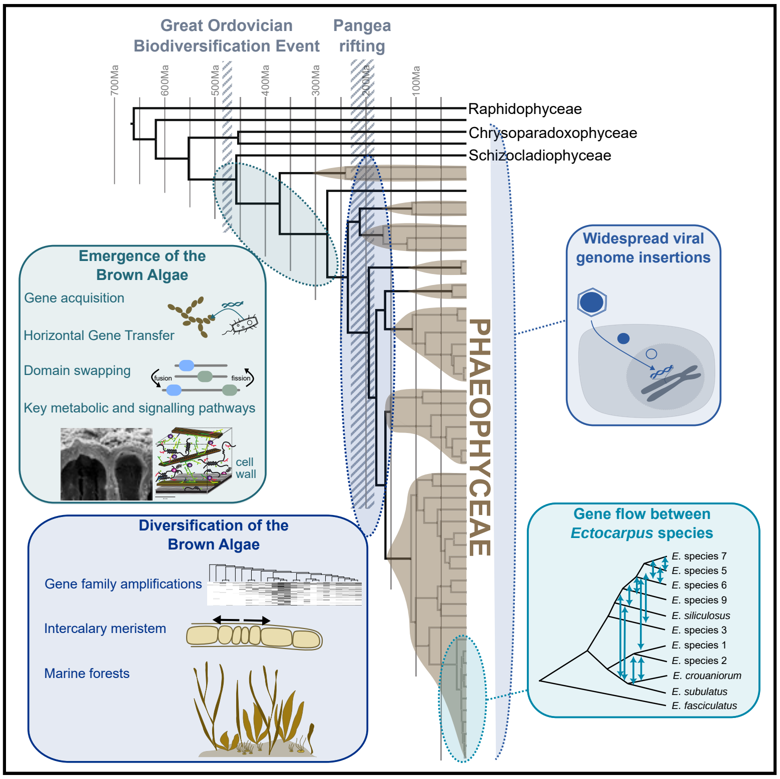
Cell

Article

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

Graphical abstract



Highlights

- An intense period of genome evolution during early emergence of the brown algae
- Gene family amplifications linked to diversification of the brown algae
- Extensive gene flow between species at the genus level in *Ectocarpus*
- Insertions of diverse *Phaeovirus* genomes are widespread in brown algae

Authors

France Denoeud, Olivier Godfroy,
Corinne Cruaud, ..., Patrick Wincker,
Jean-Marc Aury, J. Mark Cock

Correspondence

kawai@kobe-u.ac.jp (H.K.),
akirapeters@gmail.com (A.F.P.),
hsyoon2011@skku.edu (H.S.Y.),
cherve@sb-roscoff.fr (C.H.),
yenh@ysfri.ac.cn (N.Y.),
epbapteste@gmail.com (E.B.),
valero@sb-roscoff.fr (M.V.),
gabriel.markov@sb-roscoff.fr (G.V.M.),
corre@sb-roscoff.fr (E.C.),
susana.coelho@tuebingen.mpg.de
(S.M.C.),
pwincker@genoscope.cns.fr (P.W.),
jmaury@genoscope.cns.fr (J.-M.A.),
cock@sb-roscoff.fr (J.M.C.)

In brief

Comparative genomics charts the evolutionary history of the brown algal lineage, identifying an early period of accelerated genome evolution followed by diversification of the major orders, and a major impact of continuous, widespread viral genome integration.

Denoeud et al., 2024, Cell *187*, 6943–6965
November 27, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2024.10.049

CellPress

Cell

CellPress
OPEN ACCESS

Article

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

France Denoeud,[1,60] Olivier Godfroy,[2,60] Corinne Cruaud,[3,61] Svenja Heesch,[4,51,61] Zofia Nehr,[4,61] Nachida Tadrent,[1,52,61] Arnaud Couloux,[1,61] Loraine Brillet-Guéguen,[5,6,61] Ludovic Delage,[7,61] Dean Mckeown,[6,61] Taizo Motomura,[8,62] Duncan Sussfeld,[1,9,62] Xiao Fan,[10,11,62] Lisa Mazéas,[2,62] Nicolas Terrapon,[12,13,62] Josué Barrera-Redondo,[14,62] Romy Petroll,[14,62] Lauric Reynes,[15,62] Seok-Wan Choi,[16,62] Jihoon Jo,[16,62] Kavitha Uthanumallian,[17,62] Kenny Bogaert,[18,53,62] Céline Duc,[19,62] Pélagie Ratchinski,[4,62] Agnieszka Lipinska,[4,14,62] Benjamin Noel,[1,62] Eleanor A. Murphy,[20,21,62] Martin Lohr,[22,62] Ananya Khatei,[23,54,62] Pauline Hamon-Giraud,[24,62] Christophe Vieira,[25,62] Komlan Avia,[26,62] Svea Sanja Akerfors,[22] Shingo Akita,[27] Yacine Badis,[4] Tristan Barbeyron,[2] Arnaud Belcour,[24,55] 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]

(Author list continued on next page)

[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, Paris, France
[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 University, 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, Parkville, VIC, Australia
[18]Phycology Research Group, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium
[19]Nantes Université, CNRS, US2B, UMR 6286, 44000 Nantes, France
[20]University of Bristol, Bristol, UK
[21]Marine Biological Association, Plymouth, UK
[22]Johannes Gutenberg University, Mainz, Germany
[23]Algal and Microbial Biotechnology Division, Nord University, Bodø, Norway

(Affiliations continued on next page)

SUMMARY

Brown seaweeds are keystone species of coastal ecosystems, often forming extensive underwater forests, and are under considerable threat from climate change. In this study, analysis of multiple genomes has provided insights across the entire evolutionary history of this lineage, from initial emergence, through later diversification of the brown algal orders, down to microevolutionary events at the genus level. Emergence of the brown algal lineage was associated with a marked gain of new orthologous gene families, enhanced protein domain rearrangement, increased horizontal gene transfer events, and the acquisition of novel signaling molecules and key metabolic pathways, the latter notably related to biosynthesis of the alginate-based extracellular matrix, and halogen and phlorotannin biosynthesis. We show that brown algal genome diversification is tightly linked to phenotypic divergence, including changes in life cycle strategy and zoid flagellar structure. The study also showed that integration of large viral genomes has had a significant impact on brown algal genome content throughout the emergence of the lineage.

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,[15,34,35] Lars Harms,[36] William John Hatchett,[37] Bernard Henrissat,[38] Galice Hoarau,[37] Chloé Jollivet,[2] Alexander Jueterbock,[23] Ehsan Kayal,[6,56] Andrew H. Knoll,[39] Kazuhiro Kogame,[40] Arthur Le Bars,[6,41] Catherine Leblanc,[7] Line Le Gall,[9] Ronja Ley,[22] Xi Liu,[6] Steven T. LoDuca,[42] Pascal Jean Lopez,[43] Philippe Lopez,[9] Eric Manirakiza,[19] Karine Massau,[6] Stéphane Mauger,[15,57] Laetitia Mest,[4,58] Gurvan Michel,[2] Catia Monteiro,[7] Chikako Nagasato,[8] Delphine Nègre,[6,59] Eric Pelletier,[1] Naomi Phillips,[44] Philippe Potin,[7] Stefan A. Rensing,[45] Ellyn Rousselot,[19] Sylvie Rousvoal,[7] Declan Schroeder,[46] Delphine Scornet,[4] Anne Siegel,[24] Leila Tirichine,[19] Thierry Tonon,[47] Klaus Valentin,[36] Heroen Verbruggen,[28] Florian Weinberger,[48] Glen Wheeler,[21] Hiroshi Kawai,[49,63,*] Akira F. Peters,[50,63,*] Hwan Su Yoon,[16,63,*] Cécile Hervé,[2,63,*] Naihao Ye,[10,11,63,*] Eric Bapteste,[9,63,*] Myriam Valero,[15,63,*] Gabriel V. Markov,[7,63,*] Erwan Corre,[6,63,*] Susana M. Coelho,[14,63,*] Patrick Wincker,[1,63,*] Jean-Marc Aury,[1,63,*] and J. Mark Cock[4,64,*]

[24]University of Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France
[25]Research Institute for Basic Sciences, Jeju National University, Jeju 63243, Republic of Korea
[26]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France
[27]Faculty of Fisheries Sciences, Hokkaido University, Minato-cho 3-1-1, Hakodate, Hokkaido 041-8611, Japan
[28]University of Melbourne, Parkville, VIC, Australia
[29]University of Washington, Seattle, WA, USA
[30]Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, 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]Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile
[36]Alfred Wegener Institute (AWI), Bremenhaven, Germany
[37]Nord University, Bodø, Norway
[38]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs Lyngby, Denmark
[39]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[40]Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan
[41]CNRS, Institut Français de Bioinformatique, IFB-core, Évry, France
[42]Department of Geography and Geology, Eastern Michigan University, Ypsilanti, MI 48197, USA
[43]Centre National de la Recherche Scientifique, UMR BOREA MNHN/CNRS-8067/SU/IRD/Université de Caen Normandie/Université des Antilles, Plouzané, France
[44]Biology Department, Arcadia University, Glenside, PA, USA
[45]University of Freiburg, Freiburg im Breisgau, Germany
[46]University of Minnesota, St. Paul, MN, USA
[47]Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, UK
[48]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
[49]Kobe University Research Center for Inland Seas, Kobe, Japan
[50]Bezhin Rosko, 29250 Santec, France
[51]Present address: Applied Ecology & Phycology, Institute for Biosciences, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany
[52]Present address: Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université de Tours, Tours 37200, France
[53]Present address: Department of Algal Development and Evolution, Max Planck Institute for Biology, Tübingen 72076, Germany
[54]Present address: ICAR-Directorate of Coldwater Fisheries Research, Bhimtal, India
[55]Present address: Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
[56]Present address: Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA
[57]Present address: CNRS, La Rochelle Université, UMR7266, Littoral Environnement et Sociétés, La Rochelle, France

(Affiliations continued on next page)

INTRODUCTION

The brown algae (Phaeophyceae) are a lineage of complex multicellular organisms that emerged about 450 mya[1] from within a group of photosynthetic stramenopile taxa (derived from a secondary endosymbiosis involving a red alga[2]) that are either unicellular or have very simple filamentous multicellular thalli (Figure 1). The emerging brown algae acquired a number of characteristic features that are thought to have contributed to the evolutionary success of this lineage, including complex polysaccharide-based cell walls that confer protection and flexibility in the highly dynamic intertidal environment,[3] complex halogen[4] and phlorotannin[5] metabolisms that are thought to play important roles in multiple processes including defense, adhesion and cell-wall modification, and a remarkable diversity of life cycles and developmental body architectures adapted to diverse marine environments.[6] As a result of these attributes, many brown algae have become

[58]Present address: Vegenov, Saint Pol de Léon, France
[59]Present address: Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, Nantes, France
[60]These authors contributed equally
[61]These authors contributed equally
[62]These authors contributed equally
[63]Senior author
[64]Lead contact
*Correspondence: kawai@kobe-u.ac.jp (H.K.), akirapeters@gmail.com (A.F.P.), hsyoon2011@skku.edu (H.S.Y.), cherve@sb-roscoff.fr (C.H.), yenh@ysfri.ac.cn (N.Y.), epbapteste@gmail.com (E.B.), valero@sb-roscoff.fr (M.V.), gabriel.markov@sb-roscoff.fr (G.V.M.), corre@sb-roscoff.fr (E.C.), susana.coelho@tuebingen.mpg.de (S.M.C.), pwincker@genoscope.cns.fr (P.W.), jmaury@genoscope.cns.fr (J.-M.A.), cock@sb-roscoff.fr (J.M.C.)

https://doi.org/10.1016/j.cell.2024.10.049

established as key components of extensive coastal ecosystems. These seaweed-based ecosystems provide high value Earth-system-scale services, including the sequestration of several megatons of carbon per year globally, comparable to values reported for terrestrial forests,[7] but this important role of seaweed ecosystems is threatened by climate-related declines in seaweed populations worldwide.[8] 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 play a significant role in mitigating the effects of climate change.[9] 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 (https://phaeoexplorer.sb-roscoff.fr/) has generated a large dataset of genome sequences, spanning all the major orders of the Phaeophyceae.[10] This extensive genomic dataset has been analyzed here to study the origin and evolution of key genomic features during the emergence and diversification of this important group of marine organisms.

RESULTS

In-depth sequencing of brown algal genomes

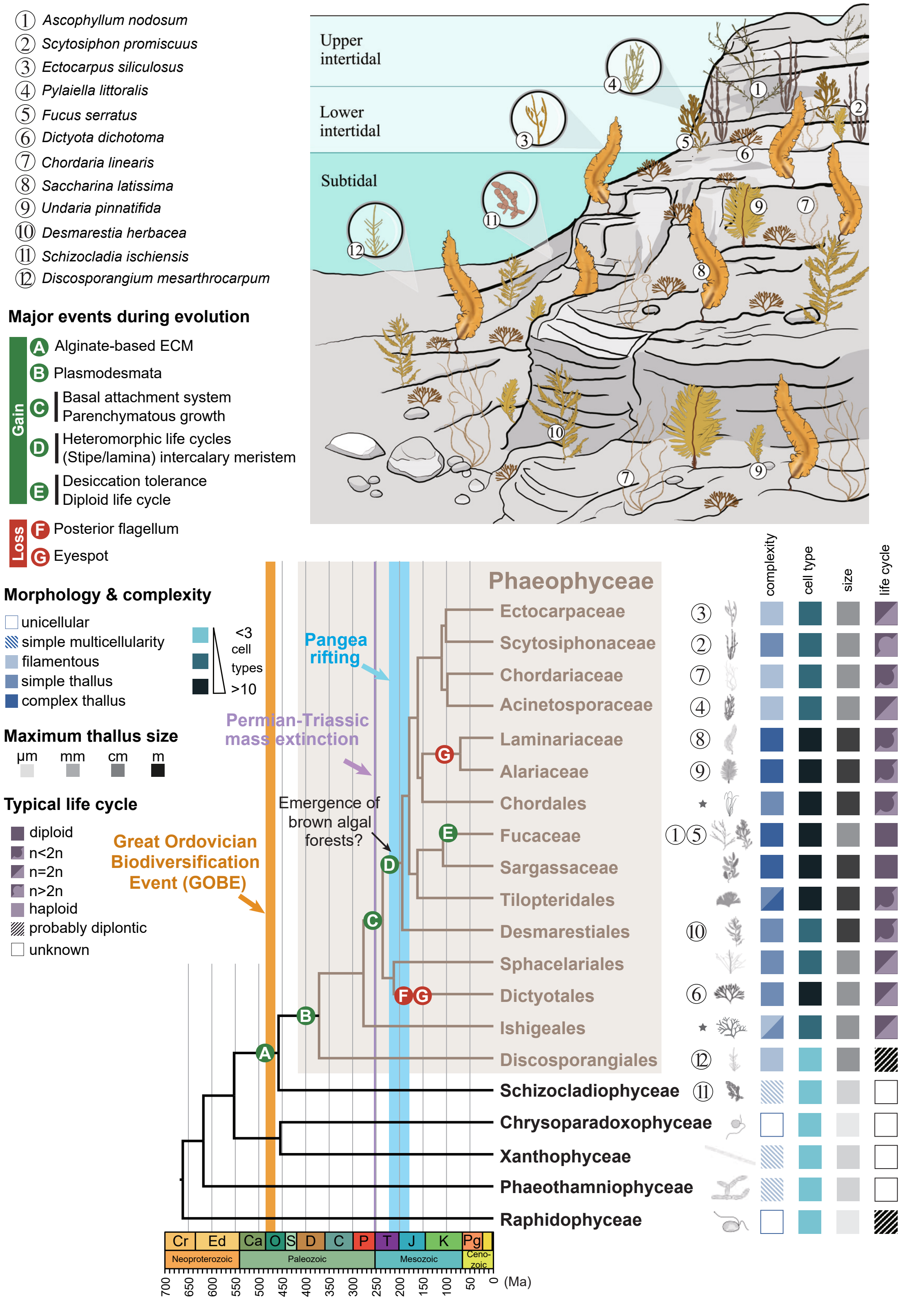
Until now, good quality genome assemblies have been obtained for only five brown algal species,[11–15] together with about 46 draft genome assemblies.[16–20] Here, we report work that has significantly expanded the genomic data available by sequencing and assembling 17 good quality genomes using long-read technology (Table S1), plus an additional 43 draft genome assemblies. These 60 genomes correspond to 40 brown algae and four closely related species, covering 16 Phaeophyceae families providing a dense coverage of this lineage (Figures S1A and S1B; Table S1A). 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, biogeography and life cycle structure (Figures 1, S1C, and S1D). The analyses carried out in this study have focused principally on a set of 21 good quality reference genomes, which include four previously published genomes (Table S1B).

Marked changes in genome content and gene structure during the emergence of the Phaeophyceae lineage

Recent evidence indicates that the brown algae emerged about 450 mya during the Great Ordovician Biodiversification Event (GOBE),[1] a conclusion that is supported by a fossil-calibrated tree built with a nuclear-gene-based phylogeny constructed using the Phaeoexplorer genome data (Figures 1 and S2A). An increase in atmospheric oxygen at the time of the GOBE, which coincided with the emergence of herbivorous marine invertebrates,[21] is likely to have created conditions conducive to the observed transition toward increased multicellular complexity during early brown algal evolution.

To investigate genomic modifications associated with the emergence and diversification of the brown algae, we first carried out a series of genome-wide analyses aimed at identifying broad trends in genome evolution over evolutionary time (Figure 2). 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 2B and S2). Similarly, a phylostratigraphy analysis indicated that 29.6% of brown algal genes cannot be traced back to ancestors outside the Phaeophyceae, with the majority of gene founder events occurring early during the emergence of the brown algae (Figures 2E and S3A; Table S2), again indicating a burst of gene birth during the emergence of this lineage. Both the Dollo analysis and the phylostratigraphy approach indicated that the gene families acquired during early brown algal evolution were significantly enriched in genes that could not be assigned to a cluster of orthologous genes (COG) category, suggesting a burst in the acquisition of genetic novelty (Figure 2G).

One of the factors underlying the marked burst of gene gain during the emergence of the brown algae was an increase in the rate of acquisition of new genes via horizontal gene transfer (HGT). A phylogeny-based search for genes potentially derived from HGT events indicated that they constitute about 1% of brown algal gene catalogs and that the novel genes were principally acquired from bacterial genomes (Figures 2F and S3B). The proportion of class-specific HGT events compared with more ancient HGT events was greater for the brown algae (33.5% of HGT events) than for the closely related taxa Xanthophyceae (*Tribonema minus*) and Raphidophyceae (*Heterosigma akashiwo*; mean of 17.1% for the two taxa, Wilcoxon $p = 0.021$), indicating that higher levels of HGT occurred during the emergence of the brown algae than in closely related taxa (Figure 2F).



(legend on next page)

The marked increase in the rate of gene gain appears to have been a key factor in the emergence of the brown algal lineage but this was not the only process that enriched brown algal genomes during this period. Domain fusions and fissions (composite genes) were prevalent during the early stages of brown algal emergence (Figures 2D and S3C), affecting about 7% of brown algal gene complements. In contrast, gene family amplifications were most prevalent at a later stage of brown algal evolution, corresponding to the diversification of the major brown algal orders during the Mesozoic (Figures 2C, S3D, and S3E; Table S3). However, the amplified gene families were significantly enriched in genes that had been gained during the emergence of the brown algae (χ^2 p = 1.04e−15; Table S1C), indicating that gene gain during the early evolution of the lineage nonetheless played a crucial role by establishing the majority of the gene families that would later undergo amplifications.

Analysis of the predicted functions of the three sets of gene families identified as having been amplified, derived from domain fusions/fissions or derived from an HGT event (Figure 2G) indicated that they were enriched in several functional categories, notably carbohydrate metabolism, signal transduction, and transcription. These functional categories may have been important in the emergence of the complex brown algal cell wall or correspond to a complexification of signaling pathways as multicellular complexity increased. Interestingly, many of the genes acquired at, or shortly after, the origin of the Phaeophyceae encode secreted or membrane proteins (Figure S3A), suggesting roles in cell-cell communication that may have been important for the emergence of complex multicellularity or as components of defense mechanisms. The acquisition of plasmodesmata by brown algae directly after their divergence from their sister taxon *Schizocladia ischiensis*[22] (Figure 1) underlines the importance of cell-cell communication from the outset of brown algal evolution.

The emergence of the brown algae also corresponded with changes in gene structure. On average, brown algal genes tend to be more intron-rich than those of the other stramenopile groups,[23] including closely related taxa (Figure S4A), with the notable exception of *Chrysoparadoxa australica* (Figure S4A). A comparison of orthologous genes indicated a phase of rapid intron acquisition just before the divergence of the Phaeophyceae and the Schizocladiophyceae, followed by a period of relative intron stability up to the present day (Figure S4B). This phase of accelerated intron acquisition coincided approximately with the periods of marked gene gain and domain reorganization discussed above and may have been an indirect consequence of increased multicellular complexity (Figure 1) due to a concomitant decrease in effective population size.[24] Once established, increased intron density may have facilitated some of the genome-wide tendencies described above, such as increased reorganization of composite genes, for example, and thereby played an important role in a context of increasing developmental complexity.[25–28]

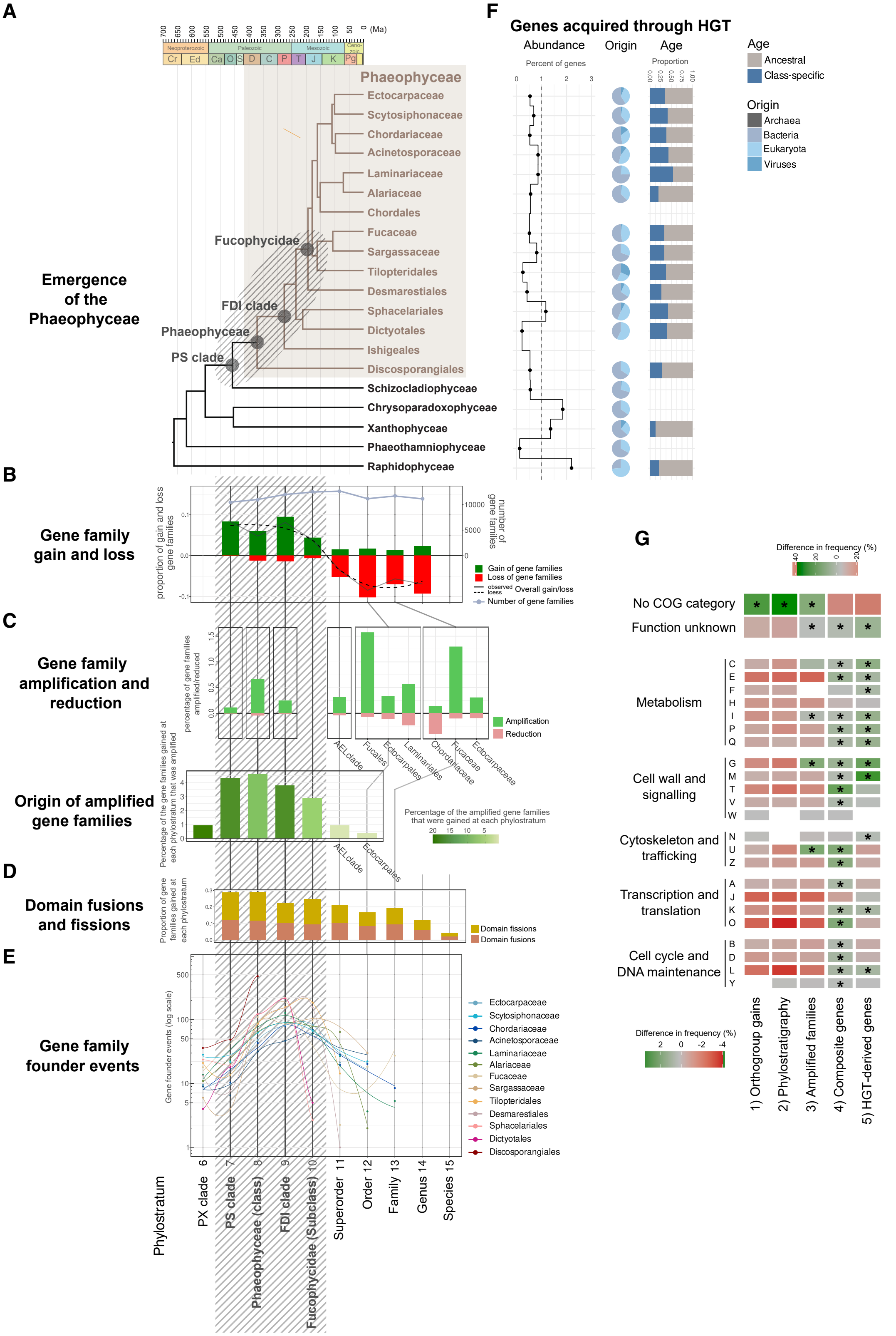
Acquisition of key metabolic and signaling pathways during the emergence of the Phaeophyceae

The success of the brown algae as an evolutionary lineage has been attributed, at least in part, to the acquisition of several key metabolic pathways, particularly those associated with cell-wall biosynthesis, and both halogen and phlorotannin metabolism.[3–5] Large complements of carbohydrate-active enzyme (CAZYme) genes (237 genes on average) were found in all brown algal orders and in their sister taxon *S. ischiensis*, but this class of gene was less abundant in the more distantly related unicellular alga *H. akashiwo* (Figures 3A, S5A, and S5B; Tables S4A and S4B). 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 3B). This analysis indicated that several key genes and gene families (mannuronan C5 epimerase [ManC5-E] and polysaccharide lyase 41 [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.[3] It is therefore interesting that ManC5-E, an enzyme whose action modulates the rigidity of the alginate polymer, appears to have been acquired very early (Figures 3B and 3C). The acquisition of ManC5-E, together with other alginate pathway enzymes such as PL41 (Figures 3A, 3B, and 3D), was probably an important evolutionary step, enabling the emergence of large, resilient substrate-anchored multicellular organisms in the highly dynamic and stressful coastal environment (Figure 1).

Vanadium-dependent haloperoxidases (vHPOs) are a central component of brown algal halogen metabolism, which has been implicated in multiple biological processes including defense, adhesion, chemical signaling, and the oxidative stress response. All three classes of brown algal vHPO (algal types I and II and bacterial-type[29–31]) appear to have been acquired early during the emergence of the Phaeophyceae (Figures 3A, S5C, and S5D; Tables S4C and S4D). Closely related stramenopile species do not possess any of these three types of haloperoxidase, with the exception of the sister taxon, *S. ischiensis*, which possesses three intermediate algal type (i.e., equidistant phylogenetically from class I and class II algal types) haloperoxidase genes (Figures 3A, S5C, and S5D). Algal type I and II vHPO

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 (maximal values for each taxa) and indicates a number of key evolutionary events that occurred during the emergence of the Phaeophyceae. 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). ECM, extracellular matrix; asterisk (*), these orders were not analyzed in this study; Cr, Cryogenian; Ed, Ediacaran; Ca, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene.
See also Figures S1, S2, S4, and S5.



(legend on next page)

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.

Gene gain may not, however, have been the proximal factor responsible for all the key metabolic innovations that occurred in the emerging brown algal lineage. Phlorotannins are characteristic brown algal polyphenolic compounds that occur in all Phaeophyceae species, with the exception of some members of the Sargassaceae. Phlorotannins are derived from phloroglucinol and brown algae possess three classes of type III polyketide synthase, two of which (PKS1 and PKS2) were acquired prior to the emergence of the Phaeophyceae and the third (PKS3) evolving much later within the Ectocarpales (Figures 3A and 4A; Table S4E). Interestingly, PKS1 proteins from different brown algal species have been shown to have different activities leading to the production of distinct metabolites,[32–34] indicating that the acquisition of novel functions by this class of enzymes may have played an important role in the emergence of the brown algal capacity to produce phlorotannins. Moreover, many stramenopile PKS type III genes encode proteins with signal peptides or signal anchors (Table S4E). For the brown algae, this feature is consistent with the cellular production site of phlorotannins and the observed transport of these compounds by physodes, secretory vesicles characteristic of brown algae.[35] 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[36–38] and indirectly suggested by *in vivo* observations colocalizing vHPOs with physode fusions at the cell periphery.[39,40] 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 perhaps with modifications of the existing PKS enzymes, triggered the emergence of new metabolic pathways leading to the production of the phlorotannin molecules characteristic of the Phaeophyceae lineage.

The acquisition of increased multicellular complexity and adaptation to new ecological niches during the early stages of brown algal evolution (i.e., during and immediately following the GOBE) is expected to have required modification and elaboration of signaling pathways. Membrane-localized signaling proteins (Figure 3A) are of particular interest in this context not only as potential mediators of intercellular signaling in a multicellular organism but also because of potential interactions with the elaborate brown algal extracellular matrices (cell walls).[3,42] A detailed analysis of the brown algal receptor kinase (RK) gene family, revealed that it actually includes two types of receptor, the previously reported leucine-rich repeat (LRR) RKs[11] and a newly discovered class of receptors with a beta-propeller extracellular domain (Figure 3A; Table S4F).

Major changes in epigenetic regulation also appear to have occurred during the emergence of the brown algae (see also supplemental information). *DNA METHYLTRANSFERASE 1* (*DNMT1*) genes were identified in *Discosporangium mesarthrocarpum* and two closely related outgroup species (*S. ischiensis* and *C. australica*) but not in other brown algal genomes, 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 3A; Table S4G). This is consistent with the reported absence of DNA methylation in the filamentous brown alga *Ectocarpus*[11] and a very low level of DNA methylation in the kelp *Saccharina japonica*[43] (which is thought to be mediated by DNMT2). Our analysis indicates that most brown algae either lack DNA methylation or exhibit very low levels of methylation and that this feature was acquired early during brown algal diversification.

Impact of morphological, life cycle, and reproductive diversification during the Mesozoic on brown algal genome evolution

A second major step in the evolutionary history of the Phaeophyceae was the rapid diversification of the major brown algal orders, which began after the origin of the Fucophycidae/Dictyotales/Ishigeales (FDI) clade, here estimated at 235.97 Ma (95% highest posterior density region [HPD]: 158.88–312.48

Figure 2. Genome-wide analyses of brown algal genome and gene content evolution

(A) Time-calibrated cladogram based on Figure S2A. The gray hatched area, which indicates key nodes corresponding to the origin and early emergence of the brown algae, is mirrored in (B)–(F).

(B) Gene family (orthogroup) gain (green) and loss (red) during the emergence and diversification of the brown algae based on a Dollo parsimony reconstruction (Figure S2B).

(C) Upper: timing of gene family amplification and reduction during the evolutionary history of the Phaeophyceae (CAFE5 analysis). Lower: time of origin (orthogroup gain, based on the Dollo parsimony reconstruction) of the 180 most strongly amplified gene families.

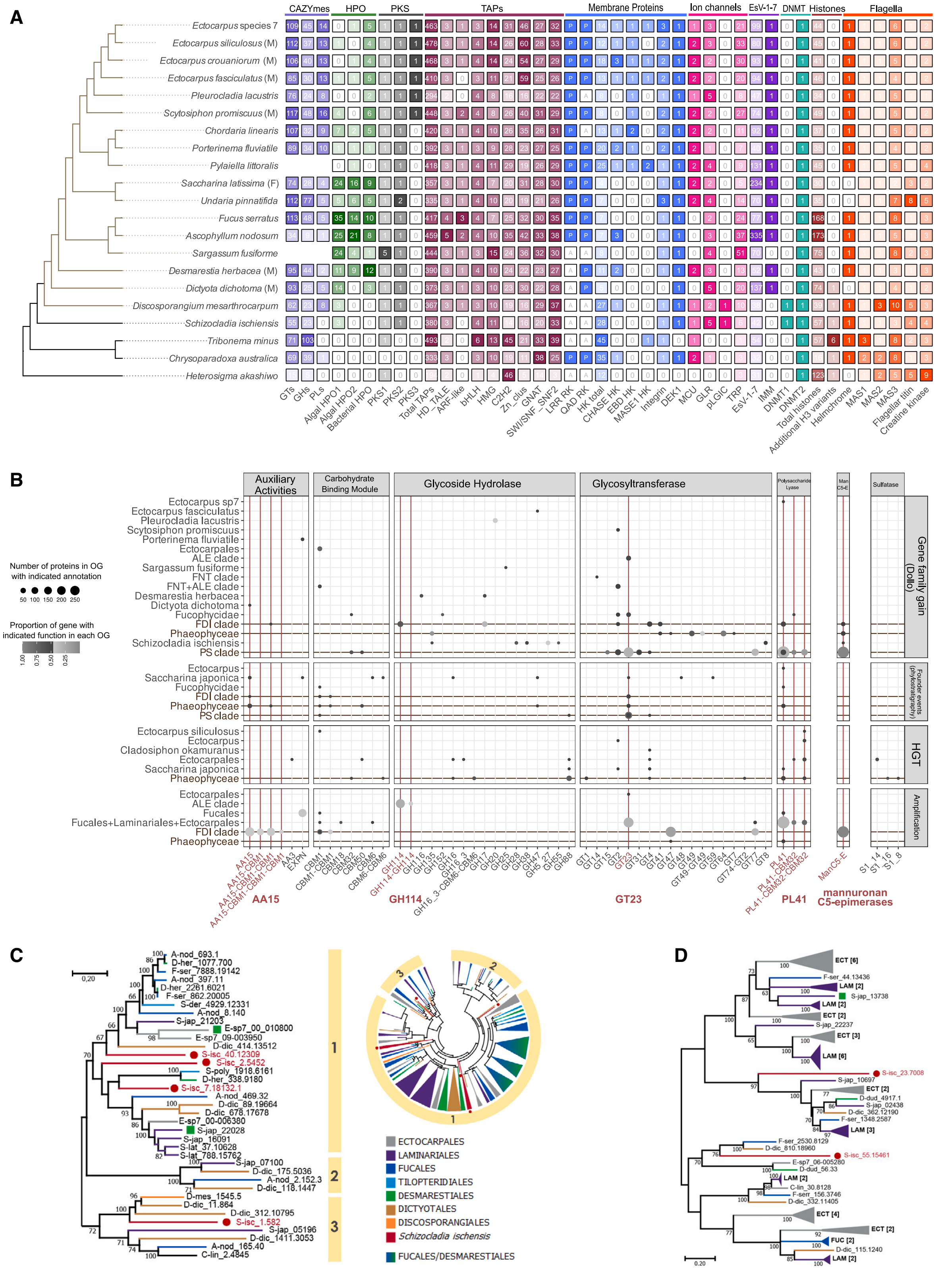
(D) Composite gene analysis. Proportions of gene families showing domain fusion (orange) or domain fission (yellow) at different age strata.

(E) Inferred gene family founder events after accounting for homology detection failure.

(F) Horizontal-gene-transfer-derived genes in orthologous groups and across species. The black trace represents the percentage of genes resulting from HGT events per species. Pie charts summarize 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 gray) and class-specific (i.e., acquired within the phylogenetic class, in blue) HGT genes.

(G) Enrichment of COG categories in sets of gene families identified as being (1) gained at the four indicated early nodes by the Dollo analysis, (2) gene founder events at the four indicated phylostrata, (3) amplified in the Phaeophyceae (180 most strongly amplified families), (4) domain fusions or fissions, and (5) HGT derived. Asterisks indicate significantly enriched categories.

FDI clade, Fucophycideae/Dictyotales/Ishigeales; PS clade, Phaeophyceae plus Schizocladiophyceae; PX clade, Phaeophyceae plus Xanthophyceae.
See also Figure S3.



(legend on next page)

mya, broadly consistent with previous work[1]; Figures 1 and S2A). This diversification closely followed the Permian-Triassic mass extinction event (which dramatically impacted marine ecosystems in which red and green algae played dominant roles) and was facilitated by Triassic marine environments that favored chlorophyll-c containing algae (e.g., high phosphate and low iron concentration), along with the appearance of new coastal niches created by Pangea rifting (Figure 1). This context would have facilitated the diversification of the brown algal lineage,[44,45] resulting in organisms that now exhibit a broad range of morphological complexity (ranging from filamentous to complex parenchymatous thalli), different types of life cycle and diverse reproductive strategies and metabolic capacities[3,6,46,47] (Figure 1). The Phaeoexplorer dataset was analyzed to identify genomic features associated with this diversification of phenotypic characteristics and to evaluate the impact on genome evolution and function.

We found indications that the diversification of life cycles, in some cases linked with the emergence of large, complex body architectures, impacted genome evolution through population genetic effects. Most brown algae have haploid-diploid life cycles involving alternation between sporophyte and gametophyte generations, the only exception being the Fucales, which have diploid life cycles. The theoretical advantages of different types of life cycle have been discussed in detail,[48] 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, 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 S5E). 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 S5E). This observation suggests that the emergence of larger, more complex brown algae may have resulted in reduced effective population sizes and consequently weaker counter-selection of non-synonymous substitutions.[24]

The diversification of the brown algae in terms of developmental complexity and life cycle structure was associated with modifications to reproductive systems, including, for example, 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 flagellum in *Dictyota dichotoma*.[49,50] Interestingly, these latter modifications are correlated with loss of the *HELMCHROME* gene, which is thought to be involved in light reception and zoid phototaxis,[41] from these species (Figures 3A and 4B). In addition, an analysis of the presence of genes for 70 high-confidence flagellar proteins[41] across eight species with different flagellar characteristics identified proteins that correlate with presence or absence of the eyespot or of the posterior flagellum (Figure 4B; Table S4H).

Brown algal diversification and the emergence of marine forests was also associated with genomic changes affecting metabolic and signaling pathways

Forests of brown algae (i.e., Laminariales, Desmarestiales, Tilopteridales, and Fucales[51]) are a key aspect of the modern marine biosphere. One of the pivotal innovations related to their emergence was a new developmental tissue, an intercalary meristem situated in the zone between the stipe and the lamina. The presence of this tissue is an ancestral state of the brown algal crown radiation (BACR) clade, and this study indicates that the intercalary meristem was acquired as early as 190 mya (Figure 1). This type of intercalary meristem would have facilitated the transition from annual to perennial life history and would, therefore, have been important for the establishment and maintenance of marine forests, particularly when upper parts of thalli are subjected to heavy grazing pressure.[10] Our results indicate that the Desmarestiales, Tilopteridales, and Fucales were all present by the early Cretaceous (Figure 1). Thus, it is possible that brown algal forests, at least at a small scale, provided both nutrients and shelter

Figure 3. 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 closely related 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. Dots represents functional family/orthogroup couples, with the size being proportional to the number of proteins annotated in the orthogroup (OG), and the color representing the proportion of the functional annotation that falls into this OG. Phaeophyceae plus Schizocladiophyceae (PS) and FDI clade, identified as gene innovation stages, are highlighted in brown. Functional categories with interesting evolutionary histories are highlighted in red.

(C) Phylogenetic tree of mannuronan C5-epimerases (ManC5-E). The phylogeny on the left, with three clusters indicated, is representative of the global view on the right.

(D) Phylogenetic tree of the polysaccharide lyase 41 (PL41) family. Green squares, biochemically characterized proteins. Brown algal sequences are color-coded in relation to their taxonomy, as indicated in (C). Schizocladiophyceae sequences 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 methyltransferase; 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, β-propeller domain; RK, membrane-localized receptor kinase; HK, histidine kinase; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor 1 domain; DEK1, defective kernel1; 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.

See also Figure S5.

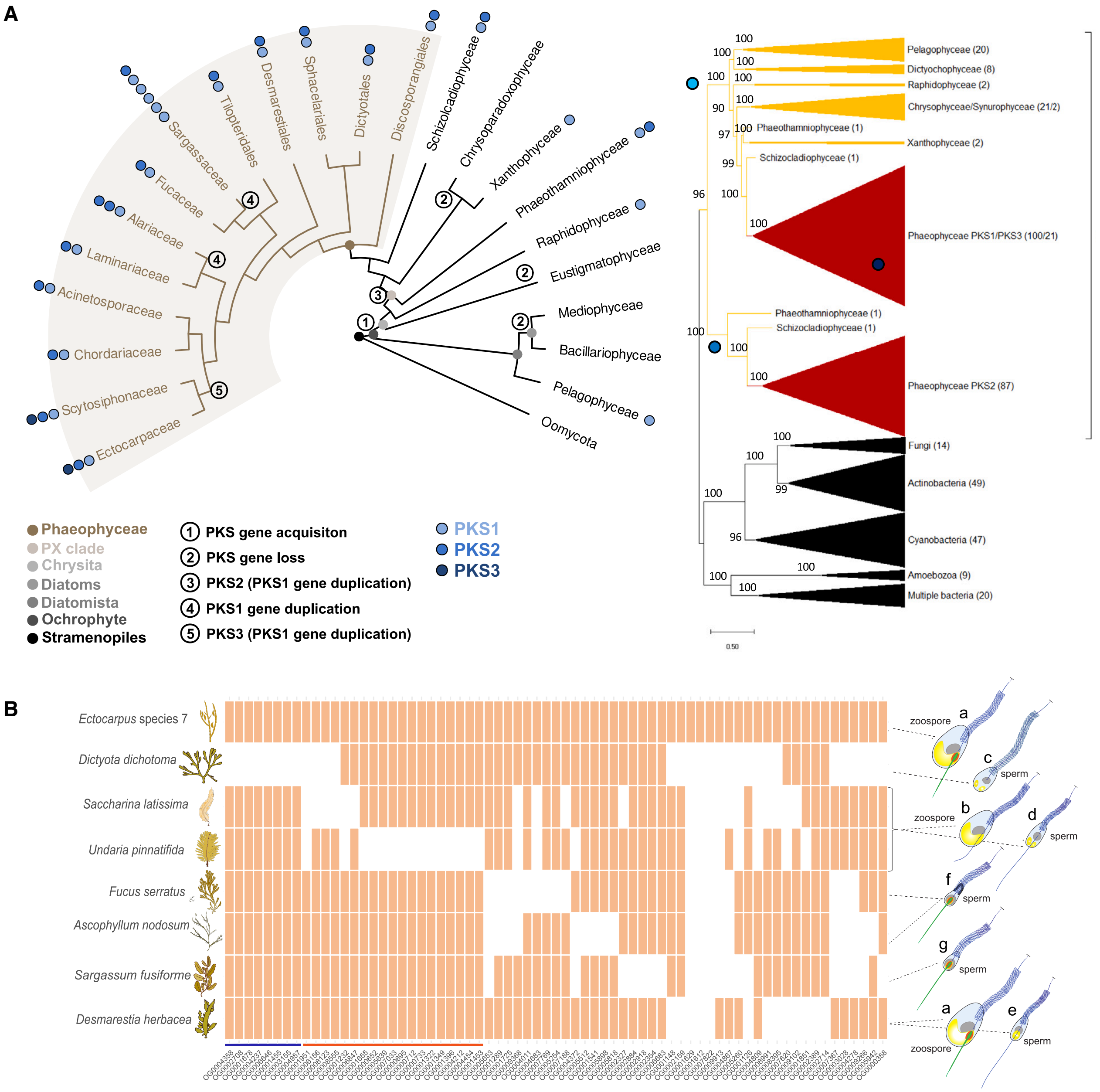


Figure 4. Evolution of key gene families during the emergence of the brown algal lineage

(A) Evolution of type III polyketide synthase (PKS) genes in the stramenopiles (left). Right: condensed view of a phylogenetic reconstruction tree of stramenopile PKS III and closely related sequences. In brackets: number of sequences identified in each phylogenetic group. Bootstrap values are indicated.

(B) Loss of orthogroups corresponding to flagellar proteome components[41] in eight brown algal species from five orders. For the zoid drawings: gray, 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. Bars below the heatmap indicate gene losses associated with loss of just the eyespot (orange) or of the entire posterior flagellum (blue).

for the marine herbivorous animals that became common during the Cretaceous Period (e.g., algae-eating echinoids, sea turtles, and euteleostean fish[52,53]).

While our estimates of kelp antiquity are earlier than those of Starko et al.,[54] they are consistent with their suggestion that Cenozoic cooling facilitated the geographic expansion of the kelp forest ecosystem. Indeed, many of the animals found today in kelp forest ecosystems originated toward the end of the Cretaceous Period, or later.[55] Currently, our understanding of Mesozoic marine noncalcified macroalgae on the basis of fossils[56–58] is too poor to provide much guidance in this regard, but documentation by Kiel et al.[55] of fossil holdfasts indicates that kelp forests were present by the late Paleogene period (~32 mya). The highly complex, multi-layered, and canopy-forming kelp forests of today, however, seem to have emerged only relatively recently, during the mid-Neogene, following the expansion of cooler water shelf environments.[54,55]

Comparative analysis of the Phaeoexplorer genome dataset identified a number of gene family expansions that potentially played important roles in the adaptation of the brown algae to their diverse niches and, more particularly, in the emergence of large, forest-forming species such as the kelps. For example, the ManC5-E family expanded markedly in the Laminariales and Fucales (Figure 3C), the two main orders that constitute extant phaeophycean forests. The capacity of ManC5-E to modify organ flexibility[3] may therefore have been an important factor for large organisms coping with the harsh hydrodynamic conditions of coastal environments.[59] 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 diversification of the brown algal lineage (Table S3), indicating that metabolic innovations affecting cell walls may have been concomitant with a complexification of associated signaling pathways.

Haloperoxidase gene families expanded independently in several brown algal orders, again with expansions being particularly marked in the Fucales and the Laminariales (Figures 3A and S5C). In the Laminariales, the algal type I family are specialized for iodine rather than bromine,[60] and this may have been an innovation that occurred specifically within the Laminariales, resulting in a halogen metabolism with an additional layer of complexity.

One of the proposed roles of halogenated molecules in brown algae is in biotic defense[4] and, clearly, an effective defense system would have been an important prerequisite for the emergence of the large, perennial organisms that constitute marine forests. Additional immunity-related families[61] that expanded during the diversification of the brown algae include five orthogroups that contain either GTPases with a central Ras of complex proteins/C-terminal of Roc 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 (Table S3).

Finally, one of the most remarkable gene family amplifications detected in this study was for proteins containing the EsV-1-7 domain, a short, cysteine-rich motif that may represent a novel class of zinc finger.[62] EsV-1-7 domain proteins are completely absent from animal and land plant genomes and most stramenopiles either have just one member (oomycetes and eustigmatophytes) or entirely lack this gene family.[62] Analysis of the Phaeoexplorer data (Figure 3A; Table S4I) indicated that the EsV-1-7 gene family started to expand in the common ancestor of the brown algae and the raphidophyte *H. akashiwo*, with 31–54 members in the non-Phaeophyceae taxa that share this ancestor. Further expansion of the family then occurred in most brown algal orders, particularly in some members of the Laminariales (234 genes in *Saccharina latissima*) and the Fucales (335 genes in *Ascophyllum nodosum*), with the genes tending to be clustered in tandem arrays (Tables S3 and S4I). These observations are consistent with the previous description of a large EsV-1-7 domain family (95 genes) in *Ectocarpus* species[7,62] and with recent observations by Nelson et al.[20] One member of this family, IMMEDIATE UPRIGHT (IMM), has been shown to play a key role in the establishment of the elaborate basal filament system of *Ectocarpus* sporophytes,[62] suggesting that EsV-1-7 domain proteins may be novel developmental regulators in brown algae. Orthologs of the *IMM* gene were found in brown algal crown group taxa and in *D. dichotoma* but not in *D. mesarthrocarpum* (Figure 3A; Table S4I), indicating that this gene originated within the EsV-1-7 gene family as the first brown algal orders started to diverge.

Recent evolutionary events within the genus *Ectocarpus*

The above analyses focused on deep-time evolutionary events related to the emergence of the Phaeophyceae and the later diversification of the brown algal orders during the Mesozoic. To complement these analyses an evaluation of relatively recent and ongoing evolutionary events in the brown algae was conducted by sequencing 22 new strains from the genus *Ectocarpus*, which originated about 19 mya (Figure S2C).

A phylogenetic tree was constructed for 11 selected *Ectocarpus* species based on 261 high-quality alignments of 1:1 orthologs (Figure 5A). The tree indicates substantial divergence between *E. fasciculatus* and two well-supported clades, designated clade 1 and clade 2. Incongruencies between the species tree and trees for individual genes indicated introgression events and/or incomplete lineage sorting across the *Ectocarpus* genus. 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 clades (Figure 5B), suggesting that gene flow has not been limited to recently diverged species pairs. These findings suggest a complex evolutionary history involving rapid divergence, hybridization, and introgression among species within the *Ectocarpus* genus, with evidence for hybridization occurring between 10.5 (for clades 1 and 2) and 3.3 mya (for *Ectocarpus* species 5 and 7) based on the fossil-calibrated tree (Figure S2C). A similar scenario has been reported for the genus *Drosophila*,[63] suggesting that recurrent hybridization and introgression among species may be a common feature associated with rapid species radiations. Major environmental changes such as the expansion of cold-water coastal areas following the green-house/cold-house Eocene-Oligocene transition (~30 mya[64]), and particularly the rapid climate destabilization and temperature drop associated with the end of the mid-Miocene thermal maximum (~15 mya[64]), may have created many new opportunities for the rapid expansion and diversification of the *Ectocarpus* genus.

Brown algal genomes contain large amounts of inserted viral sequences

A particularly striking result of this study was the identification of extensive amounts of integrated DNA sequence corresponding to large DNA viruses of the *Phaeovirus* family (Figure 6A; Table S5), which integrate into brown algal genomes as part of their lysogenic life cycles.[65] 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.

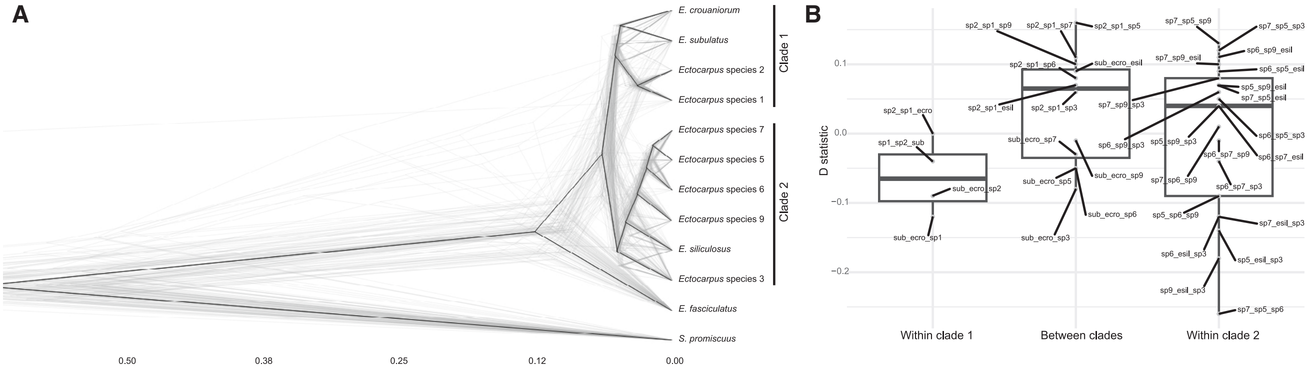


Figure 5. Evidence for gene flow within the genus *Ectocarpus*

(A) DensiTree visualisation of gene trees (gray lines) for 261 orthologs shared by 11 *Ectocarpus* species and the outgroup species *S. promiscuus*, together with the consensus species tree (black lines). All nodes of the species tree have posterior probabilities greater than 0.99.

(B) Boxplot reporting 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. *Ectocarpus fasciculatus* was defined as the outgroup.

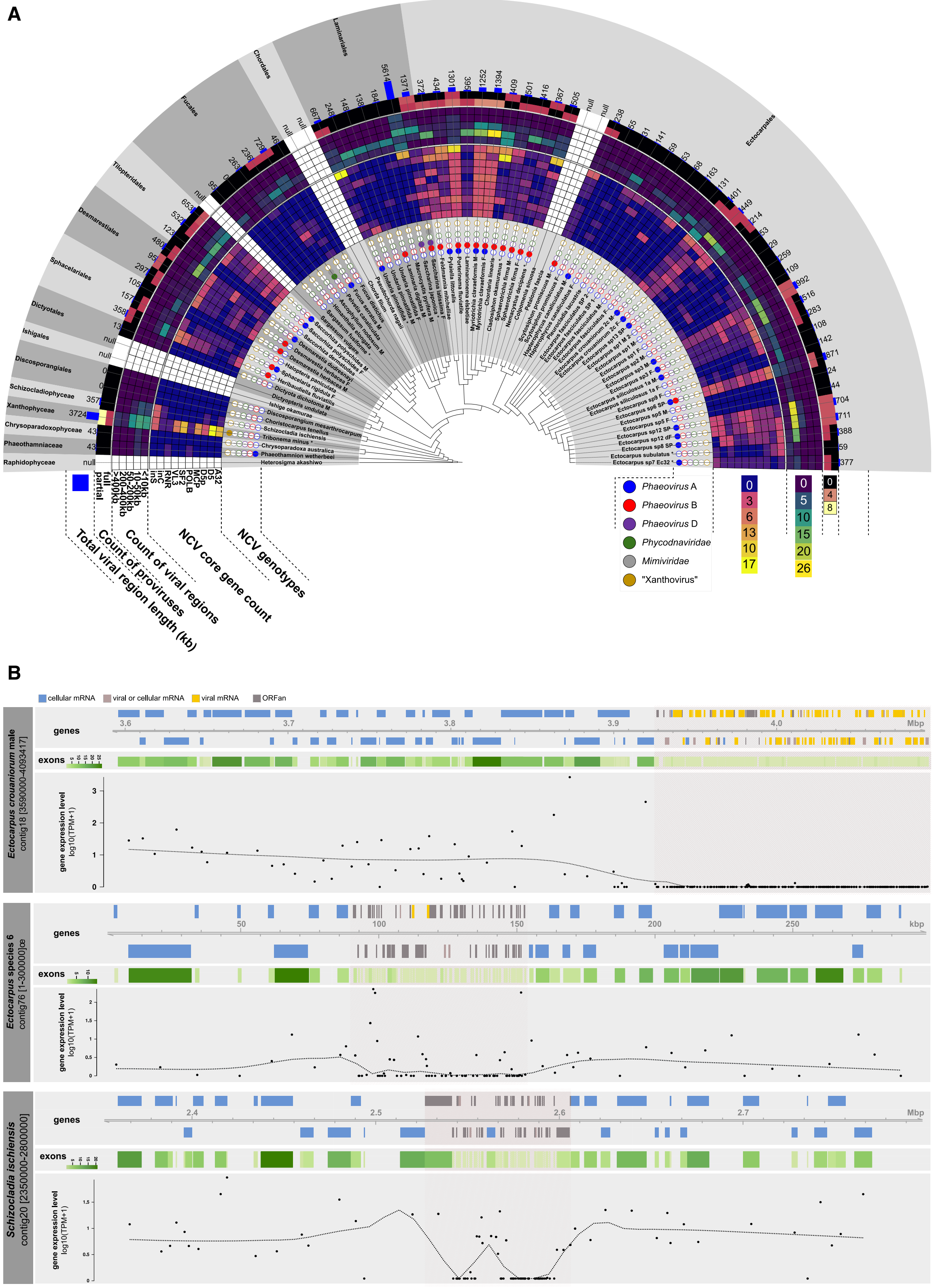
See also Figure S2.

Individual VRs ranged in size from two to 705 kbp, but the majority (81.3%) were between two and 50 kbp, while only 9% were longer than the expected minimum size (100 kbp) for an NCV genome. At least one flanking region could be identified for 40.8% of the VRs, providing direct evidence for insertion of the sequence in the algal genome (Table S5C). Figure 6B shows three examples of long VRs. Most genes in VRs are monoexonic and transcriptionally silent, as previously observed for the 310 kbp VR in the *Ectocarpus* species 7 strain Ec32 genome.[11]

On average, each of the 72 analyzed genomes contained 469 kbp of VR (with a maximum of 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 Mbp of VRs (*T. minus*, *S. latissima*, *S. japonica*, *P. fluviatile*, and *Myriotrichia clavaeformis* male and female). 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 6A; Table S5). In addition to the previously known infections in Ectocarpales[65] and Laminariales,[66] 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, and Laminariales), and the Xanthophyceae (Figure 6A; Table S5). Phylogenetic analysis of the major capsid 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 (Figures S6A and S6B). 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.[67]

The identification of integrated NCVs across almost all brown algal orders and in closely related outgroup 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 S3A). 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 6A; Table S5). 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,[68] and rates of VR loss due to meiotic segregation.[69] 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. The identification of large-scale viral genome insertion events over such a long timescale (at least 450 mya[1]) suggests that NCVs may have had a major impact on the evolution of brown algal genomes throughout the emergence of the lineage.

The widespread presence of large quantities of viral genes in brown algal genomes creates a favorable situation for recruitment of this genetic information by the algal host via HGT (provided the acquired genes confer a selective advantage[70]), but clear evidence of this type of HGT event can be difficult to obtain. However, phylogenetic evidence indicates that several *Ectocarpus* species 7 histidine kinases (HKs) were derived by HGT from viral insertions[71] and analysis of the Phaeoexplorer genomes supported this hypothesis. HKs are widespread in the



(legend on next page)

stramenopiles but several classes of membrane-localized HK were either only found in brown algae (cyclases/histidine kinases associated sensory extracellular [CHASE] domain HKs and HKs with an extracellular domain resembling an ethylene binding motif[71]) or only in brown algae and closely related taxa (membrane-associated sensor 1 [MASE1] domain HKs[71]) and appear to be absent from other stramenopile lineages (Figure 3A; Table S4J). 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 following integration of viral genomes into algal genomes (Figures 3A and S6C). Phylogenetic analysis provided further support for a HGT origin for these classes of HK (Figure S6C).

DISCUSSION

Comparative analysis of the genome resource presented in this study has provided insights into genome evolution across the entire evolutionary history of the brown algae. A period of marked genome evolution concomitant with the emergence of the brown algal lineage during the GOBE was correlated with an increase in multicellular complexity, possibly driven, at least in part, by increases in atmospheric oxygen and herbivory. During this period, the brown algae acquired key components of several metabolic pathways, notably cell-wall polysaccharide, phlorotannin, and halogen metabolisms, that were essential for their colonization of intertidal and subtidal environments. The capacity to synthesize flexible and resilient alginate-based cell walls[72] allows these organisms to resist the hydrodynamic forces of wave action,[59] whereas phlorotannins and halogen derivatives are thought to play important roles in defense.[73] 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.[74] The capacity to adhere strongly and resist both biotic and abiotic stress factors would prove essential for the success of large, sedentary multicellular organisms in these intertidal niches over evolutionary time.

The period of increased gene gain during the emergence of the brown algae was followed by a period of overall gene loss that extended up until the present day (Figure S2B). Interestingly, similar periods of ancestral gene gain followed by gene loss have also been observed for both the animal and land plant lineages,[75] indicating that this may be a common feature of multicellular eukaryotic lineages.

About 220 mya after the emergence of the brown algae, the aftermath of the Permian-Triassic mass extinction event and the initiation of Pangea rifting appear to have created favorable environments for rapid diversification of the main brown algal orders,[44,45] resulting in the emergence of a diversity of developmental, life cycle, and reproductive strategies, with correlated effects on genome evolution. During this period some orders, such as the Laminariales and Fucales, acquired characteristics such as an intercalary meristems and modified metabolic, defense, and developmental processes that are predicted to have been important prerequisites for the emergence of marine forests.

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.

Finally, one of the most surprising observations was that brown algal genomes contain many inserted viral sequences corresponding to large DNA viruses of the *Phaeovirus* family. Inserted viral sequences are widespread in eukaryotic genomes[76,77] and insertions corresponding to nucleocytoplasmic large DNA viruses have been found in green algal genomes[67,78] but the brown algal *Phaeovirus* VRs are remarkable because they are nearly ubiquitous in this lineage (being present in 67 of 69 brown algal genomes analyzed) and because individual genomes can contain several phylogenetically diverse *Phaeovirus* insertions and insertions of a broad range of different sizes. The near ubiquitous occurrence of these elements may be attributed to the capacity of phaeoviruses to insert into their hosts' genomes as part of their life cycle.

The above observations illustrate how the Phaeoexplorer genome dataset, along with the various analyses carried out in this study, can be used to link the gene content of brown algal genomes to biological processes and characteristics that have played fundamental roles during the evolution of this lineage. The establishment of this genome resource represents an important step forward for a key lineage that has remained poorly characterized at the genome level. The Phaeoexplorer dataset not only provides good quality genome assemblies for many, previously uncharacterized brown algal species but also represents a tool to explore genome function via comparative genomics approaches, adding an important evolutionary dimension to efforts to understand gene function in this lineage. The identification and analysis of key metabolic and signaling genes implicated

Figure 6. Inserted viral regions in brown algal genomes

(A) Annotated phylogeny summarizing key statistics of the presence of *Nucleocytoviricota* (NCV) sequences in the genomes of brown algae and closely related taxa. Eight genomes sourced from public databases are labeled with an asterisk. Outer layers around the 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; POLB, 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.

(B) Three examples of contigs containing large viral insertions (pink shading). Genes (colored boxes) were classified as viral, cellular (i.e., cellular organism), known proteins of unclear origin (viral or cellular) or unknown (ORFan) based on comparisons with viral and cellular protein databases (see STAR Methods). Transcript abundances are shown with a locally estimated scatterplot smoothing (LOESS) plot. Exons, exons per gene.

See also Figure S6.

in a broad range of brown algal biological functions represents an important resource for future research programs aimed at optimizing brown seaweed production in a mariculture context or at preserving and protecting natural seaweed populations in the context of climate change. Both of these approaches could potentially contribute to mitigation of the effects of climate change via multiple positive effects in terms of carbon capture, ecosystem services, and the promotion of highly sustainable cultivation practices.

To facilitate future use of this genome dataset, the annotated genomes have been made available through a website portal (https://phaeoexplorer.sb-roscoff.fr). The existing genome dataset provides very good coverage of the phylogenetic diversity of the Phaeophyceae and reasonably complete gene catalogs for each species, but future work is needed to improve further the quality of the genome assemblies described here and to add genomes for additional species, particularly members of the minor brown algal orders that are not represented in the dataset. The large proportion of genes with no predicted function in brown algal genomes is also a limitation that needs to be addressed. The recent development of CRISPR-Cas9 methodology for brown algae,[79,80] together with the other tools and resources currently available for the model brown alga *Ectocarpus*,[81] provide the means to deploy the functional genomics approaches necessary to address this question.

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, genome assemblies, and annotations, have been deposited in the European Bioinformatics Institute/European Nucleotide Archive (EBI/ENA) database under the project accession PRJEB76691 and are publicly available. Additional data and results have been deposited in the CNRS Research Data depository (https://doi.org/10.57745/9U1J85) and are publicly available.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

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 China Agriculture Research System (CARS-50), 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) and Ministry of Oceans and Fisheries (KIMST-20180430) 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, Dr. Karl Feldbausch-Stiftung, 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. Sussfeld, X.F., L. Mazéas, N. Terrapon, J.B.-R., R.P., L.R., S.-W.C., J.J., K.U., K.B., C. Duc, P.R., A.L., E.A.M., M.L., A.K., P.H.-G., C.V., S.S.A., S.A., K.A., Y.B., T. Barbeyron, A.B., S.B., A.B.-C., A. Cormier, H.C.d.C., A.D., E. Dinatale, S.D., E. Drula, J.G., L.G., A.G., M.-L.G., L.H., B.H., A.J., E.K., A.H.K., C.L., L.L.G., R.L., S.T.L., P.J.L., E.M., S.M., G.M., C.N., S.A.R., E.R., D. Schroeder, 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. Tadrent, A. Couloux, B.N., W.B., E. Denis, C.J., L. Mest, S.R., and D. Scornet; resources, O.G., A. Couloux, L.B.-G., T. Bringloe, R.A.C., C. Destombe, S.F., W.J.H., G.H., K.K., A.L.B., K.M., C.M., N.P., P.P., S.R., D. Scornet, H.V., F.W., H.K., A.F.P., M.V., E.C., and J.-M.A.; data curation, O.G., C.C., A. Couloux, 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 and editing, all authors; visualization, F.D., O.G., L.B.-G., L.D., D.M., T.M., D. Sussfeld, X.F., L. Mazéas, 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. Terrapon, M.L., R.A.C., H.C.d.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.d.C., C.B., P.P., C.H., N.Y., S.M.C., and J.M.C.; conceptualization, J.M.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL DETAILS
    - *Ascophyllum nodosum*
    - *Chordaria linearis* strain ClinC8C
    - *Choristocarpus tenellus* strain KU-1152
    - *Chrysoparadoxa australica* strain CS-1217
    - *Cladosiphon okamuranus* strain S-strain
    - *Desmarestia dudresnayi* strain DdudBR16
    - *Desmarestia herbacea* strain DmunF
    - *Desmarestia herbacea* strain DmunM
    - *Dictyota dichotoma* strain KB07f IV
    - *Dictyota dichotoma* strain ODC1387m
    - *Dictyota dichotoma* strain KB07m IV
    - *Dictyota dichotoma* strain KB07sp VI
    - *Discosporangium mesarthrocarpum* strain MT17-79
    - *Ectocarpus crouaniorum* strain Ec861
    - *Ectocarpus crouaniorum* strain Ec862
    - *Ectocarpus fasciculatus* strain Ec846
    - *Ectocarpus fasciculatus* strain Ec847
    - *Ectocarpus fasciculatus* strain EfasUO1
    - *Ectocarpus fasciculatus* strain EfasUO2
    - *Ectocarpus siliculosus* strain Ec863
    - *Ectocarpus siliculosus* strain Ec864
    - *Ectocarpus* species 1 strain Ec sil Puy CHCH Z9 G5f

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.10.049.

Received: February 19, 2024
Revised: July 20, 2024
Accepted: October 28, 2024
Published: November 20, 2024

REFERENCES

1. 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. *34*, 740–754.e4. https://doi.org/10.1016/j.cub.2023.12.069.

2. Keeling, P.J. (2009). Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Eukaryot. Microbiol. *56*, 1–8. https://doi.org/10.1111/j.1550-7408.2008.00371.x.

3. 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. Semin. Cell Dev. Biol. *134*, 112–124. https://doi.org/10.1016/j.semcdb.2022.03.005.

4. 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. https://doi.org/10.1039/c8mt00327k.

5. Schoenwaelder, M.E.A. (2008). The biology of phenolic containing vesicles. Algae *23*, 163–175. https://doi.org/10.4490/ALGAE.2008.23.3.163.

6. 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. https://doi.org/10.1016/j.pbi.2013.09.004.

7. 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. https://doi.org/10.1038/s41467-023-37385-0.

8. 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. https://doi.org/10.1016/j.cub.2011.09.028.

9. 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. https://doi.org/10.1016/j.scitotenv.2023.163699.

10. Bringloe, T.T., Starko, S., Wade, R.M., Vieira, C., Kawai, H., De Clerck, O.D., Cock, J.M., Coelho, S.M., Destombe, C., Valero, M., et al. (2020). Phylogeny and evolution of the brown algae. Crit. Rev. Plant Sci. *39*, 281–321. https://doi.org/10.1080/07352689.2020.1787679.

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

12. 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. https://doi.org/10.1038/ncomms7986.

13. 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*. Nat. Ecol. Evol. *5*, 360–368. https://doi.org/10.1038/s41559-020-01378-9.

14. 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. https://doi.org/10.3389/fgene.2020.590065.

15. 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. https://doi.org/10.1186/s12864-023-09658-x.

16. Dittami, S.M., Scornet, D., Petit, J.L., Ségurens, B., Da Silva, C., Corre, E., Dondrup, M., Glatting, K.H., König, R., Sterck, L., et al. (2009). Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. Genome Biol. *10*, R66. https://doi.org/10.1186/gb-2009-10-6-r66.

17. Wang, S., and Wu, M. (2023). The draft genome of the "golden tide" seaweed, *Sargassum horneri*: characterization and comparative analysis. Genes (Basel) *14*, 1969. https://doi.org/10.3390/genes14101969.

18. 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. https://doi.org/10.1093/dnares/dsw039.

19. 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. https://doi.org/10.1038/s41598-019-40955-2.

20. Nelson, D.R., Mystikou, A., Jaiswal, A., Rad-Menendez, C., Preston, M.J., Boever, F.D., Assal, D.C.E., Daakour, S., Lomas, M.W., Twizere, J.-C., et al. (2024). Macroalgal deep genomics illuminate multiple paths to aquatic, photosynthetic multicellularity. Mol. Plant *17*, 747–771. https://doi.org/10.1016/j.molp.2024.03.011.

21. LoDuca, S.T., Bykova, N., Wu, M., Xiao, S., and Zhao, Y. (2017). Seaweed morphology and ecology during the great animal diversification events of the Early Paleozoic: A tale of two floras. Geobiology *15*, 588–616. https://doi.org/10.1111/gbi.12244.

22. Kawai, H., Maeba, S., Sasaki, H., Okuda, K., and Henry, E.C. (2003). Schizocladia ischiensis: a new filamentous marine chromophyte belonging to a new class, Schizocladiophyceae. Protist *154*, 211–228. https://doi.org/10.1078/143446103322166518.

23. 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. https://doi.org/10.1093/molbev/msm048.

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

25. 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. https://doi.org/10.1093/molbev/msad011.

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

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

28. 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. https://doi.org/10.1093/molbev/msu083.

29. 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. https://doi.org/10.1016/j.bioorg.2012.05.003.

30. 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. https://doi.org/10.1021/acschembio.8b00041.

31. 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. https://doi.org/10.1128/AEM.02430-14.

32. Meslet-Cladière, L., Delage, L., Leroux, C.J.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. https://doi.org/10.1105/tpc.113.111336.

33. 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. https://doi.org/10.1007/s10126-010-9344-5.

34. 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. https://doi.org/10.1021/acs.jafc.1c05868.

35. Schoenwaelder, M.E.A., and Wiencke, C. (2000). Phenolic compounds in the embryo development of several Northern Hemisphere fucoids. Plant Biol. *2*, 24–33. https://doi.org/10.1055/s-2000-9178.

36. 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. https://doi.org/10.1111/j.1529-8817.2008.00642.x.

37. 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. https://doi.org/10.1021/bm0496864.

38. 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. Macromol. Biosci. *7*, 1280–1289. https://doi.org/10.1002/mabi.200700099.

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

40. 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. https://doi.org/10.1590/S1679-87592007000200007.

41. 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. https://doi.org/10.1016/j.protis.2014.07.007.

42. 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. https://doi.org/10.3390/genes12071059.

43. 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. https://doi.org/10.1111/nph.16125.

44. Knoll, A.H., Summons, R.E., Waldbauer, J.R., and Zumberge, J.E. (2007). Chapter 8. The geological succession of primary producers in the oceans. In Evolution of Primary Producers in the Sea, P.G. Falkowski and A.H. Knoll, eds. (Academic Press), pp. 133–163. https://doi.org/10.1016/B978-012370518-1/50009-6.

45. Schettino, A., and Turco, E. (2009). Breakup of Pangaea and plate kinematics of the central Atlantic and Atlas regions. Geophys. J. Int. *178*, 1078–1097. https://doi.org/10.1111/j.1365-246X.2009.04186.x.

46. 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. https://doi.org/10.1101/gr.277056.122.

47. Coelho, S.M., Mignerot, L., and Cock, J.M. (2019). Origin and evolution of sex-determination systems in the brown algae. New Phytol. *222*, 1751–1756. https://doi.org/10.1111/nph.15694.

48. Coelho, S.M., Peters, A.F., Charrier, B., Roze, D., Destombe, C., Valero, M., and Cock, J.M. (2007). Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms. Gene *406*, 152–170. https://doi.org/10.1016/j.gene.2007.07.025.

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

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

51. Fragkopoulou, E., Serrão, E.A., De Clerck, O., Costello, M.J., Araújo, M.B., Duarte, C.M., Krause-Jensen, D., and Assis, J. (2022). Global biodiversity patterns of marine forests of brown macroalgae. Glob. Ecol. Biogeogr. *31*, 636–648. https://doi.org/10.1111/geb.13450.

52. Vermeij, G.J., and Lindberg, D.R. (2000). Delayed herbivory and the assembly of marine benthic ecosystems. Paleobiology *26*, 419–430. https://doi.org/10.1666/0094-8373(2000)026<0419:DHATAO>2.0.CO;2.

53. Alfaro, M.E., Faircloth, B.C., Harrington, R.C., Sorenson, L., Friedman, M., Thacker, C.E., Oliveros, C.H., Černý, D., and Near, T.J. (2018). Explosive diversification of marine fishes at the Cretaceous-Palaeogene boundary. Nat. Ecol. Evol. *2*, 688–696. https://doi.org/10.1038/s41559-018-0494-6.

54. Starko, S., Soto Gomez, M., Darby, H., Demes, K.W., Kawai, H., Yotsukura, N., Lindstrom, S.C., Keeling, P.J., Graham, S.W., and Martone, P.T. (2019). A comprehensive kelp phylogeny sheds light on the evolution of an ecosystem. Mol. Phylogenet. Evol. *136*, 138–150. https://doi.org/10.1016/j.ympev.2019.04.012.

55. Kiel, S., Goedert, J.L., Huynh, T.L., Krings, M., Parkinson, D., Romero, R., and Looy, C.V. (2024). Early Oligocene kelp holdfasts and stepwise evolution of the kelp ecosystem in the North Pacific. Proc. Natl. Acad. Sci. USA *121*, e2317054121. https://doi.org/10.1073/pnas.2317054121.

56. Basson, P.W. (1981). Late Cretaceous alga, *Delesserites libanensis* sp. nov. Rev. Palaeobot. Palynol. *33*, 363–370. https://doi.org/10.1016/0034-6667(81)90093-2.

57. Krings, M., and Mayr, H. (2004). *Bassonia hakelensis* (Basson) nov. comb., a rare non-calcareous marine alga from the Cenomanian (Upper Cretaceous) of Lebanon. Zitteliana *44*, 105–111.

58. Barthel, K.W., and Swinburne, N.H.M. (1994). Solnhofen: a Study in Mesozoic Palaeontology (Cambridge University Press).

59. Martone, P.T., Kost, L., and Boller, M. (2012). Drag reduction in wave-swept macroalgae: alternative strategies and new predictions. Am. J. Bot. *99*, 806–815. https://doi.org/10.3732/ajb.1100541.

60. 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. https://doi.org/10.1007/s00775-005-0626-8.

61. Zambounis, A., Elias, M., Sterck, L., Maumus, F., and Gachon, C.M.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. https://doi.org/10.1093/molbev/msr296.

62. 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. https://doi.org/10.1242/dev.141523.

63. Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., Price, D.K., Waddell, P.J., Lang, M., Courtier-Orgogozo, V., et al. (2022). Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr. Biol. *32*, 111–123.e5. https://doi.org/10.1016/j.cub.2021.10.052.

64. Rohling, E.J., Yu, J., Heslop, D., Foster, G.L., Opdyke, B., and Roberts, A.P. (2021). Sea level and deep-sea temperature reconstructions suggest quasi-stable states and critical transitions over the past 40 million years. Sci. Adv. *7*, eabf5326. https://doi.org/10.1126/sciadv.abf5326.

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

66. 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. https://doi.org/10.1038/ismej.2017.130.

67. 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. https://doi.org/10.1038/s41586-020-2924-2.

68. 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. https://doi.org/10.1371/journal.pone.0086040.

69. 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. https://doi.org/10.1111/j.0022-3646.1995.00823.x.

70. Keeling, P.J. (2024). Horizontal gene transfer in eukaryotes: aligning theory with data. Nat. Rev. Genet. *25*, 416–430. https://doi.org/10.1038/s41576-023-00688-5.

71. 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. https://doi.org/10.1093/gbe/evy213.

72. Mazéas, L., Bouguerba-Collin, A., Cock, J.M., Denoeud, F., Godfroy, O., Brillet-Guéguen, L., Babbeyron, T., Lipinska, A.P., Delage, L., Corre, E., et al. (2024). Candidate genes involved in biosynthesis and degradation of the main extracellular matrix polysaccharides of brown algae and their probable evolutionary history. BMC Genom. *25*, 950. https://doi.org/10.1186/s12864-024-10811-3.

73. 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. https://doi.org/10.1016/s1369-5266(02)00273-x.

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

75. Domazet-Lošo, M., Široki, T., Šimičević, K., and Domazet-Lošo, T. (2024). Macroevolutionary dynamics of gene family gain and loss along multicellular eukaryotic lineages. Nat. Commun. *15*, 2663. https://doi.org/10.1038/s41467-024-47017-w.

76. Holmes, E.C. (2011). The evolution of endogenous viral elements. Cell Host Microbe *10*, 368–377. https://doi.org/10.1016/j.chom.2011.09.002.

77. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296. https://doi.org/10.1038/nrg3199.

78. Moniruzzaman, M., Erazo-Garcia, M.P., and Aylward, F.O. (2022). Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. Virus Evol. *8*, veac102. https://doi.org/10.1093/ve/veac102.

79. 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. https://doi.org/10.1111/nph.17525.

80. 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. *35*, 1431–1441. https://doi.org/10.1007/s10811-023-02940-1.

81. Cock, J.M. (2023). The model system *Ectocarpus*: integrating functional genomics into brown algal research. J. Phycol. *59*, 4–8. https://doi.org/10.1111/jpy.13310.

82. 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. https://doi.org/10.1093/bioinformatics/btv033.

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

84. Altschul, S.F., Madden, T.L., Schäffer, 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 Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

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

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

87. 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. J. Comput. Biol. *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

88. 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. https://doi.org/10.46471/gigabyte.15.

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

90. 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. https://doi.org/10.1038/s41587-019-0072-8.

91. 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. https://doi.org/10.1038/s41467-020-20236-7.

92. 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. https://doi.org/10.1101/gr.214270.116.

93. Aury, J.-M., and Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom. Bioinform. *3*, lqab034. https://doi.org/10.1093/nargab/lqab034.

94. 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. https://doi.org/10.7717/peerj.7359.

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

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

97. 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. https://doi.org/10.1093/bioinformatics/bts094.

98. 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 Res. *43*, D222–D226. https://doi.org/10.1093/nar/gku1221.

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

100. 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. https://doi.org/10.1038/nbt.1883.

101. 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. https://doi.org/10.1093/gigascience/giz100.

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

103. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580. https://doi.org/10.1093/nar/27.2.573.

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

105. Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

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

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

108. Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. *13*, 477–478. https://doi.org/10.1093/bioinformatics/13.4.477.

109. 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*. https://doi.org/10.7490/f1000research.1111735.1.

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

111. Stamatakis, A. (2015). Using RAxML to infer phylogenies. Curr. Protoc. Bioinformatics *51*, 6.14.1–6.14.14. https://doi.org/10.1002/0471250953.bi0614s51.

112. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.

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

114. Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912. https://doi.org/10.1093/bioinformatics/btq315.

115. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

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

117. 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. https://doi.org/10.1093/nar/gkt263.

118. 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. https://doi.org/10.1186/s13059-023-02895-z.

119. 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. https://doi.org/10.1093/nar/30.7.1575.

120. van Kempen, M., 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. https://doi.org/10.1101/2022.02.07.479398.

121. 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. https://doi.org/10.1093/molbev/msx283.

122. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*, 5691–5702. https://doi.org/10.1093/nar/gki866.

123. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.

124. 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 Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

125. 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. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

126. 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. https://doi.org/10.1186/1471-2105-11-119.

127. 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. https://doi.org/10.3390/v13020150.

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

129. 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. https://doi.org/10.1093/bioinformatics/btw006.

130. 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. https://doi.org/10.1093/molbev/mst010.

131. 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. https://doi.org/10.1093/molbev/msr121.

132. Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., and Gascuel, O. (2019). NGPhylogeny.fr: new

generation phylogenetic services for non-specialists. Nucleic Acids Res. *47*, W260–W265. https://doi.org/10.1093/nar/gkz303.

133. 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. https://doi.org/10.1093/bioinformatics/btp348.

134. 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. https://doi.org/10.3390/genes12071055.

135. Petroll, R., Varshney, D., Hiltemann, S., Finke, H., Schreiber, M., de Vries, J., and Rensing, S.A. (2024). Enhanced sensitivity of TAPscan v4 enables comprehensive analysis of streptophyte transcription factor evolution. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.13.602682.

136. Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Res. *49*, W216–W227. https://doi.org/10.1093/nar/gkab225.

137. 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. https://doi.org/10.1093/bioinformatics/btu031.

138. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (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. https://doi.org/10.1093/nar/22.22.4673.

139. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

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

141. 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. https://doi.org/10.1186/s13059-014-0550-8.

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

143. Krueger, F. (2015). Trim Galore!: A Wrapper around Cutadapt and FastQC to Consistently Apply Adapter and Quality Trimming to FastQ Files, with Extra Functionality for RRBS Data (Babraham Institute).

144. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

145. 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. https://doi.org/10.1093/bioinformatics/btt656.

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

147. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

148. 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. https://doi.org/10.1093/molbev/msv341.

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

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

151. 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. https://doi.org/10.1093/nar/gkx391.

152. 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. https://doi.org/10.1093/nar/gkh152.

153. 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. https://doi.org/10.1038/nmeth.4285.

154. 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. https://doi.org/10.1093/molbev/msx281.

155. 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. https://doi.org/10.1093/bioinformatics/btx144.

156. 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. https://doi.org/10.1093/bioinformatics/bth397.

157. 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. https://doi.org/10.1371/journal.pcbi.1003537.

158. Douglas, J., Jiménez-Silva, C.L., and Bouckaert, R. (2022). StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent. Syst. Biol. *71*, 901–916. https://doi.org/10.1093/sysbio/syac010.

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

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

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

162. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

163. 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. https://doi.org/10.1093/molbev/msaa015.

164. Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics *36*, 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

165. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

166. Wickham, H., Chang, W., and Wickham, M.H. (2016). Package 'ggplot2.' Create Elegant Data Visualisations Using the Grammar of Graphics, version 2, pp. 1–189.

167. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.

168. 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. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

169. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932. https://doi.org/10.1093/bioinformatics/btu739.

170. 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. https://doi.org/10.1093/nar/gkab1061.

171. 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. https://doi.org/10.1186/1743-422X-6-223.

172. Trgovec-Greif, L., Hellinger, H.-J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N.S., Laffy, P.W., Feichtinger, M., and Rattei, T. (2024). VOGDB—database of virus orthologous groups. Viruses *16*, 1191. https://doi.org/10.3390/v16081191.

173. 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. https://doi.org/10.1371/journal.pone.0164846.

174. 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. https://doi.org/10.1093/nar/gkac977.

175. 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 Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

176. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). Panther: a library of protein families and subfamilies indexed by function. Genome Res. *13*, 2129–2141. https://doi.org/10.1101/gr.772403.

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

178. Schlösser, U.G. (1994). SAG - Sammlung von Algenkulturen at the university of Göttingen catalogue of strains 1994. Bot. Acta *107*, 113–186. https://doi.org/10.1111/j.1438-8677.1994.tb00784.x.

179. 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. https://doi.org/10.1111/nph.14321.

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

181. Parker, B.C. (1965). Non-calcareous marine algae from California Miocene deposits. Nova Hedwigia *10*, 273.

182. 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. Mol. Phylogenet. Evol. *168*, 107408. https://doi.org/10.1016/j.ympev.2022.107408.

183. 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. https://doi.org/10.1371/journal.pbio.3000862.

184. 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. https://doi.org/10.1101/2023.07.10.548381.

185. 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. https://doi.org/10.1128/mSphereDirect.00069-18.

186. Maumus, F., Epert, A., Nogué, F., and Blanc, G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. Nat. Commun. *5*, 4268. https://doi.org/10.1038/ncomms5268.

187. 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 Res. *46*, D633–D639. https://doi.org/10.1093/nar/gkx935.

188. 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. https://doi.org/10.1371/journal.pcbi.1006146.

189. 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. https://doi.org/10.1186/1756-8935-5-7.

190. 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*, 1–106. https://doi.org/10.1111/j.0022-3646.1993.00001.x.

191. 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. *71*, 366–377. https://doi.org/10.1111/j.1365-313X.2012.04982.x.

192. 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. USA *118*, e2023575118. https://doi.org/10.1073/pnas.2023575118.

193. 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. https://doi.org/10.1098/rspb.2014.0732.

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

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

196. Subramanian, S. (2008). Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics *178*, 2429–2432. https://doi.org/10.1534/genetics.107.086405.

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

198. 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. https://doi.org/10.1007/s00239-019-09921-4.

199. 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. https://doi.org/10.1101/gr.113985.110.

200. 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 Res. *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.

201. 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. https://doi.org/10.3390/genes13061090.

202. 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 Evol. Biol. *10*, 210. https://doi.org/10.1186/1471-2148-10-210.

203. 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. https://doi.org/10.1126/science.1188021.

204. 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. https://doi.org/10.1093/molbev/msr048.

205. 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. https://doi.org/10.1534/genetics.112.145037.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Descriptions of all sequenced samples have been deposited in the EBI/ENA database	This study.	EBI/ENA project PRJEB76691
Critical commercial assays		
OmniPrep Genomic DNA Purification Kit	G Biosciences, St. Louis, MO, USA	N/A
Nucleospin Plant II midi DNA Extraction Kit	Macherey-Nagel, Düren, Germany	N/A
NEBNext DNA Modules Products	New England Biolabs, Ipswich, MA, USA	N/A
NEBNext Sample Reagent Set	New England Biolabs, Ipswich, MA, USA	N/A
Ampure XP	Beckmann Coulter Genomics, Danvers, MA, USA	N/A
Kapa Hifi Hotstart NGS library Amplification kit	Roche, Basel, Switzerland	N/A
Short Read Eliminator Kit	Pacific Biosciences, Menlo Park, CA, USA	N/A
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	N/A
TruSeq Stranded mRNA Sample Prep	Illumina	N/A
NEBNext Ultra II Directional RNA Library Prep for Illumina	New England BioLabs	N/A
Deposited data		
The sequence data generated by this project is described in Table S1.	This study.	EBI/ENA: PRJEB76691
CNRS Research Data dataset "Data for Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems"	This study.	CNRS Research Data: https://doi.org/10.57745/9U1J85
Experimental models: Organisms/strains		
The strains used for genome and transcriptome sequencing are listed in Table S1A.	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.[82]	RRID:SCR_018551 https://github.com/voutcn/megahit
MetaGene version 2008.8.19	Noguchi et al.[83]	http://metagene.cb.k.u-tokyo.ac.jp/
BLAST	Altschul et al.[84]	RRID:SCR_004870 http://blast.ncbi.nlm.nih.gov/Blast.cgi
Burrows-Wheeler Aligner	Li and Durbin[85]	RRID:SCR_010910 http://bio-bwa.sourceforge.net/
Bowtie2 version 2.3.5.1	Langmead and Salzberg[86]	RRID:SCR_016368 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SPAdes assembler version 3.8.1	Bankevich et al.[87]	RRID:SCR_000131 https://cab.spbu.ru/software/spades/
filtlong	Wick, R.	RRID:SCR_024020 https://github.com/rrwick/Filtlong
Smartdenovo	Liu et al.[88]	RRID:SCR_017622 https://github.com/ruanjue/smartdenovo

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Redbean	Ruan and Li[89]	N/A
Flye	Kolmogorov et al.[90]	RRID:SCR_017016 https://github.com/fenderglass/Flye
Necat	Chen et al.[91]	https://github.com/xiaochuanle/necat
Racon	Vaser et al.[92]	RRID:SCR_017642 https://github.com/isovic/racon
Hapo-G	Aury et al.[93]	https://www.genoscope.cns.fr/hapog/
Metabat 2	Kang et al.[94]	RRID:SCR_019134 https://bitbucket.org/berkeleylab/metabat/src/master/
SortMeRNA	Kopylova et al.[95]	RRID:SCR_014402 http://bioinfo.lifl.fr/RNA/sortmerna/
Velvet version 1.2.07	Zerbino and Birney[96]	RRID:SCR_010755 http://www.molecularevolution.org/software/genomics/velvet
Oases version 0.2.08	Schulz et al.[97]	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.[98]	N/A
Trimmomatic version 0.38 and version 0.39	Bolger et al.[99]	RRID:SCR_011848 http://www.usadellab.org/cms/index.php?page=trimmomatic
Trinity version version 2.6.5	Grabherr et al.[100]	RRID:SCR_013048 http://trinityrnaseq.sourceforge.net/
rnaSPAdes version version 3.13.1	Bushmanova et al.[101]	RRID:SCR_016992 http://cab.spbu.ru/software/rnaspades/
RepeatMasker version 4.1.0	Smit et al.[102]	RRID:SCR_012954 http://repeatmasker.org/
Tandem repeats finder	Benson et al.[103]	RRID:SCR_022193 https://github.com/Benson-Genomics-Lab/TRF
REPET	Flutre et al.[104]	N/A
BLAT	Kent[105]	RRID:SCR_011919 http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Genewise	Birney et al.[106]	RRID:SCR_015054 http://www.ebi.ac.uk/Tools/psa/genewise/
DIAMOND version 0.9.30	Buchfink et al.[107]	RRID:SCR_009457 http://www.nitrc.org/projects/diamond/
Est2Genome	Mott[108]	https://galaxy-iuc.github.io/emboss-5.0-docs/est2genome.html
Gmove	Dubarry et al.[109]	RRID:SCR_019132 http://www.genoscope.cns.fr/gmove
votingLNC	Debit, A.	https://gitlab.com/a.debit/votinglnc
AliView version 1.26	Larsson[110]	RRID:SCR_002780 https://github.com/AliView
RAxML version 8.2.	Stamatakis[111]	RRID:SCR_006086 https://github.com/stamatak/standard-RAxML
Tracer version 1.7.2	Rambaut et al.[112]	RRID:SCR_019121 https://bioweb.pasteur.fr/packages/pack@Tracer@v1.6
OrthoFinder version 2.5.2	Emms and Kelly[113]	RRID:SCR_017118 https://github.com/davidemms/OrthoFinder
Count version 9.1106	Csűös[114]	https://www.iro.umontreal.ca/∼csuros/gene_content/count.html
MUSCLE version 3.8.1551	Edgar[115]	RRID:SCR_011812 http://www.ebi.ac.uk/Tools/msa/muscle/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
OD-Seq version 1.0	Jehl et al.[116]	https://bioconductor.org/packages/release/bioc/manuals/odseq/man/odseq.pdf
HMMER3 package versions 3.1b1 and 3.3.2	Mistry et al.[117]	RRID:SCR_005305 http://hmmer.janelia.org/
GenEra	Barrera-Redondo et al.[118]	N/A
MCL	Enright et al.[119]	RRID:SCR_024109 https://micans.org/mcl/
Foldseek	Kempen et al.[120]	https://search.foldseek.com/search
CleanBlastp	Pathmanathan et al.[121]	N/A
SEED	Overbeek et al.[122]	RRID:SCR_002129 http://www.theseed.org/wiki/Home_of_the_SEED
IPR2GO	Paysan-Lafosse et al.[123]	http://www.ebi.ac.uk/interpro/search/sequence-search
eggNOG	Huerta-Cepas et al.[124]	RRID:SCR_002456 http://eggnog.embl.de
eggNOG-mapper	Cantalapiedra et al.[125]	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 version 2.6.3	Hyatt et al.[126]	RRID:SCR_011936 https://github.com/hyattpd/Prodigal
ViralRecall version 2.0	Aylward et al.[127]	https://github.com/faylward/viralrecall
esl-translate version 0.48	Rivas, E.	https://github.com/EddyRivasLab/easel/blob/master/miniapps/esl-translate.man.in
bedtools version 2.29.2	Quinlan and Hall[128]	RRID:SCR_006646 https://github.com/arq5x/bedtools2
MMseqs cluster version 13.45111	Hauser et al.[129]	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[130]	RRID:SCR_011811 http://mafft.cbrc.jp/alignment/server/
MEGA	Tamura et al.[131]	RRID:SCR_023017 https://www.megasoftware.net/
NGphylogeny platform	Lemoine et al.[132]	https://ngphylogeny.fr/.
TrimAl	Capella-Gutiérrez et al.[133]	RRID:SCR_017334 http://trimal.cgenomics.org/
TAPscan version 4	Petroll et al.[134,135]	https://plantcode.cup.uni-freiburg.de/tapscan/
Expasy web translator	Duvaud et al.[136]	RRID:SCR_024703 https://web.expasy.org/translate/
Geneious versions 11.0.5 and 11.1.5	Geneious	RRID:SCR_010519 http://www.geneious.com/
Interproscan 94.0	Jones et al.[137]	RRID:SCR_005829 http://www.ebi.ac.uk/Tools/pfa/iprscan/
Clustal 2.1	Thompson et al.[138]	RRID:SCR_001591 http://www.ebi.ac.uk/Tools/msa/clustalo/
Gblocks	Castresana[139]	RRID:SCR_015945 http://molevol.cmima.csic.es/castresana/Gblocks_server.html
Kallisto version 0.44.0.	Bray et al.[140]	RRID:SCR_016582 https://pachterlab.github.io/kallisto/about
Deseq2	Love et al.[141]	RRID:SCR_015687 https://bioconductor.org/packages/release/bioc/html/DESeq2.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FastQC	Andrews[142]	RRID:SCR_014583 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore version 0.6.5	Krueger et al.[143]	RRID:SCR_011847 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
HISAT2 version 2.1.0	Kim et al.[144]	RRID:SCR_015530 http://ccb.jhu.edu/software/hisat2/index.shtml
featureCounts	Liao et al.[145]	RRID:SCR_012919 http://bioinf.wehi.edu.au/featureCounts/
PAML version 4.9i (including MCMCTree)	Yang[146]	RRID:SCR_014932 http://abacus.gene.ucl.ac.uk/software/paml.html
phytools R package	Revell[147]	RRID:SCR_015502 https://cran.r-project.org/web/packages/phytools/index.html
VHICA package	Wallau et al.[148]	https://github.com/cran/vhica
NOVOPlasty version 3.7	Dierckxsens et al.[149]	RRID:SCR_017335 https://github.com/ndierckx/NOVOPlasty
SAMtools version 1.5	Li et al.[150]	RRID:SCR_002105 http://htslib.org/
GeSeq version 2.03	Tillich et al.[151]	RRID:SCR_017336 https://chlorobox.mpimp-golm.mpg.de/geseq.html
ARAGORN version 1.2.38	Laslett and Canback[152]	RRID:SCR_015974 http://mbio-serv2.mbioekol.lu.se/ARAGORN/
ModelFinder	Kalyaanamoorthy et al.[153]	http://www.iqtree.org/ModelFinder/
UFBoot2	Hoang et al.[154]	N/A
SynMap	Haug-Baltzell et al.[155]	https://genomevolution.org/SynMap.pl
DAGChainer	Haas et al.[156]	https://dagchainer.sourceforge.net/
CodeML	Yang et al.[146]	N/A
nwalign	Pedersen, B	https://pypi.org/project/nwalign/
BEAST version 2.7	Bouckaert et al.[157]	RRID:SCR_010228 http://beast.bio.ed.ac.uk/
StarBEAST3 version 1.1.7	Douglas et al.[158]	https://github.com/rbouckaert/starbeast3
bModelTest	Bouckaert et al.[159]	N/A
LogCombiner version 2.4.7	Bouckaert et al.[157]	N/A
TreeAnnotator version 2.4.7	Bouckaert et al.[157]	N/A
SplitsTree 4 version 4.14.6	Kloepper and Huson[160]	RRID:SCR_014734 http://www.splitstree.org/
Hectar	Gschloessl et al.[161]	https://webtools.sb-roscoff.fr/root?tool_id=abims_hectar
RShiny	R Core Team[162]	https://github.com/rstudio/shiny
IQ-TREE 2	Minh et al.[163]	https://github.com/iqtree/iqtree2
Computational analysis of gene family evolution 5 (CAFE5)	Mendes et al.[164]	https://github.com/hahnlab/CAFE5
clusterProfiler	Yu et al.[165]	RRID:SCR_016884 http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
ggplot2	Wickham et al.[166]	RRID:SCR_014601 https://cran.r-project.org/web/packages/ggplot2/index.html
tidyverse	Wickham et al.[167]	RRID:SCR_019186 https://CRAN.R-project.org/package=tidyverse
Other		
Benchmarking universal single-copy orthologue (BUSCO) analysis version 5, eukaryota_odb10	Manni et al.[168]	RRID:SCR_015008 http://busco.ezlab.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
UniRef90	Suzek et al.[169]	RRID:SCR_010646 http://www.uniprot.org/help/uniref
AlphaFold protein structure database	Varadi et al.[170]	RRID:SCR_023662 https://alphafold.ebi.ac.uk/
NCVOG database	Yutin et al.[171]	N/A
VOGDB database	Trgovec-Greif et al.[172]	https://vogdb.org/
SulfAtlas database	Barbeyron et al.[172]; Stam et al.[173]	https://sulfatlas.sb-roscoff.fr/
Pfam	Mistry et al.[174]	RRID:SCR_004726 http://pfam.xfam.org/
Panther 17.0	Thomas et al.[175]	RRID:SCR_004869 http://www.pantherdb.org/
Simple Modular Architecture Research Tool (SMART)	Letunic et al.[176]	RRID:SCR_005026 http://smart.embl.de/

EXPERIMENTAL MODEL DETAILS

Ascophyllum nodosum

Species: *Ascophyllum nodosum*
Strain: field collected sperm cells
Genotype: diploid
Sex: male
Maintenance: N/A

Chordaria linearis strain ClinC8C

Species: *Chordaria linearis*
Strain: ClinC8C
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Choristocarpus tenellus strain KU-1152

Species: *Choristocarpus tenellus*
Strain: KU-1152
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Chrysoparadoxa australica strain CS-1217

Species: *Chrysoparadoxa australica*
Strain: CS-1217
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Cladosiphon okamuranus strain S-strain

Species: *Cladosiphon okamuranus*
Strain: S-strain
Genotype: diploid
Sex: n/a
Maintenance: N/A

Desmarestia dudresnayi strain DdudBR16

Species: *Desmarestia dudresnayi*
Strain: DdudBR16
Genotype: haploid

Sex: monoicous
Maintenance: Maintained in culture

Desmarestia herbacea strain DmunF

Species: *Desmarestia herbacea*
Strain: DmunF
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Desmarestia herbacea strain DmunM

Species: *Desmarestia herbacea*
Strain: DmunM
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Dictyota dichotoma strain KB07f IV

Species: *Dictyota dichotoma*
Strain: KB07f IV
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Dictyota dichotoma strain ODC1387m

Species: *Dictyota dichotoma*
Strain: ODC1387m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Dictyota dichotoma strain KB07m IV

Species: *Dictyota dichotoma*
Strain: KB07m IV
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Dictyota dichotoma strain KB07sp VI

Species: *Dictyota dichotoma*
Strain: KB07sp VI
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Discosporangium mesarthrocarpum strain MT17-79

Species: *Discosporangium mesarthrocarpum*
Strain: MT17-79
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Ectocarpus crouaniorum strain Ec861

Species: *Ectocarpus crouaniorum*
Strain: Ec861
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus crouaniorum strain Ec862

Species: *Ectocarpus crouaniorum*
Strain: Ec862
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus fasciculatus strain Ec846

Species: *Ectocarpus fasciculatus*
Strain: Ec846
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus fasciculatus strain Ec847

Species: *Ectocarpus fasciculatus*
Strain: Ec847
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus fasciculatus strain EfasUO1

Species: *Ectocarpus fasciculatus*
Strain: EfasUO1
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus fasciculatus strain EfasUO2

Species: *Ectocarpus fasciculatus*
Strain: EfasUO2
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus siliculosus strain Ec863

Species: *Ectocarpus siliculosus*
Strain: Ec863
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus siliculosus strain Ec864

Species: *Ectocarpus siliculosus*
Strain: Ec864
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 1 strain Ec sil Puy CHCH Z9 G5f

Species: *Ectocarpus species 1*
Strain: Ec sil Puy CHCH Z9 G5f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus species 1 strain Ec sil Puy CHCH Z9 G3m

Species: *Ectocarpus* species 1
Strain: Ec sil Puy CHCH Z9 G3m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 1 strain Ec03

Species: *Ectocarpus* species 1
Strain: Ec03
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 12 strain Ec fas CH92 Nie 2f

Species: *Ectocarpus* species 12
Strain: Ec fas CH92 Nie 2f
Genotype: diploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus species 12 strain Ec fas CH92 Nie 3m

Species: *Ectocarpus* species 12
Strain: Ec fas CH92 Nie 3m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus species 13 strain EcNAP12-S#4-19m

Species: *Ectocarpus* species 13
Strain: EcNAP12-S#4-19m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus species 2 strain Ec06

Species: *Ectocarpus* species 2
Strain: Ec06
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 3 strain Ec10

Species: *Ectocarpus* species 3
Strain: Ec10
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus species 3 strain Ec11

Species: *Ectocarpus* species 3
Strain: Ec11
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 5 strain Ec13

Species: *Ectocarpus* species 5
Strain: Ec13
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus species 5 strain Ec12

Species: *Ectocarpus* species 5
Strain: Ec12
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Ectocarpus species 6 strain EcLAC-371f

Species: *Ectocarpus* species 6
Strain: EcLAC-371f
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus species 7 strain Ec32

Species: *Ectocarpus* species 7
Strain: Ec32
Genotype: haploid
Sex: male
Maintenance: N/A

Ectocarpus species 8 strain EcLAC-412m

Species: *Ectocarpus* species 8
Strain: EcLAC-412m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Ectocarpus species 9 strain EcSCA-722f

Species: *Ectocarpus* species 9
Strain: EcSCA-722f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Ectocarpus subulatus strain Bft15b

Species: *Ectocarpus subulatus*
Strain: Bft15b
Genotype: haploid
Sex: male
Maintenance: N/A

Feldmannia mitchelliae strain KU-2106 Giff mitch BNC GA

Species: *Feldmannia mitchelliae*
Strain: KU-2106 Giff mitch BNC GA
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Fucus distichus

Species: *Fucus distichus*
Strain: field collected meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

Fucus serratus

Species: *Fucus serratus*
Strain: field collected ovule cells
Genotype: diploid
Sex: female
Maintenance: N/A

Fucus serratus

Species: *Fucus serratus*
Strain: field collected sperm cells
Genotype: diploid
Sex: male
Maintenance: N/A

Halopteris paniculata strain Hal grac a UBK

Species: *Halopteris paniculata*
Strain: Hal grac a UBK
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Hapterophycus canaliculatus strain Oshoro5f

Species: *Hapterophycus canaliculatus*
Strain: Oshoro5f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Hapterophycus canaliculatus strain Oshoro7m

Species: *Hapterophycus canaliculatus*
Strain: Oshoro7m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Hapterophycus canaliculatus strain Oshoro 3F x 9M

Species: *Hapterophycus canaliculatus*
Strain: Oshoro 3F x 9M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Hapterophycus canaliculatus strain Oshoro 4F x 9M

Species: *Hapterophycus canaliculatus*
Strain: Oshoro 4F x 9M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Hapterophycus canaliculatus strain Oshoro 6F x 6M

Species: *Hapterophycus canaliculatus*
Strain: Oshoro 6F x 6M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Heribaudiella fluviatilis strain SAG. 13.90

Species: *Heribaudiella fluviatilis*
Strain: SAG. 13.90
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Heterosigma akashiwo strain CCMP452

Species: *Heterosigma akashiwo*
Strain: CCMP452
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Himanthalia elongata

Species: *Himanthalia elongata*
Strain: field meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

Laminaria digitata strain LdigPH10-18mv

Species: *Laminaria digitata*
Strain: LdigPH10-18mv
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Laminarionema elsbetiae strain ELsaHSoW15

Species: *Laminarionema elsbetiae*
Strain: ELsaHSoW15
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Macrocystis pyrifera strain P11A1

Species: *Macrocystis pyrifera*
Strain: P11A1
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Macrocystis pyrifera strain P11B4

Species: *Macrocystis pyrifera*
Strain: P11B4
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Myriotrichia clavaeformis strain Myr cla04

Species: *Myriotrichia clavaeformis*
Strain: Myr cla04
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Myriotrichia clavaeformis strain Myr cla05

Species: *Myriotrichia clavaeformis*
Strain: Myr cla05
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Myriotrichia clavaeformis strain Myr cla12

Species: *Myriotrichia clavaeformis*
Strain: Myr cla12
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Pelvetia canaliculata

Species: *Pelvetia canaliculata*
Strain: field collected meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

Phaeothamnion wetherbeei strain SAG 119.79

Species: *Phaeothamnion wetherbeei*
Strain: SAG 119.79
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Pleurocardia lacustris strain SAG 25.93

Species: *Pleurocardia lacustris*
Strain: SAG 25.93
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Porterinema fluviatile strain SAG 2381

Species: *Porterinema fluviatile*
Strain: SAG 2381
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Pylaiella littoralis strain U1.48

Species*: Pylaiella littoralis*
Strain: U1.48
Genotype: haploid
Sex: unknown
Maintenance: Maintained in culture

Pylaiella littoralis strain F24

Species: *Pylaiella littoralis*
Strain: F24
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Saccharina japonica strain Ja

Species: *Saccharina japonica*
Strain: Ja
Genotype: haploid
Sex: male
Maintenance: N/A

Saccharina latissima strain SLPER63f7

Species: *Saccharina latissima*
Strain: SLPER63f7
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Saccorhiza dermatodea strain SderLü1190fm

Species: *Saccorhiza dermatodea*
Strain: SderLü1190fm
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Saccorhiza polyschides strain SpolBR94f

Species: *Saccorhiza polyschides*
Strain: SpolBR94f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Saccorhiza polyschides strain SpolBR94m

Species: *Saccorhiza polyschides*
Strain: SpolBR94m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Saccorhiza polyschides

Species: *Saccorhiza polyschides*
Strain: field collected sample (young sporophytes ∼2-10cm)
Genotype: diploid
Sex: n/a
Maintenance: N/A

Sargassum fusiforme

Species: *Sargassum fusiforme*
Strain: unknown
Genotype: diploid
Sex: n/a
Maintenance: N/A

Schizocladia ischiensis strain KU-0333

Species: *Schizocladia ischiensis*
Strain: KU-0333
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Scytosiphon promiscuus strain 000310-Muroran-5-female

Species: *Scytosiphon promiscuus*
Strain: 000310-Muroran-5-female
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Scytosiphon promiscuus strain Ot110409-Otamoi-16-male

Species: *Scytosiphon promiscuus*
Strain: Ot110409-Otamoi-16-male
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Scytosiphon promiscuus strain SXS107

Species: *Scytosiphon promiscuus*
Strain: SXS107
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Sphacelaria rigidula strain Sph rig Cal Mo 4-1-68b

Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo 4-1-68b
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Sphacelaria rigidula strain Sph rig Cal Mo 4-1-G3b

Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo 4-1-G3b
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Sphacelaria rigidula strain Sph rig Cal Mo SP

Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo SP
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Sphaerotrichia firma strain ET2f

Species: *Sphaerotrichia firma*
Strain: ET2f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Sphaerotrichia firma strain Sfir13m

Species: *Sphaerotrichia firma*
Strain: Sfir13m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Tribonema minus strain UTEX B 3156

Species: *Tribonema minus*
Strain: UTEX B 3156
Genotype: unknown
Sex: unknown
Maintenance: N/A

Undaria pinnatifida strain Kr2015

Species: *Undaria pinnatifida*
Strain: Kr2015
Genotype: diploid
Sex: n/a
Maintenance: N/A

METHOD DETAILS

Biological material

Sequencing brown algal genomes 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. To address these problems, cultured, unialgal filamentous gametophyte material was used whenever possible (i.e. for species with haploid-diploid life cycles) and the extraction methodology was adapted for each species.

The algal strains analysed in this study are listed in Table S1A, which provides information about the sampling site for each strain. The sampling sites are shown on a world map in Figure S1D.

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 Provasoli-enriched[104] 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_medi a/01%20Basal%20Medium.pdf) micronutrients (at 20°C for *P. lacustris*) and *Phaeothamnion wetherbeei*, which was grown in MIEB12 (medium 7 in Letunic et al.[177]). Whole thallus was extracted for all species except the Fucales, where either dissected meristematic regions or released male gametes were extracted. Tissue samples were frozen in liquid nitrogen and stored at -80°C before extraction.

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 the DNA purified using Ampure XP (Beckmann Coulter Genomics, Danvers, MA, USA). Adapted fragments were amplified with 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: 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 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[82] 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[83] 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[85] (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[86] 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[87] 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. 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,[88] Redbean,[89] Flye[90] and Necat.[91] Based on the cumulative size and contiguity, we selected the best assembly for each strain. This assembly was then polished three times using Racon[92] with nanopore reads, and twice with Hapo-G[93] and Illumina PCR-free reads.

Assembly decontamination

Contigs from the short- and long-read genome assemblies were 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)[94] 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[95] and filtered out. The Illumina RNA-seq short reads from each strain were assembled using Velvet[96] version 1.2.07 and Oases[97] version 0.2.08 with kmer sizes of 61, 63 and 65 bp. BUSCO[168] 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,[98] 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 to remove poor quality sequence and adapter sequences using Trimmomatic[99] v0.39 prior to being assembled using either Trinity[100] version v2.6.5 or rnaSPAdes[101] 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. The *de novo* transcriptomes are available from the CNRS Research Data dataset (https://doi.org/10.57745/9U1J85) and from the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/).

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*)[11] and using RepBase with RepeatMasker[102] version v.4.1.0, default parameters. Tandem repeats finder (TRF)[103] was also used to mask tandem repeat duplications. In addition, transposons were annotated in ten species using REPET[104] and the transposons detected were used as a reference to mask all genomes with RepeatMasker[102] 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)[178] and UniRef90 (https://www.uniprot.org/uniref/) were aligned against each genome assembly. First, BLAT[105] 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 $\geq$ 90% of the best match score were retained. Second, the alignments were refined using Genewise[106] 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[107] (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[105] 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[108] 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.[109] 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[178] (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[168] (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.[178] The correlation between orthologous CDS lengths was higher for genomes sequenced with long reads than for genomes only sequenced with short reads (Figure S1B). 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[125] 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[137] analyses (versions v5.55-88.0, v5.51-85.0 or v5.36-75.0) to detect functional domains, Hectar[161] (v1.3) predictions of protein subcellular localisation and various DIAMOND[107] (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 $\leq 10-20$ 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 contingency 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 S1.

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 S1B (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 $\geq$1000 bp, 4) opposite strand genes, intergenic $\geq$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[162] (version 4.3.0).

Detection of long non-coding RNAs

Transcriptome data for 11 species (Table S1F), 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.[179] 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 $\geq$ 100aa were discarded. lncRNA transcripts that had significant matches in either the Pfam[174] (hmmscan e-value < 0.001) or SwissProt (BLASTp e-value < $1e^{-5}$ and similarity $\geq$ 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[166] (version 3.4.0) for plotting.

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 S1F) 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[115] version 3.8.1551 with default parameters and conserved blocks were identified with Gblocks[139] 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 S4B. 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. The number of introns per gene in brown algae and in closely-related outgroup species were compared using a contingency test (Table S1C).

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.[180] 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[110] 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[111] v.8.2.9 and the gamma model. The best-fit evolutionary model for each gene was determined using AIC.

Bayesian divergence time estimation for the brown algae

An estimation of brown algal divergence time was carried out using the 32 orthologous nuclear genes (see above and) for 51 brown algae and five non-brown species (16,185 amino acids, 56 spp.) and MCMCTree (PAML package v4.9j) with the approximate likelihood method. The WAG protein model was selected based on the AIC and BIC criteria of ModelFinder.[153] The independent clock model was selected based on previous work on the brown algal timeline by Choi et al.[1] One hundred million years was set to correspond to 1 in the MCMCTree calculation. A secondary calibration for the root was based on Choi et al.[1] using a gamma distribution of 70.2 alpha and 10.22 beta. A kelp holdfast fossil[55] was used to date the crown node of kelps with a minimum bound of 0.31, and a *Julescraneia* fossil[181] for the *Macrocystis*/*Saccharina* clade with a minimum bound of 0.13 (Figure S2A). MCMC chains were run 1.5 million generations, with the first 200,000 MCMC chains being discarded as burn-in, and the convergence of MCMC chains was checked with Tracer v1.7.2.[112] This analysis estimated that Schizocladiophyceae and brown algae diverged 457.88 Mya (95% HPD: 321.29-592.66 Ma), similar to (about 8 Mya older than) the previous estimate using plastid genes[1] and that diversification of the major brown algal linages began about 220 million years later, after the origin of DFI clade (235.97 Mya, 95% HPD: 158.88-312.48 Mya), about 12 Ma earlier than the previous estimate.[1] The fossil-calibrated phylogenetic tree for 11 *Ectocarpus* species (Figure S2C) was extracted from the brown algal tree (Figure S2A).

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[113] 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. The orthogroups contain 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[114] 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.[182]

Phylostratigraphy analysis

GenEra[118] was used to estimate gene family founder events for each genome assembly by running DIAMOND[107] 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.[119] Estimated evolutionary distances were extracted for each pair of species from the maximum likelihood species tree (substitutions/site) to calculate homology detection failure probabilities.[183] 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.[7,179] Structural alignments were performed using Foldseek[120] against the AlphaFold protein structure database.[170]

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 closely-related outgroup species (*Schizocladia ischiensis*, *Tribonema minus*, *Chrysoparadoxa australica* and *Heterosigma akashiwo*; Table S1F). 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[115] version 3.8.1551 with default parameters and the multiple alignments were filtered using OD-Seq[116] 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[117] 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.[107] 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 S3).

The amplified gene families were manually categorised into functional classes based on the output of automatic functional annotation programs (InterProScan,[137] EggNOG,[124] 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 S3).

In addition to the binomial tests, we also ran CAFE5[164] to reconstruct the history of gene family amplifications. Such reconstructions rely on a species tree and require that all gene families are present at the root of the tree. However, of the 180 amplified OGs that were strongly amplified in Phaeophyceae (see above and listed in Table S3) only 19 were present at the ancestral node. The majority (161) of the 180 families were gained during the early evolution of the lineage, most (105) at the origin of the PX clade (i.e. a collapsed node corresponding to nodes n1 and n2 in Figure S2B) or of the Phaeophyceae/FDI clades (i.e. a collapsed node corresponding to nodes n5 and n6; Figure S2B). To determine whether the 180 amplified OGs were significantly enriched in genes that were gained early during Phaeophyceae evolution (i.e. at nodes n1/n2, n4, n5/n6 in Figure S2B), a Chi-squared test was carried out using the R *chisq.test* function on a contingency table containing the proportions of OGs gained at various periods during brown algal evolution for both the amplified OGs and for the entire set of OGs as a reference dataset (Table S1C). Twelve independent CAFE5 reconstructions were carried out on the OG subsets gained at 12 different nodes (n0, n1/2, n4, n5/n6, n8, n9, n10/n11, n13, n15, n18, n19, n20), using the subtrees rooted at these nodes so that the sets of OGs gained at each node would be placed at the root of the tree for one of the 12 analyses (Figure S3E). The analysis focused on the 19 highest quality genomes (Table S1F), which is why some pairs of nodes were collapsed (e.g. nodes n1 and n2 to give n1/n2). Several parameters were tested for CAFE5: the –p option (Poisson distribution) resulted in better likelihood scores than default, but we observed a weak effect when increasing the value of lambda (–k). Consequently, all reconstructions were performed with –p (and no k, i.e. k=1) for efficiency purposes. As recommended by Mendes et al.,[164] very large gene families were discarded as these can cause the program to fail to initialize the parameters. The twelve reconstructions were then aggregated and the proportions of amplified and reduced gene families were calculated for each node (Table S3). Only results on internal nodes were considered, since leaves are more subject to artefactual amplifications/reductions due to genes being missed, fused or split in the annotations.

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[107] version 2.0.11; ‘‘very-sensitive’’ mode; e-value threshold 1e^{-5}. This raw alignment was then filtered using CleanBlastp, from the CompositeSearch suite,[121] 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.[113] 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,[184] which uses information from extant genomes to apply a Dollo parsimony model in Count,[114] 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 $\geq$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[84] 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[115] version 3.8.31 with default settings. The resulting alignments, trimmed to a minimum length of 50 amino acids using TrimAl[133] 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 $\geq$0.70 for the SH-test and aByes-test support from at least two Phaeoexplorer receptor nodes and three donor supporting nodes. Furthermore, queries that displayed significantly different amino acid compositions ($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 the e-value of the best BLAST hit in the OUTGROUP lineage (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[124] (http://eggnogdb.embl.de/#/app/home) and Pfam.[175] 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 transcripts per kilobase million (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/).

Comparative analysis of gene sets identified by genome-wide analyses of evolutionary history

Genes identified as belonging to orthogroups that were predicted to be gained at specific nodes of the phylogenetic tree based on the Dollo parsimony analysis, to belong to either significantly amplified gene families (binomial analysis) or to belong to gene families that have significantly changed in size over evolutionary time (CAFE5 analysis), to correspond to founder events (Phylostratigraphy analysis), to have been remodelled (composite gene analysis) or to have been derived from an HGT (HGT analysis) were extracted from the output of each of these analysis and aggregated in a single datatable. Correspondences were established manually between phylogenetic tree nodes and phylostrata and this information was integrated into the datatable. Counting and calculations of the frequency of events at specific time points were carried out using *ad hoc* R scripts (R version 4.4.1) and the tidyverse[167] package (version 2.0.0). Graphs were generated using the ggplot2[166] package (version 3.5.1). For each gene, a COG functional category was retrieved from the eggNOG mapper output and the COG enrichment analysis was carried out in R using the clusterProfiler[165] package (version 4.6.2) by comparing each set of gene families with the full set of gene families.

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[126] (V2.6.3, settings: default, meta) and the resulting proteins were used as queries against the NCVOG[171] and VOGDB[185] 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[127] 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[128] (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[109]), 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[129] (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. We noted that the median number of viral regions found in genomes assembled with long reads was very similar to that for genomes assembled with short reads (9 and 10, respectively). 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.

To classify genes in VRs (Figure 6B), viral sequences were removed for the NBCI RefSeq non-redundant protein database (NR) by removing proteins assigned to the "Viruses" category and by comparing the database with RVDB using BLASTp and removing any proteins that matched with an e-value cut-off of < 1-e40 to create a "virus-free NR" database. Deduced proteins were then compared with the RVDB and the virus-free NR databases using BLASTp and relative bitscores (rbitscores) were calculated by dividing the BLASTp bitscore for the best match in each database by the query protein's self-hit bitscore.[186] Self-hit scores were acquired by comparing the complete deduced proteomes with themselves using BLASTp. Proteins with a RVDB rbitscore at least 20% greater than its virus-free NR rbitscore were designated as "viral". Proteins with a virus-free NR rbitscore at least 20% greater than its RVDB rbitscore were designated as "cellular" (i.e. corresponding to a gene from a cellular organism). Ambiguous cases without a 20% differential were designated as "viral or cellular" and proteins with no significant matches were designated as ORFans (i.e. unknown proteins).

The presence of host regions flanking the viral regions was evaluated based on the ViralRecall output (Table S5C). The percentages of viral regions with two, one or zero flanking regions (longer than 2 kbp) were 25.8%, 15.0% and 59.2%, respectively (i.e. 40.8% of viral regions had at least one flanking region). Of the viral regions that had two flanking regions, 89.5%, 7.0% and 3.5% had flanking regions with a total length of >200 kbp, between 20 and 200 kbp or between 2 and 20 kbp, respectively. For the viral regions that had one flanking region, the corresponding percentages were 25.3%, 36.7% and 38.0%.

Phylogenetic analysis of viral genes

Amino acid sequences of manually-curated collections of major capsid protein (MCP) and DNA polymerase B proteins were aligned using MAFFT (v7.520, settings: –adjustdirectionaccurately –auto –maxiterate 1000) and phylogenetic trees were generated using IQ-TREE (v 2.2.2.3, settings: -m MFP -B 1000).

Metabolic networks

Genome-scale metabolic networks were reconstructed using AuCoMe[46] version 0.5.1 using the MetaCyc[187] version 26 database. A first dataset, consisting of the 60 species listed in Table S1F (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/). MDS plots were build using the vegan package, version 2.6-4 (https://github.com/vegandevs/vegan) with R 4.1.2,[162] 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 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. These reaction sets were extracted from all the networks using the Aucomana library (https://github.com/PaulineGHG/aucomana). Online wikis (phaeogem, 16bestgem and fwgem) were generated using AuReMe.[188]

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[130] 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.[131] The annotated genes are listed in Table S4B 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[173,174] (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[84] 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,[133] 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 S4C and S4D.

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*. Predicted proteomes were screened using BLASTp[84] and query sequences from *Ectocarpus* species 7 and seven other species from diverse eukaryotic taxa.

Membrane-localised proteins

Membrane protein family genes were identified either by carrying out BLASTp[84] 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 S1F), depending on the size of the gene family.

Manually-curated histidine kinase protein families were aligned with Muscle[115] before phylogenetic tree construction using IQ-TREE 2[163] (version 2.3.4) with automatic model selection and 1000 bootstraps.

Transcription-associated proteins

TAPscan v4[135] was used to analyse the transcription-associated protein (TAP) complements of 21 species. TAPscan[134] 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 S1F) 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[62] 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[62] (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*, *Chordaria linearis*, *Chrysoparadoxa australica*, *Desmarestia herbacea*, *Dictyota dichotoma*, *Discosporangium mesarthrocarpum*, *Ectocarpus crouaniorum*, *Ectocarpus fasciculatus*, *Ectocarpus siliculosus*, *Fucus serratus*, *Heterosigma akashiwo*, *Pleurocladia lacustris*, *Porterinema fluviatile*, *Pylaiella littoralis*, *Saccharina latissima*, *Sargassum fusiform*, *Schizocladia ischiensis*, *Scytosiphon promiscuus*, *Sphacelaria rigidula*, *Tribonema minus* and *Undaria pinnatifida* 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/). 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,[189] 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[137] 94.0 (www.ebi.ac.uk/interpro/search/sequence/).

Species abbreviations used in histone phylogenetic trees are: Atr, *Amborella trichopoda*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Di, *Dictyostellium 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*.

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 S1F). 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)[177] 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[138] 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 S1F) 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.[41] 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 S1F).

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,[190] 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[99] V0.38 and then mapped to the *P. fluviatile* genome using Kallisto[140] version 0.44.0. Differentially expressed genes were identified using the DESeq2 package[141] included in Bioconductor version 3.11, considering genes with an adjusted $p < 0.05$ and a $\log_2$ fold-change > 1 as differentially expressed. 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,[191] 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 S1F). Data quality was assessed with FastQC[142] 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[145] 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[141] 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, $\log_2$(fold change) ≥1, adjusted *p*-value <0.05; sporophyte-biased: mean TPM ≥1 in gametophyte and sporophyte, $\log_2$(fold change) ≤-1, adjusted *p*-value <0.05; gametophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, $\log_2$ (fold change) ≥1, adjusted *p*-value <0.05; sporophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, $\log_2$(-fold change) ≤-1, adjusted p-value <0.05; unbiased genes: mean gametophyte and sporophyte TPMs ≥1, $\log_2$(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 S1F). For the life-cycle-based assessment, the groups were: 30 haploid-diploid species and six diploid species (Table S1F). 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[146] 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.[147,162]

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[192] and also synonymous substitutions are often underestimated due to saturation of synonymous sites, which might in turn impact the omega ratios.[193] 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.[194–196] We inferred codon usage bias by estimating the effective number of codons (ENC) for each species using the enc method from the VHICA package.[148,182] 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[197]). 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.[198] 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[149] v3.7 and *rbcL* and *cox1* nucleotide sequences as seeds. Assembled genomes were checked by aligning reads using Bowtie2[86] v2.3.5.1 and processed with SAMtools[150] v1.5. Annotation of protein-coding genes was performed with GeSeq[151] v2.03. Annotation of tRNAs, tmRNAs and rRNAs was performed with ARAGORN[152] 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[130] v7 (–maxiterate 1000) and then concatenated. Alignment partitions were assigned based on genes. Each of the aligned gene sequences was trimmed with trimAl[133] v1.2 (-automated1). ML phylogenetic trees were constructed with IQ-TREE 2.[163] The protein substitution models in each gene partition were selected using ModelFinder.[153] Statistical support for tree branches was assessed with 1,000 replicates of ultrafast bootstrap (UFBoot2).[154]

Analysis of *Ectocarpus* genome synteny

Global genome synteny analysis was performed using SynMap[155] 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[199] was used as the BLAST algorithm and syntenic gene pairs were identified using DAGChainer[156] 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,[146] 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[133] and then translated back to DNA with backtranseq.[200] 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.[201]

Phylogenetic analysis of *Ectocarpus* species

Phylogenetic analysis was carried out for 11 *Ectocarpus* species plus *Scytosiphon promiscuus* as an outgroup (Table S1F). Of the 933 single-copy orthogroups identified for these 12 species, 261 high-confidence alignments were retained for gene tree and species tree inferences following the removal of low-quality alignments using BMGE.[202] Bayesian inference of the phylogeny of the *Ectocarpus* species complex was performed using BEAST[157] v2.7. The analysis was conducted under the multi-species coalescent (MSC) model, implemented in StarBEAST3[158] 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[159] 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[157] 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[203]: 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 S6. 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 the 261 orthologroups. Significance was detected using a block-jackknifing approach,[203–205] 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[160] (version 4.14.6) with 1000 bootstrap replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

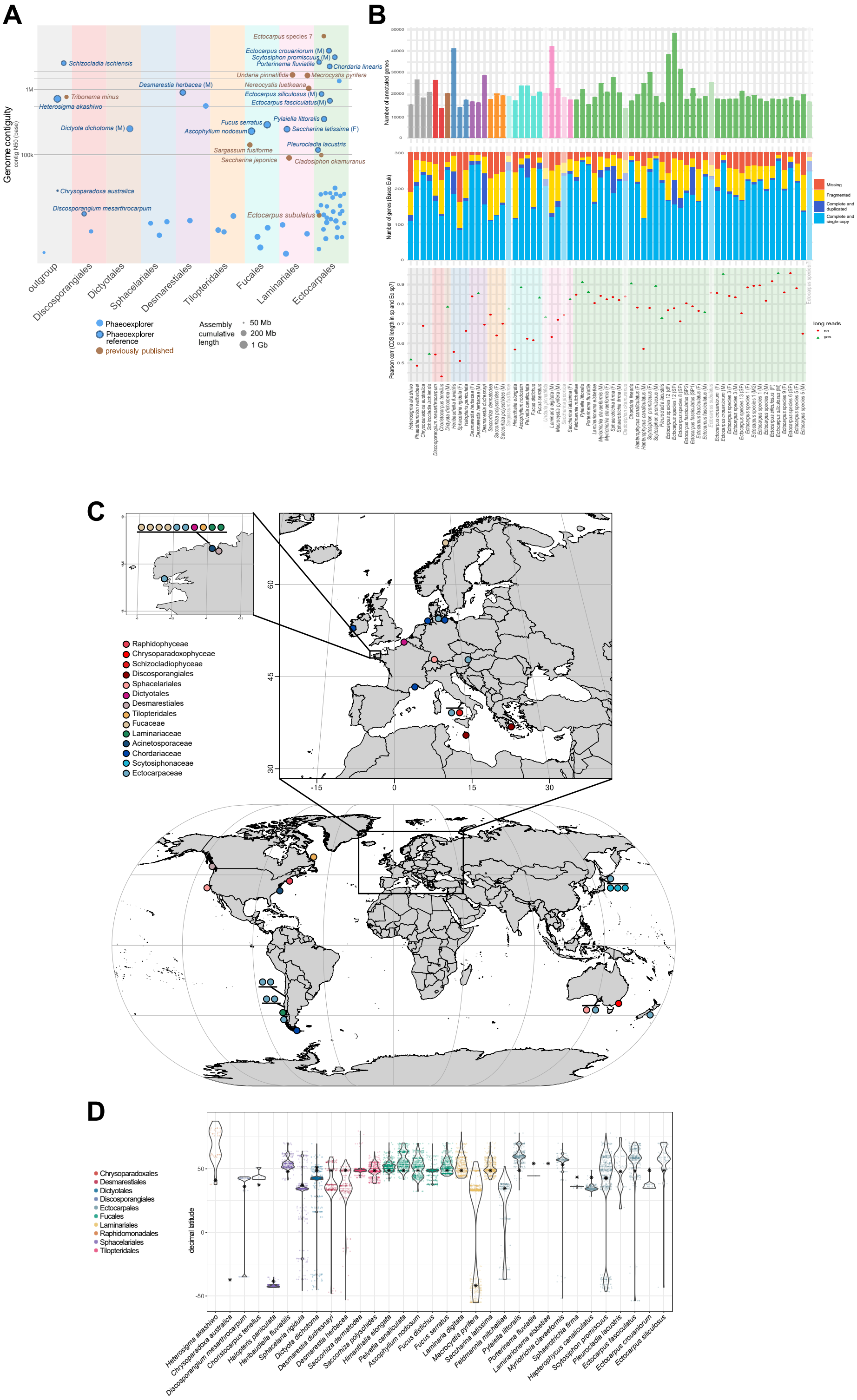
All statistical analyses are described in detail in the relevant sections of the "method details" section and the results of statistical tests are shown in the tables and figures.

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,[113] Interproscan,[137] Hectar[161] and eggNOG-mapper[125] analyses, together with the results of the various DIAMOND[107] 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 AskOmics genomic data query interface (PhaeoAskOmics), an RShiny-based transcriptomic aggregator for the model brown alga *Ectocarpus* species 7 strain Ec32, a link to genome-wide metabolic networks for the Phaeoexplorer species and a list of project-related publications.

Additional data and results have been deposited in the CNRS Research Data depository under the title "Data for Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems" (DOI: https://doi.org/10.57745/9U1J85). Dataset description: "The Phaeoexplorer project sequenced 60 genomes corresponding to 44 brown algal and sister species. This dataset corresponds to supplementary information relating to the initial annotation of the Phaeoexplorer genomes and multiple analyses of the genome data. The dataset includes additional results of the project, together with accompanying additional figures and tables, (Additional_results.tar.gz), presubmission (v0) versions of the Phaeoexplorer genome annotation (GFF) files (GFF_v0.tar.gz) and genome-wide predicted proteomes as fasta files (Proteomes_v0.tar.gz), de novo transcriptome assemblies for the Phaeoexplorer species (RNA-seq data assembled with Trinity or rnaSPAdes; de-novo-transcriptomes.tar.gz), RepeatMasker analyses of repeat sequences (RepeatMasker.tar.gz), alignment files used to generate a phylogenetic tree for the Phaeoexplorer species (PhylogeneticTree.tar.gz), alignments used to build a densitree specifically for *Ectocarpus* species (Microevolution_Ectocarpus.tar.gz), an Orthofinder-based analysis of shared orthologues (Orthogroups.tar.gz) together with a Dollo-logic-based analysis of orthogroup gain and loss during evolution (Dollo_analysis.tar.gz), a Phylostratigraphy analysis of brown algal genes (Phylostratigraphy.tar.gz), an analysis of protein functional domain fissions and fusions (CompositeGenes.tar.gz), Interproscan analyses of protein domains (InterProScan.tar.gz), Hectar predictions of protein subcellular localisations (Hectar.tar.gz), eggNOG output providing information about predicted protein functions (eggNOG.tar.gz), RNA-seq-based data on gene expression levels (mRNAexpression.tar.gz), results of a search for genes acquired via horizontal gene transfer (HGT.tar.gz), analyses of intron conservation across genomes (Introns_conservation.tar.gz), an analysis of tandem gene duplications (Tandemely_duplicated_genes.tar.gz), CAFE5 reconstruction of gene family amplifications (CAFE5.tar.gz), comparisons of CDS size with the *Ectocarpus* reference genome that were used to evaluate gene model completeness (CDS_size.tar.gz), a DESeq2 analysis of differential gene expression between the sporophyte and gametophyte generations of several brown algal species (DEG_LifeCycle.tar.gz), information about orthogroups selected to analyse the effects of morphological complexity and life cycle structure on gene evolution (Genes_selection.tar.gz). Each individual dataset contains a README file explaining its content. Detailed information about the methodology used for each analysis can be found in the STAR Methods section of the manuscript preprint (https://doi.org/10.1101/2024.02.19.579948). The majority of these analyses and datasets can also be accessed via the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/)."

Supplemental figures



(legend on next page)

Figure S1. Taxonomic diversity and assembly quality of the Phaeoexplorer genomes, and geographic localization of species and strains, related to Figure 1

(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) Number of genes annotated in each genome (upper). BUSCO scores for the predicted proteome of each genome (middle). Correlation of coding sequence (CDS) lengths for each species with the corresponding sequences from the *Ectocarpus* species 7 reference genome (lower). Previously published genomes are indicated in gray. longreads, genomes assembled using long reads. F, female; M, male.

(C) World map indicating the positions of the sampling sites for the strains sequenced in this study.

(D) North-south distributions of the species analyzed in this study. The data shown were obtained from the Global Biodiversity Information Facility (gbif.org). Asterisks indicate the sampling latitude for the sequenced strains of each species.

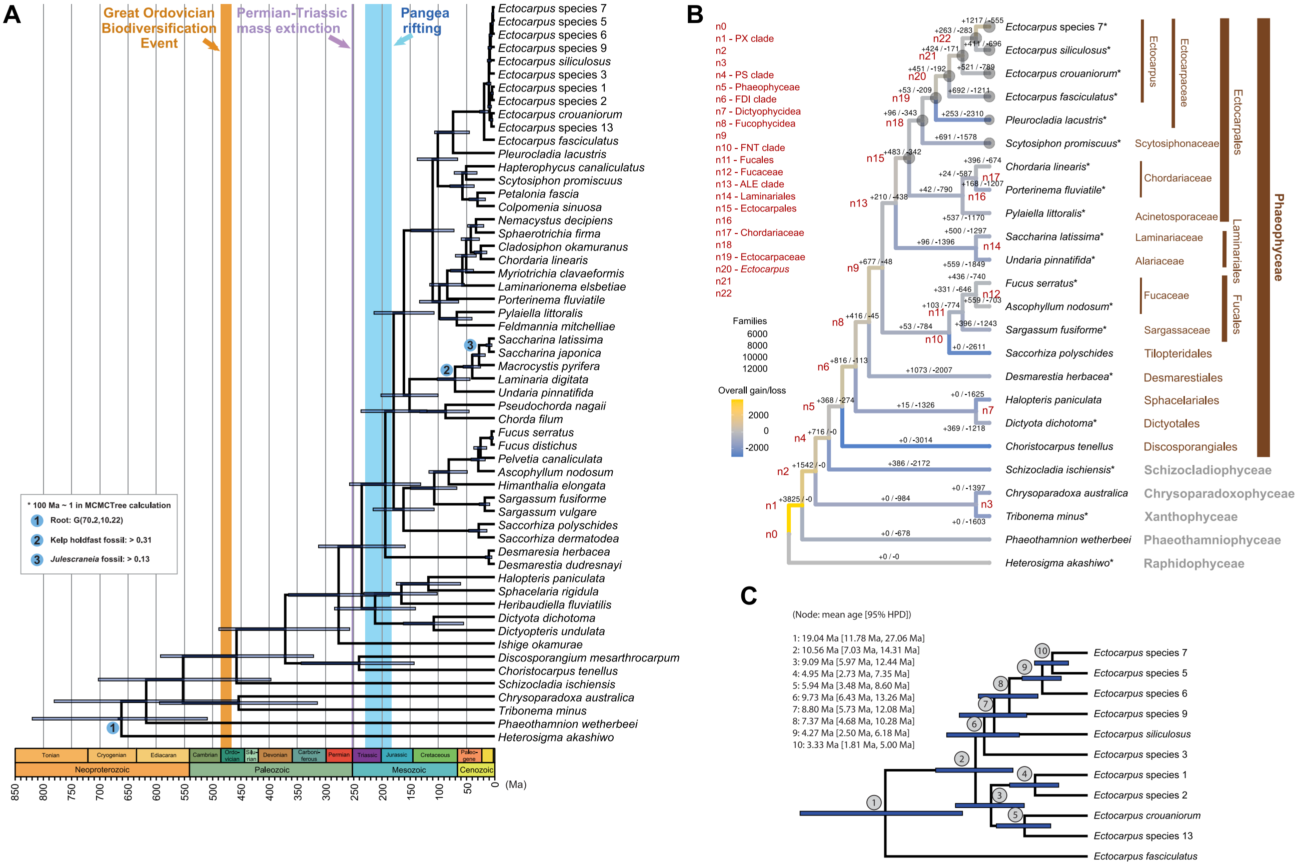
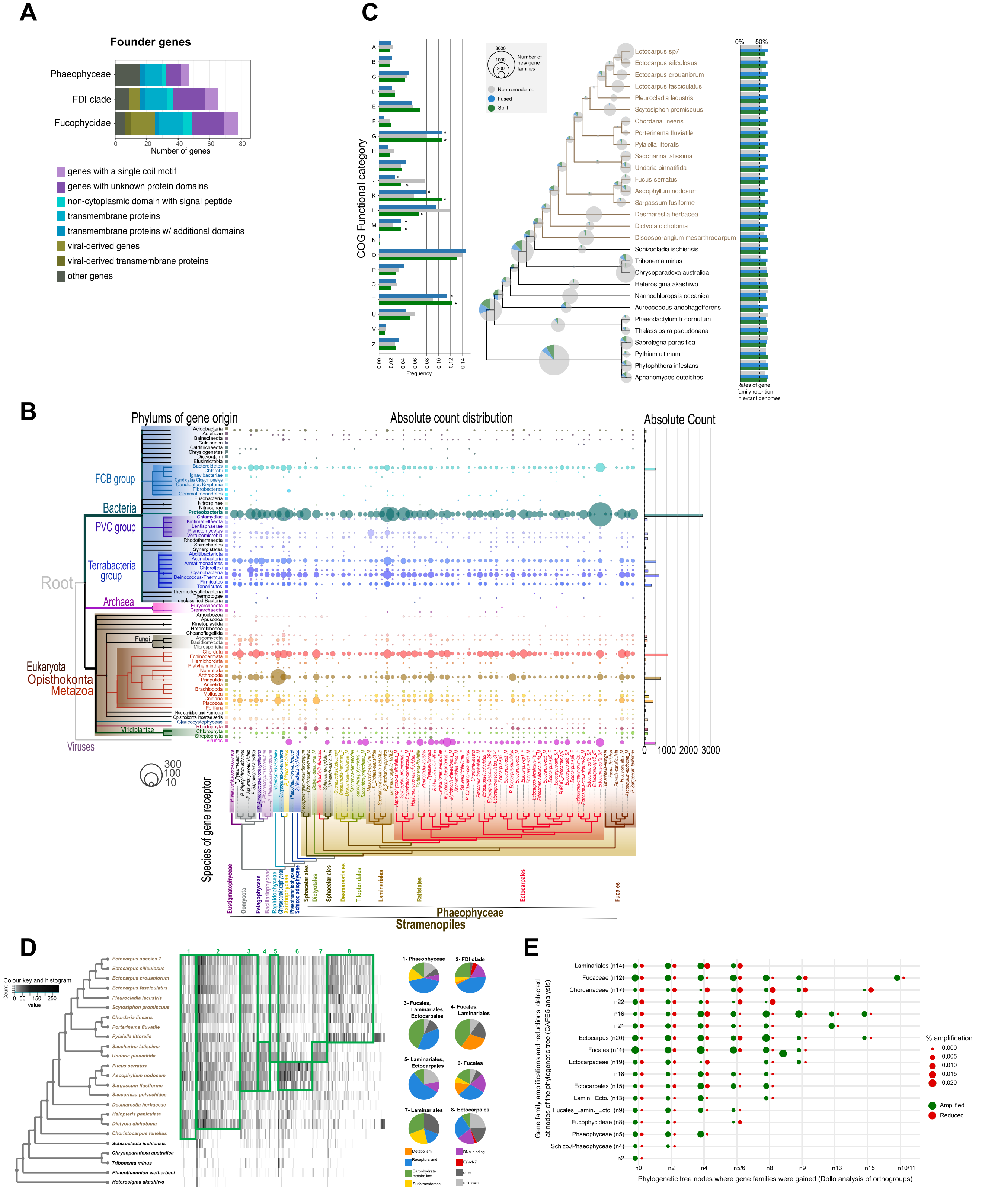


Figure S2. Fossil-calibrated time trees and Dollo logic analysis of orthogroups, related to Figures 1 and 5

(A) Fossil-calibrated maximum-likelihood phylogenetic tree of the brown algal and closely related taxa analyzed in this study. Bayesian divergence time estimation using 32 nuclear protein sequences, together with two fossil calibrations and a calibration for the root based on Choi et al.[1] (numbered blue circles). Gray bars on the nodes show 95% highest posterior density region (HPD) intervals of the node ages. The Great Ordovician Biodiversification Event (GOBE) is shown as an orange box, the Permian-Triassic mass extinction as a purple band and Pangea rifting as a blue box.

(B) Dollo-logic-based analysis of gene family gain and loss during the emergence of the brown algae. Cladogram indicating orthogroup (OG) gain and loss during the emergence of the Phaeophyceae based on Dollo parsimony analysis. Taxonomic classes, orders and families are indicated in brown (brown algae) or gray (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 colored according to overall gene family gain.

(C) Fossil-calibrated phylogenetic tree for 11 *Ectocarpus* species. Extracted from the fossil-calibrated tree shown in (A). Numbered gray bars on the nodes show 95% HPD intervals of the node ages.



(legend on next page)

Figure S3. Genome-wide analyses of gene family evolution, related to Figure 2

(A) Functional and structural features of founder genes from three taxonomic levels. FDI, Fucophycideae/Dictyotales/Ishigeales; PS, Phaeophyceae plus Schizocladiophyceae; PX, Phaeophyceae plus Xanthophyceae.

(B) Inference of HGT origins from 74 species based on monophyletic most similar homolog (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.

(C) Composite gene analysis. Phylogenetic distribution of fused (blue), split (green), and non-remodeled (gray) gene family originations across the evolution of brown algae (middle). 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: bar plot indicating the percentage of gene families retained in extant genomes among all gene families that emerged during the evolution of the species set. Left: bar plot representing the distribution of gene families in COG functional categories for functionally annotated fused, split, and non-remodeled 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 remodeled and non-remodeled gene families ($p < 0.05$, two-sided chi^2 test with Yates correction).

(D) Left: gene families (orthogroups) significantly amplified (binomial test) in the brown algae compared with outgroup taxa. Orthogroups amplified in specific groups of species are indicated by green rectangles. Right: pie charts representing the proportions of manually determined functional categories for each group of amplified gene families highlighted in the left.

(E) Plot showing the timing (phylogenetic clade) of the amplification of gene families (y axis) and the timing of the appearance (gain) of the amplified gene family (x axis, based on the Dollo analysis of orthogroups). Nodes (e.g., n2) are as indicated in Figure S2B.

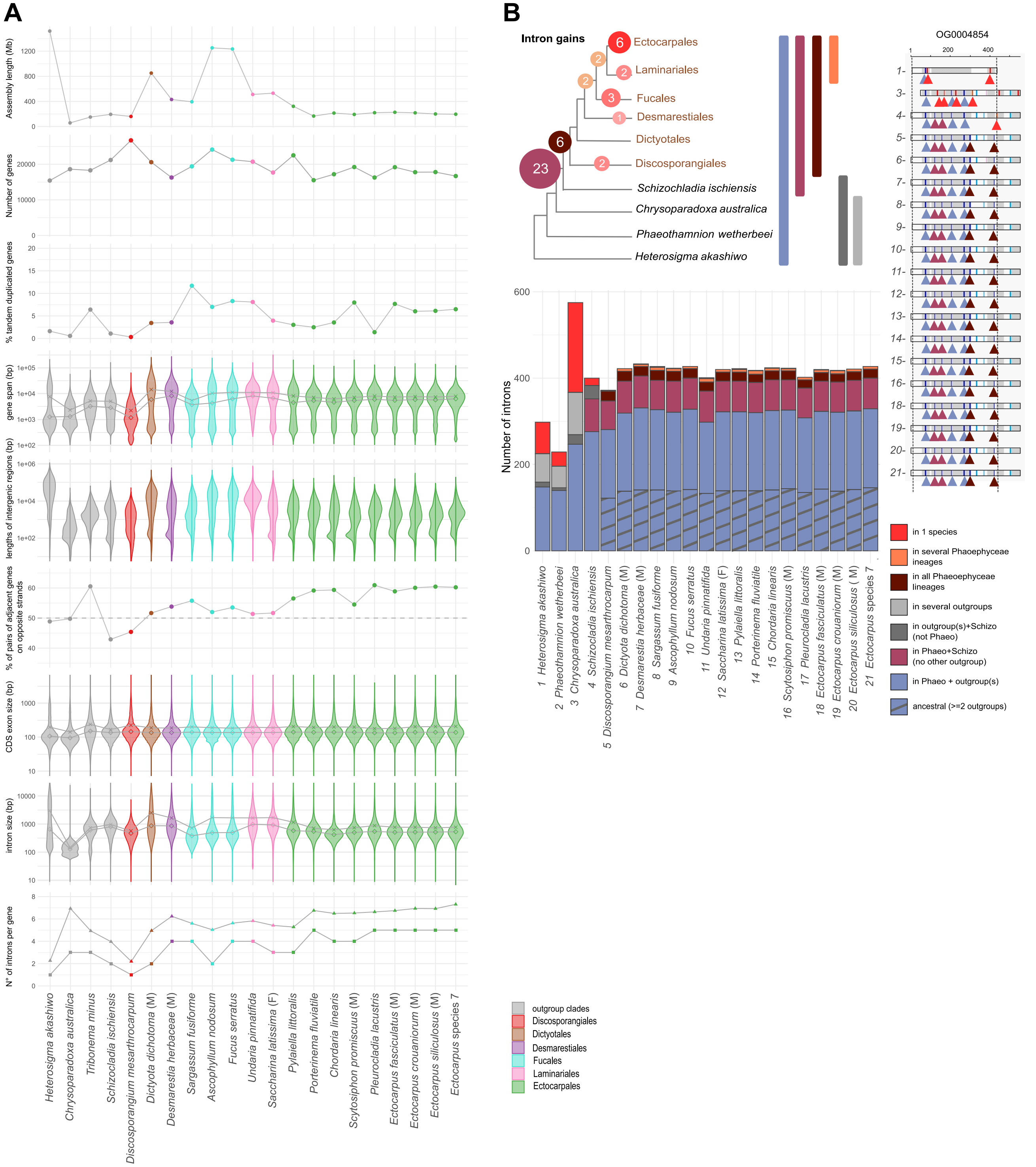
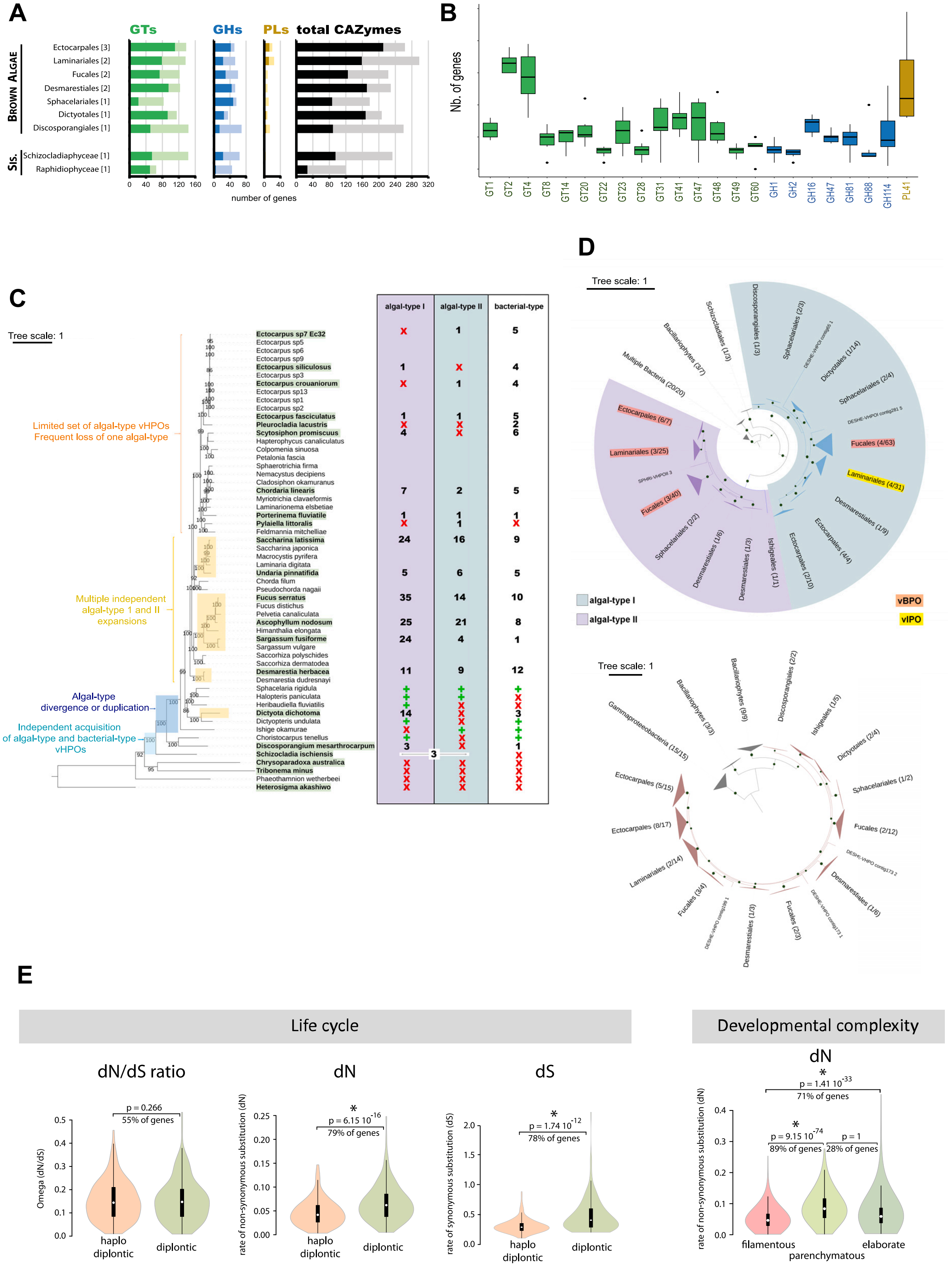


Figure S4. Structural features of the predicted genes in the Phaeoexplorer genomes, related to Figure 1

(A) Genome and gene statistics for the 21 reference genomes (Figure S1A; Table S1B). 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.

(B) Intron acquisition during the emergence of the brown algal lineage (left). The analysis examined 949 introns in 235 conserved orthologs. Colors 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. Color code and species numbering as for the left. F, female; M, male.



(legend on next page)

Figure S5. Analyses of metabolism gene families and life cycle-related gene evolution, related to Figures 1 and 3

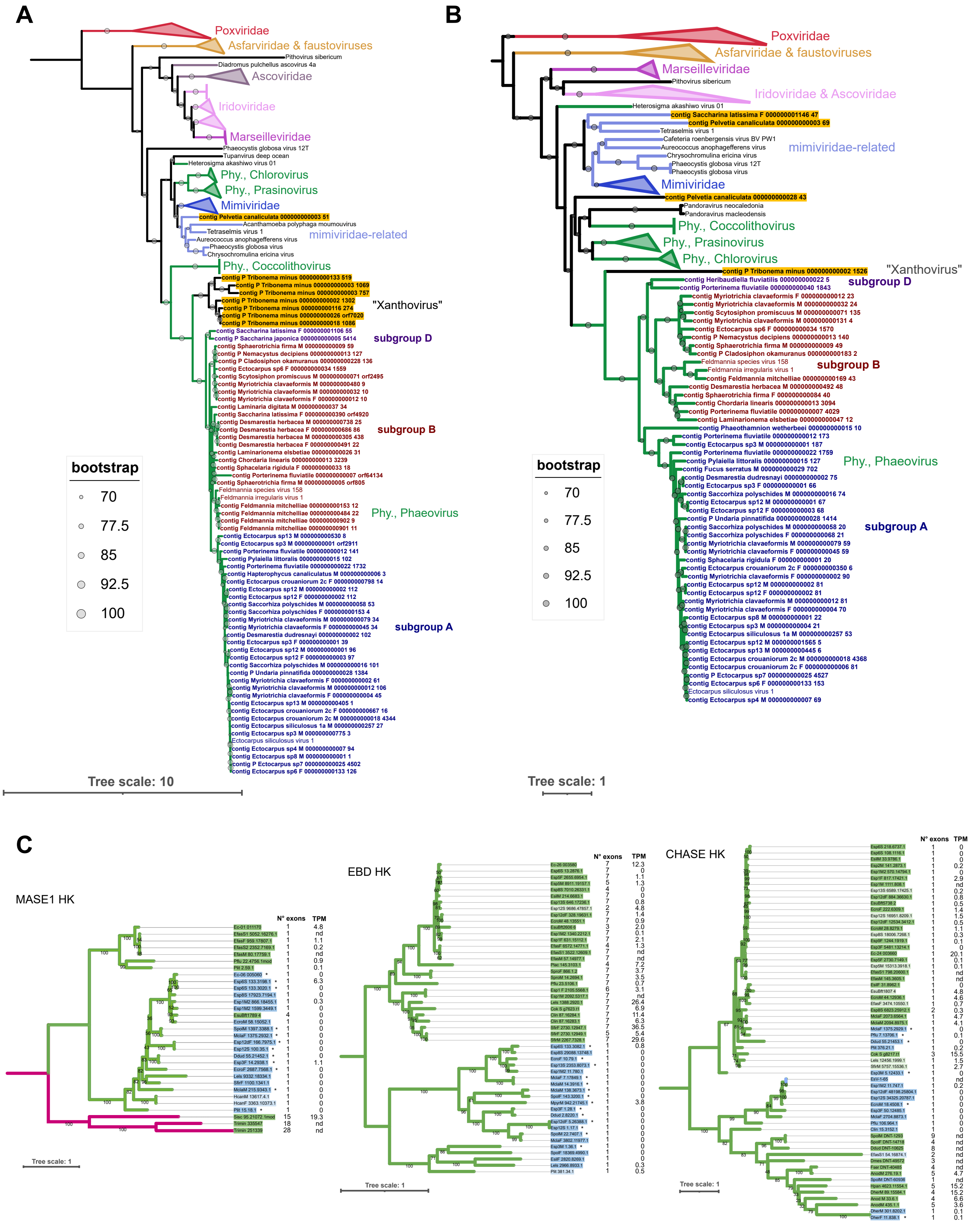
(A) Counts of numbers of genes predicted to encode glycosyltransferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs), and all CAZymes (GTs, GHs, PLs, carbohydrate esterases [CEs], auxiliary activities [AAs], carbohydrate-binding modules [CBMs]), showing numbers for both full-length proteins (dark colors) and fragments (light colors). The data are averaged by order with the number of species analyzed 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 analyzed 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 colored in blue and algal-type II vHPOs in violet. The clades that have structurally or biochemically characterized 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 1,000 bootstraps was drawn as a circular representation taking the gammaproteobacterial group as the starting point to arbitrary root the tree. Brown algal branches are colored in brown. Green dots indicate bootstrap values of between 0.7 and 1.0 (1,000 replicates).

(E) 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.


A
Poxviridae
Asfarviridae & faustoviruses
Pithovirus sibericum
Diadromus pulchellus ascovirus 4a
Ascoviridae
Iridoviridae
Marseilleviridae
Phaeocystis globosa virus 12T
Tupanvirus deep ocean
Heterosigma akashiwo virus 01
Phy., Chlorovirus
Phy., Prasinovirus
Mimiviridae
contig Pelvetia canaliculata 000000000003 51
Acanthamoeba polyphaga moumouvirus
Tetraselmis virus 1
Aureococcus anophagefferens virus
Phaeocystis globosa virus
Chrysochromulina ericina virus
mimiviridae-related
Phy., Coccolithovirus
contig P Tribonema minus 000000000133 519
contig P Tribonema minus 000000000003 1069
contig P Tribonema minus 000000000003 757
contig P Tribonema minus 000000000002 1302
contig P Tribonema minus 000000000116 274
contig P Tribonema minus 000000000026 orf7020
contig P Tribonema minus 000000000018 1086
"Xanthovirus"
contig Saccharina latissima F 000000001106 55
contig P Saccharina japonica 000000000005 5414
subgroup D
contig Sphaerotrichia firma M 000000000009 59
contig P Nemacystus decipiens 000000000013 127
contig P Cladosiphon okamuranus 000000000228 136
contig Ectocarpus sp6 F 000000000034 1559
contig Scytosiphon promiscuus M 000000000071 orf2495
contig Myriotrichia clavaeformis M 000000000480 9
contig Myriotrichia clavaeformis M 000000000032 10
contig Myriotrichia clavaeformis F 000000000012 10
contig Laminaria digitata M 000000000037 34
contig Saccharina latissima F 000000000390 orf4920
contig Desmarestia herbacea M 000000000738 25
contig Desmarestia herbacea F 000000000686 86
contig Desmarestia herbacea M 000000000305 438
contig Desmarestia herbacea F 000000000491 22
contig Laminarionema elsbetiae 000000000026 31
contig Chordaria linearis 000000000013 3239
contig Sphacelaria rigidula F 000000000033 18
contig Porterinema fluviatile 000000000007 orf64134
contig Sphaerotrichia firma M 000000000005 orf805
subgroup B
Feldmannia species virus 158
Feldmannia irregularis virus 1
contig Feldmannia mitchelliae 000000000153 12
contig Feldmannia mitchelliae 000000000484 22
contig Feldmannia mitchelliae 000000000902 9
contig Feldmannia mitchelliae 000000000901 11
Phy., Phaeovirus
contig Ectocarpus sp13 M 000000000530 8
contig Ectocarpus sp3 M 000000000001 orf2911
contig Porterinema fluviatile 000000000012 141
contig Pylaiella littoralis 000000000015 102
contig Porterinema fluviatile 000000000022 1732
contig Hapterophycus canaliculatus M 000000000006 3
contig Ectocarpus crouaniorum 2c F 000000000798 14
contig Ectocarpus sp12 M 000000000002 112
contig Ectocarpus sp12 F 000000000002 112
contig Saccorhiza polyschides M 000000000058 53
contig Saccorhiza polyschides F 000000000153 4
contig Myriotrichia clavaeformis M 000000000079 34
contig Myriotrichia clavaeformis F 000000000045 34
contig Desmarestia dudresnayi 000000000002 102
contig Ectocarpus sp3 F 000000000001 39
contig Ectocarpus sp12 M 000000000001 96
contig Ectocarpus sp12 F 000000000003 97
contig Saccorhiza polyschides M 000000000016 101
contig P Undaria pinnatifida 000000000028 1384
contig Myriotrichia clavaeformis F 000000000002 61
contig Myriotrichia clavaeformis M 000000000012 106
contig Myriotrichia clavaeformis F 000000000004 45
contig Ectocarpus sp13 M 000000000405 1
contig Ectocarpus crouaniorum 2c F 000000000667 16
contig Ectocarpus crouaniorum 2c M 000000000018 4344
contig Ectocarpus siliculosus 1a M 000000000257 27
contig Ectocarpus sp3 M 000000000775 3
Ectocarpus siliculosus virus 1
contig Ectocarpus sp4 M 000000000007 94
contig Ectocarpus sp8 M 000000000001 1
contig P Ectocarpus sp7 000000000025 4502
contig Ectocarpus sp6 F 000000000133 126
subgroup A
bootstrap
70
77.5
85
92.5
100
Tree scale: 10
B
Poxviridae
Asfarviridae & faustoviruses
Marseilleviridae
Pithovirus sibericum
Iridoviridae & Ascoviridae
Heterosigma akashiwo virus 01
contig Saccharina latissima F 000000001146 47
contig Pelvetia canaliculata 000000000003 69
Tetraselmis virus 1
Cafeteria roenbergensis virus BV PW1
Aureococcus anophagefferens virus
Chrysochromulina ericina virus
Phaeocystis globosa virus 12T
Phaeocystis globosa virus
mimiviridae-related
Mimiviridae
contig Pelvetia canaliculata 000000000028 43
Pandoravirus neocaledonia
Pandoravirus macleodensis
Phy., Coccolithovirus
Phy., Prasinovirus
Phy., Chlorovirus
contig P Tribonema minus 000000000002 1526
"Xanthovirus"
contig Heribaudiella fluviatilis 000000000022 5
contig Porterinema fluviatile 000000000040 1843
subgroup D
contig Myriotrichia clavaeformis F 000000000012 23
contig Myriotrichia clavaeformis M 000000000032 24
contig Scytosiphon promiscuus M 000000000071 135
contig Myriotrichia clavaeformis M 000000000131 4
contig Ectocarpus sp6 F 000000000034 1570
contig P Nemacystus decipiens 000000000013 140
contig Sphaerotrichia firma M 000000000009 49
contig P Cladosiphon okamuranus 000000000183 2
Feldmannia species virus 158
Feldmannia irregularis virus 1
contig Feldmannia mitchelliae 000000000169 43
subgroup B
contig Desmarestia herbacea M 000000000492 48
contig Sphaerotrichia firma F 000000000084 40
contig Chordaria linearis 000000000013 3094
contig Porterinema fluviatile 000000000007 4029
contig Laminarionema elsbetiae 000000000047 12
contig Phaeothamnion wetherbeei 000000000015 10
contig Porterinema fluviatile 000000000012 173
contig Ectocarpus sp3 M 000000000001 187
contig Porterinema fluviatile 000000000022 1759
contig Pylaiella littoralis 000000000015 127
contig Fucus serratus M 000000000029 702
contig Desmarestia dudresnayi 000000000002 75
contig Ectocarpus sp3 F 000000000001 66
Phy., Phaeovirus
contig Saccorhiza polyschides M 000000000016 74
contig Ectocarpus sp12 M 000000000001 67
contig Ectocarpus sp12 F 000000000003 68
contig P Undaria pinnatifida 000000000028 1414
contig Saccorhiza polyschides M 000000000058 20
contig Saccorhiza polyschides F 000000000068 21
contig Myriotrichia clavaeformis M 000000000079 59
contig Myriotrichia clavaeformis F 000000000045 59
contig Sphacelaria rigidula F 000000000001 20
contig Ectocarpus crouaniorum 2c F 000000000350 6
contig Myriotrichia clavaeformis F 000000000002 90
contig Ectocarpus sp12 M 000000000002 81
contig Ectocarpus sp12 F 000000000002 81
contig Myriotrichia clavaeformis M 000000000012 81
contig Myriotrichia clavaeformis F 000000000004 70
contig Ectocarpus sp8 M 000000000001 22
contig Ectocarpus sp3 M 000000000004 21
contig Ectocarpus siliculosus 1a M 000000000257 53
contig Ectocarpus sp12 M 000000001565 5
contig Ectocarpus sp13 M 000000000445 6
contig Ectocarpus crouaniorum 2c M 000000000018 4368
contig Ectocarpus crouaniorum 2c F 000000000006 81
contig P Ectocarpus sp7 000000000025 4527
contig Ectocarpus sp6 F 000000000133 153
Ectocarpus siliculosus virus 1
contig Ectocarpus sp4 M 000000000007 69
subgroup A
bootstrap
70
77.5
85
92.5
100
Tree scale: 1
C
MASE1 HK
N° exons
TPM
Tree scale: 1
EBD HK
N° exons
TPM
Tree scale: 1
CHASE HK
N° exons
TPM
Tree scale: 1

(legend on next page)

Figure S6. Phylogenetic trees of viral and putatively virus-derived genes, related to Figure 6

(A and B) *Nucleocytoviricota* major capsid protein and DNA polymerase phylogenies, respectively. The trees were both generated from aligned amino acid sequences with the Q.yeast+F+R6 model and 1,000 bootstraps in IQ-TREE. Subgroup labels refer to *Phaeovirus* genotypes. All sequences identified by this study are in bold and highlighted in yellow if they clustered outside the *Phaeovirus*. Phy, *Phycodnaviridae*.

(C) Phylogenetic trees for three classes of histidine kinase. Membrane-associated sensor1 domain (MASE1) class: brown algal genes are more closely related to viral than to closely related outgroup species genes, EsuBft1789.4 is located within a viral clade. Ethylene-binding-domain-like (EBD) class: viral-related clade limited to Ectocarpales. Cyclases/histidine kinases associated sensory extracellular domain (CHASE) class: complex pattern suggesting possible multiple HGTs. Genes are classed as algal (green or light green background label, strong or weak prediction) or inserted viral sequences (blue or light blue background label, strong or weak prediction) based on exon number, expression level and genomic context (neighboring monoexonic or multiexonic genes). See Table S4J for the gene name abbreviations. Branch colors signify brown algae (green), closely related outgroup taxa (violet) or EsV-1 genes (blue). TPM, maximum TPM; asterisks, gene located in an identified VR; nd, not determined.