nature COMMUNICATIONS

ARTICLE

Received 2 Nov 2012 | Accepted 26 Apr 2013 | Published 17 Jun 2013

DOI: 10.1038/ncomms2931 OPEN

Genome of the red alga *Porphyridium purpureum*

Debashish Bhattacharya[1], Dana C. Price[1], Cheong Xin Chan[2], Huan Qiu[1], Nicholas Rose[3], Steven Ball[4], Andreas P. M. Weber[5], Maria Cecilia Arias[4], Bernard Henrissat[6], Pedro M. Coutinho[6], Anagha Krishnan[7], Simone Zäuner[8], Shannon Morath[9], Frédérique Hilliou[10,11], Andrea Egizi[12], Marie-Mathilde Perrineau[1] & Hwan Su Yoon[13]

The limited knowledge we have about red algal genomes comes from the highly specialized extremophiles, Cyanidiophyceae. Here, we describe the first genome sequence from a mesophilic, unicellular red alga, *Porphyridium purpureum*. The 8,355 predicted genes in *P. purpureum*, hundreds of which are likely to be implicated in a history of horizontal gene transfer, reside in a genome of 19.7 Mbp with 235 spliceosomal introns. Analysis of light-harvesting complex proteins reveals a nuclear-encoded phycobiliprotein in the alga. We uncover a complex set of carbohydrate-active enzymes, identify the genes required for the methylerythritol phosphate pathway of isoprenoid biosynthesis, and find evidence of sexual reproduction. Analysis of the compact, function-rich genome of *P. purpureum* suggests that ancestral lineages of red algae acted as mediators of horizontal gene transfer between prokaryotes and photosynthetic eukaryotes, thereby significantly enriching genomes across the tree of photosynthetic life.

[1] Department of Ecology, Evolution and Natural Resources and Institute of Marine and Coastal Science, Rutgers University, New Brunswick, New Jersey 08901, USA. [2] Institute for Molecular Bioscience, and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Brisbane, Queensland 4072, Australia. [3] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey 08901, USA. [4] Unité de Glycobiologie Structurale et Fonctionnelle, UMR 8576 CNRS-USTL, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq Cedex, France. [5] Institute for Plant Biochemistry, Center of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, D-40225 Duesseldorf, Germany. [6] Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille University, CNRS UMR 7257, 13288 Marseille, France. [7] Department of Chemistry and Chemical Biology, Waksman Institute of Microbiology, Rutgers University, Piscataway, New Jersey 08854, USA. [8] Institute of Molecular Physiology and Biotechnology of Plants (IMBIO), University of Bonn, 53115 Bonn, Germany. [9] Department of Plant Biology and Pathology, Rutgers University, New Brunswick, New Jersey 08901, USA. [10] Institut National de la Recherche Agronomique, UMR 1355 Institut Sophia Agrobiotech, 0690 Sophia-Antipolis, France. [11] Centre National de la Recherche Scientifique, UMR 7254, Université de Nice Sophia Antipolis, 06903 Sophia-Antipolis, France. [12] Department of Entomology, Graduate Program in Ecology and Evolution, Rutgers University, New Brunswick, New Jersey 08901, USA. [13] Department of Biological Sciences, Sungkyunkwan University, Suwon 440-746, Korea. Correspondence and requests for materials should be addressed to D.B. (email: bhattacharya@aesop.rutgers.edu).

The red algae (Rhodophyta) form a monophyletic lineage comprising around 6,100 known species and 14,000 that are estimated to exist[1]. These taxa lack both flagella and centrioles during their life cycle[2]. Rhodophytes include many unicellular mesophilic lineages, the extremophilic Cyanidiophyceae (for example, *Cyanidioschyzon* and *Galdieria*) that live in hot springs such as in Yellowstone National Park, and economically important seaweeds such as *Gracilaria* and *Pyropia* (including *Porphyra*)[3] that are sources of agar and *nori*, respectively. Two aspects of red algal evolution are of central importance to understanding the evolution of eukaryotic phytoplankton. The first is their membership in the foundational lineage of photosynthetic eukaryotes, the supergroup Plantae (red algae, glaucophytes and green algae plus plants (also known as Archaeplastida)), whose ancestor putatively captured the canonical cyanobacterium-derived plastid[4,5]. The second is the subsequent spread of this organelle through secondary endosymbiosis to a diverse array of photosynthetic lineages collectively referred to as 'chromalveolates' (for example, diatoms, haptophytes and dinoflagellates[6,7]) that are dominant marine primary producers. For instance, red alga-derived plastids in diatoms are responsible for about 25–50% of organic carbon fixed annually in the world's oceans[8]. Secondary endosymbiosis also resulted in endosymbiotic gene transfer[9] (EGT) that relocated hundreds of former red algal genes to the nucleus of photosynthetic chromalveolates. Despite their central role in phytoplankton evolution, the genetic inventory of a unicellular lineage that may have been related to the putative donor of the 'red plastid' in chromalveolates is yet to be described. Existing rhodophyte complete genome sequences are from the thermoacidophiles *Cyanidioschyzon merolae*[10] and *Galdieria sulphuraria*[11] that have highly specialized and reduced genomes (for example, 16.5 Mbp with 5,331 protein-coding genes in *C. merolae*[10]) and from the red seaweed *Chondrus crispus*[12].

Here, we analyse the draft genome assembly from the unicellular, mesophilic red alga *Porphyridium purpureum* CCMP 1328 (referred to as *P. cruentum* in this culture collection) (Fig. 1a). This strain was isolated in 1957 from Eel Pond, Massachusetts and for this project was cultured in f/2 enriched seawater medium. We elucidate the role of horizontal gene transfer (HGT) in enriching the genome of *P. purpureum* and the extent of red algal gene sharing *via* EGT or HGT with chromalveolates and other taxa. We also analyse key aspects of red algal biology such as the evolution of proteins involved in light harvesting and metabolite transport and in starch, lipid and isoprenoid biosynthesis, and search for evidence of sexual reproduction in this species.

Results

Assembly and genome characteristics. The *P. purpureum* nuclear genome, based on the total length of the assembled contigs, was estimated to be 19.7 Mbp in size. This genome is intron-poor with 235 spliceosomal introns present in the 8,355 gene models (that is, 2.8% of genes contain introns) predicted using mRNA-seq data (see Methods for details). In comparison, 0.5% of genes in *C. merolae* contain introns[13]. The mapping of 52 million genome sequence reads to the consensus assembly showed the presence of 26,383 single-nucleotide polymorphisms in contigs with average coverage $>10\times$.

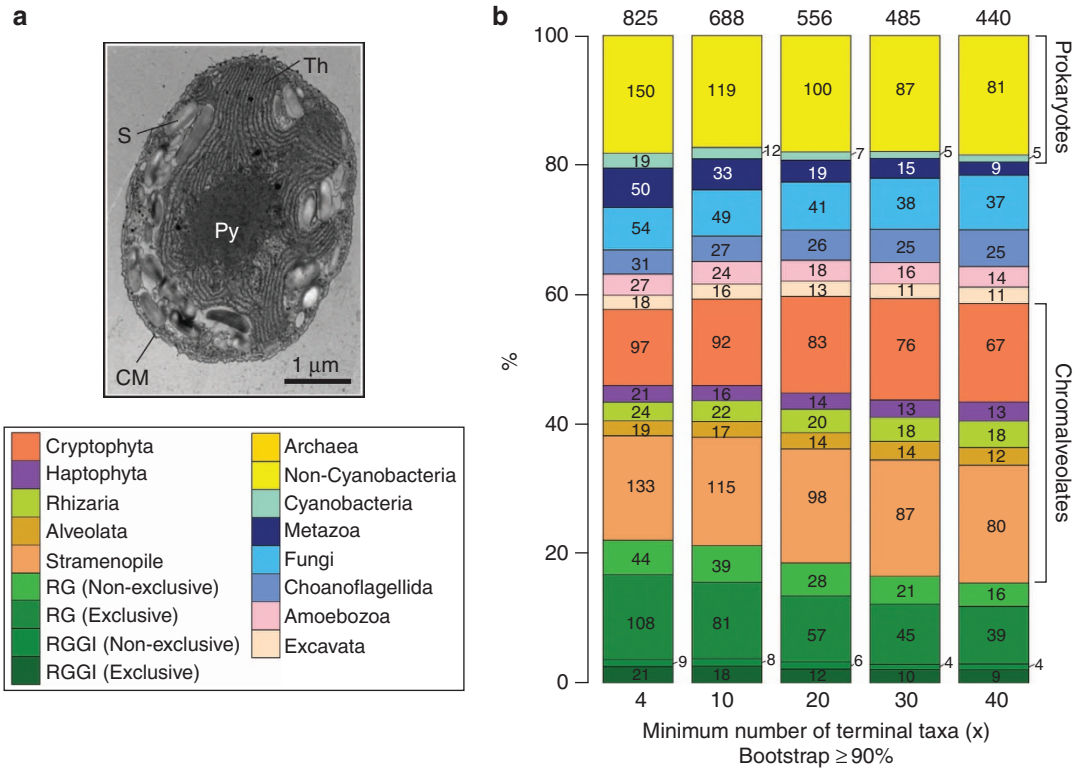

Figure 1 | Analysis of the *P. purpureum* genome. (**a**) Transmission electron microscopy image of a *P. purpureum* cell showing the central pyrenoid (Py), cell membrane (CM), starch granules (S) and plastid thylakoids (Th). (**b**) Percentage of single protein RAxML trees (raw numbers shown in the bars) that support the monophyly of *P. purpureum* (bootstrap ≥90%) solely with other Plantae members (exclusive), or in combination with non-Plantae taxa that interrupt this clade (non-exclusive). These latter groups of trees are primarily explained by red/green algal EGT into the nuclear genome of chromalveolates. For each of these algal lineages, the set of trees with different numbers of taxa (x) ≥4, ≥10, ≥20, ≥30 and ≥40 in a tree are shown. Each tree has ≥3 phyla. The Plantae-only groups are reds-greens-glaucophytes (RGGI) and reds-greens (RG).

Table 1 | Summary of over-represented gene ontology terms.

GO identifier	GO term	Type	FDR	*P*-value
0000325	Plant-type vacuole	C	5.70×10^{-2}	6.80×10^{-5}
0009536	Plastid	C	5.90×10^{-2}	9.90×10^{-5}
0044434	Chloroplast part	C	5.90×10^{-2}	1.10×10^{-4}
0003006	Developmental process involved in reproduction	P	5.90×10^{-4}	1.40×10^{-7}
0022414	Reproductive process	P	1.80×10^{-2}	8.30×10^{-6}
0000003	Reproduction	P	2.60×10^{-2}	2.00×10^{-5}
0048608	Reproductive structure development	P	2.60×10^{-2}	2.40×10^{-5}
0051704	Multi-organism process	P	5.90×10^{-2}	1.10×10^{-4}
0009793	Embryo development ending in seed dormancy	P	5.90×10^{-2}	1.20×10^{-4}
0010154	Fruit development	P	7.00×10^{-2}	1.80×10^{-4}
0048316	Seed development	P	7.00×10^{-2}	1.80×10^{-4}

C, cellular component; FDR, false discovery rate; GO, gene ontology; P, biological process.
Over- represented GO terms in Plantae-associated genes in comparison to all genes in *P. purpureum*, with the associated type shown as well as the C and P, FDR and *P*-value from Fisher's exact test.

Of these single-nucleotide polymorphisms, 94.6% were represented by two variants with an average representation of 57.5 and 42.4% for each variant, suggesting that the *P. purpureum* strain we sequenced was diploid. Previous studies have suggested both haploidy[14] and diploidy in this genus with the presence of 8–10 chromosomes[15]. Other analyses of the genome are shown in Supplementary Fig. S1, Supplementary Tables S1 and S2, and described in the Supplementary Information. The assembled genome and cDNA contigs, gene models, gene annotations, phylogenomic output and other material are available at http://cyanophora.rutgers.edu/porphyridium/.

Analysis of gene transfer. Phylogenomic analysis (Supplementary Table S3) resulted in 5,996 maximum likelihood protein trees (that is, 71.8% of the 8,355 total predicted proteins). At the stringent bootstrap threshold of $\geq$90% (or at $\geq$70 and $\geq$50%; Supplementary Fig. S2; lists of sorted phylogenies in Supplementary Tables S4–S6), about 20% of the trees supported the monophyly of red algae and other Plantae either by themselves (exclusive) or including other taxa (non-exclusive) that could have gained the Plantae gene through EGT or HGT (Fig. 1b; and Methods section). At the bootstrap support threshold of $\geq$90%, we found 440–825 trees (7–14% of total trees; 5–10% of total 8,355 proteins) that show a strong association between red algae with one other lineage. As shown in Supplementary Fig. S2, these numbers increase at the lower thresholds of $\geq$70% (766–1,310; 13–22% of total trees) and $\geq$50% (1,233–2,007; 21–33% of total trees). Of these trees (at bootstrap $\geq$90% in Fig. 1b), $\sim$40% showed sharing of red algal genes with different chromalveolate lineages either as nuclear genes or as cryptophyte nucleomorph-encoded[16] homologues (for example, 60S ribosomal protein L10A; contig 2315.2, Supplementary Data 1), and $\sim$20% with prokaryote lineages. In an independent assessment of proteins in *C. merolae*, the corresponding proportion of prokaryote-associated red algal genes in this species is smaller (12%; Supplementary Fig. S3). The majority of the *P. purpureum* trees ($>$80% in the analysis based on a bootstrap threshold $\geq$90%) show an evolutionary history that is, however, too complicated (for example, poor resolution of clades or currently too few taxa in the trees) to interpret with confidence. Taking trees containing $\geq$3 phyla and $\geq$20 terminal taxa as an indicator (middle bar; Fig. 1b), our results suggest that at least 453 genes (non-green portion; 5.4% of 8,355 proteins) in *P. purpureum* are impacted by E/HGT in their evolutionary history (773 genes at bootstrap $\geq$70%; 9.3% of 8,355). The complexity of these gene phylogenies is comparable to previous findings based on red algal transcriptome data[4,17].

Using a stringent criterion ($P \leq 0.05$ and false discovery rate ≤ 0.10), we observed no significant biases in the putative functions of *P. purpureum* proteins that are associated with non-Plantae taxa (non-green portion in Fig. 1b) in comparison with the annotated functions across the whole data set (Supplementary Fig. S4; Supplementary Tables S7–S9). Such functional biases were observed in an earlier study based solely on transcriptome data[18]. This discrepancy is likely explained by EST assembly artifacts in the earlier analysis that resulted in partial or mis-assemblies that inflated the total number of genes and their relative representation in the database. Nevertheless, among genes with phylogenies that show clear evidence of a common origin in Plantae, we found significant over-representation of gene ontology (GO) terms[19] related to reproductive and cell development (Table 1), compared with the overall gene set (see Methods). This finding suggests that these genes are more likely to be vertically inherited within Plantae, or alternatively, that radiation/innovation of these genes occurred after the divergence of this supergoup.

Given the complex evolutionary history of red algal genes found using phylogenomics combined with potential issues associated with the interpretation of results from automated pipelines[20], we conducted a manual analysis of a contig in our assembly to test the results regarding E/HGT. Contig 2035 (of size 91,179 bp) has average coverage of 623 $\times$ and encodes 42 genes, all with transcriptome evidence that show a paucity of spliceosomal introns (Supplementary Fig. S5). This *P. purpureum* genome region encodes proteins with a diversity of evolutionary histories. For example, one eukaryotic gene (contig 2035b.35) shows the expected monophyly of the lineages red algae, plants, Fungi and Metazoa in the eukaryotic tree of life (Fig. 2a), whereas the neighbouring gene (contig 2035b.36), although also of eukaryotic provenance shows a reticulate history that involves Viridiplantae and chromalveolates (Fig. 2b). Other genes on contig 2035 are apparently of bacterial origin with one that is shared by glaucophytes and chromalveolates (contig 2035b.17, Fig. 2c) and another that is shared only by red algae (contig 2035b.28, Fig. 2d). To gain a broader perspective on E/HGT, we also inspected phylogenies associated with all carbohydrate-active enzymes (CAZymes) identified in *P. purpureum* (see below for details). This revealed that of 107 CAZyme trees that could be interpreted with respect to prokaryotic or eukaryotic origin of the gene in *P. purpureum*, 41 genes (38%) had a prokaryotic provenance. Some of these genes were limited to red algae, whereas others were shared solely with Plantae and many (25 genes) had spread to chromalveolates.

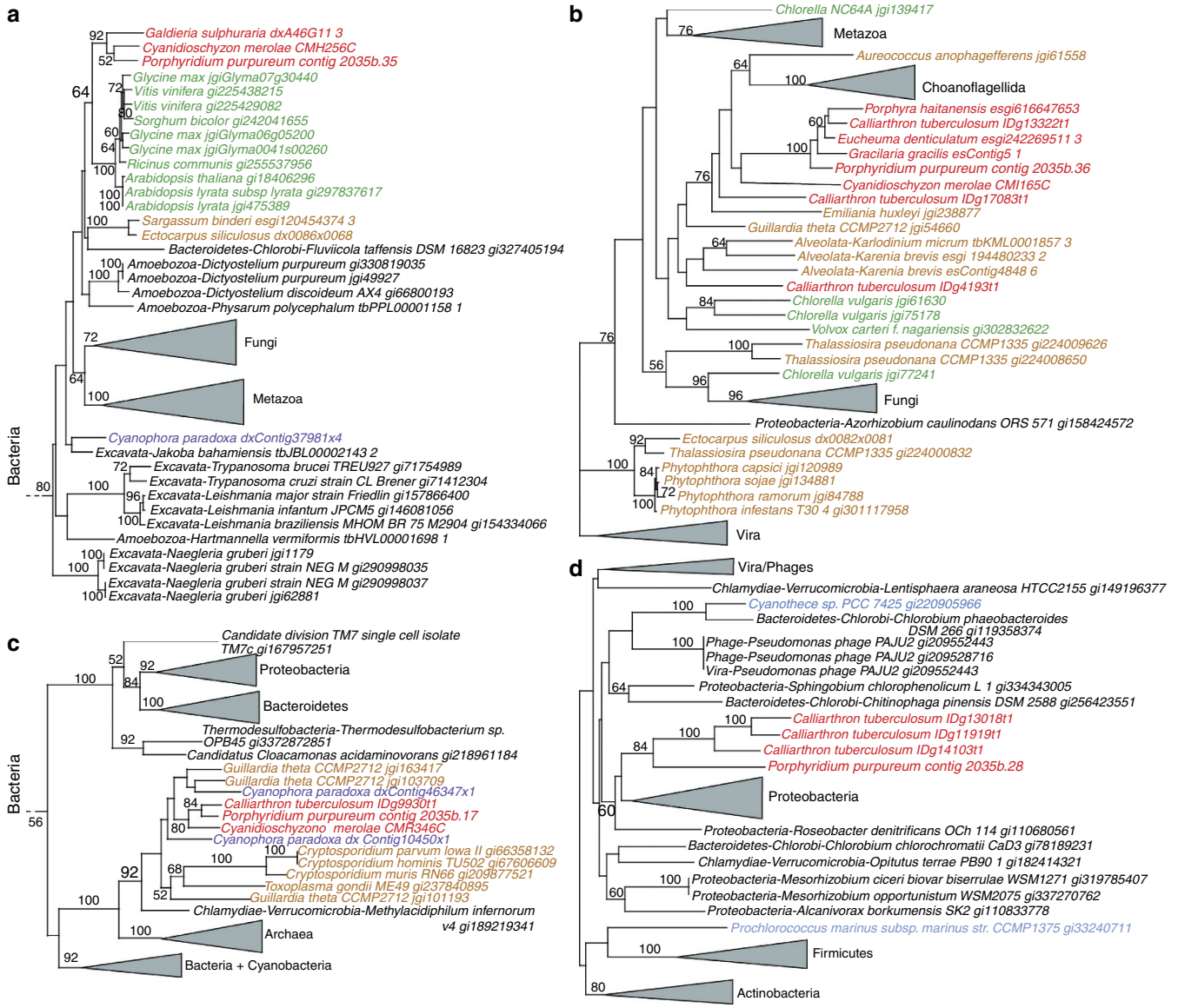

Figure 2 | Phylogenetic analysis of proteins on contig 2035b in *P. purpureum*. These are all RAxML trees (WAG + Γ + I + F model) with the results of 100 bootstrap replicates shown on the branches. (**a**) Tree inferred from a squalene monooxygenase-like protein involved in sterol biosynthesis that shows the expected monophyly of red algae and of plants within the eukaryote tree of life. (**b**) Tree inferred from a tyrosine kinase/lipopolysaccharide-modifying enzyme that shows a complex phylogenetic relationship between red algae and chromalveolates. (**c**) Tree inferred from a glycosyltransferase of bacterial origin that is consistent with the monophyly of red algae and glaucophytes and a shared history of the gene in these taxa with chromalveolates, potentially *via* secondary EGT. (**d**) Tree inferred from an unknown protein in the aminotransferase superfamily that is present only in red algae and originated through HGT presumably from a proteobacterial source. The unit of branch length in each tree is the number of substitutions per site. The GenBank GI and Joint Genome Institute (JGI) accession codes (where available) are shown after each taxon name.

Light-harvesting complex proteins in *P. purpureum*. Cyanobacteria and algae use light-harvesting proteins that contain photopigments to channel the energy gained from photons toward the chlorophyll-containing reaction centres of photosystems PSI and PSII. In many cyanobacteria, the only light-harvesting antenna proteins are phycobilisomes. In green algae and plants, all light-harvesting proteins are members of the light-harvesting complex (LHC) family[21]. In contrast, red algae are an intermediate between the two because they contain phycobilisomes, primarily associated with PSII, while also containing LHC proteins associated with PSI. *P. purpureum* was the first organism to have its phycobilisomes isolated and some of the phycobiliproteins (PBPs) and the LHC proteins have been characterized[22]. We identified seven LHC proteins in the *P. purpureum* genome (contigs 435.19, 491.7, 776.1, 2142.1, 2493.3, 3421.1 and 4406.7; see Supplementary Fig. S6A), which is consistent with previous analyses[21]. The sequence encoded on contig 491.7 was identical to the previously identified Lhcr1 from *P. purpureum* and the sequence encoded on contig 2493.3 was identical to Lhcr2 (refs 23,24). The other five LHC proteins were compared with the N-terminal fragment data from Tan *et al.*[23] (Table 2). Whereas these authors identified six unique protein bands, their sequencing results suggested that the 19.5 kDa band contained two unique proteins, which our results

confirm[23]. Therefore, all seven *P. purpureum* LHC proteins are expressed. Phylogenetic analysis of the LHC proteins showed that, as expected, the *P. purpureum* sequences grouped with other red algae and chromalveolates (Supplementary Fig. S6A).

Analysis of phycobilisome proteins showed *P. purpureum* contains alpha and beta subunits for phycoerythrin (PE) as well as several linker proteins (including L_{CM}, L_{RC}, L_C and 4 γ^{PE}) (see Supplementary Data 1). Surprisingly, we found a nuclear-encoded alpha-like PBP (Supplementary Fig. S6B). As PBP bands are not well resolved in SDS–PAGE gels[25], it is difficult to estimate from previous research the number of PBPs expressed in *P. purpureum*. The novel protein encoded on contig 2051.9 (252 aa in length) is associated in the tree with a number of cyanobacterial genes, one of which is a second α^{AP} (allophycocyanin) from *Gloeobacter violaceus* (gi37520823). Whereas this second AP-α protein was identified from the sequencing of the *G. violaceus* genome[26], analysis of the phycobilisomes did not identify any homologues in *P. purpureum* to the gi37520823 gene product. Examination of the *G. violaceus* genome shows this gene to be in an apparent operon with a bilin biosynthesis protein (MpeU-like protein) and a hypothetical protein, suggesting that if expressed it likely has a role in light harvesting (results not shown). The transcriptome data from *P. purpureum* shows extensive expression (813 mapped reads) of the novel PBP-encoding gene and examination of the protein sequence reveals a ca. 60 amino acid N-terminal extension when compared with cyanobacterial homologues. This extension appears to specify plastid targeting (scores of TargetP = 0.65, ChloroP = 0.67, Predotar = 0.26 and Wolfpsort = 14.0) and contains a phenylalanine near the N-terminus (that is, in this case, MLMFVF) that is typical for Plantae plastid targeted proteins[5]. Although lacking introns (as do most *P. purpureum* genes), these data suggest the red algal gene is a nuclear-encoded plastid-targeted PBP. A phylogenetic tree that includes all of the PBPs and core-membrane linkers (Supplementary Fig. S7) demonstrates the evolutionary relationship between the alpha and beta subunits and AP, PC, PE and the core-membrane linker and is consistent with previous data[27].

Table 2 | Comparison of molecular weights based on sequence prediction.

Contig	Weight (kDa) from Tan *et al.*[23]	Weight predicted from protein sequence (kDa)
2142.1	22.0	22.68
2493.3/Lhcr2	21.0	21.8
3421.1	19.5	22.81
435.19	19.5	23.56
4406.7	23.5	22.79
491.7/Lhcr1	23.0†	23.08
776.1	22.5‡	22.42

This was achieved using http://www.bioinformatics.org/sms/prot_mw.html
†The authors suggest that Lhcr1 (contig 491.7) is either the 23.0 or the 22.5 band. Here we assume, based on the predicted weight, that it was the 23.0 kDa band. ‡Based on the process of elimination and predicted weights, it is likely that this is the 22.5 kDa band.

Table 3 | CAZymes present in the *P. purpureum* genome.

Species	GH	GT	PL	CE	CBM
Porphyridium purpureum	31	83	0	2	40
Cyanidioschyzon merolae	21	61	0	2	16
Ostreococcus lucimarinus CCE9901	30	69	0	3	23
Micromonas sp. RCC299	41	85	0	3	30
Micromonas pusilla CCMP1545	37	77	0	2	29
Chlamydomonas reinhardtii	74	203	0	2	51
Bathycoccus prasinos RCC1005	51	172	0	5	25
Cyanophora paradoxa	84	128	3	2	24

CBM, carbohydrate-binding modules; CE, carbohydrate esterases; GH, glycoside hydrolases; GT, glycosyltransferases; PL, polysaccharide lyases.

Analysis of CAZymes and starch biosynthesis. A total of 116 putative CAZymes and 40 additional proteins containing putative carbohydrate-binding modules were identified in *P. purpureum* (Table 3) using the CAZy annotation pipeline[28]. These genes have a complex phylogenetic history (Supplementary Tables S10–S12). The genome of *P. purpureum* encodes 31 glycoside hydrolases (GH), 83 glycosyltransferases (GT) and two carbohydrate esterases, but similar to other unicellular rhodophytes and chlorophytes, lacks homologues of known polysaccharide lyases. *P. purpureum* encodes a larger number of GH and GT (114) genes when compared with *C. merolae* (82). Not surprisingly, the number of CAZy families is also 33% greater in *P. purpureum* (14 GH and 34 GT families) when compared with *C. merolae* (9 GH and 27 GT families), likely reflecting the complexity of *P. purpureum* cell-wall polysaccharides[29]. In comparison with the highly complex pathways of starch metabolism in green algae (over 30 genes) and the more diverse pathway in glaucophytes (22 genes), *P. purpureum* displays an unusually simple enzyme network consisting of 19 genes with many critical biosynthetic steps represented by single enzymes. This includes a single soluble starch synthase (GT5) that must have the remarkable property of priming polysaccharide synthesis, seeding the formation of novel granules and elongating the different size classes of chains present on amylopectin. These functions require a minimum of four distinct types of enzymes in Viridiplantae and several analogous glucan synthases in glaucophytes[30]. More exceptional and seemingly not shared with other starch accumulating red algae is the likely presence of a single isoamylase gene (contig 3410.5). Distinct isoamylase-like GH13 glycoside hydrolases are known to be involved both in starch catabolism and in the synthesis of the amylopectin crystalline structure that distinguishes starch from glycogen. This dual function seems to require several isoamylase genes in all Plantae examined thus far[30]. The presence of this single enzyme brings into question its involvement in both processes. However, absence of a second isoamylase correlates with the presence of three GH13 α-amylase candidate sequences. One of these (contig 4541.5) may have debranching activity[31]. This GH13 glycoside hydrolase is also found in some starch accumulating algae that apparently lack isoamylase genes and have a red algal plastid derived from secondary endosymbiosis. Of great interest here is the presence of another α-glucan synthase encoding gene (*GT5*), a granule-bound starch synthase that correlates with the presence of amylose in Porphyridiales (*Porphyridium* or *Rhodella*)[32], a unique feature of these unicellular algae when compared with other Rhodophyta. Other aspects of carbohydrate metabolism in *P. purpureum* are presented in the Supplementary Information.

Membrane transporters in *P. purpureum*. About 3.4% of the 8,355 predicted genes in the *P. purpureum* genome encode solute transporters, channels and pumps, which is similar to the corresponding numbers in green algae and land plants (Supplementary Data 2). Strikingly, in contrast to the currently known genomes of land plants and green and red algae, *P. purpureum* contains a putative sodium–potassium ATPase (encoded on contig 2281.11). Sodium–potassium ATPases import K^+ into cells and Na^+ out of cells at the expense of ATP, thereby keeping intracellular sodium concentrations low and potassium

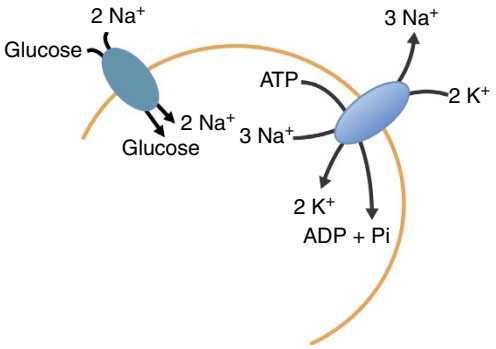

Figure 3 | Analysis of a transporter in *P. purpureum*. Schematic image showing the putative sodium–potassium ATPase and sodium:glucose cotransporter identified in the *P. purpureum* genome data.

concentrations high (Fig. 3). Thus, the pump contributes to maintaining cellular potassium and sodium homoeostasis in *P. purpureum*, which is exposed to high extracellular sodium concentrations in its environment. Furthermore, the sodium gradient across the plasma membrane that is set up by the sodium–potassium pump provides the driving force for secondary active sodium-coupled solute transporters. Indeed, the *P. purpureum* genome encodes several putatively sodium-driven transporters, such as a sodium:bicarbonate symporter (contig 2098.9), a sodium-dependent phosphate transport system (contig 2023.2), and a sodium:glucose cotransporter (Fig. 3), which is the first finding of a such a transporter in a photosynthetic organism. Interestingly, it has been recently demonstrated that *P. purpureum* can be grown to high cell densities in complete darkness with glucose as the sole carbon source[33]. In addition to the sodium:glucose symporter, four putative sucrose transporters were found in *P. purpureum* (contigs 2016.15, 2025.48, 2077.1, 3677.2). Among photosynthetic eukaryotes, sucrose transporters until now have only been reported from land plants and from the extremophilic red alga *Galdieria sulphuraria* that is able to grow heterotrophically on a range of different carbon sources[34]. It is thus possible that *P. purpureum* in addition to the monosaccharide glucose is also able to exploit disaccharides, such as sucrose, for heterotrophic growth, likely by a proton-coupled co-transport mechanism. Alternatively, because genes encoding the enzymes required for sucrose catabolism are not detectable in the *P. purpureum* genome (see Supplementary Methods), these transporters might be involved in the transport of osmotically active solutes, such as trehalose or mannosylglycerate.

Cytochrome P450 genes in *P. purpureum*. Cytochrome P450 (CYP) is one of the largest gene families with 5,100 sequences annotated in plants, 1,461 in vertebrates, 2,137 in insects, 2,960 in fungi, 1,042 in bacteria, 27 in Archaea and two in viruses[35]. Red algae are characterized by small genome sizes and therefore the species sequenced until now have a small number of *CYP* genes compared with Viridiplantae. The *P. purpureum* nuclear genome encodes 12 *CYP* genes (Supplementary Table S13), whereas *C. merolae* and *G. sulphuraria* contain 5 and 7 *CYP* genes, respectively. The *P. purpureum* CYPs contain all the conserved P450 domains (Supplementary Table S14), however, only three of these genes are orthologs of CYP clades already described: contig 3544.1 (CYP97), contig 2697.2 (CYP51) and contig 440.7 (CYP710). The remaining 9 *CYP* genes group with diverse eukaryotes in novel clades (Fig. 4a). Other aspects of CYP evolution in *P. purpureum* are presented in the Supplementary Information.

Glycerolipid biosynthesis. Membrane glycerolipid biosynthesis in *P. purpureum* follows the same path as is present in vascular plants and red algae such as *Porphyra*[17] or *C. merolae*[36] but with a few minor differences. In line with findings from other red algae, genes encoding the acetyl-CoA carboxylase subunits (*accA*, *accB* and *accD*) are encoded in the plastid, not the nuclear genome, as is the case for members of the Viridiplantae and their derived secondary endosymbionts. As also described for other red algae[17,36], *P. purpureum* lacks a plastid desaturation pathway including the gene encoding the soluble stearoyl acyl-ACP desaturase (*FAB2*) present in all members of the green lineage. This indicates that saturated C_{16}– and C_{18}– rather than monounsaturated fatty acids are exported from the plastid to the endoplasmic reticulum (ER). The first step in the biosynthesis of polyunsaturated fatty acids is catalysed by a Δ9-desaturase. The protein sequence encoded on contig 2306.6 contains an N-terminal cytochrome b_5 domain, distantly related versions of which are present in red algae and fungi, but absent in plants. *P. purpureum* is able to synthesize eicosapentaenoic acid (EPA, $20{:}5^{\Delta5,8,11,14,17}$). All required enzymes are encoded as single-copy genes in the nucleus and likely located in the ER with one exception. For the putative ω3-desaturase encoded on contig 2141.6 that catalyses the last step of EPA biosynthesis, a putative plastid transit peptide was identified and as expected, the N-terminus of the protein encoded phenylalanine residues (that is, MFAGF), indicating that ω3-desaturation takes place inside this organelle. This is in line with observations from labelling experiments[37] but contrasts with analyses of *Porphyra* EST data[17]. In contrast to findings in other red algae, some genes involved in glycerolipid synthesis appear to be present in multiple copies. Three candidate genes encoding monogalactosyl diacylglycerol transferase (MGD) and two genes encoding digalactosyl diacylglycerol transferase (DGD) homologues required for galactolipid synthesis were identified. The model plant *Arabidopsis thaliana* also contains three MGD and two DGD orthologs. Whereas MGD1 and DGD1 are constitutively active at the chloroplast inner envelope, MGD2, MGD3 and DGD2 are present at the chloroplast outer envelope following phosphate deprivation. Under these conditions, phospholipids in ER membranes are replaced by galactolipids. This phenomenon is also known from other organisms including some bacteria, but remains to be biochemically validated in *P. purpureum*.

Synthesis of sphingolipids. Sphingolipids are ubiquitous lipids that are highly enriched in the plasma membrane[38]. They are composed of an amino alcohol, the so-called sphingoid (long chain) base, amide-linked to a fatty acid. The long chain base can be further modified by the addition of complex sugar headgroups. Apart from their function as membrane building blocks, some sphingolipids also have signalling functions and are, for example, involved in cell cycle control and programmed cell death[39]. We identified all genes necessary for the formation of complex glycosphingolipids in the nuclear genome of *P. purpureum* (Fig. 4b and Supplementary Table S15). In addition to the findings in the *Porphyra* transcriptome[17], we also found a candidate for a long chain base-kinase (contig 4419.4). Unlike protein sequences from Viridiplantae, the candidate for the fatty acid α-hydroxylase (encoded on contig 522.16) contains an N-terminal cytochrome b_5 domain similar to that in fungal orthologs.

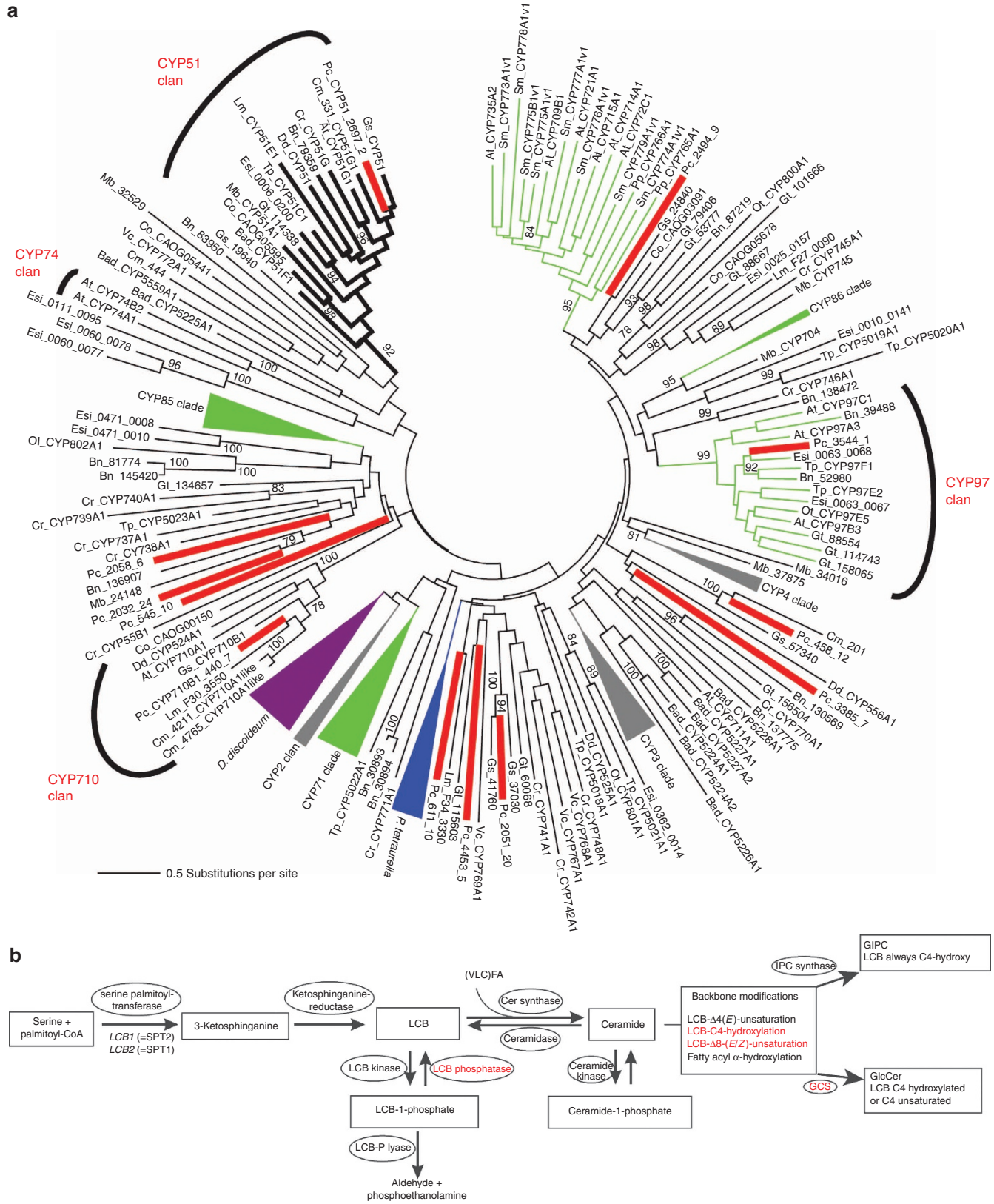

Figure 4 | Analysis of CYPs and sphingolipid metabolism in *P. purpureum*. (**a**) Maximum likelihood (RAxML; LG + Γ + I + F model) tree of CYP sequences from *P. purpureum* and other eukaryotes. Support for internal branches was assessed using 100 bootstrap replicates. The major known CYP clans are indicated. (**b**) Putative sphingolipid synthesis pathway in *P. purpureum* deduced from analyses of vascular plants (for example, *Arabidopsis*). Modifications without candidate genes in the *P. purpureum* draft assembly are indicated in red. Cer, ceramide; GCS, glucosylceramide synthase; GIPC, glucosyl inositol phosphoryl ceramide; GlcCer, glucosylceramide; IPC, inositol phosphoryl ceramide; LCB, long chain base; and VLCFA, very long chain fatty acid.

Table 4 | Identification of the meiotic toolkit genes in *P. purpureum*.

Gene	*P. purpureum* match	e-value	Top GenBank BLASTp hit
SPO11-3	Contig_2255.5	1×10^{-150}	*Ectocarpus siliculosus* topo VI subunit A
SPO11-2	Contig_3410.8	3×10^{-23}	*Oryza sativa* SPO11-2
DMC1	Contig_4438.25	2×10^{-95}	*Leishmania* DMC1
HOP1	Contig_667.2	5×10^{-23}	*Trichomonas vaginalis* meiotic synapsis protein
HOP2	Contig_4415.5	8×10^{-26}	*Glycine max* HOP2
MSH4	Contig_4404.10	2×10^{-31}	*Neosartorya fischeri* MSH4
MSH5	Contig_2159.1	2×10^{-41}	*Amphimedon queenslandica* MSH5
MND1	Contig_2121.23	5×10^{-43}	*Vitis vinifera* MND1
MER3	Contig_4401.1	6×10^{-59}	*Oryza sativa* meiotic crossover protein

Possible evidence of sexual reproduction in *P. purpureum*. Multicellular red algae within the Florideophyceae are well known for their complex triphasic life cycles. Sexual reproduction is also known in Bangiophyceae such as *Porphyra* that has a haploid gametophyte and a diploid 'conchocelis' stage[40]. Interestingly, one of the oldest multicellular eukaryotic fossils is believed to be the gametophytic stage of a Bangiophyceae, dating to rocks up to 1,200 million years old[41]. Outside of the Florideophyceae and Bangiophyceae, however, very little is known about sexual reproduction in red algae, particularly for unicellular forms such as *P. purpureum*, with most accounts suggesting the lack of sex in this lineage[42]. This is surprising because sexual reproduction is believed to be an ancient feature of eukaryotes[43,44] and relatively few lineages have completely lost this ability. To address the possibility of 'cryptic sex,' we searched the *P. purpureum* genome for the eight meiosis-specific proteins SPO11, HOP1, HOP2, MND1, DMC1, MSH4, MSH5, REC8 and MER3 (ref. 45). The presence of genes encoding a majority of these proteins is taken as evidence for meiosis and therefore sexual reproduction (for example, in *Giardia intestinalis*[43], *Trichomonas vaginalis*[45]).

Our search turned up evidence for eight of the targeted proteins (Table 4) including two paralogs of SPO11 (SPO11-2 and SPO11-3). All trees are shown in Supplementary Figs S8–S15. The presence of 8 out of 9 meiosis-specific proteins is consistent with (but does not prove) the maintenance of sexual reproduction in *P. purpureum*. Presence of all 'toolkit' proteins is not required for sexual reproduction. There are numerous examples of species known to be sexual that are missing one or more of these proteins. For example, *Drosophila melanogaster* lacks DMC1, HOP2, MND1, MSH4 and MSH5 (ref. 44). It would not be unusual if *P. purpureum* lacks REC8, because most protists, including known sexual species such as *Chlamydomonas*, also do not have this protein[44]. It is likely that in lineages missing REC8, RAD21 (the mitotic paralog) functions as well in meiosis. In the case of *P. purpureum*, the two identified contigs could be the result of a *RAD21* gene duplication, whereby one of the paralogs has assumed a meiotic function.

Discussion

Analysis of the first genome sequence from a mesophilic, unicellular red alga turned up several surprises. First, the genome is tightly packed with coding regions and is intron poor, reminiscent of a bacterial genome. There are very few large gene families, suggesting that unicellular red algal extremophiles and mesophiles may have undergone a phase of genome reduction, perhaps in an extremophilic common ancestor of all Rhodophyta. It is also interesting (if not surprising) that hundreds (5.4–9.3%) of the 8,355 *P. purpureum* genes show evidence of a reticulated evolutionary history and are likely to be implicated in E/HGT, with many more gene with phylogenies that cannot be readily interpreted using existing data. These data shed light on recent debates about the role of HGT in microbial eukaryote genome evolution, in particular whether phagotrophic and parasitic lineages are more likely to capture foreign genes than strict photoautotrophs[46,47]. In contrast to expectations, we demonstrate that anciently diverged relatives of the free-living, photosynthetic *P. purpureum* were mediators of HGT between prokaryotes and photosynthetic eukaryotes, *vis-à-vis* endosymbiosis. We have, however, no way of knowing for certain whether the red algal (or Plantae) ancestor was phagotrophic and therefore more prone to HGT, because evidence of HGT has also been described in non-phagotrophic lineages[48]. Regardless of the mechanism of ancient or more recent HGT, these data underline the fundamental importance of red algae to the evolution of eukaryotic plankton. Red algal plastids and red algal nuclear genes are now widespread in chlorophyll *c*-containing lineages such as diatoms, haptophytes and cryptophytes.

Other highlights of this genome include the finding of a nuclear-encoded PBP that apparently has a plastid function (that is, supported by the presence of a putative plastid-targeting signal), an unexpected diversity of *CYP* genes compared with green plants, and a simpler pathway for starch biosynthesis (see Supplementary Data 3 for a list of the plastid encoded genes). As red algae have the longest fossil record known among eukaryotes (1.2 billion years[41]) and the lineage contains up to 14,000 species[1], *P. purpureum* provides promise that a wealth of novel information awaits to be unearthed as additional genomes are completed from Rhodophyta.

Methods

Genome sequencing. A total of 7.4 Gbp of *P. purpureum* CCMP 1328 paired end (150 × 150 bp) genome data generated using two flow cell lanes in the Illumina GAIIx were assembled with the CLC Genomics Workbench tools (http://www.clcbio.com/products/clc-genomics-workbench/) into 4,770 contigs with a N50 of 20,296 bp. The contigs had average nucleotide coverage of 376 × (median = 56 ×) and totalled 19.7 Mbp, suggesting a genome size of ca. 20 Mbp. As with other genome studies, these estimates are subjected to further validation with better assembly of more sequence data. Thereafter, 4.1 Gbp of Illumina mRNA-seq data (150 bp x 150 bp reads) were used to train the ab-initio gene predictors (for details, see Price *et al.*[5]), resulting in a set of 8,355 weighted consensus gene structures that were used for downstream analyses.

Genome-wide analysis and phylogenomics. Repeat elements were identified using RepeatMasker (http://www.repeatmasker.org/) against the Repbase repetitive DNA elements library (version 2012-04-18) and a *de novo* repeat library elements generated using Repeat modeller (http://www.repeatmasker.org/). Duplicated genes from red algal genomes were identified using the method described in a previous study[49], except that we required aligned regions to cover >70% length of both proteins of the duplicated genes. The identity (I) cutoff for paralogous gene pairs was 30% if the total length of the aligned regions (L) was >150 amino acids. When L was ≤150 amino acids, then the minimal I was found[50] using the formula $I \geq 0.06 + 4.8L^{-0.32(1+\exp(-L/1{,}000))}$. Paralogous gene pairs were clustered into gene families. A gene was assigned to a gene family if it was paralogous to one or more of its members. Phylogenomic analysis was done based on protein sequence alignments as previously described[4,5] using MUSCLE 3.8.31 (ref. 51) (default settings) and RAxML 7.2.8 (ref. 52) (WAG + Γ model; 100 bootstrap replicates). Only trees that contained ≥3 phyla and a minimum number of taxa (N) ranging from 4 to 40 were considered, to minimize the impact of taxon sampling on this analysis. The screening for gene transfer is based on strongly supported clades

consisting of rhodophyte(s) and one other phylum, as described in an earlier study[53], where prokaryote (or when unavailable, opisthokont) sequences were used as outgroup in rooting the trees. A phylogenetic tree was considered to show non-exclusive sharing of a Plantae gene when a strongly supported clade (at the defined bootstrap support threshold) was found that comprised $\geq$90% Plantae taxa with the remainder being non-Plantae (for example, stramenopiles). This type of gene history implies putative E/HGT between the Plantae and non-Plantae taxa, but nevertheless provides support for Plantae monophyly. We did not interpret these trees as evidence of E/HGT.

Analysis of gene functional biases. Based on the annotated GO terms (http://geneontology.org/) using Blast2GO[19] (BLASTp E $\leq 10^{-5}$), we applied Fisher's exact test to assess potential functional biases in a given set of *P. purpureum* genes (test set; for example, genes associated with Plantae or non-Plantae taxa) in comparison with the annotated terms across the overall 8,355 genes (the reference set; see Supplementary Tables S7–S9), with correction for multiple testing[54]. An over- or under-representation of a GO term in the test set is statistically significant when $P \leq 0.05$ and false discovery rate ≤ 0.10.

Identification and bioinformatic analysis of CAZymes. All 8,355 putative ORFs encoded by the *P. purpureum* genome were submitted to analysis using the CAZy annotation pipeline in a two-step procedure of identification and annotation[28]. Sequences were subjected to BLASTp analysis against a library composed of the full-length proteins of the CAZy database. The hits with an *e*-value better than 0.1 were then subjected to a modular annotation procedure, that combines BLASTp against libraries of catalytic and carbohydrate-binding modules and family-specific profile Hidden Markov models (for details, see Price *et al.*[5]). The results were manually verified and completed with signal peptide, transmembrane and GPI predictions[55,56]. The fragmentary models and all models suspected of prediction errors were identified and flagged. Finally, a functional annotation step was carried out involving BLASTp comparisons against a library of modules derived from biochemically characterized enzymes[28].

Analysis of *CYP* genes. The *P. purpureum* protein models were searched by BLASTp analysis with CYP protein sequences representing the CYP2, CYP3 and CYP4 animal clades[57], and the CYP51,CYP71, CYP72, CYP74, CYP85, CYP86, CYP97, CYP710, CYP711, CYP727 and CYP746 plant clades[35]. P450 sequences from other species, including those from *Bigelowiella natans* and *Guillardia theta*[58], were identified in the same way and their CYP denomination were confirmed whenever possible using the standard P450 classification[59]. Their CYP sequences accession codes and provenances are summarized in Supplementary Table S13. The distantly related clade CYP727 and the divergent sequences CYP804A1 and CYP772A1 were excluded from the phylogenetic analysis. The CYP sequences were aligned using MUSCLE 3.8.31 (ref. 51) and a tree constructed using RAxML[52] (WAG + Γ + I model; 100 bootstrap replicates).

Microscopy. Transmission electron microscopy images were taken at the Center of Advanced Microscopy at Michigan State University, East Lansing, MI, USA on a JEOL100 CXII instrument (Japan Electron Optics Laboratories, Tokyo, Japan). For sample preparation, *P. purpureum* CCMP 1328 cells were processed as previously described[60].

References

1. Guiry, M. D. How many species of algae are there? *J. Phycol.* **48,** 1057–1063 (2012).
2. Graham, L. D. & Wilcox, L. W. *Algae* (Prentice-Hall, USA, 2000).
3. Blouin, N. A., Brodie, J. A., Grossman, A. C., Xu, P. & Brawley, S. H. Porphyra: a marine crop shaped by stress. *Trends Plant Sci.* **16,** 29–37 (2011).
4. Chan, C. X. *et al.* Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. *Curr. Biol.* **21,** 328–333 (2011).
5. Price, D. C. *et al. Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science.* **335,** 843–847 (2012).
6. Bhattacharya, D., Yoon, H. S. & Hackett, J. D. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays* **26,** 50–60 (2004).
7. Elias, M. & Archibald, J. M. Sizing up the genomic footprint of endosymbiosis. *Bioessays.* **31,** 1273–1279 (2009).
8. Falkowski, P. G. & Raven, J. A. *Aquatic photosynthesis* (Blackwell Science, 2007).
9. Martin, W. & Herrmann, R. G. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant. Physiol.* **118,** 9–17 (1998).
10. Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428,** 653–657 (2004).
11. Schönknecht, G. *et al.* Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science.* **339,** 1207–1210 (2013).
12. Collén, J. *et al.* Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc. Natl Acad. Sci. USA* **110,** 5247–5252 (2013).
13. Nozaki, H. *et al.* A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC. Biol.* **5,** 28 (2007).
14. Sivan, A., Thomas, J. C., Dubacq, J. P., van Moppes, D. & Arad, S. Protoplast fusion and genetic complementation of pigment mutations in the red microalga *Porphyridium* sp. *J. Phycol.* **31,** 167–172 (1995).
15. Schornstein, K. L. & Scott, J. Ultrastructure of cell division in the unicellular red alga *Porphyridium purpureum*. *Can. J. Bot.* **60,** 85–97 (1982).
16. Moore, C. E. & Archibald, J. M. Nucleomorph genomes. *Annu. Rev. Genet.* **43,** 251–264 (2009).
17. Chan, C. X. *et al. Porphyra* (Bangiophyceae) transcriptomes provide insights into red algal development and metabolism. *J. Phycol.* **48,** 1328–1342 (2012).
18. Chan, C. X. & Bhattacharya, D. Non-random sharing of Plantae genes. *Commun. Integr. Biol.* **4,** 361–363 (2011).
19. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).
20. Deschamps, P. & Moreira, D. Reevaluating the green contribution to diatom genomes. *Genome Biol. Evol.* **4,** 683–688 (2012).
21. Green, B. R. & Durnford, D. G. The chlorophyll-carotenoid proteins of oxygenic photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47,** p 685–714 (1996).
22. Grossman, A. R., Schaefer, M. R., Chiang, G. G. & Collier, J. L. The phycobilisome, a light-harvesting complex responsive to environmental conditions. *Microbiol. Rev.* **57,** 725–749 (1993).
23. Tan, S., Ducret, A., Aebersold, R. & Gantt, E. Red algal *LHC I* genes have similarities with both Chl a/b- and a/c-binding proteins: a 21 kDa polypeptide encoded by LhcaR2 is one of the six LHC I polypeptides. *Photosynth. Res.* **53,** 129–140 (1997).
24. Tan, S., Cunningham, F. X. & Gantt, E. LhcaR1 of the red alga *Porphyridium purpureum* encodes a polypeptide of the LHCI complex with seven potential chlorophyll a-binding residues that are conserved in most LHCs. *Plant Mol. Biol.* **33,** 157–167 (1997).
25. Redlinger, T. & Gantt, E. Phycobilisome structure of *Porphyridium purpureum*: polypeptide composition. *Plant Physiol.* **68,** 1375–1379 (1981).
26. Mendoza-Hernández, G., Pérez-Gómez, B., Krogmann, D. W., Gutiérrez-Cirlos, E. B. & Gómez-Lojero, C. Interactions of linker proteins with the phycobiliproteins in the phycobilisome substructures of *Gloeobacter violaceus*. *Photosynth. Res.* **106,** 247–261 (2010).
27. Apt, K. E., Collier, J. L. & Grossman, A. R. Evolution of the phycobiliproteins. *J. Mol. Biol.* **248,** 79–96 (1995).
28. Cantarel, B. L. *et al.* The carbohydrate-active enZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37,** D233–D238 (2009).
29. Arad, S. M. & Levy-Ontman, O. Red microalgal cell-wall polysaccharides: biotechnological aspects. *Curr. Opin. Biotechnol.* **21,** 358–364 (2010).
30. Ball, S. G., Colleoni, C., Cenci, U., Raj, J. N. & Tirtiaux, C. The evolution of the glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *J. Exp. Bot.* **62,** 1775–1801 (2011).
31. Choi, J. H. *et al.* Characterization of a novel debranching enzyme from *Nostoc punctiforme* possessing a high specificity for long branched chains. *Biochem. Biophys. Res. Commun.* **378,** 224–229 (2009).
32. Shimonaga, T. *et al.* Variation in storage α-polyglucans of red algae: amylose and semi-amylopectin types in *Porphyridium* and glycogen type in *Cyanidium*. *Mar. Biotechnol.* **9,** 192–202 (2007).
33. Oh, S. H. *et al.* Lipid production in *Porphyridium purpureum* grown under different culture conditions. *J. Biosci. Bioengineer.* **108,** 429–434 (2009).
34. Barbier, G. *et al.* Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae. *Plant. Physiol.* **137,** 460–474 (2005).
35. Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **66,** 194–211 (2011).
36. Sato, N. & Moriyama, T. Genomic and biochemical analysis of lipid biosynthesis in the unicellular rhodophyte *Cyanidioschyzon merolae*: lack of a plastidic desaturation pathway results in the coupled pathway of galactolipid synthesis. *Eukaryot. Cell.* **6,** 1006–1017 (2007).
37. Khozin, I., Adlerstein, D., Bigongo, C., Heimer, Y. M. & Cohen, Z. Elucidation of the biosynthesis of eicosapentaenoic acid in the microalga *Porphyridium purpureum* (II. Studies with radiolabeled precursors). *Plant Physiol.* **114,** 223–230 (1997).
38. Sperling, P., Franke, S., Lüthje, S. & Heinz, E. Are glucocerebrosides the predominant sphingolipids in plant plasma membranes? *Plant Physiol. Biochem.* **43,** 1031–1038 (2005).
39. Zäuner, S., Ternes, P. & Warnecke, D. Biosynthesis of sphingolipids in plants (and some of their functions). *Adv. Exp. Med. Biol.* **688,** 249–263 (2010).
40. Sahoo, D., Tang, X. & Yarish, C. *Porphyra*–the economic seaweed as a new experimental system. *Curr. Sci.* **83,** 1313–1316 (2002).

41. Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* **250,** 104–107 (1990).
42. Gantt, E. & Conti, S. F. The ultrastructure of *Porphyridium purpureum*. *J. Cell Biol.* **26,** 365–381 (1965).
43. Ramesh, M. A., Malik, S. -B. & Logsdon, Jr. J. M. A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Curr. Biol.* **15,** 185–191 (2005).
44. Schurko, A. M. & Logsdon, Jr. J. M. Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *Bioessays* **30,** 579–589 (2008).
45. Malik, S. -B., Pightling, A. W., Stefaniak, L. M., Schurko, A. M. & Logsdon Jr. J. M. An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PLoS ONE* **3,** e2879 (2008).
46. Andersson, J. O. Horizontal gene transfer between microbial eukaryotes. *Methods Mol. Biol.* **532,** 473–487 (2009).
47. Bhattacharya, D. *et al.* Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. *Sci. Rep* **2,** 356 (2012).
48. Marcet-Houben, M. & Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet.* **26,** 5–8 (2010).
49. Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P. & Li, W. H. Extent of gene duplication in the genomes of *Drosophila*, nematode and yeast. *Mol. Biol. Evol.* **19,** 256–262 (2002).
50. Rost, B. Twilight zone of protein sequence alignments. *Protein. Eng.* **12,** 85–94 (1999).
51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
52. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
53. Chan, C. X., Reyes-Prieto, A. & Bhattacharya, D. Red and green algal origin of diatom membrane transporters: insights into environmental adaptation and cell evolution. *PLoS ONE* **6,** e29138 (2011).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
55. Eisenhaber, B., Schneider, G., Wildpaner, M. & Eisenhaber, F. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *J. Mol. Biol* **337,** 243–253 (2004).
56. Käll, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–1036 (2004).
57. Feyereisen, R. Insect CYP genes and P450 enzymes. In *Insect Molecular Biology and Biochemistry* Gilbert, L. I. (ed.) 236–316 (Elsevier, Amsterdam, 2012).
58. Curtis, B. A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492,** 59–65 (2012).
59. Nelson, D. R. The cytochrome p450 homepage. *Hum. Genomics* **4,** 59–65 (2009).
60. Harris, E. H. *The Chlamydomonas sourcebook: a comprehensive guide to biology and laboratory use* (Academic Press, San Diego, 1989).

Acknowledgements

This research was supported by a grant from the National Science Foundation awarded to D.B. and H.S.Y. (0936884 and 1317114). D.B. acknowledges generous support from Rutgers University. N.R. was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate (NDSEG) program. H.S.Y. was supported by the Next-Generation BioGreen 21 Program (SSAC; 2013–PJ009525). We acknowledge the contributions of all students in the course 'Algal Genomics for Environmental and Algal Biofuel Research' at Rutgers University.

Author contributions

D.B. and H.S.Y. conceived the research in collaboration with D.C.P., C.X.C. and H.Q. D.C.P. led the genome assembly and gene prediction, C.X.C. did the phylogenomic and gene ontology term analyses, H.Q. did the repeat and gene family analyses, N.R. Analysed the light-harvesting complex data, S.B., M.C.A., B.H., P.M.C. and A.K. studied carbohydrate-active enzymes and starch metabolism, S.Z. studied lipid biosynthesis genes, A.P.M.W. Analysed the membrane transporters, S.M. studied isoprenoid biosynthesis, F.H. studied cytochrome P450 genes and A.E. and M.M.P. studied the evolution of meiosis-specific genes. D.B. wrote the paper in collaboration with the co-authors. D.B. supervised the project and is responsible for the final manuscript version.

Additional information

Accession codes: Coordinates for the draft *P. purpureum* genome and the Illumina mRNA-seq reads from this alga have been deposited at the NCBI Sequence Read Archive (SRA) with Project number BioProject ID# PRJNA189757, under the accession code SRP018727.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Bhattacharya, D. *et al.* Genome of the red alga *Porphyridium purpureum*. *Nat. Commun.* 4:1941 doi: 10.1038/ncomms2931 (2013).