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

France Denoeud[1,59], Olivier Godfroy[2,59], Corinne Cruaud[3,60], Svenja Heesch[4,51,60], Zofia Nehr[4,60], Nachida Tadrent[1,52,60], Arnaud Couloux[1,60], Loraine Brillet-Guéguen[5,6,60], Ludovic Delage[7,60], Dean Mckeown[6,60], Taizo Motomura[8,61], Duncan Sussfeld[9,1,61], Xiao Fan[10,11,61], Lisa Mazéas[2,61], Nicolas Terrapon[12,13,61], Josué Barrera-Redondo[14,61], Romy Petroll[14,61], Lauric Reynes[15,61], Seok-Wan Choi[16,61], Jihoon Jo[16,61], Kavitha Uthanumallian[17,61], Kenny Bogaert[18,53,61], Céline Duc[19,61], Pélagie Ratchinski[4,61], Agnieszka Lipinska[4,14,61], Benjamin Noel[1,61], Eleanor A. Murphy[20,21,61], Martin Lohr[22,61], Ananya Khatei[23,61], Pauline Hamon-Giraud[24,61], Christophe Vieira[25,61], Svea Sanja Akerfors[22], Shingo Akita[26], Komlan Avia[27], Yacine Badis[4], Tristan Barbeyron[2], Arnaud Belcour[24,54], Wahiba Berrabah[1], Samuel Blanquart[24], Ahlem Bouguerba-Collin[2], Trevor Bringloe[28], Rose Ann Cattolico[29], Alexandre Cormier[30], Helena Cruz de Carvalho[31,32], Romain Dallet[6], Olivier De Clerck[18], Ahmed Debit[31], Erwan Denis[1], Christophe Destombe[15], Erica Dinatale[14], Simon Dittami[7], Elodie Drula[12,13], Sylvain Faugeron[33], Jeanne Got[24], Louis Graf[16], Agnès Groisillier[19], Marie-Laure Guillemin[34,15], Lars Harms[35], William John Hatchett[36], Bernard Henrissat[37,38], Galice Hoarau[36], Chloé Jollivet[2], Alexander Jueterbock[23], Ehsan Kayal[6,55], 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[15,56], Laetitia Mest[4,57], Gurvan Michel[2], Catia Monteiro[7], Chikako Nagasato[8], Delphine Nègre[6,58], 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[35], Heroen Verbruggen[28], Florian Weinberger[48], Glen Wheeler[21], Hiroshi Kawai[49,62,*], Akira F. Peters[50,62,*], Hwan Su Yoon[16,62,*], Cécile Hervé[2,62,*], Naihao Ye[10,11,62,*], Eric Bapteste[9,62,*], Myriam Valero[15,62,*], Gabriel V. Markov[7,62,*], Erwan Corre[6,62,*], Susana M. Coelho[14,62,*], Patrick Wincker[1,62,*], Jean-Marc Aury[1,62,*] and J. Mark Cock[4,63,*]

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

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

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

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

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

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

[7]Sorbonne Université, CNRS, UMR 8227, ABIE Team, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[8]Muroran Marine Station, Hokkaido University, Muroran, Japan
[9]Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205, Sorbonne Université, CNRS, Museum
[10]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China
[11]Laboratory for Marine Fisheries Science and Food Production Processes, Laoshan Laboratory, Qingdao, Shandong 266237, China
[12]Aix Marseille Univ, CNRS, UMR 7257 AFMB, Marseille, France
[13]INRAE, USC 1408 AFMB, Marseille, France
[14]Department of Algal Development and Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076, Tübingen, Germany
[15]IRL 3614, UMR 7144, DISEEM, CNRS, Sorbonne Université, Station Biologique de Roscoff, Roscoff, 29688, France
[16]Department of Biological Sciences, Sungkyunkwan University, Suwon, 16419, Republic of Korea
[17]University of Melbourne, Australia
[18]Phycology Research Group, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium
[19]Nantes Université, CNRS, US2B, UMR 6286, F-44000 Nantes, France
[20]University of Bristol, UK
[21]Marine Biological Association, Plymouth, UK
[22]Johannes Gutenberg University, Mainz, Germany
[23]Algal and Microbial Biotechnology Division, Nord University, Norway
[24]Univ Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France
[25]Research Institute for Basic Sciences, Jeju National University, Jeju 63243, Korea
[26]Faculty of Fisheries Sciences, Hokkaido University, Minato-cho 3-1-1, Hakodate, Hokkaido, 041-8611, Japan
[27]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France
[28]University of Melbourne, Australia
[29]University of Washington, USA
[30]Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France
[31]Institut de Biologie de l'ENS (IBENS), Département de Biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[32]Université Paris Est-Créteil (UPEC), Faculté des Sciences et Technologie, 61, avenue du Général De Gaulle 94000 Créteil, France

[33]Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile
[34]Núcleo Milenio MASH, Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
[35]Alfred Wegener Institute (AWI), Bremenhaven
[36]Nord University, Norway
[37]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs Lyngby, Denmark
[38]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
[39]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/Univ. Caen Normandie/Univ. Antilles, France
[44]Biology Department, Arcadia University, USA
[45]University of Freiburg, Germany
[46]University of Minnesota, St Paul, USA
[47]Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, United Kingdom
[48]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
[49]Kobe University Research Center for Inland Seas, Kobe, Japan
[50]Bezhin Rosko, 29250 Santeg, France
[51]Current address: Applied Ecology & Phycology, Institute for Biosciences, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany
[52]Current address: Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS-Université de Tours, Tours, 37200, France
[53]Current address: Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, 72076 Tübingen
[54]Current address: Univ. Grenoble Alpes, Inria, 38000, Grenoble, France
[55]Current address: Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, United States of America
[56]Current address: CNRS, La Rochelle Université, UMR7266, Littoral Environnement et Sociétés, La Rochelle, France
[57]Current address: Vegenov, Saint Pol de Léon, France

[58]Current address: Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, France
[59]These authors contributed equally (joint first author)
[60]These authors contributed equally (joint second author)
[61]These authors contributed key analyses
[62]Senior authors
[63]Lead contact
*Correspondence: cock@sb-roscoff.fr, jmaury@genoscope.cns.fr, pwincker@genoscope.cns.fr, susana.coelho@tuebingen.mpg.de, corre@sb-roscoff.fr, gabriel.markov@sb-roscoff.fr, valero@sb-roscoff.fr, epbapteste@gmail.com, yenh@ysfri.ac.cn, cherve@sb-roscoff.fr, hsyoon2011@skku.edu, akirapeters@gmail.com, kawai@kobe-u.ac.jp

SUMMARY

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

INTRODUCTION

140 The brown algae (Phaeophyceae) are a lineage of complex multicellular organisms that emerged about
141 450 Mya[1] from within a group of photosynthetic stramenopile taxa (derived from a secondary
142 endosymbiosis involving a red alga[2]) that are either unicellular or have very simple filamentous
143 multicellular thalli (Figure 1). The emerging brown algae acquired a number of characteristic features
144 that are thought to have contributed to the evolutionary success of this lineage, including complex
145 polysaccharide-based cell walls that confer protection and flexibility in the highly dynamic intertidal
146 environment[3], complex halogen[4] and phlorotannin[5] metabolisms that are thought to play important
147 roles in multiple processes including defence, adhesion and cell wall modification, and a remarkable
148 diversity of life cycles and developmental body architectures adapted to diverse marine
149 environments[6]. As a result of these attributes, many brown algae have become established as key
150 components of extensive coastal ecosystems. These seaweed-based ecosystems provide high value
151 Earth-system-scale services, including the sequestration of several megatons of carbon per year
152 globally, comparable to values reported for terrestrial forests[7], but this important role of seaweed
153 ecosystems is threatened by climate-related declines in seaweed populations worldwide[8]. However,
154 appropriate conservation measures, coupled with the development of seaweed mariculture as a highly
155 sustainable and low impact approach to food and biomass production, could potentially reverse this
156 trend, allowing seaweeds to play a significant role in mitigating the effects of climate change[9]. To attain
157 this objective, it will be necessary to address important gaps in our knowledge of the biology and
158 evolutionary history of the brown algal lineage. For example, these seaweeds remain poorly described
159 in terms of genome sequencing due, in part, to difficulties with extracting nucleic acids. The
160 Phaeoexplorer project (https://phaeoexplorer.sb-roscoff.fr/) has generated a large dataset of genome
161 sequences, spanning all the major orders of the Phaeophyceae[10]. This extensive genomic dataset has
162 been analysed here to study the origin and evolution of key genomic features during the emergence
163 and diversification of this important group of marine organisms.
164
165 **RESULTS**
166
167 **In-depth sequencing of brown algal genomes**
168
169 Until now, good quality genome assemblies have been obtained for only five brown algal species[11–15],
170 together with about 46 draft genome assemblies[16–20]. Here we report work that has significantly
171 expanded the genomic data available by sequencing and assembling 17 good quality genomes using
172 long-read technology (Table S1), plus an additional 43 draft genome assemblies. These 60 genomes
173 correspond to 40 brown algae and four closely related species, covering 16 Phaeophyceae families
174 providing a dense coverage of this lineage (Figures S1A, S1B, Table S1). The sequenced species include

175 brown algae that occur at different levels of the intertidal and subtidal, and are representative of the
176 broad diversity of this group of seaweeds in terms of size, levels of multicellular complexity,
177 biogeography and life cycle structure (Figures 1, S2). The analyses carried out in this study have focused
178 principally on a set of 21 good quality reference genomes, which include four previously published
179 genomes (Table S1).
180
181 **Marked changes in genome content and gene structure during the emergence of the Phaeophyceae**
182 **lineage**
183
184 Recent evidence indicates that the brown algae emerged about 450 Mya during the Great Ordovician
185 Biodiversification Event (GOBE)[1], a conclusion that is supported by a fossil-calibrated tree built with a
186 nuclear-gene-based phylogeny constructed using the Phaeoexplorer genome data (Figures 1, S3). An
187 increase in atmospheric oxygen at the time of the GOBE, associated with the emergence of herbivorous
188 marine invertebrates[21], is likely to have created conditions conducive to the observed transition
189 towards increased multicellular complexity during early brown algal evolution.
190
191 To investigate genomic modifications associated with the emergence and diversification of the brown
192 algae, we first carried out a series of genome-wide analyses aimed at identifying broad trends in
193 genome evolution over evolutionary time (Figure 2). Dollo analysis of gain and loss of orthogroups (i.e.
194 gene families) indicated marked gains during early brown algal evolution followed by a broad tendency
195 to lose orthogroups later as the different brown algal orders diversified (Figures 2B, S4). Similarly, a
196 phylostratigraphy analysis indicated that 29.6% of brown algal genes cannot be traced back to
197 ancestors outside the Phaeophyceae, with the majority of gene founder events occurring early during
198 the emergence of the brown algae (Figure 2E, S5A, Table S2), again indicating a burst of gene birth
199 during the emergence of this lineage. Both the Dollo analysis and the phylostratigraphy approach
200 indicated that the gene families acquired during early brown algal evolution were significantly enriched
201 in genes that could not be assigned to a COG category, suggesting a burst in the acquisition of genetic
202 novelty (Figure 2G).
203
204 One of the factors underlying the marked burst of gene gain during the emergence of the brown algae
205 was an increase in the rate of acquisition of new genes via horizontal gene transfer (HGT). A phylogeny-
206 based search for genes potentially derived from HGT events indicated that they constitute about 1%
207 of brown algal gene catalogues and that the novel genes were principally acquired from bacterial
208 genomes (Figures 2F, S5B). The proportion of class-specific HGT events (i.e. events where a gene was
209 acquired via HGT by an ancestral acceptor species after the emergence of a taxonomic class so that

210 the gene is only found in members of that class) compared to more ancient HGT events was greater
211 for the brown algae (33.5%) than for the closely-related taxa Xanthophyceae (*Tribonema minus*) and
212 Raphidophyceae (*Heterosigma akashiwo*; mean of 17.1% for the two taxa, Wilcoxon *p*-value 0.021),
213 indicating that higher levels of HGT occurred during the emergence of the brown algae than in closely-
214 related taxa (Figure 2F).
215
216 The marked increase in the rate of gene gain appears to have been a key factor in the emergence of
217 the brown algal lineage but this was not the only process that enriched brown algal genomes during
218 this period. Domain fusions and fissions (composite genes) were prevalent during the early stages of
219 brown algal emergence (with gene fusion events being slightly more common than gene fissions;
220 Figures 2D, S5C), affecting about 7% of brown algal gene complements. In contrast, gene family
221 amplifications were most prevalent at a later stage of brown algal evolution, corresponding to the
222 diversification of the major brown algal orders during the Mesozoic (Figures 2C, S5D, S5E, Table S3),
223 and therefore did not share the same pattern of increased occurrence during early evolution of the
224 lineage. However, the amplified gene families were significantly enriched in genes that had been
225 gained during the emergence of the brown algae (χ^2 *p*-value 1.04e-15; Table S1C) indicating that gene
226 gain during the early evolution of the lineage nonetheless played a crucial role by establishing the
227 majority of the gene families that would later undergo amplifications.
228
229 Analysis of the predicted functions of the three sets of gene families identified as having been
230 amplified, derived from domain fusions/fissions or derived from an HGT event (Figure 2G) indicated
231 that they were enriched in several functional categories, notably carbohydrate metabolism, signal
232 transduction and transcription. These functional categories may have been important in the
233 emergence of the complex brown algal cell wall or correspond to a complexification of signalling
234 pathways as multicellular complexity increased. Interestingly, many of the genes acquired at, or shortly
235 after, the origin of the Phaeophyceae encode secreted or membrane proteins (Figure S5A), suggesting
236 roles in cell-cell communication that may have been important for the emergence of complex
237 multicellularity or as components of defence mechanisms. The acquisition of plasmodesmata by
238 brown algae directly after their divergence from their sister taxon *Schizocladia ischiensis*[22] (Figure 1)
239 underlines the importance of cell-cell communication from the outset of brown algal evolution.
240
241 The emergence of the brown algae also corresponded with changes in gene structure. On average,
242 brown algal genes tend to be more intron-rich than those of the other stramenopile groups[23], including
243 closely-related taxa (Figures S6A), with the notable exception of *Chrysoparadoxa australica*, whose
244 genes contain many short, C/T-rich introns (Figure S6A). A comparison of orthologous genes indicated

245 a phase of rapid intron acquisition just before the divergence of the Phaeophyceae and the
246 Schizocladiophyceae, followed by a period of relative intron stability up to the present day (Figure
247 S6B). This phase of accelerated intron acquisition coincided approximately with the periods of marked
248 gene gain and domain reorganisation discussed above and may have been an indirect consequence of
249 increased multicellular complexity (Figure 1) due to a concomitant decrease in effective population
250 size[24]. Once established, increased intron density may have facilitated some of the genome-wide
251 tendencies described above, such as increased reorganisation of composite genes for example, and
252 thereby played an important role in a context of increasing developmental complexity[25–28].
253
254 Taken together, these analyses suggest a chain of events in which GOBE increases in both atmospheric
255 oxygen and herbivory drove the emergence of the brown algae, and the acquisition of multicellular
256 complexity by this emerging lineage. The emergence of this lineage was associated with marked
257 changes in the general structure and content of their genomes, including the acquisition of new genes
258 (some via HGT) and reorganisation of protein domains within genes. The following section focuses on
259 metabolic and signalling pathways that are predicted to have played key roles during this important
260 period of emergence and early evolution of the brown algae.
261
262 **Acquisition of key metabolic and signalling pathways during the emergence of the Phaeophyceae**
263
264 The success of the brown algae as an evolutionary lineage has been attributed, at least in part, to the
265 acquisition of several key metabolic pathways, particularly those associated with cell wall biosynthesis,
266 and both halogen and phlorotannin metabolism[3–5]. AuCoMe[29] was used to provide an overview of
267 metabolic pathways across the brown algae by inferring genome-scale metabolic networks for all the
268 species in the Phaeoexplorer dataset (https://gem-aureme.genouest.org/phaeogem/; Figures S7A,
269 S7B) and also for a subset of long-read sequenced species (https://gem-
270 aureme.genouest.org/16bestgem/; Figures S7C, S7D). Comparison of the brown algal metabolic
271 networks with those of closely-related taxa allowed the identification of enzymatic activities that have
272 been gained or lost in the former compared with the latter (Figures S7E, S7F, S7G). The most robust
273 set of core reactions found mainly in brown algae is a set of 24 reactions catalysed by polyamine
274 oxidases, carotenoid and indole-3-pyruvate oxygenases, a methyladenine demethylase, a PQQ-
275 dependent sugar dehydrogenase, a xylosyl transferase (CAZyme family GT14), a ribulosamine kinase,
276 galactosylceramide sulfotransferases and phosphopentomutases (Table S4).
277
278 The identification of carbohydrate-active enzymes (CAZymes) among the AuCoMe core reaction set is
279 of particular interest because many of these enzymes have been implicated in the biosynthesis of the

280 characteristic cell walls of brown algae. Large complements of CAZYme genes (237 genes on average)
281 were found in all brown algal orders and in their sister taxon *S. ischiensis* but this class of gene was less
282 abundant in the more distantly related unicellular alga *H. akashiwo* (Figures 3A, S8A, S8B, Tables S5,
283 S6). The evolutionary history of carbohydrate metabolism gene families was investigated by combining
284 information from the genome-wide analyses of gene gain/loss, HGT and gene family amplification
285 (Figure 3B). This analysis indicated that several key genes and gene families (ManC5-E, GT23 and PL41)
286 were acquired by the common ancestor of brown algae and *S. ischiensis*, with strong evidence in some
287 cases that this occurred via HGT (PL41). Moreover, marked amplifications were detected for several
288 families (AA15, ManC5-E, GH114, GT23 and PL41), indicating that both gain and amplification of gene
289 families played important roles in the emergence of the brown algal carbohydrate metabolism gene
290 set. Alginate is a major component of brown algal cell walls and it plays an important role in conferring
291 resistance to the biomechanical effect of wave action[3]. It is therefore interesting that mannuronan C5
292 epimerase (ManC5-E), an enzyme whose action modulates the rigidity of the alginate polymer, appears
293 to have been acquired very early, by the common ancestor of brown algae and *S. ischiensis* (Figures
294 3B, 3C). The acquisition of ManC5-E, together with other alginate pathway enzymes such as PL41
295 (Figures 3A, 3B and 3D), was probably an important evolutionary step, enabling the emergence of
296 large, resilient substrate-anchored multicellular organisms in the highly dynamic and stressful coastal
297 environment (Figure 1).
298
299 Vanadium-dependent haloperoxidases (vHPOs) are a central component of brown algal halogen
300 metabolism, which has been implicated in multiple biological processes including defence, adhesion,
301 chemical signalling and the oxidative stress response. All three classes of brown algal vHPO (algal types
302 I and II and bacterial-type[30–32]) appear to have been acquired early during the emergence of the
303 Phaeophyceae (Figures 3A, S8C, S8D, Tables S7, S8). Closely-related Stramenopile species do not
304 possess any of these three types of haloperoxidase, with the exception of the sister taxon, *S. ischiensis*,
305 which possesses three intermediate algal-type (i.e. equidistant phylogenetically from class I and class
306 II algal-types) haloperoxidase genes (Figures 3A, S8C, S8D). Algal type I and II vHPO genes probably
307 diverged from an intermediate-type ancestral gene similar to the *S. ischiensis* genes early during
308 Phaeophyceae evolution. It is likely that the initial acquisitions of algal- and bacterial-type vHPOs
309 represented independent events although the presence of probable vestiges of bacterial-type vHPO
310 genes in *S. ischiensis* means that it is not possible to rule out acquisition of both types of vHPO through
311 a single event. Clearly, however, gain of vHPO genes was an important event leading to the emergence
312 of the emblematic halogen metabolism of brown algae early during their evolutionary history.
313
314 Gene gain may not, however, have been the proximal factor responsible for all the key metabolic

315 innovations that occurred in the emerging brown algal lineage. Phlorotannins are characteristic brown
316 algal polyphenolic compounds that occur in all Phaeophyceae species, with the exception of some
317 members of the Sargassaceae. Phlorotannins are derived from phloroglucinol and brown algae possess
318 three classes of type III polyketide synthase, two of which (PKS1 and PKS2) were acquired prior to the
319 emergence of the Phaeophyceae and the third (PKS3) evolving much later within the Ectocarpales
320 (Figures 3A, 4A, Table S9). Interestingly, PKS1 proteins from different brown algal species have been
321 shown to have different activities leading to the production of distinct metabolites[33–35], indicating that
322 the acquisition of novel functions by this class of enzymes may have played an important role in the
323 emergence of the brown algal capacity to produce phlorotannins. Moreover, many stramenopile PKS
324 type III genes encode proteins with signal peptides or signal anchors (Figure S8E). For the brown algae,
325 this feature is consistent with the cellular production site of phlorotannins and the observed transport
326 of these compounds by physodes, secretory vesicles characteristic of brown algae[36]. Cross-linking of
327 phlorotannins, embedded within other brown algal cell wall compounds such as alginates, has been
328 demonstrated *in vitro* through the action of vHPOs[37–39] and indirectly suggested by *in vivo* observations
329 colocalising vHPOs with physode fusions at the cell periphery[40,41]. Consequently, vHPOs are good
330 candidates for the enzymes that cross-link phlorotannins and other compounds, perhaps even for the
331 formation of covalent bonds between phloroglucinol monomers and oligomers, which could occur via
332 activation of aromatic rings through halogenation. These observations suggest that the acquisition of
333 vHPOs by the common ancestor of brown algae and *S. ischiensis*, together perhaps with modifications
334 of the existing PKS enzymes, triggered the emergence of new metabolic pathways leading to the
335 production of the phlorotannin molecules characteristic of the Phaeophyceae lineage.
336
337 The acquisition of increased multicellular complexity and adaptation to new ecological niches during
338 the early stages of brown algal evolution (i.e. during and immediately following the GOBE) is expected
339 to have required modification and elaboration of signalling pathways. Membrane-localised signalling
340 proteins (Figures 3A, S9) are of particular interest in this context not only as potential mediators of
341 intercellular signalling in a multicellular organism but also because of potential interactions with the
342 elaborate brown algal extracellular matrices (cell walls)[3,42]. A detailed analysis of the brown algal
343 receptor kinase (RK) gene family, revealed that it actually includes two types of receptor, the previously
344 reported leucine-rich repeat (LRR) RKs[11] and a newly-discovered class of receptors with a beta-
345 propeller extracellular domain (Figures 3A, S9A, Table S10). The phylogenetic distribution of RKs
346 (Figures 3A, S9A) suggests that beta-propeller-type RKs may have originated before LRR RKs in the
347 brown algae, although the presence of both beta-propeller and LRR RKs in *C. australica* (but none of
348 the other closely-related outgroup species) raises questions about the timing of the emergence of
349 these gene families. Note that oomycetes (distantly related stramenopiles) possess a phylogenetically

350 distinct family of LRR RKs[11] but not beta-propeller RKs. Diversification of receptor kinase families by
351 the acquisition of different extracellular domains is therefore a characteristic that is so far unique to
352 the three most complex multicellular lineages: animals, land plants and brown algae.
353
354 Major changes in epigenetic regulation also appear to have occurred during the emergence of the
355 brown algae (see also supplemental results). *DNMT1* genes were identified in *Discosporangium*
356 *mesarthrocarpum* and two closely-related outgroup species (*S. ischiensis* and *C. australica*) but not in
357 other brown algal genomes, indicating that the common ancestor of brown algae probably possessed
358 *DNMT1* but that this gene was lost after divergence of the Discosporangiales from other brown algal
359 taxa (Figure 3A, Table S11). This is consistent with the reported absence of DNA methylation in the
360 filamentous brown alga *Ectocarpus*[11] and a very low level of DNA methylation in the kelp *Saccharina*
361 *japonica*[43] (which is thought to be mediated by DNMT2). Our analysis indicates that most brown algae
362 either lack DNA methylation or exhibit very low levels of methylation and that this feature was
363 acquired early during brown algal diversification.
364
365 **Impact of morphological, life cycle and reproductive diversification during the Mesozoic on brown**
366 **algal genome evolution**
367
368 A second major step in the evolutionary history of the Phaeophyceae was the rapid diversification of
369 the major brown algal orders, which began after the origin of the FDI
370 (Fucophycidae/Dictyotales/Ishigeales) clade, here estimated at 235.97 Ma (95% HPD: 158.88-312.48
371 Mya, an age that is broadly consistent with previous work[1]; Figures 1, S3). This diversification closely
372 followed the Permian–Triassic mass extinction event (which dramatically impacted marine ecosystems
373 in which red and green algae played dominant roles) and was facilitated by Triassic marine
374 environments that favoured chlorophyll-c containing algae (e.g. high phosphate and low iron
375 concentration), along with the appearance of new coastal niches created by Pangea rifting (Figure 1).
376 This context would have facilitated the diversification of the brown algal lineage[44,45], resulting in
377 organisms that now exhibit a broad range of morphological complexity (ranging from filamentous to
378 complex parenchymatous thalli), different types of life cycle and diverse reproductive strategies and
379 metabolic capacities[3,6,29,46] (Figures 1, S10A). The Phaeoexplorer dataset was analysed to identify
380 genomic features associated with this diversification of phenotypic characteristics and to evaluate the
381 impact on genome evolution and function.
382
383 We found indications that the diversification of life cycles, in some cases linked with the emergence of
384 large, complex body architectures, impacted genome evolution through population genetic effects.

385 Most brown algae have haploid-diploid life cycles involving alternation between sporophyte and
386 gametophyte generations, the only exception being the Fucales, which have diploid life cycles. The
387 theoretical advantages of different types of life cycle have been discussed in detail[47] and one proposed
388 advantage of a life cycle with a haploid phase is that this allows effective purifying counter-selection
389 of deleterious alleles. When the brown algae with haploid-diploid life cycles were compared with
390 species from the Fucales, increased rates of both synonymous and non-synonymous mutation rates
391 were detected in the latter, consistent with the hypothesis that deleterious alleles are phenotypically
392 masked in species where most genes function in a diploid context (Figure S10B). Comparison of non-
393 synonymous substitution rates (dN) for genes in brown algae with different levels of morphological
394 complexity, ranging from simple filamentous thalli though parenchymatous to morphologically
395 complex, indicated significantly lower values of dN for filamentous species (Figure S10B). This
396 observation suggests that the emergence of larger, more complex brown algae may have resulted in
397 reduced effective population sizes and consequently weaker counter-selection of non-synonymous
398 substitutions[24].
399
400 For brown algae with haploid-diploid life cycles, the relative sizes of the two generations vary markedly
401 across species with, for example, the sporophyte being much larger and more developmentally
402 complex than the gametophyte in groups such as the kelps (Laminariales). To evaluate the effect of
403 this diversity on life-cycle-related gene expression, generation-biased gene (GBG) expression was
404 analysed in ten species whose haploid-diploid life cycles ranged from sporophyte dominance, through
405 sporophyte-gametophyte isomorphy to gametophyte dominance. This analysis indicated, as
406 expected[48], that the number of GBGs was inversely correlated with the degree of intergenerational
407 isomorphy (Figure S10C, Table S12). However, species with strongly heteromorphic life cycles had
408 greater numbers of both sporophyte- and gametophyte-biased genes, indicating either that gene
409 downregulation plays an important role in the establishment of heteromorphic generations or that
410 there is also cryptic complexification of the more morphologically simple generation.
411
412 The diversification of the brown algae in terms of developmental complexity and life cycle structure
413 was associated with modifications to reproductive systems, including, for example, partial or complete
414 loss of flagella from female gametes in oogamous species and more subtle modifications such as loss
415 of the eyespot in several kelps or of the entire posterior flagella in *D. dichotoma*[49,50]. Interestingly,
416 these latter modifications are correlated with loss of the *HELMCHROME* gene, which is thought to be
417 involved in light reception and zoid phototaxis[51], from these species (Figures 3A, 4B). In addition, an
418 analysis of the presence of genes for 70 high-confidence flagellar proteins[51] across eight species with

419 different flagellar characteristics identified proteins that correlate with presence or absence of the
420 eyespot or of the posterior flagellum (Figure 4B, Table S13).
421
422 **Brown algal diversification and the emergence of marine forests was also associated with genomic**
423 **changes affecting metabolic and signalling pathways**
424
425 Forests of brown algae (i.e., Laminariales, Desmarestiales, Tilopteridales and Fucales[52]) are a key
426 aspect of the modern marine biosphere. One of the pivotal innovations related to their emergence
427 was a new developmental tissue, an intercalary meristem situated in the zone between the stipe and
428 the lamina. The presence of this tissue is an ancestral state of the brown algal crown radiation (BACR)
429 clade, and the current study indicates that the intercalary meristem was acquired as early as 190 Mya
430 (Figure 1). This type of intercalary meristem would have facilitated the transition from annual to
431 perennial life history and would, therefore, have been important for the establishment and
432 maintenance of marine forests, particularly when upper parts of thalli are subjected to heavy grazing
433 pressure[10]. Our results indicate that the Desmarestiales, Tilopteridales, and Fucales were all present
434 by the early Cretaceous (Figure 1). Thus, it is possible that brown algal forests, at least at a small scale,
435 provided both nutrients and shelter for the marine herbivorous animals that became common during
436 the Cretaceous Period (e.g. algae-eating echinoids, sea turtles and euteleostean fish[53,54]).
437
438 While our estimates of kelp antiquity are earlier than those of Starko *et al*.[55], they are consistent with
439 their suggestion that Cenozoic cooling facilitated the geographic expansion of the kelp forest
440 ecosystem. Indeed, many of the animals found today in kelp forest ecosystems originated toward the
441 end of the Cretaceous Period, or later[56]. Currently, our understanding of Mesozoic marine noncalcified
442 macroalgae on the basis of fossils[57–59] is too poor to provide much guidance in this regard, but
443 documentation by Kiel *et al*.[56] of fossil holdfasts indicates that kelp forests were present by the late
444 Paleogene Period (~32 Mya). The highly complex, multi-layered and canopy-forming kelp forests of
445 today, however, seem to have emerged only relatively recently, during the mid-Neogene, following
446 the expansion of cooler water shelf environments[55,56].
447
448 Comparative analysis of the Phaeoexplorer genome dataset identified a number of gene family
449 expansions that potentially played important roles in the adaptation of the brown algae to their diverse
450 niches and, more particularly, in the emergence of large, forest-forming species such as the kelps. For
451 example, the ManC5-E family expanded markedly in the Laminariales and Fucales (Figure 3C), the two
452 main orders that constitute extant phaeophycean forests. The capacity of ManC5-E to modify organ
453 flexibility[3] may therefore have been an important factor for large organisms coping with the harsh

454 hydrodynamic conditions of coastal environments[60]. In addition, five different orthogroups containing
455 proteins with the mechanosensor wall stress-responsive component (WSC) domain were identified as
456 having increased in size during the diversification of the brown algal lineage (Table S3), indicating that
457 metabolic innovations affecting cell walls may have been concomitant with a complexification of
458 associated signalling pathways.
459
460 The haloperoxidase gene families, discussed above, expanded independently in several brown algal
461 orders, again with expansions being particularly marked in the Fucales and the Laminariales (Figures
462 3A, S8C). In contrast, with some rare exceptions, the Ectocarpales, which are mostly small, filamentous
463 algae, do not possess expanded vHPO gene families and there is evidence for multiple, independent
464 gene losses resulting in loss of one or two of the three vHPO classes in many of the species in this
465 order. In the Laminariales, the algal type I family are specialised for iodine rather than bromine[61] and
466 this may have been an innovation that occurred specifically within the Laminariales, resulting in a
467 halogen metabolism with an additional layer of complexity. However, further analysis of the enzymatic
468 activities of proteins from other orders will be necessary to determine if substrate specialisation was
469 concomitant with this specific gene family expansion.
470
471 One of the proposed roles of halogenated molecules in brown algae is in biotic defence[4] and, clearly,
472 an effective defence system would have been an important prerequisite for the emergence of the
473 large, perennial organisms that constitute marine forests. Additional immunity-related families[62] that
474 expanded during the diversification of the brown algae include five orthogroups that contain either
475 GTPases with a central Ras of complex proteins/C-terminal of Roc RocCOR domain tandem (ROCO
476 GTPases) or nucleotide-binding adaptor shared by apoptotic protease-activating factor 1, R proteins
477 and CED-4 tetratricopeptide repeat (NB-ARC-TPR) genes (Table S3). In addition, analysis of gene family
478 expansions indicated that modifications to additional biological processes such as photosynthesis,
479 carbon metabolism and signalling may have been important (Figures S11, S12, Table S3, S14, S15, S16).
480
481 Finally, one of the most remarkable gene family amplifications detected in this study was for proteins
482 containing the EsV-1-7 domain, a short, cysteine-rich motif that may represent a novel class of zinc
483 finger[63]. EsV-1-7 domain proteins are completely absent from animal and land plant genomes and most
484 stramenopiles either have just one member (oomycetes and eustigmatophytes) or entirely lack this
485 gene family[63]. Analysis of the Phaeoexplorer data (Figure 3A, Table S17) indicated that the EsV-1-7
486 gene family started to expand in the common ancestor of the brown algae and the raphidophyte *H.*
487 *akashiwo*, with 31-54 members in the non-Phaeophyceae taxa that shared this ancestor. Further
488 expansion of the family then occurred in most brown algal orders, particularly in some members of

489 the Laminariales (234 members in *Saccharina latissima*) and the Fucales (335 members in *Ascophyllum*
490 *nodosum*), with the genes tending to be clustered in tandem arrays (Tables S3, S17). These
491 observations are consistent with the previous description of a large EsV-1-7 domain family (95 genes)
492 in *Ectocarpus* species 7[63] and with recent observations by Nelson *et al*.[20]. One member of this family,
493 IMMEDIATE UPRIGHT (IMM), has been shown to play a key role in the establishment of the elaborate
494 basal filament system of *Ectocarpus* sporophytes[63], suggesting that EsV-1-7 domain proteins may be
495 novel developmental regulators in brown algae. Orthologues of the *IMM* gene were found in brown
496 algal crown group taxa and in *Dictyota dichotoma* but not in *D. mesarthrocarpum* (Figure 3A, Table
497 S17), indicating that this gene originated within the EsV-1-7 gene family as the first brown algal orders
498 started to diverge. *IMM* is therefore conserved in most brown algal taxa but its function in species
499 other than *Ectocarpus* remains to be determined.
500
501 **Recent evolutionary events within the genus *Ectocarpus***
502
503 The above analyses focused on deep-time evolutionary events related to the emergence of the
504 Phaeophyceae and the later diversification of the brown algal orders during the Mesozoic. To
505 complement these analyses an evaluation of relatively recent and ongoing evolutionary events in the
506 brown algae was conducted by sequencing 22 new strains from the genus *Ectocarpus*, which originated
507 about 19 Mya (Figure S13A).
508
509 A phylogenetic tree was constructed for 11 selected *Ectocarpus* species based on 261 high quality
510 alignments of 1:1 orthologs (Figure 5A). The tree indicates substantial divergence between *E.*
511 *fasciculatus* and two well-supported main clades, designated clade 1 and clade 2. Rates of synonymous
512 and non-synonymous substitutions, based on alignments of genes in syntenic blocks between
513 *Ectocarpus* species 7 and four other species, correlated with relative divergence times as estimated
514 from the phylogenetic tree (Table S18), supporting the tree topology. Incongruencies between the
515 species tree and trees for individual genes indicated introgression events and/or incomplete lineage
516 sorting across the *Ectocarpus* genus. This conclusion was supported by phylogenetic tree comparisons,
517 Bayesian hierarchical clustering, phylogenetic network reconstructions and PCoA analysis (Figures 5A,
518 S13B, S13C). D-statistic analysis, specifically ABBA-BABA tests, detected incongruities among species
519 quartets, indicating potential gene flow at various times during the evolution of the *Ectocarpus* genus.
520 Evidence for gene flow was particularly strong for clade 2 and there was also evidence for marked
521 exchanges between the two main clades (Figure 5B), suggesting that gene flow has not been limited
522 to recently-diverged species pairs. These findings suggest a complex evolutionary history involving
523 rapid divergence, hybridization and introgression among species within the *Ectocarpus* genus, with

524 evidence for hybridisation occurring between 10.5 Mya (for clades 1 and 2) and 3.3 Mya (for
525 *Ectocarpus* species 5 and 7) based on the fossil-calibrated tree (Figure S13A). A similar scenario has
526 been reported for the genus *Drosophila*[64], suggesting that recurrent hybridization and introgression
527 among species may be a common feature associated with rapid species radiations. Major
528 environmental changes such as the expansion of cold-water coastal areas following the green-house /
529 cold-house Eocene–Oligocene transition (~30 Mya[68]), and particularly the rapid climate destabilisation
530 and temperature drop associated with the end of the mid-Miocene thermal maximum (~15 Mya[68]),
531 may have created many new opportunities for the rapid expansion and diversification of the
532 *Ectocarpus* genus.
533
534 **Brown algal genomes contain large amounts of inserted viral sequences**
535
536 A particularly striking result of this study was the identification of extensive amounts of integrated
537 DNA sequence corresponding to large DNA viruses of the *Phaeovirus* family (Figure 6A, Table S19),
538 which integrate into brown algal genomes as part of their lysogenic life cycles[69]. Analysis of 72 genomes
539 in the Phaeoexplorer and associated public genome dataset identified a total of 792 viral regions (VRs)
540 of *Nucleocytoviricota* (NCV) origin in 743 contigs, with a combined length of 32.3 Mbp. Individual VRs
541 ranged in size from two to 705 kbp, but the majority (81.3 %) were between two and 50 kbp, whilst
542 only 9% were longer than the expected minimum size (100 kbp) for an NCV genome. On average, VRs
543 comprised 65% of their contig by length, and 72% of contigs with VRs were shorter than 100 kbp.
544 Therefore, for most VRs, their true size and genomic context is unknown due to assembly gaps and
545 short contig lengths. However, 40.8% of VRs had at least one flanking region providing direct evidence
546 for insertion of the sequence in the algal genome (Table S19). Figure 6B shows three examples of long
547 VRs. Most genes in VRs are monoexonic and transcriptionally silent, as previously observed for the 310
548 kbp VR in the *Ectocarpus* species 7 strain Ec32 genome[11].
549
550 On average, each of the 72 analysed genomes contained 469 kbp of VR (with a maximum of 5,614 kbp)
551 and only two genomes contained no VRs (both from the Discosporangiales). There were a number of
552 outlier genomes that contained more than 1 Mbp of VRs (*T. minus*, *S. latissima*, *S. japonica*, *P. fluviatile*
553 and *Myriotrichia clavaeformis* male and female). The presence of key NCV marker genes was used to
554 assess the completeness of inserted viral genomes. At least one partial provirus (a VR possessing
555 several key NCV marker genes) was present in 39 genomes, 29 of which had at least one full provirus
556 with a complete set of seven key NCV marker genes (Figure 6A, Table S19). In addition to the previously
557 known infections in Ectocarpales[69] and Laminariales[70], integrated NCV proviruses were found in all
558 Phaeophyceae orders screened, except the Discosporangiales and Dictyotales, and were also detected

559 in *T. minus* (Xanthophyceae). Moreover, NCV marker gene composition indicated that multiple
560 integrated proviruses were present in 16 genomes from multiple Phaeophyceae orders (Ectocarpales,
561 Desmarestiales, Sphacelariales, Tilopteridales, Laminariales), and the Xanthophyceae (Figure 6A, Table
562 S19). Phylogenetic analysis of the major capsid protein (MCP) and DNA polymerase genes indicated
563 that the majority of the integrated NCVs belonged to the genus *Phaeovirus*, the sole viral group known
564 to infect brown algae (Figure S14A,B). However, this analysis also revealed integrated sequences
565 corresponding to other viral groups. Viral sequences in *T. minus* belonged to a putative novel genus
566 closely related to *Phaeovirus*, for which we propose the name *Xanthovirus*. Finally, mimiviridae-related
567 VRs were identified in *S. latissima* and *Pelvetia canaliculata*, but since they are partial proviruses and
568 do not appear to possess integrase genes, they may have originated from ancient endogenization
569 events, similar to those described in chlorophytes[71].
570
571 The identification of integrated NCVs across almost all brown algal orders and in closely related
572 outgroup taxa suggests that the lysogenic life cycle strategy of phaeoviruses is ancient and that giant
573 viral genomes have been integrating into the genomes of brown algae throughout the latters'
574 evolutionary history. This conclusion was supported by the phylostratigraphic analysis, which detected
575 the appearance of many novel virus-related genes dating back to the origin of the Phaeophyceae
576 (Figure S5A). Marked differences were detected in total VR size and NCV marker gene presence across
577 the brown algal genome set, and large differences were even detected between strains from the same
578 genus (between 24 and 992 kbp of VR in different *Ectocarpus* spp. for example; Figure 6A, Table S19).
579 These differences indicate dynamic changes in VR content over evolutionary time, presumably due, at
580 least in part, to differences in rates of viral genome integration, a process that can involve multiple,
581 separate insertion events[72], and rates of VR loss due to meiotic segregation[73]. In addition, the abundant
582 presence of partial proviruses and NCV fragments in brown algal genomes indicates that inserted VRs
583 can degenerate and fragment, probably also leading to VR loss over time. The identification of large-
584 scale viral genome insertion events over such a long timescale (at least 450 My[1]) suggests that NCVs
585 may have had a major impact on the evolution of brown algal genomes throughout the emergence of
586 the lineage, making this a truly unique model system to investigate the impact of genomic exchanges
587 on genome evolution.
588
589 The widespread presence of large quantities of viral genes in brown algal genomes creates a favourable
590 situation for recruitment of this genetic information by the algal host via HGT (provided the acquired
591 genes confer a selective advantage[74]) but clear evidence of this type of HGT event can be difficult to
592 obtain. However, phylogenetic evidence indicates that several *Ectocarpus* species 7 histidine kinases
593 were derived by HGT from viral insertions[75] and analysis of the Phaeoexplorer genomes supported this

594 hypothesis. Histidine kinases (HKs) are widespread in the stramenopiles but several classes of
595 membrane-localised HK were either only found in brown algae (CHASE domain HKs and HKs with an
596 extracellular domain resembling an ethylene binding motif[75]) or only in brown algae and closely-
597 related taxa (MASE1 domain HKs[75]) and appear to be absent from other stramenopile lineages (Figures
598 3A and S9A, Table S20). These classes of HK all exhibit a patchy pattern of distribution across the brown
599 algae and are often monoexonic suggesting possible multiple acquisitions from viruses via HGT
600 following integration of viral genomes into algal genomes (Figures 3A, S9A, S14C). Phylogenetic
601 analysis provided further support for a HGT origin for these classes of HK (Figure S14C).
602
603 **DISCUSSION**
604
605 The genome resource presented and analysed in this study provides an unprecedented overview of
606 genomic diversity across the full scope of the brown algal lineage. Comparative analysis of these
607 genomes has allowed insights into brown algal genome evolution across the entire evolutionary
608 history of the group, ranging from emergence of the lineage during the GOBE through diversification
609 of the brown algal orders during the Mesozoic and up until relatively recent evolutionary events at the
610 genus level during the Miocene.
611
612 These analyses identified a period of marked genome evolution concomitant with the emergence of
613 the brown algal lineage that correlates with an increase in multicellular complexity possibly driven, at
614 least in part, by increases in atmospheric oxygen and herbivory. These early genome evolution events
615 included enhanced levels of gene gain (including via HGT) and the formation of genes with novel
616 domain combinations. These genomic changes were important in equipping the brown algal ancestor
617 with key components of several metabolic pathways, notably cell wall polysaccharide, phlorotannin
618 and halogen metabolisms, that were essential for their colonisation of intertidal and subtidal
619 environments. The capacity to synthesise flexible and resilient alginate-based cell walls allows these
620 organisms to resist the hydrodynamic forces of wave action[60], whereas phlorotannins and halogen
621 derivatives are thought to play important roles in defence[76]. There is also evidence that cell wall cross-
622 linking by phlorotannins may be important for strong adhesion to substrata, another important
623 characteristic in the dynamic intertidal and subtidal coastal environments[77]. The capacity to adhere
624 strongly and resist both biotic and abiotic stress factors would prove essential for the success of large,
625 sedentary multicellular organisms in these intertidal niches over evolutionary time. Interestingly, the
626 early genome evolution events described above also set the stage for later genome evolution events
627 in the sense that the gene families that have amplified in the brown algae are significantly enriched in
628 genes that were gained during the emergence of the lineage.

629

630 The period of increased gene gain during the emergence of the brown algae was followed by a period
631 of overall gene loss that extended up until the present day (Figure S4B). Interestingly, similar periods
632 of ancestral gene gain followed by gene loss have also been observed for both the animal and land
633 plant lineages[78], indicating that this may be a common feature of multicellular eukaryotic lineages.
634

635 About 220 MY after the emergence of the brown algae, the aftermath of the Permian–Triassic mass
636 extinction event and the initiation of Pangea rifting appear to have created favourable environments
637 for rapid diversification of the main brown algal orders[44,45], resulting in the emergence of a diversity of
638 developmental, life cycle and reproductive strategies, with correlated effects on genome evolution. In
639 some orders, such as the Laminariales and Fucales, certain characteristics acquired during this period
640 of diversification are predicted to have been important prerequisites for the emergence of marine
641 forests. These characteristics include developmental features such as an intercalary meristem but also
642 modification of metabolic, defence and developmental processes, novelties which seem to have
643 arisen, as indicated by the present study, through order-specific amplifications of key gene families.
644

645 Analysis of the genomes of multiple *Ectocarpus* species demonstrated that genomic modifications,
646 including gene gain and gene loss have continued to occur up until the present time and indicated that
647 these modifications can potentially be transmitted between species as a result of gene flow occurring
648 within a genus due to incomplete reproductive boundaries and introgression.
649

650 Finally, one of the most surprising observations was that brown algal genomes contain many inserted
651 viral sequences corresponding to large DNA viruses of the *Phaeovirus* family. Inserted viral sequences
652 are widespread in eukaryotic genomes[79,80] and insertions corresponding to nucleocytoplasmic large
653 DNA viruses have been found in green algal genomes[71,81] but the brown algal *Phaeovirus* VRs are
654 remarkable because they are nearly ubiquitous in this lineage (being present in 67 of 69 brown algal
655 genomes analysed) and because individual genomes can contain several phylogenetically diverse
656 *Phaeovirus* insertions and insertions of a broad range of different sizes. The near ubiquitous occurrence
657 of these elements may be attributed to the capacity of phaeoviruses to insert into their hosts' genomes
658 as part of their life cycle, whereas the broad range of brown algal species that harbour these elements
659 indicates that *Phaeovirus* insertions have probably been occurring since the emergence of this lineage.
660

661 The above observations illustrate how the Phaeoexplorer genome dataset, along with the various
662 analyses carried out in this study, can be used to link the gene content of brown algal genomes to
663 biological processes and characteristics that have played key 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 characterised at the genome level. The Phaeoexplorer dataset not only provides good quality genome assemblies for many, previously uncharacterised 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 signalling genes implicated in a broad range of brown algal biological functions (growth and development, adhesion, reproduction and resilience to abiotic and biotic stresses) represents an important resource for future research programs aimed at optimising 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 by the promotion of highly sustainable cultivation practises.

To facilitate future use of this genome dataset, the annotated genomes have been made available through a website portal (https://phaeoexplorer.sb-roscoff.fr), which provides multiple additional resources including genome browsers, BLAST interfaces, transcriptomic data and metabolic networks. The existing genome dataset provides very good coverage of the phylogenetic diversity of the Phaeophyceae and reasonably complete gene catalogues 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[82,83], together with the other tools and resources currently available for the model brown alga *Ectocarpus*[84], provide the means to deploy the functional genomics approaches necessary to address this question.

ACKNOWLEDGMENTS

This work was supported by the France Génomique National infrastructure project Phaeoexplorer (ANR-10-INBS-09), the European Research Council project Sexsea (638240), the Investissements d'Avenir project Idealg (ANR-10-BTBR-04-01), the European BG-01 BlueGrowth H2020 project Genialg (727892), Laoshan Laboratory grants (LSKJ202203801, LSKJ202203204), the Taishan Scholars Program and Talent Projects of Distinguished Scientific Scholars in Agriculture, the CNRS international research network DABMA (00022), the ANR projects Epicycle (ANR-19-CE20-0028-01), BrownSugar (ANR-20-

CE44-0011), HaloGene (ANR-22-CE20-0025), Seabioz (ANR-20-CE43-0013) and BrownLincs (ANR-23-CE20-0048-01), the National Research Foundation of Korea (2022R1A2B5B03002312, 2022R1A5A1031361) granted to H.S.Y., the projects Connect Talent EpiAlg Région Pays de la Loire-Nantes Métropole and Etoiles Montantes M-EpiCC Région Pays de la Loire, the MITI-funded project Algometabionte, 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.S., X.F., L.Ma., N.Te., J.B.R., R.P., L.R., S.W.C., J.J., K.U., K.B., C.Du., P.R., A.L., A.E.M., M.L., A.K., P.H.G., C.V., S.S.A., S.A., K.A., Y.B., T.Ba., A.Be., S.B., A.Bo., A.Cor., H.C.C., A.D., E.Di., S.D., E.Dr., J.G., L.G., A.G., M.L.G., L.H., B.H., A.J., E.K., A.H.K., C.B., L.L., R.L., S.T.L., P.J.L., E.M., S.M., G.M., C.N., S.A.R., E.R., D.Sch., A.S., L.T., T.T., K.V., H.V., G.W., H.K., A.F.P., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.M.A. and J.M.C; Investigation: C.C., S.H., Z.N., N.Ta., A.Cou., B.N., W.B., E.De., C.J., L.Me., S.R. and D.Sco.; Resources: O.G., A.Cou., L.B.G., T.Br., R.A.C., C.De., S.F., W.J.H., G.H., K.K., A.L.B., K.M., C.M., N.P., P.P., S.R., D.Sco., H.V., F.W., H.K., A.F.P., M.V., E.C. and J.M.A.; Data Curation: O.G., C.C., A.Cou., L.B.G., J.M.A. and J.M.C.; Writing-Original Draft: F.D., O.G., J.M.A. and J.M.C.; Writing-Review & Editing: all authors; Visualization: F.D., O.G., L.B.G., L.D., D.M., T.M., D.S., X.F., L.Ma., J.B.R., R.P., L.R., K.U., K.B., P.R., A.L., M.L., P.H.G., C.V., A.D., A.L.B., H.K., E.C. and J.M.C.; Supervision: O.G., N.Te., M.L., R.A.C., H.C.C., O.D.C., S.D., G.H., A.J., C.B., E.P., P.L., S.A.R., A.S., L.T., H.V., G.W., H.K., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.M.A. and J.M.C.; Project administration: F.D., E.B., M.V., G.V.M., E.C., S.M.C., P.W., J.M.A. and J.M.C.; Funding acquisition: H.C.C., C.B., P.P., C.H., N.Y., S.M.C. and J.M.C.; Conceptualization: J.M.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Figure legends

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

733 The upper panel indicates approximate positions in the intertidal of key species whose genomes have
734 been sequenced by the Phaeoexplorer project. The lower panel illustrates the diversity of brown algae
735 in terms of number of cell types, thallus size, morphological complexity and life cycle type (maximal
736 values for each taxa). The panel also indicates a number of key evolutionary events that occurred
737 during the emergence of the Phaeophyceae and diversification of the brown algal orders. Some
738 lineages may have secondarily lost a characteristic after its acquisition. Note that members of the
739 genus *Ishige* (Ishigeaceae) also exhibit desiccation tolerance (not shown). *, these orders were not
740 analysed in this study; Cr, Cryogenian; Ed, Ediacaran; Ca, Cambrian; O, Ordovician; S, Silurian; D,
741 Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene.
742
743 Figure 2. Genome-wide analyses of brown algal genome and gene content evolution
744 (A) Time-calibrated cladogram based on Figure S3 with Phaeophyceae in brown and outgroup species
745 in black. The grey hatched area, which indicates key nodes corresponding to the origin and early
746 emergence of the brown algae, is mirrored in (B) to (F).
747 (B) Gene family (orthogroup) gain (green) and loss (red) during the emergence and diversification of
748 the brown algae based on a Dollo parsimony reconstruction (Figure S4).
749 (C) Upper panel: timing of gene family (orthogroup) amplification and reduction during the
750 evolutionary history of the Phaeophyceae (CAFE5 analysis). Lower panel: time of origin (orthogroup
751 gain, based on the Dollo parsimony reconstruction) of the 180 most strongly amplified gene families.
752 (D) Composite gene analysis. Proportions of gene families showing domain fusion (orange) or domain
753 fission (yellow) at different age strata.
754 (E) Inferred gene family founder events after accounting for homology detection failure.
755 (F) Horizontal-gene-transfer-derived genes in orthologous groups and across species. Taxa along the
756 y-axis are as in (A). Grey bars indicate the number of HGT genes (light-grey) and the number of
757 orthogroups containing HGT genes (dark-grey) for each species. The black trace represents the
758 percentage of genes resulting from HGT events per species. Pie charts summarise the predicted origins
759 (donor taxa) of the HGT genes. The right-hand bar graph indicates the proportions of ancestral (i.e.
760 acquired before the root of the phylogenetic class, in grey) and class-specific (i.e. acquired within the
761 phylogenetic class, in blue) HGT genes.
762 (G) Enrichment of COG categories in sets of gene families identified as being 1) gained at the four
763 indicated early nodes by the Dollo analysis, 2) gene founder events at the four indicated phylostrata,
764 3) amplified in the Phaeophyceae (180 most strongly amplified families), 4) domain fusions or fissions
765 and 5) HGT-derived. Asterisks indicate significantly enriched categories compared to the entire set of
766 gene families.

767 FDI clade, Fucophycideae/Dictyotales/Ishigeales; PS clade, Phaeophyceae plus Schizocladiophyceae;
768 PX clade, Phaeophyceae plus Xanthophyceae.
769
770 Figure 3. Gene family evolution during the emergence of the brown algal lineage and a focus on
771 carbohydrate metabolism
772 (A) Variations in size for a broad range of key gene families in the brown algae and closely-related taxa.
773 Numbers indicate the size of the gene family. Note that the *S. ischiensis* algal-type HPOs appear to be
774 intermediate between classes I and II. Brown tree branches, Phaeophyceae.
775 (B) Overview of information from the orthogroup Dollo analysis, the phylostratigraphy analysis, the
776 horizontal gene transfer analysis and the gene family amplification analysis for a selection of cell-wall
777 active protein (CWAP) families. Expert functional annotations been crossed with orthogroup (OG)
778 composition, each dot represents a functional family / orthogroup couple. The size of the dot is
779 proportional to the number of proteins annotated in the OG, and the colour represents the proportion
780 of the functional annotation that falls into this OG. CWAP annotations are arranged by broad functional
781 categories along the x-axis. Phylogenetic levels considered in the genome-wide analyses are indicated
782 on the left. PS, Phaeophyceae and FDI clade, identified as gene innovation stages, are highlighted in
783 brown. Functional categories with interesting evolutionary histories are highlighted in red.
784 (C) Phylogenetic tree of mannuronan C5-epimerases (ManC5-E), a key enzyme in alginate biosynthesis.
785 The inset circle represents a global ManC5-E phylogeny with three main clusters. The phylogeny shown
786 on the left, with the three clusters indicated, is representative of the global view.
787 (D) Phylogenetic tree of the polysaccharide lyase 41 (PL41) family, a key enzyme involved in alginate
788 degradation in brown algae. The green squares indicate proteins that have been characterised
789 biochemically. Brown algal sequences are colour-coded in relation to their taxonomy, as indicated in
790 (C). Sequences belonging to the sister group Schizocladiaphyceae are shown in red and with a red
791 circle.
792 P, present; A, absent; CAZYmes, carbohydrate-active enzymes; HPO, vanadium haloperoxidase; PKS,
793 type III polyketide synthase; TAPs, Transcription-associated proteins; EsV-1-7, EsV-1-7 domain
794 proteins; DNMT, DNA methylase; GTs, glycosyltransferases; GHs, glycoside hydrolases; ARF, auxin
795 response factor-related; bHLH, basic helix-loop-helix; HMG, high mobility group; Zn-clus, zinc cluster;
796 C2H2, C2H2 zinc finger; GNAT, Gcn5-related N-acetyltransferase; SNF2, Sucrose nonfermenting 2; LRR,
797 leucine-rich repeat; QAD, b-propeller domain; RK, membrane-localised receptor kinase; HK, histidine
798 kinase; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-
799 binding-domain-like; MASE1, membrane-associated sensor1 domain; DEK1, defective kernal1; MCU,
800 mitochondrial calcium uniporter; GLR, glutamate receptor; pLGIC, pentameric ligand-gated ion
801 channel; TRP, transient receptor potential channel; IMM, IMMEDIATE UPRIGHT; H3, histone H3; MAS,

802 mastigoneme proteins; AA, auxiliary activity; ECT, Ectocarpales; LAM, Laminariales; FUC, Fucales; DES,
803 Desmarestiales.
804
805 Figure 4. Evolution of key gene families during the emergence of the brown algal lineage
806 (A) Evolution of type III polyketide synthase (PKS) genes in the stramenopiles (left panel). Light blue,
807 blue and dark blue circles correspond to the three classes of type III PKS (i.e. PKS1, PKS2 and PKS3),
808 respectively. The number of circles indicates the number of gene copies and absence of a circle
809 indicates absence of PKS genes. Coloured dots on the tree nodes indicate phylogenetic levels. Numbers
810 indicate predicted evolutionary events. Right panel: condensed view of a phylogenetic reconstruction
811 tree of stramenopile PKS III and closely-linked sequences. Red, orange and black branches correspond
812 to brown algae, other stramenopiles and bacterial/eukaryotic outgroups, respectively. In brackets:
813 number of sequences identified in each phylogenetic group. Branch bootstrap support is indicated.
814 (B) Loss of orthogroups corresponding to flagellar proteome components[51] in eight brown algal species
815 from five orders. The drawings on the right indicate the cellular structures of zoids from the eight
816 species. Grey, nucleus; yellow, chloroplast; blue, anterior flagella with mastigonemes; red, eyespot.
817 The posterior flagellum is shown either in green to indicate the presence of green autofluorescence
818 correlated with the presence of the eyespot or in blue in species without an eyespot. Note the loss of
819 the eyespot in the Laminariales species *S. latissima* and *U. pinnatifida* and the loss of the entire
820 posterior flagellum in *D. dichotoma*. Bars below the heatmap indicate gene losses associated with loss
821 of just the eyespot (orange) or of the entire posterior flagellum (blue).
822
823 Figure 5. Evidence for gene flow within the genus *Ectocarpus*
824 (A) DensiTree visualisation of gene trees (grey lines) for 261 orthologues shared by 11 *Ectocarpus*
825 species and the outgroup species *S. promiscuus,* together with the consensus species tree (black lines).
826 The species tree is strongly supported, with all nodes displaying high posterior probabilities
827 consistently reaching 0.99 or 1.
828 (B) Box plot reporting the range of D-statistic (Patterson's D) values between P2 and P3 species. Within-
829 lineage comparisons (i.e. within clades 1 and 2) and between-lineage comparisons are distinguished
830 on the x-axis. The annotation of each dot indicates species that were designated as P2 and P3. For this
831 test, *Ectocarpus fasciculatus* was defined as the outgroup.
832
833 Figure 6. Inserted viral regions in brown algal genomes
834 (A) Annotated phylogeny summarising key statistics of the presence of nucleocytoviricota (NCV)
835 sequences in the genomes of brown algae and closely-related taxa. Eight genomes that were sourced
836 from public databases are labelled with an asterisk. Outer layers around tree are as follows: 1) NCV

genotypes in each genome, 2) NCV core gene count indicates the number of copies of each viral core gene (A32, A32 packaging ATPase; D5/D5p, D5 helicase/primase; MCP, major capsid protein; 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 (coloured 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.

Supplementary figure legends

Figure S1. Taxonomic diversity and assembly quality of the Phaeoexplorer genomes, 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 panel). BUSCO scores for the predicted proteome of each genome (middle panel). Correlation of CDS lengths for each species with the corresponding sequences from the *Ectocarpus* species 7 reference genome (lower panel). Previously published genomes are indicated in grey. longreads, genomes assembled using long reads.

F, female; M, male.

Figure S2. Geographic localisation of species and strains, related to Figure 1.

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

(B) North-south distributions of the species analysed 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.

871 Figure S3. Fossil-calibrated maximum-likelihood phylogenetic tree of the brown algal and closely-
872 related taxa analysed in this study, related to Figure 1.
873 Bayesian divergence time estimation using 32 nuclear protein sequences, together with two fossil
874 calibrations and a calibration for the root based on Choi *et al*.[1] (numbered blue circles). Grey bars on
875 the nodes show 95% highest density region (HPD) intervals of the node ages. The Great Ordovician
876 Biodiversification Event (GOBE) is shown as an orange box, the Permian-Triassic mass extinction as a
877 purple band and Pangea rifting as a blue box.
878
879 Figure S4. Dollo-logic-based analysis of gene family gain and loss during the emergence of the brown
880 algae, related to Figure 2.
881 (A) Cladogram indicating orthogroup (OG) gain and loss during the emergence of the Phaeophyceae
882 based on Dollo parsimony analysis. Taxonomic classes, orders and families are indicated in brown
883 (brown algae) or grey (outgroup species) on the right. The nodes of the tree (n0-n22) are numbered in
884 red, and listed on the left with the corresponding name, if one exists. The number of OGs predicted to
885 be present at each node is indicated by the circles and the branches are coloured according to overall
886 gene family gain.
887 (B) Heat map of COG functional categories associated with OGs gained at each node of the cladogram.
888
889 Figure S5. Genome-wide analyses of gene family evolution, related to Figure 2.
890 (A) Functional and structural features of founder genes from three taxonomic levels. FDI,
891 Fucophycideae/Dictyotales/Ishigeales; PS, Phaeophyceae plus Schizocladiophyceae; PX,
892 Phaeophyceae plus Xanthophyceae.
893 (B) Inference of HGT origins from 74 species based on monophyletic most similar homologue (MMSH)
894 analysis. The tree on the left, which is derived from the NCBI taxonomy tree, indicates the source taxa
895 for HGT-derived genes. The tree at the bottom of the figure, which was constructed using single-copy
896 genes, indicates the species that have received genes by HGT. The middle part of the figure indicates
897 the number of HGT genes transferred from each source to the receiver genomes. The panel on the
898 right illustrates the number of HGTs from each phylum. The legend in the lower left corner provides a
899 reference for the circle size, which corresponds to 10, 100 or 300 HGT gene counts.
900 (C) Composite gene analysis. Phylogenetic distribution of fused (blue), split (green) and non-
901 remodelled (grey) gene family originations across the evolution of brown algae (middle panel). Pie
902 charts on each branch of the phylogeny indicate the relative contribution of gene fusion and fission to
903 the overall emergence of novel gene families, quantified by the area of the circle. Brown algal species
904 are indicated in brown and other stramenopiles in black. Note that only the topology of the species
905 tree is displayed here, without specific branch lengths. Right panel: barplot indicating the percentage

906 of gene families retained in extant genomes among all gene families that emerged during the evolution
907 of the species set. Left panel: barplot representing the distribution of gene families in COG functional
908 categories for functionally annotated fused, split, and non-remodelled orthologous groups. The
909 functional annotation assigned to an orthogroup corresponded to the most frequent functional
910 category annotated for the members of each orthogroup. Asterisks next to the bars indicate
911 statistically significant differences between remodelled and non-remodelled gene families (*p*-value
912 <0.05, two-sided Chi^2 test with Yates correction).
913 (D) Left panel: gene families (orthogroups) significantly amplified (binomial test) in the brown algae
914 compared to outgroup taxa. Orthogroups amplified in specific groups of species are indicated by green
915 rectangles. Right panel: pie charts representing the proportions of manually-determined functional
916 categories for each group of amplified gene families highlighted in the left panel.
917 (E) Plot showing the timing (phylogenetic clade) of the amplification of gene families (Y axis) and the
918 timing of the appearance (gain) of the amplified gene family (X axis, based on the Dollo analysis of
919 orthogroups). Nodes (e.g. n2) are as indicated in Figure S4.
920
921 Figure S6. Structural features of the predicted genes in the Phaeoexplorer genomes, related to Figure
922 1.
923 (A) Genome and gene statistics for the 21 reference genomes (Figure S1A and Table S1). Violin plots
924 display size distributions. Cross or triangle, mean; diamond or square, median. For intergenic regions,
925 half violins on the left correspond to intergenic distances between adjacent genes on opposite strands,
926 and half violins on the right to intergenic distances between adjacent genes on the same strand.
927 (B) Intron acquisition during the emergence of the brown algal lineage (left panel). The analysis
928 examined 949 introns in 235 conserved orthologues. Colours correspond to the species distributions
929 of introns. Hatched blue indicates introns that are shared with at least two outgroups (ancestral
930 introns). The right panel shows an example of an intron conservation profile for the orthogroup
931 OG0004854. Colour code and species numbering as for the left panel. F, female; M, male.
932
933 Figure S7. Genome-scale metabolic network analyses, related to Figure 2.
934 (A-D). Multidimensional-scaling (MDS) plots of GSMNs reconstructed with AuCoMe using the presence
935 and absence of biochemical reactions in these networks to compute the distance matrix (Jaccard
936 index) used by the MDS.
937 (A) MDS computed using the draft metabolic networks created by the first step of AuCoMe with only
938 the annotation from the genomes on all species from the Phaeoexplorer dataset.

939 (B) MDS computed using metabolic networks created after the last step of AuCoMe, i.e. after reaction
940 propagation using orthologous genes and structural verification on genome sequence, on all genomes
941 from the Phaeoexplorer dataset.
942 (C) MDS computed on draft metabolic networks from a subset of high-quality assemblies.
943 (D) MDS computed after reaction propagation on a subset of high-quality assemblies.
944 Dim., dimension.
945 (E-G). Overview of losses and gains of metabolic components for different sets of brown algal species.
946 Each row corresponds to a species and columns correspond to sets of reactions. For each row, there
947 is a coloured block if the species contains the reactions present in this block and shared with the other
948 species that also have this block in the same column. Numbers at the top, along the bottom and to the
949 right indicate the number of species that possess each reaction, the number of reactions with this
950 profile and the number of reactions per species, respectively.
951 (E) All brown algae and stramenopile outgroups.
952 (F) Selected reference quality genome assemblies.
953 (G) Comparison of freshwater and marine species with reference quality genomes.
954
955 Figure S8. Analysis of metabolism gene families, related to Figure 3.
956 (A) Counts of numbers of genes predicted to encode glycosyltransferases (GTs), glycoside hydrolases
957 (GHs), polysaccharide lyases (PLs) and all CAZymes (GTs, GHs, PLs, Carbohydrate Esterases CEs,
958 Auxiliary Activities AAs, Carbohydrate Binding Modules CBMs), showing numbers for both full-length
959 proteins (dark colours) and fragments (light colours). The data is averaged by order with the number
960 of species analysed per order in brackets.
961 (B) Number of genes for selected CAZyme families in brown algae. Only CAZyme families with at least
962 three members per genome on average, are shown. The species analysed are the same as in (A).
963 Counts include full-length proteins and fragments.
964 (C) vHPO genes identified in the 21 reference genomes based on sequence homology and active site
965 conservation. vHPO genes are indicated by a green cross or a number and absence by a red cross.
966 (D) Maximum likelihood phylogenetic trees for 259 algal-type vHPOs (left) and for bacterial-type vHPOs
967 (right). Algal-type I vHPOs are coloured in blue and algal-type II vHPOs in violet. The clades that have
968 structurally or biochemically characterised enzymes are highlighted in red for vBPOs and in yellow for
969 vIPOs. Strongly supported representative branches have been collapsed. The clade names correspond
970 either to taxa or to individual gene names. For the bacterial vHPOs, the FastTree reconstruction tool
971 with 1000 bootstraps was drawn as a circular representation taking the gammaproteobacterial group
972 as the starting point to arbitrary root the tree. Brown algal branches are coloured in brown. Green dots
973 indicate bootstrap values of between 0.7 and 1.0 (1000 replicates).

974 (E) PKS III domain structures indicating amino- (IPR001099; green) and carboxy-terminal (IPR012328;
975 brown) chalcone/stilbene synthase domains and amino-terminal signal peptide (SP; violet), signal
976 anchor (SA; light blue), SP/SA hybrid (pink) or all three (SP, SA or SP/SA; black). The number of proteins
977 in each group is indicated at the carboxy end of the protein.
978
979 Figure S9. Evolution of membrane-localised signalling proteins in the brown algae, related to Figure 3.
980 (A) Presence or absence of different membrane-localised signalling proteins in brown algae and other
981 stramenopiles. (B) Diverse domain structures of stramenopile FG-GAP domain proteins showing
982 possible evolutionary relationships. LRR, leucine-rich repeat; QAD, b-propeller domain; EBD, ethylene-
983 binding domain-like; CHASE, cyclases/histidine kinases associated sensory extracellular domain;
984 MASE1, membrane-associated sensor1 domain; HK, histidine kinase; FG-GAP, phenylalanyl-glycyl–
985 glycyl-alanyl-prolyl domain; IPT, Ig-like plexins transcription factors domain; ITGA, α-integrin domain;
986 Cad, cadherin domain; DEK1, defective kernal1; FAS, fasciculin; SP, signal peptide; TM, transmembrane
987 domain.
988
989 Figure S10. Evolution associated with the diversification of biological traits across the brown algae,
990 related to Figure 1.
991 (A) Ancestral state reconstructions for number of cell types and for morphological complexity in the
992 diploid or haploid phase of the life cycle. Left, number of different cell types in the diploid phase (DP).
993 Centre, morphological complexity in the diploid phase. Right, morphological complexity in the haploid
994 phase (HP). The colours indicate the different states and the pie charts the likelihood of these states
995 at each node.
996 (B) Rates of gene evolution in relation to life cycle structure and developmental complexity. Left, violin
997 plots showing the distribution of omega (dN/dS), rates of non-synonymous substitution (dN) and rates
998 of synonymous substitution (dS) for brown algae with haplodiplontic or diplontic life cycles. Right,
999 violin plots showing the distribution of rates of non-synonymous substitution (dN) for brown algal
1000 species with filamentous, simple parenchymatous or elaborate parenchymatous thalli. The *p*-values
1001 are for pairwise (gene-by-gene) Wilcoxon tests and the percentage of genes that exhibited the same
1002 patterns of differences in dN/dS, dN or dS values as the median values are indicated. Significant
1003 differences are indicated by an asterisk.
1004 (C) Generation-biased gene expression in relation to life cycle dimorphism. Pie diagrams indicating the
1005 proportions of sporophyte-specific (dark red), sporophyte-biased (light red), gametophyte-specific
1006 (dark blue), gametophyte-biased (light blue) and unbiased (grey) genes in species with different
1007 haploid-diploid life cycles ranging from sporophyte-dominant through isomorphic to gametophyte-
1008 dominant. sp., species.

1009

1010 Figure S11. Amplification of transcription factor gene families, related to Figure 3.

1011 (A) Ancestral state reconstruction of the sizes of three TAP families: left, C2h2 zinc finger, centre, three-
1012 amino-acid loop extension homeodomain (HD_TALE) and, right, high mobility group (HMG). The colour
1013 gradient indicates the predicted number of genes in each family across the tree.

1014 (B) Alignment of the homeodomain regions of three-amino-acid loop extension homeodomain
1015 transcription factors (TALE HD TFs). Sequences in clusters 1 and 3 are orthologous to the ORO and SAM
1016 proteins of *Ectocarpus* species 7, respectively. Cluster 2 corresponds to sequences similar to the third
1017 TALE HD TF gene of *Ectocarpus* species 7. Cluster 4 contains the sequences from closely-related
1018 outgroup taxa that could not be clearly assigned orthology to any of the *Ectocarpus* species 7 proteins.

1019

1020 Figure S12. Evolution of LHCX genes, related to Figure 3.

1021 (A) LHCX cluster genomic context in ten brown algae and *Schizocladia ischiensis*. Text above and below
1022 the line indicates the chromosome (chr) or contig (ctg) and the gene number, respectively.

1023 (B) Maximum likelihood (RAxML) phylogenetic tree of all LHCX and selected FCP protein sequences
1024 from the algal species in panel (A). Sequences are labelled with the first letters of the genus and species
1025 name, contig and gene number. Clustered LHCX genes are additionally numbered as in panel (A). Three
1026 groups of conserved LHCX paralogs encoded by unclustered, unique genes are highlighted by blue
1027 background.

1028 (C) Examples of LHCX gene clusters (LHCX genes in red) in four brown algal species where the LHCX
1029 genes are located on opposite strands of the chromosome and overlap, either partially, with some of
1030 their exons being located in the intron of the gene on the opposite strand (*C. linearis, S. latissima*), or
1031 completely, with the gene being located entirely within the intron of the gene on the opposite strand
1032 (*E. crouaniorum, S. promiscuus*).

1033

1034 Figure S13. Microevolution of genomes within the genus *Ectocarpus,* related to Figure 5.

1035 (A) Fossil-calibrated phylogenetic tree for 11 *Ectocarpus* species. Extracted from the fossil-calibrated
1036 tree shown in Figure S3. Numbered grey bars on the nodes show 95% HPD intervals of the node ages.

1037 (B) Unrooted parsimony splits network of 11 *Ectocarpus* species. The tree was generated based on 261
1038 concatenated orthologous genes using the ParsimonySplits method in Splitstree[85]. ParsimonySplits
1039 networks accommodate phylogenetic incongruities in the data by incorporating alternative branch
1040 splits. Conflicting splits are displayed as box-like structures.

1041 (C) Principal Correspondence Analysis (PCoA) of distance branching patterns (Robinson-Foulds metric)
1042 emphasizing the node for *Ectocarpus* species 5, species 6 and species 7 in the 261 gene trees. The plot
1043 displays the first two principal axes, capturing the major patterns of variation in tree topology at this

1044 node among the gene trees. Each dot represents a single gene tree, with the distance between points
1045 corresponding to the dissimilarity in their branching patterns. Dot colour indicates the cluster
1046 membership of gene trees based on the clustering analysis.
1047
1048 Figure S14. Phylogenetic trees of viral and putatively virus-derived genes, related to figure 6.
1049 (A) and (B) *Nucleocytoviricota* major capsid protein and DNA polymerase phylogenies, respectively.
1050 The trees were both generated from aligned amino acid sequences with the Q.yeast+F+R6 model and
1051 1000 bootstraps in IQ-TREE. Subgroup labels refer to *Phaeovirus* genotypes. All sequences identified
1052 by this study are in bold, and highlighted in yellow if they clustered outside the *Phaeovirus*. Phy.,
1053 *Phycodnaviridae*.
1054 (C) Phylogenetic trees for three classes of histidine kinase. Membrane-associated sensor1 domain
1055 (MASE1) class: brown algal genes are more closely related to viral than to closely-related outgroup
1056 species genes, EsuBft1789.4 is located within a viral clade. Ethylene-binding-domain-like (EBD) class:
1057 viral-related clade limited to Ectocarpales. Cyclases/histidine kinases associated sensory extracellular
1058 domain (CHASE) class: complex pattern suggesting possible multiple HGTs. Genes are classed as algal
1059 (green or light green background label, strong or weak prediction) or inserted viral sequences (blue or
1060 light blue background label, strong or weak prediction) based on exon number, expression level and
1061 genomic context (neighbouring monoexonic or multiexonic genes). See Table S20 for the gene name
1062 abbreviations. Branch colours signify brown algae (green), closely-related outgroup taxa (violet) or
1063 EsV-1 genes (blue). TPM, maximum TPM; asterisks, gene located in an identified VR; nd, not
1064 determined.
1065
1066 Figure S15. Genome assembly procedures, related to Figure 1.
1067 (A) Short read assembly procedure.
1068 (B) Long read assembly procedure.
1069
1070 Figure S16. Genome structural characteristics, related to Figure 1.
1071 (A) Ancestral state reconstruction of genome size (left in bp) and GC content (right as percent). The
1072 colour gradients indicate the genome size and GC content across the tree.
1073 (B) Cumulative sequence covered by different features for each genome assembly: masked (repeated
1074 sequence) and unmasked regions (upper graph), intergenic and gene regions (middle graph) and intron
1075 UTR and CDS (lower graph).
1076
1077 Figure S17. Genome-wide analyses of composite and HGT-derived genes, related to Figure 2.
1078 (A) Functional annotation status of fused, split and non-remodelled gene families (orthogroups).

1079 (B) Example of the emergence of a gene family with a novel domain structure just prior to the
1080 emergence of the Phaeophyceae lineage. In brown algae and *S. ischiensis* orthogroup OG0007889
1081 contains members with a novel domain structure combining domains (Interpro domains IPR041337
1082 and IPR001660) found independently in orthogroups OG0006687 and OG0001104.
1083 (C) Comparison of GC content and gene structure for different components of HGT-derived (HGT) and
1084 non-HGT-derived (core) genes. tssup, 5' untranslated region; tssdown, 3' untranslated region.
1085 (D) Proportions of COG functional categories for HGT-derived and all genes. Numbers represent the
1086 percentage of the COG category for the considered gene set.
1087
1088 Figure S18. Intron conservation and spliceosome components, related to Figure 2.
1089 (A) Intron conservation across a set of 235 conserved single-copy orthologous genes. The upper
1090 histogram shows the numbers of introns shared by the groups of species linked by black dots in the
1091 lower panel.
1092 (B) Conservation of intron phase and size of the introns from each species compared with the
1093 *Ectocarpus* species 7 reference genome.
1094 (C) Sequence logos of the donor and acceptor sites of introns from six selected species.
1095 (D) Conservation of intron positions in *Ectocarpus* species and the outgroup *S. promiscuus* (right panel
1096 and insert).
1097 (E) Phylogenetic tree of eukaryotic Lsm/Sm proteins. The tree was rooted using midpoint rooting. The
1098 colour code for the branches correspond to SmD3B (light green), SmD3A (dark green) and Lsm4 (grey).
1099 (F) Phylogenetic tree of Lsm13 and Lsm14 proteins. The tree was rooted using midpoint rooting.
1100
1101 Figure S19. Analysis of long non-coding RNA genes in nine brown algal genomes and two closely-
1102 related taxa, related to Figure 1.
1103 (A) Relative proportions of long non-coding RNA (lncRNA-like) and protein-coding (Coding-like) genes.
1104 M, male; F, female.
1105 (B) Distributions of transcript length (left), length of longest ORF (middle) and percent GC (right) for
1106 lncRNA and protein-coding (mRNA) genes.
1107
1108 Figure S20. Analysis of organellar genomes, related to Figure 4.
1109 (A) Loss of plastid genes during the evolution of the brown algae. Gene loss events were identified for
1110 five genes, with two of the genes (rbcR and rpl32) being lost more than once.
1111 (B) Maximum likelihood phylogenetic trees for the brown algae based on genes from different genomic
1112 compartments. Phylogenetic tree based on plastid genes (top), phylogenetic tree based on
1113 mitochondrial genes (middle), discordance between the nuclear and plastid phylogenies and the

1114 mitochondrial phylogeny for the Ectocarpales (bottom). Ultrafast bootstrap values are indicated for
1115 1000 replicates, with an asterisk indicating 100%.
1116
1117 Figure S21. Evolution of transcription-associated protein (TAP) families in the brown algae, related to
1118 Figure 3.
1119 (A) Expansions, contractions, gains and losses of TAP families during the emergence of the brown
1120 algae. The number of gains (dark green), losses (dark red), expansions (light green) and contractions
1121 (light red) at each node is shown and the families involved are indicated using the same colour code.
1122 The light brown box indicates the Phaeophyceae.
1123 (B) Distribution of transcription factors (TFs) across the brown algae and closely-related taxa. The
1124 bubbles indicate the relative abundance of 32 different TF families.
1125 (C) Distribution of transcriptional regulators (TRs) across the brown algae and closely-related taxa. The
1126 bubbles indicate the relative abundance of 40 different TR families.
1127
1128 Figure S22. Histone structure and evolution, related to Figure 4
1129 (A) Histone H3.1 and H3.3 isoforms differ at two residues located in the amino-terminal tail (31-32 AT
1130 for H3.1 and TA for H3.3, in green), three residues in the α2 helix of the histone fold domain (86-90
1131 GSAVL for H3.1 and STAIL for H3.3) and one residue at the carboxy-terminal position (S or A,
1132 respectively in H3.1 and H3.3). Positions refer to the mature protein without the initial methionine.
1133 (B) CenH3 proteins differ in their length and the amino acid composition of their amino-terminal tails.
1134 Strictly and highly conserved residues are depicted in black and red, respectively. The characteristic
1135 CATD (CENP-A targeting domain) is indicated by a red line. Positions refer to the mature protein
1136 without the initial methionine.
1137 (C) Phylogenetic tree of histone H3.1 and H3.3 proteins of brown algae and other eukaryotes. Brown
1138 algae, diatoms, red algae, green algae, plant, animal and additional unicellular organisms (the
1139 myxomycetes *Physarum polycephalum* and *Dictyostellium discoideum*, the ciliate *Tetrahymena*
1140 *thermopila* and the yeast *Saccharomyces cerevisiae*) are depicted in brown, red, pink, light green, dark
1141 green, blue and black, respectively. See Methods section for species abbreviations.
1142 (D) Schematic presentation of brown algal histone H1 linker proteins. The H1 variants are highly
1143 divergent across brown algae species. Strictly and highly conserved residues are depicted in black and
1144 red, respectively. Positions refer to the mature protein without the initial methionine.
1145
1146 Figure S23. Conservation and microevolution of genomes within the genus *Ectocarpus,* related to
1147 Figure 5.

1148 (A) Dotplot illustrating genome-wide synteny between the *Ectocarpus* species 7 and *E. crouaniorum*
1149 genome assemblies. Horizontal lines delimit *E. crouaniorum* contigs.
1150 (B) Gene family (orthogroup) gain and loss for the *Ectocarpus* genus based on a Dollo parsimony
1151 analysis. Cladogram branches are coloured according to the overall gain (+) or loss (-) of gene families.
1152 Grey circles on the nodes indicate the number of gene families predicted to be present.
1153 (C) Predicted functions of genes in orthogroups gained within the *Ectocarpus* genus.
1154 (D) Proportions of genes of different evolutionary ages, as estimated by phylostratigraphy analysis, on
1155 each chromosome of the *Ectocarpus* species 7 genome. Colour-coded groups of genes correspond to
1156 the following phylostratigraphy categories: 1, cellular organisms to phylum (ranks 1 to 5; orange); 2,
1157 PX clade to Phaeophyceae/*Schizocladia ischiensis* (ranks 6 to 7; yellow); 3, Phaeophyceae to subclass
1158 (ranks 8 to 10; green); 4, superorder to family (ranks 11 to 13; blue); 5, genus to species (ranks 14 to
1159 15; pink).
1160 (E) Distribution of synonymous (dS) and non-synonymous (dN) substitution rates for syntenic genes
1161 between *Ectocarpus* species 7 and *E. siliculosus*, *E. crouaniorum* or *E. fasciculatus* in relation to gene
1162 age based on the phylostratigraphy analysis. Correspondence between the colour-coded gene groups
1163 and age ranks are indicated on the left and are the same as in (D).
1164
1165 Figure S24. Evolution of signalling-related genes in the brown algae, related to Figure 3.
1166 Overview of phytohormone biosynthesis pathways and distribution of gene families associated with
1167 biosynthesis across the brown algae and other stramenopile taxa. Left, biosynthesis pathways for the
1168 phytohormones abscisic acid (ABA), brassinosteroids (BR), cytokinins (CK; including iP,
1169 isopentenyladenine; tZ, trans-zeatin; DZ, dihydrozeatin; and cZ, cis-zeatin), ethylene (Eth), gibberellins
1170 (GA), auxin (IAA), jasmonic acid (JA), salicylic acid (SA) and strigolactones (SL). Right, Presence of
1171 putative homologs of phytohormone biosynthethic enzymes in brown algae based on EggNOG v5.0
1172 gene families (root node). Colours denote z-score of the number of eggNOG family (left of heatmap)
1173 members. White denotes absence, (+) and (?) indicate that phytohormones have been reported to be
1174 present or have not been reported for brown algae (reviewed in[83,84]). ZEP, zeaxanthin epoxidase;
1175 NCED, nine-cis-epoxycarotenoid deoxygenate; SDR, short-chain dehydrogenase reductase; AAO,
1176 aldehyde oxidase; DET2, deetiolated2 (steroid 5alfa reductase); DWF4, dwarf4 (CYP90B); CPD,
1177 constitutive photomorphogenic dwarf (CYP90A); IPT, isopentenyl transferase; LOG, lonely guy (lysine
1178 de-carboxylase); SAMS, s-adenosyl methionine synthetase; ACS, ACC synthase; ACO, ACC oxidase; CPS,
1179 CDP/ent-kaurene synthase; KO, ent-kaurene oxidase; KAO, ent-kaurenoic acid oxidase; TAA,
1180 tryptophan aminotransferase; YUC, YUCCA (flavin monooxygenase); AMI, amidase; NIT, nitrilase; LOX,
1181 lipoxygenase; AOS, allene oxide synthase; AOC, allene oxide cyclase; OPR, oxo-phytodienoate
1182 reductase; ICS, isochorismate synthase; PAL, phenylalanine ammonia-lyase; D27, dwarf27 (all-trans/9-

1183 cis-B-carotene isomerase); CCD, carotenoid cleavage dioxygenase; MAX1, more axillary branches1
1184 (CYP711A).
1185
1186
1187 **Supplementary tables**
1188
1189 Table S1. List of strains used for the project, the genome assemblies generated and accession numbers
1190 for sequence data.
1191 (A) Strains used in the project.
1192 (B) Genome assemblies.
1193 (C) Statistical tests of correlations between genome assembly size and various genome features.
1194 (D) Genomic data used in the project including both genomic DNA and RNA-seq data. For the latter,
1195 DE RNA-seq indicates replicated samples that were generated by experiments aimed at measuring
1196 differential gene expression.
1197 (E) Genome assemblies and annotations generated by the project.
1198
1199 Table S2. Phylostratigraphy analysis.
1200 (A) Gene ages estimated by phylostratigraphy. (B) Gene family founder events. (C) Gene ages after the
1201 homology detection failure correction. (D) Counts of founder events after the homology detection
1202 failure correction. (E) Statistics for *Ectocarpus* species 7.
1203
1204 Table S3. Gene families (orthogroups) significantly amplified in Phaeophyceae genomes compared to
1205 closely-related taxa.
1206 OG, orthogroup; PHAEO, Phaeophyceae; FDI, Fucophycidae/Dictyotales/Ishigeales clade; LAMIN,
1207 Laminariales; ECTO, Ectocarpales.
1208
1209 Table S4. Core metabolic reactions most abundant in brown algae.
1210 Metacyc IDs and reaction names (with EC numbers) for 24 genes present in all brown algae and less
1211 than 70% of outgroup species. Gene IDs can be retrieved from the GSMN Wikis using reaction IDs.
1212
1213 Table S5. Counts of CAZYme gene family members in brown algal and closely-related taxa genomes.
1214 Asterisks indicate species that were included in Figures S8A and S8B.
1215
1216 Table S6. CAZYmes identified in brown algal and closely-related taxa genomes.
1217 ProteinID, corresponds to the locusID; Description, CAZYme gene family.

1218

1219 Table S7. Algal-type vanadium haloperoxidase proteins encoded by brown algal and closely-related
1220 taxa genomes plus representative sequences from more distant taxa.
1221

1222 Table S8. Bacterial-type vanadium haloperoxidase proteins encoded by brown algal and closely-related
1223 taxa genomes plus representative sequences from more distant taxa.
1224

1225 Table S9. Type III polyketide synthase proteins encoded by brown algal and closely-related taxa
1226 genomes plus representative sequences from more distant taxa.
1227

1228 Table S10. Receptor kinase proteins encoded by brown algal and closely-related taxa genomes.
1229 mod, modified gene model; new, new gene model.
1230

1231 Table S11. DNA methyltransferase proteins encoded by brown algal and closely-related taxa genomes.
1232 Genome coordinates and protein sequences are provided for modified (mod) gene models.
1233

1234 Table S12. Predicted reference proteomes of the ten species analysed for generation-biased gene
1235 expression indicating sporophyte- and gametophyte-biased genes.
1236

1237 Table S13. Presence of flagellar protein genes in the genomes of nine brown algal species.
1238 P, present; A, absent.
1239

1240 Table S14. Mannitol cycle enzymes encoded by brown algal and closely-related taxa genomes.
1241 M1PDH, mannitol 1-phosphate dehydrogenase; M1Pase, mannitol 1-phosphate phosphatase; HK,
1242 hexokinase; M2DH, mannitol 2-dehydrogenase.
1243

1244 Table S15. Summary of the TAPscan output with lists of annotated transcription-associated proteins
1245 (TAPs) for brown algal and closely-related taxa genomes.
1246

1247 Table S16. Ion channels encoded by brown algae and other stramenopiles.
1248 (A) Query sequences used for the screens. (B) Ion channel proteins detected. (C) Number of ion
1249 channels of each class per species. F, female; M, male; MCU, mitochondrial calcium uniporter; GLR,
1250 glutamate receptor; pLGIC, pentameric ligand-gated ion channel; TRP, transient receptor potential
1251 channel
1252

1253 Table S17. EsV-1-7 domain proteins encoded by brown algal and closely-related taxa genomes.
1254 Genome coordinates and protein sequences are provided for modified (mod) or newly created (new)
1255 gene models.
1256
1257 Table S18. Rates of synonymous and non-synonymous substitutions based on alignments of genes in
1258 syntenic blocks between *Ectocarpus* species 7 and four other *Ectocarpus* species.
1259
1260 Table S19. Inserted viral sequences in brown algal and closely-related taxa genomes.
1261 (A) Summary data for figure 5. (B) List of viral genes. (C) List of viral regions. (D) Number of viral regions
1262 detected in each genome.
1263
1264 Table S20.Histidine kinase proteins encoded by brown algal and closely-related taxa genomes.
1265 TM, transmembrane domain; CHASE, cyclases/histidine kinases associated sensory extracellular
1266 domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor1 domain.
1267
1268 Table S21. Genomes studied for the various analyses carried out within the project.
1269 Y, genome analysed.
1270
1271 Table S22. Metabolism genes lost from either *Pleurocladia lacustris* or *Porterinema fluviatile* or from
1272 both, based on GSMN analysis.
1273
1274 Table S23.Histone proteins encoded by brown algae and other stramenopiles.
1275 (A) Histone protein sequences. (B) Counts of histone proteins per genome. (C) Counts of histone genes
1276 per genome.
1277
1278 Table S24. *Porterinema fluviatile* genes differentially expressed under freshwater compared to
1279 seawater culture conditions.
1280
1281 Table S25. Tests for signatures of introgression among *Ectocarpus* species. Patterson's D-statistic
1282 (ABBA-BABA tests) was calculated for concatenated alignments of 261 ortholog genes and significance
1283 was detected using a block-jackknifing approach with a block size of 5 Kbp. *Ectocarpus fasciculatus* was
1284 used as the outgroup taxon for all ABBA-BABA tests (noted as O in the four-taxon fixed phylogeny
1285 scheme: (((P1,P2)P3)O)). All values are significant.
1286
1287 Table S26. Spliceosome components identified in brown algal and closely-related taxa genomes.

1288 M, male; F, female.
1289
1290 Table S27. Statistics for the long non-coding RNA content of 11 genomes.
1291 M, male; F, female.
1292
1293 Table S28. Characteristics of plastid and mitochondrial genomes for 33 brown algae and
1294 *Chrysoparadoxa australica*.
1295
1296 Table S29. Presence or absence of 141 shared plastid genes across the brown algae.
1297 O, present; X, absent.
1298
1299 Table S30. Sulphatase proteins encoded by brown algal and closely-related taxa genomes.
1300
1301 Table S31. Integrin proteins encoded by brown algal and closely-related taxa genomes.
1302 Genome coordinates and protein sequences are provided for modified (mod) or newly created (new)
1303 gene models. SP, signal peptide; TM, transmembrane domain; INTA, integrin alpha domain
1304
1305 Table S32. DEK1-like proteins encoded by brown algal and closely-related taxa genomes.
1306
1307 Table S33. Fasciclin proteins encoded by brown algal and closely-related taxa genomes.
1308 Genome coordinates and protein sequences are provided for modified (mod) gene models. FAS1,
1309 fasciclin domain.
1310
1311 Table S34. Tetraspanin and tetraspanin-like proteins encoded by brown algal and closely-related taxa
1312 genomes.
1313
1314 Table S35. Comparative analysis of genes that were differentially expressed in freshwater compared
1315 with seawater in *E. subulatus* and *P. fluviatile* based on whether the genes are shared orthologues or
1316 lineage-specific.
1317
1318 Table S36. Blocks of syntenic genes shared between *Ectocarpus* species 7 and four other *Ectocarpus*
1319 species.
1320
1321 Table S37. Putative components of phytohormone biosynthetic pathways in brown algae.
1322

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, S0960-9822(23)01769-4. https://doi.org/10.1016/j.cub.2023.12.069.

2. Keeling, P.J. (2009). Chromalveolates and the evolution of plastids by secondary endosymbiosis. The Journal of Eukaryotic Microbiology *56*, 1–8.

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. Seminars in Cell & Developmental Biology *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. (2008). The biology of phenolic containing vesicles. Algae *23*, 163–175.

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.

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., Clerck, O.D., Cock, J.M., Coelho, S.M., Destombe, C., Valero, M., et al. (2020). Phylogeny and Evolution of the Brown Algae. Critical Reviews in Plant Sciences *39*, 281–321. 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., Badger, J., 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*. Nature Ecology & Evolution. 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., Scornet, D., Petit, J., Ségurens, B., Da Silva, C., Corre, E., Dondrup, M., Glatting, K., 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.

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. Molecular Plant *0*. 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.

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. 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.

30. 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.

31. 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.

32. 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.

33. Meslet-Cladière, L., Delage, L., Leroux, C.J., Goulitquer, S., Leblanc, C., Creis, E., Gall, E.A., Stiger-Pouvreau, V., Czjzek, M., and Potin, P. (2013). Structure/Function Analysis of a Type III Polyketide Synthase in the Brown Alga *Ectocarpus siliculosus* Reveals a Biochemical Pathway in Phlorotannin Monomer Biosynthesis. Plant Cell *25*, 3089–3103. https://doi.org/10.1105/tpc.113.111336.

34. 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.

35. 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.

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

37. 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.

38. 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.

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

40. 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.

41. 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.

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. Geophysical Journal International *178*, 1078–1097. https://doi.org/10.1111/j.1365-246X.2009.04186.x.

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

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

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

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. 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.

52. 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. Global Ecology and Biogeography *31*, 636–648. https://doi.org/10.1111/geb.13450.

53. 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.

54. 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.

55. 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. Molecular Phylogenetics and Evolution *136*, 138–150. https://doi.org/10.1016/j.ympev.2019.04.012.

56. 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 U S A *121*, e2317054121. https://doi.org/10.1073/pnas.2317054121.

57. Basson, P.W. (1981). Late cretaceous alga, *Delesserites libanensis* sp. nov. Review of Palaeobotany and Palynology *33*, 363–370. https://doi.org/10.1016/0034-6667(81)90093-2.

58. 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, 105–111.

59. Barthel, K.W., Swinburne, N.H.M. (Nicola H.M., and Conway Morris, S. (Simon) (1994). Solnhofen : a study in Mesozoic palaeontology (Cambridge University Press).

1487 60. Martone, P.T., Kost, L., and Boller, M. (2012). Drag reduction in wave-swept macroalgae: Alternative
1488 strategies and new predictions. American Journal of Botany *99*, 806–815.
1489 https://doi.org/10.3732/ajb.1100541.

1490 61. Colin, C., Leblanc, C., Michel, G., Wagner, E., Leize-Wagner, E., Van Dorsselaer, A., and Potin, P. (2005).
1491 Vanadium-dependent iodoperoxidases in *Laminaria digitata*, a novel biochemical function diverging from
1492 brown algal bromoperoxidases. J Biol Inorg Chem *10*, 156–166.

1493 62. Zambounis, A., Elias, M., Sterck, L., Maumus, F., and Gachon, C.M. (2012). Highly dynamic exon shuffling in
1494 candidate pathogen receptors... What if brown algae were capable of adaptive immunity? Mol Biol Evol *29*,
1495 1263–1276. https://doi.org/10.1093/molbev/msr296.

1496 63. Macaisne, N., Liu, F., Scornet, D., Peters, A.F., Lipinska, A., Perrineau, M.-M., Henry, A., Strittmatter, M.,
1497 Coelho, S.M., and Cock, J.M. (2017). The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel
1498 family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development *144*, 409–
1499 418. https://doi.org/10.1242/dev.141523.

1500 64. Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., Price, D.K., Waddell, P.J., Lang,
1501 M., Courtier-Orgogozo, V., et al. (2022). Widespread introgression across a phylogeny of 155 Drosophila
1502 genomes. Current Biology *32*, 111-123.e5. https://doi.org/10.1016/j.cub.2021.10.052.

1503 65. Peters, A.F., Marie, D., Scornet, D., Kloareg, B., and Cock, J.M. (2004). Proposal of *Ectocarpus siliculosus*
1504 (Ectocarpales, Phaeophyceae) as a model organism for brown algal genetics and genomics. J Phycol *40*, 1079–
1505 1088.

1506 66. Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H.H., El Taher, A., Indermaur, A., Malinsky, M., Ricci,
1507 V., Kahmen, A., et al. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. Nature
1508 *589*, 76–81. https://doi.org/10.1038/s41586-020-2930-4.

1509 67. Stull, G.W., Qu, X.-J., Parins-Fukuchi, C., Yang, Y.-Y., Yang, J.-B., Yang, Z.-Y., Hu, Y., Ma, H., Soltis, P.S., Soltis,
1510 D.E., et al. (2021). Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution
1511 in gymnosperms. Nat Plants *7*, 1015–1025. https://doi.org/10.1038/s41477-021-00964-4.

1512 68. Rohling, E.J., Yu, J., Heslop, D., Foster, G.L., Opdyke, B., and Roberts, A.P. (2021). Sea level and deep-sea
1513 temperature reconstructions suggest quasi-stable states and critical transitions over the past 40 million years.
1514 Sci Adv *7*, eabf5326. https://doi.org/10.1126/sciadv.abf5326.

1515 69. Müller, D.G., and Knippers, R. (2011). Phaeovirus. In The Springer Index of Viruses, C. Tidona and G. Darai,
1516 eds. (Springer), pp. 1259–1263. https://doi.org/10.1007/978-0-387-95919-1_205.

1517 70. McKeown, D.A., Stevens, K., Peters, A.F., Bond, P., Harper, G.M., Brownlee, C., Brown, M.T., and Schroeder,
1518 D.C. (2017). Phaeoviruses discovered in kelp (Laminariales). ISME J *11*, 2869–2873.
1519 https://doi.org/10.1038/ismej.2017.130.

1520 71. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread
1521 endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145.
1522 https://doi.org/10.1038/s41586-020-2924-2.

1523 72. Stevens, K., Weynberg, K., Bellas, C., Brown, S., Brownlee, C., Brown, M.T., and Schroeder, D.C. (2014). A novel
1524 evolutionary strategy revealed in the phaeoviruses. PLoS One *9*, e86040.
1525 https://doi.org/10.1371/journal.pone.0086040.

1526 73. Bräutigam, M., Klein, M., Knippers, R., and Müller, D.G. (1995). Inheritance and meiotic elimination of a virus
1527 genome in the host *Ectocarpus siliculosus* (phaeophyceae). J Phycol *31*, 823–827.

1528 74. Keeling, P.J. (2024). Horizontal gene transfer in eukaryotes: aligning theory with data. Nat Rev Genet *25*, 416–
1529 430. https://doi.org/10.1038/s41576-023-00688-5.

1530 75. Kabbara, S., Hérivaux, A., Dugé de Bernonville, T., Courdavault, V., Clastre, M., Gastebois, A., Osman, M.,
1531 Hamze, M., Cock, J.M., Schaap, P., et al. (2019). Diversity and Evolution of Sensor Histidine Kinases in
1532 Eukaryotes. Genome Biol Evol *11*, 86–108. https://doi.org/10.1093/gbe/evy213.

1533 76. Potin, P., Bouarab, K., Salaün, J.-P., Pohnert, G., and Kloareg, B. (2002). Biotic interactions of marine algae.
1534 Curr Opin Plant Biol *5*, 308–317. https://doi.org/10.1016/s1369-5266(02)00273-x.

1535 77. Tarakhovskaya, E.R. (2014). Mechanisms of bioadhesion of macrophytic algae. Russ J Plant Physiol *61*, 19–25.
1536 https://doi.org/10.1134/S1021443714010154.

1537 78. Domazet-Lošo, M., Široki, T., Šimičević, K., and Domazet-Lošo, T. (2024). Macroevolutionary dynamics of gene
1538 family gain and loss along multicellular eukaryotic lineages. Nat Commun *15*, 2663.
1539 https://doi.org/10.1038/s41467-024-47017-w.

1540 79. Holmes, E.C. (2011). The Evolution of Endogenous Viral Elements. Cell Host & Microbe *10*, 368–377.
1541 https://doi.org/10.1016/j.chom.2011.09.002.

1542 80. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host
1543 biology. Nat Rev Genet *13*, 283–296. https://doi.org/10.1038/nrg3199.

1544 81. Moniruzzaman, M., Erazo-Garcia, M.P., and Aylward, F.O. (2022). Endogenous giant viruses contribute to
1545 intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. Virus Evolution *8*,
1546 veac102. https://doi.org/10.1093/ve/veac102.

1547 82. Badis, Y., Scornet, D., Harada, M., Caillard, C., Godfroy, O., Raphalen, M., Gachon, C.M.M., Coelho, S.M.,
1548 Motomura, T., Nagasato, C., et al. (2021). Targeted CRISPR-Cas9-based gene knockouts in the model brown
1549 alga *Ectocarpus*. New Phytol *231*, 2077–2091. https://doi.org/10.1111/nph.17525.

1550 83. Shen, Y., Motomura, T., Ichihara, K., Matsuda, Y., Yoshimura, K., Kosugi, C., and Nagasato, C. (2023).
1551 Application of CRISPR-Cas9 genome editing by microinjection of gametophytes of *Saccharina japonica*
1552 (Laminariales, Phaeophyceae). J Appl Phycol. https://doi.org/10.1007/s10811-023-02940-1.

1553 84. Cock, J.M. (2023). The model system *Ectocarpus*: Integrating functional genomics into brown algal research.
1554 Journal of Phycology *59*, 4–8. https://doi.org/10.1111/jpy.13310.

1555 85. Kloepper, T.H., and Huson, D.H. (2008). Drawing explicit phylogenetic networks and their integration into
1556 SplitsTree. BMC Evol Biol *8*, 22. https://doi.org/10.1186/1471-2148-8-22.

1557
1558

1559 **METHODS**
1560
1561 **KEY RESOURCES TABLE**
1562

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Descriptions of all sequenced samples have been deposited in the EBI/ENA database		EBI/ENA project PRJEB76691
Critical commercial assays		
OmniPrep Genomic DNA Purification Kit	G Biosciences, St. Louis, MO, USA	
Nucleospin Plant II midi DNA Extraction Kit	Macherey-Nagel, Düren, Germany	
NEBNext DNA Modules Products	New England Biolabs, Ipswich, MA, USA	
NEBNext Sample Reagent Set	New England Biolabs, Ipswich, MA, USA	
Ampure XP	Beckmann Coulter Genomics, Danvers, MA, USA	
Kapa Hifi Hotstart NGS library Amplification kit	Roche, Basel, Switzerland	
Short Read Eliminator Kit	Pacific Biosciences, Menlo Park, CA, USA	
1D Genomic DNA by Ligation	Oxford Nanopore Technologies Ltd, Oxford, UK	SQK-LSK109, SQK-LSK108 or SQK-LSK110
Qiagen RNeasy kit or the Macherey Nagel RNAplus kit	Macherey-Nagel, Düren, Germany	
TruSeq Stranded mRNA Sample Prep	Illumina	
NEBNext Ultra II Directional RNA Library Prep for Illumina	New England BioLabs	
Deposited data		
The sequence data generated by this project is described in Table S1.	This study.	EBI/ENA project 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.	https://doi.org/10.57745/9U1J85
Experimental models: Organisms/strains		
The strains used for genome and transcriptome sequencing are listed in Table S1D.	Culture collection references are provided where relevant.	See strain names and culture collection accessions for identifiers.
Software and algorithms		

MEGAHIT version 1.1.1	Li et al.[1]	RRID:SCR_018551 https://github.com/voutcn/megahit
MetaGene version 2008.8.19	Noguchi et al.[2]	http://metagene.cb.k.u-tokyo.ac.jp/
BLAST	Altschul et al.[3]	RRID:SCR_004870 http://blast.ncbi.nlm.nih.gov/Blast.cgi
Burrows-Wheeler Aligner	Li and Durbin[4]	RRID:SCR_010910 http://bio-bwa.sourceforge.net/
Bowtie2 version 2.3.5.1	Langmead and Salzberg[5]	RRID:SCR_016368 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SPAdes assembler version 3.8.1	Bankevich et al.[6]	RRID:SCR_000131 https://cab.spbu.ru/software/spades/
filtlong		RRID:SCR_024020 https://github.com/rrwick/Filtlong
Smartdenovo	Liu et al.[7]	RRID:SCR_017622 https://github.com/ruanjue/smartdenovo
Redbean	Ruan and Li[8]	
Flye	Kolmogorov et al.[9]	RRID:SCR_017016 https://github.com/fenderglass/Flye
Necat	Chen et al.[10]	https://github.com/xiaochuanle/necat
Racon	Vaser et al.[11]	RRID:SCR_017642 https://github.com/isovic/racon
Hapo-G	Aury et al.[12]	https://www.genoscope.cns.fr/hapog/
Metabat 2	Kang et al.[13]	RRID:SCR_019134 https://bitbucket.org/berkeleylab/metabat/src/master/
SortMeRNA	Kopylova et al.[14]	RRID:SCR_014402 http://bioinfo.lifl.fr/RNA/sortmerna/
Velvet version 1.2.07	Zerbino and Birney[15]	RRID:SCR_010755 http://www.molecularevolution.org/software/genomics/velvet
Oases version 0.2.08	Schulz et al.[16]	RRID:SCR_011896 http://www.ebi.ac.uk/~zerbino/oases/
TransDecoder	Haas, B.J.	RRID:SCR_017647 https://github.com/TransDecoder/TransDecoder
CDDsearch	Marchler-Bauer et al.[17]	

Trimmomatic version 0.38 and version 0.39	Bolger et al.[18]	RRID:SCR_011848 http://www.usadellab.org/cms/index.php?page=trimmomatic
Trinity version version 2.6.5	Grabherr et al.[19]	RRID:SCR_013048 http://trinityrnaseq.sourceforge.net/
rnaSPAdes version version 3.13.1	Bushmanova et al.[20]	RRID:SCR_016992 http://cab.spbu.ru/software/rnaspades/
RepeatMasker version 4.1.0	Smit et al.[21]	RRID:SCR_012954 http://repeatmasker.org/
Tandem repeats finder	Benson et al.[22]	RRID:SCR_022193 https://github.com/Benson-Genomics-Lab/TRF
REPET	Flutre et al.[23]	
BLAT	Kent et al.[24]	RRID:SCR_011919 http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Genewise	Birney et al.[25]	RRID:SCR_015054 http://www.ebi.ac.uk/Tools/psa/genewise/
DIAMOND version 0.9.30	Buchfink et al.[26]	RRID:SCR_009457 http://www.nitrc.org/projects/diamond/
Est2Genome	Mott et al.[27]	https://galaxy-iuc.github.io/emboss-5.0-docs/est2genome.html
Gmove	Dubarry et al.[28]	RRID:SCR_019132 http://www.genoscope.cns.fr/gmove
votingLNC		https://gitlab.com/a.debit/votinglnc
AliView version 1.26	Larsson[29]	RRID:SCR_002780 https://github.com/AliView
RAxML version 8.2.	Stamatakis[30]	RRID:SCR_006086 https://github.com/stamatak/standard-RAxML
Tracer version 1.7.2	Rambaut *et al*.[31]	RRID:SCR_019121 https://bioweb.pasteur.fr/packages/pack@Tracer@v1.6
OrthoFinder version 2.5.2	Emms and Kelly[32]	RRID:SCR_017118 https://github.com/davidemms/OrthoFinder
Count version 9.1106	Csűös[33]	https://www.iro.umontreal.ca/~csuros/gene_content/count.html
MUSCLE version 3.8.1551	Edgar[34]	RRID:SCR_011812 http://www.ebi.ac.uk/Tools/msa/muscle/

OD-Seq version 1.0	Jehl et al.[35]	https://bioconductor.org/packages/release/bioc/manuals/odseq/man/odseq.pdf
HMMER3 package versions 3.1b1 and 3.3.2	Mistry et al.[36]	RRID:SCR_005305 http://hmmer.janelia.org/
GenEra	Barrera-Redondo et al.[37]	
MCL	Enright et al.[38]	RRID:SCR_024109 https://micans.org/mcl/
Foldseek	Kempen et al.[39]	https://search.foldseek.com/search
CleanBlastp	Pathmanathan et al.[40]	
SEED		RRID:SCR_002129 http://www.theseed.org/wiki/Home_of_the_SEED
IPR2GO		http://www.ebi.ac.uk/interpro/search/sequence-search
eggNOG	Huerta-Cepas et al.[41]	RRID:SCR_002456 http://eggnog.embl.de
eggNOG-mapper	Cantalapiedra et al.[42]	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.[43]	RRID:SCR_011936 https://github.com/hyattpd/Prodigal
ViralRecall version 2.0	Aylward et al.[44]	https://github.com/faylward/viralrecall
esl-translate version 0.48		https://github.com/EddyRivasLab/easel/blob/master/miniapps/esl-translate.man.in
bedtools version 2.29.2	Quinlan and Hall[45]	RRID:SCR_006646 https://github.com/arq5x/bedtools2
MMseqs cluster version 13.45111	Hauser et al.[46]	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[47]	RRID:SCR_011811 http://mafft.cbrc.jp/alignment/server/
MEGA	Tamura et al.[48]	RRID:SCR_023017 https://www.megasoftware.net/
NGphylogeny platform		https://ngphylogeny.fr/.

TrimAl	Capella-Gutiérrez et al.[49]	RRID:SCR_017334 http://trimal.cgenomics.org/
TAPscan version 4	Petroll et al.[50,51]	https://plantcode.cup.uni-freiburg.de/tapscan/
Expasy web translator		RRID:SCR_024703 https://web.expasy.org/translate/
Geneious versions 11.0.5 and 11.1.5		RRID:SCR_010519 http://www.geneious.com/
Interproscan 94.0	Jones et al.[52]	RRID:SCR_005829 http://www.ebi.ac.uk/Tools/pfa/iprscan/
Clustal 2.1	Thompson et al.[53]	RRID:SCR_001591 http://www.ebi.ac.uk/Tools/msa/clustalo/
Gblocks	Castresana[54]	RRID:SCR_015945 http://molevol.cmima.csic.es/castresana/Gblocks_server.html
Kallisto version 0.44.0.	Bray et al.[55]	RRID:SCR_016582 https://pachterlab.github.io/kallisto/about
Deseq2	Love et al.[56]	RRID:SCR_015687 https://bioconductor.org/packages/release/bioc/html/DESeq2.html
FastQC	Andrews[57]	RRID:SCR_014583 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore version 0.6.5		RRID:SCR_011847 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
HISAT2 version 2.1.0		RRID:SCR_015530 http://ccb.jhu.edu/software/hisat2/index.shtml
featureCounts	Liao et al.[58]	RRID:SCR_012919 http://bioinf.wehi.edu.au/featureCounts/
PAML version 4.9i (including MCMCTree)	Yang[59]	RRID:SCR_014932 http://abacus.gene.ucl.ac.uk/software/paml.html
phytools R package	Revell[60]	RRID:SCR_015502 https://cran.r-project.org/web/packages/phytools/index.html
VHICA package	Wallau et al.[61]	https://github.com/cran/vhica
NOVOPlasty version 3.7	Dierckxsens et al.[62]	RRID:SCR_017335 https://github.com/ndierckx/NOVOPlasty
SAMtools version 1.5	Li et al.[63]	RRID:SCR_002105 http://htslib.org/

GeSeq version 2.03	Tillich et al.[64]	RRID:SCR_017336 https://chlorobox.mpimp-golm.mpg.de/geseq.html
ARAGORN version 1.2.38	Laslett and Canback[65]	RRID:SCR_015974 http://mbio-serv2.mbioekol.lu.se/ARAGORN/
ModelFinder	Kalyaanamoorthy et al.[66]	http://www.iqtree.org/ModelFinder/
UFBoot2	Hoang et al.[67]	
SynMap	Haug-Baltzell et al.[68]	https://genomevolution.org/SynMap.pl
DAGChainer	Haas et al.[69]	https://dagchainer.sourceforge.net/
CodeML	Yang et al.[59]	
nwalign		https://pypi.org/project/nwalign/
BEAST version 2.7	Bouckaert et al.[70]	RRID:SCR_010228 http://beast.bio.ed.ac.uk/
StarBEAST3 version 1.1.7	Douglas et al.[71]	https://github.com/rbouckaert/starbeast3
bModelTest	Bouckaert et al.[72]	
LogCombiner version 2.4.7	Bouckaert et al.[70]	
TreeAnnotator version 2.4.7	Bouckaert et al.[70]	
SplitsTree 4 version 4.14.6	Kloepper et al.[73]	RRID:SCR_014734 http://www.splitstree.org/
Hectar	Gschloessl et al.[74]	https://webtools.sb-roscoff.fr/root?tool_id=abims_hectar
RShiny		https://github.com/rstudio/shiny
IQ-TREE 2	Minh et al.[75]	https://github.com/iqtree/iqtree2
CAFE5	Mendes et al.[76]	https://github.com/hahnlab/CAFE5
Other		
BUSCO analysis version 5, eukaryota_odb10	Manni et al.[77]	RRID:SCR_015008 http://busco.ezlab.org/
UniRef90		RRID:SCR_010646 http://www.uniprot.org/help/uniref
AlphaFold protein structure database	Varadi et al.[78]	RRID:SCR_023662 https://alphafold.ebi.ac.uk/
NCVOG database	Yutin et al.[79]	
VOGDB database		https://vogdb.org/
SulfAtlas database	Barbeyron et al.[80], Stam et al.[81]11/12/24 4:16:00 PM	https://sulfatlas.sb-roscoff.fr/
Pfam	Mistry et al.[82]	RRID:SCR_004726 http://pfam.xfam.org/

Panther 17.0		RRID:SCR_004869 http://www.pantherdb.org/
Simple Modular Architecture Research Tool (SMART)	Letunic et al.[83]	RRID:SCR_005026 http://smart.embl.de/

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

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

Data and code availability

All sequence data, including DNA and RNA sequencing data, genome assemblies and annotations, have been deposited in the EBI/ENA database under the project accession PRJEB76691 and are publicly available. Additional data 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" and are publicly available. The DOI is listed in the key resources table. 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 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), 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 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/)."

This paper does not report original code.

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

METHOD DETAILS

Biological material

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 S1, which provides information about the sampling site for each strain. The sampling sites are shown on a world map in Figure S2.

1630

1631 All strains except those belonging to the Fucales were grown under laboratory conditions. The latter
1632 cannot be maintained long-term in the laboratory so field material was harvested for extractions. The
1633 haploid gametophyte generation was grown in culture for species with characterised haploid-diploid
1634 life cycles, with the exception of *Ectocarpus* strains, for which haploid partheno-sporophytes or diploid
1635 sporophytes were cultivated. All cultures were grown either in 140 mm diameter Petri dishes or in 2–
1636 10 L bottles, the latter aerated by bubbling with sterile air. Most cultures were grown in Provasoli-
1637 enriched[23] natural seawater (PES medium) under fluorescent white light (10–30 µM photons/m2·s) at
1638 13°C (or at 10°C for *Hapterophycus canaliculatus* and *Chordaria linearis* or 20°C for *Sphacelaria*
1639 *rigidula*, *Dictyota dichotoma, Schizocladia ischiensis* and *Chrysoparadoxa australica*). Exceptions
1640 included the freshwater species *Pleurocladia lacustris*, *Porterinema fluviatile* and *Heribaudiella*
1641 *fluviatilis*, which were grown in natural seawater that had been diluted to 5% with distilled water (i.e.,
1642 95% distilled water / 5% seawater) before addition of ES medium (http://sagdb.uni-
1643 goettingen.de/culture_medi a/01%20Basal%20Medium.pdf) micronutrients (at 20°C for *P. lacustris*)
1644 and *Phaeothamnion wetherbeei,* which was grown in MIEB12 (medium 7 in[84]). Whole thallus was
1645 extracted for all species except the Fucales, where either dissected meristematic regions or released
1646 male gametes were extracted. Tissue samples were frozen in liquid nitrogen and stored at -80°C before
1647 extraction.

1648

1649 **DNA extraction**

1650

1651 DNA was extracted using either the OmniPrep Genomic DNA Purification Kit (G Biosciences, St. Louis,
1652 MO, USA) or the Nucleospin Plant II midi DNA Extraction Kit (Macherey-Nagel, Düren, Germany). DNA
1653 quality was assessed using a Qubit fluorometer (Themo Fisher Scientific, Waltham, MA, USA), and
1654 fragment length was assessed by migration on a 1% agarose gel for some of the samples.

1655

1656 **Illumina library preparation and sequencing**

1657

1658 Libraries were prepared using the NEBNext DNA Modules Products (New England Biolabs, Ipswich, MA,
1659 USA) with an 'on bead' protocol developed by Genoscope, starting with 100 ng of genomic DNA. DNA
1660 was sonicated to a 100–800 bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA),
1661 end-repaired and 3'-adenylated. Illumina adapters (Bioo Scientific, Austin, TX, USA) were then added
1662 using the NEBNext Sample Reagent Set (New England Biolabs, Ipswich, MA, USA) and the DNA purified
1663 using Ampure XP (Beckmann Coulter Genomics, Danvers, MA, USA). Adapted fragments were
1664 amplified with 12 cycles of PCR using the Kapa Hifi Hotstart NGS library Amplification kit (Roche, Basel,

1665 Switzerland), followed by 0.8x AMPure XP (Beckman Coulter Genomics, Danvers, MA, USA)
1666 purification. Libraries were sequenced with Illumina MiSeq, HiSeq 4000 or NovaSeq 6000 instruments
1667 (Illumina, San Diego, CA, USA) in paired-end mode, 150 base read-length.
1668
1669 **Oxford Nanopore library preparation and sequencing**
1670
1671 Some samples were first purified using the Short Read Eliminator Kit (Pacific Biosciences, Menlo Park,
1672 CA, USA). All libraries were prepared using the protocol "1D Genomic DNA by Ligation" provided by
1673 Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK). Most of the libraries
1674 were prepared with the SQK-LSK109 kit (Oxford Nanopore Technologies), a few with the SQK-LSK108
1675 or SQK-LSK110 kits (Oxford Nanopore Technologies). Three flow cells were loaded with barcoded
1676 samples. The samples were mainly sequenced on R9.4.1 MinION or PromethION flow cells.
1677
1678 **RNA extraction, Illumina RNA-seq library preparation and sequencing**
1679
1680 RNA was extracted using either the Qiagen RNeasy kit or the Macherey Nagel RNAplus kit (Macherey-
1681 Nagel, Düren, Germany). RNA-seq libraries were prepared using the TruSeq Stranded mRNA Sample
1682 Prep (Illumina) according to the manufacturer's protocol, starting with 500 ng to 1 µg of total RNA, or
1683 using the NEBNext Ultra II Directional RNA Library Prep for Illumina (New England BioLabs) according
1684 to the manufacturer's protocol, starting with 100 ng of total RNA. The libraries were sequenced with
1685 Illumina HiSeq 2500, HiSeq 4000 or NovaSeq 6000 instruments (Illumina, San Diego, CA, USA), in
1686 paired-end mode, 150 base read-length.
1687
1688 **Assembly strategies**
1689
1690 Two assembly strategies were employed (Figure S15): one was designed for genomes exclusively
1691 sequenced using short reads with Illumina technology, while the other was designed for genomes that
1692 underwent sequencing using a combination of long and short reads, using respectively the Nanopore
1693 and Illumina technologies.
1694
1695 **Short-read-based genome assembly.** When sequencing was performed exclusively using short reads,
1696 reads corresponding to bacterial contaminants were filtered out early in the assembly process (Figure
1697 S15A) because, typically, the initial datasets were too large to run assemblers like SPAdes. To remove
1698 bacterial contaminants, an assembly based on the initial illumina dataset was first generated for each
1699 strain using a fast and non-greedy algorithm, MEGAHIT[1] version 1.1.1 with the parameters --k-min 101

1700 --k-max 131 --k-step 10. Assigning taxonomy is easier when working with contigs than with reads.
1701 Contigs exceeding 500 bp in each preliminary assembly underwent taxonomic classification based on
1702 gene models predicted using the *ab initio* software MetaGene[2] version 2008.8.19 with default
1703 parameters and then aligning proteins against UniprotKB using BLASTp (e-value <$10e^{-4}$). A
1704 superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best
1705 alignment (selected using the BLASTp score). Contigs that contained more than 50% of their genes
1706 assigned to Bacteria and with at least one gene every 10 Kbp were classified as bacterial sequences.
1707 For each strain, the initial Illumina sequencing reads were aligned against the corresponding bacterial
1708 sequences using latest version of the Burrows-Wheeler Aligner[4] (BWA) with default parameters and
1709 mapped short-reads were labelled as contaminants, and assembled for the purpose of obtaining more
1710 contiguous contigs. These bacterial contigs were then used to build a contaminant sequence database.
1711 Finally, the clean subset of reads was obtained by aligning the whole Illumina dataset against this
1712 strain-specific bacterial contig database, using Bowtie2[5] version 2.2.9 with default parameters. A final
1713 assembly was then generated for each strain using the contaminant-free read datasets and the
1714 SPAdes[6] assembler version 3.8.1 with the parameters -k 21,57,71,99,127 -m 2000 --only-assembler --
1715 careful. Genome assemblies based only on short-reads were more fragmented (N50 ranged from 3
1716 Kbp to 31 Kbp) than assemblies that used long reads but the sizes of the former were consistent with
1717 expectations.
1718
1719 **Long-read-based genome assemblies.** A subset of the strains produced DNA of both adequate quality
1720 and quantity, enabling successful long-read sequencing. In these cases, long reads were assembled
1721 directly and the detection of possible bacterial contigs was carried out after the assembly step (Figure
1722 S15B). To produce long-read-based genome assemblies we generated three samples of reads i) all
1723 reads, ii) 30X coverage of the longest reads and iii) 30X coverage of the filtlong
1724 (https://github.com/rrwick/Filtlong) highest-score reads. The three samples were used as input data
1725 for four different assemblers, Smartdenovo[7], Redbean[8], Flye[9] and Necat[10]. Based on the cumulative
1726 size and contiguity, we selected the best assembly for each strain. This assembly was then polished
1727 three times using Racon[11] with Nanopore reads, and twice with Hapo-G[12] and Illumina PCR-free reads.
1728
1729 **Assembly decontamination**
1730
1731 Contigs from the short- and long-read genome assemblies were inspected for potential bacterial
1732 sequences. This process was carried out using a combination of several analysis and tools: GC
1733 composition, read coverage, Metabat 2 (for tetramer composition and clustering)[13] and Metagene (for

1734 gene prediction and taxonomic identification, as described previously). Contigs were manually
1735 removed based on their characteristics.
1736
1737 **Transcriptome assembly**
1738
1739 Ribosomal-RNA-like reads were detected using SortMeRNA[14] and filtered out. The Illumina RNA-Seq
1740 short reads from each strain were assembled using Velvet[15] version 1.2.07 and Oases[16] version 0.2.08
1741 with kmer sizes of 61, 63 and 65 bp. BUSCO[77] analysis (v5, eukaryota_odb10) was then performed on
1742 the three resulting assemblies for each strain in order to select the best assembly, i.e. the most
1743 complete at the gene level. Reads were mapped back to the contigs with BWA-mem, and only
1744 consistent paired-end reads were retained. Uncovered regions were detected and used to identify
1745 chimeric contigs. In addition, open reading frames (ORF) and domains were identified using
1746 TransDecoder (Haas, B.J., https://github.com/TransDecoder/TransDecoder) and CDDsearch[17],
1747 respectively. Contigs were broken into uncovered regions outside ORFs and domains. In addition, read
1748 strand information was used to correctly orient RNA-Seq contigs.
1749
1750 ***De novo* transcriptomes**
1751
1752 The RNA-seq data was also used to generate *de novo* transcriptomes. For each strain, all the RNA-seq
1753 data available was cleaned to remove poor quality sequence and adapter sequences using
1754 Trimmomatic[18] v0.39 prior to being assembled using either Trinity[19] version v2.6.5 or rnaSPAdes[20]
1755 version v3.13.1. The strandness and Kmer-length parameters of the assemblers were adjusted to take
1756 into account RNA-seq read characteristics. The *de novo* transcriptomes represented an alternative
1757 source to identify and characterise genes if they were not detected in the genome assemblies. The *de*
1758 *novo* transcriptomes are available from the CNRS Research Data dataset
1759 (https://doi.org/10.57745/9U1J85) and from the Phaeoexplorer website (https://phaeoexplorer.sb-
1760 roscoff.fr/).
1761
1762 **Detection and masking of repeated sequences and transposons**
1763
1764 Prior to gene annotation, each genome assembly was masked based on the repeat library from
1765 *Ectocarpus* species 7 (formerly *Ectocarpus siliculosus*)[85] and using RepBase with RepeatMasker[21]
1766 version v.4.1.0, default parameters. Tandem repeats finder (TRF)[22] was also used to mask tandem
1767 repeat duplications. In addition, transposons were annotated in ten species using REPET[23] and the

1768 transposons detected were used as a reference to mask all genomes with RepeatMasker[21] version
1769 v4.1.0, default parameters.
1770
1771 **Gene prediction**
1772
1773 For each strain, gene prediction was performed using both homologous proteins and RNA-seq data.
1774 Proteins from *Ectocarpus* species 7 (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2)[86]
1775 and UniRef90 (https://www.uniprot.org/uniref/) were aligned against each genome assembly. First,
1776 BLAT[24] with default parameters was used to quickly localise putative genes corresponding to the
1777 *Ectocarpus* species 7 proteins. The best match and matches with a score ≥ 90% of the best match score
1778 were retained. Second, the alignments were refined using Genewise[25] with default parameters, which
1779 is more precise for intron/exon boundary detection. Alignments were retained if more than 80% of
1780 the length of the protein was aligned to the genome. To detect conserved proteins and allow detection
1781 of horizontal gene transfer, UniRef90 proteins (without *E. siliculosus* sequences) were aligned with
1782 DIAMOND[26] (v0.9.30 with parameters --evalue 0.001 --more-sensitive) to genomic regions lacking
1783 alignments with an *Ectocarpus* species 7 protein. Only the five best matches per locus were retained,
1784 based on their bitscore. Selected proteins from UniRef90 were aligned to the whole genome using
1785 Genewise as described previously, and alignments with at least 50% of the aligned protein length were
1786 retained. The assembled transcriptome for each strain was aligned to the strain's genome assembly
1787 using BLAT[24] with default parameters. For each transcript, the best match was selected based on the
1788 alignment score, with an identity greater or equal to 90%. Selected alignments were refined using
1789 Est2Genome[27] in order to precisely detect intron boundaries. Alignments were retained if more than
1790 80% of the length of the transcript was aligned to the genome with a minimal identity of 95%. Finally,
1791 the protein homologies and transcript mapping were integrated using a combiner called Gmove[28]. This
1792 tool can find coding sequences (CDSs) based on genome-located evidence without any calibration step.
1793 Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified
1794 graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for
1795 open reading frames (ORFs) consistent with the protein evidence. Translated proteins of predicted
1796 genes were then aligned against NR prot (release 19/02/2019) and the *Ectocarpus* species 7 version v2
1797 proteome[86] (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2) using DIAMOND BLASTp
1798 with parameters --evalue 10-5 --more-sensitive --unal 0. All predicted genes with significant matches
1799 (the smallest protein had to be aligned for at least 50% of its length) were retained. In addition to these
1800 genes, we also retained genes with CDS size greater than 300 bp and with a coding ratio (CDS size /
1801 mRNA size) greater or equal to 0.5.
1802

1803 **Annotation decontamination**

1804

1805 After predicting the genes, an additional analysis was carried out to detect bacterial sequences. If a
1806 contig did not contain any genes, it was analysed with MetaGene and the predicted proteins added to
1807 the gene catalogue for the purpose of detecting bacterial sequences. Proteins generated from
1808 predicted genes (Gmove plus MetaGene) were then aligned against UniprotKB using BLASTp (e-value
1809 $< 10e^{-4}$) and superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the
1810 best alignment (selected using the BLASTp score). Contigs that contained more than 80% of their genes
1811 assigned to bacteria, Archaea or viruses were classified as bacterial sequences and removed from the
1812 final assembly file. Genes belonging to these contigs were also removed from the final gene catalogue.
1813 Finally, completeness of each predicted gene catalogue was assessed using BUSCO[77] (v5.0.0;
1814 eukaryota_odb10).

1815

1816 In addition, the quality of the annotations was assessed by comparing the length of coding regions in
1817 pairs of orthologous proteins (best reciprocal hits) between each genome and *Ectocarpus* species 7,
1818 which was used as a reference because its high-quality annotation has been extensively curated[86]. The
1819 correlation between orthologous CDS lengths was higher for genomes sequenced with long reads than
1820 for genomes only sequenced with short reads (Figure S1B). This difference was probably principally
1821 due to a higher proportion of underestimated protein lengths in the latter (Table S1B) which likely
1822 corresponded to fragmented genes. The qualities of Ectocarpales genome annotations were very high
1823 (BUSCO and length of predicted CDS) even when the genomes were sequenced using only short reads,
1824 probably because their phylogenetic proximity to *Ectocarpus* species 7 facilitated the building of good
1825 quality gene models.

1826

1827 **Analyses aimed at deducing functional characteristics of predicted proteins**

1828

1829 Several different analyses of the predicted proteomes of each species were carried out to provide
1830 information about the cellular functions of the encoded proteins. These included eggNOG-mapper[42]
1831 analyses (v2.1.8 or v2.0.1, with emapperDB v5.0.2 or v4.5.1) to provide multiple functional annotations
1832 (Gene Ontology, Kyoto Encyclopedia of genes and genomes, Clusters of Orthologous Genes, Pfam),
1833 Interproscan[52] analyses (versions v5.55-88.0, v5.51-85.0 or v5.36-75.0) to detect functional domains,
1834 Hectar[74] (v1.3) predictions of protein subcellular localisation and various DIAMOND[26] (v2.0.15 vs
1835 UniRef90 2022_03, with parameter "evalue" set to $10e^{-5}$) sequence similarity searches aimed at
1836 identifying homologous proteins with functional annotations.

1837

1838 **Detection of tandemly duplicated genes**

1839

1840 Starting with the protein alignments that had been constructed to build the orthogroups, matches
1841 between proteins within the same genome with an e-value of ≤10−20 and which covered at least 80%
1842 of the smallest protein were extracted. Two genes were considered to be tandemly duplicated if they
1843 were localised on the same genomic contig separated by five or less intervening genes, regardless of
1844 their orientation. The tandemly-duplicated genes were clustered using a single linkage clustering
1845 approach. A contingency test was applied to compare the proportion of tandemly-duplicated genes in
1846 each orthogroup with the global proportion of tandemly-duplicated genes (p=0.0532792). The *p*-
1847 values are shown in table S1.

1848

1849 **Relative orientation of adjacent genes and lengths of intergenic regions**

1850

1851 For each species, the proportion of pairs of adjacent genes localized on opposite strands was compared
1852 to the expected proportion of 0.5 using a binomial test (with p=0.5). The *p*-values are shown in Table
1853 S1 (*p*-values of <0.05 correspond to cases where the proportion is significantly higher than 0.5).

1854

1855 The lengths of intergenic regions between pairs of adjacent genes located on opposite strands (i.e.
1856 divergently or convergently transcribed) were compared with the lengths of intergenic regions
1857 between genes located on the same strand (i.e. transcribed in the same direction). Contingency tables
1858 were constructed for each species using a threshold of 1000 bp for the intergenic length and the
1859 number of intergenic regions in each of four categories were counted: 1) same strand genes, intergenic
1860 <1000 bp, 2) opposite strand genes, intergenic <1000 bp, 3) same strand genes, intergenic ≥1000 bp,
1861 4) opposite strand genes, intergenic ≥1000 bp. Fisher exact tests were applied to the contingency
1862 tables (alternative hypothesis: true odds ratio is greater than 1). The *p*-values are shown in Table S1.
1863 When *p*-values are <0.05, short intergenic lengths are significantly associated with pairs of genes on
1864 opposite strands. All calculations were performed with R[87] (version 4.3.0).

1865

1866 **Detection of long non-coding RNAs**

1867

1868 Transcriptome data for 11 species (Table S21), including nine brown algal strains and two outgroup
1869 taxa, was analysed to identify lncRNAs. Any transcripts with invalid nucleotide DNA symbols were
1870 discarded and sequences shorter than 200 nucleotides were removed to avoid the detection of small
1871 RNA transcripts. The transcriptome sequences in Fasta format were analysed with votingLNC
1872 (https://gitlab.com/a.debit/votinglnc) to detect lncRNA transcripts and assign a confidence level for

1873 each transcript. A similar approach was used to detect lncRNAs in the lncPlankton database[88].
1874 VotingLNC is a meta-classifier combining the predictions of the ten most commonly used coding
1875 potential tools. Based on a majority voting ensemble procedure, the meta-tool assigns the final coding
1876 potential class to a transcript as the class label predicted most frequently by the ten classification
1877 models included in the ensemble. Alongside the majority voting class, a reliability score was calculated
1878 for each transcript. A cut-off non-coding reliability score of $p > 0.5$ was chosen to treat a transcript as
1879 lncRNA and to decrease false-positive identification. The set of transcripts predicted as lncRNA by the
1880 majority-voting procedure and having an ORF(s) encoding peptide(s) with length ≥ 100aa were
1881 discarded. lncRNA transcripts that had significant matches in either the Pfam[82] (hmmscan e-value <
1882 0.001) or SwissProt (BLASTp e-value < $1e^{-5}$ and similarity ≥ 90%) databases were removed from the
1883 dataset. Transcript length, GC content, and the length of the longest ORF were compared between
1884 lncRNAs and protein-coding RNAs. The comparison was carried out using a Wilcoxon test. R version
1885 V.4.1.2 was used for all the analyses and ggplot2 (version 3.4.0) for plotting.
1886
1887 **Intron conservation**
1888
1889 Intron positions were compared in a set of single copy genes that are conserved across all the
1890 Phaeophyceae and the outgroup species. The analysis focused on the 21 reference genomes (Table
1891 S21) and on orthogroups that occurred exactly once in at least 20 of the 21 genomes, allowing the
1892 gene to be absent from only one of the 21 genomes. In addition, orthogroups were discarded if more
1893 than three copies had been annotated in the other Phaeophyceae genomes. These filters produced a
1894 set of 235 conserved (ancestral) orthogroups. Multiple alignments were carried out for each
1895 orthogroup using MUSCLE[34] version 3.8.1551 with default parameters and conserved blocks were
1896 identified with Gblocks[54] version 0.91b with the parameters -p=t -s=n -b5=a -b2=[nsp] -b1=[nsp] -b3=6,
1897 where "nsp" is equal to 90% of the number of proteins aligned. A shell script was then used to compare
1898 intron positions in the alignments. For each intron in the multiple sequence alignment, we obtained a
1899 corresponding conservation profile listing which species contains an intron at that position. The
1900 profiles obtained for the 949 introns that are in conserved blocks of the multiple alignments are shown
1901 in Figure S6B. Both phase and length of ancestral introns (e.g. that were conserved in most
1902 Phaeophyceae and at least two sister clades) were compared to the phase and length of *Ectocarpus*
1903 species 7 introns as a reference. The same approach was used to compare intron positions across 11
1904 *Ectocarpus* species, with *Scytosiphon promiscuus* as an outgroup, by selecting 831 conserved
1905 monocopy orthogroups. The number of introns per gene in brown algae and in closely-related
1906 outgroup species were compared using a contingency test (table S1C).
1907

1908 **Phylogenomic tree of the Phaeophyceae**

1909

1910 To provide a phylogenetic framework for the analyses of the Phaeoexplorer genome dataset, the 41-
1911 species phylogenomic tree reported by Akita *et al.*[89] was updated by adding 15 additional species using
1912 the same methodology. Briefly, for the additional species, amino acid sequences were recovered for
1913 the 32 single-copy orthologous genes used to construct the published tree and these were aligned
1914 manually with the existing sequences using the alignment software AliView[29] v.1.26. The aligned
1915 sequences of the final 56 species were concatenated and maximum likelihood analysis was carried out
1916 with 10,000 rapid bootstraps using RAxML[30] v.8.2.9 and the gamma model. The best-fit evolutionary
1917 model for each gene was determined using AIC.

1918

1919 **Bayesian divergence time estimation for the brown algae**

1920

1921 An estimation of brown algal divergence time was carried out using the 32 orthologous nuclear genes
1922 (see above and[89]) for 51 brown algae and five non-brown species (16,185 amino acids, 56 spp.) and
1923 MCMCTree (PAML package v4.9j) with the approximate likelihood method. The WAG protein model
1924 was selected based on the AIC and BIC criteria of ModelFinder[66]. The independent clock model was
1925 selected based on previous work on the brown algal timeline by Choi *et al*.[90]. One hundred million
1926 years was set to correspond to 1 in the MCMCTree calculation. A secondary calibration for the root
1927 was based on Choi *et al*.[90] using a gamma distribution of 70.2 alpha and 10.22 beta. A kelp holdfast
1928 fossil[91] was used to date the crown node of kelps with a minimum bound of 0.31, and a *Julescraneia*
1929 fossil[92] for the *Macrocystis/Saccharina* clade with a minimum bound of 0.13 (Figure S3). MCMC chains
1930 were run 1.5 million generations, with the first 200,000 MCMC chains being discarded as burn-in, and
1931 the convergence of MCMC chains was checked with Tracer v1.7.2[31]. This analysis estimated that
1932 Schizocladiophyceae and brown algae diverged 457.88 Mya (95% HPD: 321.29-592.66 Ma), similar to
1933 (about 8 Mya older than) the previous estimate using plastid genes[90] and that diversification of the
1934 major brown algal linages began about 220 million years later, after the origin of DFI clade (235.97
1935 Mya, 95% HPD: 158.88-312.48 Mya), about 12 Ma earlier than the previous estimate[90]. The fossil-
1936 calibrated phylogenetic tree for 11 *Ectocarpus* species (Figure S13A) was extracted from the brown
1937 algal tree (Figure S3).

1938

1939 **Detection of orthologous groups**

1940

1941 Predicted proteins from the 60 strains sequenced in Phaeoexplorer complemented with 16 public
1942 proteomes covering the Ochrophytina subphylum and the terrestrial oomycetes were clustered using

1943 OrthoFinder[32] v2.5.2 with default parameters. This generated 56,340 orthogroups that contained
1944 90.1% of the proteins (1,415,341 of the 1,571,648). Seventy-one of the 76 strains had more than 75%
1945 of their proteins in an orthogroup shared with at least one other strain. The orthogroups contain
1946 between 2 and 6,220 proteins with a mean of 25.1 proteins and a median of three.
1947
1948 **Dollo analysis of orthogroup gain and loss**
1949
1950 An analysis of evolutionary events of gene family gain and loss was carried out on a selection of strains
1951 covering the brown algal phylogeny and sister groups as distant as the Raphidophyceae under the
1952 Dollo parsimony law using orthogroups as proxies for gene families. To limit possible problems due to
1953 the fragmentation of predicted proteins in some assemblies, we selected 24,410 orthogroups present
1954 in at least one of 17 strains that had both good quality genome assembly and good quality gene
1955 predictions. Dollo parsimony analysis was then run using Count[33] version v9.1106 based on a
1956 cladogram of a subset of 24 species representative of the Phaeoexplorer project and excluding all
1957 public outgroups more distant than *Heterosigma akashiwo*. The cladogram was based on the topology
1958 of the brown algae phylogenetic tree published by Akita *et al*.[89].
1959
1960 **Phylostratigraphy analysis**
1961
1962 GenEra[37] was used to estimate gene family founder events for each genome assembly by running
1963 DIAMOND[26] in ultra-sensitive mode against the Phaeoexplorer protein dataset and the NCBI non-
1964 redundant database. All sequence matches with e-values < 10^{-5} were treated as being homologous
1965 with the query genes in the target genomes. The NCBI taxonomy was used as an initial template to
1966 infer the evolutionary relationships of each query gene with their matches in the sequence database
1967 but taxonomic assignments within the PX clade and Phaeophyceae were then modified to reflect the
1968 evolutionary relationships that were inferred in the maximum likelihood tree. Gene families were
1969 predicted based on a clustering analysis of the query proteins against themselves using an e-value
1970 cutoff of 10^{-5} in DIAMOND and an inflation parameter of 1.5 with MCL[38]. Estimated evolutionary
1971 distances were extracted for each pair of species from the maximum likelihood species tree
1972 (substitutions/site) to calculate homology detection failure probabilities[93]. Taxonomic sampling of the
1973 species tree enabled homology detection failure tests to be carried out within the PX clade. Gene
1974 families whose ages could not be explained by homology detection failure were analysed by inspecting
1975 the functional and domain annotations for *Ectocarpus* species 7[86]. Structural alignments were
1976 performed using Foldseek[39] against the AlphaFold protein structure database[78].
1977

1978 **Detection of gene family amplifications**

1979

1980 A binomial test with a parameter of 17/21 was carried out to detect gene families (OGs) that had
1981 significantly expanded in 17 Phaeophyceae reference genomes compared with four closely-related
1982 outgroup species (*Schizocladia ischiensis*, *Tribonema minus*, *Chrysoparadoxa australica* and
1983 *Heterosigma akashiwo*; Table S21). Expanded gene families deviated significantly from the expected
1984 proportion (17/21 under the null hypothesis where there are equal gene numbers in all species).
1985 Benjamini–Hochberg FDR correction for multiple testing was then applied and 233 candidate OGs with
1986 corrected *p*-values of < 0.001 were retained. All calculations were performed with R (version 4.1.0).
1987 The set of 233 candidate OGs was then filtered to limit counting errors due to annotation artefacts
1988 (e.g. genes missed or fragmented) using the following procedure:
1989 1) A protein consensus was first deduced for each orthogroup. Protein sequences representative of all
1990 lineages were extracted and aligned using MUSCLE[34] version 3.8.1551 with default parameters and the
1991 multiple alignments were filtered using OD-Seq[35] version 1.0 to remove outlier sequences, with
1992 parameter –score set to 1.5. The consensus sequences were then extracted from the multiple
1993 alignments of non-outlier sequences using hmmemit in the HMMER3[36] package version 3.1b1 with
1994 default parameters.
1995 2) In order to estimate gene family copy number independently of the assembly and annotation
1996 processes, short read sequences for each genome were mapped onto the orthogroup consensus
1997 sequences using DIAMOND[26]. Unique matches were retained for each read and depth of coverage was
1998 calculated for each consensus orthogroup. The depth obtained for each orthogroup was normalised
1999 for each species by dividing by the depth obtained on a set of conserved single-copy genes, so that the
2000 final value obtained was representative of the gene copy number. Then, for each candidate amplified
2001 orthogroup, the average depth for the 17 Phaeophyceae species and the average depth for the four
2002 outgroup species was calculated and OGs where the depth for outgroups was more than half the depth
2003 for the Phaeophyceae were discarded. We retained 227 out of 233 orthogroups after this step.
2004 3) Finally, functional annotations were used to remove orthogroups that were likely to correspond to
2005 transposable elements. A final list of 180 OGs was retained (Table S3).

2006

2007 The amplified gene families were manually categorised into functional classes based on the output of
2008 automatic functional annotation programs (Interproscan[52], EggNOG[41], nr BLASTp) and an amplification
2009 profile was assigned to each orthogroup by identifying the taxonomic group where the amplification
2010 of the family was most marked (Table S3).

2011

2012 In addition to the binomial tests, we also ran CAFE5[76] to reconstruct the history of gene family
2013 amplifications. Such reconstructions rely on a species tree and require that all gene families are present
2014 at the root of the tree. However, of the 180 amplified OGs that were strongly amplified in
2015 Phaeophyceae (see above and listed in Table S3) only 19 were present at the ancestral node. The
2016 majority (161) of the 180 families were gained during the early evolution of the lineage, most (105) at
2017 the origin of the PX clade (i.e. a collapsed node corresponding to nodes n1 and n2 in Figure S4A) or of
2018 the Phaeophyceae/FDI clades (i.e. a collapsed node corresponding to nodes n5 and n6; Figure S4A). To
2019 determine whether the 180 amplified OGs were significantly enriched in genes that were gained early
2020 during Phaeophyceae evolution (i.e. at nodes n1/n2, n4, n5/n6 in Figure S4A), a Chi-squared test was
2021 carried out using the R *chisq.test* function on a contingency table containing the proportions of OGs
2022 gained at various periods during brown algal evolution for both the amplified OGs and for the entire
2023 set of OGs as a reference dataset (Table S1C). Twelve independent CAFE5 reconstructions were carried
2024 out on the OG subsets gained at 12 different nodes (n0, n1/2, n4, n5/n6, n8, n9, n10/n11, n13, n15,
2025 n18, n19, n20), using the subtrees rooted at these nodes so that the sets of OGs gained at each node
2026 would be placed at the root of the tree for one of the 12 analyses (Figure S5E). The analysis focused
2027 on the 19 highest quality genomes (Table S21), which is why some pairs of nodes were collapsed (e.g.
2028 nodes n1 and n2 to give n1/n2). Several parameters were tested for CAFE5: the –p option (Poisson
2029 distribution) resulted in better likelihood scores than default, but we observed a weak effect when
2030 increasing the value of lambda (–k). Consequently, all reconstructions were performed with –p (and
2031 no k, i.e. k=1) for efficiency purposes. As recommended by Mendes *et al*.[76], very large gene families
2032 were discarded as these can cause the program to fail to initialize the parameters. The twelve
2033 reconstructions were then aggregated and the proportions of amplified and reduced gene families
2034 were calculated for each node (Table S6). Only results on internal nodes were considered, since leaves
2035 are more subject to artefactual amplifications/reductions due to genes being missed, fused or split in
2036 the annotations.
2037
2038 **Composite genes**
2039
2040 The amino-acid sequences of all 530,598 genes present in the selected genomes were compared in an
2041 all-against-all pairwise alignment using DIAMOND BLASTp[26] version 2.0.11; "very-sensitive" mode; e-
2042 value threshold $1e^{-5}$. This raw alignment was then filtered using CleanBlastp, from the
2043 CompositeSearch suite[40], to remove sequence alignments with under 30% residue identity and
2044 produce the final sequence similarity network. CompositeSearch was then used on this network to
2045 identify putative composite gene families among the orthologous groups (OGs) previously computed
2046 by OrthoFinder[32]. Composite OGs containing two or more genes and having non-overlapping regions

2047 aligned to their component OGs were retained for further analysis, while singleton composite OGs and
2048 composites with overlapping component regions were discarded. A phylogeny-based approach[94],
2049 which uses information from extant genomes to apply a Dollo parsimony model in Count[33], was used
2050 to reconstruct the evolutionary events (domain fusions and fissions) that led to structural
2051 rearrangements of composite genes, allowing them to be labelled as fusion or fission events (or as
2052 complex events when sequentiality could not be clearly deduced).
2053
2054 **Horizontal gene transfer (HGT)**
2055
2056 **Dataset and experimental approach.** Uneven data collection across taxa can impact HGT
2057 identification. The phylogeny-based HGT screening approach used here requires the establishment of
2058 a comprehensive and taxonomically diverse reference dataset. The analysis focused on the
2059 Phaeoexplorer genomes using a background database called REFAL and an automated bioinformatics
2060 tool called RoutineTree, which screens for HGTs using phylogenetics. The background database was
2061 built using a starting database, GNM1157, which includes a diverse set of 17,250,679 protein
2062 sequences from 1157 genomes spanning various prokaryotic and eukaryotic lineages (540 bacteria, 45
2063 archaea, 431 Opisthokonta, 15 Rhodophyta, 83 Viridiplantae, and 43 genomes from CRASH lineages).
2064 Data from NCBI RefSeq (updated as of May 2020) and MMETSP were integrated into GNM1157 to form
2065 the background database REFAL. To enhance data quality and reduce redundancy, CD-HIT version 4.5.4
2066 was used to remove highly similar sequences (with sequence identity ≥90%) within each taxonomic
2067 order. This curation process resulted in a protein database consisting of 39.9 million sequences,
2068 representing over 7,786 taxa and providing comprehensive coverage across the diverse branches of
2069 the tree of life. To obtain the best assembled genome within a genus, the latest version was selected
2070 if multiple versions were available. In addition, the dataset was expanded by searching for genomes in
2071 other repositories such as the Joint Genome Institute. Special attention was paid to achieving balanced
2072 representation of the Rhodophyta and Viridiplantae, which are particularly crucial for HGT analysis
2073 within the Chromalveolate group. To accomplish this, protein data from six red algal transcriptomes
2074 sourced from MMETSP was added. The HGT search was applied to 72 Stramenopile genomes, including
2075 45 newly sequenced and 27 public genomes.
2076
2077 **Phylogenetic Tree Reconstruction.** The pipeline for constructing phylogenetic trees splits fasta files
2078 into individual sequence files and then carries out a search for homologous sequences, followed by
2079 multiple sequence alignment and tree-building. Nested positions within the trees were identified using
2080 artificial intelligence and hU and hBL methods were used for HGT verification. Instead of using all
2081 available sequences, sequences with the best BLAST hit scores from each kingdom, phylum, and class

2082 were used for tree construction to expedite tree-building and enhance clarity. Each gene, regardless
2083 of whether it was a copy or not, was used as a query for tree construction. To improve precision, four
2084 different methods were used for tree building: neighbour-joining, maximum parsimony, maximum
2085 likelihood and Bayesian. As a result, each node within a tree was associated with four support values.
2086 To create single-gene phylogenetic trees, a BLASTp[3] search was carried out against the background
2087 database, employing an e-value cutoff of $1e^{-05}$. For each query, the top 1,000 significant matches were
2088 sorted by bit-score in descending order as the default criterion. Matching sequences were then
2089 retrieved from the database, with a constraint of no more than three sequences per genus and no
2090 more than 12 sequences per phylum. To further refine the selection, significant matches with a query-
2091 subject alignment length of at least 120 amino acids were re-sorted based on query-subject identity in
2092 descending order. A second set of homologous sequences was then retrieved from the database
2093 following the same procedure. These two sets of homologous sequences, along with the query, were
2094 merged and aligned using MUSCLE[34] version 3.8.31 with default settings. The resulting alignments,
2095 trimmed to a minimum length of 50 amino acids using TrimAl[49] version 1.2 in automated mode (-
2096 automated1), were used to construct phylogenetic trees with FastTree version 2.1.7, with the
2097 'WAG + CAT' model and four rounds of minimum-evolution SPR moves (-spr 4) along with exhaustive
2098 ML nearest-neighbour interchanges (-mlacc 2 -slownni). Branch supports were estimated using the
2099 Shimodaira-Hasegawa (SH)-test.
2100
2101 **Inferring HGT based on tree topology.** Phylogenetic trees were examined to identify specific
2102 topologies where Phaeoexplorer query sequences were nested among other sequences, defined as a
2103 situation where two or more monophyletic clades consist of both queries and prokaryotic sequences,
2104 supported by distinct nodes within the tree. These monophyletic clades are considered to group
2105 together if they share the same set of prokaryotic sequences but differ in sequences from optional
2106 taxa. Singletons for both the donor and receptor genes were excluded to minimise contamination and
2107 recent HGT interference. To retain only robustly supported nested positions, positions were required
2108 to be multiply supported, with a minimum of ≥0.70 for the SH-test and aByes-test support from at least
2109 two Phaeoexplorer receptor nodes and three donor supporting nodes. Furthermore, queries that
2110 displayed significantly different amino acid compositions ($P < 0.05$) compared to the remaining
2111 sequences in the alignment were discarded. Queries from the CRASH category that nested among
2112 sequences from other kingdoms (supported by >70% UFBoot at one or more supporting nodes) were
2113 retained.
2114
2115 **Enhancing accuracy and establishing the timing of HGTs.** To enhance accuracy, a minimum
2116 requirement was imposed for all supporting nodes and for strongly supported nodes that indicate

2117 query-donor monophyly. To determine the timing of HGT events, temporal information, primarily
2118 derived from the timetree database, was incorporated into each node. We assigned the "smallest
2119 boundary" role to pinpoint the most recent common ancestor at the time of the HGT event. Essentially,
2120 if all descendants of a given query protein sequence can be traced back to the initial HGT event, a
2121 common ancestral node can be identified whose occurrence time can be inferred using a molecular
2122 clock approach based on archaeological and fossil evidence. The taxonomy boundaries of HGT
2123 descendants were determined by identifying the smallest ancestor shared by both the donor and
2124 receptor taxa from the monophyletic clades within the tree. By considering the emergence times of
2125 both taxa, the timing of the transfer of genes from earlier taxa to later taxa can be determined, as the
2126 reverse scenario is not considered plausible.
2127
2128 **Verification of HGTs.** Verification of HGT used the following contamination assessment criteria: i) HGT
2129 candidates were excluded if they were located in a contig where 50% of the genes had better matches
2130 with other kingdoms, ii) HGT candidates were excluded if they were located in a contig where 50% of
2131 the genes were primarily identified as HGT genes, iii) HGT candidates were excluded if one of their five
2132 closest flanking genes, both upstream and downstream, had a better match with other kingdoms. AI,
2133 hU and the hBL value were used to further validate HGT events. This process was supplemented with
2134 annotation and functional predictions for the identified HGTs.
2135 Further validation was based on the following concepts:
2136 *OUTGROUP*: This comprises all biological donors present in a tree, excluding the query species if it
2137 belongs to biological donors.
2138 *SKIP*: This includes all biological receptors (species belonging to optional taxa) in a tree, again excluding
2139 the query species if it belongs to biological receptors.
2140 *INGROUP*: This encompasses species from SKIP's upper level, excluding SKIP itself and the query
2141 species (if it belongs to biological receptors).
2142 *AI (Alien Index)*: computed for each query gene using e-values from BLAST hits:
2143 *AI = (E-value of best BLAST hit in the INGROUP lineage) / (E-value of best BLAST hit in the OUTGROUP*
2144 *lineage)*
2145 The AI score quantifies how similar queries are to their homologs in the OUTGROUP compared to
2146 homologs in the INGROUP. We apply a relatively lenient cut-off (AI > 0) for initial screening, which can
2147 be adjusted in the second screening as needed.
2148 *hU (HGT Score Support Index)*: calculated for each query gene based on the best bit scores of INGROUP
2149 vs. OUTGROUP:
2150 *hU = (Best-hit bitscore of OUTGROUP) - (Best-hit bitscore of INGROUP)*

2151 A lenient cut-off (hU > 0) is used for initial screening, with flexibility for adjustment in the second
2152 screening.
2153 *hBL (HGT Branch Length Support Index)*: calculated based on the minimum branch length to the query
2154 within INGROUP vs. OUTGROUP:
2155 *hBL = (Minimum branch length to the query within INGROUP) - (Minimum branch length to the query*
2156 *within OUTGROUP)*
2157 A lenient cut-off (hBL > 0) is applied initially, with the option for modification in the second screening.
2158 *CHE, CHS, CHBL (Consensus Hit Support)*: To mitigate the possibility that the best bit score for either
2159 INGROUP or OUTGROUP is influenced by contamination, we consider alternative matches. We
2160 introduce consensus hit support (CHE, CHS, and CHBL) to assess the reliability of AI, hU, and hBL,
2161 respectively.
2162 For example, if AI > 0, CHE evaluates the likelihood that "AI remains greater than 0" when using the e-
2163 value of each sequence in OUTGROUP instead of bbhO. A similar approach applies to CHS for hU and
2164 CHBL for hBL. This additional layer of evaluation helps ensure the robustness of the HGT verification
2165 process.
2166
2167 **Gene codon usage, functional annotation and expression.** Indices of codon usage and GC content
2168 were calculated using Codonw 1.4.4 (http://codonw.sourceforge.net). Gene functions were assigned
2169 by searching against the Gene Ontology (GO) database using blast2GO (ref blast2GO 08) and the KEGG
2170 database using blastKOALA (http://www.kegg.jp/blastkoala/) with default parameters. The full gene
2171 sets of each species were set as the background for KEGG and GO enrichment analyses by applying
2172 Student's t-test (*p*-value cutoff = 0.01). HGTs were also analysed with SEED
2173 (http://www.theseed.org/wiki/Home_of_the_SEED), IPR2GO
2174 (http://www.ebi.ac.uk/interpro/search/sequence-search), eggNOG[41]
2175 (http://eggnogdb.embl.de/#/app/home) and Pfam[82]. For each species, the differences between mean
2176 gene expression levels for HGTs and non-HGT genes with common GO terms were accessed using
2177 Student's t-test. Go terms with less than five genes in either gene category were ignored. The
2178 differences in expression dispersal (coefficient of variation: standard deviation across genes or samples
2179 / mean value) and expression specificity (frequencies of a gene to be detected as unexpressed, defined
2180 as TPM = 2, in any condition) were accessed in a similar manner. Given the variable experimental
2181 conditions associated with different transcriptome data for each species, gene expression values for a
2182 gene were used indiscriminately regardless of the conditions. Correlation tests between the codon
2183 adaptation index (CAI) and gene expression were carried out using the Spearman's rank correlation
2184 analysis tool (P. Wessa, Free Statistics Software, Office for Research Development and Education,
2185 version 1.1.23-r7, https://www.wessa.net/).

2186

2187 **Comparative analysis of gene sets identified by genome-wide analyses of evolutionary history**

2188

2189 Genes identified as belonging to orthogroups that were predicted to be gained at specific nodes of the
2190 phylogenetic tree based on the Dollo parsimony analysis, to belong to either significantly amplified
2191 gene families (binomial analysis) or to belong to gene families that have significantly changed in size
2192 over evolutionary time (CAFE5 analysis), to correspond to founder events (Phylostratigraphy analysis),
2193 to have been remodelled (composite gene analysis) or to have been derived from an HGT (HGT
2194 analysis) were extracted from the output of each of these analysis and aggregated in a single datatable.
2195 Correspondences were established manually between phylogenetic tree nodes and phylostrata and
2196 this information was integrated into the datatable. Counting and calculations of the frequency of
2197 events at specific time points were carried out using *ad hoc* R scripts (R version 4.4.1) and the tidyverse
2198 package (version 2.0.0). Graphs were generated using the ggplot2 package (version 3.5.1). For each
2199 gene, a COG functional category was retrieved from the eggNOG mapper output and the COG
2200 enrichment analysis was carried out in R using the clusterProfiler package (version 4.6.2) by comparing
2201 each set of gene families with the full set of gene families.

2202

2203 **Detection of viral genome insertions and viral regions in algal genomes**

2204

2205 To reduce the dataset size for analysis, 64 Phaeoexplorer and eight public genomes were initially
2206 filtered to retain only contigs that were more than 10 kbp in length. Gene prediction was then carried
2207 out on all contigs using Prodigal[43] (V2.6.3, settings: default, meta) and the resulting proteins were used
2208 as queries against the NCVOG[79] and VOGDB[95] databases using hmmscan (HMMER 3.3.2 with default
2209 settings). The contigs detected by hmmscan were then filtered to retain only sequences with at least
2210 one match to either viral database at a defined e-value cutoff ($1e^{-20}$ for NCVOG, and $1e^{-80}$ for VOGDB).
2211 The resulting positive 4,951 contigs were then analysed using ViralRecall[44] version 2.0 with settings -w
2212 50 -g 1 -b -f -m 2 using the built-in Nucleocytoviricota (NCV) database GVOG and a window size of 50
2213 kbp. To ensure that viral genes were not missed because they had not been annotated by Prodigal, six-
2214 frame translations of the contigs were generated using esl-translate (version 0.48 with default
2215 settings), and the resulting proteins queried against the same databases used by ViralRecall using
2216 hmmsearch (HMMER 3.3.2, settings: -E 1e-10). The ViralRecall results were then parsed using an in-
2217 house workflow. Six-frame translations were removed from the results if they overlapped (even
2218 partially) with any Prodigal gene prediction, as identified using bedtools[45] (v2.29.2; intersect). Likewise,
2219 overlapping six-frame translations and gene predictions with the same NCVOG match were removed
2220 to reduce redundancy. Based on the distance between query sequences with the same GVOG hit,

queries were flagged as frame-shifted (less than 100 bp gap), intron-containing (100-5,000 bp gap) or mono-exonic (greater than 5,000 bp gap). All queries were also checked for overlaps with multi-exonic genes that had been annotated by the Phaeoexplorer gene prediction procedure (using Gmove[28]), and flagged if they did. All queries were then filtered to retain only those that matched a set of key NCV marker genes, identified by NCVOG code (A32, D5 helicase, D5 DNA primase, MCP, DNA polymerase B, SFII and VLTF3) or some Phaeovirus integrase genes (integrase recombinase, integrase resolvase and RNR). The marker gene proteins were clustered with the protein sequences of NCVOGs using MMseqs cluster[46] (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[96]. 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).

2255 The presence of host regions flanking the viral regions was evaluated based on the ViralRecall output
2256 (Table S19C). The percentages of viral regions with two, one or zero flanking regions (longer than 2
2257 kbp) were 25.8%, 15.0% and 59.2%, respectively (i.e. 40.8% of viral regions had at least one flanking
2258 region). Of the viral regions that had two flanking regions, 89.5%, 7.0% and 3.5% had flanking regions
2259 with a total length of >200 kbp, between 20 and 200 kbp or between 2 and 20 kbp, respectively. For
2260 the viral regions that had one flanking region, the corresponding percentages were 25.3%, 36.7% and
2261 38.0%.
2262
2263 **Phylogenetic analysis of viral genes**
2264
2265 Amino acid sequences of manually-curated collections of major capsid protein (MCP) and DNA
2266 polymerase B proteins were aligned using MAFFT (v7.520, settings: --adjustdirectionaccurately --auto
2267 --maxiterate 1000) and phylogenetic trees were generated using IQ-TREE (v 2.2.2.3, settings: -m MFP
2268 -B 1000).
2269
2270 **Metabolic networks**
2271
2272 Genome-scale metabolic networks were reconstructed using AuCoMe[97] version 0.5.1 using the
2273 MetaCyc[98] version 26 database. A first dataset, consisting of the 60 species listed in Table S21 (column
2274 "Metabolic networks") plus two public diatom genomes already used in the initial AuCoMe study
2275 (*Fragilariopsis cylindrus* and *Fistulifera solaris*) was processed to build the largest possible database
2276 (phaeogem) for exploratory comparisons (https://gem-aureme.genouest.org/phaeogem/). Then, a
2277 second comparison was performed on all long-read species plus outgroups. Based on
2278 Multidimensional-scaling (MDS) analyses, the most divergent long-read species (*Choristocarpus*
2279 *tenellus*, *Laminaria digitata, Phaeothamnion wetherbeei* and the public genome of *Sargassum*
2280 *fusiforme*) were excluded to construct a 16 species dataset, balancing assembly quality and
2281 phylogenetic coverage (https://gem-aureme.genouest.org/16bestgem/). The MDS plots shown in Fig
2282 S7 were build using the vegan package, version 2.6-4 (https://github.com/vegandevs/vegan) with R
2283 4.1.2[87], using Jaccard distances. A third stricter dataset (fwgem), enriched in high-quality long-read
2284 Ectocarpales, was built to address questions related to freshwater adaptation (https://gem-
2285 aureme.genouest.org/fwgem/). A set of reactions that were overrepresented in brown algae compared
2286 to the outgroup was created (Table S4) by taking reactions present in 100% of brown algae and less
2287 than 70% of outgroups. Reactions corresponding to genes lost in freshwater species were also
2288 extracted (Table S22). These reaction sets were extracted from all the networks using the Aucomana
2289 library (https://github.com/PaulineGHG/aucomana). Online wikis (phaeogem, 16bestgem and fwgem)

2290 were generated using AuReMe[99].

2291

2292 **CAZymes**

2293

2294 CAZyme genes were identified based on shared homology with biochemically characterised proteins,
2295 either individually or as hidden Markov model (HMM) profiles. For phylogenetic analyses, proteins
2296 were aligned using MAFFT[47] with the iterative refinement method and the scoring matrix Blosum62.
2297 The alignments were manually refined and trees were constructed using the maximum likelihood
2298 approach. Alignment reliability was tested by a bootstrap analysis using 100 resamplings of the
2299 dataset. Only bootstrap values above 60% are shown. The phylogenetic trees were displayed with
2300 MEGA[48]. The annotated genes are listed in Table S6 with accession numbers.

2301

2302 **Sulfatases**

2303

2304 The sulfatases encoded by each brown algal genome were identified and assigned to their respective
2305 family and subfamily using the SulfAtlas database[80,81] (https://sulfatlas.sb-roscoff.fr/). Each predicted
2306 proteome was first submitted to the SulfAtlas HHM server (https://sulfatlas.sb-
2307 roscoff.fr/sulfatlashmm/), which allows rapid identification of sulfatase candidates and (sub)family
2308 assignment using hidden Markov model profiles for each SulfAtlas (sub)family. Each sulfatase
2309 candidate sequence was then used as a query in a BLASTp[3] search against the SulfAtlas database
2310 (https://blast.sb-roscoff.fr/sulfatlas/). Sequences with at least 50% identity with sulfatases from
2311 marine bacteria or other marine microorganisms were considered to be contaminants. Below this
2312 threshold, additional examination of the predicted gene structure and genomic context of the
2313 candidate sequence was undertaken to identify possible horizontal gene transfers.

2314

2315 **Haloperoxidases**

2316

2317 vHPO genes were identified based on sequence homology and active site conservation. Maximum
2318 likelihood phylogenetic analyses were carried out using the NGphylogeny platform at
2319 https://ngphylogeny.fr/. MAFFT was used to align vHPO sequences and alignments were automatically
2320 curated with TrimAl[49], leading to the selection of 444 informative positions from the initial 1450
2321 positions for the algal-type vHPOs and 402 informative positions from the initial 1078 positions for the
2322 bacterial-type vHPOs. Maximum likelihood trees were constructed using FastTree with the WAG+G
2323 gene model and 1000 bootstrap replicates. Maximum likelihood Newick files were formatted as

2324 circular representations using iTOL. Only bootstrap values between 0.7 and 1 were conserved. The lists
2325 of annotated vHPO genes are in Tables S7 and S8.
2326
2327 **Ion channels**
2328
2329 A search was carried out for 12 classes of ion channel in the predicted proteomes of the 21
2330 Phaeoexplorer reference genomes plus those of two diatoms, *Phaeodactylum tricornutum* and
2331 *Thalassiosira pseudonana* (Table S16). Predicted proteomes were screened using BLASTp[3] and query
2332 sequences from *Ectocarpus* species 7 and seven other species from diverse eukaryotic taxa (Table S16).
2333
2334 **Membrane-localised proteins**
2335
2336 Membrane protein family genes were identified either by carrying out BLASTp[3] searches of the
2337 predicted Phaeoexplorer proteomes using *Ectocarpus* species 7 sequences as queries or by recovering
2338 orthogroups containing the relevant *Ectocarpus* species 7 sequences as members. The BLASTp
2339 approach was used for DEK1-like calpains, fasciclins, tetraspanins, CHASE, ethylene-binding-domain-
2340 like and MASE1 domain histidine kinases whereas the orthogroup approach was used to recover other
2341 members of the histidine kinase family. Both approaches were used to search for integrins and
2342 transmembrane receptor kinases. For integrins the two methods detected exactly the same set of
2343 proteins. For receptor kinases the BLASTp and orthogroup analyses detected 99.3% and 98.3% of the
2344 269 genes, respectively. For these analyses, either the whole genome dataset was analysed or only the
2345 set of 21 reference genomes (Table S21), depending on the size of the gene family.
2346
2347 Manually-curated histidine kinase protein families were aligned with Muscle[34] before phylogenetic
2348 tree construction using IQ-TREE 2[75] (version 2.3.4) with automatic model selection and 1000
2349 bootstraps.
2350
2351 **Transcription-associated proteins**
2352
2353 TAPscan v4[51] was used to analyse the transcription-associated protein (TAP) complements of 21
2354 species. TAPscan[50] is a comprehensive tool for annotating TAPs based on the detection of highly
2355 conserved protein domains using HMM profiles with specific thresholds and coverage cut-offs.
2356 Following detection, specialised rules are applied to assign protein sequences to TAP families based
2357 on the detected domains. TAPscan v4 can assign proteins to 138 different TAP (sub)families with high
2358 accuracy.

2359

2360 **EsV-1-7 domain proteins**

2361

2362 EsV-1-7 domain proteins were identified in the 31 brown algal and sister taxa genomes (Table S21) by
2363 recovering the members of all orthogroups (with the exception of OG0000001, which is a very large
2364 OG that consisting principally of transposon sequences) that either contained one or more of a curated
2365 set of 101 EsV-1-7 domain proteins[100] for *Ectocarpus* species 7 or contained an EsV-1-7 domain protein
2366 based on a match to the Pfam EsV-1-7 motif PF19114. The recovered proteins were screened manually
2367 for the presence of at least one EsV-1-7 domain and a total of 2018 were finally identified as members
2368 of the EsV-1-7 family.

2369

2370 To identify orthologues of the EsV-1-7 protein IMMEDIATE UPRIGHT[100] (IMM), BLASTp searches of 25
2371 brown algal and four sister taxa predicted proteomes were carried out with the amino-terminal
2372 domain of the IMM protein minus the five EsV-1-7 repeats as this domain is unique to IMM. Proteins
2373 were retained as IMM orthologues if they were more similar to IMM than to the most closely-related
2374 protein in *Ectocarpus* species 7, Ec-17_002150.

2375

2376 **Histones**

2377

2378 Histone protein sequences were analysed in *Ascophyllum nodosum*, *Chordaria linearis*,
2379 *Chrysoparadoxa australica*, *Desmarestia herbacea*, *Dictyota dichotoma*, *Discosporangium*
2380 *mesarthrocarpum*, *Ectocarpus crouaniorum*, *Ectocarpus fasciculatus*, *Ectocarpus siliculosus*, *Fucus*
2381 *serratus*, *Heterosigma akashiwo, Pleurocladia lacustris*, *Porterinema fluviatile*, *Pylaiella littoralis*,
2382 *Saccharina latissima*, *Sargassum fusiform*, *Schizocladia ischiensis*, *Scytosiphon promiscuus*, *Sphacelaria*
2383 *rigidula*, *Tribonema minus* and *Undaria pinnatifida* using BLASTp against the complete predicted
2384 proteomes (https://blast.sb-roscoff.fr/phaeoexplorer/) with the histone protein sequences from the
2385 diatom *Phaeodactylum tricornutum* as queries. The genes and transcripts coding for the identified
2386 histones were then retrieved from the genomes and predicted transcripts using BLAST
2387 (https://blast.sb-roscoff.fr/phaeoexplorer/). The proteins encoded by the identified genes and
2388 transcripts were predicted with the Expasy web translator (https://web.expasy.org/translate/). All the
2389 identified histone protein sequences and the corresponding transcripts IDs are reported in Table S23.
2390 In order to identify truncated proteins or incorrect start codons, the following constraints were
2391 applied: H2A proteins must start with the SGKGKGGR sequence, H2B with AKTP, canonical H3.1 and
2392 variants H3.3 with ARTKQT and H4 with SGRGKGGKGLGKGG. For the linker histone H1, protein
2393 sequences had to be lysine-rich and sequences with incorrect start codons were determined by

2394 alignments of all identified H1 proteins. For proteins with incorrect start codons, the region upstream
2395 of the correct start codon was removed. For truncated proteins, *i.e.* proteins whose transcripts lacked
2396 either the start (no methionine) or stop codons, the protein sequence was completed based on
2397 alignment with the corresponding genomic region using the Geneious 11.0.5 software. When the
2398 sequence could not be completed, a BLAST was performed against the Phaeoexplorer *de novo*
2399 transcriptomes (https://blast.sb-roscoff.fr/phaeoexplorer_denovo/) when this data was available (this
2400 was not possible for the public genomes *T. minus*, *U. pinnatifida* and *S. fusiforme*). Based on the
2401 nomenclature established by[101], H3 histones were classified as follows: canonical H3.1 proteins
2402 harbour AT residues at positions 31-32 while histone variants H3.3 harbour TA residues, H3 proteins
2403 with other residues at positions 31-32 were named H3.4 and so on. CenH3 variants of H3 were
2404 identified by analysis with Panther 17.0 (www.pantherdb.org/tools/sequenceSearchForm.jsp?) and/or
2405 Interproscan[52] 94.0 (www.ebi.ac.uk/interpro/search/sequence/).
2406
2407 Species abbreviations used in histone phylogenetic trees are: Atr, *Amborella trichopoda*; At,
2408 *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Di, *Dictyostellium discoideum*; Dr, *Danio rerio*; Dm,
2409 *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Pp, *Physarum polycephalum*; Ppa
2410 *Physcomitrium patens*; Sc, *Saccharomyces cerevisiae*; Tm, *Tetrahymena thermophila*; Zm, *Zea mays*;
2411 Mp, *Marchantia polymorpha* subsp. *Ruderalis*; Bd, *Brachypodium distachyon*; Ccr, *Chondrus crispus*;
2412 Gs, *Galdieria sulphuraria*; Cm, *Cyanidioschyzon merolae*; Cr, *Chlamydomonas reinhardtii*; Ol,
2413 *Ostreococcus luciminarinus*; Ot, *Ostreococcus tauri*; To, Thalassiosira oceanica; Pt, *Phaeodactylum*
2414 *tricornutum*; An, *Ascophyllum nodosum*; Cl, *Chordaria linearis*; Ca, *Chrysoparadoxa australica*; Dh,
2415 *Desmarestia herbacea*; Ddi, *Dictyota dichotoma*; Dme, *Discosporangium mesarthrocarpum*; Ec,
2416 *Ectocarpus crouaniorum*; Ef, *Ectocarpus fasciculatus*; Es, *Ectocarpus siliculosus*; Fse, *Fucus serratus*; Ha,
2417 *Heterosigma akashiwo*; Pla, *Pleurocladia lacustris*; Pf, *Porterinema fluviatile*; Pli, *Pylaiella littoralis*; Sl,
2418 *Saccharina latissima*; Sf, *Sargassum fusiform*; Si, *Schizocladia ischiensis*; Sp, *Scytosiphon promiscuus*;
2419 Sri, *Sphacelaria rigidula*; Tm, *Tribonema minus*; Up, *Undaria pinnatifida*.
2420
2421 **DNA methyltransferases**
2422
2423 Searches were carried out for methyltransferases and demethylases in the predicted proteomes of 20
2424 of the high quality brown algal reference genome assemblies (based on Nanopore long-read sequence)
2425 plus the sister taxa *Chrysoparadoxa australica* and *Schizocladia ischiensis* using BLASTp (Table S21). A
2426 methyltransferase reference database was constructed by recovering sequences from NCBI, ENSEMBL
2427 and UniProtKB. Methyltransferase sequences were recovered for stramenopiles such as
2428 *Nannochloropsis gaditana*, the diatom *Phaeodactylum tricornutum*, the oomycete *Phytophthora*

2429 *infestans* and for species from more distant lineages including *Arabidopsis thaliana*, *Homo sapiens* and
2430 the fungus *Neurospora crassa*. The proteomes of the selected brown algal strains were then queried
2431 against this database using BLASTp and matches with an e-value of < 0.001, a bit score > 70, a maximum
2432 gap of 5 and percentage identity of >30% were retained. The retained matches were screened against
2433 the NCBI, UniProt and SwissProt databases to identify and remove contaminating bacterial or viral
2434 proteins. Methyltransferase domains were detected in the retained matches using the Simple Modular
2435 Architecture Research Tool (SMART)[83] domain architecture analysis and InterPro searches
2436 (https://www.ebi.ac.uk/interpro/). Sequences with methyltransferase domains were retained for
2437 further analysis. Validated brown algal methyltransferases were aligned with reference
2438 methyltransferases using Clustal[53] 2.1.
2439
2440 **Spliceosome**
2441
2442 Components of the Major Spliceosome were identified using a reference set of 147 human
2443 components (https://www.genenames.org/data/genegroup/#!/group/1518), excluding the five small
2444 nuclear RNAs (snRNAs). Including isoforms, this query set consisted of 626 proteins. These proteins
2445 were used to screen the predicted proteomes of 54 genomes (Table S21) using BLASTp and matches
2446 were retained if they had an e-value of at most $1e^{-40}$ and coverage >30%. Searches were also carried
2447 out for components of LSm and Sm complexes which have roles as scaffolds in the formation of
2448 ribonucleoprotein particles (RNPs), in the maturation of mRNAs (including splicing, such as the
2449 cytoplasmic complex LSm1-7, LSm2-8 which is part of the core U6 snRNP and other complexes
2450 important for the formation of the 3' ends of histone transcripts), in the assembly of P-Bodies and in
2451 the maintenance of telomeres.
2452
2453 **Flagella proteins**
2454
2455 A previous proteomic analysis of anterior and posterior flagella of the brown alga *Colpomenia bullosa*
2456 identified a total of 592 proteins across the two proteomes[102]. Here the *Ectocarpus* species 7
2457 orthologues of 70 of these proteins that had been detected with a very high level of confidence were
2458 used to identify the corresponding orthogroups and the presence or absence of these orthogroups
2459 was scored for seven representative species (Table S21).
2460
2461 **Detection of *Porterinema fluviatile* genes differentially expressed in freshwater and seawater**
2462

2463 Six independent cultures of *Porterinema fluviatile* were cultivated for four weeks in 140 mm Petri
2464 dishes with Provasoli-enriched culture medium[103], which was renewed every two weeks. For three
2465 Petri dishes, the culture medium was based on autoclaved natural seawater (high salinity treatment),
2466 for the other three Petri Dishes natural seawater was diluted 1:19 vol/vol with distilled water (low
2467 salinity treatment). Cultures were harvested with 40 µm nylon sieves, dried with a paper towel, and
2468 immediately frozen in liquid nitrogen. RNA extraction library construction and sequencing were carried
2469 out as described in section "RNA extraction, Illumina RNA-seq library preparation and sequencing".
2470 RNA-seq reads were cleaned with Trimmomatic[18] V0.38 and then mapped to the *P. fluviatile* genome
2471 using Kallisto[55] version 0.44.0. Differentially expressed genes were identified using the Deseq2
2472 package[56] included in Bioconductor version 3.11, considering genes with an adjusted $p < 0.05$ and a
2473 log2 fold-change > 1 as differentially expressed. The complete list of differentially expressed genes is
2474 provided in Table S24. To compare the differentially expressed genes in *P. fluviatile* with an equivalent
2475 set previously identified for *Ectocarpus subulatus* in a microarray experiment using nearly identical
2476 growth conditions[104], orthologues in the two species were detected using Orthofinder version 2.3.3.
2477 Of the 10,066 shared orthogroups, 6,606 had microarray expression data for *E. subulatus*. This
2478 information was used to classify differentially expressed genes for the two species as either shared
2479 orthologues or as lineage-specific.
2480
2481 **Identification of genes with generation-biased expression patterns**
2482
2483 RNA-seq data (two to five replicates per condition) was recovered for gametophyte and sporophyte
2484 generations of ten species (Table S21). Data quality was assessed with FastQC[57] version 0.11.9 and
2485 sequences were then trimmed with Trim Galore version 0.6.5 with the parameters --length 50, - quality
2486 24, --stringency 6, --max_n 3. The cleaned reads were mapped onto the corresponding genome for
2487 each species using HISAT2 version 2.1.0 with default options. Counting was carried out with
2488 featureCounts[58] from the subread package (version 2.0.1) on CDS features grouped by Parent.
2489 Transcript Per Kilobase Million (TPM) tables were generated for all conditions and differentially
2490 expressed genes were detected using DESeq2[56] version 1.30.1. Genes were classified into six categories
2491 based on the differential expression analysis and the TPM values: gametophyte-biased, mean TPM ≥1
2492 in gametophyte and sporophyte, log2(fold change) ≥1, adjusted *p*-value <0.05; sporophyte-biased:
2493 mean TPM ≥1 in gametophyte and sporophyte, log2(fold change) ≤-1, adjusted *p*-value <0.05;
2494 gametophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, log2(fold change) ≥1,
2495 adjusted *p*-value <0.05; sporophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte,
2496 log2(fold change) ≤-1, adjusted p-value <0.05; unbiased genes: mean gametophyte and sporophyte

2497 TPMs ≥1, log2(fold change) <1 or >-1 and/or adjusted p-value ≥0.05; unexpressed genes, mean
2498 gametophyte and sporophyte TPM <1.

2499

2500 **Life cycle and thallus architecture**

2501

2502 **Genome dataset and traits.** To study the impact of body architecture, the brown algae were divided
2503 into three categories: 22 filamentous species, eight simple parenchymatous species and 13 species
2504 with elaborate thalli (Table S21). For the life-cycle-based assessment, the groups were: 30 haploid-
2505 diploid species and six diploid species (Table S21). Body architecture information was available for 43
2506 species, and life cycle information was available for 36 species; species without body plan or life cycle
2507 information were not used in subsequent analyses. Two approaches were used to estimate selection
2508 intensity across the phylogeny, (i) a model-based method, and (ii) by evaluating codon usage bias and
2509 nucleotide composition. Two evolutionary models were used, one based on architecture and the other
2510 based on life cycle. For model-based methods the phylogeny was categorised based on the above
2511 traits, and selection intensity parameters were estimated using PAML[59] version 4.9i. Rate estimates
2512 were obtained for non-synonymous substitutions (dN), synonymous substitutions (dS) and omega
2513 (dN/dS) for the multiple sequence alignments of all genes within each orthogroup using the variable-
2514 ratio model of CODEML from PAML, which allows different omegas for different branch categories.
2515 The traits were assigned to the branches of the phylogeny using ancestral state estimation by
2516 stochastic mapping with the phytools R package[60,87].

2517

2518 **Evolutionary models to study impacts of body architecture.** To study variation in selection intensity
2519 as a function of body architecture, we devised a model with the following trait categories:
2520 filamentous/pseudoparenchymatous (simple cell division and organisation on a single plane),
2521 parenchymatous (cell division and organisation on multiple planes) and elaborate thallus (tissue
2522 differentiation). To ensure that at least 50% of the species in each category were used in the analysis,
2523 we selected orthogroups (OGs) that contained at least 11 members for filamentous, at least four
2524 members for parenchymatous and at least six members for elaborate thallus algae. Using this filter,
2525 1068 OGs were obtained, on which the model based on body architecture was fitted. Selection
2526 intensity parameters [rate of non-synonymous substitution (dN), rate of synonymous substitution (dS)
2527 and omega (dN/dS)] were estimated for the three trait categories for each gene alignment. We used
2528 the Wilcoxon signed-rank test to evaluate the statistical significance of differences between the
2529 selection intensity parameters (dN, dS and dN/dS) for each category.

2530

2531 **Evolutionary models to study the impacts of life cycle.** The impact of life cycle on molecular evolution
2532 was assessed using a model with two categories consisting of diplontic and haplodiplontic species. For
2533 this model we used 1,058 OGs that contained at least three members for diploid species and at least
2534 15 members for haploid-diploid species. Using alignments of the gene within the OGs, we estimated
2535 the selection intensity parameters for the different categories and applied the Wilcoxon signed-rank
2536 test to assess the statistical significance of differences in selection intensity between the diploid and
2537 haploid-diploid life cycles.
2538
2539 **Selection of intensity parameters.** Omega (dN/dS) provides an estimate of the ratio of substitutions
2540 at sites under selection compared to neutral sites, and is generally used to infer the strength of
2541 purifying selection. Omega needs to be interpreted with caution because not all synonymous sites are
2542 neutral[105] and also synonymous substitutions are often underestimated due to saturation of
2543 synonymous sites, which might in turn impact the omega ratios[106]. Omega values lower than one
2544 indicate substitutions are less frequent at sites under selection compared to neutral sites and are
2545 characteristic of highly conserved genes or genes evolving under strong purifying selection. As we used
2546 primarily low copy number genes in this study, the analysed genes were expected to evolve under
2547 strong purifying selection, with omega values much lower than one. Using omega for near neutral
2548 studies is challenging because near neutral sites are determined by effective population size, that is to
2549 say, sites under mild selection constraint in larger populations can behave as neutral sites in smaller
2550 populations. It is therefore difficult to infer the amount of mutation from relative values of omega. In
2551 order to obtain better insight into selection intensity, mutation accumulation was not only investigated
2552 using rates of synonymous (dS) and non-synonymous (dN) substitutions but also by estimating codon
2553 bias and nucleotide composition. Codon usage bias was used, in addition to omega, to infer selection
2554 intensity across species as the former reflects selection efficacy at synonymous sites[107–109]. We inferred
2555 codon usage bias by estimating the effective number of codons (ENC) for each species using the enc
2556 method from the VHICA package[61,89]. The effective number of codons (ENC) quantifies the extent of
2557 deviation of codon usage of a gene from equal usage of synonymous codons. For the standard genetic
2558 code, ENC values range from 20 (where a single codon is used per amino acid implying strong codon
2559 usage bias) to 61 (implies that all synonymous codons are equally used for each amino acid[110]). Low
2560 ENC indicates constrained use of codons, which potentially highlights stronger codon bias due to
2561 stronger selection at synonymous sites. As nucleotide composition can also influence codon bias, we
2562 calculated the overall GC composition, GC at the third codon position (GC3) and the theoretical
2563 expected ENC (EENC) based on GC3 using local R scripts. The lower the observed ENC (OENC, estimated
2564 from the gene sequence) relative to EENC, the stronger the influence of selection due to translation
2565 on codon usage. This was studied by estimating the difference (DENC = EENC - OENC) between the

2566 expected ENC and the observed ENC[111]. Positive DENC indicates a role for selection constraints on
2567 codon usage in addition to the influence of nucleotide composition. DENC values of zero or less
2568 indicate that codon bias is entirely driven by nucleotide composition. DENC values were used to study
2569 the influence of translation selection and nucleotide composition on codon usage bias.
2570
2571 **Assembly and analysis of organellar genomes**
2572
2573 Plastid and mitochondrial genomes were assembled *de novo* using NOVOPlasty[62] v3.7 and *rbcL* and
2574 *cox1* nucleotide sequences as seeds. Assembled genomes were checked by aligning reads using
2575 Bowtie2[5] v2.3.5.1 and processed with SAMtools[63] v1.5. Annotation of protein-coding genes was
2576 performed with GeSeq[64] v2.03. Annotation of tRNAs, tmRNAs and rRNAs was performed with
2577 ARAGORN[65] v1.2.38.
2578
2579 Maximum-likelihood (ML) phylogenetic trees were constructed using 92 plastid genomes (11 non-
2580 brown outgroup sequences) and 89 mitochondrial genomes (seven non-brown outgroup sequences).
2581 The conserved coding-region amino acid sequences of 139 plastid genes (31,159 amino acids) and 35
2582 mitochondrial genes (7,461 amino acids) were used to construct these phylogenetic trees. The
2583 sequence for each gene was aligned individually using MAFFT[47] v7 (--maxiterate 1000) and then
2584 concatenated. Alignment partitions were assigned based on genes. Each of the aligned gene sequences
2585 was trimmed with trimAl[49] v1.2 (-automated1). ML phylogenetic trees were constructed with IQ-TREE
2586 2[75]. The protein substitution models in each gene partition were selected using ModelFinder[66].
2587 Statistical support for tree branches was assessed with 1,000 replicates of ultrafast bootstrap
2588 (UFBoot2)[67].
2589
2590 **Analysis of *Ectocarpus* genome synteny**
2591
2592 Global genome synteny analysis was performed using SynMap[68] on the CoGe platform
2593 (https://genomevolution.org/coge/) with the following genomes: *Ectocarpus crouaniorum* male,
2594 *Ectocarpus fasciculatus* male, *Ectocarpus siliculosus* male, *Ectocarpus* species 7 male and *Ectocarpus*
2595 *subulatus*. SynMap identifies syntenic regions between two or more genomes using a combination of
2596 sequence similarity and collinearity algorithms. Last[112] was used as the BLAST algorithm and syntenic
2597 gene pairs were identified using DAGChainer[69] with settings "Relative Gene Order", -D = 20, -A =5.
2598 Neighbouring syntenic blocks were merged into larger blocks. Substitution rates between the synthetic
2599 CDS pairs were calculated using CodeML[59], which was also implemented in SynMap, CoGe. In detail,
2600 protein sequences were aligned using the Needleman-Wunsch algorithm implemented in nwalign

2601 (https://pypi.org/project/nwalign/) and then translated back to aligned codons. CodeML was run five
2602 times for each alignment using the default parameters and the lowest dS was retained, with the upper
2603 cutoff for dS values set at 2. *Ectocarpus* genes were grouped according to their age based on the
2604 phylostratigraphic analysis and by chromosomal location based on their chromosome position in
2605 *Ectocarpus* species 7. All plots and statistical analysis were carried out in R version v.4.3.1. Local
2606 synteny analysis was based on orthologous genes as identified by Orthofinder.
2607
2608 **Analysis of *Ectocarpus* gene evolution**
2609
2610 Protein sequence alignments were used to remove gaps with trimAl[49] and then translated back to DNA
2611 with backtranseq[113]. Only DNA fasta files with a minimum of 70 bp were retained (831 single-copy
2612 orthologs). PhyML trees were built with Geneious v11.1.5 (https://www.geneious.com). Maximum
2613 likelihood analysis was carried out to detect site specific, branch-site specific and branch specific
2614 positive selection as well as sites under negative selection, using PAML[114].
2615
2616 **Phylogenetic analysis of *Ectocarpus* species**
2617
2618 Phylogenetic analysis was carried out for 11 *Ectocarpus* species plus *Scytosiphon promiscuus* as an
2619 outgroup (Table S21). Of the 933 single-copy orthogroups identified for these 12 species, 261 high-
2620 confidence alignments were retained for gene tree and species tree inferences following the removal
2621 of low-quality alignments using BMGE[115]. Bayesian inference of the phylogeny of the *Ectocarpus*
2622 species complex was performed using BEAST[70] v2.7. The analysis was conducted under the multi-
2623 species coalescent (MSC) model, implemented in StarBEAST3[71] v1.1.7. The MSC model coestimates
2624 gene trees and the species tree within a multispecies coalescent framework, enabling the assessment
2625 of incongruences among genes with respect to the species tree. To account for substitution model
2626 uncertainty, bModelTest[72] was employed to average over a set of substitution models for each
2627 alignment. StarBEAST3 was run under both the Yule model and the strict clock model. A total of
2628 300,000,000 Markov Chain Monte Carlo (MCMC) generations were conducted, with tree states stored
2629 every 50,000 iterations. Posterior tree samples were combined, discarding the initial 10% burn-in,
2630 using LogCombiner v2.4.7. A maximum clade credibility tree was generated using TreeAnnotator[70]
2631 v2.4.7.
2632
2633 ***Ectocarpus* introgression analysis**
2634

2635 To distinguish introgression from shared ancestry, D estimates (i.e. ABBA-BABA tests) were generated
2636 from 36 four-taxon combinations[116]: four to test the level of introgression within clade 1 (i.e. *E.*
2637 *subulatus*, *E. crouaniorum*, *Ectocarpus* species 1, *Ectocarpus* species 2), 20 to test the level of
2638 introgression within clade 2 (i.e. *Ectocarpus* species 6, *Ectocarpus* species 7, *Ectocarpus* species 5,
2639 *Ectocarpus* species 9, *E. siliculosus*, *Ectocarpus* species 3) and 12 to test the level of introgression
2640 between these two clades. Tests were designed using a four-taxon fixed phylogeny (((P1,P2)P3)O),
2641 where P1 and P2 are closely related species from the same clade, P3 is a more divergent species that
2642 may have experienced admixture with one or both of the (P1,P2) taxa, and an out-group (O). *E.*
2643 *fasciculatus* was used as the out-group taxon for all ABBA-BABA tests. Details about how P1, P2 and
2644 P3 taxa were selected for each test are given in Table S25. Previous results of species tree inference
2645 were used to inform subsequent ABBA–BABA tests and to define the (((P1,P2)P3)O) phylogenies.
2646 ABBAs are sites at which the derived allele (called B) is shared between the taxa P2 and P3, whereas
2647 P1 carries the ancestral allele (called A), as defined by the outgroup while BABAs are sites at which the
2648 derived allele is shared between P1 and P3, whereas P2 carries the ancestral allele. Under incomplete
2649 lineage sorting, conflicting ABBA and BABA patterns should occur in equal frequencies, resulting in a D
2650 statistic equal to zero. Historical gene flow between P2 and P3 causes an excess of ABBA, generating
2651 positive values of D. Historical gene flow between P1 and P3 causes an excess of BABA, generating
2652 negative values of D. Patterson's D-statistic was calculated for the concatenated alignments of the 261
2653 orthologroups. Significance was detected using a block-jackknifing approach[116–118], with a block size of
2654 5 Kbp. For the jackknife procedure, one block of adjacent sites was removed n times. A Z-score was
2655 finally obtained by dividing the value of the D statistic by the standard error over n sequences of 5 Kbp.
2656 The ParimonySplits network was reconstructed for the genus *Ectocarpus* using SplitsTree 4[73] (version
2657 4.14.6) with 1000 bootstrap replicates.
2658
2659 **QUANTIFICATION AND STATISTICAL ANALYSIS**
2660
2661 All statistical analyses are described in detail in the relevant sections of the "Method details" section
2662 and the results of statistical tests are shown in the tables and figures.
2663
2664 **ADDITIONAL RESOURCES**
2665
2666 The Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr) provides access to all the annotated
2667 genome assemblies described in this study as downloadable files. The output files from the
2668 Orthofinder[32], Interproscan[52], Hectar[74] and eggNOG-mapper[42] analyses, together with the results of
2669 the various DIAMOND[26] sequence similarity analyses (see section "Analyses aimed at deducing

2670 functional characteristics of predicted proteins"), can also be downloaded. In addition, the site
2671 provides genome browser interfaces for the genomes and multiple additional tools and resources
2672 including BLAST interfaces for genomes, proteomes and *de novo* transcriptomes, various experimental
2673 protocols, an RShiny-based transcriptomic aggregator for the model brown alga *Ectocarpus* species 7
2674 strain Ec32 and a link to genome-wide metabolic networks for the Phaeoexplorer species.
2675
2676 Additional data have been deposited in the CNRS Research Data depository under the title "Data for
2677 Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key
2678 components of coastal ecosystems" (DOI: https://doi.org/10.57745/9U1J85).
2679
2680
2681 **SUPPLEMENTAL RESULTS**
2682
2683 **Structural features of brown algal genomes**
2684
2685 The brown algal genomes generated by this project exhibit differences in assembly size and GC content
2686 (Figure S16A, Table S1). The cumulative lengths of intergenic regions, the cumulative lengths of
2687 transposable elements and cumulative intron size were all correlated with genome size (Figures S6A
2688 and S16B, Table S1). An earlier analysis of the *Ectocarpus* species 7 (at that time classified as *Ectocarpus*
2689 *siliculosus*) genome found a strong tendency for adjacent genes to be located on opposite strands of
2690 the chromosome, a feature that is more typical of very small, compact genomes[85]. Analysis of the
2691 Phaeoexplorer genomes indicated that this characteristic is common to the whole FDI
2692 (Fucophycideae/Dictyotales/Ishigeales) clade but the tendency is strongest in the Ectocarpales, where
2693 the proportion of alternating gene pairs can be greater than 60% (Figure S6A, Table S1). Another
2694 feature that is shared by all brown algae, plus some of the closely-related sister taxa, is that intergenic
2695 intervals between divergently transcribed gene pairs tend to be shorter than those between tandem
2696 gene pairs (Figure S6A, Table S1). Again, this is a feature that was previously observed in *Ectocarpus*
2697 species 7 where it is associated with the presence of common nucleosome-depleted regions and
2698 histone modification peaks at the transcription start sites of the divergently transcribed genes[119].
2699 Tandemly-repeated genes are more numerous in the Phaeophyceae than in closely-related outgroup
2700 clades (Figure S6A, Table S1) and, although they only represent modest proportions of the total gene
2701 sets (around 6% for the Ectocarpales and 10-12% for the Fucales, for example), they may have played
2702 important roles in brown algal adaptation and diversification.
2703
2704 **Composite genes**

2705

2706 Composite genes (i.e. genes derived from domain fusions and fissions) tended to be retained at a
2707 higher frequency than non-composite genes during the most recent phase of brown algal evolution
2708 but a larger proportion of these genes had unknown functions than for non-composite genes (Figure
2709 S17A). Figure S17B shows an example of a novel domain association that is predicted to have arisen
2710 just prior to the emergence of the brown algal lineage.

2711

2712 **Horizontal gene transfer**

2713

2714 Genes that are predicted to be derived from HGT exhibit different ranges of codon usage bias and gene
2715 feature lengths to non-HGT-derived genes (Figure S17C), and have diverse predicted functions, with
2716 carbohydrate transport and metabolism and cell wall/membrane/envelope biogenesis being the most
2717 prevalent (Figure S17D).

2718

2719 **Brown algal introns and the spliceosome**

2720

2721 An analysis of intron conservation profiles using a set of single-copy orthologous proteins conserved
2722 across Phaeophyceae and close sister groups indicated that more than 90% of brown algal introns are
2723 shared with at least one of the sister clades, and therefore that most introns are ancient. This analysis
2724 also confirmed that a high proportion of *C. australica* introns are unique to that species, suggesting
2725 that a burst of intron gain occurred during the evolution of this taxa (Figures S6B, S18A, S18B and
2726 S18C). Analysis of the *Ectocarpus* genomes showed that intron sizes, phases and positions have been
2727 strongly conserved across the genus (Figure S18D).

2728

2729 The vast majority of splice sites in Phaeophyceae are of the canonical GT-AG type (Figure S18C) and no
2730 signatures of minor U12 introns were identified in intronic sequences, in accordance with a previous
2731 study[120]. Consistent with this observation, an analysis of spliceosome proteins using a recent reference
2732 dataset[121] found that, with the exception of ZCRB1, none of the 12 minor spliceosome genes were
2733 present in brown algal genomes (Table S26), indicating that the minor intron spliceosomal associated
2734 machinery was lost in the Phaeophyceae. As far as the Sm/Lsm spliceosome is concerned, this analysis
2735 identified lineage-specific gene duplications and loss of some components[122] (Figures S18E and S18F).

2736

2737 **Long non-coding RNAs**

2738

2739 Long non-coding RNAs (lncRNAs) play key regulatory roles in diverse eukaryotic taxa[123,124]. This class of
2740 gene has been characterised in *Ectocarpus* species 7 and *Saccharina japonica* and there is evidence
2741 that some of these genes may play regulatory roles during development and life cycle progression in
2742 brown algae[86,125,126]. Eleven transcriptomes (Table S27) were searched for lncRNAs using a meta-
2743 classifier, votingLNC, that integrates information from ten lncRNA-detection programs. This analysis
2744 indicated that brown algal genomes contain about twice as many lncRNA genes as protein-coding
2745 genes (Figure S19A). The lncRNA transcripts tend to be shorter than mRNAs and to be less GC-rich
2746 (Figure S19B).
2747
2748 **Gene loss from plastid genomes**
2749
2750 Comparison of the plastid genomes generated by the Phaeoexplorer project (Table S28) identified
2751 gene loss events during the diversification of the brown algal orders that affected five different loci
2752 (cp74, rpl32, rbcR, rpl9 and syfB). Two of these genes have been lost more than once within the lineage
2753 (Figure S20A, Table S29). Gene losses have previously been reported[127,128] for three of these genes
2754 (rpl32, rbcR and syfB).
2755
2756 **Evidence for mitochondrial introgression events between species**
2757
2758 Phylogenetic trees based on either the plastid or mitochondrial gene sets were congruent with the
2759 phylogeny based on nuclear gene data[89] at the ordinal level (Figure S20B). For the plastid genome, this
2760 congruency was also observed at the sub-ordinal level but, interestingly, a marked discordance was
2761 observed between the mitochondrial tree and the nuclear and plastid trees for the Ectocarpales,
2762 indicating possible mitochondrial introgression events between species within this order (Figure S20B),
2763 as has been observed in other eukaryotic lineages, for example *Picea*[129] and chipmunks[130].
2764
2765 **Metabolic pathway analysis using AuCoMe**
2766
2767 The robust set of 24 core reactions found mainly in brown algae by the AuCoMe analysis (see main
2768 text) included reactions catalysed by indole-3-pyruvate oxygenases (EC 1.14.13.168) and a PQQ-
2769 dependent sugar dehydrogenase (Table S4). Indole-3-pyruvate oxygenase was previously thought to
2770 be absent from brown algae and it has been proposed that its activity could be complemented by
2771 associated bacteria[131] to enable auxin biosynthesis. The identification of indole-3-pyruvate oxygenases
2772 by the current study indicates that a full auxin biosynthetic pathway may be present in brown algae.
2773 Concerning PQQ-dependent sugar dehydrogenase, the most similar bacterial protein is a membrane-

2774 bound glucose sorbosone dehydrogenase involved in vitamin C biosynthesis but the closest eukaryotic
2775 match, outside stramenopiles, is an animal hedghog-interacting protein. This suggests this the brown
2776 algal protein could either have a metabolic and/or a signalling function.
2777
2778 **Polyketide metabolism**
2779
2780 Polyketides are a group of secondary metabolites that exhibit a wide range of bioactivities. Brown
2781 algae possess three classes of type III polyketide synthase (Figures 3A, 4A and S8E, Table S9). The first
2782 class (PKS1) was acquired early in stramenopile evolution, after divergence from the oomycetes but
2783 before the divergence of Diatomista and Chrysista (Figure 4A). Acquisition of PKS1 probably involved
2784 horizontal transfer from a bacterium. Distance analysis of a maximum likelihood phylogenetic tree
2785 (Figure 4A) indicated that PKS1 was more likely to have been acquired from a cyanobacterium than
2786 from an actinobacterium, as was previously proposed[132]. The second class of type III polyketide
2787 synthase (PKS2) probably arose via a duplication of PKS1 just prior to the divergence of the brown algal
2788 lineage from the Phaeothamniophyceae (Figure 4A). *Undaria pinnatifida* is the only species where a
2789 duplication of PKS2 was observed. The PKS3 class appeared much more recently, within the
2790 Ectocarpales during the diversification of this brown algal order, again probably via a duplication of
2791 PKS1 (Figure 4A). PKS3 genes were only found in two families, the Ectocarpaceae and the
2792 Scytosiphonaceae. The enzymatic activities and functions of PKS2 and PKS3 remain to be determined
2793 but analyses of the functions of recombinant PKS1 proteins have indicated that different proteins may
2794 have different activities leading to the production of different metabolites[132–134]. For example, the
2795 *Ectocarpus* species 7 PKS1 enzyme produces phloroglucinol from malonyl-CoA alone or malonyl-CoA
2796 plus acetyl-CoA whereas the PKS1 enzymes of *Sargassum binderi* and *Sargassum fusiforme* generate
2797 4-hydroxy-6-methyl-2-pyrone. It is therefore interesting that the PKS1 family has expanded in some
2798 members of the Sargassaceae, with a maximum of five PKS1 genes detected in *S. fusiforme* (Figures 3A
2799 and 4A). These expansions may be associated with diversification of PKS1 biochemical function.
2800
2801 **Sulphatases**
2802
2803 Sulphatases remove sulphate groups from sulphated polysaccharides[81]. The brown algal enzymes
2804 appear to be of quite ancient origin, with homologues being present in other stramenopile lineages
2805 (Table S30).
2806
2807 **Carbon storage via the mannitol cycle**
2808

2809 In brown algae, the output of the Calvin cycle is used to produce the carbon storage molecule mannitol
2810 via the mannitol cycle[135]. The genes that encode the four enzymes of the mannitol cycle are not only
2811 conserved in all brown algae but were also found in taxa as distant as *H. akashiwo* (Table S14),
2812 indicating that the mannitol cycle was acquired before brown algae diverged from closely-related
2813 stramenopile taxa. The mannitol cycle genes include a linked pair of *M1PDH3* and *M1Pase1* genes
2814 which may therefore represent the ancestral core of this metabolic cycle. These pairs of genes are
2815 head-to-head and adjacent in brown algae with an intergenic region ranging from about 500 bp
2816 (Ectocarpales) to 10 kbp (*Fucus serratus*). The organisation of this gene pair differs in sister groups
2817 with, for example, a tandem arrangement and an intervening gene in *S. ischiensis* and two *M1PDH3* /
2818 *M1Pase1* gene pairs in *H. akashiwo* with tail-to-head and head-to-tail arrangements and long (150-200
2819 kbp) intergenic regions. Several species within the Ectocarpales and Laminariales have acquired an
2820 additional *M1PDH* gene perhaps allowing greater capacity or finer regulation of carbon storage
2821 through the mannitol cycle (Table S14).
2822
2823 **LHCX light-harvesting proteins**
2824
2825 In diatoms, LHCX proteins play an important role in the photoprotection of the photosynthetic
2826 apparatus under conditions of excess light[136]. Chromosome six of the *Ectocarpus* species 7 genome is
2827 known to contain a cluster of 14 LHCX genes interspersed with genes coding for proteins unrelated to
2828 photosynthesis[85,137]. Analysis of the Phaeoexplorer genomes indicated that LHCX genes were present
2829 in this genomic context before the divergence of the Desmarestiales but that the expansion of the
2830 gene cluster only began in a common ancestor of the Laminariales and Ectocarpales (Figure S12A). The
2831 general features of the cluster are conserved in the genomes of *Ectocarpus* species 7, *Ectocarpus*
2832 *crouaniorum*, *Ectocarpus fasciculatus*, *Scytosiphon promiscuus*, *Porterinema fluviatile* and *C. linearis*
2833 but in *C. linearis* the cluster appears to have undergone a partial rearrangement and contains only
2834 three LHCX copies whereas the clusters in all other species include at least seven LHCX genes. In *S.*
2835 *latissima*, the genes of the cluster were found in three short contigs that may be adjacent on the same
2836 chromosome. LHCX gene clusters are also present in *D. dichotoma* (4 genes on contig 3024, 3 on contig
2837 458 and 2 on contig 1592) of the early branching order Dictyotales and in *S. ischiensis* (4 plus 2 genes
2838 on contig 28 and 2 genes on contig 64), but the genomic context of these clusters differs from that of
2839 the cluster found in *Ectocarpus* spp., indicating independent expansions of the LHCX family in these
2840 taxa. LHCX sequences from *S. ischiensis*, *D. dichotoma* and the genus *Ectocarpus* grouped into
2841 independent, species/genus-specific clusters in a phylogenetic analysis (Figure S12B), further
2842 supporting independent expansions of the LHCX gene family in these taxa. The LHCX clusters probably
2843 evolved by gene duplication and ectopic gene conversion. The involvement of gene conversion is

2844 indicated by the higher intraspecific identity of LHCX sequences between paralogs than between
2845 putative orthologs (based on synteny) compared across species (Figures S12A, S12B). The large gene
2846 cluster that contains LHCX4 to LHCX7 commonly contains overlapping LHCX genes on opposite strands
2847 of the chromosome, with some or all of the exons of one gene being located within an intron of the
2848 other (Figure S12C). Changes in gene dosage in response to environmental selection (as observed, for
2849 example, in studies by[138,139]) may have been a driver of the LHCX gene family expansion.
2850
2851 **Transcription factors**
2852
2853 An analysis of transcription-associated proteins (TAPs) indicated that the range of gene families and
2854 the relative sizes of each gene family are relatively constant across the brown algae and their close
2855 sister species (Figures 3A, S21A, S21B and S21C, Table S15), suggesting that the emergence of the
2856 brown algae did not involve major changes to the transcription machinery. For example, large families
2857 of C2H2, Zn_clus, GNAT and SWI_SNF_SNF2 genes are a common feature in both brown algal and sister
2858 taxa genomes.
2859
2860 The number of TAP genes has remained relatively stable during the diversification of the brown algae
2861 (Figures 3A, S21B and S21C). This contrasts with what has been observed for the green lineage, where
2862 the size of TAP complement has been shown to be broadly correlated with morphological
2863 complexity[140]. Although the overall TAP complement has been stable, variations were detected in the
2864 sizes of some specific TAP families, including variations that correlated with the degree of
2865 morphological complexity. A broadly positive correlation was observed for the three-amino-acid
2866 length extension homeodomain (TALE HD) transcription factor (TF) families for example (Figure S11A,
2867 Table S15). These observations suggest that increased multicellular complexity in the brown algae may
2868 have required different combinations of TFs rather than simply increasing numbers. Two TALE HD TFs
2869 (OUROBOROS and SAMSARA) have been shown to be necessary for the deployment of the sporophyte
2870 developmental program in *Ectocarpus* species 7[141]. The *Ectocarpus* species 7 TALE HD TF family consists
2871 of three genes which are conserved across the brown algae[141] but we detected a modest expansion of
2872 the gene family in some members of the Fucales, which have acquired one or two additional
2873 paralogues of the third TALE HD TF (Figure S11B).
2874
2875 **Ion channels**
2876
2877 The Phaeoexplorer dataset allowed the origins of the major families of brown algal ion channels to be
2878 traced (Table S16). Most classes of brown algal ion channel are also found in other stramenopile

2879 groups, such as IP_3 receptors, which were earlier shown to mediate calcium signalling in *Fucus*[142], and
2880 Nav/Cav four-domain voltage-gated calcium channels, which are associated with rapid signalling
2881 processes and the regulation of flagella motion[143]. Pennate diatoms and angiosperms, which both lack
2882 flagellated gametes, do not possess Nav/Cav channels[144]. Land plants and diatoms also lack IP_3
2883 receptors. All stramenopile lineages, including brown algae, have OSCAs, which are mechanosensitive
2884 channels that have been implicated in calcium signalling during osmoregulation[145,146], and transient
2885 receptor potential (TRP) channels (although the latter were not detected in *H. akashiwo*). In brown
2886 algae the only ubiquitous ligand-gated ion channels are the glutamate receptors (GLR). P2X receptors
2887 are absent and the pentameric ligand-gated ion channel (pLGIC, related to acetylcholine and GABA
2888 receptors in animals) was only found in *D. mesarthrocarpum*. EukCat, a novel class of single domain
2889 voltage-gated calcium channel recently discovered in diatoms[147], appears to be absent from brown
2890 algae and other stramenopiles.
2891
2892 The Mitochondrial Calcium Uniporter (MCU) plays an important role in the uptake of calcium into
2893 mitochondria and is highly conserved across the eukaryotes. It is therefore surprising that the Fucales,
2894 like diatoms[148], have lost this transporter (Table S16) and it is unclear how Fucales and diatoms
2895 regulate mitochondrial calcium, as alternative mechanisms are not known. Cyclic nucleotide gated
2896 channels, which are only present at low copy number, also appear to have been lost from several
2897 brown algal lineages, including kelps. In contrast, TRP channels are found in all brown algal genomes
2898 and the gene family is massively expanded in some lineages (Table S16). In animals, TRP channels are
2899 primarily involved in sensory functions (e.g. touch, temperature) and ligand-based activation[149], so the
2900 brown algal channels may play a role in sensing the environment. TRP channels have been lost in land
2901 plants, which show an expansion of CNGCs and GLRs[150].
2902
2903 **Membrane-localised signalling proteins**
2904
2905 Brown algae have integrin-like molecules[151], and this feature was found to be shared with several sister
2906 stramenopile lineages including pelagophytes. A single integrin-like gene was found in close sister
2907 groups but this gene was duplicated early in brown algal evolution, with a further recent duplication
2908 in *Ectocarpus* species 7 and *U. pinnatifida* and gene loss in several other brown algae (Figures 3A and
2909 S9A, Table S31). Proteins with an FG-GAP domain (a component of integrins) were found in several
2910 other stramenopile lineages but these proteins did not have the characteristic integrin domain
2911 structure and their evolutionary relationship to brown algal integrin-like proteins is uncertain (Figure
2912 S9B).
2913

2914 Brown algae and closely-related sister groups also possess a single gene that is very similar to the land
2915 plant membrane-localised calpain domain protein DEK1 (Figures 3A and S9A, Table S32), which is
2916 thought to act as a mechanosensor in land plants by activating the RMA channel[152].
2917
2918 Fasciclin (FAS) domain proteins are found throughout the stramenopiles but only brown algae have
2919 membrane-localised proteins with four extracellular FAS domains analogous to the membrane-
2920 localised fasciclin proteins in animals that interact with the extracellular matrix (Figure S9A, Table
2921 S33)[153].
2922
2923 All stramenopiles, including brown algae, appear to have a single tetraspanin gene plus a gene that
2924 encodes a stramenopile-specific eight transmembrane domain tetraspanin-like protein (Figure S9A,
2925 Table S34).
2926
2927 **Flagellar proteins**
2928
2929 Most stramenopiles possess motile cells with two heteromorphic flagella, although there are some
2930 exceptions such as the male gametes of pennate diatoms. Searches for the presence of flagellar
2931 proteins, based on proteomic inventories of *Colpomenia bullosa* anterior and posterior flagella[102],
2932 indicated that highly-conserved proteins such as the mastigoneme proteins MAS1, MAS2 and MAS3, a
2933 flagellar titin protein, which is thought to be involved in connecting mastigonemes to the flagellar
2934 axoneme, and flagellar creatine kinase are present in all brown algae and sister taxa (Figures 3A and
2935 4B, Table S13). In contrast, the HELMCHROME protein, which is thought to be involved in light
2936 reception and zoid phototaxis[102], was only found in brown algae and some closely-related sister taxa,
2937 the most distant being *C. australica* (Figure 3A).
2938
2939 **Histone proteins**
2940
2941 The emergence of the brown algae also appears to have been associated with genomic events that
2942 may have had an important impact on the regulation of chromatin function. Most brown algal
2943 genomes code for about 20 histone proteins but the number of histone genes was correlated with
2944 genome size. For example, *U. pinnatifida* and *A. nodosum*, which have quite large genomes (634 and
2945 1,253 Mbp, respectively) have 36 and 173 histone genes, respectively (Table S23). As in animal
2946 genomes[154], the majority of brown algal histone genes occur in clusters. For example, *Chordaria*
2947 *linearis* and *U. pinnatifida* both have three contigs containing 80-90% of their histone genes, whereas
2948 *Saccharina latissima* and *S. ischiensis* each have a single contig which contains ~40% of their histone

2949 genes. As in other eukaryotic lineages, histone H4 is highly conserved due to its structural role in
2950 binding H3, H2A and H2B, and no variants were found in brown algae. All Phaeophyceae possess H3.3
2951 and CenH3 but additional H3 variants were only found in the Discosporangiales and Dictyotales and in
2952 sister taxa as distant as the raphidophytes (Figures 3A and S22A). The centromeric variant CenH3 is
2953 highly divergent across the brown algae, particularly in the N-terminal tail (Figure S22B), as reported
2954 in other lineages[155]. In contrast to what has been observed in humans and land plants[156,157], H3.3 is
2955 encoded by a single gene in brown algae. H3.1 is encoded by several genes as in the majority of
2956 organisms. The brown algal H3.1 and H3.3 clades are distinct from those of animals and land plants
2957 indicating independent amplifications (Figure S22C). Linker H1 variants are highly divergent across the
2958 eukaryotes and this is also true for brown algae, with *A. nodosum,* for example, possessing 11 different
2959 H1 proteins (Figure S22D, Table S23). Brown algae with a genome of less than 1 Gbp possess ~10 H2B-
2960 coding genes but *F. serratus*, *A. nodosum* and the sister taxon *H. akashiwo* have at least three times
2961 more. Finally, brown algae not only possess H2A variants such as the H2A.X but also three novel H2A
2962 variants that only exist in Phaeophyceae and the sister taxon Schizocladiophyceae (Rousselot et al,
2963 unpublished).
2964
2965 **Histone post-translational modification**
2966
2967 An analysis of eight histone post-translational modifications in *Ectocarpus* species 7 indicated that the
2968 majority had similar genomic and genic distributions to those observed in animals and land plants,
2969 indicating that they have similar functional roles[119]. However, dimethylation of lysine 79 of histone H3
2970 (H3K79me2) has a very different distribution pattern in *Ectocarpus* compared to animals. In animals,
2971 H3K79me2 is associated with the 5' ends of transcribed genes[158] whereas in *Ectocarpus* it occurs as
2972 extended blocks, which can span several genes and appears to be associated, at least indirectly, with
2973 down-regulation of gene expression[119]. The H3K79me2 mark is written by DOT1-class DNA methylases,
2974 which are encoded by multiple genes in brown algae. Dollo analysis identified gain of a new DOT1
2975 orthogroup (OG0008134) in the common ancestor of the Phaeophyceae and the two close sister taxa
2976 *S. ischiensis* and *Phaeothamnion wetherbeei*, indicating a possible correlation between this unusual
2977 H3K79me2 distribution pattern and the appearance of a novel DOT1 protein.
2978
2979 **Genome changes associated with adaptation to freshwater environments**
2980
2981 A small number of brown algal species have adapted to freshwater habitats, with each adaptation
2982 event occurring independently. The Phaeoexplorer dataset includes genomes for four freshwater
2983 species in addition to the previously sequenced *Ectocarpus subulatus*[159]. Adaptation to a freshwater

2984 habitat was correlated with loss of orthogroups (Figure S4A) and clear reductions in the sizes of some
2985 key gene families such as haloperoxidases (Figures 3A and S8C). In terms of gene gain, 12 orthogroups
2986 were found to be consistently over-represented when the genomes of each of the five freshwater
2987 species were compared with those of related seawater species, including orthogroups with genes
2988 involved in membrane transport and signalling. A transcriptomic analysis of *P. fluviatile* grown in either
2989 seawater or freshwater conditions identified 1,328 differentially-expressed genes (DEGs), with the
2990 genes upregulated in freshwater being enriched in DNA replication regulation and transcription
2991 functions (Table S24). Comparison of the differentially regulated gene set with a similar set generated
2992 for *E. subulatus* indicated that both sets of DEGs were enriched in lineage-specific genes, suggesting
2993 that, to a large extent, each adaptation to freshwater involved lineage-specific innovations (Table S35).
2994 Furthermore, a focused comparative analysis of metabolic networks (https://gem-
2995 aureme.genouest.org/fwgem/; Figure S7G) indicated that both *P. fluviatile* and *Pleurocladia lacustris*
2996 have lost alginate lyases and a glycosyltransferase GT18, related to polysaccharide metabolism (Table
2997 S22). This is in line with previous work showing specific cell wall remodelling in *E. subulatus*[160]. *P.*
2998 *fluviatile* and *P. lacustris* have also lost genes encoding ubiquitin-conjugating enzymes and this may
2999 have impacted protein degradation and thus osmolarity, or be linked to the preferential activation of
3000 ubiquitin-independent regulatory mechanisms in situations of osmotic stress[145]. In addition to the
3001 shared gene losses, both species have independently lost genes. The strictly freshwater *P. lacustris* has
3002 lost several phosphatases, peptidases and aldolases, perhaps due to a reduced need to produce
3003 osmolytes, while the brackish water *P. fluviatile* has lost several enzymes related to phenol metabolism
3004 and a magnesium transporter. More surprisingly, the two species have lost different sets of enzymes
3005 involved in RNA modifications, which may either indicate convergent acquisition of different
3006 regulatory mechanisms, or may simply be markers of phylogenetic divergence.
3007
3008 **Comparative analysis of *Ectocarpus* genomes**
3009
3010 The genus *Ectocarpus* includes 16 cryptic species[161]. Twenty-two new strains of *Ectocarpus*,
3011 representing 12 new species, were sequenced in this study. Comparison of high continuity *Ectocarpus*
3012 assemblies indicated that *Ectocarpus* spp. genomes share extensive synteny (Figure S23A, Table S36).
3013
3014 The genus *Ectocarpus* belongs to the order Ectocarpales, which emerged late in brown algal evolution
3015 but has become the most species-rich brown algal order, with over 750 species[162]. The Ectocarpales
3016 exhibited the same general phenomenon of orthogroup loss as the other Phaeophyceae orders during
3017 the diversification of the brown algae. However, this trend was not uniform within the Ectocarpales
3018 and a period of orthogroup gain is predicted to have occurred during the early emergence of the genus

3019 *Ectocarpus*, analogous to the burst of gene gain observed during the early emergence of the brown
3020 algal lineage (Figure S23B). The sets of both gained and conserved orthogroups were significantly
3021 enriched in transcription-related functions (Figure S23C). An analysis of the *Ectocarpus* species 7
3022 reference genome orthologues indicated that rapidly-evolving, taxonomically-restricted genes (Group
3023 5 based on the phylostratigraphy analysis: Table S2) are most abundant on the sex chromosome
3024 (Figure S23D), an observation that is consistent with other recent findings[163]. These *Ectocarpus* genus-
3025 specific genes exhibit higher dN and dS values than genes that the genus shares with other brown
3026 algae (Figure S23E, Table S18).
3027
3028 **Hormone biosynthesis pathways**
3029
3030 Brown algal genomes contain homologues of components of several phytohormone biosynthetic
3031 pathways, including indole-3-acetic acid (the auxin IAA), cytokinins, brassinosteroids, ethylene and
3032 gibberellins (Figure S24, Table S37), with the exception of tryptophan aminotransferase (TAA; auxin
3033 biosynthesis) and copalyl diphosphate/ent-kaurene synthase and ent-kaurene oxidase (CPS/KO;
3034 gibberellin biosynthesis). The presence of these genes is consistent with the identification of the
3035 corresponding hormones[164,165] and of brassinosteroid-like molecules[166] in brown algae. As far as
3036 abscisic acid (ABA) is concerned, brown algae have the capacity to synthesise oxidised cleaved
3037 carotenoids, supporting the idea that they may generate ABA-like signalling molecules[167]. Similarly,
3038 OPDA, an intermediate of the jasmonic acid pathway, has been identified in brown algae, raising the
3039 possibility that this molecule itself may have a signalling role similar to OPDA-like molecules in
3040 hornworts[168]. The putative brown algal phytohormone biosynthetic genes are also present in the
3041 genomes of sister taxa and therefore appear to have been acquired well before the emergence of the
3042 brown algae. However, an expansion of the flavin monooxygenase (YUC) gene family (which mediates
3043 a rate limiting step in IAA production in land plants) appears to have occurred in the common ancestor
3044 of the Phaeophyceae, *S. ischiensis, C. australica* and *T. minus* (Figure S24). Several gene families that
3045 are putatively associated with the biosynthesis of phytohormones such as ABA, brassinolide, ethylene,
3046 gibberellins, jasmonic acid and strigolactones have expanded in the Fucales (Figure S24).
3047
3048
3049 **References**
3050

3051 1. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for
3052 large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

3053 2. Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome
3054 shotgun sequences. Nucleic Acids Research *34*, 5623–5630. https://doi.org/10.1093/nar/gkl723.

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

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

5. 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.

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

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

8. 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.

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

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

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

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

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

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

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

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

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

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

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

3096 20. Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). rnaSPAdes: a *de novo* transcriptome
3097 assembler and its application to RNA-Seq data. GigaScience *8*, giz100.
3098 https://doi.org/10.1093/gigascience/giz100.

3099 21. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker. http://repeatmasker.org.

3100 22. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*, 573–
3101 580.

3102 23. Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification
3103 in *de novo* annotation approaches. PLoS One *6*, e16526. https://doi.org/10.1371/journal.pone.0016526.

3104 24. Kent, W. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656–664.

3105 25. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995.
3106 https://doi.org/10.1101/gr.1865504.

3107 26. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using
3108 DIAMOND. Nat Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

3109 27. Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput
3110 Appl Biosci *13*, 477–478.

3111 28. Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C.D., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M.
3112 (2016). Gmove a tool for eukaryotic gene predictions using various evidences. F1000Research *5*.
3113 https://doi.org/10.7490/f1000research.1111735.1.

3114 29. Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets.
3115 Bioinformatics *30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531.

3116 30. Stamatakis, A. (2015). Using RAxML to Infer Phylogenies. Curr Protoc Bioinformatics *51*, 6.14.1-14.
3117 https://doi.org/10.1002/0471250953.bi0614s51.

3118 31. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in
3119 Bayesian Phylogenetics Using Tracer 1.7. Syst Biol *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.

3120 32. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics.
3121 Genome Biol *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

3122 33. Csűös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood.
3123 Bioinformatics *26*, 1910–1912. https://doi.org/10.1093/bioinformatics/btq315.

3124 34. Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic
3125 Acids Res *32*, 1792–1797.

3126 35. Jehl, P., Sievers, F., and Higgins, D.G. (2015). OD-seq: outlier detection in multiple sequence alignments. BMC
3127 Bioinformatics *16*, 269. https://doi.org/10.1186/s12859-015-0702-1.

3128 36. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3
3129 and convergent evolution of coiled-coil regions. Nucleic Acids Res *41*, e121.
3130 https://doi.org/10.1093/nar/gkt263.

3131 37. Barrera-Redondo, J., Lotharukpong, J.S., Drost, H.-G., and Coelho, S.M. (2023). Uncovering gene-family
3132 founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol
3133 *24*, 54. https://doi.org/10.1186/s13059-023-02895-z.

38. 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.

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

40. 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.

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

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

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. Petroll, R., Varshney, D., Hiltemann, S., Finke, H., Schreiber, M., Vries, J. de, 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 https://doi.org/10.1101/2024.07.13.602682.

52. 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.

3177 53. Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple
3178 sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
3179 Nucleic Acids Res *22*, 4673–4680.

3180 54. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic
3181 analysis. Mol Biol Evol *17*, 540–552.

3182 55. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification.
3183 Nat Biotechnol *34*, 525–527. https://doi.org/10.1038/nbt.3519.

3184 56. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-
3185 seq data with DESeq2. Genome Biol *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

3186 57. Andrews, S. (2016). FastQC A Quality Control tool for High Throughput Sequence Data.
3187 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

3188 58. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning
3189 sequence reads to genomic features. Bioinformatics *30*, 923–930.
3190 https://doi.org/10.1093/bioinformatics/btt656.

3191 59. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586–1591.
3192 https://doi.org/10.1093/molbev/msm088.

3193 60. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods
3194 in Ecology and Evolution *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

3195 61. Wallau, G.L., Capy, P., Loreto, E., Le Rouzic, A., and Hua-Van, A. (2016). VHICA, a New Method to Discriminate
3196 between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within *Drosophila*.
3197 Mol Biol Evol *33*, 1094–1109. https://doi.org/10.1093/molbev/msv341.

3198 62. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes
3199 from whole genome data. Nucleic Acids Res *45*, e18. https://doi.org/10.1093/nar/gkw955.

3200 63. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and
3201 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools.
3202 Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

3203 64. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq -
3204 versatile and accurate annotation of organelle genomes. Nucleic Acids Res *45*, W6–W11.
3205 https://doi.org/10.1093/nar/gkx391.

3206 65. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in
3207 nucleotide sequences. Nucleic Acids Res *32*, 11–16. https://doi.org/10.1093/nar/gkh152.

3208 66. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast
3209 model selection for accurate phylogenetic estimates. Nat Methods *14*, 587–589.
3210 https://doi.org/10.1038/nmeth.4285.

3211 67. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the
3212 Ultrafast Bootstrap Approximation. Mol Biol Evol *35*, 518–522. https://doi.org/10.1093/molbev/msx281.

3213 68. Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., and Lyons, E. (2017). SynMap2 and SynMap3D:
3214 web-based whole-genome synteny browsers. Bioinformatics *33*, 2197–2198.
3215 https://doi.org/10.1093/bioinformatics/btx144.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

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

75. 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.

76. 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.

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

78. 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.

79. 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.

80. 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.

81. 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.

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

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

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

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

86. 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.

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

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

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

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

91. 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 U S A *121*, e2317054121. https://doi.org/10.1073/pnas.2317054121.

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

93. 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.

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

95. 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.

96. 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.

97. 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.

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

3300 99. Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O.,
3301 Guillaudeux, N., et al. (2018). Traceability, reproducibility and wiki-exploration for "à-la-carte"
3302 reconstructions of genome-scale metabolic models. PLoS Comput. Biol. *14*, e1006146.
3303 https://doi.org/10.1371/journal.pcbi.1006146.

3304 100. Macaisne, N., Liu, F., Scornet, D., Peters, A.F., Lipinska, A., Perrineau, M.-M., Henry, A., Strittmatter, M.,
3305 Coelho, S.M., and Cock, J.M. (2017). The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel
3306 family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development *144*, 409–
3307 418. https://doi.org/10.1242/dev.141523.

3308 101. Talbert, P.B., Ahmad, K., Almouzni, G., Ausió, J., Berger, F., Bhalla, P.L., Bonner, W.M., Cande, W.Z.,
3309 Chadwick, B.P., Chan, S.W.L., et al. (2012). A unified phylogeny-based nomenclature for histone variants.
3310 Epigenetics Chromatin *5*, 7. https://doi.org/10.1186/1756-8935-5-7.

3311 102. Fu, G., Nagasato, C., Oka, S., Cock, J.M., and Motomura, T. (2014). Proteomics Analysis of Heterogeneous
3312 Flagella in Brown Algae (Stramenopiles). Protist *165*, 662–675. https://doi.org/10.1016/j.protis.2014.07.007.

3313 103. Starr, R.C., and Zeikus, J.A. (1993). UTEX-The culture collection of algae at the University of Texas at
3314 Austin 1993 list of cultures. J Phycol *29 (Suppl.)*, 1–106.

3315 104. Dittami, S.M., Gravot, A., Goulitquer, S., Rousvoal, S., Peters, A.F., Bouchereau, A., Boyen, C., and Tonon,
3316 T. (2012). Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to
3317 low salinities in *Ectocarpus* (brown algae). Plant J. https://doi.org/10.1111/j.1365-313X.2012.04982.x.

3318 105. Rahman, S., Kosakovsky Pond, S.L., Webb, A., and Hey, J. (2021). Weak selection on synonymous codons
3319 substantially inflates dN/dS estimates in bacteria. Proc Natl Acad Sci U S A *118*, e2023575118.
3320 https://doi.org/10.1073/pnas.2023575118.

3321 106. Duchêne, S., Holmes, E.C., and Ho, S.Y.W. (2014). Analyses of evolutionary dynamics in viruses are
3322 hindered by a time-dependent bias in rate estimates. Proc Biol Sci *281*, 20140732.
3323 https://doi.org/10.1098/rspb.2014.0732.

3324 107. Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent"
3325 sites in *Drosophila* DNA. Genetics *139*, 1067–1076. https://doi.org/10.1093/genetics/139.2.1067.

3326 108. Akashi, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift.
3327 Gene *205*, 269–278. https://doi.org/10.1016/s0378-1119(97)00400-9.

3328 109. Subramanian, S. (2008). Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic
3329 Genomes. Genetics *178*, 2429–2432. https://doi.org/10.1534/genetics.107.086405.

3330 110. Wright, F. (1990). The 'effective number of codons' used in a gene. Gene *87*, 23–29.
3331 https://doi.org/10.1016/0378-1119(90)90491-9.

3332 111. Forcelloni, S., and Giansanti, A. (2020). Evolutionary Forces and Codon Bias in Different Flavors of
3333 Intrinsic Disorder in the Human Proteome. J Mol Evol *88*, 164–178. https://doi.org/10.1007/s00239-019-
3334 09921-4.

3335 112. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic
3336 sequence comparison. Genome Res. *21*, 487–493. https://doi.org/10.1101/gr.113985.110.

3337 113. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov,
3338 A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids
3339 Research *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.

114. 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.

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

116. 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.

117. 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.

118. 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.

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

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

121. 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.

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

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

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

125. 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.

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

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

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

3382 129. Ran, J.-H., Shen, T.-T., Liu, W.-J., Wang, P.-P., and Wang, X.-Q. (2015). Mitochondrial introgression and
3383 complex biogeographic history of the genus *Picea*. Mol Phylogenet Evol *93*, 63–76.
3384 https://doi.org/10.1016/j.ympev.2015.07.020.

3385 130. Sarver, B.A.J., Herrera, N.D., Sneddon, D., Hunter, S.S., Settles, M.L., Kronenberg, Z., Demboski, J.R.,
3386 Good, J.M., and Sullivan, J. (2021). Diversification, Introgression, and Rampant Cytonuclear Discordance in
3387 Rocky Mountains Chipmunks (Sciuridae: *Tamias*). Syst Biol *70*, 908–921.
3388 https://doi.org/10.1093/sysbio/syaa085.

3389 131. Dittami, S.M., Barbeyron, T., Boyen, C., Cambefort, J., Collet, G., Delage, L., Gobet, A., Groisillier, A.,
3390 Leblanc, C., Michel, G., et al. (2014). Genome and metabolic network of "Candidatus Phaeomarinobacter
3391 ectocarpi" Ec32, a new candidate genus of Alphaproteobacteria frequently associated with brown algae.
3392 Front Genet *5*, 241. https://doi.org/10.3389/fgene.2014.00241.

3393 132. Meslet-Cladière, L., Delage, L., Leroux, C.J., Goulitquer, S., Leblanc, C., Creis, E., Gall, E.A., Stiger-
3394 Pouvreau, V., Czjzek, M., and Potin, P. (2013). Structure/Function Analysis of a Type III Polyketide Synthase in
3395 the Brown Alga *Ectocarpus siliculosus* Reveals a Biochemical Pathway in Phlorotannin Monomer Biosynthesis.
3396 Plant Cell *25*, 3089–3103. https://doi.org/10.1105/tpc.113.111336.

3397 133. Baharum, H., Morita, H., Tomitsuka, A., Lee, F.C., Ng, K.Y., Rahim, R.A., Abe, I., and Ho, C.L. (2011).
3398 Molecular cloning, modeling, and site-directed mutagenesis of type III polyketide synthase from *Sargassum*
3399 *binderi* (Phaeophyta). Mar Biotechnol (NY) *13*, 845–856. https://doi.org/10.1007/s10126-010-9344-5.

3400 134. 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.
3401 (2021). A Type III Polyketide Synthase (SfuPKS1) Isolated from the Edible Seaweed Sargassum fusiforme
3402 Exhibits Broad Substrate and Catalysis Specificity. J Agric Food Chem *69*, 14643–14649.
3403 https://doi.org/10.1021/acs.jafc.1c05868.

3404 135. Michel, G., Tonon, T., Scornet, D., Cock, J.M., and Kloareg, B. (2010). Central and storage carbon
3405 metabolism of the brown alga *Ectocarpus siliculosus*: insights into the origin and evolution of storage
3406 carbohydrates in Eukaryotes. New Phytol *188*, 67–81. https://doi.org/10.1111/j.1469-8137.2010.03345.x.

3407 136. Buck, J.M., Sherman, J., Bártulos, C.R., Serif, M., Halder, M., Henkel, J., Falciatore, A., Lavaud, J.,
3408 Gorbunov, M.Y., Kroth, P.G., et al. (2019). Lhcx proteins provide photoprotection via thermal dissipation of
3409 absorbed light in the diatom *Phaeodactylum tricornutum*. Nat Commun *10*, 4167.
3410 https://doi.org/10.1038/s41467-019-12043-6.

3411 137. Dittami, S.M., Michel, G., Collen, J., Boyen, C., and Tonon, T. (2010). Chlorophyll-binding proteins
3412 revisited - a multigenic family of light-harvesting and stress proteins from a brown algal perspective. BMC
3413 Evol Biol *10*, 365. https://doi.org/10.1186/1471-2148-10-365.

3414 138. Hess, K., Oliverio, R., Nguyen, P., Le, D., Ellis, J., Kdeiss, B., Ord, S., Chalkia, D., and Nikolaidis, N. (2018).
3415 Concurrent action of purifying selection and gene conversion results in extreme conservation of the major
3416 stress-inducible Hsp70 genes in mammals. Sci Rep *8*, 5082. https://doi.org/10.1038/s41598-018-23508-x.

3417 139. Hanikenne, M., Kroymann, J., Trampczynska, A., Bernal, M., Motte, P., Clemens, S., and Krämer, U.
3418 (2013). Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental
3419 adaptation. PLoS Genet *9*, e1003707. https://doi.org/10.1371/journal.pgen.1003707.

3420 140. Lang, D., and Rensing, S.A. (2015). The Evolution of Transcriptional Regulation in the Viridiplantae and
3421 its Correlation with Morphological Complexity. In Evolutionary Transitions to Multicellular Life: Principles and
3422 mechanisms Advances in Marine Genomics., I. Ruiz-Trillo and A. M. Nedelcu, eds. (Springer Netherlands), pp.
3423 301–333. https://doi.org/10.1007/978-94-017-9642-2_15.

3424 141. Arun, A., Coelho, S.M., Peters, A.F., Bourdareau, S., Pérès, L., Scornet, D., Strittmatter, M., Lipinska, A.P.,
3425 Yao, H., Godfroy, O., et al. (2019). Convergent recruitment of TALE homeodomain life cycle regulators to direct

3426 sporophyte development in land plants and brown algae. Elife *8*, e43101.
3427 https://doi.org/10.7554/eLife.43101.

3428 142. Roberts, S., and Brownlee, C. (1995). Calcium influx, fertilisation potential and egg activation in *Fucus*
3429 *serratus*. Zygote *3*, 191–197.

3430 143. Fujiu, K., Nakayama, Y., Yanagisawa, A., Sokabe, M., and Yoshimura, K. (2009). *Chlamydomonas CAV2*
3431 encodes a voltage- dependent calcium channel required for the flagellar waveform conversion. Curr Biol *19*,
3432 133–139. https://doi.org/10.1016/j.cub.2008.11.068.

3433 144. Verret, F., Wheeler, G., Taylor, A.R., Farnham, G., and Brownlee, C. (2010). Calcium channels in
3434 photosynthetic eukaryotes: implications for evolution of calcium-based signalling. New Phytol *187*, 23–43.
3435 https://doi.org/10.1111/j.1469-8137.2010.03271.x.

3436 145. Yuan, F., Yang, H., Xue, Y., Kong, D., Ye, R., Li, C., Zhang, J., Theprungsirikul, L., Shrift, T., Krichilsky, B., et
3437 al. (2014). OSCA1 mediates osmotic-stress-evoked Ca2+ increases vital for osmosensing in *Arabidopsis*.
3438 Nature *514*, 367–371. https://doi.org/10.1038/nature13593.

3439 146. Murthy, S.E., Dubin, A.E., Whitwam, T., Jojoa-Cruz, S., Cahalan, S.M., Mousavi, S.A.R., Ward, A.B., and
3440 Patapoutian, A. (2018). OSCA/TMEM63 are an Evolutionarily Conserved Family of Mechanically Activated Ion
3441 Channels. Elife *7*, e41844. https://doi.org/10.7554/eLife.41844.

3442 147. Helliwell, K.E., Chrachri, A., Koester, J.A., Wharam, S., Verret, F., Taylor, A.R., Wheeler, G.L., and
3443 Brownlee, C. (2019). Alternative Mechanisms for Fast Na+/Ca2+ Signaling in Eukaryotes via a Novel Class of
3444 Single-Domain Voltage-Gated Channels. Curr Biol *29*, 1503-1511.e6.
3445 https://doi.org/10.1016/j.cub.2019.03.041.

3446 148. Pittis, A.A., Goh, V., Cebrian-Serrano, A., Wettmarshausen, J., Perocchi, F., and Gabaldón, T. (2020).
3447 Discovery of EMRE in fungi resolves the true evolutionary history of the mitochondrial calcium uniporter. Nat
3448 Commun *11*, 4031. https://doi.org/10.1038/s41467-020-17705-4.

3449 149. Ramsey, I.S., Delling, M., and Clapham, D.E. (2006). An introduction to TRP channels. Annu Rev Physiol
3450 *68*, 619–647. https://doi.org/10.1146/annurev.physiol.68.040204.100431.

3451 150. Wheeler, G.L., and Brownlee, C. (2008). Ca2+ signalling in plants and green algae--changing channels.
3452 Trends Plant Sci *13*, 506–514. https://doi.org/10.1016/j.tplants.2008.06.004.

3453 151. Kloareg, B., Badis, Y., Cock, J.M., and Michel, G. (2021). Role and Evolution of the Extracellular Matrix in
3454 the Acquisition of Complex Multicellularity in Eukaryotes: A Macroalgal Perspective. Genes *12*, 1059.
3455 https://doi.org/10.3390/genes12071059.

3456 152. Tran, D., Galletti, R., Neumann, E.D., Dubois, A., Sharif-Naeini, R., Geitmann, A., Frachisse, J.-M., Hamant,
3457 O., and Ingram, G.C. (2017). A mechanosensitive Ca2+ channel activity is dependent on the developmental
3458 regulator DEK1. Nat Commun *8*, 1009. https://doi.org/10.1038/s41467-017-00878-w.

3459 153. Seifert, G.J. (2018). Fascinating Fasciclins: A Surprisingly Widespread Family of Proteins that Mediate
3460 Interactions between the Cell Exterior and the Cell Surface. International Journal of Molecular Sciences *19*,
3461 1628. https://doi.org/10.3390/ijms19061628.

3462 154. Marzluff, W., and Duronio, R. (2002). Histone mRNA expression: multiple levels of cell cycle regulation
3463 and important developmental consequences. Curr Opin Cell Biol *14*, 692–699.

3464 155. Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly
3465 evolving DNA. Science *293*, 1098–1102. https://doi.org/10.1126/science.1062939.

3466 156. Shi, L., Wen, H., and Shi, X. (2017). The Histone Variant H3.3 in Transcriptional Regulation and Human
3467 Disease. J Mol Biol *429*, 1934–1945. https://doi.org/10.1016/j.jmb.2016.11.019.

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

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

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

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

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

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

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

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

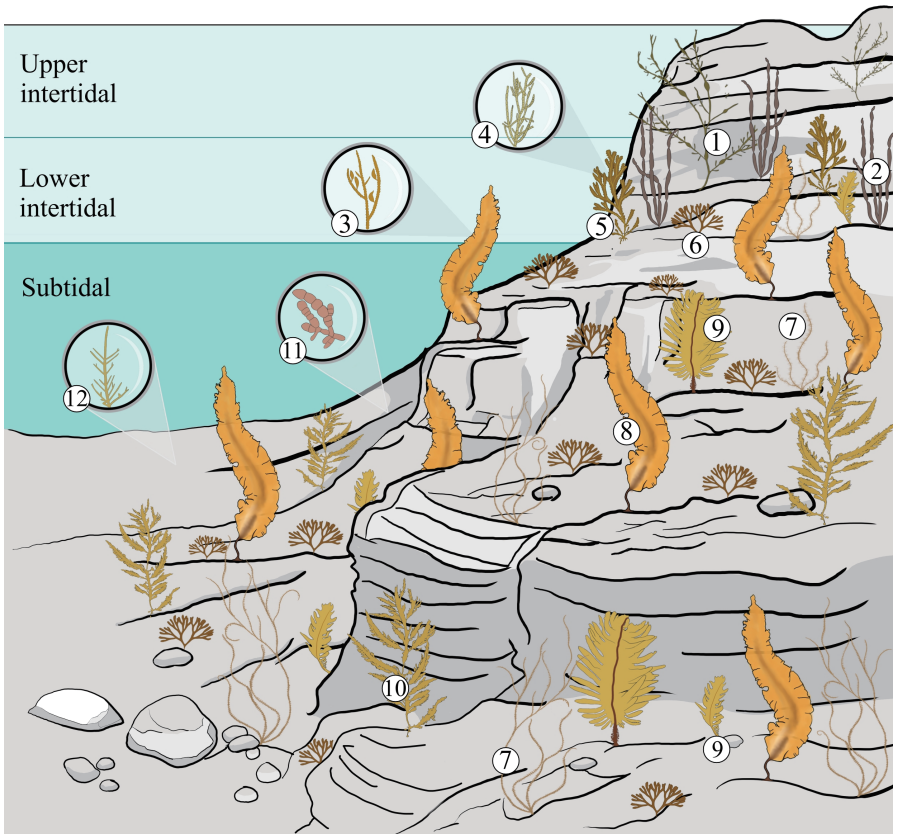
165. Lu, Y., and Xu, J. (2015). Phytohormones in microalgae: a new opportunity for microalgal biotechnology? Trends Plant Sci *20*, 273–282. https://doi.org/10.1016/j.tplants.2015.01.006.

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

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

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

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



Major events during evolution

Gain

- **A** Alginate-based ECM
- **B** Plasmodesmata
- **C** Basal attachment system; Parenchymatous growth
- **D** Heteromorphic life cycles (Stipe/lamina) intercalary meristem
- **E** Desiccation tolerance; Diploid life cycle

Loss

- **F** Posterior flagellum
- **G** Eyespot

Morphology & complexity

- unicellular
- simple multicellularity
- filamentous
- simple thallus
- complex thallus

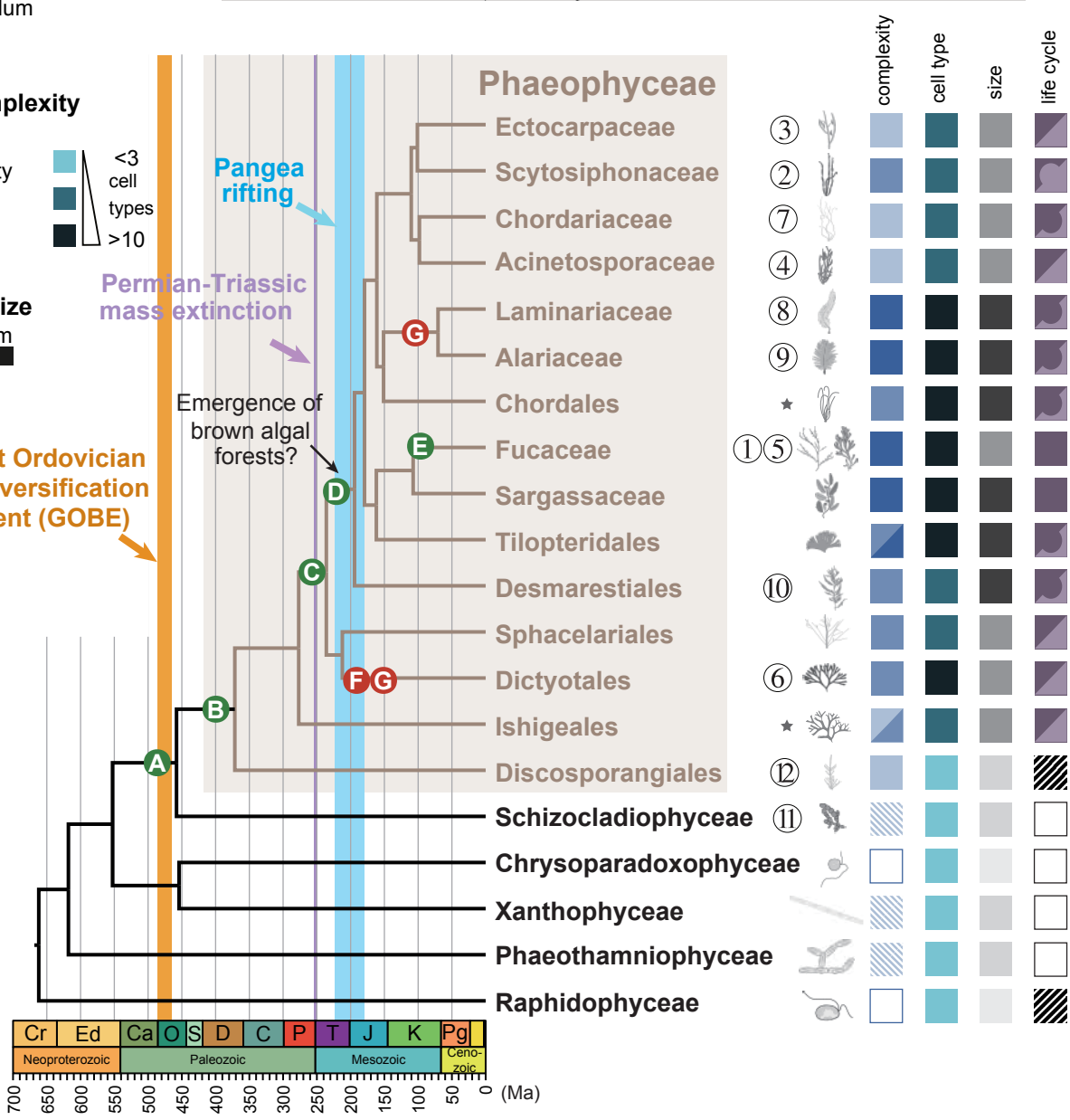
<3 cell types
>10

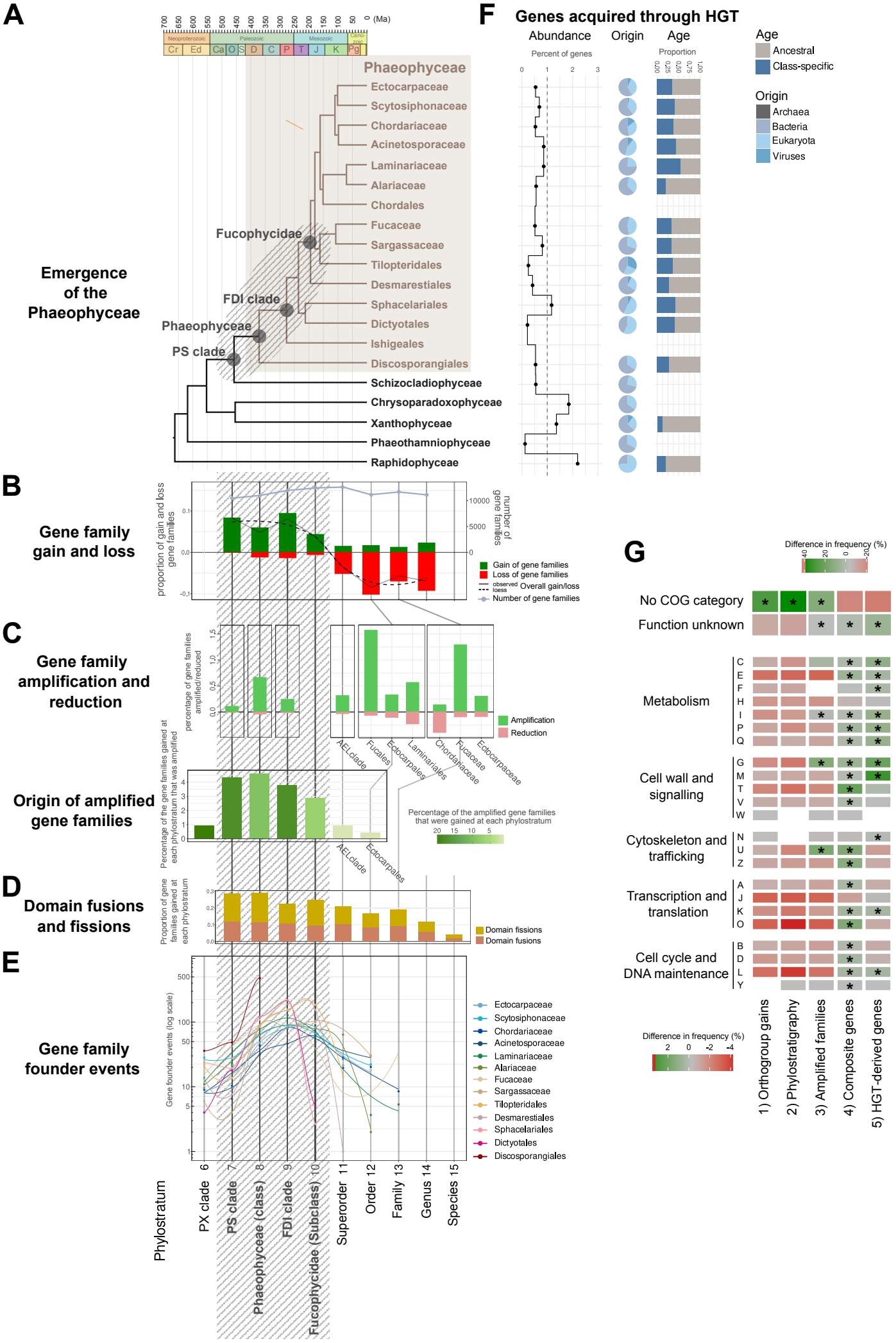
Maximum thallus size

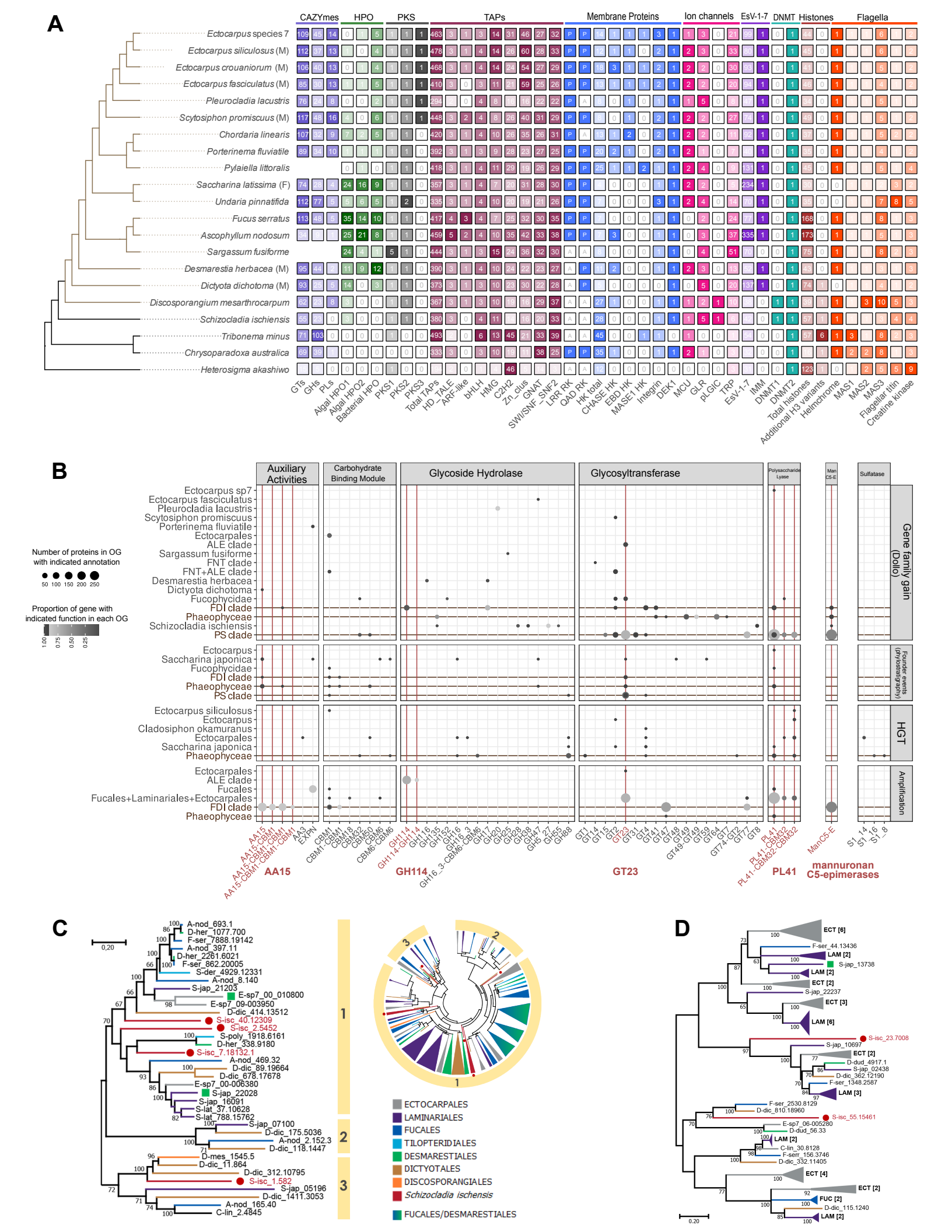
µm mm cm m

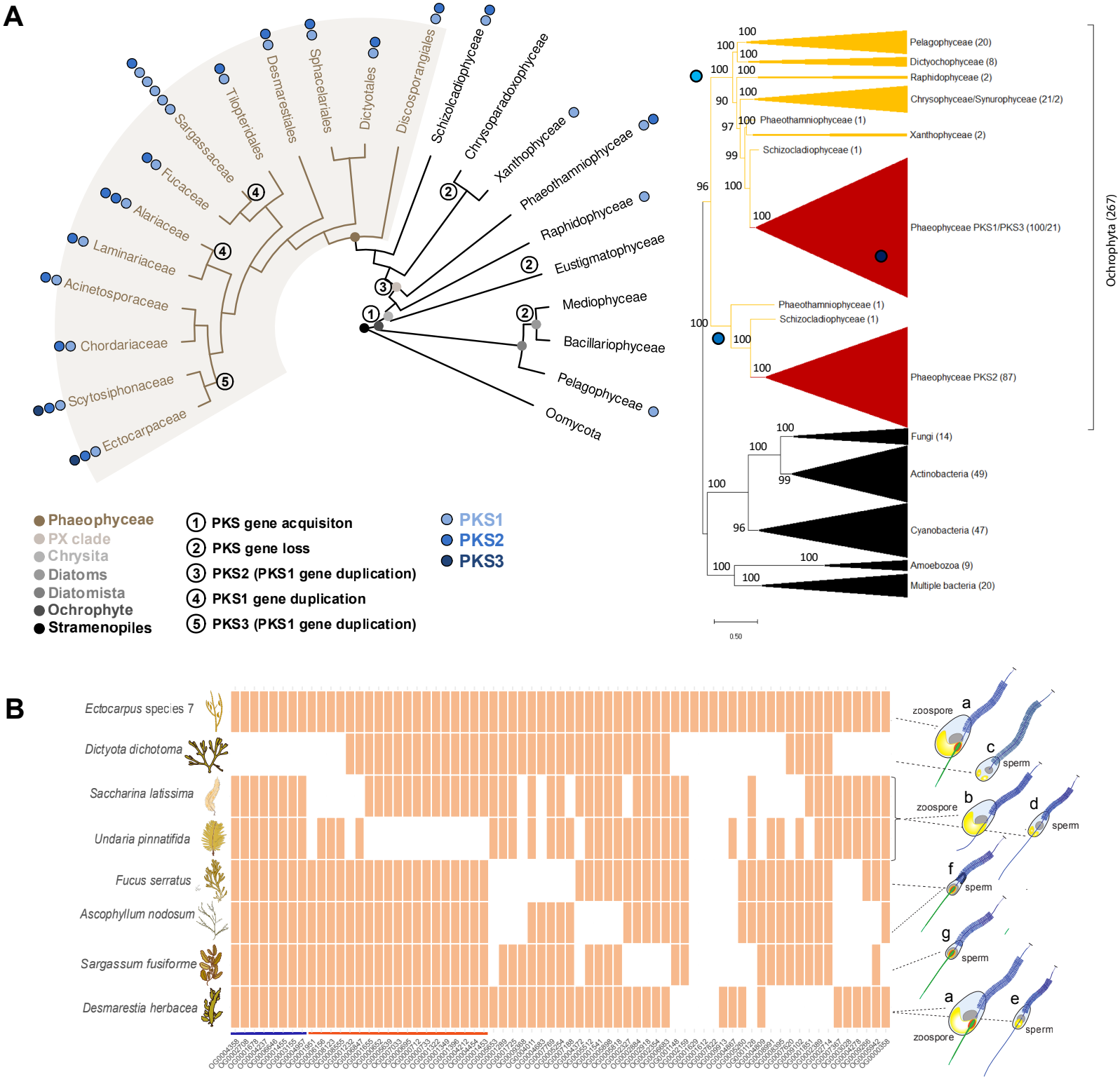
Typical life cycle

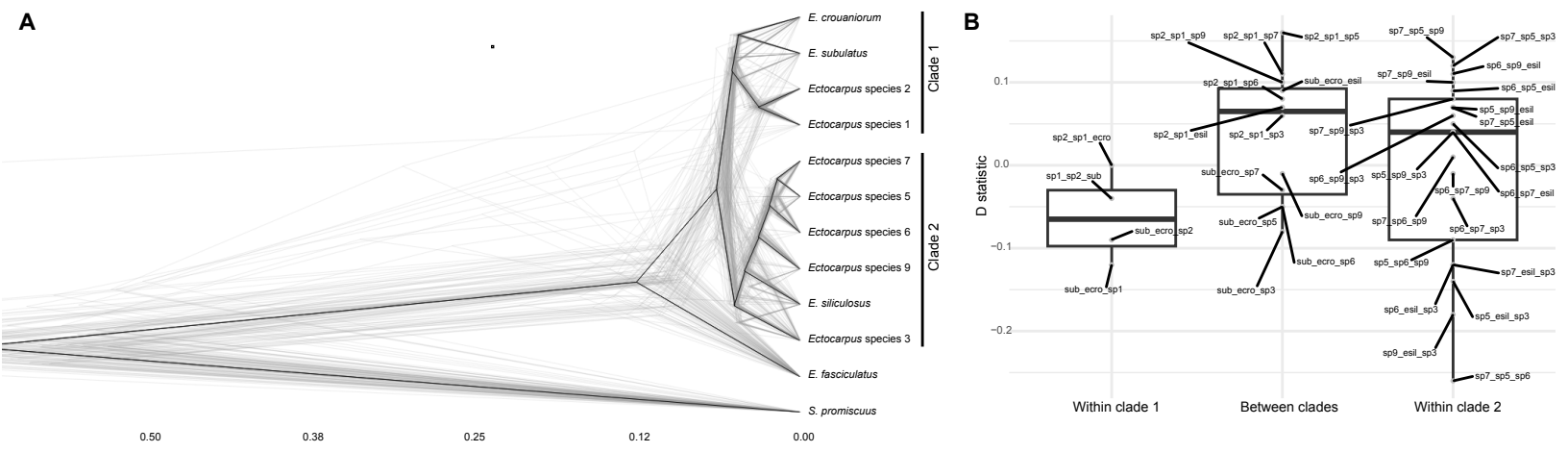
- diploid
- n<2n
- n=2n
- n>2n
- haploid
- probably diplontic
- unknown

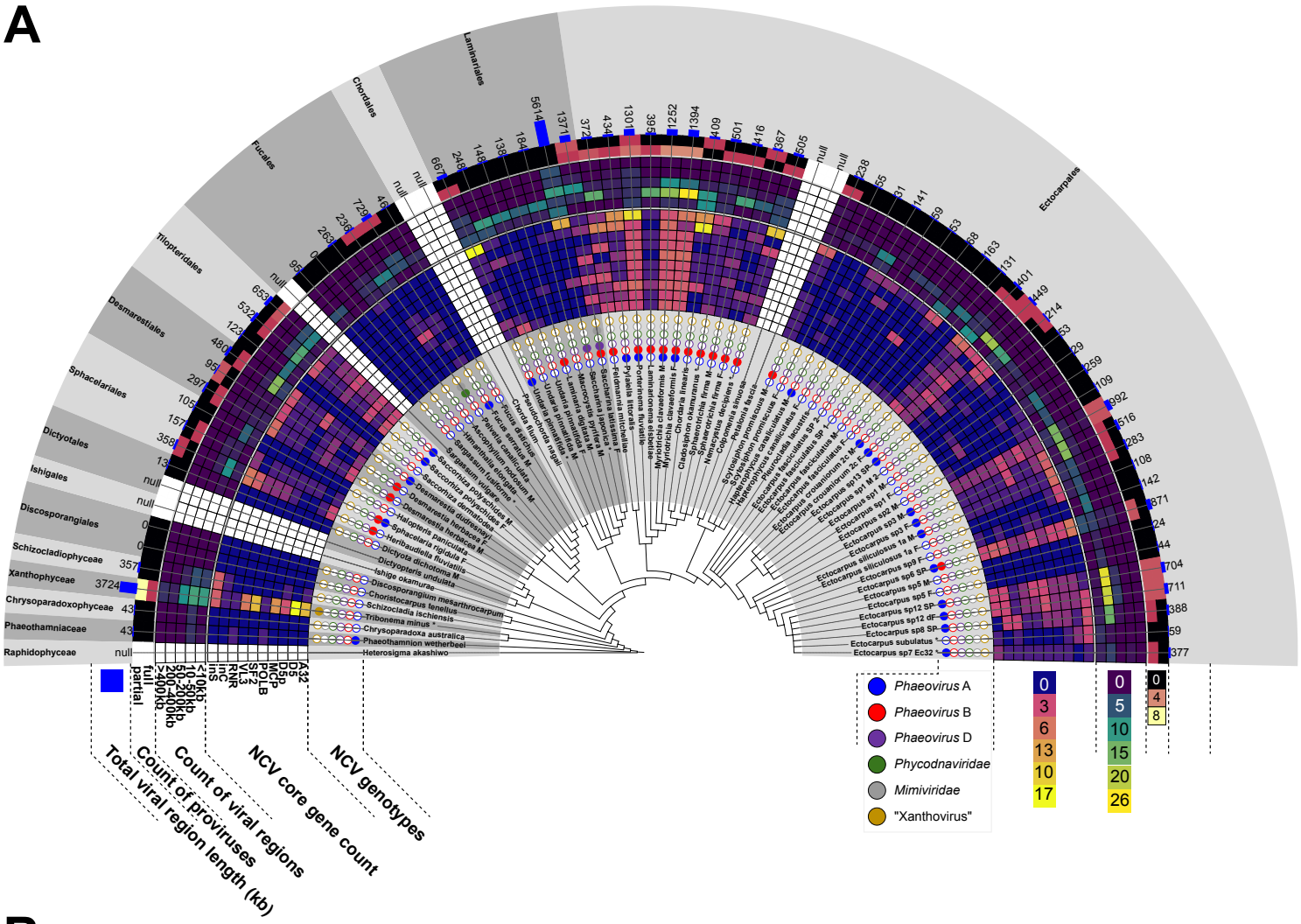


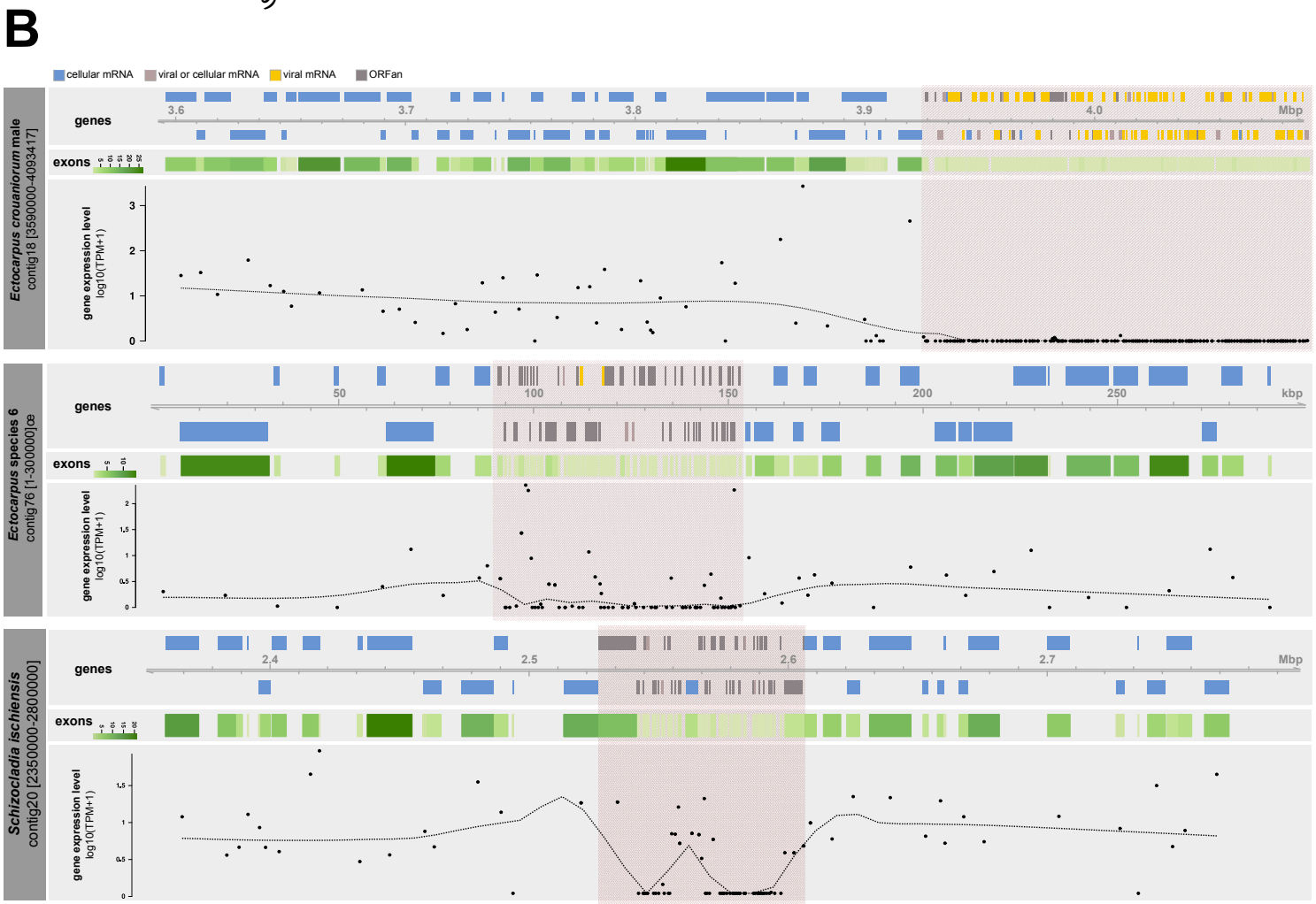

A
Phaeophyceae
Ectocarpaceae
Scytosiphonaceae
Chordariaceae
Acinetosporaceae
Laminariaceae
Alariaceae
Chordales
Fucaceae
Sargassaceae
Tilopteridales
Desmarestiales
Sphacelariales
Dictyotales
Ishigeales
Discosporangiales
Schizocladiophyceae
Chrysoparadoxophyceae
Xanthophyceae
Phaeothamniophyceae
Raphidophyceae
Fucophycidae
FDI clade
Phaeophyceae
PS clade
Emergence of the Phaeophyceae
B
Gene family gain and loss
proportion of gain and loss gene families
number of gene families
Gain of gene families
Loss of gene families
Overall gain/loss
Number of gene families
C
Gene family amplification and reduction
percentage of gene families amplified/reduced
Amplification
Reduction
AELclade
Fucales
Ectocarpales
Laminariales
Chordariaceae
Fucaceae
Ectocarpaceae
Origin of amplified gene families
Percentage of the gene families gained at each phylostratum that was amplified
Percentage of the amplified gene families that were gained at each phylostratum
D
Domain fusions and fissions
Proportion of gene families gained at each phylostratum
Domain fissions
Domain fusions
E
Gene family founder events
Gene founder events (log scale)
Phylostratum
PX clade
PS clade
Phaeophyceae (class)
FDI clade
Fucophycidae (Subclass)
Superorder
Order
Family
Genus
Species
F
Genes acquired through HGT
Abundance
Origin
Age
Percent of genes
Proportion
Age
Ancestral
Class-specific
Origin
Archaea
Bacteria
Eukaryota
Viruses
G
Difference in frequency (%)
No COG category
Function unknown
Metabolism
Cell wall and signalling
Cytoskeleton and trafficking
Transcription and translation
Cell cycle and DNA maintenance
1) Orthogroup gains
2) Phylostratigraphy
3) Amplified families
4) Composite genes
5) HGT-derived genes


A
CAZYmes
HPO
PKS
TAPs
Membrane Proteins
Ion channels
EsV-1-7
DNMT
Histones
Flagella
Ectocarpus species 7
Ectocarpus siliculosus (M)
Ectocarpus crouaniorum (M)
Ectocarpus fasciculatus (M)
Pleurocladia lacustris
Scytosiphon promiscuus (M)
Chordaria linearis
Porterinema fluviatile
Pylaiella littoralis
Saccharina latissima (F)
Undaria pinnatifida
Fucus serratus
Ascophyllum nodosum
Sargassum fusiforme
Desmarestia herbacea (M)
Dictyota dichotoma (M)
Discosporangium mesarthrocarpum
Schizocladia ischiensis
Tribonema minus
Chrysoparadoxa australica
Heterosigma akashiwo
GTs
GHs
PLs
Algal HPO1
Algal HPO2
Bacterial HPO
PKS1
PKS2
PKS3
Total TAPs
HD_TALE
ARF-like
bHLH
HMG
C2H2
Zn_clus
GNAT
SWI/SNF_SNF2
LRR RK
QAD RK
HK total
CHASE HK
EBD HK
MASE1 HK
Integrin
DEK1
MCU
GLR
pLGIC
TRP
EsV-1-7
IMM
DNMT1
DNMT2
Total histones
Additional H3 variants
Helmchrome
MAS1
MAS2
MAS3
Flagellar titin
Creatine kinase
B
Auxiliary Activities
Carbohydrate Binding Module
Glycoside Hydrolase
Glycosyltransferase
Polysaccharide Lyase
Man C5-E
Sulfatase
Gene family gain (Dollo)
Founder events (phylostratigraphy)
HGT
Amplification
Number of proteins in OG with indicated annotation
50 100 150 200 250
Proportion of gene with indicated function in each OG
1.00 0.75 0.50 0.25
Ectocarpus sp7
Ectocarpus fasciculatus
Pleurocladia lacustris
Scytosiphon promiscuus
Porterinema fluviatile
Ectocarpales
ALE clade
Sargassum fusiforme
FNT clade
FNT+ALE clade
Desmarestia herbacea
Dictyota dichotoma
Fucophycidae
FDI clade
Phaeophyceae
Schizocladia ischiensis
PS clade
Ectocarpus
Saccharina japonica
Fucophycidae
FDI clade
Phaeophyceae
PS clade
Ectocarpus siliculosus
Ectocarpus
Cladosiphon okamuranus
Ectocarpales
Saccharina japonica
Phaeophyceae
Ectocarpales
ALE clade
Fucales
Fucales+Laminariales+Ectocarpales
FDI clade
Phaeophyceae
AA15
AA15-CBM1
AA15-CBM1-CBM1
AA15-CBM1-CBM1-CBM1
AA3
EXPN
CBM1
CBM1-CBM1
CBM18
CBM32
CBM50
CBM6
CBM6-CBM6
GH114
GH114-GH114
GH116
GH135
GH152
GH16
GH16_3
GH16_3-CBM6-CBM6
GH17
GH20
GH25
GH28
GH38
GH47
GH5_27
GH55
GH88
GT1
GT14
GT15
GT2
GT23
GT31
GT4
GT41
GT47
GT48
GT49
GT49-GT49
GT59
GT64
GT7
GT74-GT2
GT77
GT8
PL41
PL41-CBM32
PL41-CBM32-CBM32
ManC5-E
S1_14
S1_16
S1_8
AA15
GH114
GT23
PL41
mannuronan C5-epimerases
C
0,20
A-nod_693.1
D-her_1077.700
F-ser_7888.19142
A-nod_397.11
D-her_2261.6021
F-ser_862.20005
S-der_4929.12331
A-nod_8.140
S-jap_21203
E-sp7_00_010800
E-sp7_09-003950
D-dic_414.13512
S-isc_40.12309
S-isc_2.5452
S-poly_1918.6161
D-her_338.9180
S-isc_7.18132.1
A-nod_469.32
D-dic_89.19664
D-dic_678.17678
E-sp7_00-006380
S-jap_22028
S-jap_16091
S-lat_37.10628
S-lat_788.15762
S-jap_07100
D-dic_175.5036
A-nod_2.152.3
D-dic_118.1447
D-mes_1545.5
D-dic_11.864
D-dic_312.10795
S-isc_1.582
S-jap_05196
D-dic_1411.3053
A-nod_165.40
C-lin_2.4845
1
2
3
ECTOCARPALES
LAMINARIALES
FUCALES
TILOPTERIDIALES
DESMARESTIALES
DICTYOTALES
DISCOSPORANGIALES
Schizocladia ischensis
FUCALES/DESMARESTIALES
D
ECT [6]
F-ser_44.13436
LAM [2]
S-jap_13738
LAM [2]
ECT [2]
S-jap_22237
ECT [3]
LAM [6]
S-isc_23.7008
S-jap_10697
ECT [2]
D-dud_4917.1
S-jap_02438
D-dic_362.12190
F-ser_1348.2587
LAM [3]
F-ser_2530.8129
D-dic_810.18960
S-isc_55.15461
E-sp7_06-005280
D-dud_56.33
LAM [2]
C-lin_30.8128
F-serr_156.3746
D-dic_332.11405
ECT [4]
ECT [2]
FUC [2]
D-dic_115.1240
LAM [2]
0.20


A
Sargassaceae
Fucaceae
Alariaceae
Laminariaceae
Acinetosporaceae
Chordariaceae
Scytosiphonaceae
Ectocarpaceae
Tilopteridales
Desmarestiales
Sphacelariales
Dictyotales
Discosporangiales
Schizolcadiophyceae
Chrysoparadoxophyceae
Xanthophyceae
Phaeothamniophyceae
Raphidophyceae
Eustigmatophyceae
Mediophyceae
Bacillariophyceae
Pelagophyceae
Oomycota
Pelagophyceae (20)
Dictyochophyceae (8)
Raphidophyceae (2)
Chrysophyceae/Synurophyceae (21/2)
Phaeothamniophyceae (1)
Xanthophyceae (2)
Schizocladiophyceae (1)
Phaeophyceae PKS1/PKS3 (100/21)
Phaeothamniophyceae (1)
Schizocladiophyceae (1)
Phaeophyceae PKS2 (87)
Fungi (14)
Actinobacteria (49)
Cyanobacteria (47)
Amoebozoa (9)
Multiple bacteria (20)
Ochrophyta (267)
0.50
Phaeophyceae
PX clade
Chrysita
Diatoms
Diatomista
Ochrophyte
Stramenopiles
① PKS gene acquisiton
② PKS gene loss
③ PKS2 (PKS1 gene duplication)
④ PKS1 gene duplication
⑤ PKS3 (PKS1 gene duplication)
PKS1
PKS2
PKS3
B
Ectocarpus species 7
Dictyota dichotoma
Saccharina latissima
Undaria pinnatifida
Fucus serratus
Ascophyllum nodosum
Sargassum fusiforme
Desmarestia herbacea
zoospore
sperm


A
E. crouaniorum
E. subulatus
Ectocarpus species 2
Ectocarpus species 1
Ectocarpus species 7
Ectocarpus species 5
Ectocarpus species 6
Ectocarpus species 9
E. siliculosus
Ectocarpus species 3
E. fasciculatus
S. promiscuus
Clade 1
Clade 2
0.50
0.38
0.25
0.12
0.00
B
D statistic
0.1
0.0
−0.1
−0.2
sp2_sp1_sp9
sp2_sp1_sp7
sp2_sp1_sp5
sp2_sp1_sp6
sub_ecro_esil
sp2_sp1_esil
sp2_sp1_sp3
sp7_sp9_sp3
sp2_sp1_ecro
sp1_sp2_sub
sub_ecro_sp2
sub_ecro_sp1
sub_ecro_sp7
sub_ecro_sp5
sub_ecro_sp9
sub_ecro_sp6
sub_ecro_sp3
sp7_sp5_sp9
sp7_sp5_sp3
sp6_sp9_esil
sp7_sp9_esil
sp6_sp5_esil
sp5_sp9_esil
sp7_sp5_esil
sp6_sp9_sp3
sp5_sp9_sp3
sp6_sp5_sp3
sp6_sp7_sp9
sp6_sp7_esil
sp7_sp6_sp9
sp6_sp7_sp3
sp5_sp6_sp9
sp7_esil_sp3
sp6_esil_sp3
sp5_esil_sp3
sp9_esil_sp3
sp7_sp5_sp6
Within clade 1
Between clades
Within clade 2


A
Total viral region length (kb)
Count of proviruses
Count of viral regions
NCV core gene count
NCV genotypes
Phaeovirus A
Phaeovirus B
Phaeovirus D
Phycodnaviridae
Mimiviridae
"Xanthovirus"


B
cellular mRNA
viral or cellular mRNA
viral mRNA
ORFan
genes
exons
gene expression level log10(TPM+1)
Ectocarpus crouaniorum male contig18 [3590000-4093417]
Ectocarpus species 6 contig76 [1-300000]ce
Schizocladia ischiensis contig20 [2350000-2800000]