Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems

France Denoeud[1,57], Olivier Godfroy[2,57], Corinne Cruaud[3,58], Svenja Heesch[4,49,58], Zofia Nehr[4,58], Nachida Tadrent[1,50,58], Arnaud Couloux[1,58], Loraine Brillet-Guéguen[5,6,58], Ludovic Delage[7,58], Dean Mckeown[6,58], Taizo Motomura[8,59], Duncan Sussfeld[9,1,59], Xiao Fan[10,11,59], Lisa Mazéas[2,59], Nicolas Terrapon[12,13,59], Josué Barrera-Redondo[14,59], Romy Petroll[14,59], Lauric Reynes[15,59], Seok-Wan Choi[16,59], Jihoon Jo[16,59], Kavitha Uthanumallian[17,59], Kenny Bogaert[18,51,59], Céline Duc[19,59], Pélagie Ratchinski[4,59], Agnieszka Lipinska[4,14,59], Benjamin Noel[1,59], Eleanor A. Murphy[20,21,59], Martin Lohr[22,59], Ananya Khatei[23,59], Pauline Hamon-Giraud[24,59], Christophe Vieira[25,59], Svea Sanja Akerfors[22], Shingo Akita[26], Komlan Avia[27], Yacine Badis[4], Tristan Barbeyron[2], Arnaud Belcour[24,52], Wahiba Berrabah[1], Samuel Blanquart[24], Ahlem Bouguerba-Collin[2], Trevor Bringloe[28], Rose Ann Cattolico[29], Alexandre Cormier[30], Helena Cruz de Carvalho[31,32], Romain Dallet[6], Olivier De Clerck[18], Ahmed Debit[31], Erwan Denis[1], Christophe Destombe[15], Erica Dinatale[14], Simon Dittami[7], Elodie Drula[12,13], Sylvain Faugeron[33], Jeanne Got[24], Louis Graf[16], Agnès Groisillier[19], Marie-Laure Guillemin[34,15], Lars Harms[35], William John Hatchett[36], Bernard Henrissat[37,38], Galice Hoarau[36], Chloé Jollivet[2], Alexander Jueterbock[23], Ehsan Kayal[6,53], Kazuhiro Kogame[39], Arthur Le Bars[6,40], Catherine Leblanc[7], Ronja Ley[22], Xi Liu[6], Pascal Jean Lopez[41], Philippe Lopez[9], Eric Manirakiza[19], Karine Massau[6], Stéphane Mauger[15,54], Laetitia Mest[4,55], Gurvan Michel[2], Catia Monteiro[7], Chikako Nagasato[8], Delphine Nègre[6,56], Eric Pelletier[1], Naomi Phillips[42], Philippe Potin[7], Stefan A. Rensing[43], Ellyn Rousselot[19], Sylvie Rousvoal[7], Declan Schroeder[44], Delphine Scornet[4], Anne Siegel[24], Leila Tirichine[19], Thierry Tonon[45], Klaus Valentin[35], Heroen Verbruggen[28], Florian Weinberger[46], Glen Wheeler[21], Hiroshi Kawai[47,60,*], Akira F. Peters[48,60,*], Hwan Su Yoon[16,60,*], Cécile Hervé[2,60,*], Naihao Ye[10,11,60,*], Eric Bapteste[9,60,*], Myriam Valero[15,60,*], Gabriel V. Markov[7,60,*], Erwan Corre[6,60,*], Susana M. Coelho[14,60,*], Patrick Wincker[1,60,*], Jean-Marc Aury[1,60,*] and J. Mark Cock[4,61,*]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, Evry, 91057, France

[2]Sorbonne Université, CNRS, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France

[3]Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, 91057, France

[4]Sorbonne Université, CNRS, Algal Genetics Group, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France

[5]CNRS, UMR 8227, Laboratory of Integrative Biology of Marine Models, Sorbonne Université, Station Biologique de Roscoff, Roscoff, France

[6]CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Roscoff, France

[7]Sorbonne Université, CNRS, UMR 8227, ABIE Team, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France

[8]Muroran Marine Station, Hokkaido University, Muroran, Japan

[9]Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205, Sorbonne Université, CNRS, Museum

[10]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China

[11]Laboratory for Marine Fisheries Science and Food Production Processes, Laoshan Laboratory, Qingdao, Shandong 266237, China

[12]Aix Marseille Univ, CNRS, UMR 7257 AFMB, Marseille, France

[13]INRAE, USC 1408 AFMB, Marseille, France

[14]Department of Algal Development and Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076, Tübingen, Germany

[15]IRL 3614, UMR 7144, DISEEM, CNRS, Sorbonne Université, Station Biologique de Roscoff, Roscoff, 29688, France

[16]Department of Biological Sciences, Sungkyunkwan University, Suwon, 16419, Republic of Korea

[17]University of Melbourne, Australia

[18]Phycology Research Group, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium

[19]Nantes Université, CNRS, US2B, UMR 6286, F-44000 Nantes, France

[20]University of Bristol, UK

[21]Marine Biological Association, Plymouth, UK

[22]Johannes Gutenberg University, Mainz, Germany

[23]Algal and Microbial Biotechnology Division, Nord University, Norway

[24]Univ Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France

[25]Research Institute for Basic Sciences, Jeju National University, Jeju 63243, Korea

[26]Faculty of Fisheries Sciences, Hokkaido University, Minato-cho 3-1-1, Hakodate, Hokkaido, 041-8611, Japan

[27]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France

[28]University of Melbourne, Australia

[29]University of Washington, USA

[30]Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France

[31]Institut de Biologie de l'ENS (IBENS), Département de Biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[32]Université Paris Est-Créteil (UPEC), Faculté des Sciences et Technologie, 61, avenue du Général De Gaulle 94000 Créteil, France

[33]Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile

[34]Núcleo Milenio MASH, Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile

[35]Alfred Wegener Institute (AWI), Bremenhaven

[36]Nord University, Norway

[37]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs Lyngby, Denmark

[38]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[39]Biological Sciences, Faculty of Science, Hokkaido University, Sapporo, 060-0810, Japan

[40]CNRS, Institut Français de Bioinformatique, IFB-core, Évry, France

[41]Centre National de la Recherche Scientifique, UMR BOREA MNHN/CNRS-8067/SU/IRD/Univ. Caen Normandie/Univ. Antilles, France

[42]Biology Department, Arcadia University, USA

[43]University of Freiburg, Germany

[44]Minneapolis-St Paul University, USA

[45]Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, United Kingdom

[46]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

[47]Kobe University Research Center for Inland Seas, Kobe, Japan

[48]Bezhin Rosko, 29250 Santeg, France

[49]Current address: Applied Ecology & Phycology, Institute for Biosciences, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany

[50]Current address: Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS-Université de Tours, Tours, 37200, France

[51]Current address: Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, 72076 Tübingen

[52]Current address: Univ. Grenoble Alpes, Inria, 38000, Grenoble, France

[53]Current address: Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, United States of America

[54]Current address: CNRS, La Rochelle Université, UMR7266, Littoral Environnement et Sociétés, La Rochelle, France

[55]Current address: Vegenov, Saint Pol de Léon, France

[56]Current address: Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, France

[57]These authors contributed equally (joint first author)

[58]These authors contributed equally (joint second author)

[59]These authors contributed key analyses

[60]Senior authors

[61]Lead contact

*Correspondence: cock@sb-roscoff.fr, jmaury@genoscope.cns.fr, pwincker@genoscope.cns.fr, susana.coelho@tuebingen.mpg.de, corre@sb-roscoff.fr, gabriel.markov@sb-roscoff.fr, valero@sb-roscoff.fr, epbapteste@gmail.com, yenh@ysfri.ac.cn, cherve@sb-roscoff.fr, hsyoon2011@skku.edu, akirapeters@gmail.com, kawai@kobe-u.ac.jp

SUMMARY

Brown seaweeds are keystone species of coastal ecosystems, often forming extensive underwater forests, that are under considerable threat from climate change. Despite their ecological and evolutionary importance, this phylogenetic group, which is very distantly related to animals and land plants, is still poorly characterised at the genome level. Here we analyse 60 new genomes that include species from all the major brown algal orders. Comparative analysis of these genomes indicated the occurrence of several major events coinciding approximately with the emergence of the brown algal lineage. These included marked gain of new orthologous gene families, enhanced protein domain rearrangement, horizontal gene transfer events and the acquisition of novel signalling molecules and metabolic pathways. The latter include enzymes implicated in processes emblematic of the brown algae such as biosynthesis of the alginate-based extracellular matrix, and halogen and phlorotannin biosynthesis. These early genomic innovations enabled the adaptation of brown algae to their intertidal habitats. The subsequent diversification of the brown algal orders tended to involve loss of gene families, and genomic features were identified that correlated with the emergence of differences in life cycle strategy, flagellar structure and halogen metabolism. We show that integration of large viral genomes has had a significant impact on brown algal genome content and propose that this process has persisted throughout the evolutionary history of the lineage. Finally, analysis of microevolutionary patterns within the genus *Ectocarpus* indicated that deep gene flow between species may be an important factor in genome evolution on more recent timescales.

INTRODUCTION

Brown algae correspond to the class Phaeophyceae within the stramenopile lineage. Many of these organisms are key components of extensive coastal ecosystems that provide high value ecosystem services, including the sequestration of several megatons of carbon per year globally, comparable to values reported for terrestrial forests[1]. The important role of kelp ecosystems is threatened by climate-related declines in seaweed populations worldwide[2]. However, appropriate conservation measures, coupled with the development of seaweed mariculture as a highly sustainable and low impact approach to food and biomass production, could potentially reverse this trend, allowing seaweeds to emerge as a tool with a significant role in mitigating the effects of climate change[3]. To attain this objective, it will be necessary to address important gaps in our knowledge of the biology and evolutionary history of the brown algal lineage. For example, these seaweeds remain poorly described in terms of genome sequencing due, in part, to difficulties with extracting nucleic acids. The Phaeoexplorer project has generated 60 new genomes corresponding to 40 species of brown algae and four close sister species (Table S1). The 40 brown algal species include representatives of 16 families, spanning all the major orders of the Phaeophyceae[4]. The sequenced species include brown algae that occur at different levels of the intertidal and subtidal, and are representative of the broad diversity of this group of seaweeds in terms of size, levels of multicellular complexity and life cycle structure (Figure 1). This extensive genomic dataset has been analysed to study the origin and evolution of key genomic features during the emergence and diversification of this important group of marine organisms.

RESULTS

1) In-depth sequencing of the Phaeophyceae lineage

Sequencing the genomes of brown algae has been hampered by the significant challenges involved, including inherent problems with growing brown algae, the presence of molecules that interfere with sequencing reactions and complex associations with microbial symbionts. Consequently, despite their modest genome sizes and accessible haploid stages, complete genome sequences have only been obtained for nine species to date[5–13]. Here we report work that has significantly expanded the genomic data available by sequencing and assembling 60 new genomes for 40 brown algae and four closely related species, covering 16 Phaeophyceae families (Figure 2A, Table S1). The Phaeoexplorer dataset (which can be accessed at https://phaeoexplorer.sb-roscoff.fr) includes a set of good quality assemblies complemented with lower quality draft assemblies that have been used to sample additional strains from across the lineage (Figures 2A and S1A, Table S1). Seventeen of the new

genomes were sequenced using long-read technology, primarily Oxford Nanopore (Table S1). The remainder were assembled using short-read data and, although these assemblies consist of shorter contigs, they provide a dense coverage of the Phaeophyceae lineage and have facilitated gene-level comparative analysis. Analyses of genome content and evolution have, however, focused principally on a set of 21 good quality reference genomes, which include four previously published genomes (Table S1). In addition to broadly sampling brown algal taxa, the study included extensive sampling of 12 species from the genus *Ectocarpus*, providing a detailed survey of diversity and microevolution events within this genus (Table S1). Finally, in addition to the nuclear genomes, the Phaeoexplorer data allowed the assembly and annotation of 34 sets of plastid and mitochondrial genomes (Table S2).

The brown algal genomes exhibit marked differences in assembly size and GC content (Figure S1B). Comparison of structural features across the set of nuclear genome assemblies indicated not only that that the cumulative lengths of intergenic regions and transposable elements were correlated with genome size but that there was also a correlation with cumulative intron size (Figures 2B and S1C, Table S1). An earlier analysis of the *Ectocarpus* species 7 (at that time classified as *Ectocarpus siliculosus*) genome found a strong tendency for adjacent genes to be located on opposite strands of the chromosome, a feature that is more typical of very small, compact genomes[5]. Analysis of the Phaeoexplorer genomes indicated that this characteristic is common to the whole FDI (Fucophycideae/Dictyotales/Ishigeales) clade but the tendency is strongest in the Ectocarpales, where the proportion of alternating gene pairs can be greater than 60% (Figure 2B). Another feature that is shared by all brown algae, plus some of the closely-related sister taxa, is that intergenic intervals between divergently transcribed gene pairs tend to be shorter than those between tandem gene pairs (Figure 2B). Again, this is a feature that was previously observed in *Ectocarpus* species 7 where it is associated with the presence of common nucleosome-depleted regions and histone modification peaks at the transcription start sites of the divergently transcribed genes[14]. Clusters of tandemly-repeated genes are more numerous in the Phaeophyceae than in sister clades (Figure 2B). These clusters only represent modest proportions of the total gene sets (around 6% for the Ectocarpales and 10-12% for the Fucales, for example) but the amplified families may have played important roles in brown algal adaptation and diversification.

On average, brown algal genes tend to be more intron-rich than those of the other stramenopile groups[15], including closely-related sister species (Figure 2B), with the notable exception of *Chrysoparadoxa australica*, whose genes contain many short, C/T-rich introns (Figure 2B). The vast majority of splice sites in Phaeophyceae are of the canonical GT-AG type and no signatures of minor

U12 introns were identified in intronic sequences, in accordance with a previous study[16]. Consistent with this observation, an analysis of spliceosome proteins using a recent reference dataset[17] found that, with the exception of ZCRB1, none of the 12 minor spliceosome genes were present in brown algal genomes (Table S3), indicating that the minor intron spliceosomal associated machinery was lost in the Phaeophyceae. As far as the Sm/Lsm spliceosome is concerned, this analysis identified lineage-specific gene duplications and loss of some components[18] (Figures S2A and S2B).

An analysis of intron conservation profiles using a set of single-copy orthologous proteins conserved across Phaeophyceae and close sister groups confirmed that a high proportion of *C. australica* introns are unique to that species, suggesting that a burst of intron gain occurred during its evolution (Figures 2C, S2C, S2D and S2E). In Phaeophyceae species, more than 90% of introns are shared with at least one of the sister clades, indicating that most introns are ancient. Very few of the introns that have been gained during Phaeophyceae evolution are unique to one order or species, with most being conserved across all brown algal lineages. This observation indicates that a phase of intron acquisition occurred at around the time of the emergence of the Phaeophyceae (principally before the divergence of the Phaeophyceae and *Schizocladia ischiensis*) followed by a period of intron stability up to the present day (Figure 2C). The phase of intron acquisition was possibly linked to increased multicellular complexity (Figure 1) and concomitant decreases in effective population sizes[19]. Increased intron density may have facilitated some of the genome-wide tendencies that will be described below, such as increased reorganisation of composite genes for example. Even if the mechanisms that initially led to an increased prevalence of introns are likely to have been neutral[19], this increase will have provided scope for alternative splicing and domain reorganisation that may have played an important role in the acquisition of developmental complexity[17,20–22].

Long non-coding RNAs (lncRNAs) play key regulatory roles in diverse eukaryotic taxa[23,24]. This class of gene has been characterised in *Ectocarpus* species 7 and *Saccharina japonica* and there is evidence that some of these genes may play regulatory roles during development and life cycle progression in brown algae[25–27]. Eleven transcriptomes (Table S4) were searched for lncRNAs using a meta-classifier, votingLNC, that integrates information from ten lncRNA-detection programs. This analysis indicated that brown algal genomes contain about twice as many lncRNA genes as protein-coding genes (Figure S2F). The lncRNA transcripts tend to be shorter than mRNAs and to be less GC-rich (Figure S2F).

2) Evolutionary events at the origin of the brown algal lineage

Marked gene gain and gene rearrangement during the emergence of the Phaeophyceae lineage

The brown algae are a lineage of complex multicellular organisms that emerged from within a group of stramenopile taxa that are either unicellular or have very simple filamentous multicellular thalli (Figure 1). To investigate events that occurred during the emergence of the brown algal lineage, we first carried out a series of genome-wide analyses aimed at identifying broad trends in genome evolution (Figure 3). This comparative analysis of the Phaeoexplorer dataset identified several marked changes in genome structure and content coinciding approximately with emergence of the Phaeophyceae. Dollo analysis of gain and loss of orthogroups (i.e. gene families) indicated marked gains during early brown algal evolution followed by a broad tendency to lose orthogroups later as the different brown algal orders diversified (Figures 3B and S3A). The orthogroups gained during early evolution of the brown algal lineage are enriched in genes related to signalling, transcription and carbohydrate metabolism (Figures 3B and S3B). Similarly, a phylostratigraphy analysis indicated that 29.6% of brown algal genes could not be traced back to ancestors outside the Phaeophyceae, with the majority of gene founder events occurring early during the emergence of the brown algae (Figure 3A, Table S5), again indicating a burst of gene birth during the emergence of this lineage. Interestingly, many of the genes acquired at the origin of the Phaeophyceae or later encode secreted or membrane proteins suggesting possible roles in cell-cell communication that may have been important for the emergence of complex multicellularity or perhaps as components of defence mechanisms (Figure 3A). In addition, comparison of the Phaeophyceae with the four close sister species identified 180 orthogroups that had significantly expanded in brown algae, with predicted functions predominantly related to metabolism, particularly carbohydrate metabolism, and signal transduction (Figure 3C, Table S6).

Analysis of composite genes (domain fusions and fissions) indicated that domain reorganisation was also prevalent during the early stages of brown algal emergence, with gene fusion events being slightly more common than gene fissions (Figure 3E), representing about 6% to 7% of brown algal gene complements. Composite genes tended to be retained at a higher frequency than non-composite genes during the most recent phase of brown algal evolution, but a larger proportion of these genes had unknown functions than for non-composite genes (Figure S4A). Functions enriched in the composite gene set include carbohydrate metabolism, transcription, cell wall biogenesis and signal transduction (Figure 3E). Figure S4B shows an example of a novel domain association that is predicted to have arisen just prior to the emergence of the brown algal lineage.

A potential source of novel genetic information during the emergence of the brown algae could have been horizontal gene transfer. A phylogeny-based search for genes potentially derived from horizontal gene transfer (HGT) events indicated that these genes constitute about 1% of brown algal gene catalogues and are predicted to have been principally acquired from bacterial genomes (Figures 3D and S4C). The proportion of class-specific HGT events compared to more ancient HGT events was greater for the brown algae (33.5%) than for the sister taxa Xanthophyceae (*Tribonema minus*) and Raphidophyceae (*Heterosigma akashiwo*; mean of 17.1% for the two sister taxa, Wilcoxon *p*-value 0.02105), indicating that higher levels of HGT occurred during the emergence of the brown algae than for the sister taxa (Figure 3D). This difference is consistent with the results of the phylostratigraphy analysis, which indicated a peak of gene founder events directly following the emergence of the Phaeophyceae lineage. Genes that are predicted to be derived from HGT exhibit different ranges of codon usage bias and gene feature lengths to non-HGT-derived genes (Figure S4D), and have diverse predicted functions, with carbohydrate transport and metabolism and cell wall/membrane/envelope biogenesis being the most prevalent (Figure S4E).

AuCoMe[28] was used to provide an overview of metabolic pathways across the brown algae by inferring genome-scale metabolic networks for all the species in the Phaeoexplorer dataset (https://gem-aureme.genouest.org/phaeogem/; Figures S5A and S5B) and also for a subset of long-read sequenced species (https://gem-aureme.genouest.org/16bestgem/; Figures S5C and S5D). Comparison of the brown algal metabolic networks with those of the sister taxa allowed the identification of enzymatic activities that have been gained or lost in the former compared with the latter (Figures S5E, S5F and S5G). The most robust set of core reactions found mainly in brown algae is a set of 23 reactions, including polyamine oxidases, carotenoid and indole-3-pyruvate oxygenases, a methyladenine demethylase, an endo-1,3-alpha-glucosidase, a xylosyl transferase, a ribulosamine kinase, several galactosylceramide sulfotransferases and phosphopentomutases (Table S7). This set of robust brown algal core reactions is consistent with the Dollo and phylostratigraphic analyses, which both indicated that genes involved in carbohydrate metabolism and signalling were prevalent among the genes gained by the brown algal lineage.

Acquisition of key metabolic pathways during the emergence of the Phaeophyceae lineage

Several emblematic features of brown algae, such as their characteristic cell walls and their novel halogen and phlorotannin metabolisms, can be linked to genomic events that occurred during the early stages of brown algal emergence. For example, large complements of carbohydrate-active enzyme (CAZYme) genes (237 genes on average) were found in all brown algal orders and in its sister

taxon *S. ischiensis* but this class of gene was less abundant in the more distantly related unicellular alga *H. akashiwo* (Figures 4A, S6A and S6B, Tables S8 and S9). The evolutionary history of carbohydrate metabolism gene families was investigated by combining information from the genome-wide analyses of gene gain/loss, HGT and gene family amplification (Figure 4B). This analysis indicated that several key genes and gene families (ManC5-E, GT23 and PL41) were acquired by the common ancestor of brown algae and *S. ischiensis*, with strong evidence in some cases that this occurred via HGT (PL41). Moreover, marked amplifications were detected for several families (AA15, ManC5-E, GH114, GT23 and PL41), indicating that both gain and amplification of gene families played important roles in the emergence of the brown algal carbohydrate metabolism gene set. Alginate is a major component of brown algal cell walls and it plays an important role in conferring resistance to the biomechanical effect of wave action[29]. It is therefore interesting that mannuronan C5 epimerase (ManC5-E), an enzyme whose action modulates the rigidity of the alginate polymer, appears to have been acquired very early, by the common ancestor of brown algae and *S. ischiensis* (Figures 4B and 4C). The acquisition of ManC5-E, together with other alginate pathway enzymes such as PL41 (Figures 4A, 4B and 4D), was probably an important evolutionary step, enabling the emergence of large, resilient substrate-anchored multicellular organisms in the hostile coastal environment (Figure 1). The sulphatases that remove sulphate groups from sulphated polysaccharides[30] appear to be of more ancient origin, with homologues being present in other stramenopile lineages (Table S10).

Brown algae are able to produce a broad range of halogenated molecules and these molecules are thought to play important roles in multiple processes including defence, adhesion and cell wall modification[31]. Vanadium-dependent haloperoxidases (vHPOs) are a central component of these reactions and all three classes of brown algal vHPO (algal types I and II and bacterial-type[32–34]) appear to have been acquired early during the emergence of the Phaeophyceae (Figures 4A, S6C and S6D, Tables S11 and S12). Closely-related sister species do not possess any of these three types of haloperoxidase, with the exception of the closest sister taxon, *S. ischiensis*, which possesses three algal-type haloperoxidase genes (Figure 4A). These sequences appear to be intermediate forms as they are equally distant, phylogenetically, from the class I and class II algal-types (Figures 4A, S6C and S6D). Algal type I and II vHPO genes probably diverged from an intermediate-type ancestral gene similar to the *S. ischiensis* genes early during Phaeophyceae evolution. It is likely that the initial acquisitions of algal- and bacterial-type vHPOs represented independent events although the presence of probable vestiges of bacterial-type vHPO genes in *S. ischiensis* means that it is not possible to rule out acquisition of both types of vHPO through a single event.

Polyketides are a group of secondary metabolites that exhibit a wide range of bioactivities. Brown algae possess three classes of type III polyketide synthase (Figures 4A, 5A and S6E, Tables S13). The first class (PKS1) was acquired early in stramenopile evolution, after divergence from the oomycetes but before the divergence of Diatomista and Chrysista (Figure 5A). Acquisition of PKS1 probably involved horizontal transfer from a bacterium. Distance analysis of a maximum likelihood phylogenetic tree (Figure 5A) indicated that PKS1 was more likely to have been acquired from a cyanobacterium than from an actinobacterium, as was previously proposed[35]. The second class of type III polyketide synthase (PKS2) probably arose via a duplication of PKS1 just prior to the divergence of the brown algal lineage from the Phaeothamniophyceae (Figure 5A). *Undaria pinnatifida* is the only species where a duplication of PKS2 was observed. The PKS3 class appeared much more recently, within the Ectocarpales during the diversification of this brown algal order, again probably via a duplication of PKS1 (Figure 5A). PKS3 genes were only found in two families, the Ectocarpaceae and the Scytosiphonaceae. The enzymatic activities and functions of PKS2 and PKS3 remain to be determined but analyses of the functions of recombinant PKS1 proteins have indicated that different proteins may have different activities leading to the production of different metabolites[35–37]. For example, the *Ectocarpus* species 7 PKS1 enzyme produces phloroglucinol from malonyl-CoA alone or malonyl-CoA plus acetyl-CoA whereas the PKS1 enzymes of *Sargassum binderi* and *Sargassum fusiforme* generate 4-hydroxy-6-methyl-2-pyrone. It is therefore interesting that the PKS1 family has expanded in some members of the Sargassaceae, with a maximum of five PKS1 genes detected in *S. fusiforme* (Figures 4A and 5A). These expansions may be associated with diversification of PKS1 biochemical function. Many of the stramenopile PKS III genes encode proteins with signal peptides or signal anchors (Figure S6E). For the brown algae, this feature is consistent with the cellular production site of phlorotannins (which are derived from phloroglucinol) and the observed transport of these compounds by physodes, secretory vesicles characteristic of brown algae[38]. More specifically, polymerization of phloroglucinol gives rise to higher molecular weight compounds with C-C and/or C-O bonds, such as fucols, phloroethols or fucophloroethols, which constitute the phlorotannin family of compounds and are unique to brown algae. Phlorotannins are a common feature of all brown algae with the exception of a few Sargassaceae such as *Cystoseira tamariscifolia* and *Ericaria selaginoides*[39,40]. Cross-linking of phlorotannins, embedded within other brown algal cell wall compounds such as alginates, has been demonstrated *in vitro* through the action of vHPOs[41–43] and indirectly suggested by *in vivo* observations colocalising vHPOs with physode fusions at the cell periphery[44,45]. Consequently, vHPOs are good candidates for the enzymes that cross-link phlorotannins and other compounds, perhaps even for the formation of covalent bonds between phloroglucinol monomers and oligomers, which could occur via activation of aromatic rings through halogenation. These observations suggest that the acquisition of vHPOs by the common

ancestor of brown algae and *S. ischiensis*, together with existing PKS enzymes, triggered the emergence of new metabolic pathways leading to the production of the phlorotannin molecules characteristic of the Phaeophyceae lineage.

In brown algae, the output of the Calvin cycle is used to produce the carbon storage molecule mannitol via the mannitol cycle[46]. The genes that encode the four enzymes of the mannitol cycle are not only conserved in all brown algae but were also found in sister taxa as distant as *H. akashiwo* (Table S14), indicating that the mannitol cycle was acquired before brown algae diverged from closely-related stramenopile taxa. The mannitol cycle genes include a linked pair of *M1PDH3* and *M1Pase1* genes which may therefore represent the ancestral core of this metabolic cycle. These pairs of genes are head-to-head and adjacent in brown algae with an intergenic region ranging from about 500 bp (Ectocarpales) to 10 kbp (*Fucus serratus*). The organisation of this gene pair differs in sister groups with, for example, a tandem arrangement and an intervening gene in *S. ischiensis* and two *M1PDH3* / *M1Pase1* gene pairs in *H. akashiwo* with tail-to-head and head-to-tail arrangements and long (150-200 kbp) intergenic regions.

Evolution of novel cellular and signalling components during the emergence of the Phaeophyceae lineage

Most stramenopiles possess motile cells with two heteromorphic flagella, although there are some exceptions such as the male gametes of pennate diatoms. Searches for the presence of flagellar proteins, based on proteomic inventories of *Colpomenia bullosa* anterior and posterior flagella[47], indicated that highly-conserved proteins such as the mastigoneme proteins MAS1, MAS2 and MAS3, a flagellar titin protein, which is thought to be involved in connecting mastigonemes to the flagellar axoneme, and flagellar creatine kinase are present in all brown algae and sister taxa (Figures 4A and 5B, Table S15). In contrast, the helmchrome protein, which is thought to be involved in light reception and zoid phototaxis[47], was only found in brown algae and some closely-related sister taxa, the most distant being *C. australica* (Figure 4A).

An analysis of transcription-associated proteins indicated that the range of gene families and the relative sizes of each gene family are relatively constant across the brown algae and their close sister species (Figures 4A, 5C, S7A and S7B, Table S16), suggesting that the emergence of the brown algae did not involve major changes to the transcription machinery. For example, large families of C2H2, Zn_clus, GNAT and SWI_SNF_SNF2 genes are a common feature in both brown algal and sister taxa genomes.

Brown algal genomes contain homologues of components of several phytohormone biosynthetic pathways, including indole-3-acetic acid (the auxin IAA), cytokinins, brassinosteroids, ethylene and gibberellins (Figure S7C, Table S17), with the exception of tryptophan aminotransferase (TAA; auxin biosynthesis) and copalyl diphosphate/ent-kaurene synthase and ent-kaurene oxidase (CPS/KO; gibberellin biosynthesis). The presence of these genes is consistent with the identification of the corresponding hormones[48,49] and of brassinosteroid-like molecules[50] in brown algae. As far as abscisic acid (ABA) is concerned, brown algae have the capacity to synthesise oxidised cleaved carotenoids, supporting the idea that they may generate ABA-like signalling molecules[51]. Similarly, OPDA, an intermediate of the jasmonic acid pathway, has been identified in brown algae, raising the possibility that this molecule itself may have a signalling role similar to OPDA-like molecules in hornworts[52]. The putative brown algal phytohormone biosynthetic genes are also present in the genomes of sister taxa and therefore appear to have been acquired well before the emergence of the brown algae. However, an expansion of the flavin monooxygenase (YUC) gene family (which mediates a rate limiting step in IAA production in land plants) appears to have occurred in the common ancestor of the Phaeophyceae, *S. ischiensis, C. australica* and *T. minus* (Figure S7C).

Membrane-localised signalling proteins (Figures 4A and S7D) are of interest in brown algae not only as potential mediators of intercellular signalling in these complex multicellular organisms but also because of potential interactions with their elaborate extracellular matrices (cell walls)[29,53]. A detailed analysis of the brown algal receptor kinase (RK) gene family, revealed that it actually includes two types of receptor kinase, the previously reported leucine-rich repeat (LRR) RKs[5] and a newly-discovered class of receptors with a beta-propeller extracellular domain (Figures 4A and S7D, Table S18). The phylogenetic distribution of RKs (Figures 4A and S7D) suggests that beta-propeller-type RKs may have originated before LRR RKs in the brown algae although the presence of both beta-propeller and LRR RKs in *C. australica* (but none of the other sister species analysed) raises questions about the timing of the emergence of the brown algal RK family. Note that oomycetes (distantly related stramenopiles) possess a phylogenetically distinct family of LRR RKs[5] but not beta-propeller RKs. Diversification of receptor kinase families by the acquisition of different extracellular domains is therefore a characteristic that is so far unique to the three most complex multicellular lineages: animals, land plants and brown algae. Histidine kinases (HKs) are widespread in the stramenopiles but several classes of membrane-localised HK were either only found in brown algae (HKs with an extracellular domain resembling an ethylene binding motif[54]) or only in brown algae and close sister groups (CHASE and MASE1 domain HKs[54]) and appear to be absent from other stramenopile lineages (Figures 4A and S7D, Table S19). These classes of HK all exhibit a patchy pattern of distribution across

the brown algae and are often monoexonic suggesting possible multiple acquisitions from viruses via HGT (Figures 4A and S7D). Brown algae have integrin-like molecules[53], and this feature was found to be shared with several sister stramenopile lineages including pelagophytes. A single integrin-like gene was found in close sister groups but this gene was duplicated early in brown algal evolution, with a further recent duplication in *Ectocarpus* species 7 and *U. pinnatifida* and gene loss in several other brown algae (Figures 4A and S7D, Table S20). Proteins with an FG-GAP domain (a component of integrins) were found in several other stramenopile lineages but these proteins did not have the characteristic integrin domain structure and their evolutionary relationship to brown algal integrin-like proteins is uncertain (Figure S7D). Brown algae and closely-related sister groups also possess a single gene that is very similar to the land plant membrane-localised calpain domain protein DEK1 (Figures 4A and S7D, Table S21), which is thought to act as a mechanosensor in land plants by activating the RMA channel[55]. Fasciclin (FAS) domain proteins are found throughout the stramenopiles but only brown algae have membrane-localised proteins with four extracellular FAS domains analogous to the membrane-localised fasciclin proteins in animals that interact with the extracellular matrix (Figure S7D, Table S22)[56]. Finally, all stramenopiles, including brown algae, appear to have a single tetraspanin gene plus a gene that encodes a stramenopile-specific eight transmembrane domain tetraspanin-like protein (Figure S7D, Table S23).

The Phaeoexplorer dataset allowed the origins of the major families of brown algal ion channels to be traced (Table S24). Most classes of brown algal ion channel are also found in other stramenopile groups, such as IP_3 receptors, which were earlier shown to mediate calcium signalling in *Fucus*[57], and Nav/Cav four-domain voltage-gated calcium channels, which are associated with rapid signalling processes and the regulation of flagella motion[58]. Pennate diatoms and angiosperms, which both lack flagellated gametes, do not possess Nav/Cav channels[59]. Land plants and diatoms also lack IP_3 receptors. All stramenopile lineages, including brown algae, have OSCAs, which are mechanosensitive channels that have been implicated in calcium signalling during osmoregulation[60,61], and transient receptor potential (TRP) channels (although the latter were not detected in *H. akashiwo*). In brown algae the only ubiquitous ligand-gated ion channels are the glutamate receptors (GLR). P2X receptors are absent and the pentameric ligand-gated ion channel (pLGIC, related to acetylcholine and GABA receptors in animals) was only found in *D. mesarthrocarpum*. EukCat, a novel class of single domain voltage-gated calcium channel recently discovered in diatoms[62], appears to be absent from brown algae and other stramenopiles.

The EsV-1-7 domain is a short, cysteine-rich motif that may represent a novel class of zinc finger[63]. The EsV-1-7 domain protein IMMEDIATE UPRIGHT (IMM) has been shown to play a key role in the

establishment of the elaborate basal filament system of *Ectocarpus* sporophytes[63]. The *IMM* gene is part of a gene family, which has just one member in oomycetes and eustigmatophytes but was shown to have expanded to over 90 members in the *Ectocarpus* species 7 genome[63]. Therefore, although EsV-1-7 domain proteins are completely absent from animal and land plant genomes, they are likely to be an important family of regulatory molecules in the brown algae. Analysis of the Phaeoexplorer data (Figure 4A, Table S25) indicated that the EsV-1-7 gene family had already started to expand in the common ancestor of the brown algae and the raphidophyte *H. akashiwo*, with 31-54 members in the sister taxa that shared this ancestor. Further expansion of the family then occurred in most brown algal orders, with the largest number of genes (335 members) being detected in *Ascophyllum nodosum*. EsV-1-7 domain genes tend to be clustered in tandem arrays in brown algal genomes (Table S6 and S25). The *IMM* gene was found in brown algal crown group taxa and in *Dictyota dichotoma* but not in *D. mesarthrocarpum* (Figure 4A, Table S25), indicating that this gene originated within the EsV-1-7 gene family as the first brown algal orders started to diverge. *IMM* is therefore conserved in most brown algal taxa but its function in species other than *Ectocarpus* remains to be determined.

The emergence of the brown algae also appears to have been associated with genomic events that may have had an important impact on the regulation of chromatin function. Most brown algal genomes code for about 20 histone proteins but the number of histone genes was correlated with genome size. For example, *U. pinnatifida* and *A. nodosum*, which have quite large genomes (634 and 1,253 Mbp, respectively) have 36 and 173 histone genes, respectively (Table S26). As in animal genomes[64], the majority of brown algal histone genes occur in clusters. For example, *Chordaria linearis* and *U. pinnatifida* both have three contigs containing 80-90% of their histone genes, whereas *Saccharina latissima* and *S. ischiensis* each have a single contig which contains ~40% of their histone genes. As in other eukaryotic lineages, histone H4 is highly conserved due to its structural role in binding H3, H2A and H2B, and no variants were found in brown algae. All Phaeophyceae possess H3.3 and CenH3 but additional H3 variants were only found in the Discosporangiales and Dictyotales and in sister taxa as distant as the raphidophytes (Figures 4A and S8A). The centromeric variant CenH3 is highly divergent across the brown algae, particularly in the N-terminal tail (Figure S8B), as reported in other lineages[65]. In contrast to what has been observed in humans and land plants[66,67], H3.3 is encoded by a single gene in brown algae. H3.1 is encoded by several genes as in the majority of organisms. The brown algal H3.1 and H3.3 clades are distinct from those of animals and land plants indicating independent amplifications (Figure S8C). Linker H1 variants are highly divergent across the eukaryotes and this is also true for brown algae, with *A. nodosum,* for example, possessing 11 different H1 proteins (Figure S8D, Table S26). Brown algae with a genome of less than 1 Gbp

possess ~10 H2B-coding genes but *F. serratus*, *A. nodosum* and the sister taxon *H. akashiwo* have at least three times more. Finally, brown algae not only possess H2A variants such as the H2A.X but also three novel H2A variants that only exist in Phaeophyceae and the sister taxon Schizocladiophyceae (Rousselot et al, unpublished).

An analysis of eight histone post-translational modifications in *Ectocarpus* species 7 indicated that the majority had similar genomic and genic distributions to those observed in animals and land plants, indicating that they have similar functional roles[14]. However, dimethylation of lysine 79 of histone H3 (H3K79me2) has a very different distribution pattern in *Ectocarpus* compared to animals. In animals, H3K79me2 is associated with the 5' ends of transcribed genes[68] whereas in *Ectocarpus* it occurs as extended blocks, which can span several genes and appears to be associated, at least indirectly, with down-regulation of gene expression[14]. The H3K79me2 mark is written by DOT1-class DNA methylases, which are encoded by multiple genes in brown algae. Dollo analysis identified gain of a new DOT1 orthogroup (OG0008134) in the common ancestor of the Phaeophyceae and the two close sister taxa *S. ischiensis* and *Phaeothamnion wetherbeei*, indicating a possible correlation between this unusual H3K79me2 distribution pattern and the appearance of a novel DOT1 protein.

DNA methylation was not detected in the filamentous brown alga *Ectocarpus*[5] but occurs at a low level in the kelp *Saccharina japonica*[25]. Both species lack the canonical DNA methylase DNMT1 and methylation is thought to be mediated by DNMT2 in *S. japonica.* Interestingly, analysis of the Phaeoexplorer dataset identified *DNMT1* genes in *Discosporangium mesarthrocarpum* and two sister species *S. ischiensis* and *C. australica* indicating that the common ancestor of brown algae probably possessed *DNMT1* but that this gene was lost after divergence of the Discosporangiales from other brown algal taxa (Figure 4A, Table S27). Consequently, most brown algae are expected to either lack DNA methylation or to exhibit very low levels of methylation.

Overall, these analyses indicated that the emergence of the brown algal lineage coincided with several marked genomic changes including periodic bursts of the emergence of new gene families over evolutionary time, increased domain reorganisation and the appearance of several key families including haloperoxidases, membrane-localised signalling molecules and ion channels.

3) Diversification and adaptation of the brown algae to intertidal ecosystems

Morphological and life cycle diversification was correlated with changes in gene expression and gene evolution rates

Since the establishment of the brown algal lineage, approximately 450 Mya[69], the Phaeophyceae have diversified and adapted to multiple niches in predominantly coastal, marine environments. This diversification has resulted in organisms with a broad range of morphological complexity ranging from filamentous to complex parenchymatous thalli, with the more complex thalli containing a greater number of different cell types (Figures 1 and S9A). Additional highly variable characteristics of the brown algae include the structures of their life cycles, their diverse reproductive strategies and their metabolic pathways[28,29,70,71]. The Phaeoexplorer dataset was analysed to search for genomic features associated with these diverse phenotypic characteristics. Most brown algae have haploid-diploid life cycles involving alternation between sporophyte and gametophyte generations but the relative sizes of the two generations varies markedly across species with, for example, the sporophyte being much larger and more developmentally complex than the gametophyte in groups such as the kelps. Analysis of generation-biased gene (GBG) expression in ten species whose haploid-diploid life cycles ranged from sporophyte dominance, through sporophyte-gametophyte isomorphy to gametophyte dominance indicated, as expected[72], that the number of GBGs was inversely correlated with the degree of intergenerational isomorphy (Figure S9B, Table S28). However, species with very heteromorphic life cycles had greater numbers of both sporophyte- and gametophyte-biased genes, indicating either that gene downregulation plays an important role in the establishment of heteromorphic generations or that there is also cryptic complexification of the more morphologically simple generation.

We also found indications that life cycle variation and the emergence of large, complex bodyplans in the brown algae may have impacted genome evolution through population genetic effects. The theoretical advantages of different types of life cycle have been discussed in detail[73] and one proposed advantage of a life cycle with a haploid phase is that this allows effective purifying counter-selection of deleterious alleles. When the brown algae with haploid-diploid life cycles were compared with species from the Fucales (the only brown algal order with diploid life cycles) increased rates of both synonymous and non-synonymous mutation rates were detected in the latter, consistent with the hypothesis that deleterious alleles are phenotypically masked in species where most genes function in a diploid context (Figure S9C). Comparison of non-synonymous substitution rates (dN) for genes in brown algae with different levels of morphological complexity, ranging from simple filamentous thalli though parenchymatous to morphologically complex, indicated significantly lower values of dN for filamentous species (Figure S9C). This observation suggests that the larger, more complex brown algae may have smaller effective population sizes and consequently weaker counter-selection of non-synonymous substitution rates[19].

Diversification of flagellated cells was correlated with changes in the flagellar proteome

Diversification of brown algal reproductive systems has been associated with multiple changes to flagellated cells, including partial or complete loss of flagella from female gametes in oogamous species and more subtle modifications such as loss of the eyespot in several kelps or of the entire posterior flagella in *D. dichotoma*[74,75]. Interestingly, the latter modifications are correlated with loss of the helmchrome gene from these species (Figures 4A and 5B). An analysis of the presence of genes for 70 high-confidence flagellar proteins across eight species with different flagellar characteristics identified proteins that correlate with presence or absence of the eyespot or of the posterior flagellum (Figure 5B, Table S15).

Diversification of brown algal metabolic and signalling pathways

The metabolic pathways of brown algae have diversified as different species have adapted to different niches. For example, the ManC5-E family, which is involved in alginate modification, markedly expanded in the Laminariales and Fucales (Figure 4C). ManC5-E converts β-D-mannuronate to α-L-guluronate allowing the formation of a rigid gel in the presence of calcium. In addition, five different orthogroups containing proteins with the mechanosensor wall stress-responsive component (WSC) domain were identified as having increased in size during the emergence of the brown algal lineage (Table S6), indicating a complexification of carbohydrate metabolism and perhaps signalling pathways. Polysaccharide metabolism gene amplification events may reflect, at least in part, fine tuning to optimally adapt brown algal cell walls to specific features of their ecological niche and to modulate cell wall characteristics for different organs and cell types within a complex bodyplan. Similarly, several species within the Ectocarpales and Laminariales have acquired an additional *M1PDH* gene perhaps allowing greater capacity or finer regulation of carbon storage through the mannitol cycle (Table S14). Several gene families that are putatively associated with the biosynthesis of phytohormones such as ABA, brassinolide, ethylene, gibberellins, jasmonic acid and strigolactones have expanded in the Fucales (Figure S7C).

In diatoms, LHCX proteins play an important role in the photoprotection of the photosynthetic apparatus under conditions of excess light[76]. Chromosome six of the *Ectocarpus* species 7 genome is known to contain a cluster of 14 LHCX genes interspersed with genes coding for proteins unrelated to photosynthesis[5,77]. Analysis of the Phaeoexplorer genomes indicated that LHCX genes were present in this genomic context before the divergence of the Desmarestiales but that the expansion

of the gene cluster only began in a common ancestor of the Laminariales and Ectocarpales (Figure S10A). The general features of the cluster are conserved in the genomes of *Ectocarpus* species 7, *Ectocarpus crouaniorum*, *Ectocarpus fasciculatus*, *Scytosiphon promiscuus*, *Porterinema fluviatile* and *C. linearis* but in *C. linearis* the cluster appears to have undergone a partial rearrangement and contains only three LHCX copies whereas the clusters in all other species include at least seven LHCX genes. In *S. latissima*, the genes of the cluster were found in three short contigs that may be adjacent on the same chromosome. LHCX gene clusters are also present in *D. dichotoma* (4 genes on contig 3024, 3 on contig 458 and 2 on contig 1592) of the early branching order Dictyotales and in *S. ischiensis* (4 plus 2 genes on contig 28 and 2 genes on contig 64), but the genomic context of these clusters differs from that of the cluster found in *Ectocarpus* spp., indicating independent expansions of the LHCX family in these taxa. LHCX sequences from *S. ischiensis*, *D. dichotoma* and the genus *Ectocarpus* grouped into independent, species/genus-specific clusters in a phylogenetic analysis (Figure S10B), further supporting independent expansions of the LHCX gene family in these taxa. The LHCX clusters probably evolved by gene duplication and ectopic gene conversion. The involvement of gene conversion is indicated by the higher intraspecific identity of LHCX sequences between paralogs than between putative orthologs (based on synteny) compared across species (Figure S10A and S10B). The large gene cluster that contains LHCX4 to LHCX7 commonly contains overlapping LHCX genes on opposite strands of the chromosome, with some or all of the exons of one gene being located within an intron of the other (Figure S10C). Changes in gene dosage in response to environmental selection (as observed, for example, in studies by[78,79]) may have been a driver of the LHCX gene family expansion.

Comparison of the plastid genomes generated by the Phaeoexplorer project (Table S2) identified gene loss events during the diversification of the brown algal orders that affected five different loci. Two of these genes have been lost more than once within the lineage (Figure S10D, Table S29). Gene losses have previously been reported[80,81] for three of these genes (rpl32, rbcR and sylB). Phylogenetic trees based on either the plastid or mitochondrial gene sets were congruent with the phylogeny based on nuclear gene data[82] at the ordinal level (Figure S10E). For the plastid genome, this congruency was also observed at the sub-ordinal level but, interestingly, a marked discordance was observed between the mitochondrial tree and the nuclear and plastid trees for the Ectocarpales, indicating possible mitochondrial introgression events between species within this order (Figure S10E), as has been observed in other eukaryotic lineages, for example *Picea*[83] and chipmunks[84].

An emblematic characteristic of brown algae is their capacity to accumulate halogens and their complex halogen metabolisms, particularly for groups like the Laminariales and Fucales[31].

Comparative analysis identified several instances of independent expansions of the haloperoxidase gene families within orders, with expansions being particularly marked in the Fucales and the Laminariales (Figures 4A and S6C). In contrast, with some rare exceptions, the Ectocarpales do not possess expanded vHPO gene families and there is evidence for multiple, independent gene losses leading to loss of one or two of the three vHPO classes in many species. In the Laminariales, the algal type I family are specialised for iodine rather than bromine[85] and this may have been an innovation that occurred specifically within the Laminariales but further analysis of the enzymatic activities of proteins from other orders will be necessary to determine if substrate specialisation was concomitant with this specific gene family expansion.

The number of transcription-associated protein (TAP) genes has remained relatively stable during the diversification of the brown algae (Figures 4A, S7A and S7B). This contrasts with what has been observed for the green lineage, where size of TAP complement has been shown to be broadly correlated with morphological complexity[86]. Although the overall TAP complement has been stable, variations were detected in the sizes of some specific TAP families, including variations that correlated with the degree of morphological complexity. A broadly positive correlation was observed for the three-amino-acid length extension homeodomain (TALE HD) transcription factor (TF) families for example (Figure S11A, Table S16). These observations suggest that increased multicellular complexity in the brown algae may have required different combinations of TFs rather than simply increasing numbers. Two TALE HD TFs (OUROBOROS and SAMSARA) have been shown to be necessary for the deployment of the sporophyte developmental program in *Ectocarpus* species 7[87]. The *Ectocarpus* species 7 TALE HD TF family consists of three genes which are conserved across the brown algae[87] but we detected a modest expansion of the gene family in some members of the Fucales, which have acquired one or two additional paralogues of the third TALE HD TF (Figure S11B).

The Mitochondrial Calcium Uniporter (MCU) plays an important role in the uptake of calcium into mitochondria and is highly conserved across the eukaryotes. It is therefore surprising that the Fucales, like diatoms[88], have lost this transporter (Table S24) and it is unclear how Fucales and diatoms regulate mitochondrial calcium, as alternative mechanisms are not known. Cyclic nucleotide gated channels, which are only present at low copy number, also appear to have been lost from several brown algal lineages, including kelps. In contrast, TRP channels are found in all brown algal genomes and the gene family is massively expanded in some lineages (Table S24). In animals, TRP channels are primarily involved in sensory functions e.g. touch, temperature and ligand-based activation[89], so the brown algal channels may play a role in sensing the environment. TRP channels have been lost in land plants, which show an expansion of CNGCs and GLRs[90].

Five orthogroups containing GTPases with a central Ras of complex proteins/C-terminal of Roc RocCOR domain tandem (ROCO GTPases) or nucleotide-binding adaptor shared by apoptotic protease-activating factor 1, R proteins and CED-4 tetratricopeptide repeat (NB-ARC-TPR) genes increased in size during brown algal evolution (Table S6). A role has been proposed for these gene families in biotic defence responses[91] so the diversification of the gene families may have conferred important immunity characteristics.

Genome changes associated with adaptation to freshwater environments

A small number of brown algal species have adapted to freshwater habitats, with each adaptation event occurring independently. The Phaeoexplorer dataset includes genomes for four freshwater species in addition to the previously sequenced *Ectocarpus subulatus*[9]. Adaptation to a freshwater habitat was correlated with loss of orthogroups (Figure S3A) and clear reductions in the sizes of some key gene families such as haloperoxidases (Figures 4A and S6C). In terms of gene gain, 12 orthogroups were found to be consistently over-represented when the genomes of each of the five freshwater species were compared with those of related seawater species, including orthogroups with genes involved in membrane transport and signalling. A transcriptomic analysis of *P. fluviatile* grown in either seawater or freshwater conditions identified 1,328 differentially-expressed genes (DEGs), with the genes upregulated in freshwater being enriched in DNA replication regulation and transcription functions (Table S30). Comparison of the differentially regulated gene set with a similar set generated for *E. subulatus* indicated that both sets of DEGs were enriched in lineage-specific genes, suggesting that, to a large extent, each adaptation to freshwater involved lineage-specific innovations (Table S31). Furthermore, a focused comparative analysis of metabolic networks (https://gem-aureme.genouest.org/fwgem/; Figure S5G) indicated that both *P. fluviatile* and *Pleurocladia lacustris* have lost alginate lyases and a glycosyltransferase GT18, related to polysaccharide metabolism (Table S32). This is in line with previous work showing specific cell wall remodelling in *E. subulatus*[92]. *P. fluviatile* and *P. lacustris* have also lost genes encoding ubiquitin-conjugating enzymes and this may have impacted protein degradation and thus osmolarity, or be linked to the preferential activation of ubiquitin-independent regulatory mechanisms in situations of osmotic stress[85]. In addition to the shared gene losses, both species have independently lost genes. The strictly freshwater *P. lacustris* has lost several phosphatases, peptidases and aldolases, perhaps due to a reduced need to produce osmolytes, while the brackish water *P. fluviatile* has lost several enzymes related to phenol metabolism and a magnesium transporter. More surprisingly, the two species have lost different sets of enzymes involved in RNA modifications, which may either indicate

convergent acquisition of different regulatory mechanisms, or may simply be markers of phylogenetic divergence.

Brown algal genomes contain large amounts of inserted viral sequences

Finally, we found evidence that the genetic content of brown algal genomes can be significantly impacted as a result of large DNA viruses of the Phaeovirus family integrating into the genomes as part of their lysogenic life cycles (Figure 6, Table S33). An analysis of 72 genomes in the Phaeoexplorer and associated public genome dataset identified a total of 792 viral regions (VRs) of *Nucleocytoviricota* (NCV) origin in 743 contigs, with a combined length of 32.3 Mbp. Individual VRs ranged in size from two to 705 kbp, but the majority (81.3 %) were between two and 50 kbp, whilst only 9% were longer than the expected minimum size (100 kbp) for an NCV genome. On average, VRs comprised 65% of their contig by length, and 72% of contigs with VRs were shorter than 100 kbp. Therefore, for most VRs, their true size and genomic context is unknown due to assembly gaps and short contig lengths. On average, each genome contained 469 kbp of VR (ranging from less than one to 5,614 kbp) and only two genomes contained no VRs (both from the Discosporangiales). There were a number of outlier genomes that contained more than 1,000 kbp of VRs (*T. minus*, *S. latissima* female, *S. japonica*, *P. fluviatile* and *Myriotrichia clavaeformis* male and female). The presence of key NCV marker genes was used to assess the completeness of inserted viral genomes. At least one partial provirus (a VR possessing several key NCV marker genes) was present in 39 genomes, 29 of which had at least one full provirus with a complete set of seven key NCV marker genes (Figure 6, Table S33). In addition to the previously known infections in Ectocarpales[93] and Laminariales[94], integrated NCV proviruses were found in all Phaeophyceae orders screened, except the Discosporangiales and Dictyotales, and were also detected in *T. minus* (Xanthophyceae). Moreover, NCV marker gene composition indicated that multiple integrated proviruses were present in 16 genomes from multiple Phaeophyceae orders (Ectocarpales, Desmarestiales, Sphacelariales, Tilopteridales, Laminariales), and the Xanthophyceae (Figure 6, Table S33). Phylogenetic analysis of the major coat protein (MCP) and DNA polymerase genes indicated that the majority of the integrated NCVs belonged to the genus *Phaeovirus*, the sole viral group known to infect brown algae. However, this analysis also revealed integrated sequences corresponding to other viral groups. Viral sequences in *T. minus* belonged to a putative novel genus closely related to *Phaeovirus*, for which we propose the name *Xanthovirus*. Finally, mimiviridae-related VRs were identified in *S. latissima* and *Pelvetia canaliculata*, but since they are partial proviruses and do not appear to possess integrase genes, they may have originated from ancient endogenization events, similar to those described in chlorophytes[95]. The identification of integrated NCV across almost all brown algal orders and in

closely related sister taxa suggests that the lysogenic life cycle strategy of phaeoviruses is ancient and that giant viral genomes have been integrating into the genomes of brown algae throughout the latters' evolutionary history. This conclusion was supported by the phylostratigraphic analysis, which detected the appearance of many novel virus-related genes dating back to the origin of the Phaeophyceae (Figure 3A). Marked differences were detected in total VR size and NCV marker gene presence across the brown algal genome set, and large differences were even detected between strains from the same genus (between 24 and 992 kbp of VR in different *Ectocarpus* spp. for example; Figure 6, Table S33). These differences indicate dynamic changes in VR content over evolutionary time, presumably due, at least in part, to differences in rates of viral genome integration, a process that can involve multiple, separate insertion events[96], and rates of VR loss due to meiotic segregation[97]. In addition, the abundant presence of partial proviruses and NCV fragments in brown algal genomes indicates that inserted VRs can degenerate and fragment, probably also leading to VR loss over time. It remains to be determined what mechanisms are responsible for eliminating and fragmenting inserted phaeoviruses. The identification of large-scale viral genome insertion events over such a long timescale (at least 450 My[69]) suggests that NCVs may have had a major impact on the evolution of brown algal genomes throughout the emergence of the lineage, making this a truly unique model system to investigate the impact of genomic exchanges on genome evolution.

In conclusion, changes in the gene content of brown algal genomes can be correlated with the diversification of the lineage and adaptation to diverse aquatic environments, a process that has involved changes in life cycle structure, the evolution of different levels of multicellular complexity and modifications to metabolic networks. There are also indications that the changes associated with adaptation have reciprocally impacted the evolution of the species' genomes via population genetic mechanisms. Finally, during the diversification of the brown algae, genome content has been significantly impacted by the insertion of phaeoviral sequences.

4) Ongoing evolutionary processes: microevolution and speciation

To investigate short-term and ongoing evolutionary events in the brown algae, we sequenced 22 new strains from the genus *Ectocarpus*. *Ectocarpus* belongs to the order Ectocarpales, which emerged late in brown algal evolution but has become the most species-rich brown algal order, with over 750 species[98]. Therefore, although more morphologically complex organisms have evolved in other orders such as the Laminariales and the Fucales, the Ectocarpales are nonetheless of particular interest because they are a highly successful order in terms of species numbers. The genus

Ectocarpus is a species complex in which 16 cryptic species have been identified[99]. Comparison of the high continuity *Ectocarpus* assemblies indicated that *Ectocarpus* spp. genomes share extensive synteny (Figure S12A, Table S34). In addition, intron sizes, phases and positions are strongly conserved across *Ectocarpus* species (Figure S12B). The Ectocarpales exhibited the same general phenomenon of orthogroup loss as the other Phaeophyceae orders. However, this trend was not uniform within the Ectocarpales and a period of orthogroup gain was predicted during the early emergence of the genus *Ectocarpus* (Figure 7A). The sets of both gained and conserved orthogroups were significantly enriched in transcription-related functions (Figure S12C). Analysis of the *Ectocarpus* species 7 reference genome orthologues indicated that rapidly-evolving, taxonomically-restricted genes (Group 5 based on the phylostratigraphy analysis: Table S5) are most abundant on the sex chromosome (Figure 7B), an observation that is consistent with other recent findings[100]. These *Ectocarpus* genus-specific genes exhibit higher dN and dS values than genes that the genus shares with other brown algae (Figure 7C, Table S35).

A phylogenetic tree was constructed for 11 selected *Ectocarpus* species based on 257 high quality alignments of 1:1 orthologs (Figures 7D). This species tree is strongly supported, with all nodes displaying high posterior probabilities consistently reaching 0.99 or 1. The tree indicates substantial divergence between *E. fasciculatus* and two well-supported main clades, designated clade 1 and clade 2. Rates of synonymous and non-synonymous substitutions, based on alignments of genes in syntenic blocks between *Ectocarpus* species 7 and four other species correlated with relative divergence times as estimated from the phylogenetic tree (Table S35), supporting the tree topology. Incongruencies between the species tree and trees for individual genes indicated introgression events and/or incomplete lineage sorting across the *Ectocarpus* genus. This conclusion was supported by phylogenetic tree comparisons, Bayesian hierarchical clustering, phylogenetic network reconstructions and PCoA analysis (Figures 7D, S12D and S12E). D-statistic analysis, specifically ABBA-BABA tests, detected incongruities among species quartets, indicating potential gene flow at various times during the evolution of the *Ectocarpus* genus. Evidence for gene flow was particularly strong for clade 2 and there was also evidence for marked exchanges between the two main clades (Figure 7E), suggesting that gene flow has not been limited to closely-related species pairs. These findings suggest a complex evolutionary history involving rapid divergence, hybridization, and introgression among species within the *Ectocarpus* genus.

In conclusion, analysis of the genus *Ectocarpus* identified a number of evolutionary processes including gene gain and gene loss, emergence of new genes and gene flow within the genus underlining that genome evolution in the brown algae is an ongoing, dynamic process.

DISCUSSION

The genome resource presented in this study has provided an unprecedented overview of genomic diversity across the full scope of the brown algal lineage. Comparative analysis of these genomes has provided insights into brown algal genome evolution at multiple time-scales ranging from the early emergence of the lineage to recent evolutionary events at the genus level. These analyses identified a period of marked genome evolution concomitant with the emergence of the brown algal lineage involving enhanced levels of gene gain and the formation of genes with novel domain combinations. These early events were important in equipping the brown algal ancestor with key components of several metabolic pathways that were essential for their colonisation of intertidal and subtidal environments leading to the formation of the extensive underwater forests present worldwide today. These pathways notably include cell wall polysaccharide, phlorotannin and halogen metabolisms. The capacity to synthesise flexible and resilient alginate-based cell walls allows these organisms to resist the hydrodynamic forces of wave action[101], whereas phlorotannins and halogen derivatives are thought to play important roles in defence[102]. There is also evidence that cell wall cross-linking by phlorotannins may be important for strong adhesion to substrata, another important characteristic in the dynamic intertidal and subtidal coastal environments[103]. The capacity to adhere strongly and resist both biotic and abiotic stress factors would have been essential for the success of large, sedentary multicellular organisms in these intertidal niches over evolutionary time.

Comparative analyses also identified changes in gene content and patterns of gene evolution associated with the diversification of the different brown algal orders that resulted in marked differences in life cycle structure, multicellular complexity and differences regarding specific biological features such as zoid flagellar structure. Notably, diversification of the brown algae was associated with modifications to the contents of several important gene families, including CAZYmes, LHCX, haloperoxidases, transcription factors and ion channels. Particularly marked changes were observed in the genomes of freshwater species, which have made a major transition from their ancestral marine habitats. Analysis of the genomes of multiple *Ectocarpus* species demonstrated that genomic modifications, including gene gain and gene loss have continued to occur up until the present time and indicated that these modifications can potentially be transmitted between species as a result of gene flow occurring within a genus due to incomplete reproductive boundaries and introgression.

Overall, these analyses identified multiple processes that have contributed to brown algal genome evolution at different scales of evolutionary time. In addition to the evolution of taxonomically-restricted genes, we found evidence that genes have been acquired by HGT and also that domain shuffling has played an important role in the generation of genic novelty. Infection of brown algae by large DNA viruses of the Phaeovirus family has resulted in the introduction of a remarkable quantity of additional genetic material into brown algal genomes and this large quantity of integrated viral genes represents a considerable potential source for the evolution of novel genes through horizontal gene transfer.

The comparative genomic approaches used here have characterised changes in gene content and structure over the course of the evolutionary history of the brown algae and many of these changes have been correlated with the emergence of key biological characteristics of the lineage. With comprehensive tools and resources currently available for the model brown alga *Ectocarpus*[104] plus the recent development of CRISPR-Cas9 methodology for brown algae[105,106], it will now be possible to apply functional genomic approaches to validate and further investigate the roles of these genes in brown algal biological processes.

In addition, the Phaeoexplorer brown algal genome dataset represents an important new resource for brown algal research that will continue to be exploited using comparative genomic approaches coupled with gene function analyses. To facilitate future use of this genome resource, the annotated genomes have been made available through a website portal (https://phaeoexplorer.sb-roscoff.fr), which provides multiple additional resources including genome browsers, BLAST interfaces, transcriptomic data and metabolic networks.

ACKNOWLEDGMENTS

This work was supported by the France Génomique National infrastructure project Phaeoexplorer (ANR-10-INBS-09), the European Research Council project Sexsea (638240), the Investissements d'Avenir project Idealg (ANR-10-BTBR-04-01), the European BG-01 BlueGrowth H2020 project Genialg (727892), Laoshan Laboratory grants (LSKJ202203801, LSKJ202203204), the Taishan Scholars Program and Talent Projects of Distinguished Scientific Scholars in Agriculture, the CNRS international research network DABMA (00022), the ANR projects Epicycle (ANR-19-CE20-0028-01), BrownSugar (ANR-20-CE44-0011), HaloGene (ANR-22-CE20-0025), Seabioz (ANR-20-CE43-0013) and BrownLincs (ANR-23-CE20-0048-01), the National Research Foundation of Korea (2022R1A2B5B03002312, 2022R1A5A1031361) granted to H.S.Y., the projects Connect Talent EpiAlg

Région Pays de la Loire-Nantes Métropole and Etoiles Montantes M-EpiCC Région Pays de la Loire, the MITI-funded project Algometabionte, the CNRS and Sorbonne University. We are grateful to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), which is part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing both help and computing and storage resources.

AUTHOR CONTRIBUTIONS

Software: L.B.G., R.D., A.L.B., X.L., D.N. and E.C.; Formal analysis: F.D., O.G., L.D., D.M., T.M., D.S., X.F., L.Ma., N.Te., J.B.R., R.P., L.R., S.W.C., J.J., K.U., K.B., C.Du., P.R., A.L., A.E.M., M.L., A.K., P.H.G., C.V., S.S.A., S.A., K.A., Y.B., T.Ba., A.Be., S.B., A.Bo., A.Cor., H.C.C., A.D., E.Di., S.D., E.Dr., J.G., L.G., A.G., M.L.G., L.H., B.H., A.J., E.K., C.B., R.L., P.J.L., E.M., S.M., G.M., C.N., S.A.R., E.R., D.Sch., A.S., L.T., T.T., K.V., H.V., G.W., H.K., A.F.P., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.M.A. and J.M.C; Investigation: C.C., S.H., Z.N., N.Ta., A.Cou., B.N., W.B., E.De., C.J., L.Me., S.R. and D.Sco.; Resources: O.G., A.Cou., L.B.G., T.Br., R.A.C., C.De., S.F., W.J.H., G.H., K.K., A.L.B., K.M., C.M., N.P., P.P., S.R., D.Sco., H.V., F.W., H.K., A.F.P., M.V., E.C. and J.M.A.; Data Curation: O.G., C.C., A.Cou., L.B.G., J.M.A. and J.M.C.; Writing-Original Draft: F.D., O.G., J.M.A. and J.M.C.; Writing-Review & Editing: all authors; Visualization: F.D., O.G., L.B.G., L.D., D.M., T.M., D.S., X.F., L.Ma., J.B.R., R.P., L.R., K.U., K.B., P.R., A.L., M.L., P.H.G., C.V., A.D., A.L.B., H.K., E.C. and J.M.C.; Supervision: O.G., N.Te., M.L., R.A.C., H.C.C., O.D.C., S.D., G.H., A.J., C.B., E.P., P.L., S.A.R., A.S., L.T., H.V., G.W., H.K., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.M.A. and J.M.C.; Project administration: F.D., E.B., M.V., G.V.M., E.C., S.M.C., P.W., J.M.A. and J.M.C.; Funding acquisition: H.C.C., C.B., P.P., C.H., N.Y., S.M.C. and J.M.C.; Conceptualization: J.M.C.

Figure legends

Figure 1. Ecology, diversity and evolutionary features of the brown algae

The upper panel indicates approximate positions in the intertidal of key species whose genomes have been sequenced by the Phaeoexplorer project. The lower panel illustrates the diversity of brown algae in terms of number of cell types, thallus size, morphological complexity and life cycle type (maximal values for each taxa). The panel also indicates a number of key evolutionary events that occurred during the emergence of the Phaeophyceae and diversification of the brown algal orders.

Some lineages may have secondarily lost a characteristic after its acquisition. Note that members of the genus *Ishige* (Ishigeaceae) also exhibit desiccation tolerance (not shown).

Figure 2. Taxonomic diversity and assembly quality of the Phaeoexplorer genomes and structural features of the predicted genes

(A) Taxonomic distribution and assembly quality (contig N50) of the Phaeoexplorer genome dataset (blue) and previously published brown algal genomes (brown). "Reference" quality Phaeoexplorer genomes are circled in black.

(B) Genome and gene statistics for the 21 reference genomes (panel (A) and Table S1). Violin plots display size distributions. Cross or triangle, mean; diamond or square, median. For intergenic regions, half violins on the left correspond to intergenic distances between adjacent genes on opposite strands, and half violins on the right to intergenic distances between adjacent genes on the same strand.

(C) Intron acquisition during the emergence of the brown algal lineage (left panel). Colours correspond to the species distributions of introns. Hatched blue indicates introns that are shared with at least two outgroups (ancestral introns). The right panel shows an example of an intron conservation profile for the orthogroup OG0004854. Colour code and species numbering as for the left panel.

TDG, tandem duplicated genes; F, female; M, male.

Figure 3. Genome-wide analyses of brown algal genome and gene content evolution

(A) Inferred gene family founder events in seven brown algal lineages before (dashed lines) and after (solid lines) accounting for homology detection failure (upper panel). Shared peaks of taxonomically-restricted gene emergence were detected at three taxonomic levels (Phaeophyceae, FDI clade and Fucophycidae) after accounting for homology detection failure. Lower panel: functional and structural features of founder genes from three taxonomic levels. FDI, Fucophycideae/Dictyotales/Ishigeales; PS, Phaeophyceae plus Schizocladiophyceae; PX, Phaeophyceae plus Xanthophyceae.

(B) Dollo parsimony reconstruction of gene family (orthogroup) gain (green) and loss (red) during the emergence of the brown algae (upper panel). Circles indicate the numbers of orthogroups predicted to be present at each node or leaf. The tree is a cladogram with Phaeophyceae in brown and outgroup species in black. Lower panel: circles correspond to significantly enriched GO terms in the set of orthogroups gained at a specific node of the phylogenetic tree (based on the Dollo analysis) compared to the entire set of orthogroups. Functional categories have been grouped according to

COG functional categories. Circle size corresponds to the number of orthogroups and colour to the fold-change enrichment of the GO-term.

(C) Gene families (orthogroups) significantly amplified in the brown algae compared to outgroup taxa (upper panel). Species along the y-axis are as in (B). Orthogroups amplified in specific groups of species are indicated by green rectangles. The numbers correspond with the numbered pie charts in the lower panel. Lower panel: pie charts representing the proportions of manually-determined functional categories for each group of amplified gene families highlighted in the upper panel.

(D) Horizontal-gene-transfer-derived genes in orthologous groups and across species. Species along the y-axis are as in (B). Grey bars indicate the number of HGT genes (light-grey) and the number of orthogroups containing HGT genes (dark-grey) for each species. The black trace represents the percentage of genes resulting from HGT events per species. Pie charts summarise the predicted origins (donor taxa) of the HGT genes. The right-hand bar graph indicates the proportions of ancestral (i.e. acquired before the root of the phylogenetic class, in grey) and class-specific (i.e. acquired within the phylogenetic class, in blue) HGT genes.

(E) Composite gene analysis. Phylogenetic distribution of fused (blue), split (green) and non-remodelled (grey) gene family originations across the evolution of brown algae (middle panel). Pie charts on each branch of the phylogeny indicate the relative contribution of gene fusion and fission to the overall emergence of novel gene families, quantified by the area of the circle. Brown algal species are indicated in brown and other stramenopiles in black. Note that only the topology of the species tree is displayed here, without specific branch lengths. Right panel: barplot indicating the percentage of gene families retained in extant genomes among all gene families that emerged during the evolution of the species set. Left panel: barplot representing the distribution of gene families in COG functional categories for functionally annotated fused, split, and non-remodelled orthologous groups. The functional annotation assigned to an orthogroup corresponded to the most frequent functional category annotated for the members of each orthogroup. Asterisks next to the bars indicate statistically significant differences between remodelled and non-remodelled gene families (p-value <0.05, two-sided Chi^2 test with Yates correction). COG functional categories as in (B).

Figure 4. Gene family evolution during the emergence of the brown algal lineage and a focus on carbohydrate metabolism

(A) Variations in size for a broad range of key gene families in the brown algae and sister taxa. Numbers indicate the size of the gene family. Note that the *S. ischiensis* algal-type HPOs appear to be intermediate between classes I and II. Brown tree branches, Phaeophyceae.

(B) Overview of information from the orthogroup Dollo analysis, the phylostratigraphy analysis, the horizontal gene transfer analysis and the gene family amplification analysis for a selection of cell-wall

active protein (CWAP) families. Expert functional annotations been crossed with orthogroup (OG) composition, each dot represents a couple functional family / orthogroup. The size of the dot is proportional to the number of proteins annotated in the OG, and the colour represents the proportion of the functional annotation that falls into this OG. CWAP annotations are arranged by broad functional categories along the x-axis. Phylogenetic levels considered in the genome-wide analyses are indicated on the right. PS, Phaeophyceae and FDI clade, identified as gene innovation stages, are highlighted in dark-brown. Functional categories with interesting evolutionary histories are highlighted in light red.

(C) Phylogenetic tree of mannuronan C5-epimerases (ManC5-E), a key enzyme in alginate biosynthesis. The inset circle represents a global ManC5-E phylogeny with three main clusters. The phylogeny shown on the left, with the three clusters indicated, is representative of the global view.

(D) Phylogenetic tree of the polysaccharide lyase 41 (PL41) family, a key enzyme involved in the alginate degradation in brown algae. The green squares indicate sequences that have been characterised biochemically. Brown algal sequences are colour-coded in relation to their taxonomy, as indicated in (C). Sequences belonging to the sister group Schizocladiaphyceae are shown in red and with a red circle.

P, present; A, absent; CAZYmes, carbohydrate-active enzymes; HPO, vanadium haloperoxidase; PKS, type III polyketide synthase; TAPs, Transcription-associated proteins; EsV-1-7, EsV-1-7 domain proteins; DNMT, DNA methylase; GTs, glycosyltransferases; GHs, glycoside hydrolases; ARF, auxin response factor-related; bHLH, basic helix-loop-helix; HMG, high mobility group; Zn-clus, zinc cluster; C2H2, C2H2 zinc finger; GNAT, Gcn5-related N-acetyltransferase; SNF2, Sucrose nonfermenting 2; LRR, leucine-rich repeat; QAD, b-propeller domain; RK, membrane-localised receptor kinase; HK, histidine kinase; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor1 domain; DEK1, defective kernal1; MCU, mitochondrial calcium uniporter; GLR, glutamate receptor; pLGIC, pentameric ligand-gated ion channel; TRP, transient receptor potential channel; IMM, IMMEDIATE UPRIGHT; H3, histone H3; MAS, mastigoneme proteins; AA, auxiliary activity; ECT, Ectocarpales; LAM, Laminariales; FUC, Fucales; DES, Desmarestiales.

Figure 5. Evolution of key gene families during the emergence of the brown algal lineage

(A) Evolution of type III polyketide synthase (PKS) genes in the stramenopiles (left panel). Light blue, blue and dark blue circles correspond to the three classes of type III PKS (i.e. PKS1, PKS2 and PKS3), respectively. The number of circles indicates the number of gene copies and absence of a circle indicates absence of PKS genes. Coloured dots on the tree nodes indicate phylogenetic levels. Numbers indicate predicted evolutionary events. Right panel: condensed view of a phylogenetic

reconstruction tree of stramenopile PKS III and closely-linked sequences. Red, orange and black branches correspond to brown algae, other stramenopiles and bacterial/eukaryotic outgroups, respectively. In brackets: number of sequences identified in each phylogenetic group. Branch bootstrap support is indicated.

(B) Loss of orthogroups corresponding to flagellar proteome components[47] in eight brown algal species from five orders. The drawings on the right indicate the cellular structures of zoids from the eight species. Grey, nucleus; yellow, chloroplast; blue, anterior flagella with mastigonemes; red, eyespot. The posterior flagellum is shown either in green to indicate the presence of green autofluorescence correlated with the presence of the eyespot or in blue in species without an eyespot. Note the loss of the eyespot in the Laminariales species *S. latissima* and *U. pinnatifida* and the loss of the entire posterior flagellum in *D. dichotoma*. Bars below the heatmap indicate gene losses associated with loss of just the eyespot (orange) or of the entire posterior flagellum (blue).

(C) Expansions, contractions, gains and losses of transcription-associated protein (TAP) families during the emergence of the brown algae. The number of gains (dark green), losses (dark red), expansions (light green) and contractions (light red) at each node is shown and the families involved are indicated using the same colour code. The light brown box indicates the Phaeophyceae.

Figure 6. Annotated phylogeny summarising key statistics of the presence of nucleocytoviricota (NCV) sequences in the genomes of brown algae and sister taxa

Eight genomes that were sourced from public databases are labelled with an asterisk. Outer layers around tree are as follows: 1) NCV genotypes in each genome, 2) NCV core gene count indicates the number of copies of each viral core gene (A32, A32 packaging ATPase; D5/D5p, D5 helicase/primase; MCP, major capsid protein; poB, DNA polymerase B; SF2, superfamily 2 helicase; VL3, very late transcription factor 3; RNR, ribonucleotide reductase; inC, integrase recombinase; inS, integrase resolvase), 3) Count of viral regions is the number of viral regions within each size range category as indicated, 4) Count of proviruses is the estimated number of complete or partial integrated viral genomes in a genome, 5) Total viral region length is the sum of the lengths in kbp of all viral regions within a genome. The outermost layer indicates the taxonomic class or order of the host clades.

Figure 7. Evolution of genomes within the genus *Ectocarpus*

(A) Gene family (orthogroup) gain and loss at the origin of the *Ectocarpus* genus based on a Dollo parsimony analysis. Cladogram branches are coloured according to the overall gain (+) or loss (-) of gene families. Grey circles on the nodes represent the number of gene families present.

(B) Proportions of genes of different evolutionary ages, as estimated by phylostratigraphy analysis, on each chromosome of the *Ectocarpus* species 7 genome. Colour-coded groups of genes correspond

to the following phylostratigraphy categories: 1, cellular organisms to phylum (ranks 1 to 5; orange); 2, PX clade to Phaeophyceae/*Schizocladia ischiensis* (ranks 6 to 7; yellow); 3, Phaeophyceae to subclass (ranks 8 to 10; green); 4, superorder to family (ranks 11 to 13; blue); 5, genus to species (ranks 14 to 15; pink).

(C) Distribution of synonymous (dS) and non-synonymous (dN) substitution rates for syntenic genes between *Ectocarpus* species 7 and *E. siliculosus*, *E. crouaniorum* or *E. fasciculatus* in relation to gene age based on the phylostratigraphy analysis. Correspondence between the colour-coded gene groups and age ranks are indicated on the left and are the same as in (B).

(D) DensiTree visualisation of gene trees for 257 orthologues shared by 11 *Ectocarpus* species and the outgroup species *S. promiscuus,* together with the consensus species tree. Individual gene trees and the consensus species tree are depicted by grey lines and a black line, respectively.

(E) Violin plot reporting the range of D-statistic (Patterson's D) values between P2 and P3 species. Within-lineage comparisons (i.e. within clades 1 and 2) and between-lineage comparisons are distinguished on the x-axis. The annotation of each dot indicates species that were designated as P2 and P3. For this test, *Ectocarpus fasciculatus* was defined as the outgroup.

Supplementary figure legends

Figure S1. Genome structure statistics and quality control, related to Figure 2.

(A) Number of genes annotated in each genome (upper panel). BUSCO scores for the predicted proteome of each genome (middle panel). Correlation of CDS lengths for each species with the corresponding sequences from the *Ectocarpus* species 7 reference genome (lower panel). Previously published genomes are indicated in grey. longreads, genome assembled using long reads.

(B) Ancestral state reconstruction of genome size (left in bp) and GC content (right as percent). The colour gradients indicate the genome size and GC content across the tree.

(C) Cumulative sequence covered by different features for each genome assembly: masked (repeated sequence) and unmasked regions (upper graph), intergenic and gene regions (middle graph) and intron UTR and CDS (lower graph).

Figure S2. Spliceosome components and intron conservation, related to Figure 2.

(A) Phylogenetic tree of eukaryotic Lsm/Sm proteins. The tree was rooted using midpoint rooting. The colour code for the branches correspond to SmD3B (light green), SmD3A (dark green) and Lsm4 (grey).

(B) Phylogenetic tree of Lsm13 and Lsm14 proteins. The tree was rooted using midpoint rooting.

(C) Intron conservation across a set of 235 conserved single-copy orthologous genes. The upper histogram shows the numbers of introns shared by the groups of species linked by black dots in the lower panel.

(D) Conservation of intron phase and size of the introns from each species compared with the *Ectocarpus* species 7 reference genome.

(E) Sequence logos of the donor and acceptor sites of introns from six selected species.

(F) Analysis of long non-coding RNA genes in nine brown algal genomes and two sister taxa. Upper panel: relative proportions of long non-coding RNA (lncRNA-like) and protein-coding (Coding-like) genes. Lower panel: distributions of transcript length (left), length of longest ORF (middle) and percent GC (right) for lncRNA and protein-coding (mRNA) genes. M, male; F, female.

Figure S3. Dollo-logic-based analysis of gene family gain and loss during the emergence of the brown algae, related to Figure 3.

(A) Complete version of the Dollo analysis in Figure 3B. Cladogram indicating orthogroup (OG) gain and loss during the emergence of the Phaeophyceae based on Dollo analysis. Taxonomic classes, orders or families of the species are indicated in brown (brown algae) or grey (outgroup species) on the right. The nodes of the tree (n0-n22) are numbered in red, and listed on the left with the corresponding name, if one exists. The number of OGs predicted to be present at each node is indicated by the circles and the branches are coloured according to overall gene family gain.

(B) Heat map of COG functional categories associated with OGs gained at each node of the cladogram.

Figure S4. Genome-wide analyses of gene family evolution, related to Figure 3.

(A) Functional annotation status of fused, split, and non-remodelled gene families (orthogroups).

(B) Example of the emergence of a gene family with a novel domain structure just prior to the emergence of the Phaeophyceae lineage. In brown algae and *S. ischiensis* orthogroup OG0007889 contains members with a novel domain structure combining domains (Interpro domains IPR041337 and IPR001660) found independently in orthogroups OG0006687 and OG0001104.

(C) Inference of HGT origins from 74 species based on monophyletic most similar homologue (MMSH) analysis. The tree on the left, which is derived from the NCBI taxonomy tree, indicates the source taxa for HGT-derived genes. The tree at the bottom of the figure, which was constructed using single-copy genes, indicates the species that have received genes by HGT. The middle part of the figure indicates the number of HGT genes transferred from each source to the receiver genomes. The panel on the right illustrates the number of HGTs from each phylum. The legend in the lower left corner provides a reference for the circle size, which corresponds to 10, 100 or 300 HGT gene counts.

(D) Comparison of GC content and gene structure for different components of HGT-derived (HGT) and non-HGT-derived (core) genes. tssup, 5' untranslated region; tssdown, 3' untranslated region.
(E) Pie charts comparing proportions of COG functional categories for HGT-derived (right) and all (left) genes. Numbers represent the percentage of the COG category for the considered gene set.

Figure S5. Genome-scale metabolic network analyses, related to Figure 3.
(A-D). Multidimensional-scaling (MDS) plots of GSMNs reconstructed with AuCoMe using the presence and absence of biochemical reactions in these networks to compute the distance matrix (Jaccard index) used by the MDS.
(A) MDS computed using the draft metabolic networks created by the first step of AuCoMe with only the annotation from the genomes on all species from the Phaeoexplorer dataset.
(B) MDS computed using metabolic networks created after the last step of AuCoMe, i.e. after reaction propagation using orthologous genes and structural verification on genome sequence, on all genomes from the Phaeoexplorer dataset.
(C) MDS computed on draft metabolic networks from a subset of high quality assemblies.
(D) MDS computed after reaction propagation on a subset of high quality assemblies.
Dim., dimension.
(E-G). Overview of losses and gains of metabolic components for different sets of brown algal species. Each row corresponds to a species and columns correspond to sets of reactions. For each row, there is a coloured block if the species contains the reactions present in this block and shared with the other species that also have this block in the same column. Numbers at the top, along the bottom and to the right indicate the number of species that possess each reaction, the number of reactions with this profile and the number of reactions per species, respectively.
(E) All brown algae and stramenopile outgroups.
(F) Selected reference quality genome assemblies.
(G) Comparison of freshwater and marine species with reference quality genomes.

Figure S6. Analysis of metabolism gene families, related to Figure 4.
(A) Counts of numbers of genes predicted to encode glycosyltransferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs) and all CAZymes (GTs, GHs, PLs, Carbohydrate Esterases CEs, Auxiliary Activities AAs, Carbohydrate Binding Modules CBMs), showing numbers for both full-length proteins (dark colours) and fragments (light colours). The data is averaged by order with the number of species analysed per order in brackets.

(B) Number of genes for selected CAZyme families in brown algae. Only CAZyme families with at least three members per genome on average, are shown. The species analysed are the same as in (A). Counts include full-length proteins and fragments.

(C) vHPO genes identified in the 21 reference genomes based on sequence homology and active site conservation. vHPO genes are indicated by a green cross or a number and absence by a red cross.

(D) Maximum likelihood phylogenetic trees for 259 algal-type vHPOs (left) and for bacterial-type vHPOs (right). Algal-type I vHPOs are coloured in blue and algal-type II vHPOs in violet. The clades that have structurally or biochemically characterised enzymes are highlighted in red for vBPOs and in yellow for vIPOs. Strongly supported representative branches have been collapsed. The clade names correspond either to taxa or to individual gene names. For the bacterial vHPOs, the FastTree reconstruction tool with 1000 bootstraps was drawn as a circular representation taking the gammaproteobacterial group as the starting point to arbitrary root the tree. Brown algal branches are coloured in brown. Green dots indicate bootstrap values of between 0.7 and 1.0 (1000 replicates).

(E) PKS III domain structures indicating amino- (IPR001099; green) and carboxy-terminal (IPR012328; brown) chalcone/stilbene synthase domains and amino-terminal signal peptide (SP; violet), signal anchor (SA; light blue), SP/SA hybrid (pink) or all three (SP, SA or SP/SA; black). The number of proteins in each group is indicated at the carboxy end of the protein.

Figure S7. Evolution of signalling-related genes in the brown algae, related to Figure 4.

(A) Distribution of transcription factors (TFs) across the brown algae and sister taxa. The bubbles indicate the relative abundance of 32 different TF families.

(B) Distribution of transcriptional regulators (TRs) across the brown algae and sister taxa. The bubbles indicate the relative abundance of 40 different TR families.

(C) Overview of phytohormone biosynthesis pathways and distribution of gene families associated with biosynthesis across the brown algae and other stramenopile taxa. Left, biosynthesis pathways for the phytohormones abscisic acid (ABA), brassinosteroids (BR), cytokinins (CK; including iP, isopentenyladenine; tZ, trans-zeatin; DZ, dihydrozeatin; and cZ, cis-zeatin), ethylene (Eth), gibberellins (GA), auxin (IAA), jasmonic acid (JA), salicylic acid (SA) and strigolactones (SL). Right, Presence of putative homologs of phytohormone biosynthethic enzymes in brown algae based on EggNOG v5.0 gene families (root node). Colours denote z-score of the number of eggNOG family (left of heatmap) members. White denotes absence, (+) and (?) indicate that phytohormones have been reported to be present or have not been reported for brown algae (reviewed in[48,49]). ZEP, zeaxanthin epoxidase; NCED, nine-cis-epoxycarotenoid deoxygenate; SDR, short-chain dehydrogenase reductase; AAO, aldehyde oxidase; DET2, deetiolated2 (steroid 5alfa reductase); DWF4, dwarf4

(CYP90B); CPD, constitutive photomorphogenic dwarf (CYP90A); IPT, isopentenyl transferase; LOG, lonely guy (lysine de-carboxylase); SAMS, s-adenosyl methionine synthetase; ACS, ACC synthase; ACO, ACC oxidase; CPS, CDP/ent-kaurene synthase; KO, ent-kaurene oxidase; KAO, ent-kaurenoic acid oxidase; TAA, tryptophan aminotransferase; YUC, YUCCA (flavin monooxygenase); AMI, amidase; NIT, nitrilase; LOX, lipoxygenase; AOS, allene oxide synthase; AOC, allene oxide cyclase; OPR, oxo-phytodienoate reductase; ICS, isochorismate synthase; PAL, phenylalanine ammonia-lyase; D27, dwarf27 (all-trans/9-cis-B-carotene isomerase); CCD, carotenoid cleavage dioxygenase; MAX1, more axillary branches1 (CYP711A).

(D) Presence or absence of different membrane-localised signalling proteins in brown algae and other stramenopiles (left). The inset (right) indicates the diverse domain structures of stramenopile FG-GAP domain proteins showing possible evolutionary relationships. LRR, leucine-rich repeat; QAD, b-propeller domain; EBD, ethylene-binding domain-like; CHASE, cyclases/histidine kinases associated sensory extracellular domain; MASE1, membrane-associated sensor1 domain; HK, histidine kinase; FG-GAP, phenylalanyl-glycyl–glycyl-alanyl-prolyl domain; IPT, Ig-like plexins transcription factors domain; ITGA, α-integrin domain; Cad, cadherin domain; DEK1, defective kernal1; FAS, fasciculin; SP, signal peptide; TM, transmembrane domain.

Figure S8. Histone structure and evolution, related to Figure 4

(A) Histone H3.1 and H3.3 isoforms differ at two residues located in the amino-terminal tail (31-32 AT for H3.1 and TA for H3.3, in green), three residues in the α2 helix of the histone fold domain (86-90 GSAVL for H3.1 and STAIL for H3.3) and one residue at the carboxy-terminal position (S or A, respectively in H3.1 and H3.3). Positions refer to the mature protein without the initial methionine.

(B) CenH3 proteins differ in their length and the amino acid composition of their amino-terminal tails. Strictly and highly conserved residues are depicted in black and red, respectively. The characteristic CATD (CENP-A targeting domain) is indicated by a red line. Positions refer to the mature protein without the initial methionine.

(C) Phylogenetic tree of histone H3.1 and H3.3 proteins of brown algae and other eukaryotes. Brown algae, diatoms, red algae, green algae, plant, animal and unicellular (the myxomycetes *Physarum polycephalum* and *Dictyostellium discoideum*, the ciliate *Tetrahymena thermopila* and the yeast *S. cerevisiae*) are depicted in brown, red, pink, light green, dark green, blue and black, respectively. Atr, *Amborella trichopoda*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Di, *D. discoideum*; Dr, *Danio rerio*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Pp, *Physarum polycephalum*; Ppa *Physcomitrium patens*; Sc, *Saccharomyces cerevisiae*; Tm, *Tetrahymena thermophila*; Zm, *Zea mays*; Mp, *Marchantia polymorpha* subsp. *Ruderalis*; Bd, *Brachypodium distachyon*; Ccr, *Chondrus crispus*; Gs, *Galdieria sulphuraria*; Cm, *Cyanidioschyzon merolae*; Cr,

Chlamydomonas reinhardtii; Ol, *Ostreococcus luciminarinus*; Ot, *Ostreococcus tauri*; To, *Thalassiosira oceanica*; Pt, *Phaeodactylum tricornutum*; An, *Ascophyllum nodosum*; Cl, *Chordaria linearis*; Ca, *Chrysoparadoxa australica*; Dh, *Desmarestia herbacea*; Ddi, *Dictyota dichotoma*; Dme, *Discosporangium mesarthrocarpum*; Ec, *Ectocarpus crouaniorum*; Ef, *Ectocarpus fasciculatus*; Es, *Ectocarpus siliculosus*; Fse, *Fucus serratus*; Ha, *Heterosigma akashiwo*; Pla, *Pleurocladia lacustris*; Pf, *Porterinema fluviatile*; Pli, *Pylaiella littoralis*; Sl, *Saccharina latissima*; Sf, *Sargassum fusiform*; Si, *Schizocladia ischiensis*; Sp, *Scytosiphon promiscuus*; Sri, *Sphacelaria rigidula*; Tm, *Tribonema minus*; Up, *Undaria pinnatifida*.

(D) Schematic presentation of brown algal histone H1 linker proteins. The H1 variants are highly divergent across brown algae species. Strictly and highly conserved residues are depicted in black and red, respectively. Positions refer to the mature protein without the initial methionine.

Figure S9. Evolution associated with diversification of biological traits across the brown algae, related to Figure 2.

(A) Ancestral state reconstructions for number of cell types and for morphological complexity in the diploid or haploid phase of the life cycle. Left, number of different cell types in the diploid phase (DP). Centre, morphological complexity in the diploid phase. Right, morphological complexity in the haploid phase (HP). The colours indicate the different states and the pie charts the likelihood of these states at each node.

(B) Generation-biased gene expression in relation to life cycle dimorphism. Pie diagrams indicating the proportions of sporophyte-specific (dark red), sporophyte-biased (light red), gametophyte-specific (dark blue), gametophyte-biased (light blue) and unbiased (grey) genes in species with different haploid-diploid life cycles ranging from sporophyte-dominant through isomorphic to gametophyte-dominant.

(C) Rates of gene evolution in relation to life cycle structure and developmental complexity. Left, violin plots showing the distribution of omega (dN/dS), rates of non-synonymous substitution (dN) and rates of synonymous substitution (dS) for brown algae with haplodiplontic or diplontic life cycles. Right, violin plots showing the distribution of rates of non-synonymous substitution (dN) for brown algal species with filamentous, simple parenchymatous or elaborate parenchymatous thalli. The *p*-values are for pairwise (gene-by-gene) Wilcoxon tests and the percentage of genes that exhibited the same patterns of differences in dN/dS, dN or dS values as the median values are indicated. Significant differences are indicated by an asterisk.

Figure S10. Evolution of photosynthesis genes and organellar genomes, related to Figure 4.

(A) LHCX cluster genomic context in ten brown algae and *Schizocladia ischiensis*. Text above and below the line indicates the chromosome (chr) or contig (ctg) and the gene number, respectively.
(B) Maximum likelihood (RAxML) phylogenetic tree of all LHCX and selected FCP protein sequences from the algal species in panel (A). Sequences are labelled with the first letters of the genus and species name, contig and gene number. Clustered LHCX genes are additionally numbered as in panel (A). Three groups of conserved LHCX paralogs encoded by unclustered, unique genes are highlighted by blue background.
(C) Examples of LHCX gene clusters (LHCX genes in red) in four brown algal species where the LHCX genes are located on opposite strands of the chromosome and overlap, either partially, with some of their exons being located in the intron of the gene on the opposite strand (*C. linearis, S. latissima*), or completely, with the gene being located entirely within the intron of the gene on the opposite strand (*E. crouaniorum, S. promiscuus*).
(D) Loss of plastid genes during the evolution of the brown algae. Gene loss events were identified for five genes, with two of the genes (rbcR and rpl32) being lost more than once.
(E) Maximum likelihood phylogenetic trees for the brown algae based on genes from different genomic compartments. Phylogenetic tree based on plastid genes (top), phylogenetic tree based on mitochondrial genes (middle), discordance between the nuclear and plastid phylogenies and the mitochondrial phylogeny for the Ectocarpales (bottom). Ultrafast bootstrap values are indicated for 1000 replicates, with an asterisk indicating 100%.

Figure S11. Amplification of transcription factor gene families, related to Figure 4.
(A) Ancestral state reconstruction of the sizes of three TAP families: left, C2h2 zinc finger, centre, three-amino-acid loop extension homeodomain (HD_TALE) and, right, high mobility group (HMG). The colour gradient indicates the predicted number of genes in each family across the tree.
(B) Alignment of the homeodomain regions of three-amino-acid loop extension homeodomain transcription factors (TALE HD TFs). Sequences in clusters 1 and 3 are orthologous to the ORO and SAM proteins of *Ectocarpus* species 7, respectively. Cluster 2 corresponds to sequences similar to the third TALE HD TF gene of *Ectocarpus* species 7. Cluster 4 contains the sequences from sister taxa that could not be clearly assigned orthology to any of the *Ectocarpus* species 7 proteins.

Figure S12. Conservation and microevolution of genomes within the genus *Ectocarpus,* related to Figure 7.
(A) Dotplot illustrating genome-wide synteny between the *Ectocarpus* species 7 and *E. crouaniorum* genome assemblies. Horizontal lines delimit *E. crouaniorum* contigs.

(B) Conservation of intron positions in *Ectocarpus* species and the outgroup *S. promiscuus* (right panel and insert).

(C) Predicted functions of genes in orthogroups gained within the *Ectocarpus* genus.

(D) Unrooted parsimony splits network of 11 *Ectocarpus* species. The tree was generated based on 257 concatenated orthologous genes representing a total 274,196 nucleotides using the ParsimonySplits method in Splitstree v.4.14.6[107]. ParsimonySplits networks accommodate phylogenetic incongruities in the data by incorporating alternative branch splits. Conflicting splits are displayed as box-like structures.

(E) Principal coordinate analysis (PCoA) of distance branching patterns (Kendall-Colijn metric) for gene trees. The colour of each dot signifies the number of positively selected codons identified in orthologues with a significance level of $P < 0.01$ by phylogenetic analysis by maximum likelihood (PAML) analyses: red for zero selected codons, green for between one and four selected codons and blue for at least five selected codons.

Figure S13. Genome assembly procedures, related to Figure 2.

(A) Short read assembly procedure.

(B) Long read assembly procedure.

Supplementary tables

Table S1. List of strains used for the project, the genome assemblies generated and accession numbers for sequence data.

(A) Strains used in the project.

(B) Genome assemblies.

(C) Statistical tests of correlations between genome assembly size and various genome features.

Table S2. Characteristics of plastid and mitochondrial genomes for 33 brown algae and *Chrysoparadoxa australica*.

Table S3. Spliceosome components identified in brown algal and sister taxa genomes.

M, male; F, female.

Table S4. Statistics for the long non-coding RNA content of 11 genomes.

M, male; F, female.

Table S5. Phylostratigraphy analysis.

(A) Gene ages estimated by phylostratigraphy. (B) Gene family founder events. (C) Gene ages after the homology detection failure correction. (D) Counts of founder events after the homology detection failure correction. (E) Statistics for *Ectocarpus* species 7.

Table S6. Gene families (orthogroups) significantly amplified in Phaeophyceae genomes compared to sister taxa.

OG, orthogroup; PHAEO, Phaeophyceae; FDI, FDI clade; LAMIN, Laminariales; ECTO, Ectocarpales.

Table S7. Core metabolic reactions most abundant in brown algae.

Metacyc IDs and reaction names (with EC numbers) for 24 genes present in all brown algae and less than 70% of outgroup species. Gene IDs can be retrieved from the GSMN Wikis using reaction IDs.

Table S8. Counts of CAZYme gene family members brown algal and sister taxa genomes.

Asterisks indicate species that were included in Figures S6A and S6B.

Table S9. CAZYmes identified in brown algal and sister taxa genomes.

ProteinID, corresponds to the locusID; Description, CAZYme gene family.

Table S10. Sulphatase proteins encoded by brown algal and sister taxa genomes.

Table S11. Algal-type vanadium haloperoxidase proteins encoded by brown algal and sister taxa genomes plus representative sequences from more distant taxa.

Table S12. Bacterial-type vanadium haloperoxidase proteins encoded by brown algal and sister taxa genomes plus representative sequences from more distant taxa.

Table S13. Type III polyketide synthase proteins encoded by brown algal and sister taxa genomes plus representative sequences from more distant taxa.

Table S14. Mannitol cycle enzymes encoded by brown algal and sister taxa genomes.

M1PDH, mannitol 1-phosphate dehydrogenase; M1Pase, mannitol 1-phosphate phosphatase; HK, hexokinase; M2DH, mannitol 2-dehydrogenase.

Table S15. Presence of flagellar protein genes in the genomes of nine brown algal species.

P, present; A, absent.

Table S16. Summary of the TAPscan output with lists of annotated transcription-associated proteins (TAPs) for brown algal and sister taxa genomes.

Table S17. Putative components of phytohormone biosynthetic pathways in brown algae.

Table S18. Receptor kinase proteins encoded by brown algal and sister taxa genomes.
mod, modified gene model; new, new gene model.

Table S19.Histidine kinase proteins encoded by brown algal and sister taxa genomes.
TM, transmembrane domain; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor1 domain.

Table S20. Integrin proteins encoded by brown algal and sister taxa genomes.
Genome coordinates and protein sequences are provided for modified (mod) or newly created (new) gene models. SP, signal peptide; TM, transmembrane domain; INTA, integrin alpha domain

Table S21. DEK1-like proteins encoded by brown algal and sister taxa genomes.

Table S22. Fasciclin proteins encoded by brown algal and sister taxa genomes.
Genome coordinates and protein sequences are provided for modified (mod) gene models. FAS1, fasciclin domain.

Table S23. Tetraspanin and tetraspanin-like proteins encoded by brown algal and sister taxa genomes.

Table S24. Ion channels encoded by brown algae and other stramenopiles.
(A) Query sequences used for the screens. (B) Ion channel proteins detected. (C) Number of ion channels of each class per species. F, female; M, male; MCU, mitochondrial calcium uniporter; GLR, glutamate receptor; pLGIC, pentameric ligand-gated ion channel; TRP, transient receptor potential channel

Table S25. EsV-1-7 domain proteins encoded by brown algal and sister taxa genomes.

Genome coordinates and protein sequences are provided for modified (mod) or newly created (new) gene models.

Table S26.Histone proteins encoded by brown algae and other stramenopiles.
(A) Histone protein sequences. (B) Counts of histone proteins per genome. (C) Counts of histone genes per genome.

Table S27. DNA methyltransferase proteins encoded by brown algal and sister taxa genomes.
Genome coordinates and protein sequences are provided for modified (mod) gene models.

Table S28. Predicted reference proteomes of the ten species analysed for generation-biased gene expression indicating sporophyte- and gametophyte-biased genes.

Table S29. Presence or absence of 141 shared plastid genes across the brown algae.
O, present; X, absent.

Table S30. *Porterinema fluviatile* genes differentially expressed under freshwater compared to seawater culture conditions.

Table S31. Comparative analysis of genes that were differentially expressed in freshwater compared with seawater in *E. subulatus* and *P. fluviatile* based on whether the genes are shared orthologues or lineage-specific.

Table S32. Metabolism genes lost from either *Pleurocladia lacustris* or *Porterinema fluviatile* or from both, based on GSMN analysis.

Table S33. Inserted viral sequences in brown algal and sister taxa genomes.
(A) Summary data for figure 5. (B) List of viral genes. (C) List of viral regions.

Table S34. Blocks of syntenic genes shared between *Ectocarpus* species 7 and four other *Ectocarpus* species.

Table S35. Rates of synonymous and non-synonymous substitutions based on alignments of genes in syntenic blocks between *Ectocarpus* species 7 and four other *Ectocarpus* species.

Table S36. Genomes studied for the various analyses carried out within the project.

Y, genome analysed.

Table S37. Tests for signatures of introgression among *Ectocarpus* species. Patterson's D-statistic (ABBA-BABA tests) was calculated for concatenated alignments of 257 ortholog genes (~274 Kbp) and significance was detected using a block-jackknifing approach with a block size of 5 Kbp. *Ectocarpus fasciculatus* was used as the out-group taxon for all ABBA-BABA tests (noted as O in the four-taxon fixed phylogeny scheme: (((P1,P2)P3)O)). All values are significant.

References

1. Eger, A.M., Marzinelli, E.M., Beas-Luna, R., Blain, C.O., Blamey, L.K., Byrnes, J.E.K., Carnell, P.E., Choi, C.G., Hessing-Lewis, M., Kim, K.Y., et al. (2023). The value of ecosystem services in global marine kelp forests. Nat Commun *14*, 1894. 10.1038/s41467-023-37385-0.

2. Wernberg, T., Russell, B.D., Thomsen, M.S., Gurgel, C.F.D., Bradshaw, C.J.A., Poloczanska, E.S., and Connell, S.D. (2011). Seaweed communities in retreat from ocean warming. Curr Biol *21*, 1828–1832. 10.1016/j.cub.2011.09.028.

3. Ross, F.W.R., Boyd, P.W., Filbee-Dexter, K., Watanabe, K., Ortega, A., Krause-Jensen, D., Lovelock, C., Sondak, C.F.A., Bach, L.T., Duarte, C.M., et al. (2023). Potential role of seaweeds in climate change mitigation. Sci Total Environ *885*, 163699. 10.1016/j.scitotenv.2023.163699.

4. Bringloe, T.T., Starko, S., Wade, R.M., Vieira, C., Kawai, H., Clerck, O.D., Cock, J.M., Coelho, S.M., Destombe, C., Valero, M., et al. (2020). Phylogeny and Evolution of the Brown Algae. Critical Reviews in Plant Sciences *39*, 281–321. 10.1080/07352689.2020.1787679.

5. Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J., Badger, J., et al. (2010). The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature *465*, 617–621. 10.1038/nature09016.

6. Ye, N., Zhang, X., Miao, M., Fan, X., Zheng, Y., Xu, D., Wang, J., Zhou, L., Wang, D., Gao, Y., et al. (2015). *Saccharina* genomes provide novel insight into kelp biology. Nat Commun *6*, 6986. 10.1038/ncomms7986.

7. Nishitsuji, K., Arimoto, A., Iwai, K., Sudo, Y., Hisata, K., Fujie, M., Arakaki, N., Kushiro, T., Konishi, T., Shinzato, C., et al. (2016). A draft genome of the brown alga, *Cladosiphon okamuranus*, S-strain: a platform for future studies of "mozuku" biology. DNA Res. *23*, 561–570. 10.1093/dnares/dsw039.

8. Nishitsuji, K., Arimoto, A., Higa, Y., Mekaru, M., Kawamitsu, M., Satoh, N., and Shoguchi, E. (2019). Draft genome of the brown alga, *Nemacystus decipiens*, Onna-1 strain: Fusion of genes involved in the sulfated fucan biosynthesis pathway. Sci Rep *9*, 4607. 10.1038/s41598-019-40955-2.

9. Dittami, S.M., Corre, E., Brillet-Guéguen, L., Lipinska, A.P., Pontoizeau, N., Aite, M., Avia, K., Caron, C., Cho, C.H., Collén, J., et al. (2020). The genome of *Ectocarpus subulatus* - A highly stress-tolerant brown alga. Mar Genomics, 100740. 10.1016/j.margen.2020.100740.

10. Graf, L., Shin, Y., Yang, J.H., Choi, J.W., Hwang, I.K., Nelson, W., Bhattacharya, D., Viard, F., and Yoon, H.S. (2021). A genome-wide investigation of the effect of farming and human-mediated introduction on the ubiquitous seaweed *Undaria pinnatifida*. Nature Ecology & Evolution. 10.1038/s41559-020-01378-9.

11. Wang, S., Lin, L., Shi, Y., Qian, W., Li, N., Yan, X., Zou, H., and Wu, M. (2020). First Draft Genome Assembly of the Seaweed *Sargassum fusiforme*. Front Genet *11*, 590065. 10.3389/fgene.2020.590065.

12. Wang, S., and Wu, M. (2023). The Draft Genome of the "Golden Tide" Seaweed, *Sargassum horneri*: Characterization and Comparative Analysis. Genes (Basel) *14*, 1969. 10.3390/genes14101969.

13. Diesel, J., Molano, G., Montecinos, G.J., DeWeese, K., Calhoun, S., Kuo, A., Lipzen, A., Salamov, A., Grigoriev, I.V., Reed, D.C., et al. (2023). A scaffolded and annotated reference genome of giant kelp (*Macrocystis pyrifera*). BMC Genomics *24*, 543. 10.1186/s12864-023-09658-x.

14. Bourdareau, S., Tirichine, L., Lombard, B., Loew, D., Scornet, D., Wu, Y., Coelho, S.M., and Cock, J.M. (2021). Histone modifications during the life cycle of the brown alga *Ectocarpus*. Genome Biology *22*, 12.

15. Roy, S.W., and Penny, D. (2007). A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. Mol Biol Evol *24*, 1447–1457. 10.1093/molbev/msm048.

16. Larue, G.E., and Roy, S.W. (2023). Where the minor things are: a pan-eukaryotic survey suggests neutral processes may explain much of minor intron evolution. Nucleic Acids Research *51*, 10884–10908. 10.1093/nar/gkad797.

17. Vosseberg, J., Stolker, D., von der Dunk, S.H.A., and Snel, B. (2023). Integrating Phylogenetics With Intron Positions Illuminates the Origin of the Complex Spliceosome. Mol Biol Evol *40*, msad011. 10.1093/molbev/msad011.

18. Veretnik, S., Wills, C., Youkharibache, P., Valas, R.E., and Bourne, P.E. (2009). Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. PLoS Comput Biol *5*, e1000315. 10.1371/journal.pcbi.1000315.

19. Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404. 10.1126/science.1089370.

20. Yang, P., Wang, D., and Kang, L. (2021). Alternative splicing level related to intron size and organism complexity. BMC Genomics *22*, 853. 10.1186/s12864-021-08172-2.

21. Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463. 10.1038/nature08909.

22. Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and Urrutia, A.O. (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol. Biol. Evol. *31*, 1402–1413. 10.1093/molbev/msu083.

23. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol *22*, 96–118. 10.1038/s41580-020-00315-9.

24. Csorba, T., Questa, J.I., Sun, Q., and Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. Proc Natl Acad Sci U S A *111*, 16160–16165. 10.1073/pnas.1419030111.

25. Fan, X., Han, W., Teng, L., Jiang, P., Zhang, X., Xu, D., Li, C., Pellegrini, M., Wu, C., Wang, Y., et al. (2020). Single-base methylome profiling of the giant kelp *Saccharina japonica* reveals significant differences in DNA methylation to microalgae and plants. New Phytol. *225*, 234–249. 10.1111/nph.16125.

26. Cormier, A., Avia, K., Sterck, L., Derrien, T., Wucher, V., Andres, G., Monsoor, M., Godfroy, O., Lipinska, A., Perrineau, M.-M., et al. (2017). Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. New Phytol. *214*, 219–232. 10.1111/nph.14321.

27. Yang, X., Li, L., Wang, X., Yao, J., and Duan, D. (2020). Non-Coding RNAs Participate in the Regulation of CRY-DASH in the Growth and Early Development of *Saccharina japonica* (Laminariales, Phaeophyceae). Int J Mol Sci *21*, 309. 10.3390/ijms21010309.

28. Belcour, A., Got, J., Aite, M., Delage, L., Collén, J., Frioux, C., Leblanc, C., Dittami, S.M., Blanquart, S., Markov, G.V., et al. (2023). Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. Genome Res. *33*, 972–987. 10.1101/gr.277056.122.

29. Mazéas, L., Yonamine, R., Barbeyron, T., Henrissat, B., Drula, E., Terrapon, N., Nagasato, C., and Hervé, C. (2023). Assembly and synthesis of the extracellular matrix in brown algae. Seminars in Cell & Developmental Biology *134*, 112–124. 10.1016/j.semcdb.2022.03.005.

30. Stam, M., Lelièvre, P., Hoebeke, M., Corre, E., Barbeyron, T., and Michel, G. (2023). SulfAtlas, the sulfatase database: state of the art and new developments. Nucleic Acids Res *51*, D647–D653. 10.1093/nar/gkac977.

31. Küpper, F.C., and Carrano, C.J. (2019). Key aspects of the iodine metabolism in brown algae: a brief critical review. Metallomics *11*, 756–764. 10.1039/c8mt00327k.

32. Wischang, D., Radlow, M., Schulz, H., Vilter, H., Viehweger, L., Altmeyer, M.O., Kegler, C., Herrmann, J., Müller, R., Gaillard, F., et al. (2012). Molecular cloning, structure, and reactivity of the second bromoperoxidase from *Ascophyllum nodosum*. Bioorg Chem *44*, 25–34. 10.1016/j.bioorg.2012.05.003.

33. Radlow, M., Czjzek, M., Jeudy, A., Dabin, J., Delage, L., Leblanc, C., and Hartung, J. (2018). X-ray Diffraction and Density Functional Theory Provide Insight into Vanadate Binding to Homohexameric Bromoperoxidase II and the Mechanism of Bromide Oxidation. ACS Chem Biol *13*, 1243–1259. 10.1021/acschembio.8b00041.

34. Fournier, J.-B., Rebuffet, E., Delage, L., Grijol, R., Meslet-Cladière, L., Rzonca, J., Potin, P., Michel, G., Czjzek, M., and Leblanc, C. (2014). The Vanadium Iodoperoxidase from the marine flavobacteriaceae species *Zobellia galactanivorans* reveals novel molecular and evolutionary features of halide specificity in the vanadium haloperoxidase enzyme family. Appl Environ Microbiol *80*, 7561–7573. 10.1128/AEM.02430-14.

35. Meslet-Cladière, L., Delage, L., Leroux, C.J., Goulitquer, S., Leblanc, C., Creis, E., Gall, E.A., Stiger-Pouvreau, V., Czjzek, M., and Potin, P. (2013). Structure/Function Analysis of a Type III

Polyketide Synthase in the Brown Alga *Ectocarpus siliculosus* Reveals a Biochemical Pathway in Phlorotannin Monomer Biosynthesis. Plant Cell *25*, 3089–3103. 10.1105/tpc.113.111336.

36. Baharum, H., Morita, H., Tomitsuka, A., Lee, F.C., Ng, K.Y., Rahim, R.A., Abe, I., and Ho, C.L. (2011). Molecular cloning, modeling, and site-directed mutagenesis of type III polyketide synthase from *Sargassum binderi* (Phaeophyta). Mar Biotechnol (NY) *13*, 845–856. 10.1007/s10126-010-9344-5.

37. Zhao, D.-S., Hu, Z.-W., Dong, L.-L., Wan, X.-J., Wang, S., Li, N., Wang, Y., Li, S.-M., Zou, H.-X., and Yan, X. (2021). A Type III Polyketide Synthase (SfuPKS1) Isolated from the Edible Seaweed Sargassum fusiforme Exhibits Broad Substrate and Catalysis Specificity. J Agric Food Chem *69*, 14643–14649. 10.1021/acs.jafc.1c05868.

38. Schoenwaelder, M.E.A., and Wiencke, C. (2000). Phenolic Compounds in the Embryo Development of Several Northern Hemisphere Fucoids. Plant Biology *2*, 24–33.

39. Jégou, C., Kervarec, N., Cérantola, S., Bihannic, I., and Stiger-Pouvreau, V. (2015). NMR use to quantify phlorotannins: the case of Cystoseira tamariscifolia, a phloroglucinol-producing brown macroalga in Brittany (France). Talanta *135*, 1–6. 10.1016/j.talanta.2014.11.059.

40. Jégou, C., Connan, S., Bihannic, I., Cérantola, S., Guérard, F., and Stiger-Pouvreau, V. (2021). Phlorotannin and Pigment Content of Native Canopy-Forming Sargassaceae Species Living in Intertidal Rockpools in Brittany (France): Any Relationship with Their Vertical Distribution and Phenology? Mar Drugs *19*, 504. 10.3390/md19090504.

41. Salgado, L.T., Cinelli, L.P., Viana, N.B., Tomazetto de Carvalho, R., De Souza Mourão, P.A., Teixeira, V.L., Farina, M., and Filho, A.G.M.A. (2009). A VANADIUM BROMOPEROXIDASE CATALYZES THE FORMATION OF HIGH-MOLECULAR-WEIGHT COMPLEXES BETWEEN BROWN ALGAL PHENOLIC SUBSTANCES AND ALGINATES(1). J Phycol *45*, 193–202. 10.1111/j.1529-8817.2008.00642.x.

42. Berglin, M., Delage, L., Potin, P., Vilter, H., and Elwing, H. (2004). Enzymatic cross-linking of a phenolic polymer extracted from the marine alga Fucus serratus. Biomacromolecules *5*, 2376–2383.

43. Bitton, R., Berglin, M., Elwing, H., Colin, C., Delage, L., Potin, P., and Bianco-Peled, H. (2007). The influence of halide-mediated oxidation on algae-born adhesives. Macromolecular Bioscience *7*, 1280–1289.

44. Arnold, T.M., and Targett, N.M. (2003). To grow and defend: lack of tradeoffs for brown algal phlorotannins. Oikos *100*, 406–408. 10.1034/j.1600-0706.2003.11680.x.

45. Salgado, L.T., Tomazetto, R., Cinelli, L.P., Farina, M., and Amado Filho, G.M. (2007). The influence of brown algae alginates on phenolic compounds capability of ultraviolet radiation absorption in vitro. Braz. j. oceanogr. *55*, 145–154.

46. Michel, G., Tonon, T., Scornet, D., Cock, J.M., and 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. 10.1111/j.1469-8137.2010.03345.x.

47. Fu, G., Nagasato, C., Oka, S., Cock, J.M., and Motomura, T. (2014). Proteomics Analysis of Heterogeneous Flagella in Brown Algae (Stramenopiles). Protist *165*, 662–675. 10.1016/j.protis.2014.07.007.

48. Stirk, W.A., and Van Staden, J. (2014). Chapter Five - Plant Growth Regulators in Seaweeds: Occurrence, Regulation and Functions. In Advances in Botanical Research Sea Plants., N. Bourgougnon, ed. (Academic Press), pp. 125–159. 10.1016/B978-0-12-408062-1.00005-6.

49. Lu, Y., and Xu, J. (2015). Phytohormones in microalgae: a new opportunity for microalgal biotechnology? Trends Plant Sci *20*, 273–282. 10.1016/j.tplants.2015.01.006.

50. Hamdy, A.-H.A., Aboutabl, E.A., Sameer, S., Hussein, A.A., Díaz-Marrero, A.R., Darias, J., and Cueto, M. (2009). 3-Keto-22-epi-28-nor-cathasterone, a brassinosteroid-related metabolite from *Cystoseira myrica*. Steroids *74*, 927–930. 10.1016/j.steroids.2009.06.008.

51. Nègre, D., Aite, M., Belcour, A., Frioux, C., Brillet-Guéguen, L., Liu, X., Bordron, P., Godfroy, O., Lipinska, A.P., Leblanc, C., et al. (2019). Genome-Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae *Saccharina japonica* and *Cladosiphon okamuranus*. Antioxidants (Basel) *8*. 10.3390/antiox8110564.

52. Chini, A., Monte, I., Zamarreño, A.M., García-Mina, J.M., and Solano, R. (2023). Evolution of the jasmonate ligands and their biosynthetic pathways. New Phytologist *238*, 2236–2246. 10.1111/nph.18891.

53. Kloareg, B., Badis, Y., Cock, J.M., and Michel, G. (2021). Role and Evolution of the Extracellular Matrix in the Acquisition of Complex Multicellularity in Eukaryotes: A Macroalgal Perspective. Genes *12*, 1059. 10.3390/genes12071059.

54. Kabbara, S., Hérivaux, A., Dugé de Bernonville, T., Courdavault, V., Clastre, M., Gastebois, A., Osman, M., Hamze, M., Cock, J.M., Schaap, P., et al. (2019). Diversity and Evolution of Sensor Histidine Kinases in Eukaryotes. Genome Biol Evol *11*, 86–108. 10.1093/gbe/evy213.

55. Tran, D., Galletti, R., Neumann, E.D., Dubois, A., Sharif-Naeini, R., Geitmann, A., Frachisse, J.-M., Hamant, O., and Ingram, G.C. (2017). A mechanosensitive Ca2+ channel activity is dependent on the developmental regulator DEK1. Nat Commun *8*, 1009. 10.1038/s41467-017-00878-w.

56. Seifert, G.J. (2018). Fascinating Fasciclins: A Surprisingly Widespread Family of Proteins that Mediate Interactions between the Cell Exterior and the Cell Surface. International Journal of Molecular Sciences *19*, 1628. 10.3390/ijms19061628.

57. Roberts, S., and Brownlee, C. (1995). Calcium influx, fertilisation potential and egg activation in *Fucus serratus*. Zygote *3*, 191–197.

58. Fujiu, K., Nakayama, Y., Yanagisawa, A., Sokabe, M., and Yoshimura, K. (2009). *Chlamydomonas CAV2* encodes a voltage- dependent calcium channel required for the flagellar waveform conversion. Curr Biol *19*, 133–139. 10.1016/j.cub.2008.11.068.

59. Verret, F., Wheeler, G., Taylor, A.R., Farnham, G., and Brownlee, C. (2010). Calcium channels in photosynthetic eukaryotes: implications for evolution of calcium-based signalling. New Phytol *187*, 23–43. 10.1111/j.1469-8137.2010.03271.x.

60. Yuan, F., Yang, H., Xue, Y., Kong, D., Ye, R., Li, C., Zhang, J., Theprungsirikul, L., Shrift, T., Krichilsky, B., et al. (2014). OSCA1 mediates osmotic-stress-evoked Ca2+ increases vital for osmosensing in *Arabidopsis*. Nature *514*, 367–371. 10.1038/nature13593.

61. Murthy, S.E., Dubin, A.E., Whitwam, T., Jojoa-Cruz, S., Cahalan, S.M., Mousavi, S.A.R., Ward, A.B., and Patapoutian, A. (2018). OSCA/TMEM63 are an Evolutionarily Conserved Family of Mechanically Activated Ion Channels. Elife *7*, e41844. 10.7554/eLife.41844.

62. Helliwell, K.E., Chrachri, A., Koester, J.A., Wharam, S., Verret, F., Taylor, A.R., Wheeler, G.L., and Brownlee, C. (2019). Alternative Mechanisms for Fast Na+/Ca2+ Signaling in Eukaryotes via a Novel Class of Single-Domain Voltage-Gated Channels. Curr Biol *29*, 1503-1511.e6. 10.1016/j.cub.2019.03.041.

63. Macaisne, N., Liu, F., Scornet, D., Peters, A.F., Lipinska, A., Perrineau, M.-M., Henry, A., Strittmatter, M., Coelho, S.M., and Cock, J.M. (2017). The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development *144*, 409–418. 10.1242/dev.141523.

64. Marzluff, W., and Duronio, R. (2002). Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences. Curr Opin Cell Biol *14*, 692–699.

65. Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science *293*, 1098–1102. 10.1126/science.1062939.

66. Shi, L., Wen, H., and Shi, X. (2017). The Histone Variant H3.3 in Transcriptional Regulation and Human Disease. J Mol Biol *429*, 1934–1945. 10.1016/j.jmb.2016.11.019.

67. Okada, T., Endo, M., Singh, M.B., and Bhalla, P.L. (2005). Analysis of the histone H3 gene family in *Arabidopsis* and identification of the male-gamete-specific variant AtMGH3. Plant J *44*, 557–568. 10.1111/j.1365-313X.2005.02554.x.

68. Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.-E., Chen, J., Lazar, M.A., et al. (2008). DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. Mol. Cell. Biol. *28*, 2825–2839. 10.1128/MCB.02076-07.

69. Choi, S.-W., Graf, L., Choi, J.W., Jo, J., Boo, G.H., Kawai, H., Choi, C.G., Xiao, S., Knoll, A.H., Andersen, R.A., et al. (2024). Ordovician origin and subsequent diversification of the brown algae. Curr Biol, S0960-9822(23)01769-4. 10.1016/j.cub.2023.12.069.

70. Cock, J.M., Godfroy, O., Macaisne, N., Peters, A.F., and Coelho, S.M. (2014). Evolution and regulation of complex life cycles: a brown algal perspective. Curr Opin Plant Biol *17*, 1–6.

71. Coelho, S.M., Mignerot, L., and Cock, J.M. (2019). Origin and evolution of sex-determination systems in the brown algae. New Phytologist *222*, 1751–1756.

72. Lipinska, A.P., Serrano-Serrano, M.L., Cormier, A., Peters, A.F., Kogame, K., Cock, J.M., and Coelho, S.M. (2019). Rapid turnover of life-cycle-related genes in the brown algae. Genome Biol. *20*, 35. 10.1186/s13059-019-1630-6.

73. Coelho, S., Peters, A.F., Charrier, B., Destombe, C., Valero, M., and Cock, J. (2007). Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms. Gene *406*, 152–170.

74. Kawai, H. (1992). A summary of the Morphology of Chloroplasts and Flagellated Cells in the Phaeophyceae. Algae *7*, 33–43.

75. Kinoshita, N., Nagasato, C., and Motomura, T. (2017). Phototaxis and chemotaxis of brown algal swarmers. J Plant Res *130*, 443–453. 10.1007/s10265-017-0914-8.

76. Buck, J.M., Sherman, J., Bártulos, C.R., Serif, M., Halder, M., Henkel, J., Falciatore, A., Lavaud, J., Gorbunov, M.Y., Kroth, P.G., et al. (2019). Lhcx proteins provide photoprotection via thermal dissipation of absorbed light in the diatom *Phaeodactylum tricornutum*. Nat Commun *10*, 4167. 10.1038/s41467-019-12043-6.

77. Dittami, S.M., Michel, G., Collen, J., Boyen, C., and Tonon, T. (2010). Chlorophyll-binding proteins revisited - a multigenic family of light-harvesting and stress proteins from a brown algal perspective. BMC Evol Biol *10*, 365. 10.1186/1471-2148-10-365.

78. Hess, K., Oliverio, R., Nguyen, P., Le, D., Ellis, J., Kdeiss, B., Ord, S., Chalkia, D., and Nikolaidis, N. (2018). Concurrent action of purifying selection and gene conversion results in extreme conservation of the major stress-inducible Hsp70 genes in mammals. Sci Rep *8*, 5082. 10.1038/s41598-018-23508-x.

79. Hanikenne, M., Kroymann, J., Trampczynska, A., Bernal, M., Motte, P., Clemens, S., and Krämer, U. (2013). Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. PLoS Genet *9*, e1003707. 10.1371/journal.pgen.1003707.

80. Starko, S., Bringloe, T.T., Gomez, M.S., Darby, H., Graham, S.W., and Martone, P.T. (2021). Genomic Rearrangements and Sequence Evolution across Brown Algal Organelles. Genome Biol Evol *13*, evab124. 10.1093/gbe/evab124.

81. Liu, F., Jin, Z., Wang, Y., Bi, Y., and Melton, J.T. (2017). Plastid Genome of Dictyopteris divaricata (Dictyotales, Phaeophyceae): Understanding the Evolution of Plastid Genomes in Brown Algae. Mar Biotechnol (NY) *19*, 627–637. 10.1007/s10126-017-9781-5.

82. Akita, S., Vieira, C., Hanyuda, T., Rousseau, F., Cruaud, C., Couloux, A., Heesch, S., Cock, J.M., and Kawai, H. (2022). Providing a phylogenetic framework for trait-based analyses in brown algae: Phylogenomic tree inferred from 32 nuclear protein-coding sequences. Molecular Phylogenetics and Evolution *168*, 107408. 10.1016/j.ympev.2022.107408.

83. Ran, J.-H., Shen, T.-T., Liu, W.-J., Wang, P.-P., and Wang, X.-Q. (2015). Mitochondrial introgression and complex biogeographic history of the genus *Picea*. Mol Phylogenet Evol *93*, 63–76. 10.1016/j.ympev.2015.07.020.

84. Sarver, B.A.J., Herrera, N.D., Sneddon, D., Hunter, S.S., Settles, M.L., Kronenberg, Z., Demboski, J.R., Good, J.M., and Sullivan, J. (2021). Diversification, Introgression, and Rampant Cytonuclear Discordance in Rocky Mountains Chipmunks (Sciuridae: *Tamias*). Syst Biol *70*, 908–921. 10.1093/sysbio/syaa085.

85. Colin, C., Leblanc, C., Michel, G., Wagner, E., Leize-Wagner, E., Van Dorsselaer, A., and Potin, P. (2005). Vanadium-dependent iodoperoxidases in *Laminaria digitata*, a novel biochemical function diverging from brown algal bromoperoxidases. J Biol Inorg Chem *10*, 156–166.

86. Lang, D., and Rensing, S.A. (2015). The Evolution of Transcriptional Regulation in the Viridiplantae and its Correlation with Morphological Complexity. In Evolutionary Transitions to

Multicellular Life: Principles and mechanisms Advances in Marine Genomics., I. Ruiz-Trillo and A. M. Nedelcu, eds. (Springer Netherlands), pp. 301–333. 10.1007/978-94-017-9642-2_15.

87. Arun, A., Coelho, S.M., Peters, A.F., Bourdareau, S., Pérès, L., Scornet, D., Strittmatter, M., Lipinska, A.P., Yao, H., Godfroy, O., et al. (2019). Convergent recruitment of TALE homeodomain life cycle regulators to direct sporophyte development in land plants and brown algae. Elife *8*, e43101. 10.7554/eLife.43101.

88. Pittis, A.A., Goh, V., Cebrian-Serrano, A., Wettmarshausen, J., Perocchi, F., and Gabaldón, T. (2020). Discovery of EMRE in fungi resolves the true evolutionary history of the mitochondrial calcium uniporter. Nat Commun *11*, 4031. 10.1038/s41467-020-17705-4.

89. Ramsey, I.S., Delling, M., and Clapham, D.E. (2006). An introduction to TRP channels. Annu Rev Physiol *68*, 619–647. 10.1146/annurev.physiol.68.040204.100431.

90. Wheeler, G.L., and Brownlee, C. (2008). Ca2+ signalling in plants and green algae--changing channels. Trends Plant Sci *13*, 506–514. 10.1016/j.tplants.2008.06.004.

91. Zambounis, A., Elias, M., Sterck, L., Maumus, F., and Gachon, C.M. (2012). Highly dynamic exon shuffling in candidate pathogen receptors... What if brown algae were capable of adaptive immunity? Mol Biol Evol *29*, 1263–1276. 10.1093/molbev/msr296.

92. Siméon, A., Kridi, S., Kloareg, B., and Hervé, C. (2020). Presence of Exogenous Sulfate Is Mandatory for Tip Growth in the Brown Alga *Ectocarpus subulatus*. Front Plant Sci *11*, 1277. 10.3389/fpls.2020.01277.

93. Müller, D.G., and Knippers, R. (2011). Phaeovirus. In The Springer Index of Viruses, C. Tidona and G. Darai, eds. (Springer), pp. 1259–1263. 10.1007/978-0-387-95919-1_205.

94. McKeown, D.A., Stevens, K., Peters, A.F., Bond, P., Harper, G.M., Brownlee, C., Brown, M.T., and Schroeder, D.C. (2017). Phaeoviruses discovered in kelp (Laminariales). ISME J *11*, 2869–2873. 10.1038/ismej.2017.130.

95. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145. 10.1038/s41586-020-2924-2.

96. Stevens, K., Weynberg, K., Bellas, C., Brown, S., Brownlee, C., Brown, M.T., and Schroeder, D.C. (2014). A novel evolutionary strategy revealed in the phaeoviruses. PLoS One *9*, e86040. 10.1371/journal.pone.0086040.

97. Bräutigam, M., Klein, M., Knippers, R., and Müller, D.G. (1995). Inheritance and meiotic elimination of a virus genome in the host *Ectocarpus siliculosus* (phaeophyceae). J Phycol *31*, 823–827.

98. Guiry, M., and Guiry, G. (2024). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway.

99. Montecinos, A.E., Couceiro, L., Peters, A.F., Desrut, A., Valero, M., and Guillemin, M.-L. (2017). Species delimitation and phylogeographic analyses in the *Ectocarpus* subgroup *siliculosi* (Ectocarpales, Phaeophyceae). J. Phycol. *53*, 17–31. 10.1111/jpy.12452.

100. Barrera-Redondo, J., Lipinska, A.P., Liu, P., Dinatale, E., Cossard, G., Bogaert, K., Hoshino, M., Avia, K., Leiria, G., Avdievich, E., et al. (2024). Origin and evolutionary trajectories of brown algal sex chromosomes. Preprint at bioRxiv, 10.1101/2024.01.15.575685 10.1101/2024.01.15.575685.

101. Martone, P.T., Kost, L., and Boller, M. (2012). Drag reduction in wave-swept macroalgae: Alternative strategies and new predictions. American Journal of Botany *99*, 806–815. 10.3732/ajb.1100541.

102. Potin, P., Bouarab, K., Salaün, J.-P., Pohnert, G., and Kloareg, B. (2002). Biotic interactions of marine algae. Curr Opin Plant Biol *5*, 308–317. 10.1016/s1369-5266(02)00273-x.

103. Tarakhovskaya, E.R. (2014). Mechanisms of bioadhesion of macrophytic algae. Russ J Plant Physiol *61*, 19–25. 10.1134/S1021443714010154.

104. Cock, J.M. (2023). The model system *Ectocarpus*: Integrating functional genomics into brown algal research. Journal of Phycology *59*, 4–8. 10.1111/jpy.13310.

105. Badis, Y., Scornet, D., Harada, M., Caillard, C., Godfroy, O., Raphalen, M., Gachon, C.M.M., Coelho, S.M., Motomura, T., Nagasato, C., et al. (2021). Targeted CRISPR-Cas9-based gene knockouts in the model brown alga *Ectocarpus*. New Phytol *231*, 2077–2091. 10.1111/nph.17525.

106. Shen, Y., Motomura, T., Ichihara, K., Matsuda, Y., Yoshimura, K., Kosugi, C., and Nagasato, C. (2023). Application of CRISPR-Cas9 genome editing by microinjection of gametophytes of *Saccharina japonica* (Laminariales, Phaeophyceae). J Appl Phycol. 10.1007/s10811-023-02940-1.

107. Kloepper, T.H., and Huson, D.H. (2008). Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol *8*, 22. 10.1186/1471-2148-8-22.

METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Descriptions of all sequenced samples have been deposited in the EBI/ENA database		EBI/ENA project PRJEB72149
Critical commercial assays		
OmniPrep Genomic DNA Purification Kit	G Biosciences, St. Louis, MO, USA	
Nucleospin Plant II midi DNA Extraction Kit	Macherey-Nagel, Düren, Germany	
NEBNext DNA Modules Products	New England Biolabs, Ipswich, MA, USA	
NEBNext Sample Reagent Set	New England Biolabs, Ipswich, MA, USA	
Ampure XP	Beckmann Coulter Genomics, Danvers, MA, USA	
Kapa Hifi Hotstart NGS library Amplification kit	Roche, Basel, Switzerland	
Short Read Eliminator Kit	Pacific Biosciences, Menlo Park, CA, USA	
1D Genomic DNA by Ligation	Oxford Nanopore Technologies Ltd, Oxford, UK	SQK-LSK109, SQK-LSK108 or SQK-LSK110
Qiagen RNeasy kit or the Macherey Nagel RNAplus kit	Macherey-Nagel, Düren, Germany	
TruSeq Stranded mRNA Sample Prep	Illumina	
NEBNext Ultra II Directional RNA Library Prep for Illumina	New England BioLabs	
Deposited data		
The sequence data generated by this project has been deposited in the EBI/ENA database	This study.	EBI/ENA project PRJEB72149
Experimental models: Organisms/strains		
The strains used for genome and transcriptome sequencing are listed in Table S1.	Culture collection references are provided where relevant.	See strain names and culture collection accessions for identifiers.
Software and algorithms		
MEGAHIT version 1.1.1	Li et al.[1]	RRID:SCR_018551 https://github.com/voutcn/megahit
MetaGene version 2008.8.19	Noguchi et al.[2]	http://metagene.cb.k.u-tokyo.ac.jp/

BLAST	Altschul et al.[3]	RRID:SCR_004870 http://blast.ncbi.nlm.nih.gov/Blast.cgi
Burrows-Wheeler Aligner	Li and Durbin[4]	RRID:SCR_010910 http://bio-bwa.sourceforge.net/
Bowtie2 v2.3.5.1	Langmead and Salzberg[5]	RRID:SCR_016368 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SPAdes assembler version 3.8.1	Bankevich et al.[6]	RRID:SCR_000131 https://cab.spbu.ru/software/spades/
filtlong		RRID:SCR_024020 https://github.com/rrwick/Filtlong
Smartdenovo	Liu et al.[7]	RRID:SCR_017622 https://github.com/ruanjue/smartdenovo
Redbean	Ruan and Li[8]	
Flye	Kolmogorov et al.[9]	RRID:SCR_017016 https://github.com/fenderglass/Flye
Necat	Chen et al.[10]	https://github.com/xiaochuanle/necat
Racon	Vaser et al.[11]	RRID:SCR_017642 https://github.com/isovic/racon
Hapo-G	Aury et al.[12]	https://www.genoscope.cns.fr/hapog/
Metabat 2	Kang et al.[13]	RRID:SCR_019134 https://bitbucket.org/berkeleylab/metabat/src/master/
SortMeRNA	Kopylova et al.[14]	RRID:SCR_014402 http://bioinfo.lifl.fr/RNA/sortmerna/
Velvet version 1.2.07	Zerbino and Birney[15]	RRID:SCR_010755 http://www.molecularevolution.org/software/genomics/velvet
Oases version 0.2.08	Schulz et al.[16]	RRID:SCR_011896 http://www.ebi.ac.uk/~zerbino/oases/
TransDecoder	Haas, B.J.	RRID:SCR_017647 https://github.com/TransDecoder/TransDecoder
CDDsearch	Marchler-Bauer et al.[17]	
Trimmomatic v0.38 and v0.39	Bolger et al.[18]	RRID:SCR_011848 http://www.usadellab.org/cms/index.php?page=trimmomatic
Trinity version v2.6.5	Grabherr et al.[19]	RRID:SCR_013048 http://trinityrnaseq.sourceforge.net/

rnaSPAdes version v3.13.1	Bushmanova et al.[20]	RRID:SCR_016992 http://cab.spbu.ru/software/rnaspades/
RepeatMasker version v.4.1.0	Smit et al.[21]	RRID:SCR_012954 http://repeatmasker.org/
Tandem repeats finder	Benson et al.[22]	RRID:SCR_022193 https://github.com/Benson-Genomics-Lab/TRF
REPET	Flutre et al.[23]	
BLAT	Kent et al.[24]	RRID:SCR_011919 http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Genewise	Birney et al.[25]	RRID:SCR_015054 http://www.ebi.ac.uk/Tools/psa/genewise/
DIAMOND v0.9.30	Buchfink et al.[26]	RRID:SCR_009457 http://www.nitrc.org/projects/diamond/
Est2Genome	Mott et al.[27]	https://galaxy-iuc.github.io/emboss-5.0-docs/est2genome.html
Gmove	Dubarry et al.[28]	RRID:SCR_019132 http://www.genoscope.cns.fr/gmove
votingLNC		https://gitlab.com/a.debit/votinglnc
AliView v.1.26	Larsson[29]	RRID:SCR_002780 https://github.com/AliView
RAxML v.8.2.	Stamatakis[30]	RRID:SCR_006086 https://github.com/stamatak/standard-RAxML
OrthoFinder v2.5.2	Emms and Kelly[31]	RRID:SCR_017118 https://github.com/davidemms/OrthoFinder
Count v9.1106	Csűös[32]	https://www.iro.umontreal.ca/~csuros/gene_content/count.html
MUSCLE version 3.8.1551	Edgar[33]	RRID:SCR_011812 http://www.ebi.ac.uk/Tools/msa/muscle/
OD-Seq version 1.0	Jehl et al.[34]	https://bioconductor.org/packages/release/bioc/manuals/odseq/man/odseq.pdf
HMMER3 package versions 3.1b1 and 3.3.2	Mistry et al.[35]	RRID:SCR_005305 http://hmmer.janelia.org/
GenEra	Barrera-Redondo et al.[36]	
MCL	Enright et al.[37]	RRID:SCR_024109 https://micans.org/mcl/

Foldseek	Kempen et al.[38]	https://search.foldseek.com/search
CleanBlastp	Pathmanathan et al.[39]	
SEED		RRID:SCR_002129 http://www.theseed.org/wiki/Home_of_the_SEED
IPR2GO		http://www.ebi.ac.uk/interpro/search/sequence-search
eggNOG	Huerta-Cepas et al.[40]	RRID:SCR_002456 http://eggnog.embl.de
eggNOG-mapper	Cantalapiedra et al.[41]	RRID:SCR_021165 http://eggnog-mapper.embl.de
Spearman's rank correlation analysis tool version 1.1.23-r7	P. Wessa, Free Statistics Software, Office for Research Development and Education	https://www.wessa.net/
Prodigal V2.6.3	Hyatt et al.[42]	RRID:SCR_011936 https://github.com/hyattpd/Prodigal
ViralRecall version 2.0	Aylward et al.[43]	https://github.com/faylward/viralrecall
esl-translate version 0.48		https://github.com/EddyRivasLab/easel/blob/master/miniapps/esl-translate.man.in
bedtools v2.29.2	Quinlan and Hall[44]	RRID:SCR_006646 https://github.com/arq5x/bedtools2
mmseqs cluster version 13.45111	Hauser et al.[45]	RRID:SCR_008184 https://github.com/eturro/mmseq#mmseq-transcript-and-gene-level-expression-analysis-using-multi-mapping-rna-seq-reads
MAFFT v7	Katoh and Standley[46]	RRID:SCR_011811 http://mafft.cbrc.jp/alignment/server/
MEGA	Tamura et al.[47]	RRID:SCR_023017 https://www.megasoftware.net/
NGphylogeny platform		https://ngphylogeny.fr/.
TrimAl	Capella-Gutiérrez et al.[48]	RRID:SCR_017334 http://trimal.cgenomics.org/
TAPscan	Petroll et al.[49]	https://plantcode.cup.uni-freiburg.de/tapscan/
Expasy web translator		RRID:SCR_024703 https://web.expasy.org/translate/
Geneious versions 11.0.5 and 11.1.5		RRID:SCR_010519 http://www.geneious.com/

Interproscan 94.0	Jones et al.[50]	RRID:SCR_005829 http://www.ebi.ac.uk/Tools/pfa/iprscan/
Clustal 2.1	Thompson et al.[51]	RRID:SCR_001591 http://www.ebi.ac.uk/Tools/msa/clustalo/
Gblocks	Castresana[52]	RRID:SCR_015945 http://molevol.cmima.csic.es/castresana/Gblocks_server.html
Kallisto version 0.44.0.	Bray et al.[53]	RRID:SCR_016582 https://pachterlab.github.io/kallisto/about
Deseq2	Love et al.[54]	RRID:SCR_015687 https://bioconductor.org/packages/release/bioc/html/DESeq2.html
FastQC	Andrews[55]	RRID:SCR_014583 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore version 0.6.5		RRID:SCR_011847 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
HISAT2 version 2.1.0		RRID:SCR_015530 http://ccb.jhu.edu/software/hisat2/index.shtml
featureCounts	Liao et al.[56]	RRID:SCR_012919 http://bioinf.wehi.edu.au/featureCounts/
PAML v 4.9i	Yang[57]	RRID:SCR_014932 http://abacus.gene.ucl.ac.uk/software/paml.html
phytools R package	Revell[58]	RRID:SCR_015502 https://cran.r-project.org/web/packages/phytools/index.html
VHICA package	Wallau et al.[59]	https://github.com/cran/vhica
NOVOPlasty v3.7	Dierckxsens et al.[60]	RRID:SCR_017335 https://github.com/ndierckx/NOVOPlasty
SAMtools v1.5	Li et al.[61]	RRID:SCR_002105 http://htslib.org/
GeSeq v2.03	Tillich et al.[62]	RRID:SCR_017336 https://chlorobox.mpimp-golm.mpg.de/geseq.html
ARAGORN v1.2.38	Laslett and Canback[63]	RRID:SCR_015974 http://mbio-serv2.mbioekol.lu.se/ARAGORN/
ModelFinder	Kalyaanamoorthy et al.[64]	http://www.iqtree.org/ModelFinder/
UFBoot2	Hoang et al.[65]	

SynMap	Haug-Baltzell et al.[66]	https://genomevolution.org/SynMap.pl
DAGChainer	Haas et al.[67]	https://dagchainer.sourceforge.net/
CodeML	Yang et al.[57]	
nwalign		https://pypi.org/project/nwalign/
BEAST v2.7	Bouckaert et al.[68]	RRID:SCR_010228 http://beast.bio.ed.ac.uk/
StarBEAST3 v1.1.7	Douglas et al.[69]	https://github.com/rbouckaert/starbeast3
bModelTest	Bouckaert et al.[70]	
LogCombiner v2.4.7	Bouckaert et al.[68]	
TreeAnnotator v2.4.7	Bouckaert et al.[68]	
SplitsTree 4	Kloepper et al.[71]	RRID:SCR_014734 http://www.splitstree.org/
Hectar	Gschloessl et al.[72]	https://webtools.sb-roscoff.fr/root?tool_id=abims_hectar
RShiny		https://github.com/rstudio/shiny
IG-TREE 2	Minh et al.[73]	https://github.com/iqtree/iqtree2
Other		
BUSCO analysis v5, eukaryota_odb10	Manni et al.[74]	RRID:SCR_015008 http://busco.ezlab.org/
UniRef90		RRID:SCR_010646 http://www.uniprot.org/help/uniref
AlphaFold protein structure database	Varadi et al.[75]	RRID:SCR_023662 https://alphafold.ebi.ac.uk/
NCVOG database	Yutin et al.[76]	
VOGDB database		https://vogdb.org/
SulfAtlas database	Barbeyron et al.[77], Stam et al.[78] 21/02/2024 08:45:00	https://sulfatlas.sb-roscoff.fr/
Pfam	Mistry et al.[79]	RRID:SCR_004726 http://pfam.xfam.org/
Panther 17.0		RRID:SCR_004869 http://www.pantherdb.org/
Simple Modular Architecture Research Tool (SMART)	Letunic et al.[80]	RRID:SCR_005026 http://smart.embl.de/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, J. Mark Cock (cock@sb-roscoff.fr).

Materials availability

All the laboratory-cultivated strains grown to provide material for genome sequencing can be accessed via the Roscoff Culture Collection (https://www.roscoff-culture-collection.org).

Data and code availability

All sequence data, including DNA and RNA sequencing data, have been deposited in the EBI/ENA database under the following project PRJEB72149 and will be publicly available as of the date of publication.

This paper does not report original code.

Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Biological material

The algal strains analysed in this study are listed in Table S1A. All strains except those belonging to the Fucales were grown under laboratory conditions. The latter cannot be maintained long-term in the laboratory so field material was harvested for extractions. The haploid gametophyte generation was grown in culture for species with characterised haploid-diploid life cycles, with the exception of *Ectocarpus* strains, for which haploid partheno-sporophytes or diploid sporophytes were cultivated. All cultures were grown either in 140 mm diameter Petri dishes or in 2–10 L bottles, the latter aerated by bubbling with sterile air. Most cultures were grown in Provosoli-enriched[23] natural seawater (PES medium) under fluorescent white light (10–30 µM photons/m2·s) at 13°C (or at 10°C for *Hapterophycus canaliculatus* and *Chordaria linearis* or 20°C for *Sphacelaria rigidula, Dictyota dichotoma, Schizocladia ischiensis* and *Chrysoparadoxa australica*). Exceptions included the freshwater species *Pleurocladia lacustris*, *Porterinema fluviatile* and *Heribaudiella fluviatilis*, which were grown in natural seawater that had been diluted to 5% with distilled water (i.e., 95% distilled

water / 5% seawater) before addition of ES medium (http://sagdb.uni-goettingen.de/culture_media/01%20Basal%20Medium.pdf) micronutrients (at 20°C for *P. lacustris*) and *Phaeothamnion wetherbeei,* which was grown in MIEB12 (medium 7 in[81]). Whole thallus was extracted for all species except the Fucales, where either dissected meristematic regions or released male gametes were extracted.

DNA extraction

DNA was extracted using either the OmniPrep Genomic DNA Purification Kit (G Biosciences, St. Louis, MO, USA) or the Nucleospin Plant II midi DNA Extraction Kit (Macherey-Nagel, Düren, Germany). DNA quality was assessed using a Qubit fluorometer (Themo Fisher Scientific, Waltham, MA, USA), and fragment length was assessed by migration on a 1% agarose gel for some of the samples.

Illumina library preparation and sequencing

Libraries were prepared using the NEBNext DNA Modules Products (New England Biolabs, Ipswich, MA, USA) with an 'on bead' protocol developed by Genoscope, starting with 100 ng of genomic DNA. DNA was sonicated to a 100–800 bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA), end-repaired and 3□-adenylated. Illumina adapters (Bioo Scientific, Austin, TX, USA) were then added using the NEBNext Sample Reagent Set (New England Biolabs, Ipswich, MA, USA) and purified using Ampure XP (Beckmann Coulter Genomics, Danvers, MA, USA). Adapted fragments were amplified by 12 cycles of PCR using the Kapa Hifi Hotstart NGS library Amplification kit (Roche, Basel, Switzerland), followed by 0.8x AMPure XP (Beckman Coulter Genomics, Danvers, MA, USA) purification. Libraries were sequenced with Illumina MiSeq, HiSeq 4000 or NovaSeq 6000 instruments (Illumina, San Diego, CA, USA) in paired-end mode, 150 base read-length.

Oxford Nanopore library preparation and sequencing

Some samples were first purified using the Short Read Eliminator Kit (Pacific Biosciences, Menlo Park, CA, USA). All libraries were prepared using the protocol "1D Genomic DNA by Ligation" provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK). Most of the libraries were prepared with the SQK-LSK109 kit (Oxford Nanopore Technologies), a few with the SQK-LSK108 or SQK-LSK110 kits (Oxford Nanopore Technologies). Three flow cells were loaded with barcoded samples. The samples were mainly sequenced on R9.4.1 MinION or PromethION flow cells.

RNA extraction, Illumina RNA-seq library preparation and sequencing

RNA was extracted using either the Qiagen RNeasy kit or the Macherey Nagel RNAplus kit (Macherey-Nagel, Düren, Germany). RNA-seq libraries were prepared using the TruSeq Stranded mRNA Sample Prep (Illumina) according to the manufacturer's protocol, starting with 500 ng to 1 µg of total RNA, or using the NEBNext Ultra II Directional RNA Library Prep for Illumina (New England BioLabs) according to the manufacturer's protocol, starting with 100 ng of total RNA. The libraries were sequenced with Illumina HiSeq 2500, HiSeq 4000 or NovaSeq 6000 instruments (Illumina, San Diego, CA, USA), in paired-end mode, 150 base read-length.

Assembly strategies

Two assembly strategies were employed (Figure S13): one was designed for genomes exclusively sequenced using short reads with Illumina technology, while the other was designed for genomes that underwent sequencing using a combination of long and short reads, using respectively the Nanopore and Illumina technologies.

Short-read-based genome assembly. When sequencing was performed exclusively using short reads, reads corresponding to bacterial contaminants were filtered out early in the assembly process (Figure S13A) because, typically, the initial datasets were too large to run assemblers like SPAdes. To remove bacterial contaminants, an assembly based on the initial illumina dataset was first generated for each strain using a fast and non-greedy algorithm, MEGAHIT[1] version 1.1.1 with the parameters --k-min 101 --k-max 131 --k-step 10. Assigning taxonomy is easier when working with contigs than with reads. Contigs exceeding 500 bp in each preliminary assembly underwent taxonomic classification based on gene models predicted using the *ab initio* software MetaGene[2] version 2008.8.19 with default parameters and then aligning proteins against UniprotKB using BLASTp (e-value $<10e^{-4}$). A superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 50% of their genes assigned to Bacteria and with at least one gene every 10 Kbp were classified as bacterial sequences. For each strain, the initial Illumina sequencing reads were aligned against the corresponding bacterial sequences using latest version of the Burrows-Wheeler Aligner[4] (BWA) with default parameters and mapped short-reads were labelled as contaminants, and assembled for the purpose of obtaining more contiguous contigs. These bacterial contigs were then used to build a contaminant sequence database. Finally, the clean subset of reads was obtained by aligning the whole Illumina dataset against this strain-specific bacterial contig database, using Bowtie2[5] version 2.2.9 with default

parameters. A final assembly was then generated for each strain using the contaminant-free read datasets and the SPAdes[6] assembler version 3.8.1 with the parameters -k 21,57,71,99,127 -m 2000 --only-assembler --careful. Genome assemblies based only on short-reads were more fragmented (N50 ranged from 3 Kbp to 31 Kbp) than assemblies that used long reads but the sizes of the former were consistent with expectations.

Long-read-based genome assemblies. A subset of the strains produced DNA of both adequate quality and quantity, enabling successful long-read sequencing. In these cases, long reads were assembled directly and the detection of possible bacterial contigs was carried out after the assembly step (Figure S13B). To produce long-read-based genome assemblies we generated three samples of reads i) all reads, ii) 30X coverage of the longest reads and iii) 30X coverage of the filtlong (https://github.com/rrwick/Filtlong) highest-score reads. The three samples were used as input data for four different assemblers, Smartdenovo[7], Redbean[8], Flye[9] and Necat[10]. Based on the cumulative size and contiguity, we selected the best assembly for each strain. This assembly was then polished three times using Racon[11] with Nanopore reads, and twice with Hapo-G[12] and Illumina PCR-free reads.

Assembly decontamination

Contigs from the short- and long-read genome assemblies were then inspected for potential bacterial sequences. This process was carried out using a combination of several analysis and tools: GC composition, read coverage, Metabat 2 (for tetramer composition and clustering)[13] and Metagene (for gene prediction and taxonomic identification, as described previously). Contigs were manually removed based on their characteristics.

Transcriptome assembly

Ribosomal-RNA-like reads were detected using SortMeRNA[14] and filtered out. The Illumina RNA-Seq short reads from each strain were assembled using Velvet[15] version 1.2.07 and Oases[16] version 0.2.08 with kmer sizes of 61, 63 and 65 bp. BUSCO[74] analysis (v5, eukaryota_odb10) was then performed on the three resulting assemblies for each strain in order to select the best assembly, i.e. the most complete at the gene level. Reads were mapped back to the contigs with BWA-mem, and only consistent paired-end reads were retained. Uncovered regions were detected and used to identify chimeric contigs. In addition, open reading frames (ORF) and domains were identified using TransDecoder (Haas, B.J., https://github.com/TransDecoder/TransDecoder) and CDDsearch[17],

respectively. Contigs were broken into uncovered regions outside ORFs and domains. In addition, read strand information was used to correctly orient RNA-Seq contigs.

De novo transcriptomes

The RNA-seq data was also used to generate *de novo* transcriptomes. For each strain, all the RNA-seq data available was cleaned for sequencing quality and presence of adapter sequences using Trimmomatic[18] v0.39 prior to being assembled using either Trinity[19] version v2.6.5 or rnaSPAdes[20] version v3.13.1. The strandness and Kmer-length parameters of the assemblers were adjusted to take into account RNA-seq read characteristics. The *de novo* transcriptomes represented an alternative source to identify and characterise genes if they were not detected in the genome assemblies.

Detection and masking of repeated sequences and transposons

Prior to gene annotation, each genome assembly was masked based on the repeat library from *Ectocarpus* species 7 (formerly *Ectocarpus siliculosus*)[82] and using RepBase with RepeatMasker[21] version v.4.1.0, default parameters. Tandem repeats finder (TRF)[22] was also used to mask tandem repeat duplications. In addition, transposons were annotated in ten species using REPET[23] and the transposons detected were used as a reference to mask all genomes with RepeatMasker[21] version v4.1.0, default parameters.

Gene prediction

For each strain, gene prediction was performed using both homologous proteins and RNA-seq data. Proteins from *Ectocarpus* species 7 (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2)[83] and UniRef90 (https://www.uniprot.org/uniref/) were aligned against each genome assembly. First, BLAT[24] with default parameters was used to quickly localise putative genes corresponding to the *Ectocarpus* species 7 proteins. The best match and matches with a score ≥ 90% of the best match score were retained. Second, the alignments were refined using Genewise[25] with default parameters, which is more precise for intron/exon boundary detection. Alignments were retained if more than 80% of the length of the protein was aligned to the genome. To detect conserved proteins and allow detection of horizontal gene transfer, UniRef90 proteins (without *E. siliculosus* sequences) were aligned with DIAMOND[26] v0.9.30 with parameters --evalue 0.001 --more-sensitive to genomic regions lacking alignments with an *Ectocarpus* species 7 protein. Only the five best matches per locus were

retained, based on their bitscore. Selected proteins from UniRef90 were aligned to the whole genome using Genewise as described previously, and alignments with at least 50% of the aligned protein length were retained. The assembled transcriptome for each strain was aligned to the strain's genome assembly using BLAT[24] with default parameters. For each transcript, the best match was selected based on the alignment score, with an identity greater or equal to 90%. Selected alignments were refined using Est2Genome[27] in order to precisely detect intron boundaries. Alignments were retained if more than 80% of the length of the transcript was aligned to the genome with a minimal identity of 95%. Finally, the protein homologies and transcript mapping were integrated using a combiner called Gmove[28]. This tool can find coding sequences (CDSs) based on genome-located evidence without any calibration step. Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for open reading frames (ORFs) consistent with the protein evidence. Translated proteins of predicted genes were then aligned against NR prot (release 19/02/2019) and the *Ectocarpus* species 7 version v2 proteome[83] (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2) using DIAMOND BLASTp with parameters --evalue 10-5 --more-sensitive --unal 0. All predicted genes with significant matches (the smallest protein had to be aligned for at least 50% of its length) were retained. In addition to these genes, we also retained genes with CDS size greater than 300 bp and with a coding ratio (CDS size / mRNA size) greater or equal to 0.5.

Annotation decontamination

After predicting the genes, an additional analysis was carried out to detect bacterial sequences. If a contig did not contain any genes, it was analysed with MetaGene and the predicted proteins added to the gene catalogue for the purpose of detecting bacterial sequences. Proteins generated from predicted genes (Gmove plus MetaGene) were then aligned against UniprotKB using BLASTp (e-value < $10e^{-4}$) and superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 80% of their genes assigned to bacteria, Archaea or viruses were classified as bacterial sequences and removed from the final assembly file. Genes belonging to these contigs were also removed from the final gene catalogue. Finally, completeness of each predicted gene catalogue was assessed using BUSCO[74] (v5.0.0; eukaryota_odb10).

In addition, the quality of the annotations was assessed by comparing the length of coding regions in pairs of orthologous proteins (best reciprocal hits) between each genome and *Ectocarpus* species 7,

which was used as a reference because its high-quality annotation has been extensively curated[83]. The correlation between orthologous CDS lengths was higher for genomes sequenced with long reads than for genomes only sequenced with short reads (Figure S1A). This difference was probably principally due to a higher proportion of underestimated protein lengths in the latter (Table S1B) which likely corresponded to fragmented genes. The qualities of Ectocarpales genome annotations were very high (BUSCO and length of predicted CDS) even when the genomes were sequenced using only short reads, probably because their phylogenetic proximity to *Ectocarpus* species 7 facilitated the building of good quality gene models.

Analyses aimed at deducing functional characteristics of predicted proteins

Several different analyses of the predicted proteomes of each species were carried out to provide information about the cellular functions of the encoded proteins. These included eggNOG-mapper[41] analyses (v2.1.8 or v2.0.1, with emapperDB v5.0.2 or v4.5.1) to provide multiple functional annotations (Gene Ontology, Kyoto Encyclopedia of genes and genomes, Clusters of Orthologous Genes, Pfam), Interproscan[50] analyses (versions v5.55-88.0, v5.51-85.0 or v5.36-75.0) to detect functional domains, Hectar[72] (v1.3) predictions of protein subcellular localisation and various DIAMOND[26] (v2.0.15 vs UniRef90 2022_03, with parameter "evalue" set to $10e^{-5}$) sequence similarity searches aimed at identifying homologous proteins with functional annotations.

Detection of tandemly duplicated genes

Starting with the protein alignments that had been constructed to build the orthogroups, matches between proteins within the same genome with an e-value of ≤10–20 and which covered at least 80% of the smallest protein were extracted. Two genes were considered to be tandemly duplicated if they were localised on the same genomic contig separated by five or less intervening genes, regardless of their orientation. The tandemly-duplicated genes were clustered using a single linkage clustering approach. A binomial test was applied to compare the proportion of tandemly-duplicated genes in each orthogroup with the global proportion of tandemly-duplicated genes (p=0.0532792). The *p*-values are shown in table S6.

Relative orientation of adjacent genes and lengths of intergenic regions

For each species, the proportion of pairs of adjacent genes localized on opposite strands was compared to the expected proportion of 0.5 using a binomial test (with p=0.5). The *p*-values are

shown in Table S1 (*p*-values of <0.05 correspond to cases where the proportion is significantly higher than 0.5).

The lengths of intergenic regions between pairs of adjacent genes located on opposite strands (i.e. divergently or convergently transcribed) were compared with the lengths of intergenic regions between genes located on the same strand (i.e. transcribed in the same direction). Contingency tables were constructed for each species using a threshold of 1000 bp for the intergenic length and the number of intergenic regions in each of four categories were counted: 1) same strand genes, intergenic <1000 bp, 2) opposite strand genes, intergenic <1000 bp, 3) same strand genes, intergenic ≥1000 bp, 4) opposite strand genes, intergenic ≥1000 bp. Fisher exact tests were applied to the contingency tables (alternative hypothesis: true odds ratio is greater than 1). The *p*-values are shown in Table S1. When *p*-values are <0.05, short intergenic lengths are significantly associated with pairs of genes on opposite strands. All calculations were performed with R[91] (version 4.3.0).

Detection of long non-coding RNAs

Transcriptome data for 11 species (Table S36), including nine brown algal strains and two outgroup taxa, was analysed to identify lncRNAs. Any transcripts with invalid nucleotide DNA symbols were discarded and sequences shorter than 200 nucleotides were removed to avoid the detection of small RNA transcripts. The transcriptome sequences in Fasta format were analysed with votingLNC (https://gitlab.com/a.debit/votinglnc) to detect lncRNA transcripts and assign a confidence level for each transcript. A similar approach was used to detect lncRNAs in the lncPlankton database[84]. VotingLNC is a meta-classifier combining the predictions of the ten most commonly used coding potential tools. Based on a majority voting ensemble procedure, the meta-tool assigns the final coding potential class to a transcript as the class label predicted most frequently by the ten classification models included in the ensemble. Alongside the majority voting class, a reliability score was calculated for each transcript. A cut-off non-coding reliability score of $p > 0.5$ was chosen to treat a transcript as lncRNA and to decrease false-positive identification. The set of transcripts predicted as lncRNA by the majority-voting procedure and having an ORF(s) encoding peptide(s) with length ≥ 100aa were discarded. lncRNA transcripts that had significant matches in either the Pfam[79] (hmmscan e-value < 0.001) or SwissProt (BLASTp e-value < $1e^{-5}$ and similarity ≥ 90%) databases were removed from the dataset. Transcript length, GC content, and the length of the longest ORF were compared between lncRNAs and protein-coding RNAs. The comparison was carried out using a Wilcoxon test. R version V.4.1.2 was used for all the analyses and ggplot2 (V.3.4.0) for plotting.

Phylogenomic tree of the Phaeophyceae

To provide a phylogenetic framework for the analyses of the Phaeoexplorer genome dataset, the 41-species phylogenomic tree reported by Akita *et al.*[85] was updated by adding 15 additional species using the same methodology. Briefly, for the additional species, amino acid sequences were recovered for the 32 single-copy orthologous genes used to construct the published tree and these were aligned manually with the existing sequences using the alignment software AliView[29] v.1.26. The aligned sequences of the final 56 species were concatenated and maximum likelihood analysis was carried out with 10,000 rapid bootstraps using RAxML[30] v.8.2.9 and the gamma model. The best-fit evolutionary model for each gene was determined using AIC.

Detection of orthologous groups

Predicted proteins from the 60 strains sequenced in Phaeoexplorer complemented with 16 public proteomes covering the Ochrophytina subphylum and the terrestrial oomycetes were clustered using OrthoFinder[31] v2.5.2 with default parameters. This generated 56,340 orthogroups that contained 90.1% of the proteins (1,415,341 of the 1,571,648). Seventy-one of the 76 strains had more than 75% of their proteins in an orthogroup shared with at least one other strain. Those orthogroups contained between 2 and 6,220 proteins with a mean of 25.1 proteins and a median of three.

Dollo analysis of orthogroup gain and loss

An analysis of evolutionary events of gene family gain and loss was carried out on a selection of strains covering the brown algal phylogeny and sister groups as distant as the Raphidophyceae under the Dollo parsimony law using orthogroups as proxies for gene families. To limit possible problems due to the fragmentation of predicted proteins in some assemblies, we selected 24,410 orthogroups present in at least one of 17 strains that had both good quality genome assembly and good quality gene predictions. Dollo parsimony analysis was then run using Count[32] version v9.1106 based on a cladogram of a subset of 24 species representative of the Phaeoexplorer project and excluding all public outgroups more distant than *Heterosigma akashiwo*. The cladogram was based on the topology of the brown algae phylogenetic tree published by Akita *et al.*[85].

Intron conservation

Intron positions were compared in a set of single copy genes that are conserved across all the Phaeophyceae and the outgroup species. The analysis focused on the 21 reference genomes (Table S36) and on orthogroups that occurred exactly once in at least 20 of the 21 genomes, allowing the gene to be absent from only one of the 21 genomes. In addition, orthogroups were discarded if more than three copies had been annotated in the other Phaeophyceae genomes. These filters produced a set of 235 conserved (ancestral) orthogroups. Multiple alignments were carried out for each orthogroup using MUSCLE[33] version 3.8.1551 with default parameters and conserved blocks were identified with Gblocks[52] version 0.91b with the parameters -p=t -s=n -b5=a -b2=[nsp] -b1=[nsp] -b3=6, where "nsp" is equal to 90% of the number of proteins aligned. A shell script was then used to compare intron positions in the alignments. For each intron in the multiple sequence alignment, we obtained a corresponding conservation profile listing which species contains an intron at that position. The profiles obtained for the 949 introns that are in conserved blocks of the multiple alignments are shown in Figure 2C. Both phase and length of ancestral introns (e.g. that were conserved in most Phaeophyceae and at least two sister clades) were compared to the phase and length of *Ectocarpus* species 7 introns as a reference. The same approach was used to compare intron positions across 11 *Ectocarpus* species, with *Scytosiphon promiscuus* as an outgroup, by selecting 831 conserved monocopy orthogroups.

Detection of gene family amplifications

A binomial test with a parameter of 17/21 was carried out to detect gene families (OGs) that had significantly expanded in 17 Phaeophyceae reference genomes compared with four outgroup species (Table S36). Expanded gene families deviated significantly from the expected proportion (17/21 under the null hypothesis where there are equal gene numbers in all species). Benjamini–Hochberg FDR correction for multiple testing was then applied and 233 candidate OGs with corrected *p*-values of < 0.001 were retained. All calculations were performed with R (version 4.1.0).

The set of 233 candidate OGs was then filtered to limit counting errors due to annotation artefacts (e.g. genes missed or fragmented) using the following procedure:

1) A protein consensus was first deduced for each orthogroup. Protein sequences representative of all lineages were extracted and aligned using MUSCLE[33] version 3.8.1551 with default parameters and the multiple alignments were filtered using OD-Seq[34] version 1.0 to remove outlier sequences, with parameter –score set to 1.5. The consensus sequences were then extracted from the multiple alignments of non-outlier sequences using hmmemit in the HMMER3[35] package version 3.1b1 with default parameters.

2) In order to estimate gene family copy number independently of the assembly and annotation processes, short read sequences for each genome were mapped onto the orthogroup consensus sequences using DIAMOND[26]. Unique matches were retained for each read and depth of coverage was calculated for each consensus orthogroup. The depth obtained for each orthogroup was normalised for each species by dividing by the depth obtained on a set of conserved single-copy genes, so that the final value obtained was representative of the gene copy number. Then, for each candidate amplified orthogroup, the average depth for the 17 Phaeophyceae species and the average depth for the four outgroup species was calculated and OGs where the depth for outgroups was more than half the depth for the Phaeophyceae were discarded. We retained 227 out of 233 orthogroups after this step.
3) Finally, functional annotations were used to remove orthogroups that were likely to correspond to transposable elements. A final list of 180 OGs was retained (Table S6).

The amplified gene families were manually categorised into functional classes based on the output of automatic functional annotation programs (Interproscan[50], EggNOG[40], nr BLASTp) and an amplification profile was assigned to each orthogroup by identifying the taxonomic group where the amplification of the family was most marked (Table S6).

Phylostratigraphy analysis

GenEra[36] was used to estimate gene family founder events for each genome assembly by running DIAMOND[26] in ultra-sensitive mode against the Phaeoexplorer protein dataset and the NCBI non-redundant database. All sequence matches with e-values < 10^{-5} were treated as being homologous with the query genes in the target genomes. The NCBI taxonomy was used as an initial template to infer the evolutionary relationships of each query gene with their matches in the sequence database but taxonomic assignments within the PX clade and Phaeophyceae were then modified to reflect the evolutionary relationships that were inferred in the maximum likelihood tree. Gene families were predicted based on a clustering analysis of the query proteins against themselves using an e-value cutoff of 10^{-5} in DIAMOND and an inflation parameter of 1.5 with MCL[37]. Estimated evolutionary distances were extracted for each pair of species from the maximum likelihood species tree (substitutions/site) to calculate homology detection failure probabilities[86]. Taxonomic sampling of the species tree enabled homology detection failure tests to be carried out within the PX clade. Gene families whose ages could not be explained by homology detection failure were analysed by inspecting the functional and domain annotations for *Ectocarpus* species 7[83]. Structural alignments were performed using Foldseek[38] against the AlphaFold protein structure database[75].

Composite genes

The amino-acid sequences of all 530,598 genes present in the selected genomes were compared in an all-against-all pairwise alignment using DIAMOND BLASTp[26] version 2.0.11; "very-sensitive" mode; e-value threshold $1e^{-5}$. This raw alignment was then filtered using CleanBlastp, from the CompositeSearch suite[39], to remove sequence alignments with under 30% residue identity and produce the final sequence similarity network. CompositeSearch was then used on this network to identify putative composite gene families among the orthologous groups (OGs) previously computed by OrthoFinder[31]. Composite OGs containing two or more genes and having non-overlapping regions aligned to their component OGs were retained for further analysis, while singleton composite OGs and composites with overlapping component regions were discarded. A phylogeny-based approach[87], which uses information from extant genomes to apply a Dollo parsimony model in Count[32], was used to reconstruct the evolutionary events (domain fusions and fissions) that led to structural rearrangements of composite genes, allowing them to be labelled as fusion or fission events (or as complex events when sequentiality could not be clearly deduced).

Horizontal gene transfer (HGT)

Dataset and experimental approach. Uneven data collection across taxa can impact HGT identification. The phylogeny-based HGT screening approach used here requires the establishment of a comprehensive and taxonomically diverse reference dataset. The analysis focused on the Phaeoexplorer genomes using a background database called REFAL and an automated bioinformatics tool called RoutineTree, which screens for HGTs using phylogenetics. The background database was built using a starting database, GNM1157, which includes a diverse set of 17,250,679 protein sequences from 1157 genomes spanning various prokaryotic and eukaryotic lineages (540 bacteria, 45 archaea, 431 Opisthokonta, 15 Rhodophyta, 83 Viridiplantae, and 43 genomes from CRASH lineages). Data from NCBI RefSeq (updated as of May 2020) and MMETSP were integrated into GNM1157 to form the background database REFAL. To enhance data quality and reduce redundancy, CD-HIT version 4.5.4 was used to remove highly similar sequences (with sequence identity ≥90%) within each taxonomic order. This curation process resulted in a protein database consisting of 39.9 million sequences, representing over 7,786 taxa and providing comprehensive coverage across the diverse branches of the tree of life. To obtain the best assembled genome within a genus, the latest version was selected if multiple versions were available. In addition, the dataset was expanded by searching for genomes in other repositories such as the Joint Genome Institute. Special attention was

paid to achieving balanced representation of the Rhodophyta and Viridiplantae, which are particularly crucial for HGT analysis within the Chromalveolate group. To accomplish this, protein data from six red algal transcriptomes sourced from MMETSP was added. The HGT search was applied to 72 Stramenopile genomes, including 45 newly sequenced and 27 public genomes.

Phylogenetic Tree Reconstruction. The pipeline for constructing phylogenetic trees splits fasta files into individual sequence files and then carries out a search for homologous sequences, followed by multiple sequence alignment and tree-building. Nested positions within the trees were identified using artificial intelligence and hU and hBL methods were used for HGT verification. Instead of using all available sequences, sequences with the best BLAST hit scores from each kingdom, phylum, and class were used for tree construction to expedite tree-building and enhance clarity. Each gene, regardless of whether it was a copy or not, was used as a query for tree construction. To improve precision, four different methods were used for tree building: neighbour-joining, maximum parsimony, maximum likelihood and Bayesian. As a result, each node within a tree was associated with four support values. To create single-gene phylogenetic trees, a BLASTp[3] search was carried out against the background database, employing an e-value cutoff of $1e^{-05}$. For each query, the top 1,000 significant matches were sorted by bit-score in descending order as the default criterion. Matching sequences were then retrieved from the database, with a constraint of no more than three sequences per genus and no more than 12 sequences per phylum. To further refine the selection, significant matches with a query-subject alignment length of at least 120 amino acids were re-sorted based on query-subject identity in descending order. A second set of homologous sequences was then retrieved from the database following the same procedure. These two sets of homologous sequences, along with the query, were merged and aligned using MUSCLE[33] version 3.8.31 with default settings. The resulting alignments, trimmed to a minimum length of 50 amino acids using TrimAl[48] version 1.2 in automated mode (-automated1), were used to construct phylogenetic trees with FastTree version 2.1.7, with the 'WAG+CAT' model and four rounds of minimum-evolution SPR moves (-spr 4) along with exhaustive ML nearest-neighbour interchanges (-mlacc 2 -slownni). Branch supports were estimated using the Shimodaira-Hasegawa (SH)-test.

Inferring HGT based on tree topology. Phylogenetic trees were examined to identify specific topologies where Phaeoexplorer query sequences were nested among other sequences, defined as a situation where two or more monophyletic clades consist of both queries and prokaryotic sequences, supported by distinct nodes within the tree. These monophyletic clades are considered to group together if they share the same set of prokaryotic sequences but differ in sequences from optional taxa. Singletons for both the donor and receptor genes were excluded to minimise contamination

and recent HGT interference. To retain only robustly supported nested positions, positions were required to be multiply supported, with a minimum of ≥0.70 for the SH-test and aBayes-test support from at least two Phaeoexplorer receptor nodes and three donor supporting nodes. Furthermore, queries that displayed significantly different amino acid compositions ($P < 0.05$) compared to the remaining sequences in the alignment were discarded. Queries from the CRASH category that nested among sequences from other kingdoms (supported by >70% UFBoot at one or more supporting nodes) were retained.

Enhancing accuracy and establishing the timing of HGTs. To enhance accuracy, a minimum requirement was imposed for all supporting nodes and for strongly supported nodes that indicate query-donor monophyly. To determine the timing of HGT events, temporal information, primarily derived from the timetree database, was incorporated into each node. We assigned the "smallest boundary" role to pinpoint the most recent common ancestor at the time of the HGT event. Essentially, if all descendants of a given query protein sequence can be traced back to the initial HGT event, a common ancestral node can be identified whose occurrence time can be inferred using a molecular clock approach based on archaeological and fossil evidence. The taxonomy boundaries of HGT descendants were determined by identifying the smallest ancestor shared by both the donor and receptor taxa from the monophyletic clades within the tree. By considering the emergence times of both taxa, the timing of the transfer of genes from earlier taxa to later taxa can be determined, as the reverse scenario is not considered plausible.

Verification of HGTs. Verification of HGT used the following contamination assessment criteria: i) HGT candidates were excluded if they were located in a contig where 50% of the genes had better matches with other kingdoms, ii) HGT candidates were excluded if they were located in a contig where 50% of the genes were primarily identified as HGT genes, iii) HGT candidates were excluded if one of their five closest flanking genes, both upstream and downstream, had a better match with other kingdoms. AI, hU and the hBL value were used to further validate HGT events. This process was supplemented with annotation and functional predictions for the identified HGTs.

Further validation was based on the following concepts:

OUTGROUP: This comprises all biological donors present in a tree, excluding the query species if it belongs to biological donors.

SKIP: This includes all biological receptors (species belonging to optional taxa) in a tree, again excluding the query species if it belongs to biological receptors.

INGROUP: This encompasses species from SKIP's upper level, excluding SKIP itself and the query species (if it belongs to biological receptors).

AI (Alien Index): computed for each query gene using e-values from BLAST hits:

AI = (E-value of best BLAST hit in the INGROUP lineage) / (E-value of best BLAST hit in the OUTGROUP lineage)

The AI score quantifies how similar queries are to their homologs in the OUTGROUP compared to homologs in the INGROUP. We apply a relatively lenient cut-off (AI > 0) for initial screening, which can be adjusted in the second screening as needed.

hU (HGT Score Support Index): calculated for each query gene based on the best bit scores of INGROUP vs. OUTGROUP:

hU = (Best-hit bitscore of OUTGROUP) - (Best-hit bitscore of INGROUP)

A lenient cut-off (hU > 0) is used for initial screening, with flexibility for adjustment in the second screening.

hBL (HGT Branch Length Support Index): calculated based on the minimum branch length to the query within INGROUP vs. OUTGROUP:

hBL = (Minimum branch length to the query within INGROUP) - (Minimum branch length to the query within OUTGROUP)

A lenient cut-off (hBL > 0) is applied initially, with the option for modification in the second screening.

CHE, CHS, CHBL (Consensus Hit Support): To mitigate the possibility that the best bit score for either INGROUP or OUTGROUP is influenced by contamination, we consider alternative matches. We introduce consensus hit support (CHE, CHS, and CHBL) to assess the reliability of AI, hU, and hBL, respectively.

For example, if AI > 0, CHE evaluates the likelihood that "AI remains greater than 0" when using the e-value of each sequence in OUTGROUP instead of bbhO. A similar approach applies to CHS for hU and CHBL for hBL. This additional layer of evaluation helps ensure the robustness of the HGT verification process.

Gene codon usage, functional annotation and expression. Indices of codon usage and GC content were calculated using Codonw 1.4.4 (http://codonw.sourceforge.net). Gene functions were assigned by searching against the Gene Ontology (GO) database using blast2GO (ref blast2GO 08) and the KEGG database using blastKOALA (http://www.kegg.jp/blastkoala/) with default parameters. The full gene sets of each species were set as the background for KEGG and GO enrichment analyses by applying Student's t-test (*p*-value cutoff = 0.01). HGTs were also analysed with SEED (http://www.theseed.org/wiki/Home_of_the_SEED), IPR2GO (http://www.ebi.ac.uk/interpro/search/sequence-search), eggNOG[40] (http://eggnogdb.embl.de/#/app/home) and Pfam[79]. For each species, the differences between

mean gene expression levels for HGTs and non-HGT genes with common GO terms were accessed using Student's t-test. Go terms with less than five genes in either gene category were ignored. The differences in expression dispersal (coefficient of variation: standard deviation across genes or samples / mean value) and expression specificity (frequencies of a gene to be detected as unexpressed, defined as TPM = 2, in any condition) were accessed in a similar manner. Given the variable experimental conditions associated with different transcriptome data for each species, gene expression values for a gene were used indiscriminately regardless of the conditions. Correlation tests between the codon adaptation index (CAI) and gene expression were carried out using the Spearman's rank correlation analysis tool (P. Wessa, Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7, https://www.wessa.net/).

Detection of viral genome insertions and viral regions in algal genomes

To reduce the dataset size for analysis, 64 Phaeoexplorer and eight public genomes were initially filtered to retain only contigs that were more than 10 kbp in length. Gene prediction was then carried out on all contigs using Prodigal[42] (V2.6.3, settings: default, meta) and the resulting proteins were used as queries against the NCVOG[76] and VOGDB[88] databases using hmmscan (HMMER 3.3.2 with default settings). The contigs detected by hmmscan were then filtered to retain only sequences with at least one match to either viral database at a defined e-value cutoff ($1e^{-20}$ for NCVOG, and $1e^{-80}$ for VOGDB). The resulting positive 4,951 contigs were then analysed using ViralRecall[43] version 2.0 with settings -w 50 -g 1 -b -f -m 2 using the built-in Nucleocytoviricota (NCV) database GVOG and a window size of 50 kbp. To ensure that viral genes were not missed because they had not been annotated by Prodigal, six-frame translations of the contigs were generated using esl-translate (version 0.48 with default settings), and the resulting proteins queried against the same databases used by ViralRecall using hmmsearch (HMMER 3.3.2, settings: -E 1e-10). The ViralRecall results were then parsed using an in-house workflow. Six-frame translations were removed from the results if they overlapped (even partially) with any Prodigal gene prediction, as identified using bedtools[44] (v2.29.2; intersect). Likewise, overlapping six-frame translations and gene predictions with the same NCVOG match were removed to reduce redundancy. Based on the distance between query sequences with the same GVOG hit, queries were flagged as frame-shifted (less than 100 bp gap), intron-containing (100-5,000 bp gap) or mono-exonic (greater than 5,000 bp gap). All queries were also checked for overlaps with multi-exonic genes that had been annotated by the Phaeoexplorer gene prediction procedure (using Gmove[28]), and flagged if they did. All queries were then filtered to retain only those that matched a set of key NCV marker genes, identified by NCVOG code (A32, D5 helicase, D5 DNA primase, MCP, DNA polymerase B, SFII and VLTF3) or some Phaeovirus integrase

genes (integrase recombinase, integrase resolvase and RNR). The marker gene proteins were clustered with the protein sequences of NCVOGs using mmseqs cluster[45] (version 13.45111 with settings --min-seq-id 0.3 -c 0.8). Finally, the parsed results of the NCV marker gene set identified by the ViralRecall screen were manually curated, retaining only those queries with varying combinations of the following properties: placement within a viral region as identified by ViralRecall, similar hmmsearch results (score and e-value) and gene length to that of known NCV genes, not part of a multi-exonic gene, lack of Pfam HMM matches to cellular domains sharing homology to the marker gene (specific to certain marker genes), and clustered with an NCVOG in the mmseqs analysis. The marker gene content of the viral regions was manually assessed to estimate the number of complete or partial inserted viruses in each genome. VRs were considered to be complete proviruses if they contained all seven of the key NCV marker genes listed above. VRs were classed as partial proviruses if they only contained a subset of the seven key NCV marker genes, the presence of the MCP and DNA polymerase B genes being particularly strong indicators of a partial provirus.

Metabolic networks

Genome-scale metabolic networks were reconstructed using AuCoMe[89] version 0.5.1 using the MetaCyc[90] version 26 database. A first dataset, consisting of the 60 species listed in Table S36 (column "Metabolic networks") plus two public diatom genomes already used in the initial AuCoMe study (*Fragilariopsis cylindrus* and *Fistulifera solaris*) was processed to build the largest possible database (phaeogem) for exploratory comparisons (https://gem-aureme.genouest.org/phaeogem/). Then, a second comparison was performed on all long-read species plus outgroups. Based on Multidimensional-scaling (MDS) analyses, the most divergent long-read species (*Choristocarpus tenellus*, *Laminaria digitata, Phaeothamnion wetherbeei* and the public genome of *Sargassum fusiforme*) were excluded to construct a 16 species dataset, balancing assembly quality and phylogenetic coverage (https://gem-aureme.genouest.org/16bestgem/). The MDS plots presented on Fig S5 were build using the vegan package, version 2.6-4 (https://github.com/vegandevs/vegan) with R 4.1.2[91], using Jaccard distances. A third stricter dataset (fwgem), enriched in high-quality long-read Ectocarpales, was built to address questions related to freshwater adaptation (https://gem-aureme.genouest.org/fwgem/). A set of reactions that were overrepresented in brown algae compared to the outgroup was created (Table S7) by taking reactions present in 100% of brown algae and less than 70% of outgroups. Reactions corresponding to genes lost in freshwater species were also extracted (Table S32). These reaction sets were extracted from all the networks using the Aucomana library (https://github.com/AuReMe/aucomana). Online wikis (phaeogem, 16bestgem and fwgem) were generated using AuReMe[92].

CAZymes

CAZyme genes were identified based on shared homology with biochemically characterised proteins, either individually or as hidden Markov model (HMM) profiles. For phylogenetic analyses, proteins were aligned using MAFFT[46] with the iterative refinement method and the scoring matrix Blosum62. The alignments were manually refined and trees were constructed using the maximum likelihood approach. Alignment reliability was tested by a bootstrap analysis using 100 resamplings of the dataset. Only bootstrap values above 60% are shown. The phylogenetic trees were displayed with MEGA[47]. The annotated genes are listed in Table S9 with accession numbers.

Sulfatases

The sulfatases encoded by each brown algal genome were identified and assigned to their respective family and subfamily using the SulfAtlas database[77,78] (https://sulfatlas.sb-roscoff.fr/). Each predicted proteome was first submitted to the SulfAtlas HHM server (https://sulfatlas.sb-roscoff.fr/sulfatlashmm/), which allows rapid identification of sulfatase candidates and (sub)family assignment using hidden Markov model profiles for each SulfAtlas (sub)family. Each sulfatase candidate sequence was then used as a query in a BLASTp[3] search against the SulfAtlas database (https://blast.sb-roscoff.fr/sulfatlas/). Sequences with at least 50% identity with sulfatases from marine bacteria or other marine microorganisms were considered to be contaminants. Below this threshold, additional examination of the predicted gene structure and genomic context of the candidate sequence was undertaken to identify possible horizontal gene transfers.

Haloperoxidases

vHPO genes were identified based on sequence homology and active site conservation. Maximum likelihood phylogenetic analyses were carried out using the NGphylogeny platform at https://ngphylogeny.fr/. MAFFT was used to align vHPO sequences and alignments were automatically curated with TrimAl[48], leading to the selection of 444 informative positions from the initial 1450 positions for the algal-type vHPOs and 402 informative positions from the initial 1078 positions for the bacterial-type vHPOs. Maximum likelihood trees were constructed using FastTree with the WAG+G gene model and 1000 bootstrap replicates. Maximum likelihood Newick files were

formatted as circular representations using iTOL. Only bootstrap values between 0.7 and 1 were conserved. The lists of annotated vHPO genes are in Tables S11 and S12.

Ion channels

A search was carried out for 12 classes of ion channel in the predicted proteomes of the 21 Phaeoexplorer reference genomes plus those of two diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Table S24). Predicted proteomes were screened using BLASTp[3] and query sequences from *Ectocarpus* species 7 and seven other species from diverse eukaryotic taxa (Table S24).

Membrane-localised proteins

Membrane protein family genes were identified either by carrying out BLASTp[3] searches of the predicted Phaeoexplorer proteomes using *Ectocarpus* species 7 sequences as queries or by recovering orthogroups containing the relevant *Ectocarpus* species 7 sequences as members. The BLASTp approach was used for DEK1-like calpains, fasciclins, tetraspanins, CHASE, ethylene-binding-domain-like and MASE1 domain histidine kinases whereas the orthogroup approach was used to recover other members of the histidine kinase family. Both approaches were used to search for integrins and transmembrane receptor kinases. For integrins the two methods detected exactly the same set of proteins. For receptor kinases the BLASTp and orthogroup analyses detected 99.3% and 98.3% of the 269 genes, respectively. For these analyses, either the whole genome dataset was analysed or only the set of 21 reference genomes (Table S36), depending on the size of the gene family.

Transcription-associated proteins

TAPscan v4 (Petroll et al., unpublished) was used to analyse the transcription-associated protein (TAP) complements of 21 species. TAPscan[49] is a comprehensive tool for annotating TAPs based on the detection of highly conserved protein domains using HMM profiles with specific thresholds and coverage cut-offs. Following detection, specialised rules are applied to assign protein sequences to TAP families based on the detected domains. TAPscan v4 can assign proteins to 138 different TAP (sub)families with high accuracy.

EsV-1-7 domain proteins

EsV-1-7 domain proteins were identified in the 31 brown algal and sister taxa genomes (Table S36) by recovering the members of all orthogroups (with the exception of OG0000001, which is a very large OG that consisting principally of transposon sequences) that either contained one or more of a curated set of 101 EsV-1-7 domain proteins[93] for *Ectocarpus* species 7 or contained an EsV-1-7 domain protein based on a match to the Pfam EsV-1-7 motif PF19114. The recovered proteins were screened manually for the presence of at least one EsV-1-7 domain and a total of 2018 were finally identified as members of the EsV-1-7 family.

To identify orthologues of the EsV-1-7 protein IMMEDIATE UPRIGHT[93] (IMM), BLASTp searches of 25 brown algal and four sister taxa predicted proteomes were carried out with the amino-terminal domain of the IMM protein minus the five EsV-1-7 repeats as this domain is unique to IMM. Proteins were retained as IMM orthologues if they were more similar to IMM than to the most closely-related protein in *Ectocarpus* species 7, Ec-17_002150.

Histones

Histone protein sequences were analysed in *Ascophyllum nodosum* (An), *Chordaria linearis* (Cl), *Chrysoparadoxa australica* (Ca), *Desmarestia herbacea* (Dh), *Dictyota dichotoma* (Ddi), *Discosporangium mesarthrocarpus* (Dme), *Ectocarpus crouaniorum* (Ec), *Ectocarpus fasciculatus* (Ef), *Ectocarpus siliculosus* (Es), *Fucus serratus* (Fse), *Heterosigma akashiwo* (Ha), *Pleurocladia lacustris* (Pla), *Porterinema fluviatile* (Pf), *Pylaiella littoralis* (Pli), *Saccharina latissima* (Sl), *Sargassum fusiform* (Sf), *Schizocladia ischiensis* (Si), *Scytosiphon promiscuus* (Sp), *Sphacelaria rigidula* (Sri), *Tribonema minus* (Tm) and *Undaria pinnatifida* (Up) using BLASTp against the complete predicted proteomes (https://blast.sb-roscoff.fr/phaeoexplorer/) with the histone protein sequences from the diatom *Phaeodactylum tricornutum* as queries. The genes and transcripts coding for the identified histones were then retrieved from the genomes and predicted transcripts using BLAST (https://blast.sb-roscoff.fr/phaeoexplorer/). The proteins encoded by the identified genes and transcripts were predicted with the Expasy web translator (https://web.expasy.org/translate/). All the identified histone protein sequences and the corresponding transcripts IDs are reported in Table S26. In order to identify truncated proteins or incorrect start codons, the following constraints were applied: H2A proteins must start with the SGKGKGGR sequence, H2B with AKTP, canonical H3.1 and variants H3.3 with ARTKQT and H4 with SGRGKGGKGLGKGG. For the linker histone H1, protein sequences had to be lysine-rich and sequences with incorrect start codons were determined by alignments of all identified H1 proteins. For proteins with incorrect start codons, the region upstream of the correct

start codon was removed. For truncated proteins, *i.e.* proteins whose transcripts lacked either the start (no methionine) or stop codons, the protein sequence was completed based on alignment with the corresponding genomic region using the Geneious 11.0.5 software. When the sequence could not be completed, a BLAST was performed against the Phaeoexplorer *de novo* transcriptomes (https://blast.sb-roscoff.fr/phaeoexplorer_denovo/) when this data was available (this was not possible for the public genomes *T. minus*, *U. pinnatifida* and *S. fusiforme*). Based on the nomenclature established by[94], H3 histones were classified as follows: canonical H3.1 proteins harbour AT residues at positions 31-32 while histone variants H3.3 harbour TA residues, H3 proteins with other residues at positions 31-32 were named H3.4 and so on. CenH3 variants of H3 were identified by analysis with Panther 17.0 (www.pantherdb.org/tools/sequenceSearchForm.jsp?) and/or Interproscan[50] 94.0 (www.ebi.ac.uk/interpro/search/sequence/).

DNA methyltransferases

Searches were carried out for methyltransferases and demethylases in the predicted proteomes of 20 of the high quality brown algal reference genome assemblies (based on Nanopore long-read sequence) plus the sister taxa *Chrysoparadoxa australica* and *Schizocladia ischiensis* using BLASTp (Table S36). A methyltransferase reference database was constructed by recovering sequences from NCBI, ENSEMBL and UniProtKB. Methyltransferase sequences were recovered for stramenopiles such as *Nannochloropsis gaditana*, the diatom *Phaeodactylum tricornutum*, the oomycete *Phytophthora infestans* and for species from more distant lineages including *Arabidopsis thaliana*, *Homo sapiens* and the fungus *Neurospora crassa*. The proteomes of the selected brown algal strains were then queried against this database using BLASTp and matches with an e-value of < 0.001, a bit score > 70, a maximum gap of 5 and percentage identity of >30% were retained. The retained matches were screened against the NCBI, UniProt and SwissProt databases to identify and remove contaminating bacterial or viral proteins. Methyltransferase domains were detected in the retained matches using the Simple Modular Architecture Research Tool (SMART)[80] domain architecture analysis and InterPro searches (https://www.ebi.ac.uk/interpro/). Sequences with methyltransferase domains were retained for further analysis. Validated brown algal methyltransferases were aligned with reference methyltransferases using Clustal[51] 2.1.

Spliceosome

Components of the Major Spliceosome were identified using a reference set of 147 human components (https://www.genenames.org/data/genegroup/#!/group/1518), excluding the five small

nuclear RNAs (snRNAs). Including isoforms, this query set consisted of 626 proteins. These proteins were used to screen the predicted proteomes of 54 genomes (Table S36) using BLASTp and matches were retained if they had an e-value of at most $1e^{-40}$ and coverage >30%. Searches were also carried out for components of LSm and Sm complexes which have roles as scaffolds in the formation of ribonucleoprotein particles (RNPs), in the maturation of mRNAs (including splicing, such as the cytoplasmic complex LSm1-7, LSm2-8 which is part of the core U6 snRNP and other complexes important for the formation of the 3' ends of histone transcripts), in the assembly of P-Bodies and in the maintenance of telomeres.

Flagella proteins

A previous proteomic analysis of anterior and posterior flagella of the brown alga *Colpomenia bullosa* identified a total of 592 proteins across the two proteomes[95]. Here the *Ectocarpus* species 7 orthologues of 70 of these proteins that had been detected with a very high level of confidence were used to identify the corresponding orthogroups and the presence or absence of these orthogroups was scored for seven representative species (Table S36).

Detection of *Porterinema fluviatile* genes differentially expressed in freshwater and seawater

Six independent cultures of *Porterinema fluviatile* were cultivated for four weeks in 140 mm Petri dishes with Provasoli-enriched culture medium[96], which was renewed every two weeks. For three Petri dishes, the culture medium was based on autoclaved natural seawater (high salinity treatment), for the other three Petri Dishes natural seawater was diluted 1:19 vol/vol with distilled water (low salinity treatment). Cultures were harvested with 40 µm nylon sieves, dried with a paper towel, and immediately frozen in liquid nitrogen. RNA extraction library construction and sequencing were carried out as described in section "RNA extraction, Illumina RNA-seq library preparation and sequencing". RNA-seq reads were cleaned with Trimmomatic[18] V0.38 and then mapped to the *P. fluviatile* genome using Kallisto[53] version 0.44.0. Differentially expressed genes were identified using the Deseq2 package[54] included in Bioconductor version 3.11, considering genes with an adjusted $p < 0.05$ and a log2 fold-change > 1 as differentially expressed. The complete list of differentially expressed genes is provided in Table S30. To compare the differentially expressed genes in *P. fluviatile* with an equivalent set previously identified for *Ectocarpus subulatus* in a microarray experiment using nearly identical growth conditions[97], orthologues in the two species were detected using Orthofinder version 2.3.3. Of the 10,066 shared orthogroups, 6,606 had microarray expression

data for *E. subulatus*. This information was used to classify differentially expressed genes for the two species as either shared orthologues or as lineage-specific.

Identification of genes with generation-biased expression patterns

RNA-seq data (two to five replicates per condition) was recovered for gametophyte and sporophyte generations of ten species (Table S36). Data quality was assessed with FastQC[55] version 0.11.9 and sequences were then trimmed with Trim Galore version 0.6.5 with the parameters --length 50, -quality 24, --stringency 6, --max_n 3. The cleaned reads were mapped onto the corresponding genome for each species using HISAT2 version 2.1.0 with default options. Counting was carried out with featureCounts[56] from the subread package (version 2.0.1) on CDS features grouped by Parent. Transcript Per Kilobase Million (TPM) tables were generated for all conditions and differentially expressed genes were detected using DESeq2[54] version 1.30.1. Genes were classified into six categories based on the differential expression analysis and the TPM values: gametophyte-biased, mean TPM ≥1 in gametophyte and sporophyte, log2(fold change) ≥1, adjusted *p*-value <0.05; sporophyte-biased: mean TPM ≥1 in gametophyte and sporophyte, log2(fold change) ≤-1, adjusted *p*-value <0.05; gametophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, log2(fold change) ≥1, adjusted *p*-value <0.05; sporophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, log2(fold change) ≤-1, adjusted *p*-value <0.05; unbiased genes: mean gametophyte and sporophyte TPMs ≥1, log2(fold change) <1 or >-1 and/or adjusted *p*-value ≥0.05; unexpressed genes, mean gametophyte and sporophyte TPM <1.

Life cycle and thallus architecture

Genome dataset and traits. To study the impact of body architecture, the brown algae were divided into three categories: 22 filamentous species, eight simple parenchymatous species and 13 species with elaborate thalli (Table S36). For the life-cycle-based assessment, the groups were: 30 haploid-diploid species and six diploid species (Table S36). Body architecture information was available for 43 species, and life cycle information was available for 36 species; species without body plan or life cycle information were not used in subsequent analyses. Two approaches were used to estimate selection intensity across the phylogeny, (i) a model-based method, and (ii) by evaluating codon usage bias and nucleotide composition. Two evolutionary models were used, one based on architecture and the other based on life cycle. For model-based methods the phylogeny was categorised based on the above traits, and selection intensity parameters were estimated using PAML[57] version 4.9i. Rate

estimates were obtained for non-synonymous substitutions (dN), synonymous substitutions (dS) and omega (dN/dS) for the multiple sequence alignments of all genes within each orthogroup using the variable-ratio model of CODEML from PAML, which allows different omegas for different branch categories. The traits were assigned to the branches of the phylogeny using ancestral state estimation by stochastic mapping with the phytools R package[58,91].

Evolutionary models to study impacts of body architecture. To study variation in selection intensity as a function of body architecture, we devised a model with the following trait categories: filamentous/pseudoparenchymatous (simple cell division and organisation on a single plane), parenchymatous (cell division and organisation on multiple planes) and elaborate thallus (tissue differentiation). To ensure that at least 50% of the species in each category were used in the analysis, we selected orthogroups (OGs) that contained at least 11 members for filamentous, at least four members for parenchymatous and at least six members for elaborate thallus algae. Using this filter, 1068 OGs were obtained, on which the model based on body architecture was fitted. Selection intensity parameters [rate of non-synonymous substitution (dN), rate of synonymous substitution (dS) and omega (dN/dS)] were estimated for the three trait categories for each gene alignment. We used the Wilcoxon signed-rank test to evaluate the statistical significance of differences between the selection intensity parameters (dN, dS and dN/dS) for each category.

Evolutionary models to study the impacts of life cycle. The impact of life cycle on molecular evolution was assessed using a model with two categories consisting of diplontic and haplodiplontic species. For this model we used 1,058 OGs that contained at least three members for diploid species and at least 15 members for haploid-diploid species. Using alignments of the gene within the OGs, we estimated the selection intensity parameters for the different categories and applied the Wilcoxon signed-rank test to assess the statistical significance of differences in selection intensity between the diploid and haploid-diploid life cycles.

Selection of intensity parameters. Omega (dN/dS) provides an estimate of the ratio of substitutions at sites under selection compared to neutral sites, and is generally used to infer the strength of purifying selection. Omega needs to be interpreted with caution because not all synonymous sites are neutral[98] and also synonymous substitutions are often underestimated due to saturation of synonymous sites, which might in turn impact the omega ratios[99]. Omega values lower than one indicate substitutions are less frequent at sites under selection compared to neutral sites and are characteristic of highly conserved genes or genes evolving under strong purifying selection. As we used primarily low copy number genes in this study, the analysed genes were expected to evolve

under strong purifying selection, with omega values much lower than one. Using omega for near neutral studies is challenging because near neutral sites are determined by effective population size, that is to say, sites under mild selection constraint in larger populations can behave as neutral sites in smaller populations. It is therefore difficult to infer the amount of mutation from relative values of omega. In order to obtain better insight into selection intensity, mutation accumulation was not only investigated using rates of synonymous (dS) and non-synonymous (dN) substitutions but also by estimating codon bias and nucleotide composition. Codon usage bias was used, in addition to omega, to infer selection intensity across species as the former reflects selection efficacy at synonymous sites[100–102]. We inferred codon usage bias by estimating the effective number of codons (ENC) for each species using the enc method from the VHICA package[59,85]. The effective number of codons (ENC) quantifies the extent of deviation of codon usage of a gene from equal usage of synonymous codons. For the standard genetic code, ENC values range from 20 (where a single codon is used per amino acid implying strong codon usage bias) to 61 (implies that all synonymous codons are equally used for each amino acid[103]). Low ENC indicates constrained use of codons, which potentially highlights stronger codon bias due to stronger selection at synonymous sites. As nucleotide composition can also influence codon bias, we calculated the overall GC composition, GC at the third codon position (GC3) and the theoretical expected ENC (EENC) based on GC3 using local R scripts. The lower the observed ENC (OENC, estimated from the gene sequence) relative to EENC, the stronger the influence of selection due to translation on codon usage. This was studied by estimating the difference (DENC = EENC - OENC) between the expected ENC and the observed ENC[104]. Positive DENC indicates a role for selection constraints on codon usage in addition to the influence of nucleotide composition. DENC values of zero or less indicate that codon bias is entirely driven by nucleotide composition. DENC values were used to study the influence of translation selection and nucleotide composition on codon usage bias.

Assembly and analysis of organellar genomes

Plastid and mitochondrial genomes were assembled *de novo* using NOVOPlasty[60] v3.7 and *rbcL* and *cox1* nucleotide sequences as seeds. Assembled genomes were checked by aligning reads using Bowtie2[5] v2.3.5.1 and processed with SAMtools[61] v1.5. Annotation of protein-coding genes was performed with GeSeq[62] v2.03. Annotation of tRNAs, tmRNAs and rRNAs was performed with ARAGORN[63] v1.2.38.

Maximum-likelihood (ML) phylogenetic trees were constructed using 92 plastid genomes (11 non-brown outgroup sequences) and 89 mitochondrial genomes (seven non-brown outgroup sequences).

The conserved coding-region amino acid sequences of 139 plastid genes (31,159 amino acids) and 35 mitochondrial genes (7,461 amino acids) were used to construct these phylogenetic trees. The sequence for each gene was aligned individually using MAFFT[46] v7 (--maxiterate 1000) and then concatenated. Alignment partitions were assigned based on genes. Each of the aligned gene sequences was trimmed with trimAl[48] v1.2 (-automated1). ML phylogenetic trees were constructed with IQ-TREE 2[73]. The protein substitution models in each gene partition were selected using ModelFinder[64]. Statistical support for tree branches was assessed with 1,000 replicates of ultrafast bootstrap (UFBoot2)[65].

Analysis of *Ectocarpus* genome synteny

Global genome synteny analysis was performed using SynMap[66] on the CoGe platform (https://genomevolution.org/coge/) with the following genomes: *Ectocarpus crouaniorum* male, *Ectocarpus fasciculatus* male, *Ectocarpus siliculosus* male, *Ectocarpus* species 7 male and *Ectocarpus subulatus*. SynMap identifies syntenic regions between two or more genomes using a combination of sequence similarity and collinearity algorithms. Last[105] was used as the BLAST algorithm and syntenic gene pairs were identified using DAGChainer[67] with settings "Relative Gene Order", -D = 20, -A =5. Neighbouring syntenic blocks were merged into larger blocks. Substitution rates between the synthetic CDS pairs were calculated using CodeML[57], which was also implemented in SynMap, CoGe. In detail, protein sequences were aligned using the Needleman-Wunsch algorithm implemented in nwalign (https://pypi.org/project/nwalign/) and then translated back to aligned codons. CodeML was run five times for each alignment using the default parameters and the lowest dS was retained, with the upper cutoff for dS values set at 2. *Ectocarpus* genes were grouped according to their age based on the phylostratigraphic analysis and by chromosomal location based on their chromosome position in *Ectocarpus* species 7. All plots and statistical analysis were carried out in R version v.4.3.1. Local synteny analysis was based on orthologous genes as identified by Orthofinder.

Analysis of *Ectocarpus* gene evolution

Protein sequence alignments were used to remove gaps with trimAl[48] and then translated back to DNA with backtranseq[106]. Only DNA fasta files with a minimum of 70 bp were retained (831 single-copy orthologs). PhyML trees were built with Geneious v11.1.5 (https://www.geneious.com). Maximum likelihood analysis was carried out to detect site specific, branch-site specific and branch specific positive selection as well as sites under negative selection, using PAML[107].

Phylogenetic analysis of *Ectocarpus* species

Phylogenetic analysis was carried out for 12 *Ectocarpus* species (Table S36). Of the 933 single-copy orthogroups identified for these 12 species, 257 high-confidence alignments were retained for gene tree and species tree inferences following the removal of low-quality alignments using BMGE[108]. Bayesian inference of the phylogeny of the *Ectocarpus* species complex was performed using BEAST[68] v2.7. The analysis was conducted under the multi-species coalescent (MSC) model, implemented in StarBEAST3[69] v1.1.7. The MSC model coestimates gene trees and the species tree within a multispecies coalescent framework, enabling the assessment of incongruences among genes with respect to the species tree. To account for substitution model uncertainty, bModelTest[70] was employed to average over a set of substitution models for each alignment. StarBEAST3 was run under both the Yule model and the strict clock model. A total of 300,000,000 Markov Chain Monte Carlo (MCMC) generations were conducted, with tree states stored every 50,000 iterations. Posterior tree samples were combined, discarding the initial 10% burn-in, using LogCombiner v2.4.7. A maximum clade credibility tree was generated using TreeAnnotator[68] v2.4.7.

Ectocarpus introgression analysis

To distinguish introgression from shared ancestry, D estimates (i.e. ABBA-BABA tests) were generated from 36 four-taxon combinations[109]: four to test the level of introgression within clade 1 (i.e. *E. subulatus*, *E. crouaniorum*, *Ectocarpus* species 1, *Ectocarpus* species 2), 20 to test the level of introgression within clade 2 (i.e. *Ectocarpus* species 6, *Ectocarpus* species 7, *Ectocarpus* species 5, *Ectocarpus* species 9, *E. siliculosus*, *Ectocarpus* species 3) and 12 to test the level of introgression between these two clades. Tests were designed using a four-taxon fixed phylogeny (((P1,P2)P3)O), where P1 and P2 are closely related species from the same clade, P3 is a more divergent species that may have experienced admixture with one or both of the (P1,P2) taxa, and an out-group (O). *E. fasciculatus* was used as the out-group taxon for all ABBA-BABA tests. Details about how P1, P2 and P3 taxa were selected for each test are given in Table S37. Previous results of species tree inference were used to inform subsequent ABBA–BABA tests and to define the (((P1,P2)P3)O) phylogenies. ABBAs are sites at which the derived allele (called B) is shared between the taxa P2 and P3, whereas P1 carries the ancestral allele (called A), as defined by the outgroup while BABAs are sites at which the derived allele is shared between P1 and P3, whereas P2 carries the ancestral allele. Under incomplete lineage sorting, conflicting ABBA and BABA patterns should occur in equal frequencies, resulting in a D statistic equal to zero. Historical gene flow between P2 and P3 causes an excess of ABBA, generating positive values of D. Historical gene flow between P1 and P3 causes an excess of

BABA, generating negative values of D. Patterson's D-statistic was calculated for the concatenated alignments of 257 ortholog genes (~274 Kbp). Significance was detected using a block-jackknifing approach[109–111], with a block size of 5 Kbp. For the jackknife procedure, one block of adjacent sites was removed n times. A Z-score was finally obtained by dividing the value of the D statistic by the standard error over n sequences of 5 Kbp. The ParimonySplits network was reconstructed for the genus *Ectocarpus* using SplitsTree 4[71] with 1000 bootstrap replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses are described in detail in the relevant sections of the "Method details" section and the results of statistical tests are shown in the figures with full descriptions in the legends.

ADDITIONAL RESOURCES

The Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr) provides access to all the annotated genome assemblies described in this study as downloadable files. The output files from the Orthofinder[31], Interproscan[50], Hectar[72] and eggNOG-mapper[41] analyses, together with the results of the various DIAMOND[26] sequence similarity analyses (see section "Analyses aimed at deducing functional characteristics of predicted proteins"), can also be downloaded. In addition, the site provides genome browser interfaces for the genomes and multiple additional tools and resources including BLAST interfaces for genomes, proteomes and *de novo* transcriptomes, various experimental protocols, an RShiny-based transcriptomic aggregator for the model brown alga *Ectocarpus* species 7 strain Ec32 and a link to genome-wide metabolic networks for the Phaeoexplorer species.

Supplementary references

1. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

2. Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research *34*, 5623–5630. 10.1093/nar/gkl723.

3. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389–3402.

4. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.

5. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.

6. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology *19*, 455. 10.1089/cmb.2012.0021.

7. Liu, H., Wu, S., Li, A., Ruan, J., Wu, S., Li, A., and Ruan, J. (2021). SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte *2021*, 1–9. 10.46471/gigabyte.15.

8. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat Methods *17*, 155–158. 10.1038/s41592-019-0669-3.

9. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol *37*, 540–546. 10.1038/s41587-019-0072-8.

10. Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Dai, Q., Bray, T., Wang, Y.-X., Xing, J.-F., Huang, Z.-J., Wang, D.-P., et al. (2021). Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun *12*, 60. 10.1038/s41467-020-20236-7.

11. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746. 10.1101/gr.214270.116.

12. Aury, J.-M., and Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genomics and Bioinformatics *3*, lqab034. 10.1093/nargab/lqab034.

13. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. 10.7717/peerj.7359.

14. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217. 10.1093/bioinformatics/bts611.

15. Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829. 10.1101/gr.074492.107.

16. Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086–1092. 10.1093/bioinformatics/bts094.

17. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Research *43*, D222–D226. 10.1093/nar/gku1221.

18. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. 10.1093/bioinformatics/btu170.

19. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644–652. 10.1038/nbt.1883.

20. Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience *8*, giz100. 10.1093/gigascience/giz100.

21. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker. http://repeatmasker.org.

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

23. Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One *6*, e16526. 10.1371/journal.pone.0016526.

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

25. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995. 10.1101/gr.1865504.

26. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods *18*, 366–368. 10.1038/s41592-021-01101-x.

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

28. Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C.D., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M. (2016). Gmove a tool for eukaryotic gene predictions using various evidences. F1000Research *5*. 10.7490/f1000research.1111735.1.

29. Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics *30*, 3276–3278. 10.1093/bioinformatics/btu531.

30. Stamatakis, A. (2015). Using RAxML to Infer Phylogenies. Curr Protoc Bioinformatics *51*, 6.14.1-14. 10.1002/0471250953.bi0614s51.

31. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol *20*, 238. 10.1186/s13059-019-1832-y.

32. Csűös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912. 10.1093/bioinformatics/btq315.

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

34. Jehl, P., Sievers, F., and Higgins, D.G. (2015). OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics *16*, 269. 10.1186/s12859-015-0702-1.

35. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res *41*, e121. 10.1093/nar/gkt263.

36. Barrera-Redondo, J., Lotharukpong, J.S., Drost, H.-G., and Coelho, S.M. (2023). Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol *24*, 54. 10.1186/s13059-023-02895-z.

37. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575–1584. 10.1093/nar/30.7.1575.

38. Kempen, M. van, Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Preprint at bioRxiv, 10.1101/2022.02.07.479398 10.1101/2022.02.07.479398.

39. Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2018). CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. Mol Biol Evol *35*, 252–255. 10.1093/molbev/msx283.

40. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research *47*, D309–D314. 10.1093/nar/gky1085.

41. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution *38*, 5825–5829. 10.1093/molbev/msab293.

42. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. 10.1186/1471-2105-11-119.

43. Aylward, F.O., and Moniruzzaman, M. (2021). ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. Viruses *13*, 150. 10.3390/v13020150.

44. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

45. Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323–1330. 10.1093/bioinformatics/btw006.

46. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772–780. 10.1093/molbev/mst010.

47. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731–2739. 10.1093/molbev/msr121.

48. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. 10.1093/bioinformatics/btp348.

49. Petroll, R., Schreiber, M., Finke, H., Cock, J.M., Gould, S.B., and Rensing, S.A. (2021). Signatures of Transcription Factor Evolution and the Secondary Gain of Red Algae Complexity. Genes *12*, 1055. 10.3390/genes12071055.

50. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. 10.1093/bioinformatics/btu031.

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

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

53. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol *34*, 525–527. 10.1038/nbt.3519.

54. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-8.

55. Andrews, S. (2016). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

56. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. 10.1093/bioinformatics/btt656.

57. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586–1591. 10.1093/molbev/msm088.

58. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution *3*, 217–223. 10.1111/j.2041-210X.2011.00169.x.

59. Wallau, G.L., Capy, P., Loreto, E., Le Rouzic, A., and Hua-Van, A. (2016). VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within *Drosophila*. Mol Biol Evol *33*, 1094–1109. 10.1093/molbev/msv341.

60. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res *45*, e18. 10.1093/nar/gkw955.

61. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

62. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res *45*, W6–W11. 10.1093/nar/gkx391.

63. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res *32*, 11–16. 10.1093/nar/gkh152.

64. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods *14*, 587–589. 10.1038/nmeth.4285.

65. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol *35*, 518–522. 10.1093/molbev/msx281.

66. Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics *33*, 2197–2198. 10.1093/bioinformatics/btx144.

67. Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics *20*, 3643–3646. 10.1093/bioinformatics/bth397.

68. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol *10*, e1003537. 10.1371/journal.pcbi.1003537.

69. Douglas, J., Jiménez-Silva, C.L., and Bouckaert, R. (2022). StarBeast3: Adaptive Parallelized Bayesian Inference under the Multispecies Coalescent. Syst Biol *71*, 901–916. 10.1093/sysbio/syac010.

70. Bouckaert, R.R., and Drummond, A.J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC Evol Biol *17*, 42. 10.1186/s12862-017-0890-6.

71. Kloepper, T.H., and Huson, D.H. (2008). Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol *8*, 22. 10.1186/1471-2148-8-22.

72. Gschloessl, B., Guermeur, Y., and Cock, J. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinf *9*, 393.

73. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530–1534. 10.1093/molbev/msaa015.

74. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution *38*, 4647–4654. 10.1093/molbev/msab199.

75. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res *50*, D439–D444. 10.1093/nar/gkab1061.

76. Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J *6*, 223. 10.1186/1743-422X-6-223.

77. Barbeyron, T., Brillet-Guéguen, L., Carré, W., Carrière, C., Caron, C., Czjzek, M., Hoebeke, M., and Michel, G. (2016). Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. PLOS ONE *11*, e0164846. 10.1371/journal.pone.0164846.

78. Stam, M., Lelièvre, P., Hoebeke, M., Corre, E., Barbeyron, T., and Michel, G. (2023). SulfAtlas, the sulfatase database: state of the art and new developments. Nucleic Acids Res *51*, D647–D653. 10.1093/nar/gkac977.

79. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research *49*, D412–D419. 10.1093/nar/gkaa913.

80. Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Research *49*, D458–D460. 10.1093/nar/gkaa937.

81. Schlösser, U.G. (1994). SAG - Sammlung von Algenkulturen at the University of Göttingen Catalogue of Strains 1994. Botanica Acta *107*, 113–186. 10.1111/j.1438-8677.1994.tb00784.x.

82. Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J., Badger, J., et al. (2010). The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature *465*, 617–621. 10.1038/nature09016.

83. Cormier, A., Avia, K., Sterck, L., Derrien, T., Wucher, V., Andres, G., Monsoor, M., Godfroy, O., Lipinska, A., Perrineau, M.-M., et al. (2017). Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. New Phytol. *214*, 219–232. 10.1111/nph.14321.

84. Debit, A., Vincens, P., Bowler, C., and Carvalho, H.C. de (2023). LncPlankton V1.0: a comprehensive collection of plankton long non-coding RNAs. Preprint at bioRxiv, 10.1101/2023.11.03.565479 10.1101/2023.11.03.565479.

85. Akita, S., Vieira, C., Hanyuda, T., Rousseau, F., Cruaud, C., Couloux, A., Heesch, S., Cock, J.M., and Kawai, H. (2022). Providing a phylogenetic framework for trait-based analyses in brown algae: Phylogenomic tree inferred from 32 nuclear protein-coding sequences. Molecular Phylogenetics and Evolution *168*, 107408. 10.1016/j.ympev.2022.107408.

86. Weisman, C.M., Murray, A.W., and Eddy, S.R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol *18*, e3000862. 10.1371/journal.pbio.3000862.

87. Mulhair, P.O., Moran, R.J., Pathmanathan, J.S., Sussfeld, D., Creevey, C.J., Siu-Ting, K., Whelan, F.J., Pisani, D., Constantinides, B., Pelletier, E., et al. (2023). Bursts of novel composite gene families at major nodes in animal evolution. Preprint at bioRxiv, 10.1101/2023.07.10.548381 10.1101/2023.07.10.548381.

88. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. (2018). A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. mSphere *3*, e00069-18. 10.1128/mSphereDirect.00069-18.

89. Belcour, A., Got, J., Aite, M., Delage, L., Collén, J., Frioux, C., Leblanc, C., Dittami, S.M., Blanquart, S., Markov, G.V., et al. (2023). Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. Genome Res. *33*, 972–987. 10.1101/gr.277056.122.

90. Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Research *46*, D633–D639. 10.1093/nar/gkx935.

91. R Core Team (2018). R: A language and environment for statistical computing.

92. Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O., Guillaudeux, N., et al. (2018). Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLoS Comput. Biol. *14*, e1006146. 10.1371/journal.pcbi.1006146.

93. Macaisne, N., Liu, F., Scornet, D., Peters, A.F., Lipinska, A., Perrineau, M.-M., Henry, A., Strittmatter, M., Coelho, S.M., and Cock, J.M. (2017). The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development *144*, 409–418. 10.1242/dev.141523.

94. Talbert, P.B., Ahmad, K., Almouzni, G., Ausió, J., Berger, F., Bhalla, P.L., Bonner, W.M., Cande, W.Z., Chadwick, B.P., Chan, S.W.L., et al. (2012). A unified phylogeny-based nomenclature for histone variants. Epigenetics Chromatin *5*, 7. 10.1186/1756-8935-5-7.

95. Fu, G., Nagasato, C., Oka, S., Cock, J.M., and Motomura, T. (2014). Proteomics Analysis of Heterogeneous Flagella in Brown Algae (Stramenopiles). Protist *165*, 662–675. 10.1016/j.protis.2014.07.007.

96. Starr, R.C., and Zeikus, J.A. (1993). UTEX-The culture collection of algae at the University of Texas at Austin 1993 list of cultures. J Phycol *29 (Suppl.)*, 1–106.

97. Dittami, S.M., Gravot, A., Goulitquer, S., Rousvoal, S., Peters, A.F., Bouchereau, A., Boyen, C., and Tonon, T. (2012). Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to low salinities in *Ectocarpus* (brown algae). Plant J. 10.1111/j.1365-313X.2012.04982.x.

98. Rahman, S., Kosakovsky Pond, S.L., Webb, A., and Hey, J. (2021). Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. Proc Natl Acad Sci U S A *118*, e2023575118. 10.1073/pnas.2023575118.

99. Duchêne, S., Holmes, E.C., and Ho, S.Y.W. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc Biol Sci *281*, 20140732. 10.1098/rspb.2014.0732.

100. Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics *139*, 1067–1076. 10.1093/genetics/139.2.1067.

101. Akashi, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene *205*, 269–278. 10.1016/s0378-1119(97)00400-9.

102. Subramanian, S. (2008). Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes. Genetics *178*, 2429–2432. 10.1534/genetics.107.086405.

103. Wright, F. (1990). The 'effective number of codons' used in a gene. Gene *87*, 23–29. 10.1016/0378-1119(90)90491-9.

104. Forcelloni, S., and Giansanti, A. (2020). Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. J Mol Evol *88*, 164–178. 10.1007/s00239-019-09921-4.

105. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493. 10.1101/gr.113985.110.

106. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research *50*, W276–W279. 10.1093/nar/gkac240.

107. Steffen, R., Ogoniak, L., Grundmann, N., Pawluchin, A., Soehnlein, O., and Schmitz, J. (2022). paPAML: An Improved Computational Tool to Explore Selection Pressure on Protein-Coding Sequences. Genes *13*, 1090. 10.3390/genes13061090.

108. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology *10*, 210. 10.1186/1471-2148-10-210.

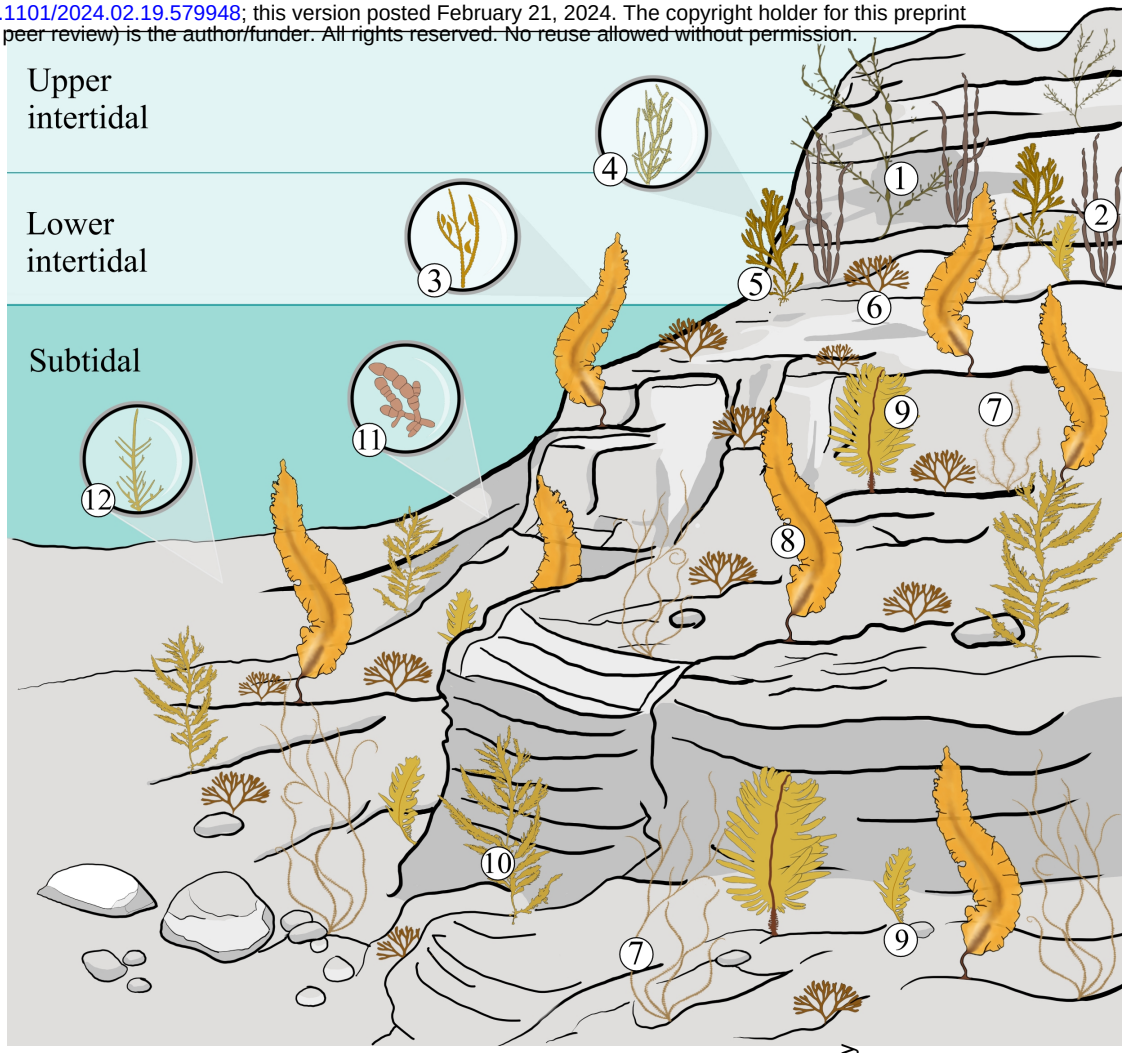
109. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722. 10.1126/science.1188021.

110. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol Biol Evol *28*, 2239–2252. 10.1093/molbev/msr048.

111. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. 10.1534/genetics.112.145037.

Figure 1

① *Ascophyllum nodosum*
② *Scytosiphon promiscuus*
③ *Ectocarpus siliculosus*
④ *Pylaiella littoralis*
⑤ *Fucus serratus*
⑥ *Dictyota dichotoma*
⑦ *Chordaria linearis*
⑧ *Saccharina latissima*
⑨ *Undaria pinnatifida*
⑩ *Desmarestia herbacea*
⑪ *Schizocladia ischiensis*
⑫ *Discosporangium mesarthocarpum*



Major events during evolution

Gain
- A Alginate-based ECM
- B Plasmodesmata
- C Basal attachment system
  Parenchymatous growth
- D Heteromorphic life cycles
- E Desiccation tolerance
  Diploid life cycle

Loss
- F Lateral flagellum
- G Eyespot

Morphology & complexity

- unicellular
- simple multicellularity
- filamentous
- pseudo-parenchymatous
- complex thallus

<3 cell types
>10

Maximum thallus size

µm mm cm m

Typical life cycle

- diploid
- n<2n
- n=2n
- n>2n
- haploid
- probably diplontic
- probably haplodiplontic
- unknown

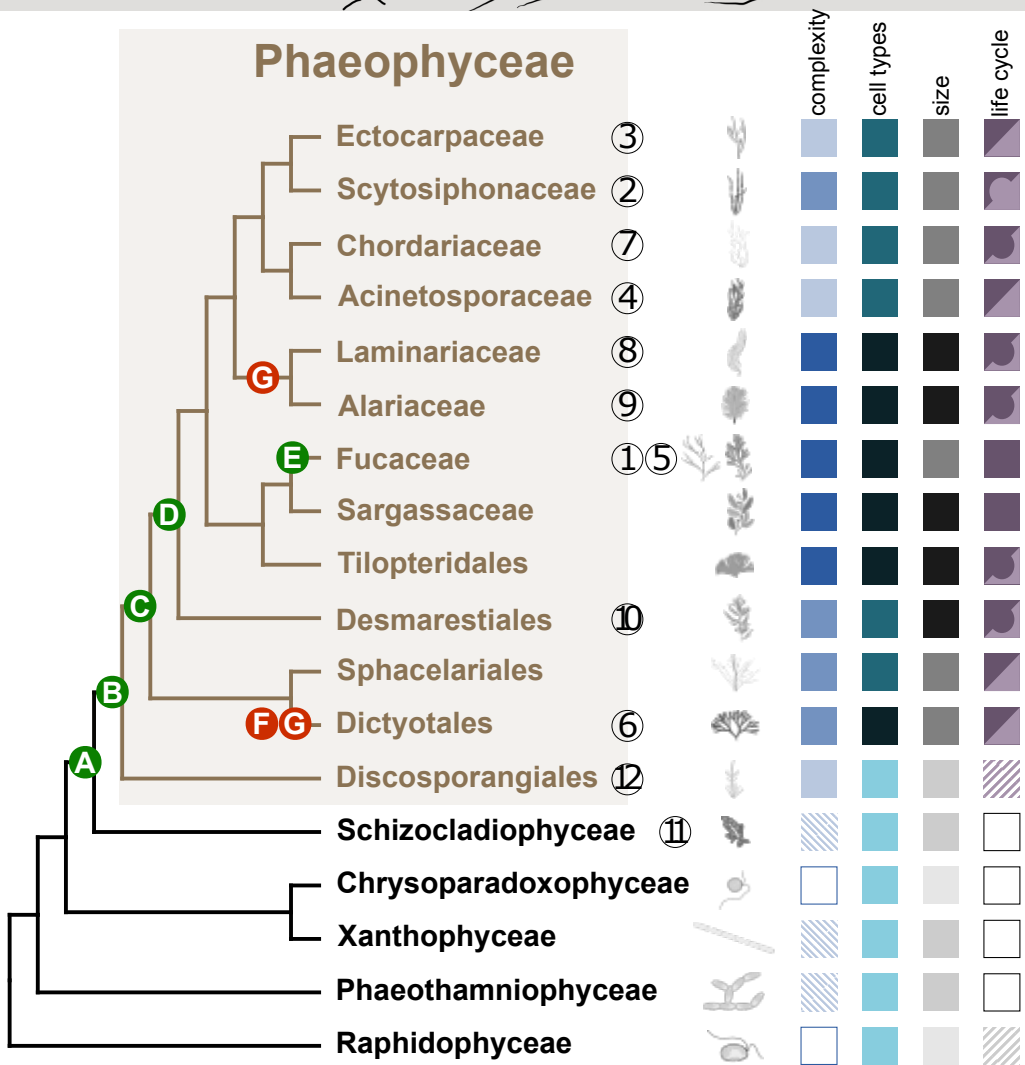


Figure 2

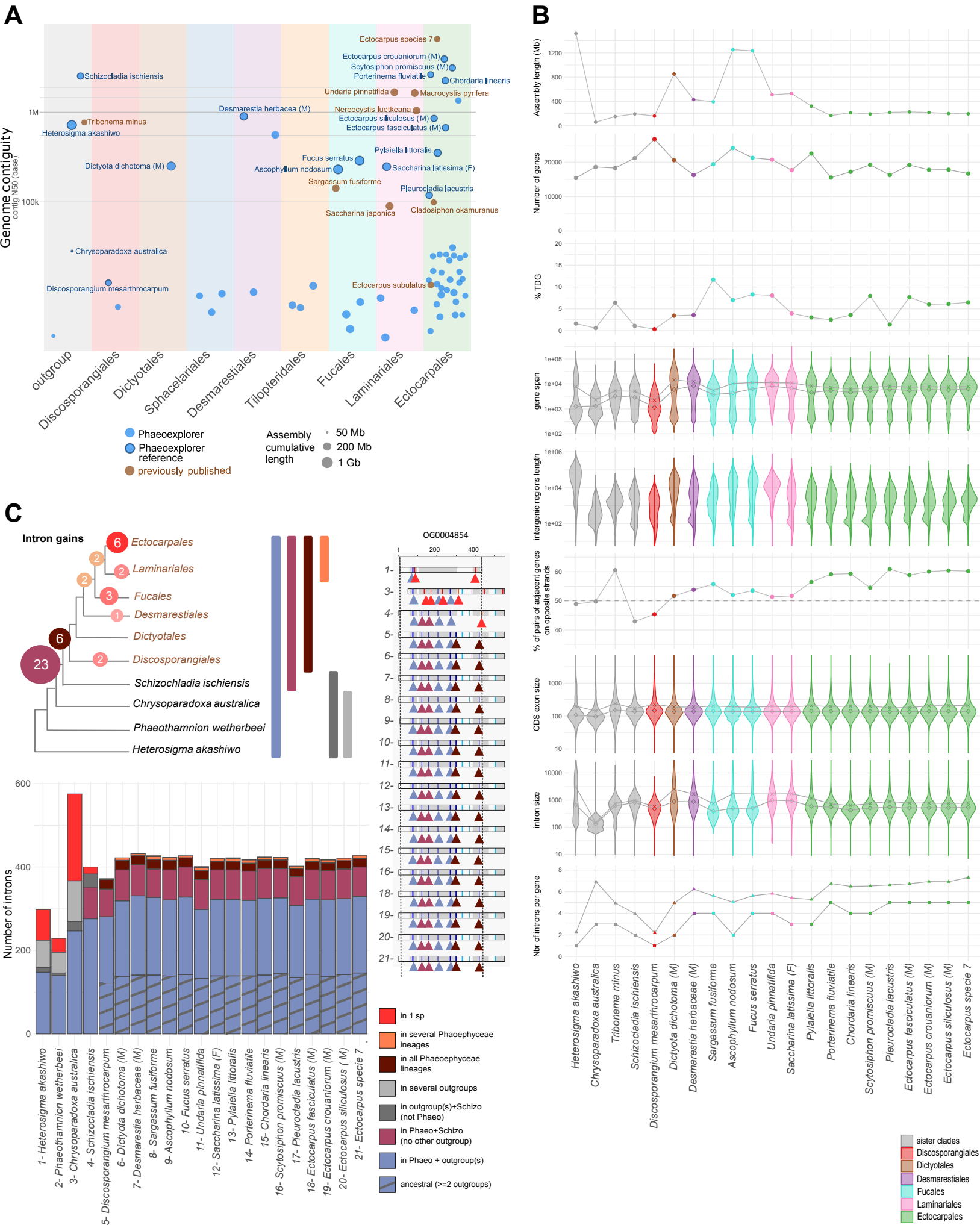
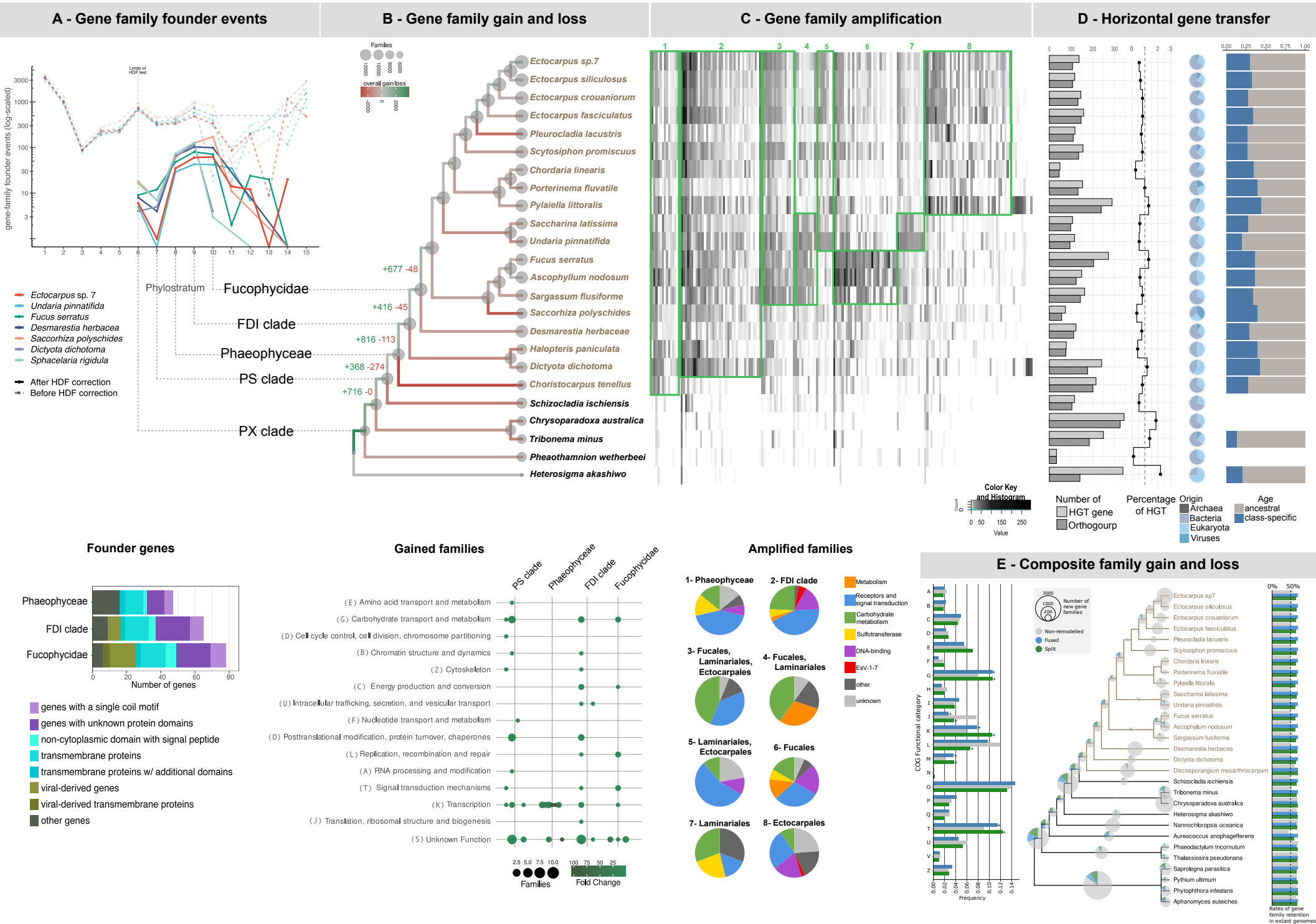


Figure 3


A - Gene family founder events
B - Gene family gain and loss
C - Gene family amplification
D - Horizontal gene transfer
gene-family founder events (log-scaled)
Limits of HDF test
Phylostratum
Fucophycidae
FDI clade
Phaeophyceae
PS clade
PX clade
Ectocarpus sp. 7
Undaria pinnatifida
Fucus serratus
Desmarestia herbacea
Saccorhiza polyschides
Dictyota dichotoma
Sphacelaria rigidula
After HDF correction
Before HDF correction
Families
overall gain/loss
+677 -48
+416 -45
+816 -113
+368 -274
+716 -0
Ectocarpus sp.7
Ectocarpus siliculosus
Ectocarpus crouaniorum
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Chordaria linearis
Porterinema fluvatile
Pylaiella littoralis
Saccharina latissima
Undaria pinnatifida
Fucus serratus
Ascophyllum nodosum
Sargassum flusiforme
Saccorhiza polyschides
Desmarestia herbaceae
Halopteris paniculata
Dictyota dichotoma
Choristocarpus tenellus
Schizocladia ischiensis
Chrysoparadoxa australica
Tribonema minus
Pheaothamnion wetherbeei
Heterosigma akashiwo
Color Key and Histogram
Count
Value
Number of HGT gene
Orthogourp
Percentage of HGT
Origin
Archaea
Bacteria
Eukaryota
Viruses
Age
ancestral
class-specific

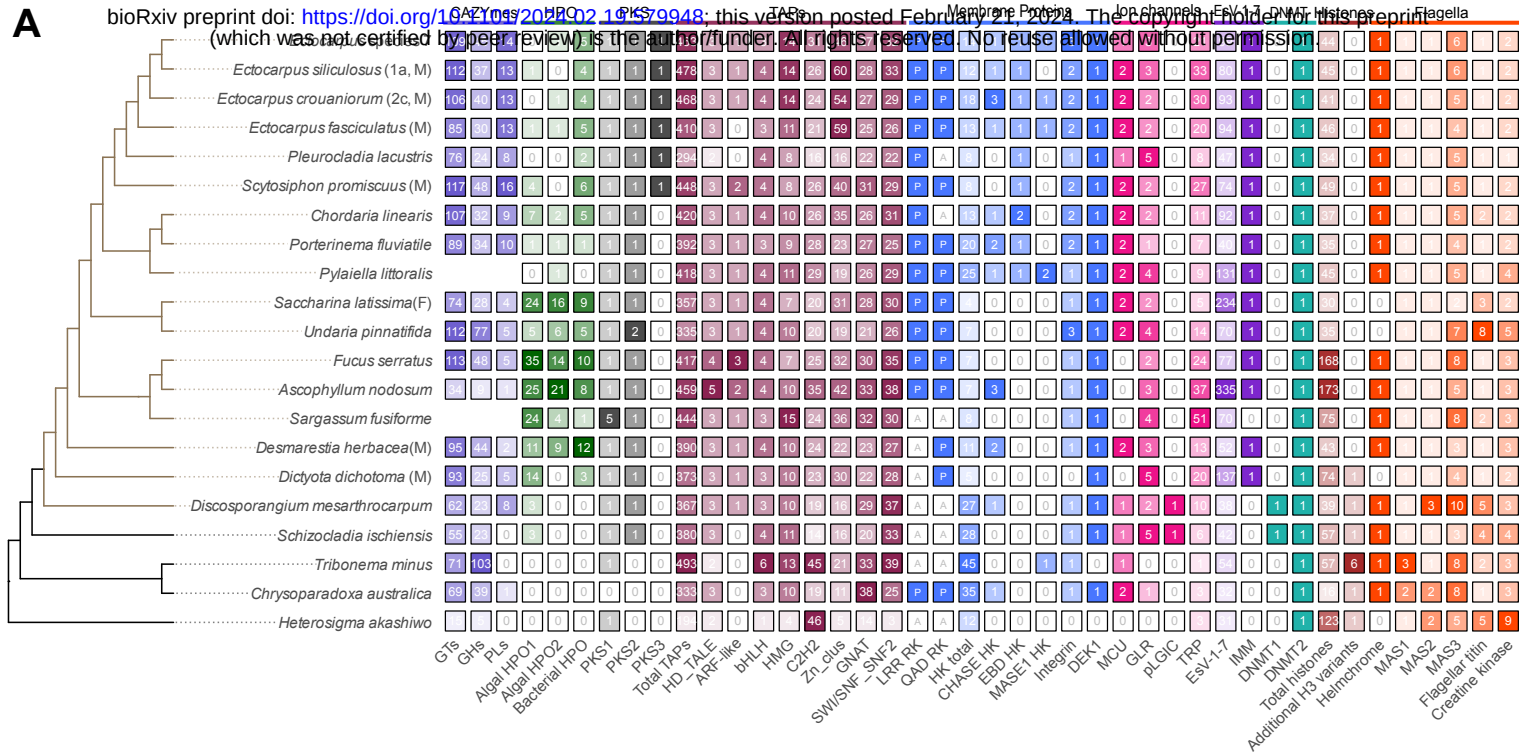
Founder genes
Phaeophyceae
FDI clade
Fucophycidae
Number of genes
genes with a single coil motif
genes with unknown protein domains
non-cytoplasmic domain with signal peptide
transmembrane proteins
transmembrane proteins w/ additional domains
viral-derived genes
viral-derived transmembrane proteins
other genes

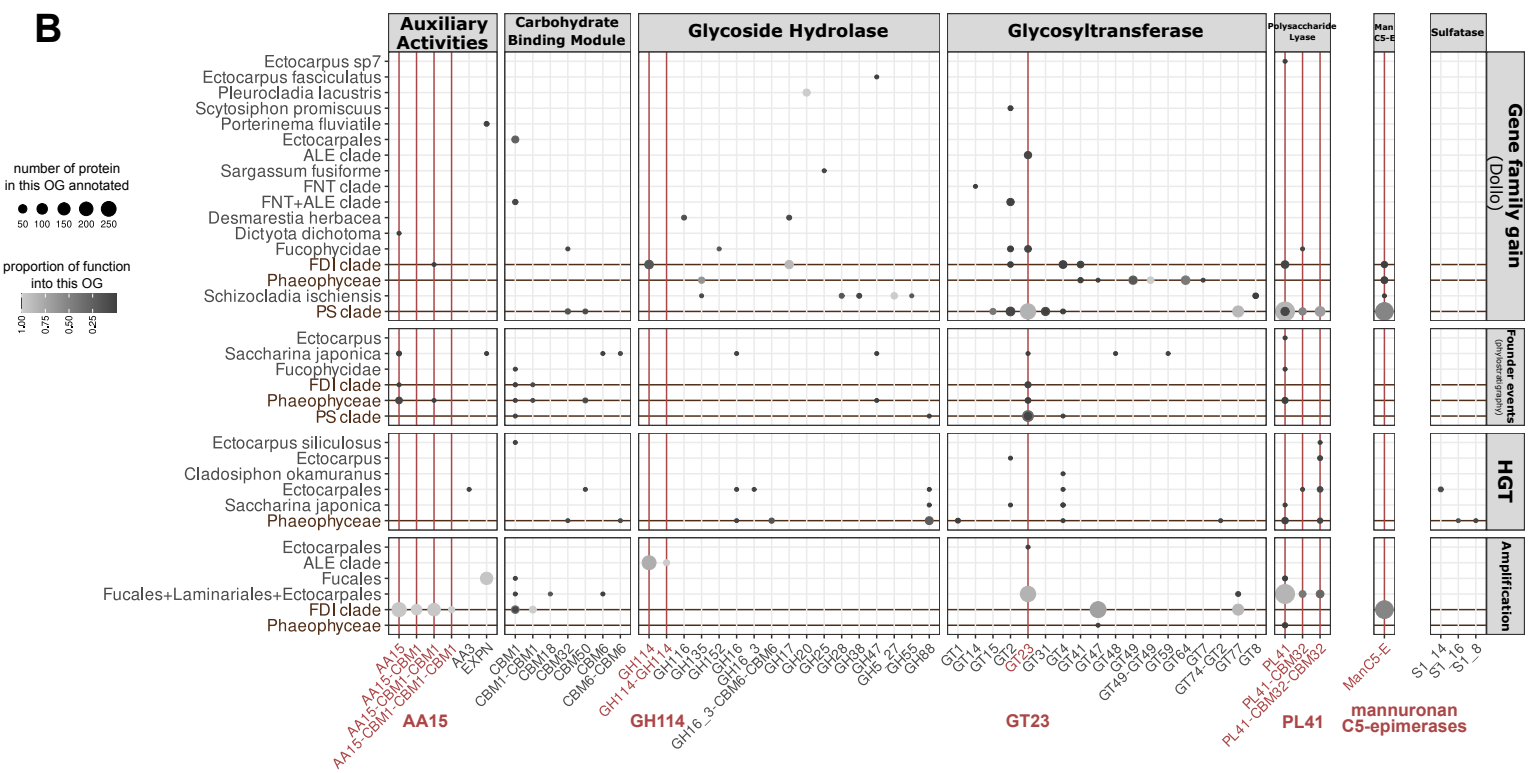
Gained families
PS clade
Phaeophyceae
FDI clade
Fucophycidae
(E) Amino acid transport and metabolism
(G) Carbohydrate transport and metabolism
(D) Cell cycle control, cell division, chromosome partitioning
(B) Chromatin structure and dynamics
(Z) Cytoskeleton
(C) Energy production and conversion
(U) Intracellular trafficking, secretion, and vesicular transport
(F) Nucleotide transport and metabolism
(O) Posttranslational modification, protein turnover, chaperones
(L) Replication, recombination and repair
(A) RNA processing and modification
(T) Signal transduction mechanisms
(K) Transcription
(J) Translation, ribosomal structure and biogenesis
(S) Unknown Function
Families
Fold Change

Amplified families
1- Phaeophyceae
2- FDI clade
3- Fucales, Laminariales, Ectocarpales
4- Fucales, Laminariales
5- Laminariales, Ectocarpales
6- Fucales
7- Laminariales
8- Ectocarpales
Metabolism
Receptors and signal transduction
Carbohydrate metabolism
Sulfotransferase
DNA-binding
EsV-1-7
other
unknown

E - Composite family gain and loss
COG Functional category
Frequency
Number of new gene families
Non-remodelled
Fused
Split
Ectocarpus sp7
Ectocarpus siliculosus
Ectocarpus crouaniorum
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Chordaria linearis
Porterinema fluviatile
Pylaiella littoralis
Saccharina latissima
Undaria pinnatifida
Fucus serratus
Ascophyllum nodosum
Sargassum fusiforme
Desmarestia herbacea
Dictyota dichotoma
Discosporangium mesarthrocarpum
Schizocladia ischiensis
Tribonema minus
Chrysoparadoxa australica
Heterosigma akashiwo
Nannochloropsis oceanica
Aureococcus anophagefferens
Phaeodactylum tricornutum
Thalassiosira pseudonana
Saprolegna parasitica
Pythium ultimum
Phytophthora infestans
Aphanomyces euteiches
Rates of gene family retention in extant genomes

Figure 4





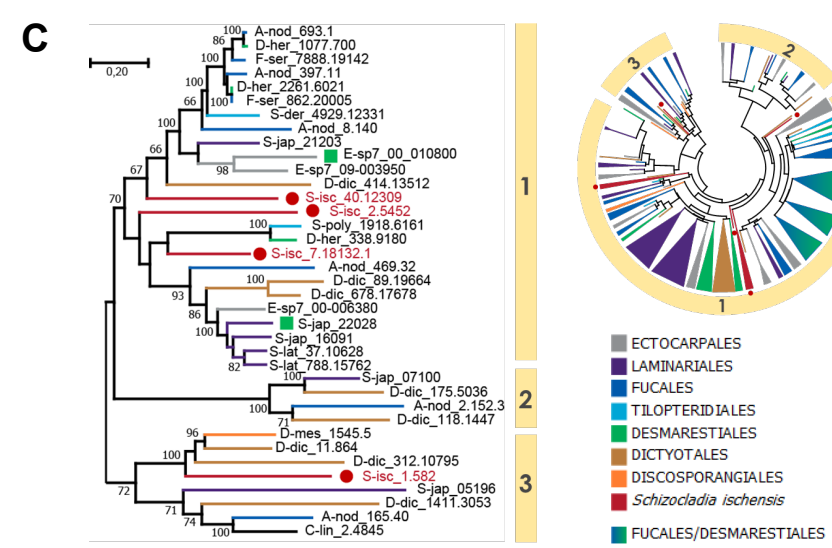
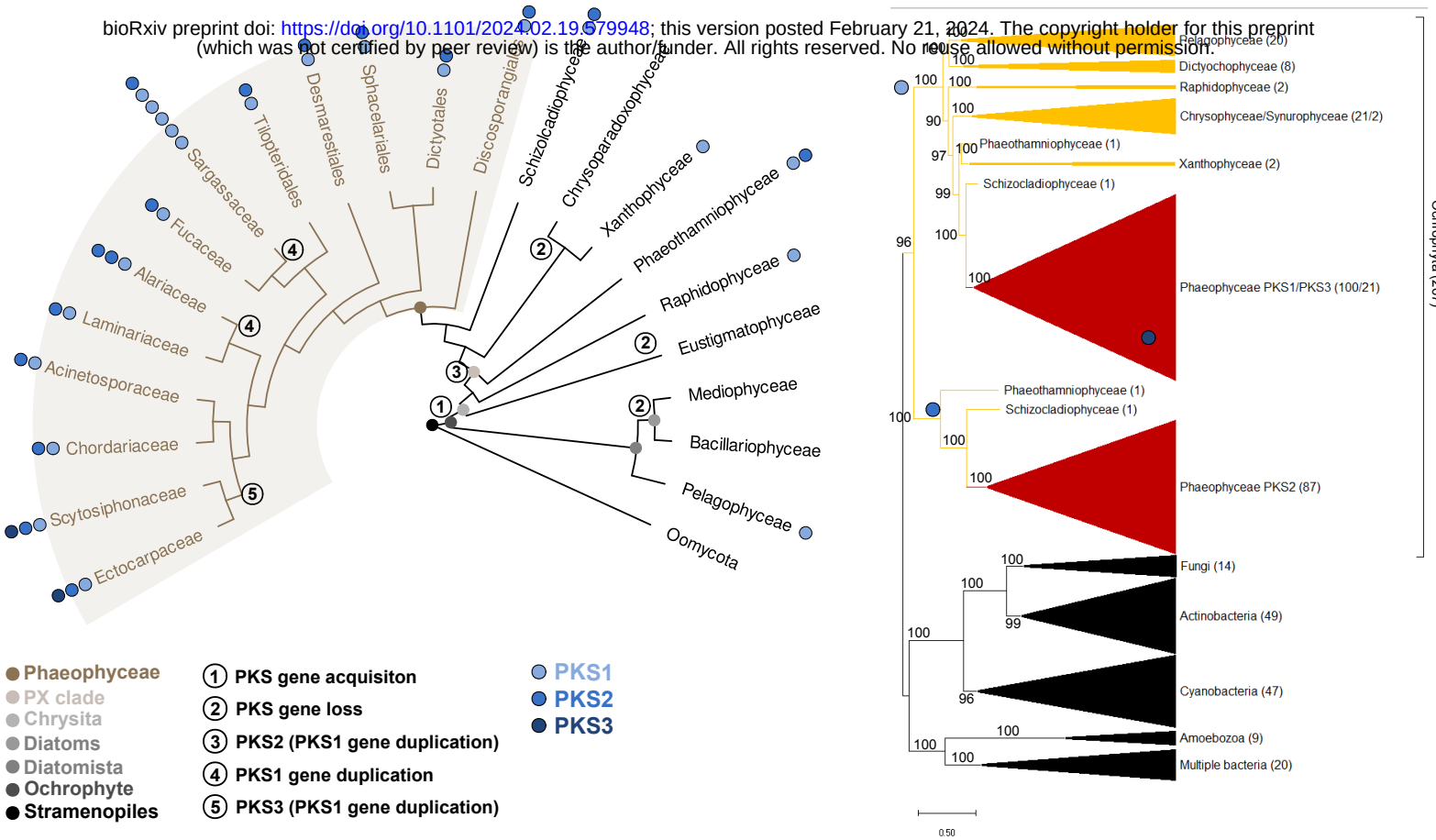
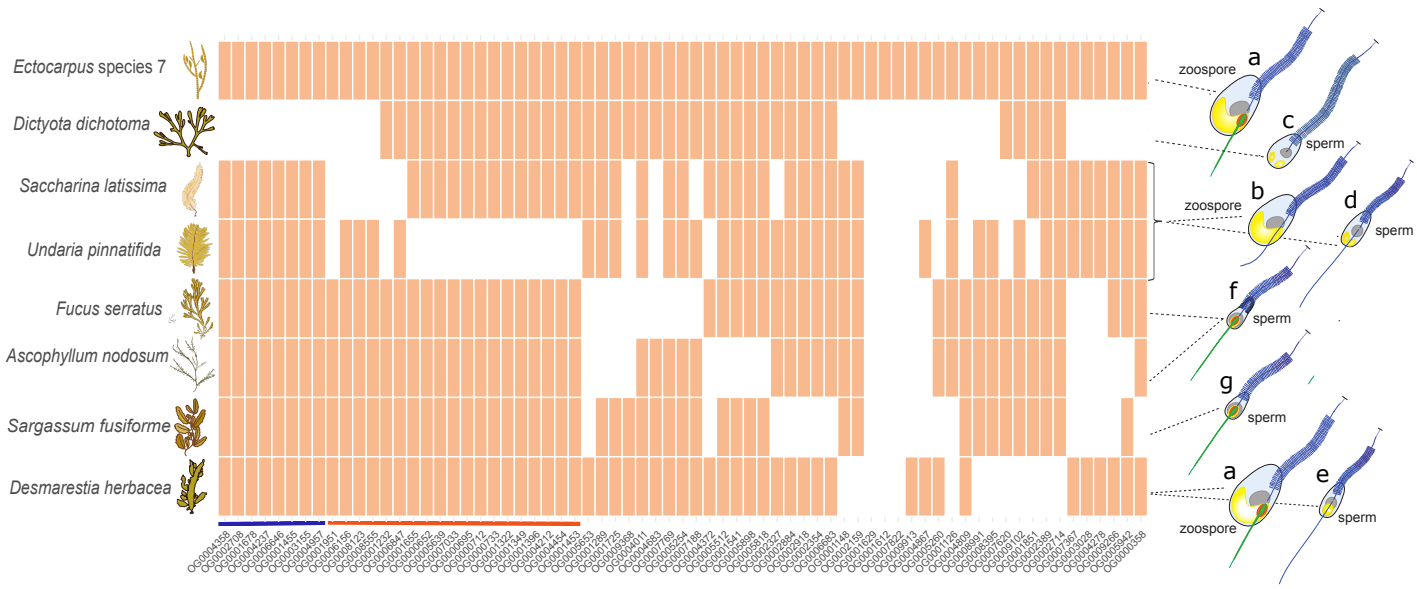


Figure 5

A



B



C

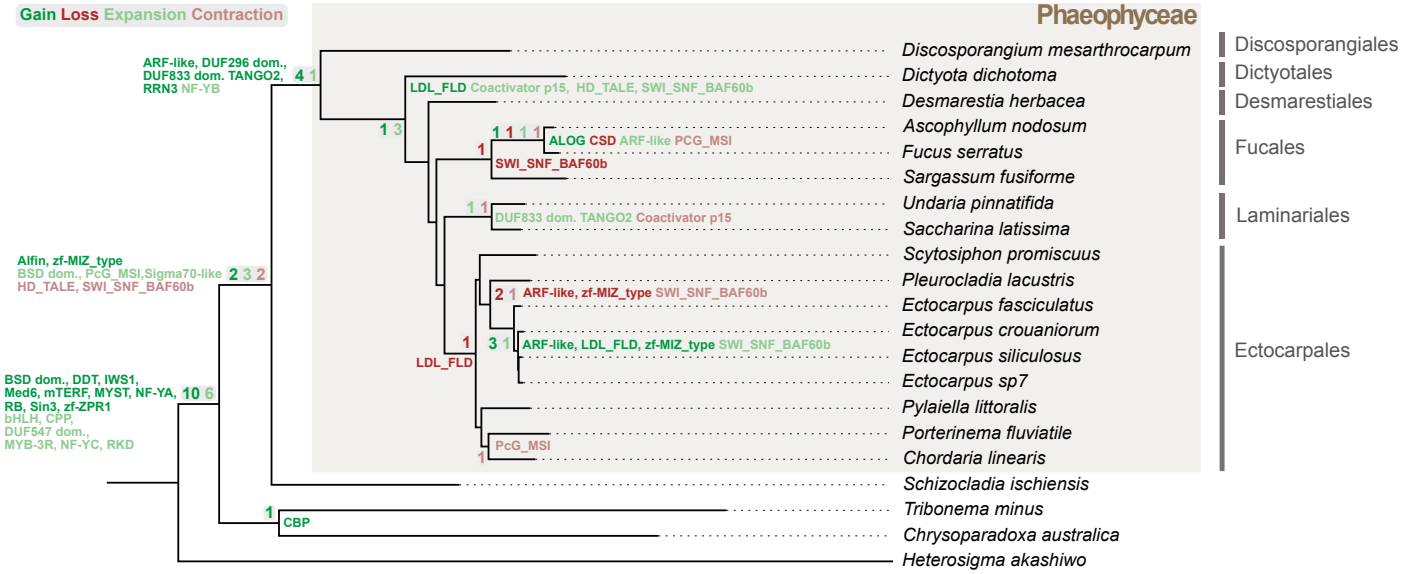


Figure 6

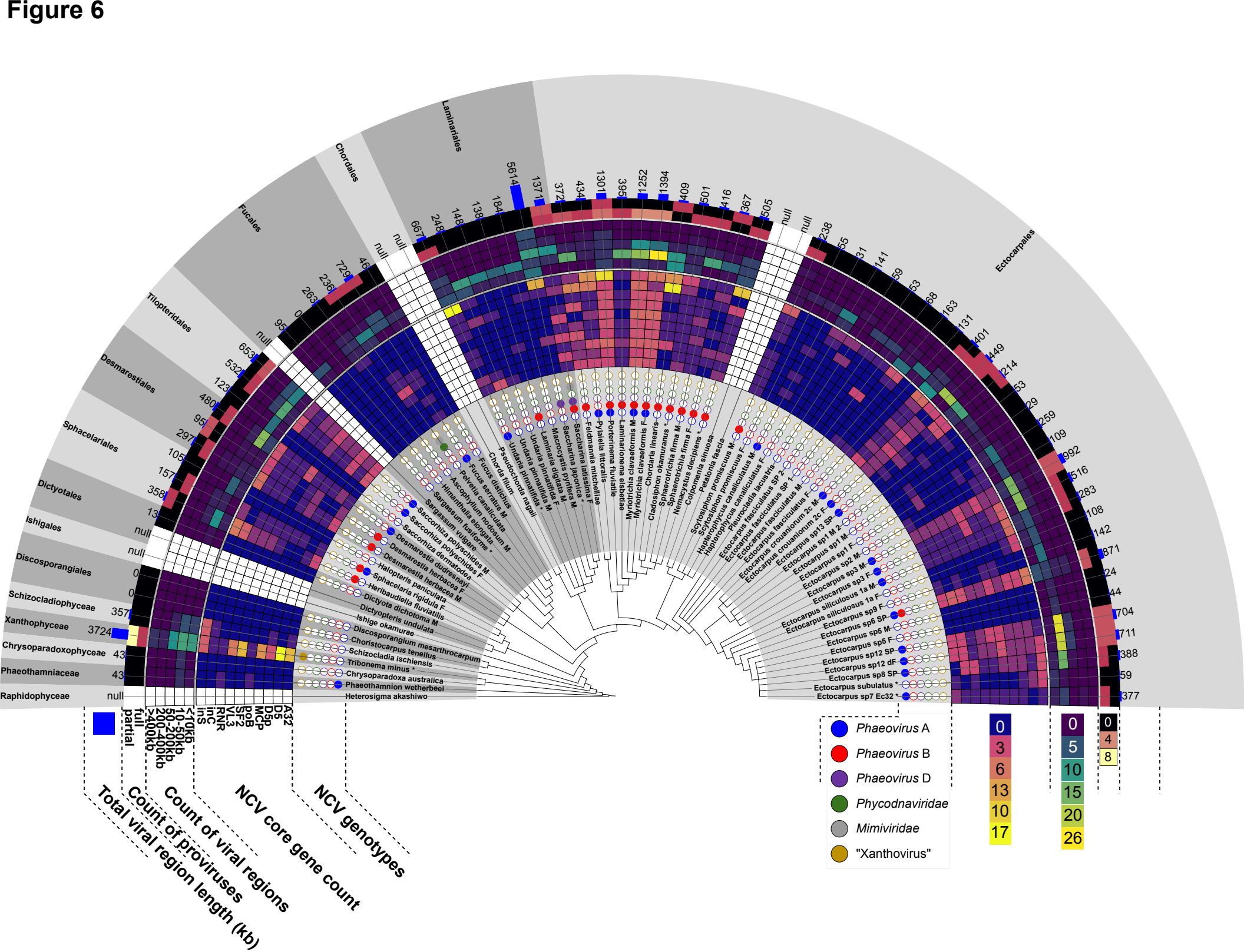

Laminariales
Chordales
Fucales
Tilopteridales
Desmarestiales
Sphacelariales
Dictyotales
Ishigales
Discosporangiales
Schizocladiophyceae
Xanthophyceae
Chrysoparadoxophyceae
Phaeothamniaceae
Raphidophyceae
Ectocarpales
Saccharina japonica *
Saccharina latissima F
Feldmannia mitchelliae
Pylaiella littoralis
Porterinema fluviatile
Laminarionema elsbetiae
Myriotrichia clavaeformis M
Myriotrichia clavaeformis F
Chordaria linearis
Cladosiphon okamuranus *
Sphaerotrichia firma M
Sphaerotrichia firma F
Nemacystus decipiens *
Colpomenia sinuosa
Petalonia fascia
Scytosiphon promiscuus M
Scytosiphon promiscuus F
Hapterophycus canaliculatus M
Hapterophycus canaliculatus F
Pleurocladia lacustris
Ectocarpus fasciculatus SP 2
Ectocarpus fasciculatus M
Ectocarpus fasciculatus F
Ectocarpus crouaniorum 2c M
Ectocarpus crouaniorum 2c F
Ectocarpus sp13 SP
Ectocarpus sp1 M 2
Ectocarpus sp1 M
Ectocarpus sp1 F
Ectocarpus sp2 M
Ectocarpus sp3 M
Ectocarpus sp3 F
Ectocarpus siliculosus 1a M
Ectocarpus siliculosus 1a F
Ectocarpus sp9 F
Ectocarpus sp6 SP
Ectocarpus sp5 M
Ectocarpus sp5 F
Ectocarpus sp12 SP
Ectocarpus sp12 dF
Ectocarpus sp8 SP
Ectocarpus subulatus *
Ectocarpus sp7 Ec32 *
Undaria pinnatifida M
Undaria pinnatifida F
Laminaria digitata M
Macrocystis pyrifera M
Pseudochorda nagaii
Chorda filum
Fucus distichus
Fucus serratus M
Pelvetia canaliculata
Ascophyllum nodosum M
Himanthalia elongata *
Sargassum fusiforme *
Sargassum vulgare
Saccorhiza polyschides M
Saccorhiza polyschides F
Saccorhiza dermatodea
Desmarestia dudresnayi
Desmarestia herbacea M
Desmarestia herbacea F
Halopteris paniculata
Sphacelaria rigidula F
Heribaudiella fluviatilis
Dictyota dichotoma M
Dictyopteris undulata
Ishige okamurae
Discosporangium mesarthrocarpum
Choristocarpus tenellus
Schizocladia ischiensis
Tribonema minus *
Chrysoparadoxa australica
Phaeothamnion wetherbeei
Heterosigma akashiwo
Phaeovirus A
Phaeovirus B
Phaeovirus D
Phycodnaviridae
Mimiviridae
"Xanthovirus"
full
partial
<10kb
10-50kb
50-200kb
200-400kb
>400kb
inC
inS
RNR
VL3
SF2
PolB
MCP
D5p
D5
A32
Total viral region length (kb)
Count of proviruses
Count of viral regions
NCV core gene count
NCV genotypes

Figure 7

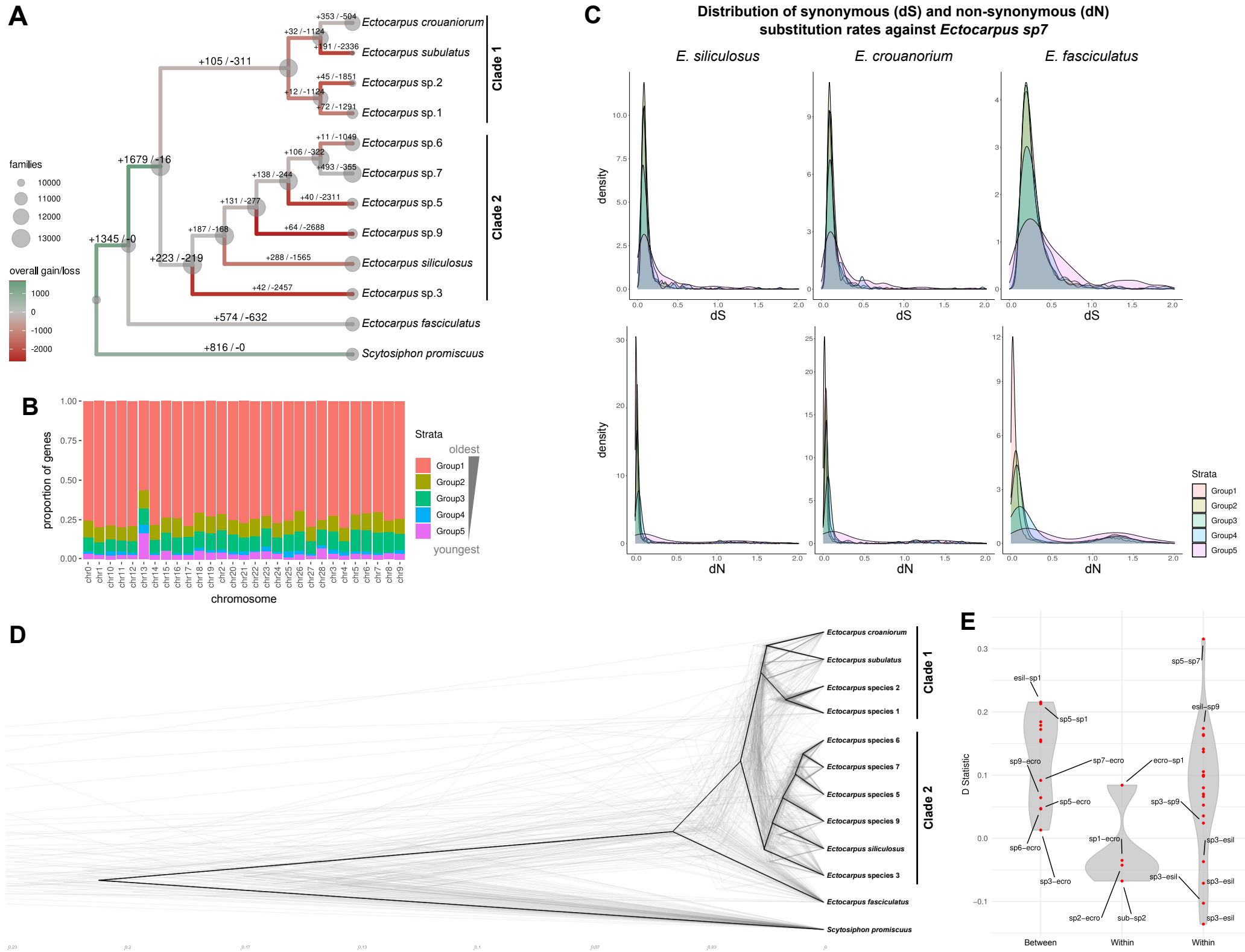

A
families
10000
11000
12000
13000
overall gain/loss
1000
0
-1000
-2000
+353 / -504
Ectocarpus crouaniorum
+32 / -1124
+191 / -2336
Ectocarpus subulatus
+105 / -311
+45 / -1851
Ectocarpus sp.2
+12 / -1124
+72 / -1291
Ectocarpus sp.1
Clade 1
+11 / -1049
Ectocarpus sp.6
+106 / -322
+493 / -355
Ectocarpus sp.7
+1679 / -16
+138 / -244
+40 / -2311
Ectocarpus sp.5
+131 / -277
+64 / -2688
Ectocarpus sp.9
+187 / -168
+1345 / -0
+288 / -1565
Ectocarpus siliculosus
+223 / -219
+42 / -2457
Ectocarpus sp.3
Clade 2
+574 / -632
Ectocarpus fasciculatus
+816 / -0
Scytosiphon promiscuus
B
proportion of genes
1.00
0.75
0.50
0.25
0.00
chr0 chr1 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr2 chr20 chr21 chr22 chr23 chr24 chr25 chr26 chr27 chr28 chr3 chr4 chr5 chr6 chr7 chr8 chr9
chromosome
Strata
oldest
Group1
Group2
Group3
Group4
Group5
youngest
C
Distribution of synonymous (dS) and non-synonymous (dN) substitution rates against Ectocarpus sp7
E. siliculosus
E. crouanorium
E. fasciculatus
density
dS
dN
Strata
Group1
Group2
Group3
Group4
Group5
D
Ectocarpus croaniorum
Ectocarpus subulatus
Ectocarpus species 2
Ectocarpus species 1
Clade 1
Ectocarpus species 6
Ectocarpus species 7
Ectocarpus species 5
Ectocarpus species 9
Ectocarpus siliculosus
Ectocarpus species 3
Clade 2
Ectocarpus fasciculatus
Scytosiphon promiscuus
E
D Statistic
esil–sp1
sp5–sp1
sp9–ecro
sp7–ecro
sp5–ecro
sp6–ecro
sp3–ecro
sp5–sp7
esil–sp9
ecro–sp1
sp3–sp9
sp1–ecro
sp3–esil
sp2–ecro
sub–sp2
Between clades
Within clade1
Within clade2

Figure S1

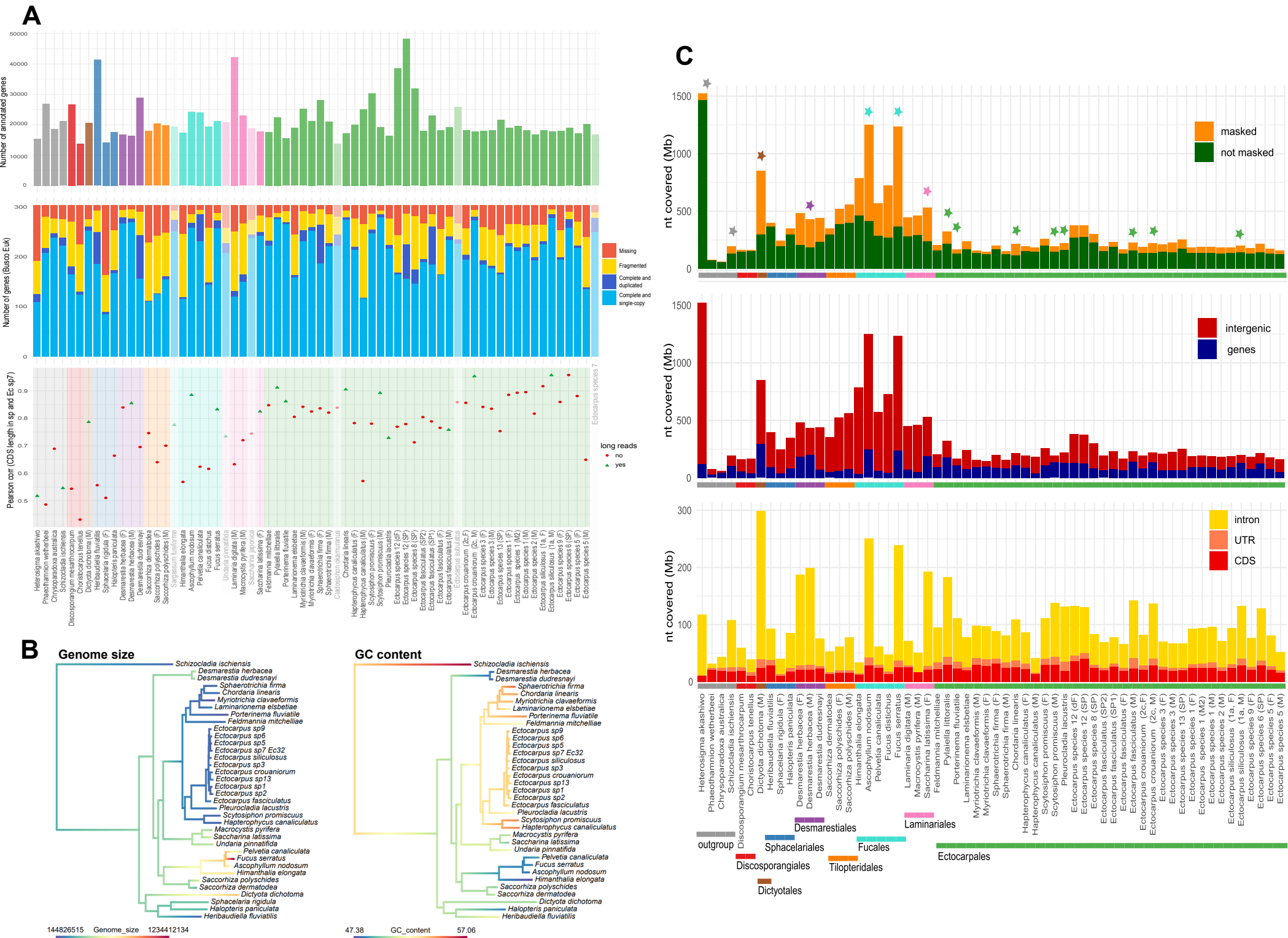

A
Number of annotated genes
Number of genes (Busco Euk)
Missing
Fragmented
Complete and duplicated
Complete and single-copy
Pearson corr (CDS length in sp and Ec sp7)
long reads
no
yes
B
Genome size
Schizocladia ischiensis
Desmarestia herbacea
Desmarestia dudresnayi
Sphaerotrichia firma
Chordaria linearis
Myriotrichia clavaeformis
Laminarionema elsbetiae
Porterinema fluviatile
Feldmannia mitchelliae
Ectocarpus sp9
Ectocarpus sp6
Ectocarpus sp5
Ectocarpus sp7 Ec32
Ectocarpus siliculosus
Ectocarpus sp3
Ectocarpus crouaniorum
Ectocarpus sp13
Ectocarpus sp1
Ectocarpus sp2
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Hapterophycus canaliculatus
Macrocystis pyrifera
Saccharina latissima
Undaria pinnatifida
Pelvetia canaliculata
Fucus serratus
Ascophyllum nodosum
Himanthalia elongata
Saccorhiza polyschides
Saccorhiza dermatodea
Dictyota dichotoma
Sphacelaria rigidula
Halopteris paniculata
Heribaudiella fluviatilis
144826515 Genome_size 1234412134
GC content
47.38 GC_content 57.06
C
nt covered (Mb)
masked
not masked
intergenic
genes
intron
UTR
CDS
outgroup
Desmarestiales
Laminariales
Sphacelariales
Fucales
Discosporangiales
Tilopteridales
Dictyotales
Ectocarpales

Figure S2

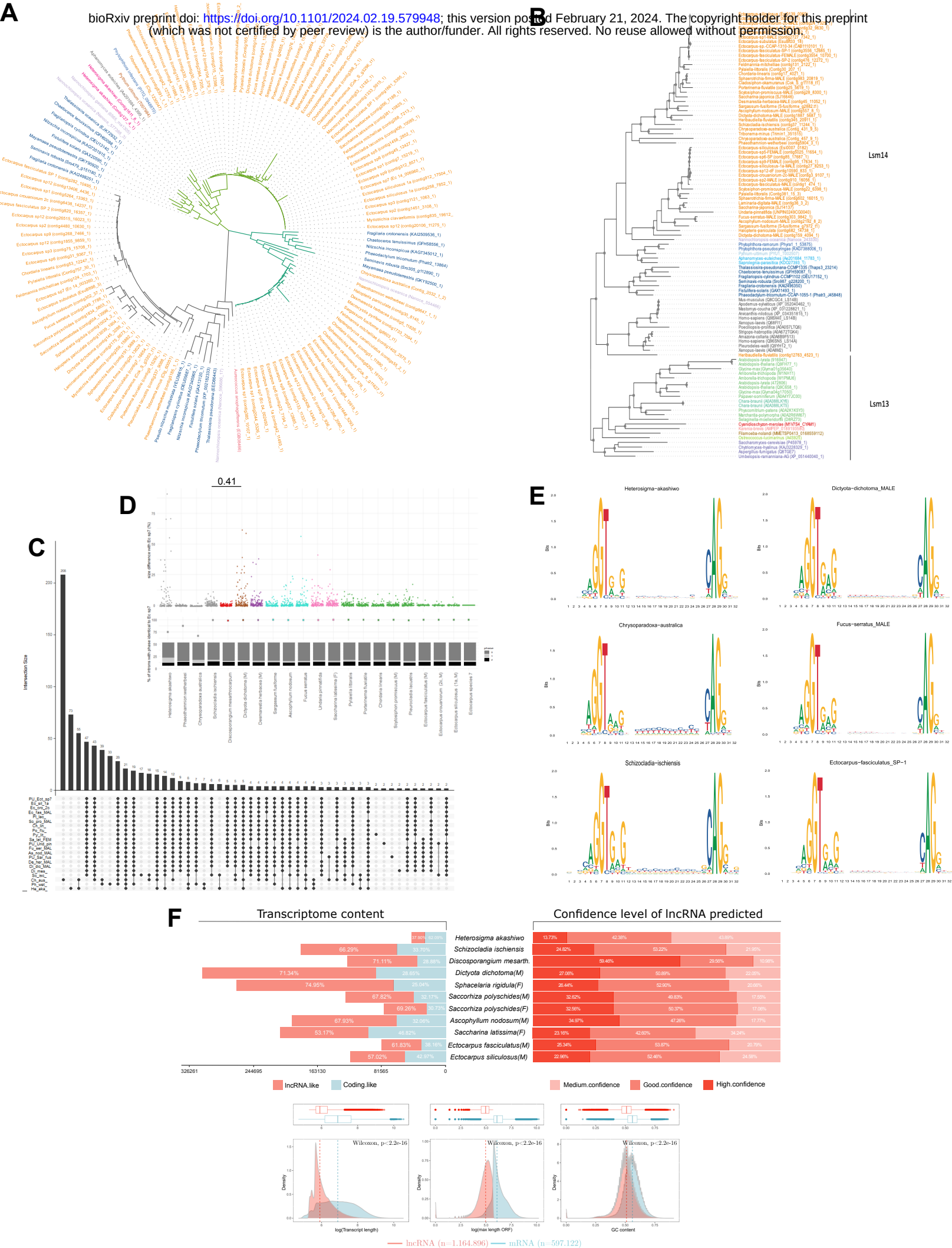
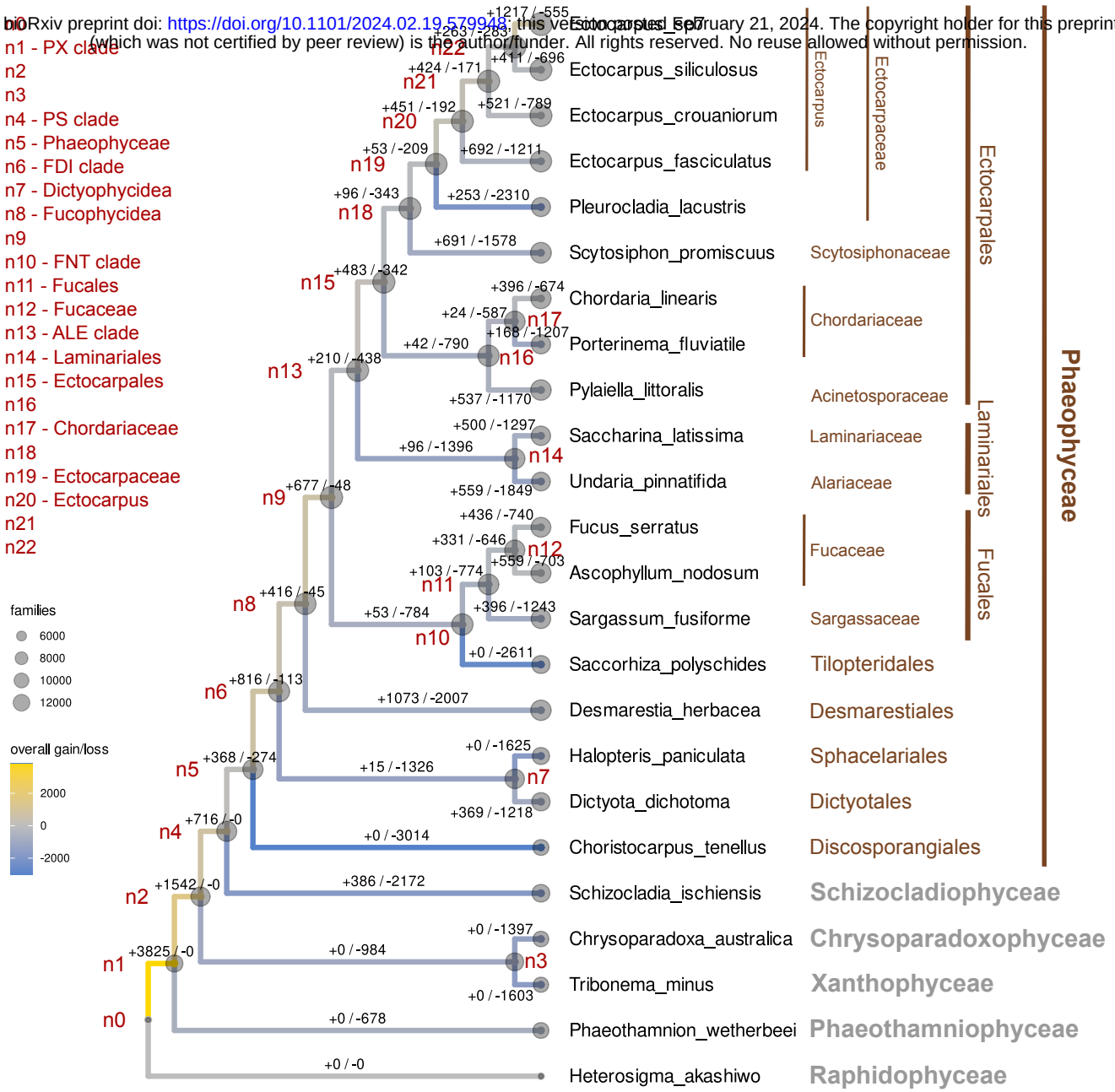


Figure S3

A



B

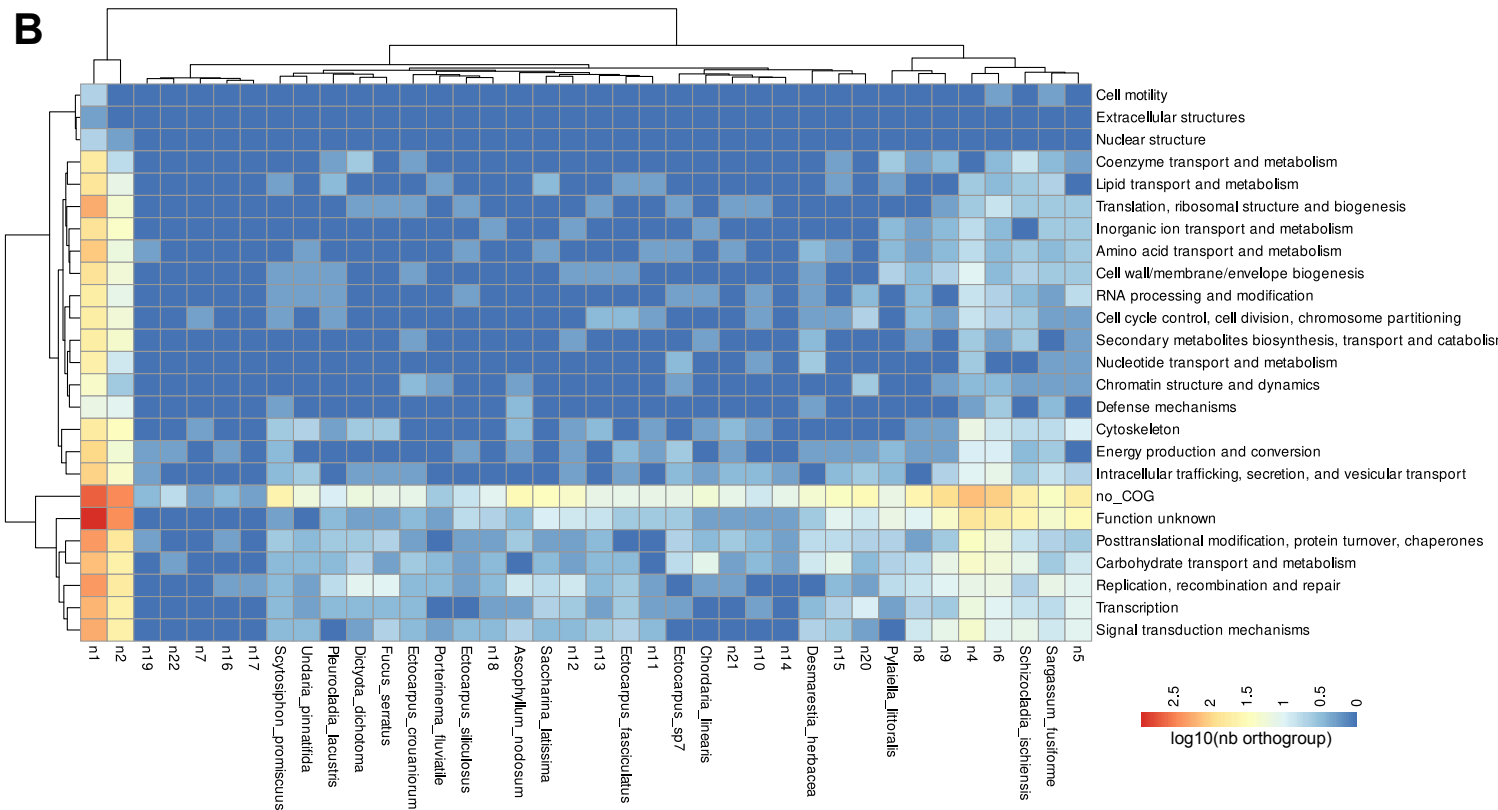


Figure S4

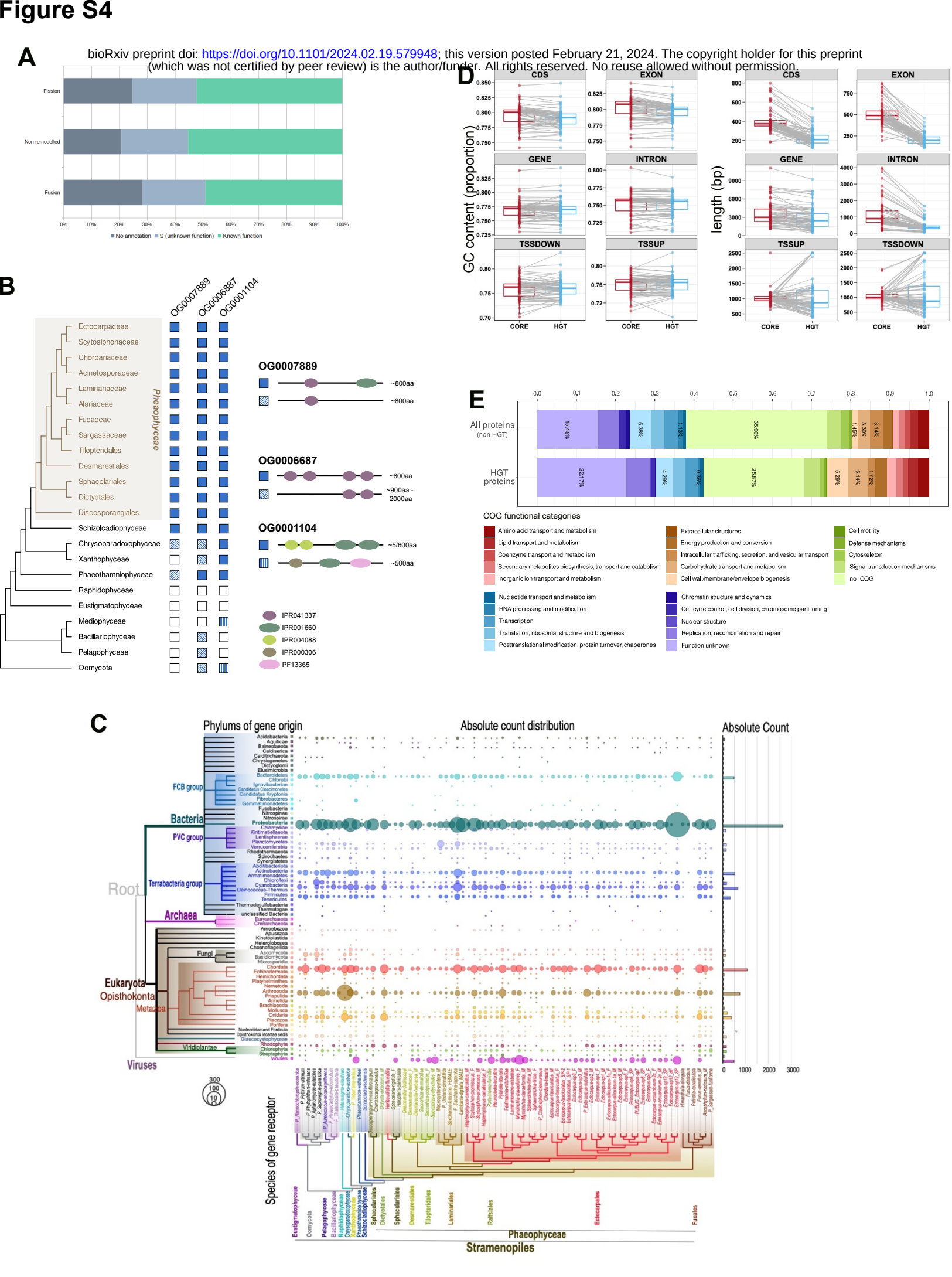
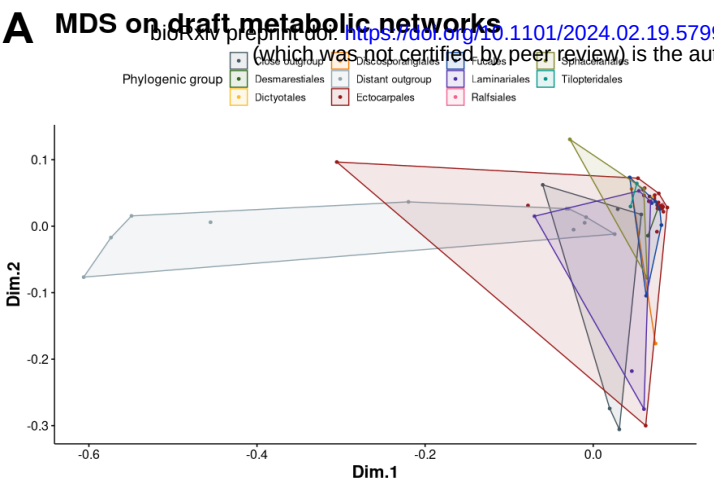
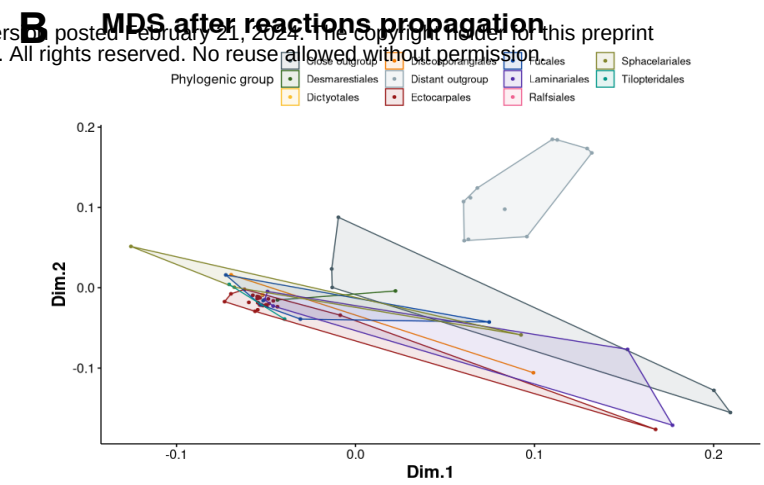
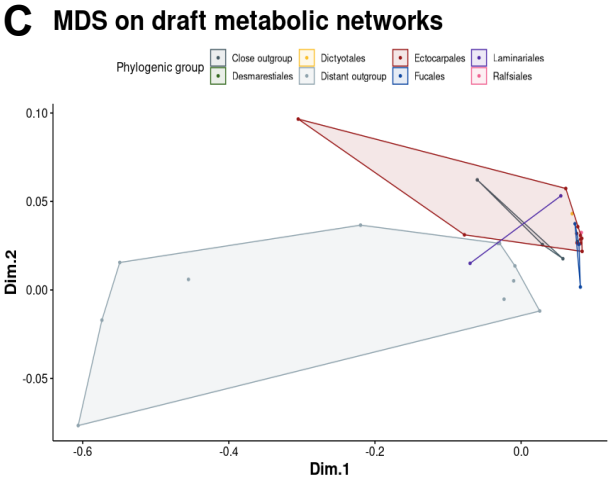
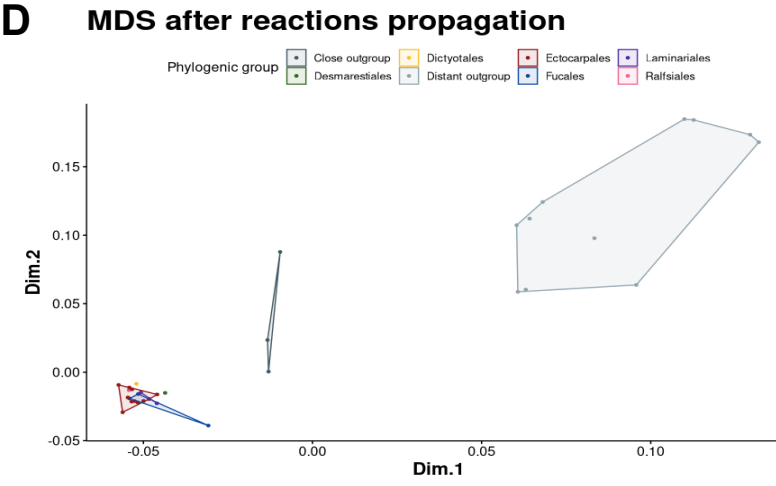


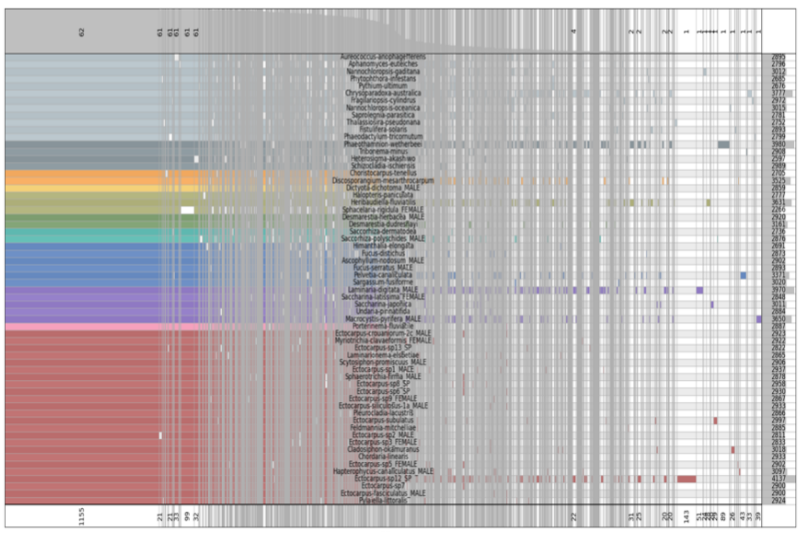
Figure S5

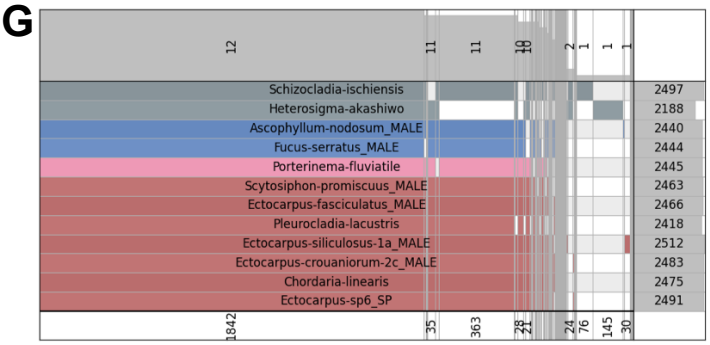












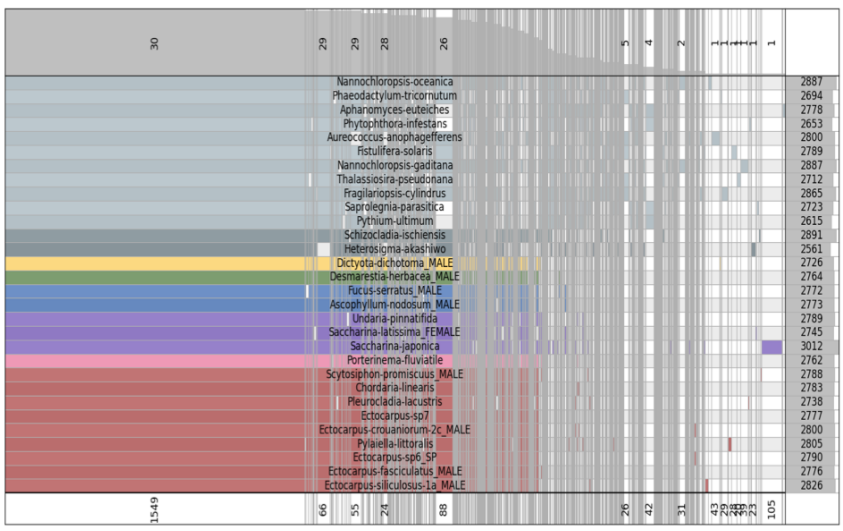
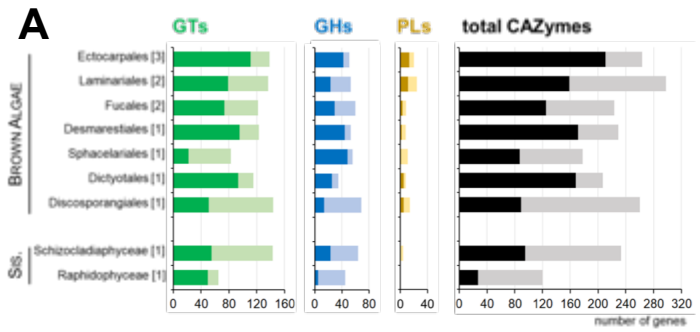
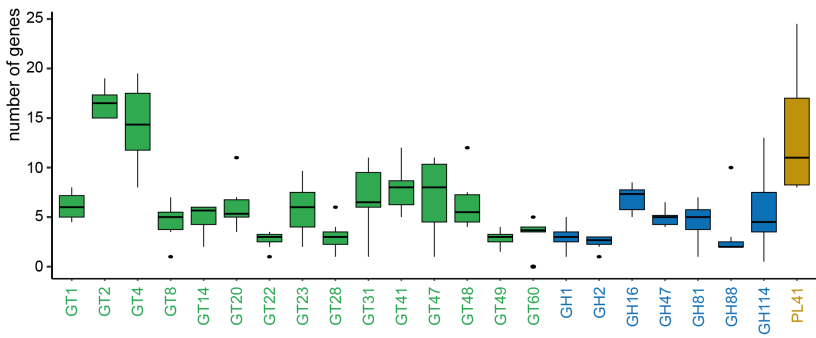
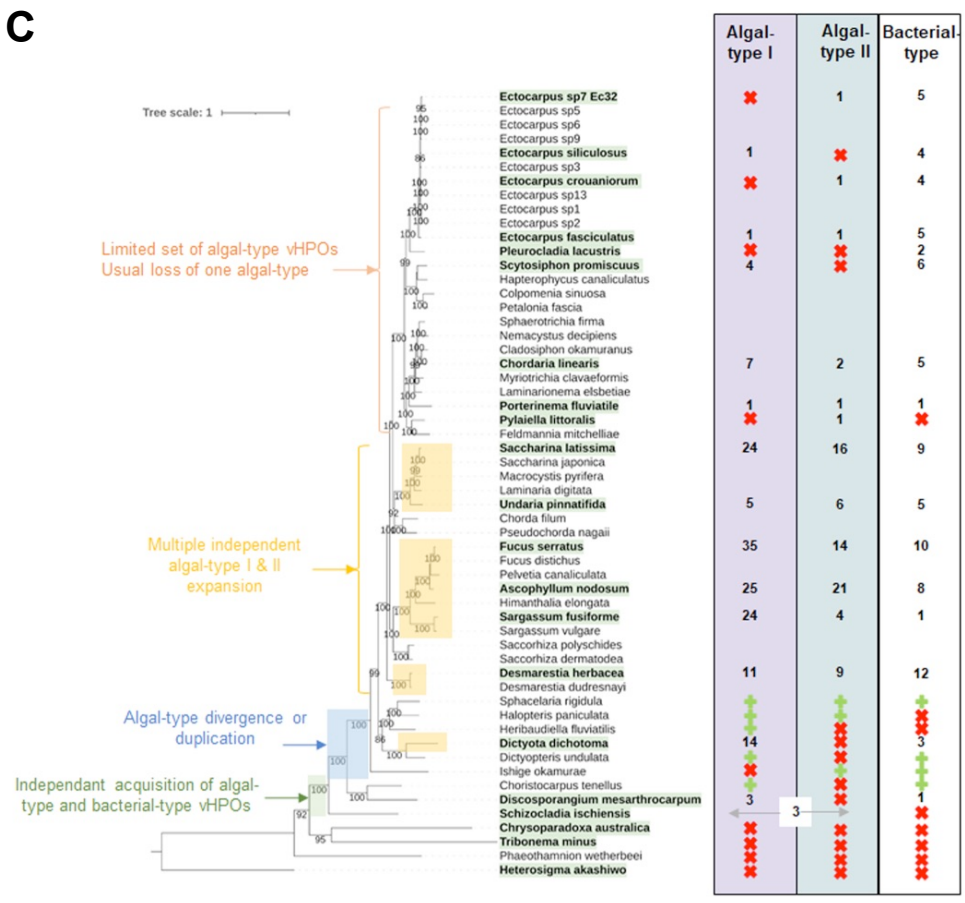
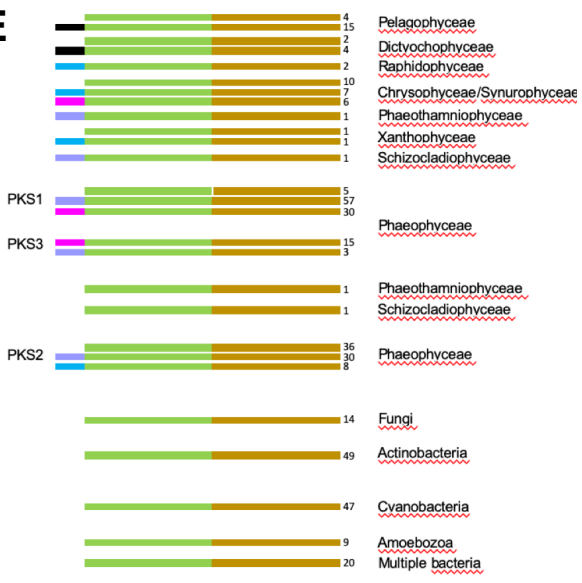


Figure S6


A
GTs
GHs
PLs
total CAZymes
BROWN ALGAE
Ectocarpales [3]
Laminariales [2]
Fucales [2]
Desmarestiales [1]
Sphacelariales [1]
Dictyotales [1]
Discosporangiales [1]
Sis.
Schizocladiaphyceae [1]
Raphidophyceae [1]
0 40 80 120 160
0 40 80
0 40
0 40 80 120 160 200 240 280 320
number of genes


B
number of genes
0 5 10 15 20 25
GT1 GT2 GT4 GT8 GT14 GT20 GT22 GT23 GT28 GT31 GT41 GT47 GT48 GT49 GT60 GH1 GH2 GH16 GH47 GH81 GH88 GH114 PL41


C
Tree scale: 1
Limited set of algal-type vHPOs
Usual loss of one algal-type
Multiple independent
algal-type I & II
expansion
Algal-type divergence or
duplication
Independant acquisition of algal-
type and bacterial-type vHPOs
Ectocarpus sp7 Ec32
Ectocarpus sp5
Ectocarpus sp6
Ectocarpus sp9
Ectocarpus siliculosus
Ectocarpus sp3
Ectocarpus crouaniorum
Ectocarpus sp13
Ectocarpus sp1
Ectocarpus sp2
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Hapterophycus canaliculatus
Colpomenia sinuosa
Petalonia fascia
Sphaerotrichia firma
Nemacystus decipiens
Cladosiphon okamuranus
Chordaria linearis
Myriotrichia clavaeformis
Laminarionema elsbetiae
Porterinema fluviatile
Pylaiella littoralis
Feldmannia mitchelliae
Saccharina latissima
Saccharina japonica
Macrocystis pyrifera
Laminaria digitata
Undaria pinnatifida
Chorda filum
Pseudochorda nagaii
Fucus serratus
Fucus distichus
Pelvetia canaliculata
Ascophyllum nodosum
Himanthalia elongata
Sargassum fusiforme
Sargassum vulgare
Saccorhiza polyschides
Saccorhiza dermatodea
Desmarestia herbacea
Desmarestia dudresnayi
Sphacelaria rigidula
Halopteris paniculata
Heribaudiella fluviatilis
Dictyota dichotoma
Dictyopteris undulata
Ishige okamurae
Choristocarpus tenellus
Discosporangium mesarthrocarpum
Schizocladia ischiensis
Chrysoparadoxa australica
Tribonema minus
Phaeothamnion wetherbeei
Heterosigma akashiwo
Algal-type I
Algal-type II
Bacterial-type


E
Pelagophyceae
Dictyochophyceae
Raphidophyceae
Chrysophyceae/Synurophyceae
Phaeothamniophyceae
Xanthophyceae
Schizocladiophyceae
PKS1
PKS3
Phaeophyceae
Phaeothamniophyceae
Schizocladiophyceae
PKS2
Phaeophyceae
Fungi
Actinobacteria
Cyanobacteria
Amoebozoa
Multiple bacteria

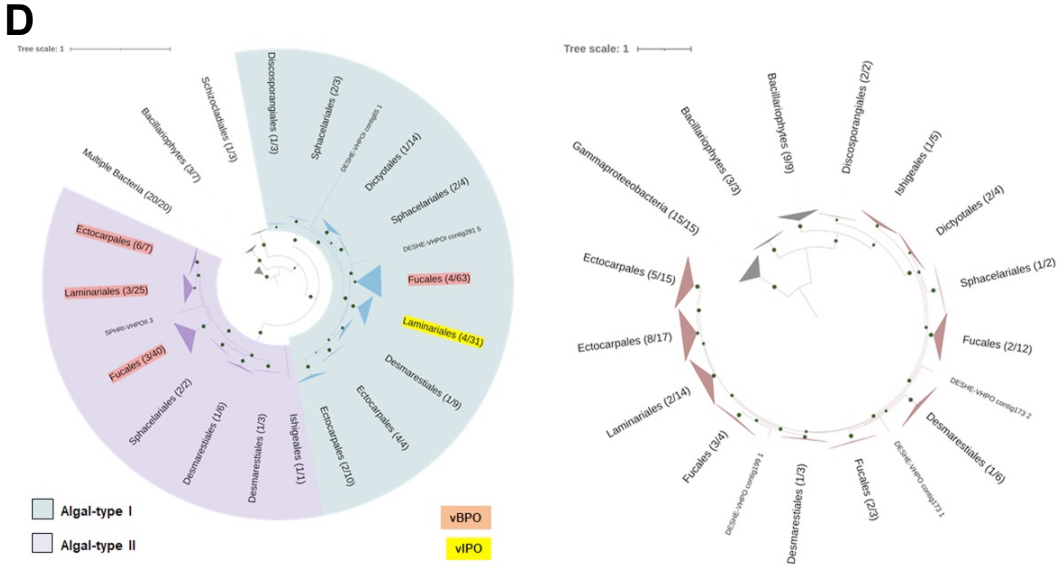

D
Tree scale: 1
Discosporangiales (1/3)
Sphacelariales (2/3)
Dictyotales (1/14)
Sphacelariales (2/4)
Fucales (4/63)
Laminariales (4/31)
Desmarestiales (1/9)
Ectocarpales (4/4)
Ectocarpales (2/10)
Ishigeales (1/1)
Desmarestiales (1/3)
Desmarestiales (1/6)
Sphacelariales (2/2)
Fucales (3/40)
Laminariales (3/25)
Ectocarpales (6/7)
Multiple Bacteria (20/20)
Bacillariophytes (3/7)
Schizocladiales (1/3)
Algal-type I
Algal-type II
vBPO
vIPO
Tree scale: 1
Bacillariophytes (9/9)
Discosporangiales (2/2)
Ishigeales (1/5)
Dictyotales (2/4)
Sphacelariales (1/2)
Fucales (2/12)
Desmarestiales (1/6)
Fucales (2/3)
Desmarestiales (1/3)
Fucales (3/4)
Laminariales (2/14)
Ectocarpales (8/17)
Ectocarpales (5/15)
Gammaproteeobacteria (15/15)
Bacillariophytes (3/3)

Figure S7

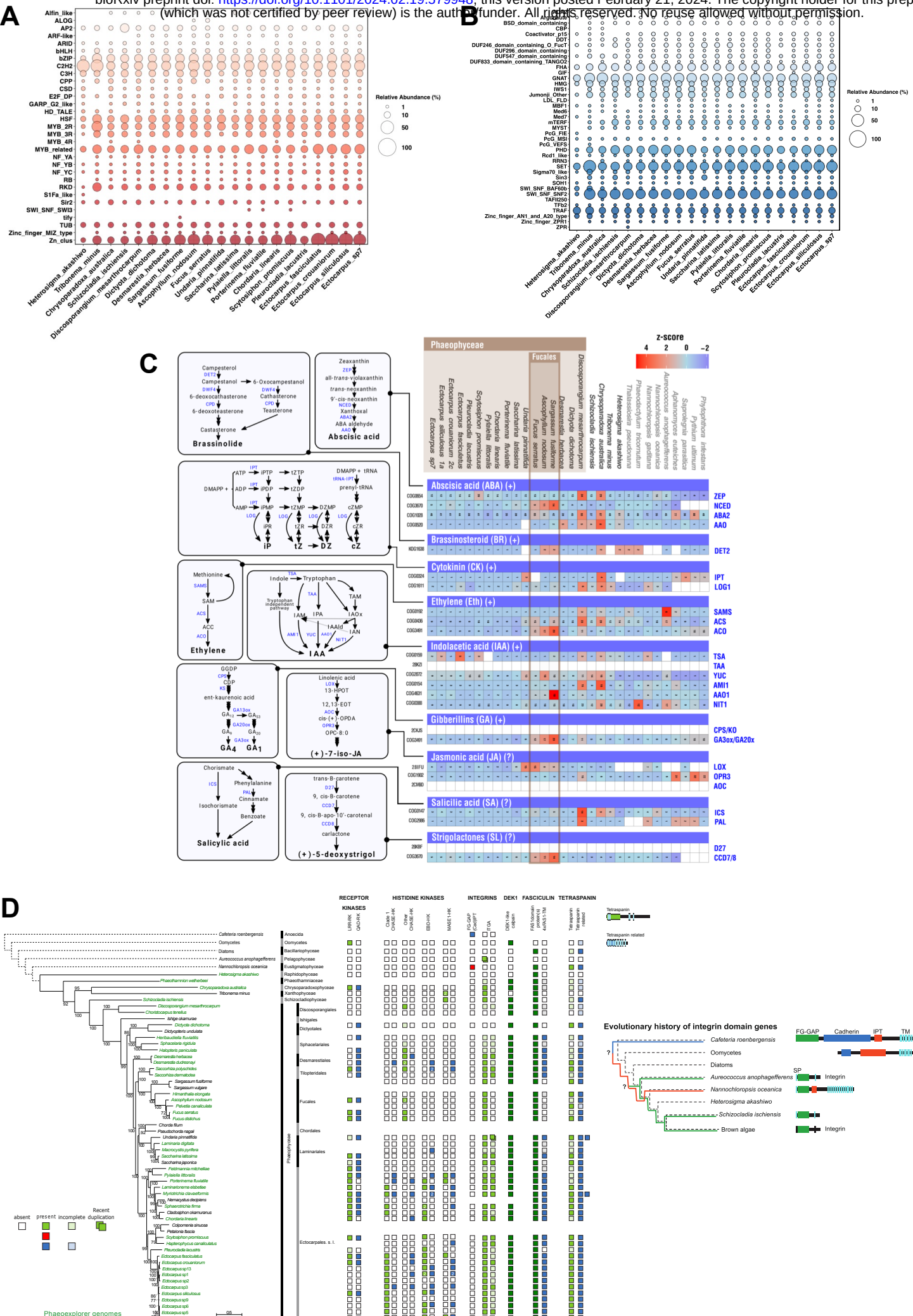


Figure S8

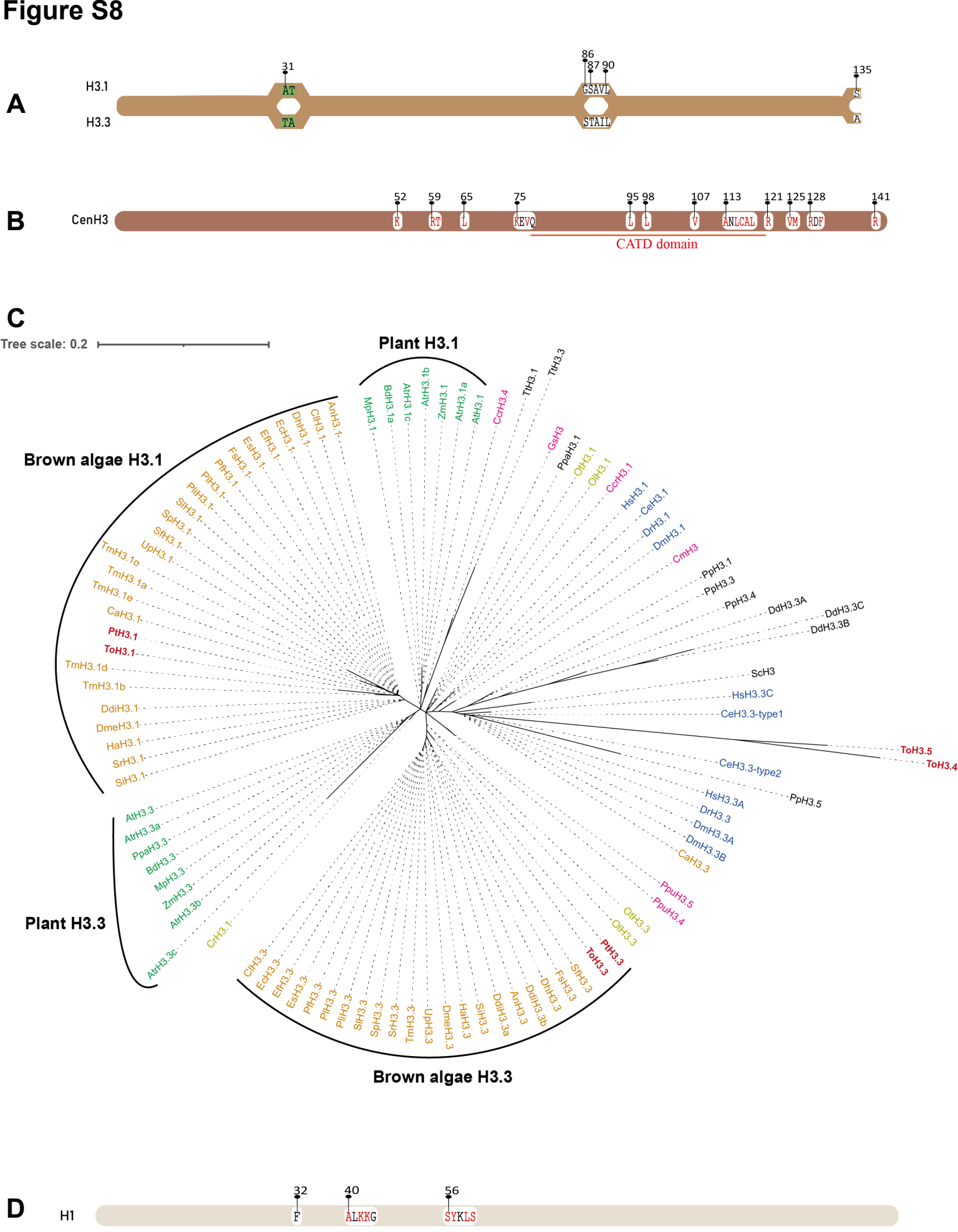
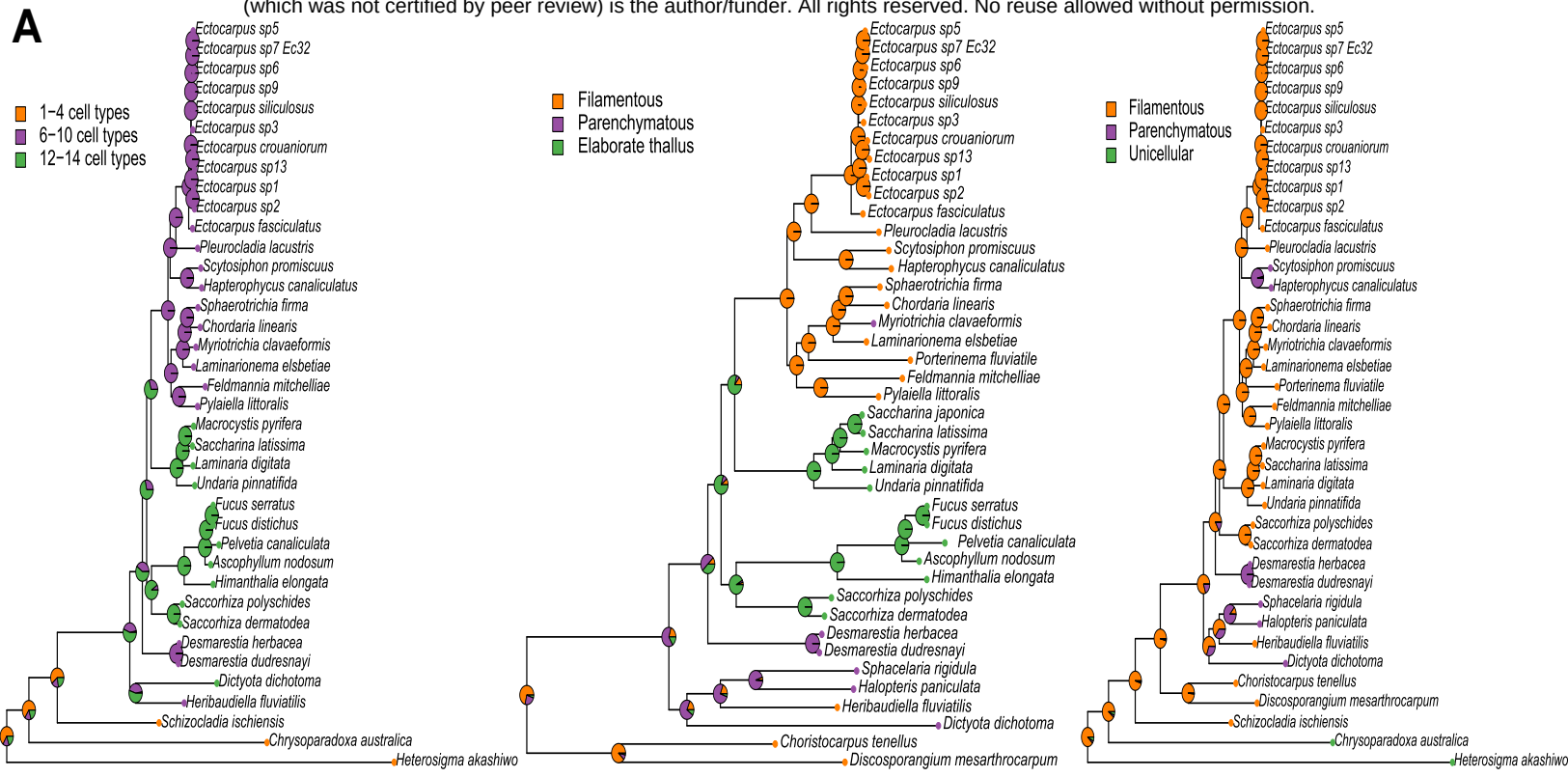
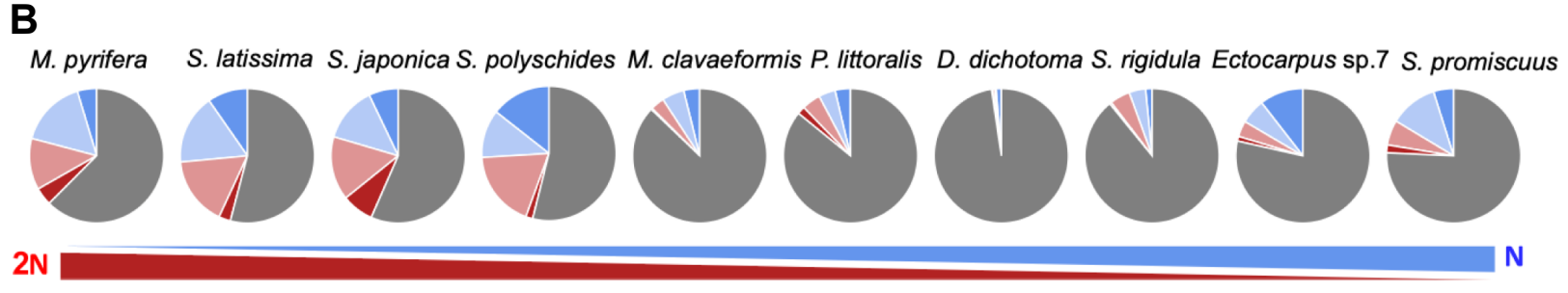

A
H3.1
H3.3
31
AT
TA
86
87 90
GSAVL
STAIL
135
S
A
B
CenH3
52
59
65
75
95 98
107
113
121 125 128
141
K
RT
L
KEVQ
L
L
V
ANLCAL
R
VM
RDF
R
CATD domain
C
Tree scale: 0.2
Plant H3.1
Brown algae H3.1
Plant H3.3
Brown algae H3.3
MpH3.1
BdH3.1a
AtrH3.1c
AtrH3.1b
ZmH3.1
AtrH3.1a
AtH3.1
CcrH3.4
TtH3.1
TtH3.3
GsH3
PpaH3.1
OtH3.1
OlH3.1
CcrH3.1
HsH3.1
CeH3.1
DrH3.1
DmH3.1
CmH3
PpH3.1
PpH3.3
PpH3.4
DdH3.3A
DdH3.3C
DdH3.3B
ScH3
HsH3.3C
CeH3.3-type1
ToH3.5
ToH3.4
CeH3.3-type2
PpH3.5
HsH3.3A
DrH3.3
DmH3.3A
DmH3.3B
CaH3.3
PpuH3.5
PpuH3.4
OtH3.3
OlH3.3
PtH3.3
ToH3.3
SfH3.3
FsH3.3
DhH3.3
DdiH3.3b
AnH3.3
DdiH3.3a
SlH3.3
HaH3.3
DmeH3.3
UpH3.3
TmH3.3
SrH3.3
SpH3.3
SlH3.3
PlH3.3
PfH3.3
EsH3.3
EfH3.3
EcH3.3
ClH3.3
CrH3.1
AtrH3.3c
AtrH3.3b
ZmH3.3
MpH3.3
BdH3.3
PpaH3.3
AtrH3.3a
AtH3.3
SiH3.1
SrH3.1
HaH3.1
DmeH3.1
DdiH3.1
TmH3.1b
TmH3.1d
ToH3.1
PtH3.1
CaH3.1
TmH3.1e
TmH3.1a
TmH3.1c
UpH3.1
SfH3.1
SpH3.1
SlH3.1
PlH3.1
PfH3.1
FsH3.1
EsH3.1
EfH3.1
EcH3.1
DhH3.1
ClH3.1
AnH3.1
D
H1
32
40
56
F
ALKKG
SYKLS

Figure S9





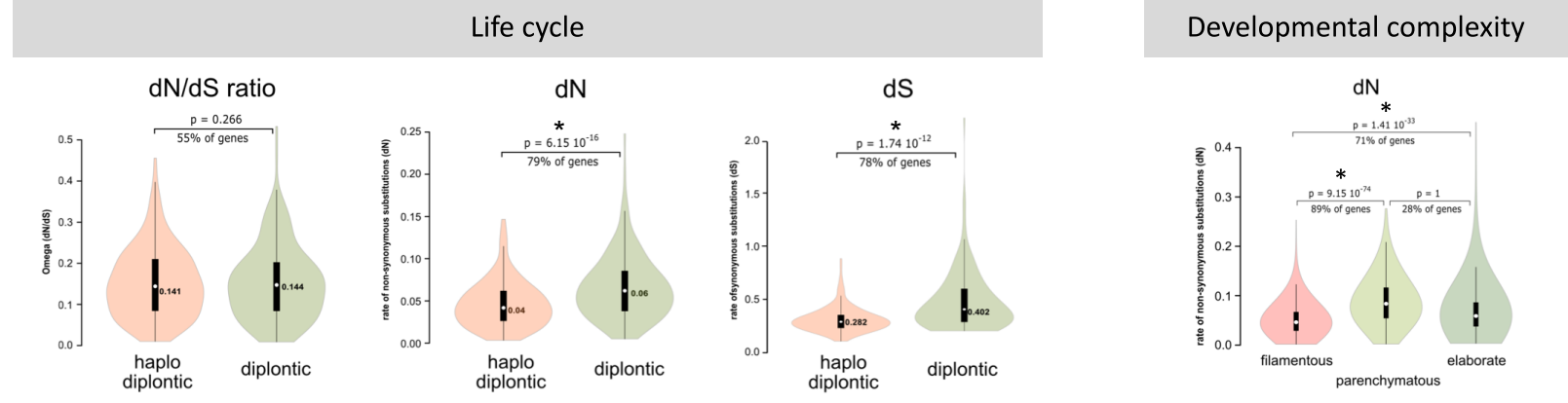


Figure S10

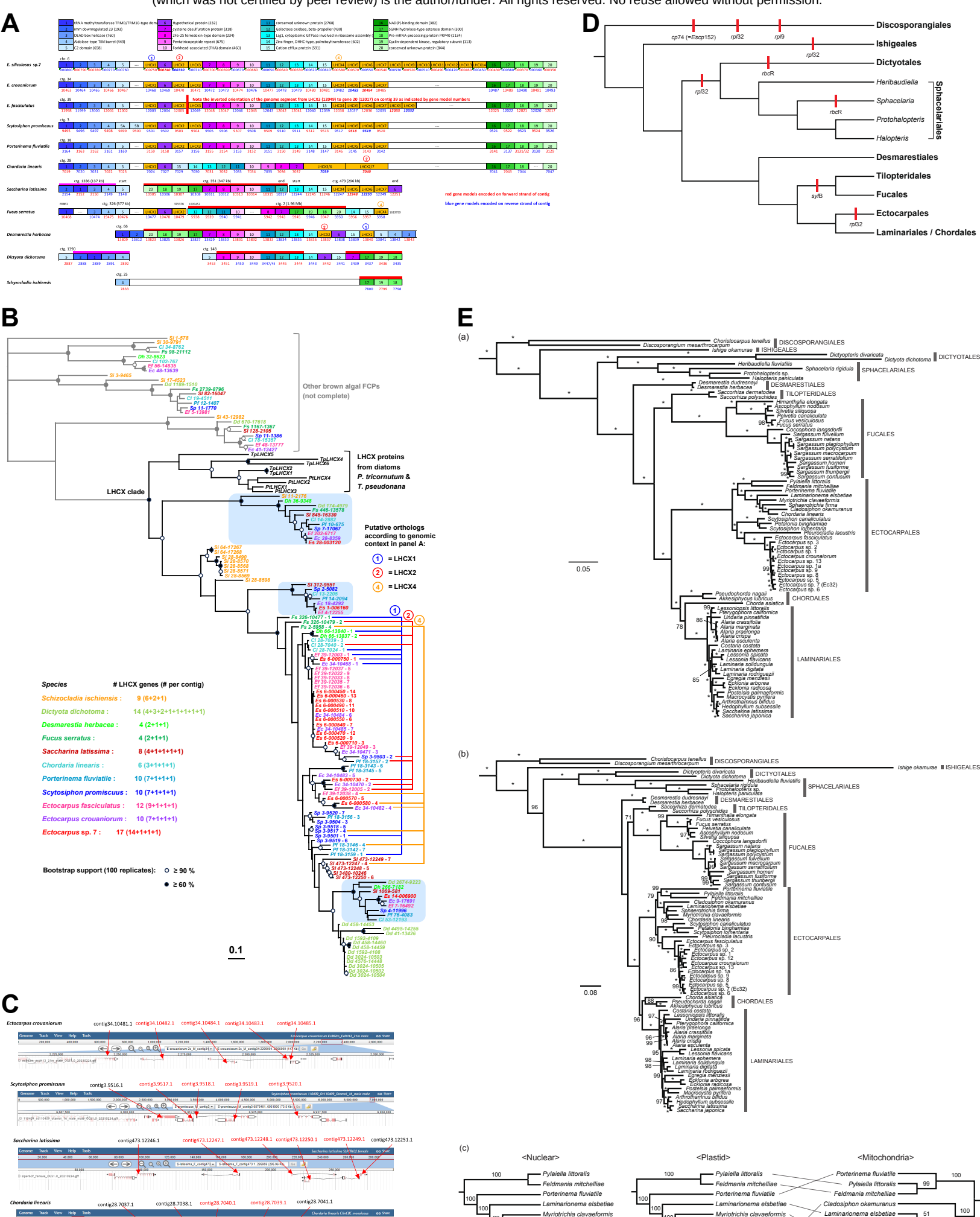


Figure S11

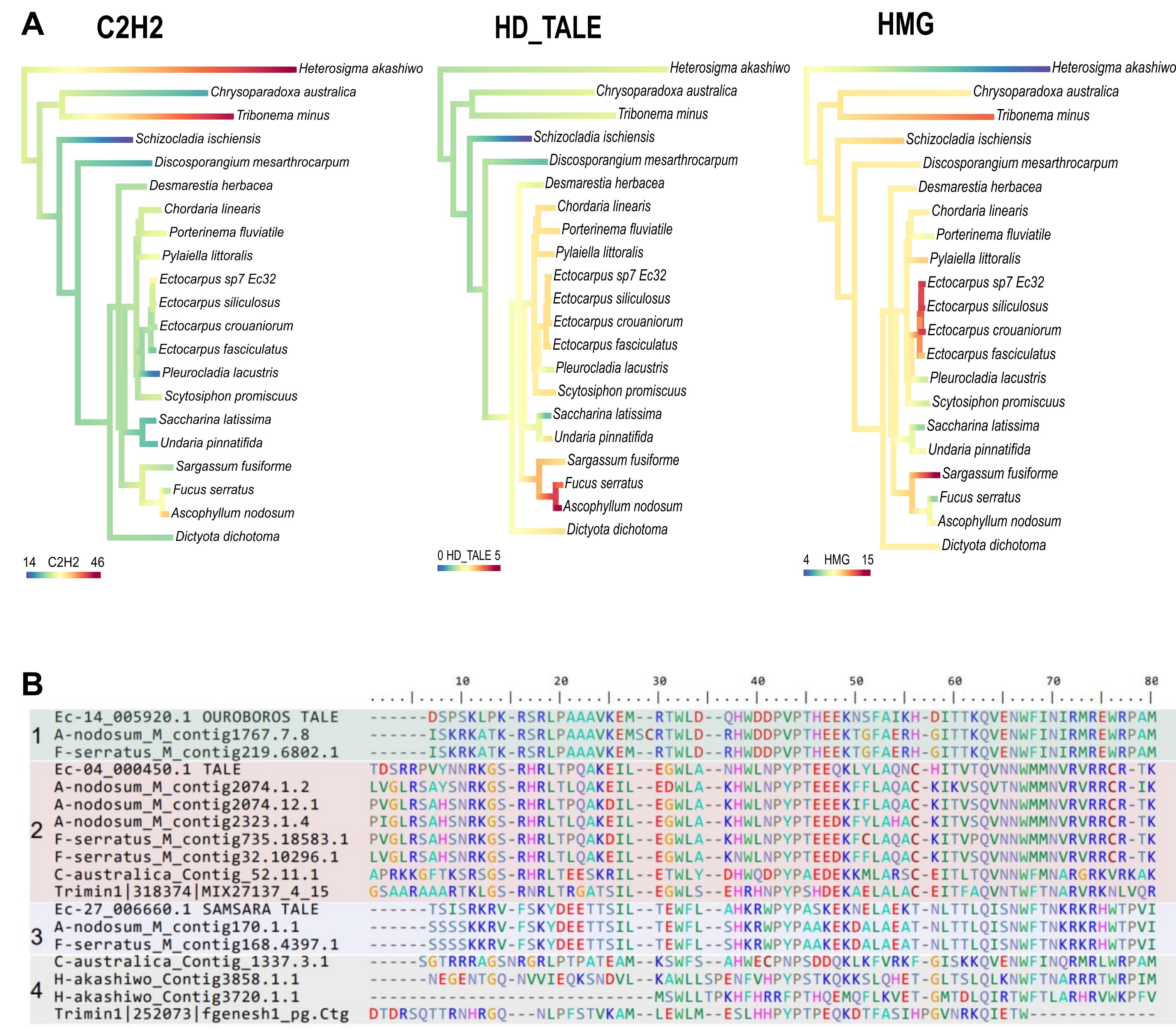
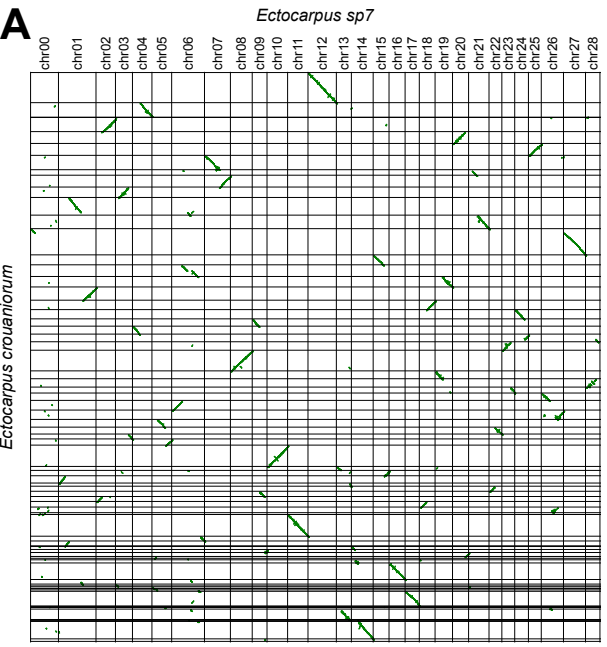
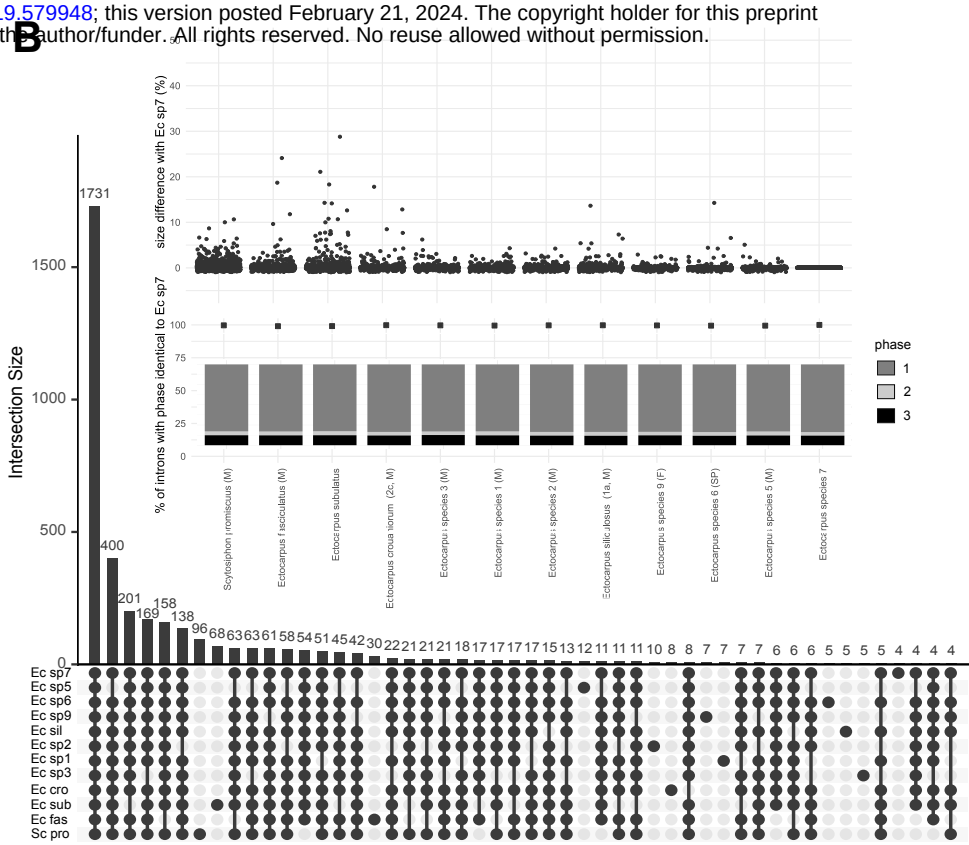
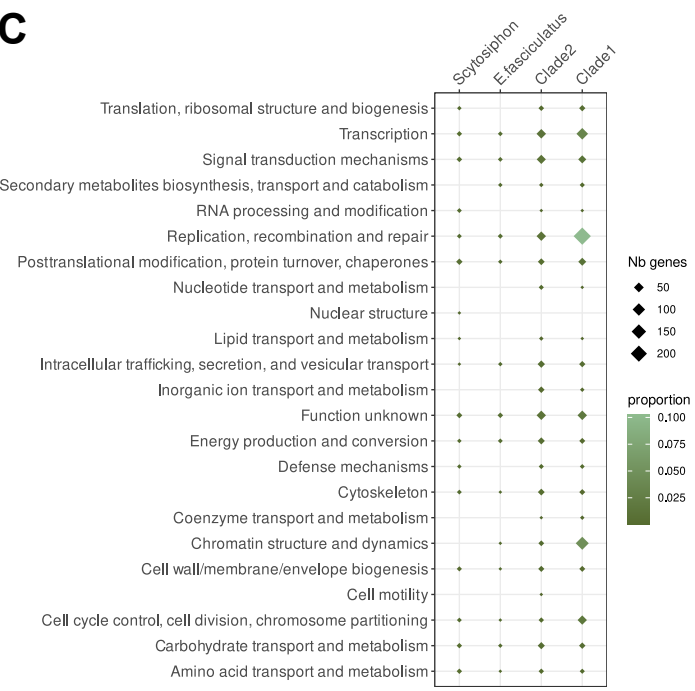
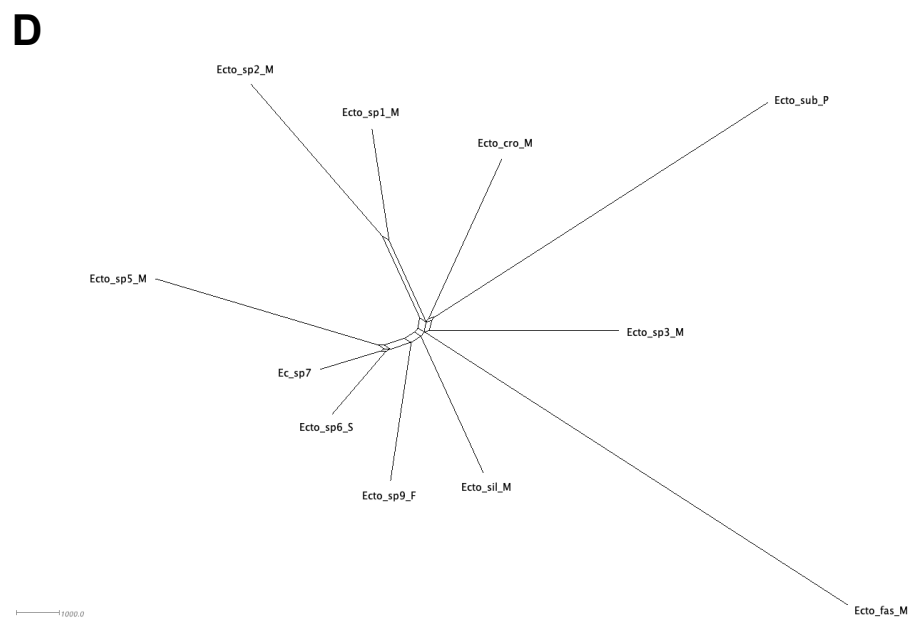

A
C2H2
HD_TALE
HMG
Heterosigma akashiwo
Chrysoparadoxa australica
Tribonema minus
Schizocladia ischiensis
Discosporangium mesarthrocarpum
Desmarestia herbacea
Chordaria linearis
Porterinema fluviatile
Pylaiella littoralis
Ectocarpus sp7 Ec32
Ectocarpus siliculosus
Ectocarpus crouaniorum
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Saccharina latissima
Undaria pinnatifida
Sargassum fusiforme
Fucus serratus
Ascophyllum nodosum
Dictyota dichotoma
14 C2H2 46
0 HD_TALE 5
4 HMG 15
B
10 20 30 40 50 60 70 80
1
Ec-14_005920.1 OUROBOROS TALE ------DSPSKLPK-RSRLPAAAVKEM--RTWLD--QHWDDPVPTHEEKNSFAIKH-DITTKQVENWFINIRMREWRPAM
A-nodosum_M_contig1767.7.8 ------ISKRKATK-RSRLPAAAVKEMSCRTWLD--RHWDDPVPTHEEKTGFAERH-GITTKQVENWFINIRMREWRPAM
F-serratus_M_contig219.6802.1 ------ISKRKATK-RSRLPAAAVKEM--RTWLD--RHWDDPVPTHEEKTGFAERH-GITTKQVENWFINIRMREWRPAM
2
Ec-04_000450.1 TALE TDSRRPVYNNRKGS-RHRLTPQAKEIL--EGWLA--NHWLNPYPTEEQKLYLAQNC-HITVTQVNNWMMNVRVRRCR-TK
A-nodosum_M_contig2074.1.2 LVGLRSAYSNRKGS-RHRLTLQAKEIL--EDWLA--KHWLNPYPTEEEKFFLAQAC-KIKVSQVTNWMMNVRVRRCR-IK
A-nodosum_M_contig2074.12.1 PVGLRSAHSNRKGS-RHRLTPQAKDIL--EGWLA--KHWLNPYPTEEEKIFLAQAC-KITVSQVNNWMMNVRVRRCR-TK
A-nodosum_M_contig2323.1.4 PIGLRSAHSNRKGS-RHRLTLQAKEIL--EGWLA--KHWLNPYPTEEDKFYLAHAC-KITVSQVNNWMMNVRVRRCR-TK
F-serratus_M_contig735.18583.1 PVGLRSAHSNRKGS-RHRLTPQAKDIL--EGWLA--KHWLNPYPTEEEKFCLAQAC-KITVPQVNNWMMNVRVRRCR-TK
F-serratus_M_contig32.10296.1 LVGLRSAHSNRKGS-RHRLTLQAKEIL--EGWLA--KNWLNPYPTEEDKFFLAQAC-KITVSQVNNWMMNVRVRRCR-TK
C-australica_Contig_52.11.1 APRKKGFTKSRSGS-RHRLTEESKRIL--ETWLY--DHWQDPYPAEDEKKMLARSC-EITLTQVNNWFMNARGRKVRKAK
Trimin1|318374|MIX27137_4_15 GSAARAAARTKLGS-RNRLTRGATSIL--EGWLS--EHRHNPYPSHDEKAELALAC-EITFAQVNTWFTNARVRKNLVQR
3
Ec-27_006660.1 SAMSARA TALE ------TSISRKRV-FSKYDEETTSIL--TEWFL--AHKRWPYPASKEKNELAEKT-NLTTLQISNWFTNKRKRHWTPVI
A-nodosum_M_contig170.1.1 ------SSSSKKRV-FSKYDEETTSIL--TEWFL--SHKRWPYPAAKEKDALAEAT-NLTTLQISNWFTNKRKRHWTPVI
F-serratus_M_contig168.4397.1 ------SSSSKKRV-FSKYDEETTSIL--TEWFL--SHKRWPYPAAKEKDALAEAT-NLTTLQISNWFTNKRKRHWTPVI
4
C-australica_Contig_1337.3.1 -----SGTRRRAGSNRGRLPTPATEAM--KSWFS--AHWECPNPSDDQKLKFVRKF-GISKKQVENWFINQRMRLWRPAM
H-akashiwo_Contig3858.1.1 ------NEGENTGQ-NVVIEQKSNDVL--KAWLLSPENFVHPYPSTKQKKSLQHET-GLTSLQLKNWFTNARRRTWRPIM
H-akashiwo_Contig3720.1.1 --------------------------MSWLLTPKHFHRRFPTHQEMQFLKVET-GMTDLQIRTWFTLARHRVWKPFV
Trimin1|252073|fgenesh1_pg.Ctg DTDRSQTTRNHRGQ---NLPFSTVKAM--LEWLM--ESLHHPYPTPEQKDTFASIHPGVNRKQIETW-----------

Figure S12









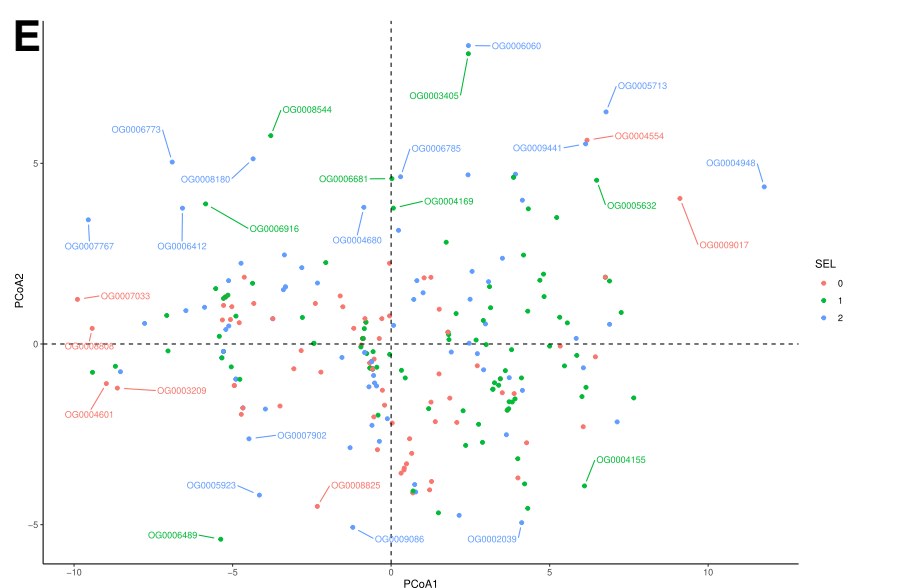


Figure S13