

1 The genome of *Ectocarpus subulatus* – a 2 highly stress-tolerant brown alga

3 Simon M. Dittami^{1*}, Erwan Corre², Loraine Brillet-Guéguen^{1,2}, Agnieszka P. Lipinska¹, Noé
4 Pontoizeau^{1,2}, Meziane Aite³, Komlan Avia^{1,4}, Christophe Caron^{2†}, Chung Hyun Cho⁵, Jonas Collén¹,
5 Alexandre Cormier¹, Ludovic Delage¹, Sylvie Doubleau⁶, Clémence Frioux³, Angélique Gobet¹,
6 Irene González-Navarrete⁷, Agnès Groisillier¹, Cécile Hervé¹, Didier Jollivet⁸, Hetty KleinJan¹,
7 Catherine Leblanc¹, Xi Liu², Dominique Marie⁸, Gabriel V. Markov¹, André E. Minoche^{7,9}, Misharl
8 Monsoor², Pierre Pericard², Marie-Mathilde Perrineau¹, Akira F. Peters¹⁰, Anne Siegel³, Amandine
9 Siméon¹, Camille Trottier³, Hwan Su Yoon⁵, Heinz Himmelbauer^{7,9,11}, Catherine Boyen¹, Thierry
10 Tonon^{1,12}

11
12 ¹ Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique
13 de Roscoff, 29680 Roscoff, France

14 ² CNRS, Sorbonne Université, FR2424, ABiMS platform, Station Biologique de Roscoff, 29680,
15 Roscoff, France

16 ³ Institute for Research in IT and Random Systems - IRISA, Université de Rennes 1, France

17 ⁴ Université de Strasbourg, INRA, SVQV UMR-A 1131, F-68000 Colmar, France

18 ⁵ Department of Biological Sciences, Sungkyunkwan University, Suwon 16419, Republic of Korea

19 ⁶ IRD, UMR DIADE, 911 Avenue Agropolis, BP 64501, 34394 Montpellier, France

20 ⁷ Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr.
21 Aiguader 88, Barcelona, 08003 Spain

22 ⁸ Sorbonne Université, CNRS, Adaptation and Diversity in the Marine Environment (ADME),
23 Station Biologique de Roscoff (SBR), 29680 Roscoff, France

24 ⁹ Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

25 ¹⁰ Bezhin Rosko, 40 Rue des Pêcheurs, 29250 Santec, France

26 ¹¹ Department of Biotechnology, University of Natural Resources and Life Sciences (BOKU),
27 Vienna, 1190 Vienna, Austria

28 ¹² Centre for Novel Agricultural Products, Department of Biology, University of York, Heslington,
29 York, YO10 5DD, United Kingdom.

30 † Deceased

31

32 * Correspondence: simon.dittami@sb-roscoff.fr, phone +33 29 82 92 362, fax +33 29 82 92 324.

33

34 **Abstract**

35

36 Brown algae are multicellular photosynthetic stramenopiles that colonize marine rocky shores
37 worldwide. *Ectocarpus* sp. Ec32 has been established as a genomic model for brown algae. Here we
38 present the genome and metabolic network of the closely related species, *Ectocarpus subulatus*
39 Kützinger, which is characterized by high abiotic stress tolerance. Since their separation, both strains
40 show new traces of viral sequences and the activity of large retrotransposons, which may also be
41 related to the expansion of a family of chlorophyll-binding proteins. Further features suspected to
42 contribute to stress tolerance include an expanded family of heat shock proteins, the reduction of
43 genes involved in the production of halogenated defence compounds, and the presence of fewer cell
44 wall polysaccharide-modifying enzymes. Overall, *E. subulatus* has mainly lost members of gene
45 families down-regulated in low salinities, and conserved those that were up-regulated in the same
46 condition. However, 96% of genes that differed between the two examined *Ectocarpus* species, as
47 well as all genes under positive selection, were found to encode proteins of unknown function. This
48 underlines the uniqueness of brown algal stress tolerance mechanisms as well as the significance of
49 establishing *E. subulatus* as a comparative model for future functional studies.

50

51 Introduction

52 Brown algae (Phaeophyceae) are multicellular photosynthetic organisms that are successful
53 colonizers of rocky shores in the world's oceans. In many places they constitute the dominant
54 vegetation in the intertidal zone, where they have adapted to multiple stressors including strong
55 variations in temperature, salinity, irradiation, and mechanical stress (wave action) over the tidal
56 cycle¹. In the subtidal environment, brown algae form kelp forests that harbor highly diverse
57 communities². They are also harvested as food or for industrial purposes, such as the extraction of
58 alginates³. The worldwide annual harvest of brown algae has reached 10 million tons in 2014 and is
59 constantly growing⁴. Brown algae share some basic photosynthetic machinery with land plants, but
60 their plastids derived from a secondary or tertiary endosymbiosis event with a red alga, and they
61 belong to an independent lineage of eukaryotes, the stramenopiles⁵. This phylogenetic background,
62 together with their distinct habitat, contributes to the fact that brown algae have evolved numerous
63 unique metabolic pathways, life cycle features, and stress tolerance mechanisms.

64 To enable functional studies of brown algae, strain Ec32 of the small filamentous alga *Ectocarpus*
65 sp. has been established as a genetic and genomic model⁶⁻⁸. This strain was formerly described as
66 *Ectocarpus siliculosus*, but has since been shown to belong to an independent clade by molecular
67 methods^{9,10}. More recently, three additional brown algal genomes, that of the kelp species
68 *Saccharina japonica*¹¹, that of *Cladosiphon okamuranus*¹², and that of *Nemacystus decipiens*¹³, have
69 been characterized. Comparisons between these four genomes have allowed researchers to obtain a
70 first overview of the unique genomic features of brown algae, as well as a glimpse of the genetic
71 diversity within this group. However, given the evolutionary distance between these algae, it is
72 difficult to link genomic differences to physiological differences and possible adaptations to their
73 lifestyle. To be able to generate more accurate hypotheses on the role of particular genes and
74 genomic features for adaptive traits, a common strategy is to compare closely related strains and
75 species that differ only in a few genomic features. The genus *Ectocarpus* is particularly well suited
76 for such comparative studies because it comprises a wide range of morphologically similar but
77 genetically distinct strains and species that have adapted to different marine and brackish water
78 environments^{9,14-16}. One species within this group, *Ectocarpus subulatus* Kützinger¹⁰, comprises
79 isolates highly resistant to elevated temperature¹⁷ and low salinity. A strain of this species was even
80 isolated from freshwater¹⁸, constituting one of the handful of known marine-freshwater transitions in
81 brown algae¹⁹.

82 Here we present the draft genome and metabolic network of a strain of *E. subulatus*, establishing the
83 genomic basis for its use as a comparative model to study stress tolerance mechanisms, and in
84 particular low salinity tolerance, in brown algae. Similar strategies have been successfully employed
85 in terrestrial plants, where “extremophile” relatives of model- or economically relevant species have
86 been sequenced to explore new stress tolerance mechanisms in the green lineage²⁰⁻²⁵. The study of
87 the *E. subulatus* genome, and subsequent comparative analysis with other brown algal genomes, in
88 particular that of *Ectocarpus* sp. Ec32, provides insights into the dynamics of *Ectocarpus* genome
89 evolution and divergence, and highlights important adaptive processes, such as a potentially
90 retrotransposon driven expansion of the family of chlorophyll-binding proteins with subsequent
91 diversification. Most importantly, our analyses underline that most of the observed differences
92 between the examined species of *Ectocarpus* correspond to proteins with yet unknown functions.

93 Results

94 Sequencing and assembly of the *E. subulatus* genome

95 A total of 34.7Gb of paired-end read data and of 28.8Gb of mate-pair reads (corresponding to 45
96 million non-redundant mate-pairs) were acquired (Supporting Information Table S1). The final
97 genome assembly size of strain Bft15b was 227Mb (Table 1), and we also obtained 123Mb of
98 bacterial contigs corresponding predominantly to *Alphaproteobacteria* (50%, with the dominant
99 genera *Roseobacter* 8% and *Hyphomonas* 5%), followed by *Gammaproteobacteria* (18%), and
100 *Flavobacteria* (13%). The mean sequencing coverage of mapped reads was 67X for the paired-end
101 library, and the genomic coverage was 6.9, 14.4, and 30.4X for the 3kb, 5kb, and 10kb mate-pair
102 libraries, respectively. RNA-seq experiments yielded 8.8Gb of RNA-seq data, of which 96.6%
103 (Bft15b strain in seawater), 87.6% (freshwater strain in seawater), and 85.3% (freshwater strain in
104 freshwater) aligned with the final genome assembly of the Bft15b strain.

105 Gene prediction and annotation

106 The number of predicted proteins in *E. subulatus* was 60% higher than that predicted for Ec32
107 (Table 1), mainly due to the presence of mono-exonic genes, many of which corresponded to
108 transposases, which were not removed from our predictions, but had been manually removed from
109 the Ec32 genome. Ninety-eight percent of the gene models were supported by at least one associated
110 RNA-seq read, and 92% were supported by at least ten reads, with lowly-expressed (<10 reads)
111 genes being generally shorter (882 vs 1,403 bases), and containing fewer introns (2.6 vs 5.7). In
112 7.3% of all predicted proteins we detected a signal peptide, and 3.7% additionally contained an
113 'ASAFAP'-motif (Supporting Information Table S2) indicating that they are likely targeted to the
114 plastid²⁶. Overall the BUSCO²⁷ analyses indicate that the *E. subulatus* genome is 86% complete
115 (complete and fragmented genes) and 91% when not considering proteins also absent from all other
116 currently sequenced brown algae (Table 1).

117 Repeated elements

118 Thirty percent of the *E. subulatus* genome consisted of repeated elements. The most abundant groups
119 of repeated elements were large retrotransposon derivatives (LARDs), followed by long terminal
120 repeats (LTRs, predominantly Copia and Gypsy), and long and short interspersed nuclear elements
121 (LINEs, Figure 1A). The overall distribution of sequence identity levels within superfamilies showed
122 two peaks, one at an identity level of 78-80%, and one at 96-100% (Figure 1C). An examination of
123 transposon conservation at the level of individual families revealed a few families that follow this
124 global bimodal distribution (e.g. TIR B343 or LARD B204), while the majority exhibited a
125 unimodal distribution with peaks either at high (e.g. LINE R15) or at lower identity levels (e.g.
126 LARD B554) (Figure 1C). Terminal repeat retrotransposons in miniature (TRIM) and LARDs, both
127 non-autonomous groups of retrotransposons, were among the most conserved families. A detailed
128 list of transposons is provided in Supporting Information Table S3. In line with previous
129 observations carried out in *Ectocarpus* sp. Ec32, no methylation was detected in the *E. subulatus*
130 genomic DNA.

131 **Organelar genomes**

132 Plastid and mitochondrial genomes from *E. subulatus* have 95.5% and 91.5% sequence identity with
133 their *Ectocarpus* sp. Ec32 counterparts in the conserved regions respectively. Only minor structural
134 differences were observed between organelar genomes of both *Ectocarpus* genomes, as detailed in
135 Supporting Information Text S1.

136 **Global comparison of predicted proteomes**

137 **Metabolic network-based comparisons**

138 Similar to the network previously obtained for *Ectocarpus* sp. Ec32²⁸, the *E. subulatus* Bft15b
139 metabolic network comprised 2,074 metabolic reactions and 2,173 metabolites in 464 pathways,
140 which can be browsed at <http://gem-aureme.irisa.fr/sububftgem>. In total, 2,445 genes associated with
141 at least one metabolic reaction, and 215 pathways were complete (Figure 2). Comparisons between
142 both networks were carried out on a pathway level (Supporting Information Text S1, Section
143 “Metabolic network-based comparisons”), but no pathways were found to be truly specific to either
144 Ec32 and/or Bf15b.

145 **Genes under positive selection**

146 Out of the 2,311 orthogroups with single-copy orthologs that produced high quality alignments, 172
147 gene pairs (7.4%) exhibited dN/dS ratios > 0.5 (Supporting Information Table S4). Among these,
148 only eleven (6.4%) were found to fit significantly better with the model allowing for positive
149 selection in the *Ectocarpus* branch. These genes are likely to have been under positive selection, and
150 two of them contained a signal peptide targeting the plastid. All of them are genes specific to the
151 brown algal lineage with unknown function, and only two genes contained protein domains related
152 to a biochemical function (one oxidoreductase-like domain, and one protein prenyltransferase, alpha
153 subunit). However, all of them were expressed at least in *E. subulatus* Bft15b. There was no trend
154 for these genes to be located in specific regions of the genome (all except two for *Ectocarpus* sp.
155 Ec32 were on different scaffolds) and none of the genes were located in the pseudoautosomal region
156 of the sex chromosome.

157 **Genes specific to either *Ectocarpus* genome, and expanded genes and gene families**

158 After manual curation based on tblastn searches to eliminate artefacts arising from differences in the
159 gene predictions, 184 expanded gene clusters and 1,611 predicted proteins were found to be specific
160 to *E. subulatus* compared to *Ectocarpus* sp., while 449 clusters were expanded and 689 proteins
161 were found specifically in the latter (Figure 2, Supporting Information Table S5). This is far less
162 than the 2,878 and 1,093 unique clusters found for a recent comparison of *N. decipiens* and *C.*
163 *okamuranus*¹³. Gene set enrichment analyses revealed no GO categories to be significantly over-
164 represented among the genes unique to or expanded in *E. subulatus* Bft15b, but several categories
165 were over-represented among the genes and gene families specific to or expanded in the *Ectocarpus*
166 sp. Ec32 strain. Many were related either to signalling pathways or to the membrane and transporters
167 (Figure 2), but it is difficult to distinguish between the effects of a potentially incomplete genome
168 assembly and true gene losses in Bft15b. In the manual analyses we therefore focussed on the genes
169 specific to and expanded in *E. subulatus*.

170 Among the 1,611 *E. subulatus*-specific genes, 1,436 genes had no homologs (e-value < 1e-5) in the
171 UniProt database as of May 20th 2016: they could thus, at this point in time, be considered lineage-
172 specific and had no function associated to them. Among the remaining 175 genes, 145 had hits (e-
173 value < 1e-5) in *Ectocarpus* sp. Ec32, *i.e.* they likely correspond to multi-copy genes that had
174 diverged prior to the separation of *Ectocarpus* and *S. japonica*, and for which the *Ectocarpus* sp.
175 Ec32 and *S. japonica* orthologs were lost. Thirteen genes had homology only with uncharacterized
176 proteins or were too dissimilar from characterized proteins to deduce hypothetical functions; another
177 eight probably corresponded to short viral sequences integrated into the algal genome
178 (EsuBft1730_2, EsuBft4066_3, EsuBft4066_2, EsuBft284_15, EsuBft43_11, EsuBft551_12,
179 EsuBft1883_2, EsuBft4066_4), and one (EsuBft543_9) was related to a retrotransposon. Two
180 adjacent genes (EsuBft1157_4, EsuBft1157_5) were also found in diatoms and may be related to the
181 degradation of cellobiose and the transport of the corresponding sugars. Two genes, EsuBft1440_3
182 and EsuBft1337_8, contained conserved motifs (IPR023307 and SSF56973) typically found in toxin
183 families. Two more (EsuBft1006_6 and EsuBft308_11) exhibited low similarities to animal and
184 fungal transcription factors, and the last (EsuBft36_20 and EsuBft440_20) consisted almost
185 exclusively of short repeated sequences of unknown function (“ALEW” and
186 “GAAASGVAGGAVVVNG”, respectively). In total, 1.7% contained a signal peptide targeting the
187 plastid, *i.e.* significantly less than the 3.7% in the entire dataset (Fisher exact test, $p < 0.0001$).

188 The large majority of *Ectocarpus* sp. Ec32-specific proteins (511) also corresponded to proteins of
189 unknown function without matches in public databases. Ninety-seven proteins were part of the *E.*
190 *siliculosus* virus-1 (EsV-1) inserted into the Ec32 genome and the remaining 81 proteins were poorly
191 annotated, usually only via the presence of a domain. Examples are ankyrin repeat-containing
192 domain proteins (12), Zinc finger domain proteins (6), proteins containing wall sensing component
193 (WSC) domains (3), protein kinase-like proteins (3), and Notch domain proteins (2).

194 Regarding the 184 clusters of expanded genes in *E. subulatus*, 139 (1,064 proteins) corresponded to
195 proteins with unknown function, 98% of which were found only in *Ectocarpus*. Furthermore, nine
196 clusters (202 proteins) represented sequences related to transposons predicted in both genomes, and
197 eight clusters (31 proteins) were similar to known viral sequences. Only 28 clusters (135 proteins)
198 could be roughly assigned to biological functions (Table 2). They comprised proteins potentially
199 involved in modification of the cell-wall structure (including sulfation), in transcriptional regulation
200 and translation, in cell-cell communication and signalling, as well as a few stress response proteins,
201 notably a set of HSP20s, and several proteins of the light-harvesting complex (LHC) potentially
202 involved in non-photochemical quenching. Only 0.6% of all genes expanded in Bft15b contained a
203 signal peptide targeting the plastid, *i.e.* significantly less than the 3.7% in the entire dataset (Fisher
204 exact test, $p < 0.0001$).

205 Striking examples of likely expansions in *Ectocarpus* sp. Ec32 or reduction in *E. subulatus* Bft15b
206 were different families of serine-threonine protein kinase domain proteins present in 16 to 25 copies
207 in Ec32 compared to only 5 or 6 in Bft15b, Kinesin light chain-like proteins (34 vs. 13 copies), two
208 clusters of Notch region containing proteins (11 and 8 vs. 2 and 1 copies), a family of unknown
209 WSC domain containing proteins (8 copies vs. 1), putative regulators of G-protein signalling (11 vs.
210 4 copies), as well as several expanded clusters of unknown and viral proteins. However, these results

211 need to be taken with caution because the *E. subulatus* Bft15b genome was less complete than that
212 of *Ectocarpus* sp. Ec32.

213 **Correlation with gene expression patterns**

214 To assess whether genomic adaptations in *E. subulatus* Bft15b were located preferentially in genes
215 that are known to be responsive to salinity stress, we compared expanded gene families to previously
216 available expression data obtained for a freshwater strain of *E. subulatus* grown in freshwater vs
217 seawater²⁹. This analysis revealed that genes that were down-regulated in response to low salinity
218 were significantly over-represented among the gene families expanded in *Ectocarpus* sp. Ec32 or
219 reduced in *E. subulatus* Bft15b, (42% of genes vs 26% for all genes; Fischer exact test $p=0.0002$),
220 while genes that were upregulated in response to low salinity were significantly under-represented
221 (25% vs 33%; Fischer exact test $p=0.006$; Figure 3, Supporting Information Table S6). This
222 indicates that *E. subulatus* Bft15b has mainly lost members of gene families that were generally
223 down-regulated in low salinities, and conserved those that were upregulated in this condition.

224 **Targeted manual annotation of specific pathways**

225 In addition to the global analyses carried out above, genes related to cell wall metabolism, sterol
226 metabolism, polyamine and central carbon metabolism, algal defence metabolites, transporters, and
227 abiotic stress response were manually examined and annotated, because, based on literature studies,
228 these functions could be expected to explain the physiological differences between *E. subulatus*
229 Bft15b and *Ectocarpus* sp. Ec32. Overall the differences between both *Ectocarpus* strains with
230 respect to these genes were minor; a detailed description of these results is available in Supporting
231 Information Text S1 and Supporting Information Table S7, and a brief overview of the main
232 differences is presented below.

233 Regarding gene families reduced in *E. subulatus* Bft15b or expanded in *Ectocarpus* sp. Ec32, the *E.*
234 *subulatus* genome encoded only 320 WSC-domain containing proteins, vs. 444 in *Ectocarpus* sp..
235 Many of these genes were down-regulated in response to low salinity, (61% of the WSC domain
236 containing genes with available expression data; Fischer exact test, $p=0.0004$) while only 7% were
237 upregulated (Fischer exact test, $p\text{-value}=0.0036$). In yeast, WSC domain proteins may act as cell
238 surface mechanosensors and activate the intracellular cell wall integrity signalling cascade in
239 response to hypo-osmotic shock³⁰. Whether or not they have similar functions in brown algae,
240 however, remains to be established. Furthermore, we found fewer aryl sulfotransferase, tyrosinases,
241 potential bromoperoxidases, and thyroid peroxidases in the *E. subulatus* genome compared to
242 *Ectocarpus* sp., and it entirely lacks haloalkane dehalogenases (Supporting Information Text S1). All
243 of these enzymes are involved in the production of polyphenols and halogenated defence
244 compounds, suggesting that *E. subulatus* may be investing less energy in defence, although a
245 potential bias induced by differences in the assembly completeness cannot be excluded here.

246 Regarding gene families expanded in *E. subulatus* Bft15b or reduced in *Ectocarpus* sp. Ec32, we
247 detected differences with respect to a few “classical” stress response genes. Notably an HSP20
248 protein was present in three copies in the genome of *E. subulatus* and only one copy in *Ectocarpus*
249 sp.. We also found a small group of LHCX-family chlorophyll-binding proteins (CBPs) as well as a
250 larger group belonging to the LHCF/LHCR family that have probably undergone a recent expansion
251 in *E. subulatus* (Figure 4). Some of the proteins appeared to be truncated (marked with asterisks),

252 but all of them were associated with RNA-seq reads, suggesting that they may be functional. A
253 number of these proteins were also flanked by LTR-like sequences. CBPs have been reported to be
254 up-regulated in response to abiotic stress in stramenopiles^{31,32}, including *Ectocarpus*³³, probably as a
255 way to deal with excess light energy when photosynthesis is affected.

256 Discussion

257 Here we present the draft genome and metabolic network of *E. subulatus* strain Bft15b, a brown alga
258 which, compared to *Ectocarpus* sp. Ec32, is characterized by high abiotic stress tolerance^{10,17}. Based
259 on time-calibrated molecular trees, both species separated roughly 16 Mya²⁹, *i.e.* slightly before the
260 split between *Arabidopsis thaliana* and *Thellungiella salsuginea* (7-12 Mya)³⁴. This split was
261 probably followed by an adaptation of *E. subulatus* to highly fluctuating and low salinity habitats¹⁹.

262 Traces of recent transposon activity and integration of viral sequences

263 The *E. subulatus* Bft15b genome is only approximately 6% (flow cytometry) to 23% (genome
264 assembly) larger than that of *Ectocarpus* sp. Ec32, and no major genomic rearrangements or
265 duplications were detected. However, we observed traces of recent transposon activity, especially
266 from LTR transposons, which is in line with the absence of DNA methylation. Bursts in transposon
267 activity have been identified as one potential driver of local adaptation and speciation in other model
268 systems such as salmon³⁵ or land plants^{34,36}. Furthermore, LTRs are known to mediate the
269 retrotransposition of individual genes, leading to the duplication of the latter³⁷. In *E. subulatus*
270 Bft15b, only a few expansions of gene families were observed since the separation from *Ectocarpus*
271 sp. Ec32, and only in the case of the recent expansion of the LHCR family were genes flanked by a
272 pair of LTR-like sequences. These elements lacked both the group antigen (GAG) and reverse
273 transcriptase (POL) proteins, which implies that, if retro-transposition was the mechanism
274 underlying the expansion of this group of proteins, it would have depended on other active
275 transposable elements to provide these activities.

276 A second factor that has shaped the *Ectocarpus* genomes were viruses. Viral infections are a
277 common phenomenon in Ectocarpales³⁸, and a well-studied example is the *Ectocarpus siliculosus*
278 virus-1 (EsV-1)³⁹. It was found to be present latently in several strains of *Ectocarpus* sp. closely
279 related to strain Ec32, and has also been found integrated in the genome of the latter, although it is
280 not expressed⁷. As previously indicated by comparative genome hybridization experiments⁴⁰, the *E.*
281 *subulatus* Bft15b genome does not contain a complete EsV-1 like insertion, although a few shorter
282 EsV-1-like proteins were found. Thus, the EsV-1 integration observed in *Ectocarpus* sp. Ec32 has
283 likely occurred after the split with *E. subulatus*, and the biological consequences of this insertion
284 remain to be explored.

285 Few classical stress response genes but no transporters involved in adaptation

286 One aim of this study was to identify genes that may potentially be responsible for the high abiotic
287 stress and salinity tolerance of *E. subulatus*. Similar studies on genomic adaptation to changes in
288 salinity or to drought in terrestrial plants have previously highlighted genes generally involved in
289 stress tolerance to be expanded in “extremophile” organisms. Examples are the expansion of
290 catalase, glutathione reductase, and heat shock protein families in desert poplar²⁴, arginine
291 metabolism in jujube⁴¹, or genes related to cation transport, abscisic acid signalling, and wax

292 production in *T. salsuginea*³⁴. In our study, we found that gene families reduced in *E. subulatus*
293 Bft15b compared to the marine *Ectocarpus* sp. Ec32 model have previously been shown to be
294 repressed in response to stress, whereas gene families up-regulated in response to stress had a higher
295 probability of being conserved. However, there are only few signs of known stress response gene
296 families among them, notably the two additional HSP20 proteins and an expanded family of CBPs.
297 *E. subulatus* Bft15b also has a slightly reduced set of genes involved in the production of
298 halogenated defence compounds that may be related to its habitat preference: it is frequently found
299 in brackish and even freshwater environments with low availability of halogens. It also specializes in
300 habitats with high levels of abiotic stress compared to most other brown algae, and may thus invest
301 less energy in defence against biotic stressors.

302 Another anticipated adaptation to life in varying salinities lies in modifications of the cell wall.
303 Notably, the content of sulfated polysaccharides is expected to play a crucial role as these
304 compounds are present in all marine plants and algae, but absent in their freshwater relatives^{42,43}.
305 The fact that we found only small differences in the number of encoded sulfatases and
306 sulfotransferases indicates that the absence of sulfated cell-wall polysaccharides previously observed
307 in *E. subulatus* in low salinities⁴⁴ is probably a regulatory effect or simply related to the lack of
308 sulfate in low salinity. This is also coherent with the wide distribution of *E. subulatus* in marine,
309 brackish water, and freshwater environments.

310 Finally, transporters have previously been described as a key element in plant adaptation to different
311 salinities⁴⁵. Similar results have also been obtained for *Ectocarpus* in a study of quantitative trait loci
312 (QTLs) associated with salinity and temperature tolerance⁴⁶. In our study, however, we found no
313 indication of genomic differences related to transporters between the two species. This observation
314 corresponds to previous physiological experiments indicating that *Ectocarpus*, unlike many
315 terrestrial plants, responds to strong changes in salinity as an osmoconformer rather than an
316 osmoregulator, *i.e.* it allows the intracellular salt concentration to adjust to values close to the
317 external medium rather than keeping the intracellular ion composition constant³³.

318 **Species-specific genes of unknown function are likely to play a dominant role in** 319 **adaptation**

320 In addition to genes that may be directly involved in the adaptation to the environment, we found
321 several gene clusters containing domains potentially involved in cell-cell signalling that were
322 expanded in the *Ectocarpus* sp. Ec32 genome (Table 2), *e.g.* a family of ankyrin repeat-containing
323 domain proteins⁴⁷. These observed differences may be, in part, responsible for the existing pre-
324 zygotic reproductive barrier between the two examined species of *Ectocarpus*⁴⁸.

325 The vast majority of genomic differences between the two investigated species of *Ectocarpus*,
326 however, corresponds to proteins of entirely unknown functions. All of the 11 gene pairs under
327 positive selection were unknown genes taxonomically restricted to brown algae. Of the 1,611 *E.*
328 *subulatus* Bft15b-specific genes, 88% were unknown. Most of these genes were expressed and are
329 thus likely to correspond to true genes; their absence from the *Ectocarpus* sp. Ec32 genome was also
330 confirmed at the nucleotide level. A large part of the mechanisms that underlie the adaptation to
331 different ecological niches in *Ectocarpus* may, therefore, lie in these genes of unknown function.
332 This can be partly explained by the fact that still only few brown algal genomes have been

333 sequenced, and that currently most of our knowledge on the function of their proteins is based on
334 studies in model plants, animals, yeast, or bacteria, which have evolved independently from
335 stramenopiles for over 1 billion years⁴⁹. They differ from land plants even in otherwise highly
336 conserved aspects, for instance in their life cycles, cell walls, and primary metabolism⁵⁰. Substantial
337 contributions of lineage-specific genes to the evolution of organisms and the development of
338 innovations have also been described for animal models⁵¹, and studies in basal metazoans
339 furthermore indicate that they are essential for species-specific adaptive processes⁵².

340 Despite the probable importance of these unknown genes for local adaptation, *Ectocarpus* may still
341 heavily rely on classical stress response genes for abiotic stress tolerance. Many of the gene families
342 known to be related to stress response in land plants (including transporters and genes involved in
343 cell wall modification), and for which no significant differences in gene contents were observed,
344 have previously been reported to be strongly regulated in response to environmental stress in
345 *Ectocarpus*^{29,33,53}. This high transcriptomic plasticity is probably one of the features that allow
346 *Ectocarpus* to thrive in a wide range of environments, and may form the basis for its capacity to
347 further adapt to “extreme environments” such as freshwater¹⁸.

348 Conclusion and future work

349 We have shown that since the separation of *E. subulatus* and *Ectocarpus sp.* Ec32, both genomes
350 have been shaped partially by the activity of viruses and transposons, particularly large
351 retrotransposons. Over this period of time, *E. subulatus* has adapted to environments with high
352 abiotic variability including brackish water and even freshwater. We have identified a few genes that
353 likely contribute to this adaptation, including HSPs, CBPs, a reduction of genes involved in
354 halogenated defence compounds, or some changes in cell wall polysaccharide-modifying enzymes.
355 However, the majority of genes that differ between the two examined *Ectocarpus* species or that
356 may be under positive selection encode proteins of unknown function. This underlines the
357 fundamental differences that exist between brown algae and terrestrial plants or other lineages of
358 algae. Studies as the present one, *i.e.* without strong *a priori* assumptions about the mechanisms
359 involved in adaptation, are therefore essential to start elucidating the specificities of this lineage as
360 well as the various functions of the unknown genes.

361 Methods

362 **Biological material.** Haploid male parthenosporophytes of *E. subulatus* strain Bft15b (Culture
363 Collection of Algae and Protozoa CCAP accession 1310/34), isolated in 1978 by Dieter G. Müller in
364 Beaufort, North Carolina, USA, were grown in 14 cm (100 ml) Petri Dishes in Provasoli-enriched
365 seawater⁵⁴ under a 14/10 daylight cycle at 14°C. Strains were examined by light microscopy (800X
366 magnification, phase contrast) to ensure that they were free of contaminating eukaryotes, but did still
367 contain some alga-associated bacteria. Approximately 1 g fresh weight of algal culture was dried on
368 a paper towel and immediately frozen in liquid nitrogen. For RNA-seq experiments, in addition to
369 Bft15b, a second strain of *E. subulatus*, the diploid freshwater strain CCAP 1310/196 isolated from
370 Hopkins River Falls, Australia¹⁸, was included. One culture was grown as described above for
371 Bft15b, and for a second culture, seawater was diluted 20-fold with distilled water prior to the
372 addition of Provasoli nutrients²⁹ (culture condition referred to as freshwater).

373 **Flow cytometry** experiments to measure nuclear DNA contents were carried out as previously
374 described⁵⁵, except that young sporophyte tissue was used instead of gametes. Samples of the
375 genome-sequenced *Ectocarpus* sp. strain Ec32 (CCAP accession 1310/4 from San Juan de Marcona,
376 Peru) were run in parallel as a size reference.

377 **DNA and RNA** were extracted using a phenol-chloroform-based protocol⁵⁶. For **DNA sequencing**,
378 four Illumina libraries were prepared and sequenced on a HiSeq2000: one paired-end library
379 (Illumina TruSeq DNA PCR-free LT Sample Prep kit #15036187, sequenced with 2x100 bp read
380 length), and three mate-pair libraries with span sizes of 3kb, 5kb, and 10kb respectively (Nextera
381 Mate Pair Sample Preparation Kit; sequenced with 2x50bp read length). One poly-A enriched RNA-
382 seq library was generated for each of the three aforementioned cultures according to the Illumina
383 TruSeq Stranded mRNA Sample Prep kit #15031047 protocol and sequenced with 2x50 bp read
384 length.

385 The degree of **DNA methylation** was examined by HPLC on CsCl-gradient purified DNA⁵⁶ from
386 three independent cultures per strain as previously described⁵⁷.

387 Redundancy of mate-pairs (MPs) was reduced to mitigate the negative effect of redundant chimeric
388 MPs during scaffolding. To this means, mate-pair reads were aligned with bwa-0.6.1 to a
389 preliminary *E. subulatus* Bft15b draft assembly calculated from paired-end data only. Mate-pairs
390 that did not map with both reads were removed, and for the remaining pairs, read-starts were
391 obtained by parsing the cigar string using Samtools and a custom Pearl script. Mate-pairs with
392 redundant mapping coordinates were removed for the final **assembly**, which was carried out using
393 SOAPDenovo2⁵⁸. Scaffolding was then carried out using SSPACE basic 2.0⁵⁹ (trim length up to 5
394 bases, minimum 3 links to scaffold contigs, minimum 15 reads to call a base during an extension)
395 followed by a run of GapCloser (part of the SOAPDenovo package, default settings). A dot plot of
396 syntenic regions between *E. subulatus* Bft15b and *Ectocarpus* sp. Ec32 was generated using D-
397 Genies 1.2.0⁶⁰. Given the high degree of synteny observed (Supporting Information Text S1),
398 additional scaffolding was carried out using MeDuSa and the *Ectocarpus* sp. Ec32 genome as
399 reference⁶¹. This super-scaffolding method assumes that both genome structures are be similar.
400 Annotations were generated first for version 1 of the Bft15b genome and then transferred to the new
401 scaffolds of version 2 using the ALLMAPS⁶² liftover function. Both the assemblies with (V2) and
402 without (V1) MeDuSa scaffolding have been made available. RNA-seq reads were cleaned using
403 Trimmomatic (default settings), and a second Bft15b genome-guided assembly was performed with
404 Tophat2 and with Cufflinks. Sequencing coverage was calculated based on mapped algal reads only,
405 and for mate-pair libraries the genomic coverage was calculated as number of unique algal mate-
406 pairs * span size / assembly size.

407 As cultures were not treated with antibiotics prior to DNA extraction, **bacterial scaffolds were**
408 **removed** from the final assembly using the taxoblast pipeline⁶³. Every scaffold was cut into
409 fragments of 500 bp, and these fragments were aligned (blastn, e-value cutoff 0.01) against the
410 GenBank non-redundant nucleotide (nt) database. Scaffolds for which more than 90% of the
411 alignments were with bacterial sequences were removed from the assembly (varying this threshold
412 between 30 and 95% resulted in only very minor differences in the final assembly). Finally, we ran
413 the Anvi'o v5 pipeline to identify any remaining contaminant bins (both bacterial and eukaryote)

414 based on G/C and kmer contents as well as coverage⁶⁴. “Contaminant” scaffolds were submitted to
415 the MG-Rast server to obtain an overview of the taxa present in the sample⁶⁵. They are available at
416 <http://application.sb-roscoff.fr/blast/subulatus/download.html>.

417 **Repeated elements** were searched for *de novo* using TEdenovo and annotated using TEannot with
418 default parameters. LTR-like sequences were predicted by the LTR-harvest pipeline⁶⁶. These tools
419 are part of the REPET pipeline⁶⁷, of which version 2.5 was used for our dataset.

420 **BUSCO 2.0** analyses²⁷ were run on the servers of the IPlant Collaborative⁶⁸ with the general
421 eukaryote database as a reference and default parameters and the predicted proteins as input.

422 **Plastid and mitochondrial genomes** of *E. subulatus* Bft15b, were manually assembled based on
423 scaffolds 416 and 858 respectively, using the published organellar genomes of *Ectocarpus* sp. Ec32
424 (accessions NC_013498.1, NC_030223.1) as a guide^{7,69,70}. Genes were manually annotated based on
425 the result of homology searches with *Ectocarpus* sp. Ec32 using a bacterial genetic code (11) and
426 based on ORF predictions using ORF finder. Ribosomal RNA sequences were identified by
427 RNAmmer⁷¹ for the plastid and MITOS⁷² for the plastid, and tRNAs or other small RNAs were
428 identified using ARAGORN⁷³ and tRNAscan-SE⁷⁴. In the case of the mitochondrial genome, the
429 correctness of the manual assembly was verified by PCR where manual and automatic assemblies
430 diverged.

431 Putative **protein-coding sequences** were identified using Eugene 4.1c⁷⁵. Assembled RNA-seq reads
432 were mapped against the assembled genome using GenomeThreader 1.6.5, and all available proteins
433 from the Swiss-Prot database as well as predicted proteins from the *Ectocarpus* sp. Ec32 genome⁷
434 were aligned to the genome using KLAST⁷⁶. Both aligned *de novo*-assembled transcripts and
435 proteins were provided to Eugene for gene prediction, which was run with the parameter set
436 previously optimized for the *Ectocarpus* sp. Ec32 genome⁷. The subcellular localization of the
437 proteins was predicted using SignalP version 4.1⁷⁷ and the ASAFIND software version 1.1.5²⁶.

438 **For functional annotation**, predicted proteins were submitted to InterProScan and compared to the
439 Swiss-Prot database by BlastP search (e-value cutoff 1e-5), and the results imported to Blast2GO⁷⁸
440 The genome and all automatic annotations were imported into Apollo^{79,80} for manual curation.
441 During manual curation sequences were aligned with characterized reference sequences from
442 suitable databases (e.g. CAZYME, TCDB, SwissProt) using BLAST, and the presence of
443 InterProScan domains necessary for the predicted enzymatic function was manually verified.

444 The *E. subulatus* Bft15b **genome-scale metabolic model** reconstruction was carried out as
445 previously described²⁸ by merging an annotation-based reconstruction obtained with Pathway
446 Tools⁸¹ and an orthology-based reconstruction based on the *Arabidopsis thaliana* metabolic network
447 AraGEM⁸² using Pantograph⁸³. A final step of gap-filling was then carried out using the Meneco
448 tool⁸⁴. The entire reconstruction pipeline is available via the AuReMe workspace⁸⁵. For pathway-
449 based analyses, pathways that contained only a single reaction or that were less than 50% complete
450 were not considered.

451 **Functional comparisons of gene contents** were based primarily on orthologous clusters of genes
452 shared with version 2 of the *Ectocarpus* sp. Ec32 genome⁸⁶ as well as the *S. japonica* (Areschoug)

453 genome¹¹. They were determined by the OrthoFinder software version 0.7.1⁸⁷. To identify genes
454 specific to either of the *Ectocarpus* genomes, we examined all proteins that were not part of a multi-
455 species cluster and verified their absence in the other genome by tblastn searches (threshold e-value
456 of 1e-10). Only genes without tblastn hit that encoded proteins of at least 50 amino acids were
457 further examined. A second approach consisted in identifying clusters of genes that were expanded
458 or reduced in either of the two *Ectocarpus* genomes based on the Orthofinder results. Blast2GO 3.1⁷⁸
459 was then used to identify significantly enriched GO terms among the genes specific to either
460 *Ectocarpus* genome or the expanded/reduced gene families (Fischer's exact test with FDR correction
461 FDR<0.05). These different sets of genes were also examined manually for function, genetic
462 context, GC content, and EST coverage (to ensure the absence of contaminants).

463 The search for **genes under positive selection** was based on a previous analysis in other brown
464 algae⁸⁸. Therefore, Orthofinder analyses were expanded to include also *Macrocystis pyrifera*,
465 *Scytosiphon lomentaria*⁸⁸, and *Cladosiphon okamuranus*¹². Rates of non-synonymous to
466 synonymous substitution (ω =dN/dS) were searched for in clusters of single-copy orthologs. Protein
467 sequences were aligned with Tcoffee⁸⁹ (M-Coffee mode), translated back to nucleotide using
468 Pal2Nal⁹⁰, and curated with Gblocks⁹¹ (-t c -b4 20) or manually when necessary. Sequences that
469 produced a gapless alignment that exceeded 100bp were retained for pairwise dN/dS analysis
470 between *Ectocarpus* strains using CodeML (F3x4 model of codon frequencies, runmode = -2) of the
471 PAML4 suite⁹². Orthogroups for which the pairwise dN/dS ratio between *Ectocarpus* species
472 exceeded 0.5, which were not saturated (dS < 1), and which contained single-copy orthologs in at
473 least two other species were used to perform positive selection analysis with CodeML (PAML4,
474 F3x4 model of codon frequencies): branch-site models were used to estimate dN/dS values by site
475 and among branches in the species tree generated for each orthogroup. The branch leading to the
476 genus *Ectocarpus* was selected as a 'foreground branch', allowing different values of dN/dS among
477 sites in contrast to the remaining branches that shared the same distribution of ω . Two alternative
478 models were tested for the foreground branch: H1 allowing the dN/dS to exceed 1 for a proportion of
479 sites (positive selection), and H0 constraining dN/dS<1 for all sites (neutral and purifying selection).
480 A likelihood ratio test was then performed for the two models (LRT=2×(lnLH1-lnLH0)) and genes
481 for which H1 fitted the data significantly better (p<0.05) were identified as evolving under positive
482 selection.

483 **Phylogenetic analyses** were carried out for gene families of particular interest. For chlorophyll-
484 binding proteins (CBPs), reference sequences were obtained from a previous study⁹³, and aligned
485 together with *E. subulatus* Bft15b and *S. japonica* CBPs using MAFFT (G-INS-i)⁹⁴. Alignments
486 were then manually curated, conserved positions selected in Jalview⁹⁵, and maximum likelihood
487 analyses carried out using PhyML 3.0⁹⁶, the LG substitution model, 1000 bootstrap replicates, and an
488 estimation of the gamma distribution parameter. The resulting phylogenetic tree was visualized
489 using MEGA7⁹⁷.

490 **Acknowledgements**

491 We would like to thank Philippe Potin, Mark Cock, Susanna Coelho, Florian Maumus, and Olivier
492 Panaud for helpful discussions, as well as Gwendoline Andres for help setting up the Jbrowse
493 instance. This work was funded partially by ANR project IDEALG (ANR-10-BTBR-04)

494 “Investissements d’Avenir, Biotechnologies-Bioressources”, the European Union’s Horizon 2020
495 research and innovation Programme under the Marie Skłodowska-Curie grant agreement number
496 624575 (ALFF), and the CNRS Momentum call. Sequencing was performed at the Genomics Unit of
497 the Centre for Genomic Regulation (CRG), Barcelona, Spain.

498 **Author contributions**

499 Conceived the study: SMD, AP, AS, HH, CB, TT. Provided materials: AFP, APL. Performed
500 experiments: SMD, SD, IGN, DM, MMP. Analysed data: SMD, APL, EC, LBG, NP, MA, KA,
501 CHC, JC, AC, LD, SD, CF, AGo, AGr, CH, DJ, HK, XL, GVM, AEM, MM, PP, MMP, ASim, CT,
502 HSY, TT. Wrote the manuscript: SMD, KA, APL, JC, LD, CH, Ago, AGr, GVM, ASim, TT.
503 Revised and approved of the final manuscript: all authors.

504 **Additional Information**

505 **Competing interests**

506 The authors declare no competing interest.

507 **Data availability**

508 Sequence data (genomic and transcriptomic reads) were submitted to the European Nucleotide
509 Archive (ENA) under project accession number PRJEB25230 using the EMBLmyGFF3 script⁹⁸. A
510 JBrowse⁹⁹ instance comprising the most recent annotations is available via the server of the Station
511 Biologique de Roscoff (<http://mmo.sb-roscoff.fr/jbrowseEsu>). The reconstructed metabolic network
512 of *E. subulatus* is available at <http://gem-aureme.irisa.fr/sububftgem>. Additional resources and
513 annotations including a blast server are available at <http://application.sb-roscoff.fr/project/subulatus/index.html>. The complete set of manual annotations is provided in
514 Supporting Information Table S7.
515

516 **References**

- 517 1. Davison, I. R. & Pearson, G. A. Stress tolerance in intertidal seaweeds. *J. Phycol.* **32**, 197–
518 211 (1996).
- 519 2. Steneck, R. S. *et al.* Kelp forest ecosystems: Biodiversity, stability, resilience and future.
520 *Environmental Conservation* **29**, 436–459 (2002).
- 521 3. McHugh, D. J. A guide to the seaweed industry. *FAO Fish. Tech. Pap. (FAO, Rome, Italy)*
522 (2003).
- 523 4. Food and Agriculture Organization of the United Nations, F. Global production statistics
524 1950-2014. (2016). Available at: <http://www.fao.org/fishery/statistics/global-production/en>.
525 (Accessed: 16th September 2016)
- 526 5. Archibald, J. M. The puzzle of plastid evolution. *Curr. Biol.* **19**, R81-8 (2009).
- 527 6. Peters, A. F., Marie, D., Scornet, D., Kloareg, B. & Cock, J. M. Proposal of *Ectocarpus*
528 *siliculosus* (Ectocarpales, Phaeophyceae) as a model organism for brown algal genetics and
529 genomics. *J. Phycol.* **40**, 1079–1088 (2004).
- 530 7. Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity
531 in brown algae. *Nature* **465**, 617–21 (2010).
- 532 8. Heesch, S. *et al.* A sequence-tagged genetic map for the brown alga *Ectocarpus siliculosus*

- 533 provides large-scale assembly of the genome sequence. *New Phytol.* **188**, 42–51 (2010).
534 9. Stache-Crain, B., Müller, D. G. & Goff, L. J. Molecular systematics of *Ectocarpus* and
535 *Kuckuckia* (Ectocarpales, Phaeophyceae) inferred from phylogenetic analysis of nuclear- and
536 plastid-encoded DNA sequences. *J. Phycol.* **33**, 152–168 (1997).
537 10. Peters, A. F., Coucerio, L., Tsiamis, K., Küpper, F. C. & Valero, M. Barcoding of cryptic
538 stages of marine brown algae isolated from incubated substratum reveals high diversity.
539 *Cryptogam. Algal.* **36**, 3–29 (2015).
540 11. Ye, N. *et al.* *Saccharina* genomes provide novel insight into kelp biology. *Nat. Commun.* **6**,
541 6986 (2015).
542 12. Nishitsuji, K. *et al.* A draft genome of the brown alga, *Cladosiphon okamuranus*, S-strain: a
543 platform for future studies of ‘mozuku’ biology. *DNA Res.* dsw039 (2016).
544 doi:10.1093/dnares/dsw039
545 13. Nishitsuji, K. *et al.* Draft genome of the brown alga, *Nemacystus decipiens*, Onna-1 strain:
546 Fusion of genes involved in the sulfated fucan biosynthesis pathway. *Sci. Rep.* **9**, 4607
547 (2019).
548 14. Montecinos, A. E. *et al.* Species delimitation and phylogeographic analyses in the *Ectocarpus*
549 subgroup *siliculosi* (Ectocarpales, Phaeophyceae). *J. Phycol.* **53**, 17–31 (2017).
550 15. Harvey, W. H. *Phycologia britannica, or, a history of British sea-weeds: containing coloured*
551 *figures, generic and specific characters, synonymes, and descriptions of all the species of*
552 *algae inhabiting the shores of the British Islands.* (Reeve & Benham, 1848).
553 16. Kützing, F. T. *Phycologia generalis oder Anatomie, Physiologie und Systemkunde der Tange.*
554 (F.A. Brockhaus, 1843).
555 17. Bolton, J. J. Ecoclimatic variation in *Ectocarpus siliculosus* (Phaeophyceae) with respect to
556 temperature growth optima and survival limits. *Mar. Biol.* **73**, 131–138 (1983).
557 18. West, J. & Kraft, G. *Ectocarpus siliculosus* (Dillwyn) Lyngb. from Hopkins River Falls,
558 Victoria - the first record of a freshwater brown alga in Australia. *Muelleria* **9**, 29–33 (1996).
559 19. Dittami, S. M., Heesch, S., Olsen, J. L. & Collén, J. Transitions between marine and
560 freshwater environments provide new clues about the origins of multicellular plants and
561 algae. *J. Phycol.* **53**, 731–745 (2017).
562 20. Oh, D.-H., Dassanayake, M., Bohnert, H. J. & Cheeseman, J. M. Life at the extreme: lessons
563 from the genome. *Genome Biol.* **13**, 241 (2012).
564 21. Dittami, S. M. & Tonon, T. Genomes of extremophile crucifers: new platforms for
565 comparative genomics and beyond. *Genome Biol.* **13**, 166 (2012).
566 22. Dassanayake, M. *et al.* The genome of the extremophile crucifer *Thellungiella parvula*. *Nat.*
567 *Genet.* **43**, 913–918 (2011).
568 23. Amtmann, A. Learning from evolution: *Thellungiella* generates new knowledge on essential
569 and critical components of abiotic stress tolerance in plants. *Mol. Plant* **2**, 3–12 (2009).
570 24. Ma, T. *et al.* Genomic insights into salt adaptation in a desert poplar. *Nat. Commun.* **4**, 2797
571 (2013).
572 25. Zeng, X. *et al.* The draft genome of Tibetan hulless barley reveals adaptive patterns to the
573 high stressful Tibetan Plateau. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1095–1100 (2015).
574 26. Gruber, A., Roca, G., Kroth, P. G., Armbrust, E. V. & Mock, T. Plastid proteome prediction
575 for diatoms and other algae with secondary plastids of the red lineage. *Plant J.* **81**, 519–28
576 (2015).
577 27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
578 BUSCO: assessing genome assembly and annotation completeness with single-copy
579 orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
580 28. Prigent, S. *et al.* The genome-scale metabolic network of *Ectocarpus siliculosus* (EctoGEM):
581 a resource to study brown algal physiology and beyond. *Plant J.* **80**, 367–381 (2014).
582 29. Dittami, S. M. *et al.* Towards deciphering dynamic changes and evolutionary mechanisms
583 involved in the adaptation to low salinities in *Ectocarpus* (brown algae). *Plant J.* **71**, 366–377

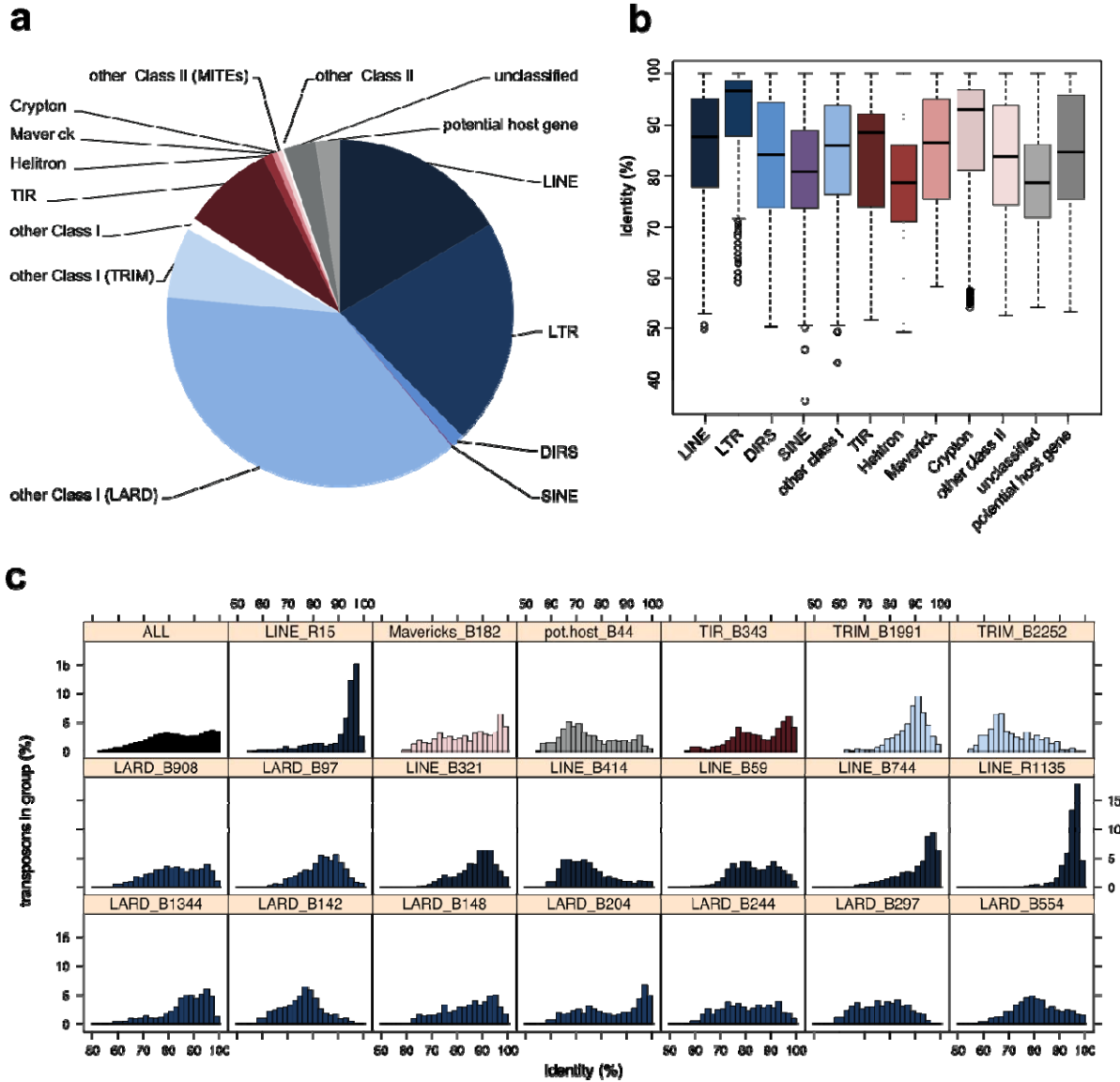
- 584 (2012).
- 585 30. Gualtieri, T., Ragni, E., Mizzi, L., Fascio, U. & Popolo, L. The cell wall sensor Wsc1p is
586 involved in reorganization of actin cytoskeleton in response to hypo-osmotic shock in
587 *Saccharomyces cerevisiae*. *Yeast* **21**, 1107–1120 (2004).
- 588 31. Dong, H.-P. *et al.* High light stress triggers distinct proteomic responses in the marine diatom
589 *Thalassiosira pseudonana*. *BMC Genomics* **17**, 994 (2016).
- 590 32. Zhu, S.-H. & Green, B. R. Photoprotection in the diatom *Thalassiosira pseudonana*: Role of
591 LI818-like proteins in response to high light stress. *Biochim. Biophys. Acta - Bioenerg.* **1797**,
592 1449–1457 (2010).
- 593 33. Dittami, S. M. *et al.* Global expression analysis of the brown alga *Ectocarpus siliculosus*
594 (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to
595 abiotic stress. *Genome Biol.* **10**, R66 (2009).
- 596 34. Wu, H.-J. *et al.* Insights into salt tolerance from the genome of *Thellungiella salsuginea*.
597 *Proc. Natl. Acad. Sci. U. S. A.* **109**, 12219–24 (2012).
- 598 35. de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. Bursts and horizontal evolution
599 of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* **8**, 422
600 (2007).
- 601 36. Hu, T. T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size
602 change. *Nat. Genet.* **43**, 476–81 (2011).
- 603 37. Tan, S. *et al.* LTR-mediated retroposition as a mechanism of RNA-based duplication in
604 metazoans. *Genome Res.* **26**, 1663–1675 (2016).
- 605 38. Müller, D. G., Kapp, M. & Knippers, R. Viruses in marine brown algae. in **50**, 49–67
606 (Academic Press, 1998).
- 607 39. Delaroque, N. *et al.* The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1
608 genome. *Virology* **287**, 112–132 (2001).
- 609 40. Dittami, S. M. *et al.* Microarray estimation of genomic inter-strain variability in the genus
610 *Ectocarpus* (Phaeophyceae). *BMC Mol. Biol.* **12**, 2 (2011).
- 611 41. Liu, M.-J. *et al.* The complex jujube genome provides insights into fruit tree biology. *Nat.*
612 *Commun.* **5**, 5315 (2014).
- 613 42. Kloareg, B. & Quatrano, R. S. Structure of the cell-walls of marine-algae and
614 ecophysiological functions of the matrix polysaccharides. *Ocean. Mar Biol* **26**, 259–315
615 (1988).
- 616 43. Popper, Z. A. *et al.* Evolution and diversity of plant cell walls: from algae to flowering plants.
617 *Annu. Rev. Plant Biol.* **62**, 567–90 (2011).
- 618 44. Torode, T. A. *et al.* Monoclonal antibodies directed to fucoidan preparations from brown
619 algae. *PLoS One* **10**, e0118366 (2015).
- 620 45. Rao, A. Q. *et al.* Genomics of salinity tolerance in plants. in *Plant Genomics* (ed.
621 Abdurakhmonov, I. Y.) 273–299 (InTech, 2016). doi:10.5772/63361
- 622 46. Avia, K. *et al.* High-density genetic map and identification of QTLs for responses to
623 temperature and salinity stresses in the model brown alga *Ectocarpus*. *Sci. Rep.* **7**, 43241
624 (2017).
- 625 47. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. The ankyrin repeat as molecular
626 architecture for protein recognition. *Protein Sci.* **13**, 1435–1448 (2004).
- 627 48. Lipinska, A. P., Van Damme, E. J. M. & De Clerck, O. Molecular evolution of candidate
628 male reproductive genes in the brown algal model *Ectocarpus*. *BMC Evol. Biol.* **16**, 5 (2016).
- 629 49. Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. A molecular timeline
630 for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* **21**, 809–18 (2004).
- 631 50. Charrier, B. *et al.* Development and physiology of the brown alga *Ectocarpus siliculosus*:
632 two centuries of research. *New Phytol.* **177**, 319–32 (2008).
- 633 51. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**,
634 692–702 (2011).

- 635 52. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just
636 orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–
637 413 (2009).
- 638 53. Ritter, A. *et al.* Transcriptomic and metabolomic analysis of copper stress acclimation in
639 *Ectocarpus siliculosus* highlights signaling and tolerance mechanisms in brown algae. *BMC*
640 *Plant Biol.* **14**, 116 (2014).
- 641 54. Starr, R. C. & Zeikus, J. A. UTEX - the culture collection of algae at the University of Texas at
642 Austin: 1993 list of cultures. *J. Phycol.* **29**, 1–106 (1993).
- 643 55. Bothwell, J. H., Marie, D., Peters, A. F., Cock, J. M. & Coelho, S. M. Role of
644 endoreduplication and apomeiosis during parthenogenetic reproduction in the model brown
645 alga *Ectocarpus*. *New Phytol.* **188**, 111–21 (2010).
- 646 56. Le Bail, A. *et al.* Normalisation genes for expression analyses in the brown alga model
647 *Ectocarpus siliculosus*. *BMC Mol. Biol.* **9**, 75 (2008).
- 648 57. Rival, A. *et al.* Variations in genomic DNA methylation during the long-term in vitro
649 proliferation of oil palm embryogenic suspension cultures. *Plant Cell Rep.* **32**, 359–368
650 (2013).
- 651 58. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo
652 assembler. *Gigascience* **1**, 18 (2012).
- 653 59. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-
654 assembled contigs using SSPACE. *Bioinformatics* **27**, 578–9 (2011).
- 655 60. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient
656 and simple way. *PeerJ* **6**, e4958 (2018).
- 657 61. Bosi, E. *et al.* MeDuSa: a multi-draft based scaffold. *Bioinformatics* **31**, 2443–2451 (2015).
- 658 62. Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.*
659 **16**, 3 (2015).
- 660 63. Dittami, S. M. & Corre, E. Detection of bacterial contaminants and hybrid sequences in the
661 genome of the kelp *Saccharina japonica* using Taxoblast. *PeerJ* **5**, e4073 (2017).
- 662 64. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data.
663 *PeerJ* **3**, e1319 (2015).
- 664 65. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic
665 phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
- 666 66. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for
667 de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
- 668 67. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element
669 diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
- 670 68. Goff, S. A. *et al.* The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant*
671 *Sci.* **2**, 34 (2011).
- 672 69. Delage, L. *et al.* In silico survey of the mitochondrial protein uptake and maturation systems
673 in the brown alga *Ectocarpus siliculosus*. *PLoS One* **6**, e19540 (2011).
- 674 70. Le Corguillé, G. *et al.* Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus*
675 *vesiculosus*: further insights on the evolution of red-algal derived plastids. *BMC Evol. Biol.* **9**,
676 253 (2009).
- 677 71. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
678 *Nucleic Acids Res.* **35**, 3100–8 (2007).
- 679 72. Bernt, M. *et al.* MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol.*
680 *Phylogenet. Evol.* **69**, 313–9 (2013).
- 681 73. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes
682 in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–6 (2004).
- 683 74. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web
684 servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686–9 (2005).
- 685 75. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr.*

- 686 *Bioinform.* **3**, 11 (2008).
- 687 76. Nguyen, V. H. & Lavenier, D. PLAST: parallel local alignment search tool for database
688 comparison. *BMC Bioinformatics* **10**, 329 (2009).
- 689 77. Nielsen, H. Predicting Secretory Proteins with SignalP. in *Protein Function Prediction* (ed.
690 Daisuke Kihara) 59–73 (Springer, 2017). doi:10.1007/978-1-4939-7015-5_6
- 691 78. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO
692 suite. *Nucleic Acids Res.* **36**, 3420–35 (2008).
- 693 79. Lee, E. *et al.* Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*
694 **14**, R93 (2013).
- 695 80. Dunn, N. *et al.* GMOD/Apollo: Apollo2.0.8(JB#d3827c). *Zenodo* (2017).
696 doi:10.5281/ZENODO.1063658
- 697 81. Karp, P. D. *et al.* Pathway Tools version 19.0 update: software for pathway/genome
698 informatics and systems biology. *Brief. Bioinform.* **17**, 877–890 (2016).
- 699 82. de Oliveira Dal’Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M. & Nielsen,
700 L. K. AraGEM, a genome-scale reconstruction of the primary metabolic network in
701 *Arabidopsis*. *Plant Physiol.* **152**, 579–89 (2010).
- 702 83. Loira, N., Zhukova, A. & Sherman, D. J. Pantograph: A template-based method for genome-
703 scale metabolic model reconstruction. *J. Bioinform. Comput. Biol.* **13**, 1550006 (2015).
- 704 84. Prigent, S. *et al.* Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded
705 Genome-Wide Metabolic Networks. *PLOS Comput. Biol.* **13**, e1005276 (2017).
- 706 85. Aite, M. *et al.* Traceability, reproducibility and wiki-exploration for “à-la-carte”
707 reconstructions of genome-scale metabolic models. *PLOS Comput. Biol.* **14**, e1006146
708 (2018).
- 709 86. Cormier, A. *et al.* Re-annotation, improved large-scale assembly and establishment of a
710 catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. *New*
711 *Phytol.* **214**, 219–232 (2017).
- 712 87. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
713 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
714 (2015).
- 715 88. Lipinska, A. P. *et al.* Rapid turnover of life-cycle-related genes in the brown algae. *Genome*
716 *Biol.* **20**, 35 (2019).
- 717 89. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein
718 and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*
719 **39**, W13-7 (2011).
- 720 90. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence
721 alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-12
722 (2006).
- 723 91. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and
724 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–77 (2007).
- 725 92. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–
726 91 (2007).
- 727 93. Dittami, S. M., Michel, G., Collén, J., Boyen, C. & Tonon, T. Chlorophyll-binding proteins
728 revisited--a multigenic family of light-harvesting and stress proteins from a brown algal
729 perspective. *BMC Evol. Biol.* **10**, 365 (2010).
- 730 94. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple
731 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–66 (2002).
- 732 95. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview
733 Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**,
734 1189–91 (2009).
- 735 96. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large
736 phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

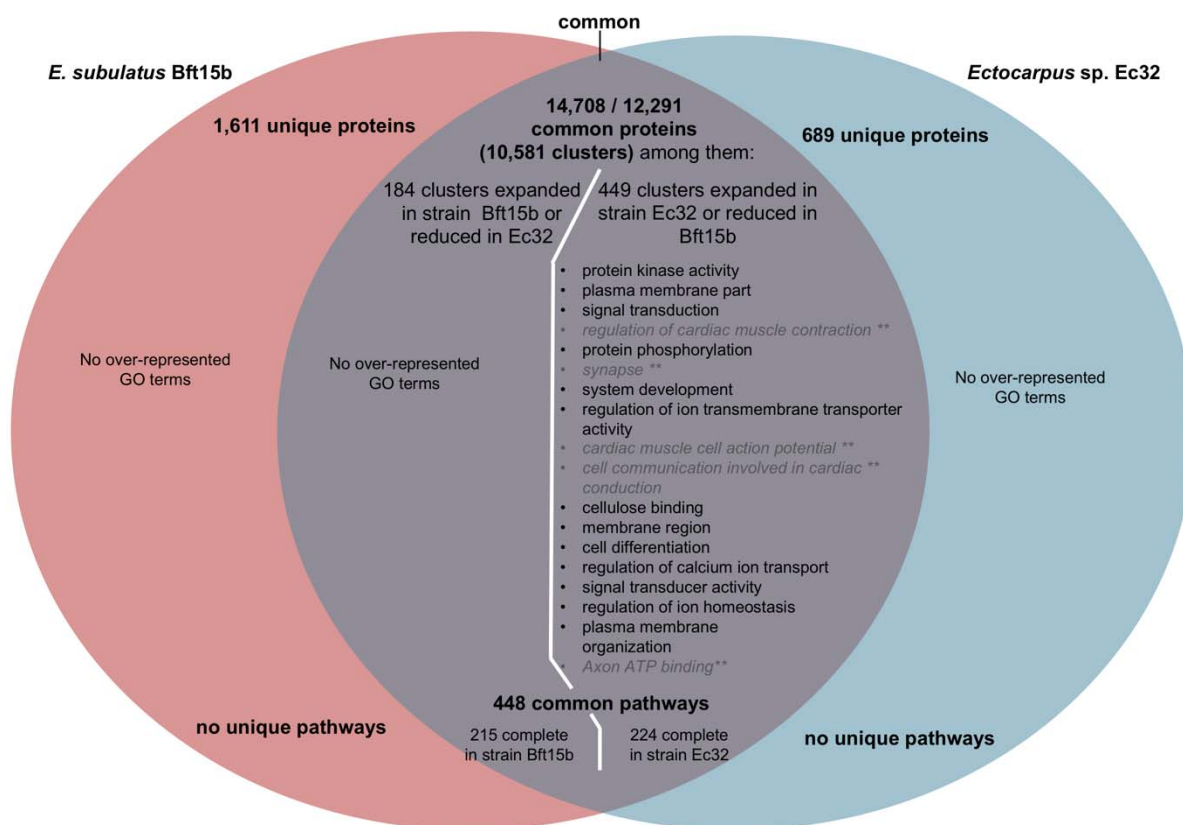
- 737 97. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis
 738 Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
 739 98. Dainat, J. & Gourelé, H. NBISweden/EMBLmyGFF3: EMBLmyGFF3-1.2.2. *Zenodo* (2018).
 740 doi:10.5281/ZENODO.1208446
 741 99. Skinner, M. E., Uzilov, A. V, Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a next-
 742 generation genome browser. *Genome Res.* **19**, 1630–8 (2009).
 743

744 **Figures**



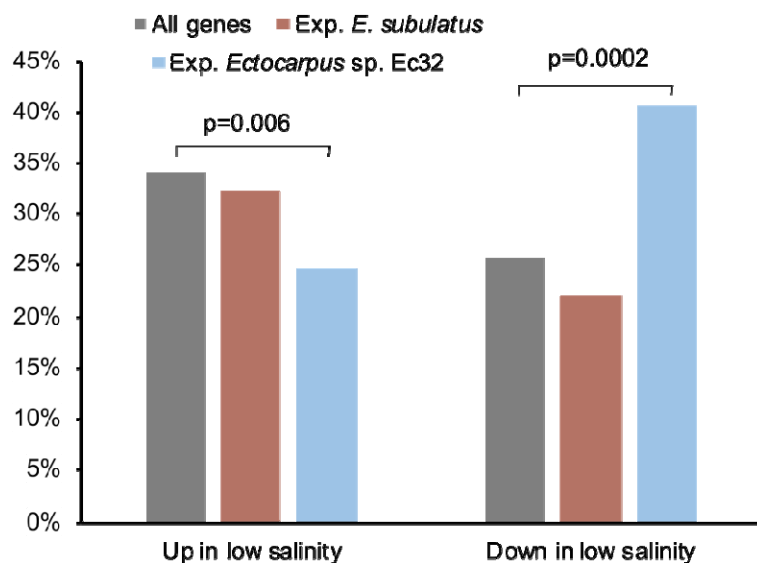
745
 746 **Figure 1:** Repeated elements identified within the genome of *E. subulatus* Bft15b. A) Number of
 747 transposons detected in the different superfamilies; B) Boxplot of sequence identity levels for the
 748 detected superfamilies; and C) Distribution of sequence identities in all and the 20 most abundant
 749 transposon families.

750



751

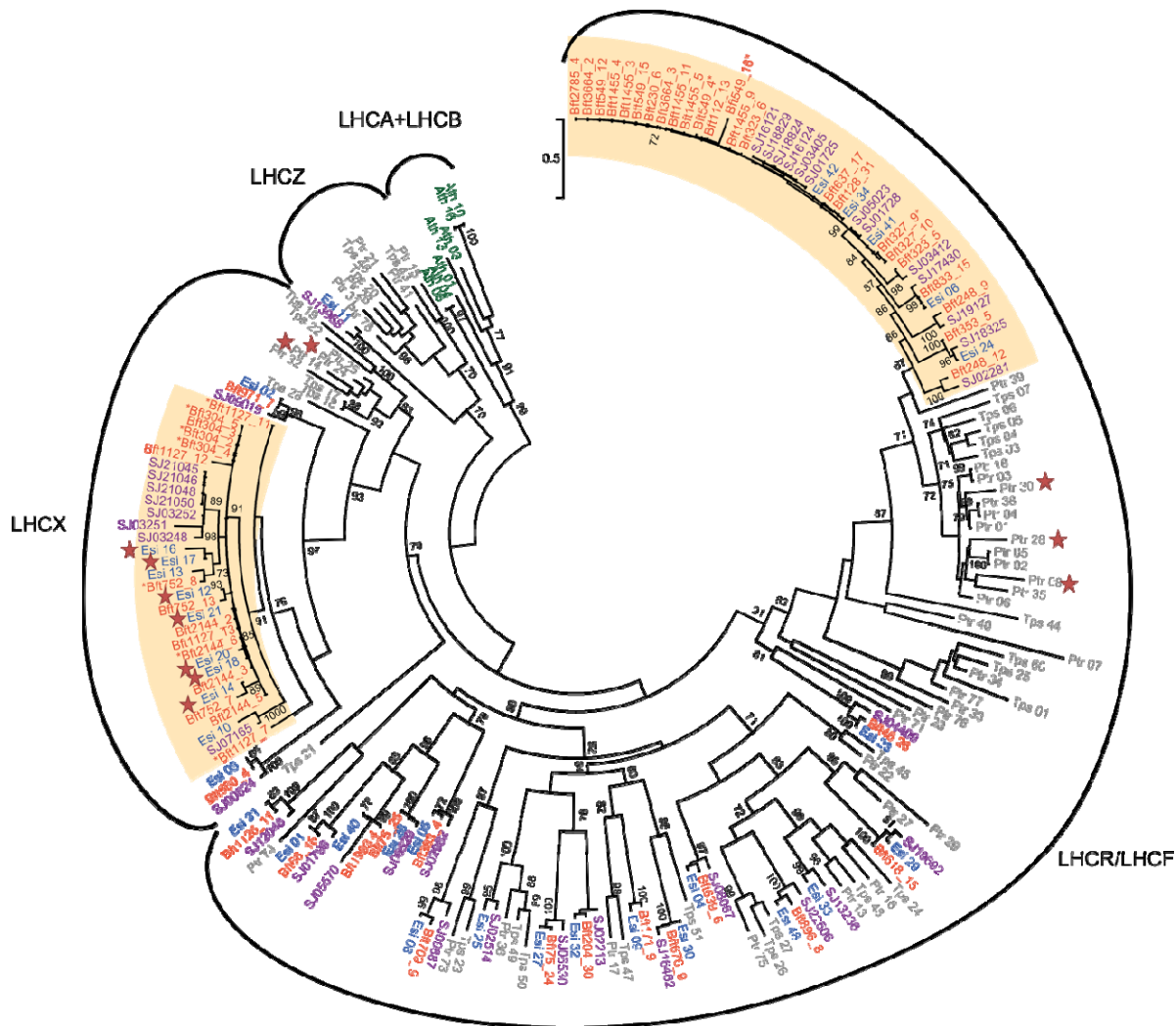
752 **Figure 2:** Comparison of gene content and metabolic capacities of *E. subulatus* Bft15b and
 753 *Ectocarpus* sp. Ec32. The top part of the Venn diagram displays the number of predicted proteins
 754 and protein clusters unique and common to both genomes in the OrthoFinder analysis. The middle
 755 part shows GO annotations significantly enriched ($FDR \leq 0.05$) among these proteins. For the
 756 common clusters, the diagram also contains the results of gene set enrichment analyses for
 757 annotations found among clusters expanded in *E. subulatus* Bft15b and those expanded in
 758 *Ectocarpus* sp. Ec32. Functional annotations not directly relevant to the functioning of *Ectocarpus*
 759 or shown to be false positives are shown in grey and italics. The bottom part shows the comparison
 760 of both genomes in terms of their metabolic pathways.



761

762 **Figure 3:** Percentage of significantly (FDR<0.05) up- and down-regulated genes in *E. subulatus* in
763 response to low salinity (5% seawater). Grey bars are values obtained for all genes with expression
764 data (n=6,492), while brown and blue bars include only genes belonging to gene families expanded
765 in *E. subulatus* Bft15b (n=99) or *Ectocarpus* sp. Ec32 (n=202), respectively (“Exp.” stands for
766 expanded). P-values correspond to the result of a Fisher exact test. Gene expression data were
767 obtained from previous microarray experiments²⁹. Please refer to Supporting information Table S6
768 for additional data.

769



770

771 **Figure 4:** Maximum likelihood tree of chlorophyll binding proteins (CBPs) sequences in *E.*
772 *subulatus* Bft15b (orange) *Ectocarpus* sp. Ec32 (blue), *S. japonica* (purple), and diatoms
773 (*Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, grey). Support values correspond to the
774 percentage of bootstrap support from 1000 replicate runs, only values $\geq 70\%$ are shown. *A. thaliana*
775 sequences (green) were added as outgroup. Accessions for *E. subulatus* Bft15b are given without the
776 Esu prefix; for *Ectocarpus* sp. Ec32, diatoms and *A. thaliana*, see⁹³. Stars indicate genes that have
777 been previously shown to be stress-induced⁹³, asterisks next to the protein names indicate incomplete
778 proteins. Probable expansions in *E. subulatus* Bft15b are indicated by an orange background.

779

780 **Tables**

781 **Table 1:** Assembly statistics of available brown algal genomes. PE = paired-end, MP = mate-pair,
782 n.d. = not determined

	<i>E. subulatus</i> Bft15b	<i>Ectocarpus</i> sp. Ec32 ⁷	<i>S. japonica</i> ¹¹	<i>C. okamuranus</i> ¹²	<i>N. decipiens</i> ¹³
Sequencing strategy	Illumina (PE+MP)	Sanger+Bac libraries	Illumina PE+PacBio	Illumina (PE+MP)	Illumina (PE+MP)
Genome size estimate (flow cytometry)	226	214 ^{6*}	545	140	n. d.
Genome size (assembled)	242 Mb	196 Mb	537 Mb	130 Mb	154 Mb
Genomic Coverage	119 X	11 X [#]	178 X	100 X	420 X
G/C contents	54%	53%	50%	54%	56%
Number of scaffolds >2kb	1,757	1,561	6,985	541	685
Scaffold N50 (kb)	510 kb	497 kb	254 kb	416 kb	1,863 kb
Number of predicted genes	25,893	17,418	18,733	13,640	15,156
Mean number of exons per gene	5.4	8.0	6.5	9.3	11.2
Repetitive elements	30%	30% ^{##}	40%	4.1%	8.8%
BUSCO genome completeness (complete+fragmented)	86% (91% ^{*#})	94% (99% ^{*#})	91% (96% ^{*#})	88% (93% ^{*#})	92% (97% ^{*#})
BUSCO Fragmented proteins	13.5%	7.4%	14.2%	11.9%	5.6%

783 ^{##} 23% according to ⁷, but 30% when re-run with the current version (2.5) of the REPET pipeline.

784 ^{*#} not considering proteins absent from all three brown algal genomes.

785

786 **Table 2:** Clusters of orthologous genes identified by OrthoFinder as expanded in the genome of *E.*
 787 *subulatus* Bft15b or reduced in *Ectocarpus* sp. Ec32, after manual identification of false positives,
 788 and removal of clusters without functional annotation or related to transposon or viral sequences.

Cluster(s)	# Ec32	# Bft15b	Putative annotation or functional domain
<i>Cell-wall related proteins</i>			
OG0000597	1	3	Peptidoglycan-binding domain
OG0000284, -782, -118	6	12	Carbohydrate-binding WSC domain
OG0000889	1	2	Cysteine desulfuration protein
OG0000431	1	3	Galactose-3-O-sulfotransferase (partial)
<i>Transcriptional regulation and translation</i>			
OG0000785	1	2	AN1-type zinc finger protein
OG0000059	4	10	C2H2 zinc finger protein
OG0000884	1	2	Zinc finger domain
OG0000766	1	2	DNA-binding SAP domain
OG0000853	1	2	RNA binding motif protein
OG0000171	1	6	Helicase
OG0000819	1	2	Fungal transcriptional regulatory protein domain
OG0000723	1	2	Translation initiation factor eIF2B
OG0000364	2	3	Ribosomal protein S15
OG0000834	1	2	Ribosomal protein S13
<i>Cell-cell communication and signaling</i>			
OG0000967	1	2	Ankyrin repeat-containing domain
OG0000357	2	3	Regulator of G protein signaling domain
OG0000335	2	3	Serine/threonine kinase domain
OG0000291	2	3	Protein kinase
OG0000185	3	4	Octicosapeptide/Phox/Bem1p domain
<i>Others</i>			
OG0000726	1	3	HSP20
OG0000104	1	9	Light harvesting complex protein
OG0000277	3	3	Major facilitator superfamily transporter
OG0000210	2	4	Cyclin-like domain
OG0000721	1	2	Myo-inositol 2-dehydrogenase
OG0000703	1	2	Short-chain dehydrogenase
OG0000749	1	2	Putative Immunophilin
OG0000463	1	3	Zinc-dependent metalloprotease with notch domain

789