Da Costa MS (2010) Diversity, biological roles and biosynthetic pathways for sugar-glycerate containing compatible solutes in bacteria and archaea. *Environ Microbiol* 13:2056–2077.

199. Kremer B (1980) Taxonomic implications of algal photoassimilate patterns. *Br Phycol J* 15:399–409.

200. Roberts E, Roberts AW (2009) A cellulose synthase (Cesa) gene from the red alga Porphyra yezoensis (Rhodophyta). *J Phycol* 45:203–212.

201. Matthews PR, Schindler M, Howles P, Arioli T, Williamson RE (2010) A CESA from Griffithsia monilis (Rhodophyta, Florideophyceae) has a family 48 carbohydrate-binding module. *J Exp Bot* 61:4461–4468.

202. Michel G et al. (2001) The kappa-carrageenase of P. carrageenovora features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases. *Structure* 9:513–525.

203. Rees D (1969) Conformational analysis of polysaccharides. Part II. Alternating copolymers of the agar–carrageenan–chondroitin type by model building in the computer with calculation of helical parameters. *J Chem Soc B*:217–226.

204. Knutsen S, Myslabodski D, Larsen B (1994) A modified system of nomenclature for red algal galactans. *Bot Mar* 37:163–169.

205. Usov A. (1998) Structural analysis of red seaweed galactans of agar and carrageenan groups. *Food Hydrocolloids* 12:301–308.

206. Genicot-Joncour S et al. (2009) The cyclization of the 3,6-anhydro-galactose ring of iota-carrageenan is catalyzed by two D-galactose-2,6-sulfurylases in the red alga Chondrus crispus. *Plant Physiol* 151:1609–1616.

207. Schauss SJ et al. (1995) Characterization of bovine tracheobronchial phenol sulphotransferase cDNA and detection of mRNA regulation by cortisol. *Biochem J* 311:209–217.

208. Barnes S, Buchina ES, King RJ, McBurnett T, Taylor KB (1989) Bile acid sulfotransferase I from rat liver sulfates bile acids and 3-hydroxy steroids: purification, N-terminal amino acid sequence, and kinetic properties. *J Lipid Res* 30:529–540.

209. Sakakibara Y et al. (1998) Molecular cloning, expression, and functional characterization of novel mouse sulfotransferases. *Biochem Biophys Res Commun* 247:681–686.

210. Nash a R et al. (1988) Oestrogen sulfotransferase: molecular cloning and sequencing of cDNA for the bovine placental enzyme. *Aust J Biol Sci* 41:507–516.

211. Takahashi S et al. (2009) Molecular cloning, expression and characterization of a novel mouse SULT6 cytosolic sulfotransferase. *J Biol Chem* 146:399–405.

212. Piotrowski M et al. (2004) Desulfoglucosinolate sulfotransferases from Arabidopsis thaliana catalyze the final step in the biosynthesis of the glucosinolate core structure. *J Biol Chem* 279:50717–50725.

213. Ito Y, Habuchi O (2000) Purification and characterization of N-acetylgalactosamine 4-sulfate 6-O-sulfotransferase from the squid cartilage. *J Biol Chem* 275:34728–34736.

214. Liu J, Shworak NW, Fritze LM, Edelberg JM, Rosenberg RD (1996) Purification of heparan sulfate D-glucosaminyl 3-O-sulfotransferase. *J Biol Chem* 271:27072–27082.

215. Rivera-Marrero C, Ritzenthaler JD, Newburn S, Roman J, Cummings RD (2002) Molecular cloning and expression of a novel glycolipid sulfotransferase in Mycobacterium tuberculosis. *Microbiology* 148:783–92.

216. Fukuta M et al. (1995) Molecular cloning and expression of chick chondrocyte chondroitin 6-sulfotransferase. *J Biol Chem* 270:18575–18580.

217. Mazany KD, Peng T, Watson CE, Tabas I, Williams KJ (1998) Human chondroitin 6-sulfotransferase: cloning, gene structure, and chromosomal localization. *Biochim Biophys Acta* 1407:92–97.

218. Fukuta M et al. (1997) Molecular cloning and characterization of human keratan sulfate Gal-6-sulfotransferase. *J Biol Chem* 272:32321–32328.

219. Ong E, Yeh JC, Ding Y, Hindsgaul O, Fukuda M (1998) Expression cloning of a human sulfotransferase that directs the synthesis of the HNK-1 glycan on the neural cell adhesion molecule and glycolipids. *J Biol Chem* 273:5190–5195.

220. Hiraoka N, Misra A, Belot F, Hindsgaul O, Fukuda M (2001) Molecular cloning and expression of two distinct human N-acetylgalactosamine 4-O-sulfotransferases that transfer sulfate to GalNAc beta 1-->4GlcNAc beta 1-->R in both N- and O-glycans. *Glycobiology* 11:495–504.

221. Kang HG, Evers MR, Xia G, Baenziger JU, Schachner M (2001) Molecular cloning and expression of an N-acetylgalactosamine-4-O-sulfotransferase that transfers sulfate to terminal and non-terminal beta 1,4-linked N-acetylgalactosamine. *J Biol Chem* 276:10861–10869.

222. Mikami T, Mizumoto S, Kago N, Kitagawa H, Sugahara K (2003) Specificities of three distinct human chondroitin/dermatan N-acetylgalactosamine 4-O-sulfotransferases demonstrated using partially desulfated dermatan sulfate as an acceptor: implication of differential roles in dermatan sulfate biosynthesis. *J Biol Chem* 278:36115–36127.

223. Evers MR, Xia G, Kang HG, Schachner M, Baenziger JU (2001) Molecular cloning and characterization of a dermatan-specific N-acetylgalactosamine 4-O-sulfotransferase. *J Biol Chem* 276:36344–36353.

224. Honke K et al. (2001) Molecular cloning and characterization of a human beta-Gal-3'-sulfotransferase that acts on both type 1 and type 2 (Gal beta 1-3/1-4GlcNAc-R) oligosaccharides. *J Biol Chem* 276:267–274.

225. Honke K et al. (1997) Molecular cloning and expression of cDNA encoding human 3'-phosphoadenylylsulfate:galactosylceramide 3'-sulfotransferase. *J Biol Chem* 272:4864–4868.

226. Hanson SR, Best MD, Wong C-H (2004) Sulfatases: structure, mechanism, biological activity, inhibition, and synthetic utility. *Angewandte Chemie* 43:5736–5763.

227. Kahnert A, Kertesz MA (2000) Characterization of a sulfur-regulated oxygenative alkylsulfatase from Pseudomonas putida S-313. *J Biol Chem* 275:31661–31667.

228. Davison J, Brunel F, Phanopoulos A, Prozzi D, Terpstra P (1992) Cloning and sequencing of Pseudomonas genes determining sodium dodecyl sulfate biodegradation. *Gene* 114:19–24.

229. Barbeyron T, Potin P, Richard C, Collin O, Kloareg B (1995) Arylsulphatase from Alteromonas carrageenovora. *Microbiology* 141:2897–2904.

230. Ball SG, Morell MK (2003) From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule. *Ann Rev Plant Biol* 54:207–233.

231. Zeeman S, Smith S, Smith A (2007) The diurnal metabolism of leaf starch. *Biochem J* 401:13–28.

232. Ball S, Colleoni C, Cenci U, Raj JN, Tirtiaux C (2011) The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *J Exp Bot* 62:1775–801.

233. Fettke J et al. (2009) Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. *J Exp Bot* 60:2907–2922.

234. Viola R, Nyvall P, Pedersén M (2001) The unique features of starch metabolism in red algae. *Proc R Soc London B* 268:1417–1422.

235. Wilson W a et al. (2010) Regulation of glycogen metabolism in yeast and bacteria. *FEMS Microbiol Rev* 34:952–985.

236. Karsten U, West J (1993) Ecophysiological studies on six species of the mangrove red algal genus Caloglossa. *Funct Plant Biol* 20:729–739.

237. Karsten U et al. (1992) Mannitol in the red algal genus Caloglossa (Harvey) J. Agardh. *J Plant Physiol* 140:292–297.

238. Mostaert A, Karsten U, King R (1995) Inorganic ions and mannitol in the red alga Caloglossa leprieurii (Ceramiales, Rhodophyta): response to salinity change. *Phycologia* 34:501–507.

239. Eggert A, Raimund S, Daele K Van Den, Karsten U (2006) Biochemical characterization of mannitol metabolism in the unicellular red alga Dixoniella grisea (Rhodellophyceae). *Eur J Phycol* 41:405–413.

240. Eggert A, Raimund S, Michalik D, West J, Karsten U (2007) Ecophysiological performance of the primitive red alga Dixoniella grisea (Rhodellophyceae) to irradiance, temperature and salinity stress: growth responses and the osmotic role of mannitol. *Phycologia* 46:22–28.

241. Rousvoal S et al. (2011) Mannitol-1-phosphate dehydrogenase activity in Ectocarpus siliculosus, a key role for mannitol synthesis in brown algae. *Planta* 233:261–273.

242. Martone PT et al. (2009) Discovery of lignin in seaweed reveals convergent evolution of cell-wall architecture. *Curr Biol* 19:169–175.

243. Xu Z et al. (2009) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. *BMC Bioinformatics* 10 Suppl 1:S3.

244. Li X, Chapple C (2010) Understanding lignification: challenges beyond monolignol biosynthesis. *Plant Physiol* 154:449–452.

245. Weng J-K, Chapple C (2010) The origin and evolution of lignin biosynthesis. *New Phytol* 187:273–85.

246. Persson B, Hedlund J, Jörnvall H (2008) Medium- and short-chain dehydrogenase/reductase gene and protein families□: the MDR superfamily. *Cell Mol Life Sci* 65:3879–3894.

247. Ma Q-H (2007) Characterization of a cinnamoyl-CoA reductase that is associated with stem development in wheat. *J Exp Bot* 58:2011–2021.

248. Amsler CD (2009) *Algal chemical ecology* (Springer Verlag).

249. Emiliani G, Fondi M, Fani R, Gribaldo S (2009) A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. *Biol direct* 4:7.

250. La Barre S, Potin P, Leblanc C, Delage L (2010) The halogenated metabolism of brown algae (Phaeophyta), its biological importance and its environmental significance. *Mar Drugs* 8:988–1010.

251. Flodin C (1999) 4-Hydroxybenzoic acid: a likely precursor of 2,4,6-tribromophenol in Ulva lactuca. *Phytochemistry* 51:249–255.

252. Güven KC, Percot A, Sezik E (2010) Alkaloids in marine algae. *Mar drugs* 8:269–284.

253. Singh SP, Kumari S, Rastogi RP, Singh KL, Sinha RP (2008) Mycosporine-like amino acids (MAAs): chemical structure, biosynthesis and significance as UV-absorbing/screening compounds. *Indian J Exp Biol* 46:7–17.

254. Kräbs G, Watanabe M, Wiencke C (2004) A monochromatic action spectrum for the photoinduction of the UV-absorbing mycosporine-like amino acid shinorine in the red alga Chondrus crispus. *Photochem Photobiol* 79:515–519.

255. Balskus EP, Walsh CT (2010) The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. *Science* 329:1653–1656.

256. Pearson WR (2005) Phylogenies of glutathione transferase families. *Meth enzymol* 401:186–204.

257. Thom R, Dixon DP, Edwards R, Cole DJ, Lapthorn a J (2001) The structure of a zeta class glutathione S-transferase from Arabidopsis thaliana: characterisation of a GST with novel active-site architecture and a putative role in tyrosine catabolism. *J Mol Biol* 308:949–962.

258. Mueller L, Goodman C, Silady R (2000) AN9, a petunia glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein. *Plant Physiol* 123:1561–1570.

259. Gong H, Jiao Y, Hu W, Pua E (2005) Expression of glutathione-S-transferase and its role in plant growth and development in vivo and shoot morphogenesis in vitro. *Plant mol biol* 57:53–66.

260. Hervé C, De Franco P-O, Groisillier A, Tonon T, Boyen C (2008) New members of the glutathione transferase family discovered in red and brown algae. *Biochem J* 412:535–544.

261. Board PG (2011) The omega-class glutathione transferases: structure, function, and genetics. *Drug Metab Rev* 43:226–235.

262. Franco P-O De, Rousvoal S, Tonon T, Boyen C (2009) Whole genome survey of the glutathione transferase family in the brown algal model Ectocarpus siliculosus. *Marine genomics* 1:135–148.

263. Frova C (2006) Glutathione transferases in the genomics era: new insights and perspectives. *Biomol Eng* 23:149–169.

264. Morel F, Aninat C (2011) The glutathione transferase kappa family. *Drug metabolism reviews* 43:281–291. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21428694 [Accessed November 30, 2011].

265. Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B (2000) Membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG). A widespread protein superfamily. *Am J Respir Crit Care Med* 161:S20–524.

266. Bouarab K, Potin P, Correa J, Kloareg B (1999) Sulfated oligosaccharides mediate the interaction between a marine red alga and its green algal pathogenic endophyte. *Plant Cell* 11:1635–1650.

267. Boller T, Felix G (2009) A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu Rev Plant Biol* 60:379–406.

268. Dangl JL, Jones JDG (2001) Plant pathogens and integrated defence responses to infection. *Nature* 411:826–833.

269. Cuming AC (2009) Plant-pathogen interactions: a view from the evolutionary basement. *New Phytol* 183:237–239.

270. Kumar H, Kawai T, Akira S (2009) Pathogen recognition in the innate immune response. *Biochem J* 420:1–16.

271. Ting JP-Y, Williams KL (2005) The CATERPILLER family: an ancient family of immune/apoptotic proteins. *Clin Immunol* 115:33–37.

272. Han BW, Herrin BR, Cooper MD, Wilson I a (2008) Antigen recognition by variable lymphocyte receptors. *Science* 321:1834–1837.

273. Povelones M, Waterhouse RM, Kafatos FC, Christophides GK (2009) Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. *Science* 324:258–261.

274. Li D, Roberts R (2001) WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell Mol Life Sci* 58:2085– 2097.

275. Buist G, Steen A, Kok J, Kuipers OP (2008) LysM, a widely distributed protein motif for binding to (peptido)glycans. *Mol Microbiol* 68:838–47.

276. Radutoiu S et al. (2003) Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* 425:585–592.

277. Weinberger F, Friedlander M (2000) Response of Gracilaria conferta (Rhodophyta) to oligoagars results in defense against agar-degrading epiphytes. *J Phycol* 36:1079–1086.

278. Paul C, Pohnert G (2011) Production and role of volatile halogenated compounds from marine algae. *Nat Prod Rep* 28:186–195.

279. O'Brien PJ (2000) Peroxidases. *Chem-Biol Interact* 129:113–139.

280. Hervé C, Tonon T, Collén J, Corre E, Boyen C (2006) NADPH oxidases in Eukaryotes: red algae provide new hints! *Curr Genet* 49:190–204.

281. Bernroitner M, Zamocky M, Furtmüller PG, Peschek G a, Obinger C (2009) Occurrence, phylogeny, structure, and function of catalases and peroxidases in cyanobacteria. *J Exp Bot* 60:423–440.

282. Tamura K et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739.

283. Janssen DB, Dinkla IJT, Poelarends GJ, Terpstra P (2005) Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol* 7:1868–1882.

284. Chovancova E, Kosinski J, Bujnicki J, Damborsky J (2007) Phylogenetic analysis of haloalkane dehalogenases. *Proteins* 316:305–316.

285. Omura T (2010) Structural diversity of cytochrome P450 enzyme system. *J Biochem* 147:297–306.

286. Nelson DR (2011) Progress in tracing the evolutionary paths of cytochrome P450. *Biochim Biophys Acta* 1814:14–18.

287. Lepesheva GI, Waterman MR (2004) CYP51--the omnipotent P450. *Mol Cell Endocrinol* 215:165–170.

288. Kim J, Smith JJ, Tian L, Dellapenna D (2009) The evolution and function of carotenoid hydroxylases in Arabidopsis. *Plant Cell Physiol* 50:463–479.

289. Schubert N, García-Mendoza E, Pacheco-Ruiz I (2006) Carotenoid composition of marine red algae. *J P* 42:1208–1216.

290. Morikawa T, Saga H, Hashizume H, Ohta D (2009) CYP710A genes encoding sterol C22-desaturase in Physcomitrella patens as molecular evidence for the evolutionary conservation of a sterol biosynthetic pathway in plants. *Planta* 229:1311–1322.

291. Morikawa T et al. (2006) Cytochrome P450 CYP710A encodes the sterol C-22 desaturase in Arabidopsis and tomato. *Plant Cell* 18:1008–1022.

292. Davison IR, Pearson GA (2008) Stress tolerance in intertidal seaweeds. *J Phycol* 32:197–211.