Shaping Brown Algal Genomes: Evolution and Regulation of Transposable Elements

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Erica Dinatale

aus Saronno, Italien

Tübingen

2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	07.05.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Susanna M. Coelho
2. Berichterstatter/-in:	Prof. Dr. Ralf J. Sommer

Acknowledgments

At the end of this intellectually demanding and deeply formative journey, I would like to express my sincere gratitude to everyone who contributed to my development as a scientist and, in many ways, as a person.

First and foremost, I am deeply grateful to my advisors, Prof. Dr. Susana M. Coelho and Prof. Dr. Hajk-Georg Drost. Thank you for your scientific vision, for sharing your enthusiasm for science with me, and for your constant encouragement throughout these years. Your high standards, critical thinking, and trust have shaped both this thesis and my approach to science. I am especially thankful for the intellectual freedom you granted me to explore and develop my own ideas, and for your kind guidance during the more challenging moments of this journey. I would also like to express my very special thanks to Dr. Rory Craig and Dr. Claudia Martinho, co-authors of my first first-author paper. Thank you for teaching me how to think and work as a scientist, for sitting next to me during complicated analyses and moments when I felt stuck in my writing, and for teaching me everything I know about transposable elements and small RNAs. Your mentorship went far beyond scientific training. It has been a privilege to work so closely with you, and I deeply miss our daily exchanges now that you have both started new chapters elsewhere in the world. I also sincerely thank the members of my thesis advisory committee, Prof. Dr. Marja Timmermans and Dr. Luisa Pallares, for their insightful suggestions and constructive feedback, which significantly strengthened this work. My gratitude extends to the research support team at the Max Planck Institute. Jeannette, Sibylle, and Susan, thank you for your patience, for reviewing documents countless times, and for your unwavering support of students. I would likely still be (unsuccessfully) navigating administrative webpages without your help. This thesis would not have been possible without the stimulating and collaborative environment of the Coelho and Drost lab. I am grateful to Heike for her essential administrative support, and to Agnes, Aga, Alex, Andrea, Carole, Cecile, Daniel, Doro, Fabian, Jia Xuan, Liping, Masa, Michael, Min, Morgane, Rèmy, Siva, Vika and Viktor for scientific discussions, shared ideas, and daily exchanges that made research even more enjoyable.

Beyond the lab, I am deeply thankful to the colleagues and friends in Tübingen who made these years intellectually enriching and personally meaningful. Whether through discussions on evolutionary biology, manuscript revisions, conference preparations, or simply shared coffee breaks, your presence made a difference.

Thank you, Alice, Josué, Pengfei, Rita, Romy and Sodai for your friendship, for pushing me and making me laugh, for making the lab a safe space. Andrea, thank you for being my rock time and again, and for reminding me I can do this. Anna B. and Anna N., thank you for shining around me and making me shine, thank you for your passion; you inspire me. Thank you Adrian, Linda, Lisa, Radhika, Miriam, Christian,

Marcos, Natalie for the incredible food, the laughters and your constant support. Alena, Kitty, Juli, Leo, Moritz, Linda, Paula and Pietje, thank you for keeping me grounded.

Finally, I would like to thank my friends and family in Italy, who have supported me unconditionally despite the distance and the frequent question of what exactly I was doing so far away. Your patience, understanding, and emotional strength carried me through the inevitable doubts and uncertainties of a PhD. Thank you for being my foundation. Sorry Angiulì for chosing Germany and yes, I will find a job closer to home. Thank you to Tübingen for being a home during these important years.

Completing this thesis marks the end of an important chapter and the beginning of a new one. I am profoundly grateful to everyone who has been part of this journey. I will carry these years and **all of you** with me forever.

Table of contents

Einleitung ____ I

Introduction ____ III

List of publications ____ V

List of abbreviations ____ VI

1. Background ____ 1

1.1 Brown algae in a box ____ 1

1.2 An “interesting type of chromosomal behavior” ____ 2

1.3 Transposable elements influence the host genome ____ 7

1.3 The host genome regulates transposable elements activity ____ 8

2. Objectives of this thesis and expected outcome ____ 11

3. Contributions ____ 13

3.1 Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems ____ 13

3.2 Origin and evolutionary trajectories of brown algal sex chromosomes ____ 15

3.3 Characterization of the transposable element landscape shaping the Ectocarpus genome ____ 17

4. Discussion ____ 19

4.1 The diverse mobilome of brown algae ____ 19

4.2 TE-enrichment on the sex chromosome ____ 20

4.3 Intronic TEs and their relationship to genic sequence ____ 22

4.4 Regulation and repression of transposable elements in Ectocarpus ____ 23

5. Conclusion ____ 26

Bibliography ____ 27

Appendix ____ 33

Einleitung

Obwohl ihre Entdeckung im Jahr 1983 mit dem Nobelpreis ausgezeichnet wurde, sind Transposons (TEs) noch immer nicht vollständig verstanden und werden in den meisten Genomstudien vernachlässigt. Seit ihrer Entdeckung durch Barbara McClintock in den 1940er und 1950er Jahren hat sich die Wahrnehmung dieser mobilen genetischen Sequenzen in der Wissenschaft grundlegend gewandelt, und die daraus resultierenden Forschungsergebnisse haben unser Verständnis der Genomfunktion und -evolution maßgeblich beeinflusst. Wie bereits von Barbara McClintock beschrieben, wird immer deutlicher, dass TEs eine zentrale Rolle bei der Genomplastizität und anderen Evolutionsprozessen in eukaryotischen Genomen spielen: Sie beeinflussen die Genexpression, die Geschlechtsbestimmung und die phänotypische Variation. Diese Beobachtungen basieren jedoch oft auf gut untersuchten pflanzlichen oder tierischen Modellorganismen. In unserer Arbeit untersuchen wir das Ausmaß der Vielfalt und die Rolle von TEs in den Genomen von Braunalgen. Diese Gruppe von Organismen hat sich über eine Milliarde Jahre lang unabhängig von Tieren und Pflanzen entwickelt und stellt die drittkomplexeste evolutionäre Linie auf dem Planeten dar. Braunalgen weisen eine faszinierende Vielfalt an Entwicklungsmustern, Lebenszyklen und Geschlechtsbestimmungssystemen auf, was sie zu einem spannenden Thema für die Weiterentwicklung unseres Verständnisses der TE-Dynamik macht.

TEs in Braunalgen wurden bereits für den Modellorganismus *Ectocarpus* beschrieben (Cock et al. 2010), aber der Einfluss von TEs auf Genomsequenzen wurde noch nicht im Detail untersucht, sodass noch keine Informationen darüber vorliegen, wie TEs mit benachbarten oder Wirtsgenen interagieren und wie ihre Transposition im Genom moduliert wird. In derselben Studie wurde ein Zusammenhang zwischen small RNAs (sRNAs) und TEs erwähnt, ein Detail, das das Wissen über die vernachlässigbaren 5mC-Methylierungsgrade in Braunalgen erweitert und somit eine Richtung für weitere Studien auf dem Gebiet der TE-Regulierung vorgibt. Später beschrieben Bourdareau et al. 2021 und Gueno et al. 2022 Histonmarker im *Ectocarpus*-Genom und erwähnten H3K79me2 als Histonmarker, der mit einer verringerten Genexpression und TEs assoziiert ist. Diese Studien befassten sich jedoch nicht mit Transposons als

Hauptziel, sodass es möglich ist, unser Wissen in dieser Hinsicht mit einigen technischen Verbesserungen zu verfeinern. Darüber hinaus wurden kürzlich 69 Braunalgen-Genome erstellt und zur Verfügung gestellt (Denoeud et al. 2024), darunter eine qualitativ hochwertigere Version des *Ectocarpus*-Genoms (Liu et al. 2024). Daher können viele fehlende Teile zu diesem Puzzle hinzugefügt werden.

In diesem Zusammenhang ist es das Ziel meines Projekts, die TE-Population in *Ectocarpus* detailliert zu beschreiben, um eine zuverlässige Ressource zu schaffen, die zur Ergänzung oder Verbesserung der TE-Annotation in anderen verfügbaren Braunalgen-Genomen verwendet werden kann. Darüber hinaus untersucht diese Arbeit das Verhalten und die Regulation von TEs innerhalb des Wirtsgenoms, wobei der Schwerpunkt auf den Mechanismen liegt, die die Genomstabilität aufrechterhalten. Die Bemühungen zur Erreichung der oben genannten Ziele sind in drei Publikationen zusammengefasst, darunter eine als Erstautor, die meine Dissertation bilden. Die folgende Abbildung (Fig. 1) fasst den Inhalt der Dissertation zusammen und gibt einen Überblick über meine Ziele und wo diese in meiner Publikationsliste zu finden sind.

Introduction

Despite their discovery being worthy of the Nobel Prize in 1983, transposable elements (TEs) are still not fully understood and are neglected in most genomic studies. Since their discovery by Barbara McClintock in the 1940s and 1950s, the scientific community's perception of these mobile genetic sequences has undergone major paradigm shifts, and the resulting research has had a major impact on our understanding of genome function and evolution. As this great scientist foresaw, it has become increasingly clear that TEs play a central role in genome plasticity and other evolutionary processes in eukaryotic genomes: they influence gene expression, sex determination and phenotypic variation. However, these observations are often based on well-studied plant or animal model organisms. In our work, we are investigating the extent of diversity and the role of TEs in the genomes of brown algae. This group of organisms has evolved independently of animals and plants for over a billion years and represents the third most complex evolutionary lineage on the planet. Brown algae exhibit a fascinating array of developmental patterns, life cycles and sex determination systems, making them an exciting subject for advancing our understanding of TE dynamics.

TEs in brown algae have previously been described for the model species *Ectocarpus* (Cock et al. 2010), but the impact of TEs on genome sequences has not been studied in detail, leaving a lack of information on how TEs interact with neighboring or host genes and how their transposition is modulated in the genome. The same study mentioned a correlation between small RNA (sRNA) and TEs, a detail that adds to the knowledge regarding the negligible levels of 5mC methylation in brown algae, thus providing a direction for further studies in the field of TE regulation. Later on, (Bourdareau et al. 2021) and (Gueno et al. 2022) described histone markers in the *Ectocarpus* genome, mentioning H3K79me2 as a histone marker associated with reduced gene expression and transposable elements. These studies, however, did not address transposons as the main target, so it is possible to refine our knowledge in this regard with some technical improvements. In addition, 69 brown algae genomes have recently been generated and made available (Denoeud et al. 2024), including a higher-quality version of the *Ectocarpus* genome (Liu et al. 2024). Therefore, many missing pieces can be added to this puzzle.

In this context, the aim of my project is to describe the TE population in *Ectocarpus* in detail, in order to generate a reliable resource that can be used to complement or improve TE annotation in other available brown algal genomes. Furthermore, this work explores TE behavior and regulation within the host genome, focusing on the mechanisms that maintain genome stability. The efforts made to achieve the objectives listed above are summarized in three publications, including one first-authorship, which constitute my thesis. The figure below (Fig. 1) summarizes the thesis content and gives an overview of my aims and where to address them in my list of publications.

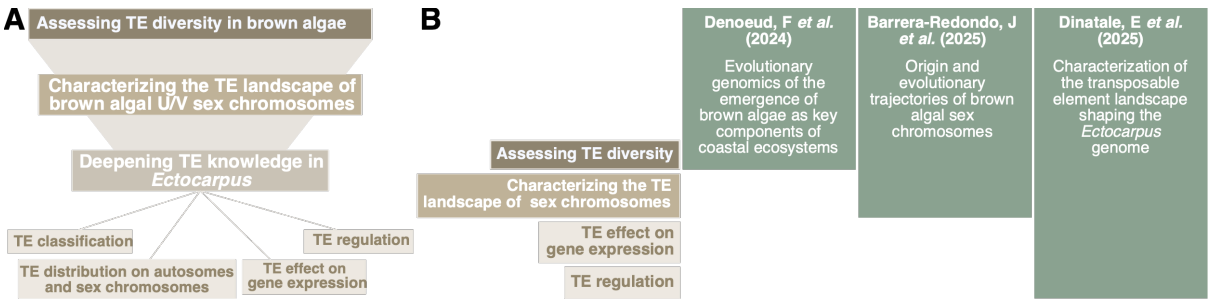


Figure 1 Thesis research aims' structure (A) and their relationship to the research articles (B)

List of publications

†Shared authorship

1. France Denoeud, Olivier Godfroy, Corinne Cruaud, Svenja Heesch, Zofia Nehr, Nachida Tadrent, Arnaud Couloux, Loraine Brillet-Guéguen, Ludovic Delage, Dean Mckeown, Taizo Motomura, Duncan Sussfeld, Xiao Fan, Lisa Mazéas, Nicolas Terrapon, Josué Barrera-Redondo, Romy Petroll, Lauric Reynes, Seok-Wan Choi, Jihoon Jo, Kavitha Uthanumallian, Kenny Bogaert, Céline Duc, Pélagie Ratchinski, Agnieszka Lipinska, Benjamin Noel, Eleanor A. Murphy, Martin Lohr, Ananya Khatei, Pauline Hamon-Giraud, Christophe Vieira, Komlan Avia, Svea Sanja Akerfors, Shingo Akita, Yacine Badis, Tristan Barbeyron, Arnaud Belcour, Wahiba Berrabah, Samuel Blanquart, Ahlem Bouguerba-Collin, Trevor Bringloe, Rose Ann Cattolico, Alexandre Cormier, Helena Cruz de Carvalho, Romain Dallet, Olivier De Clerck, Ahmed Debit, Erwan Denis, Christophe Destombe, **Erica Dinatale**, Simon Dittami, Elodie Drula, Sylvain Faugeron, Jeanne Got, Louis Graf, Agnès Groisillier, Marie-Laure Guillemin, Lars Harms, William John Hatchett, Bernard Henrissat, Galice Hoarau, Chloé Jollivet, Alexander Jueterbock, Ehsan Kayal, Andrew H. Knoll, Kazuhiro Kogame, Arthur Le Bars, Catherine Leblanc, Line Le Gall, Ronja Ley, Xi Liu, Steven T. LoDuca, Pascal Jean Lopez, Philippe Lopez, Eric Manirakiza, Karine Massau, Stéphane Mauger, Laetitia Mest, Gurvan Michel, Catia Monteiro, Chikako Nagasato, Delphine Nègre, Eric Pelletier, Naomi Phillips, Philippe Potin, Stefan A. Rensing, Ellyn Rousselot, Sylvie Rousvoal, Declan Schroeder, Delphine Scornet, Anne Siegel, Leila Tirichine, Thierry Tonon, Klaus Valentin, Heroen Verbruggen, Florian Weinberger, Glen Wheeler, Hiroshi Kawai, Akira F. Peters, Hwan Su Yoon, Cécile Hervé, Naihao Ye, Eric Bapteste, Myriam Valero, Gabriel V. Markov, Erwan Corre, Susana M. Coelho, Patrick Wincker, Jean-Marc Aury, J. Mark Cock. 2024. **Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems.** *Cell,* 187, 6943–6965. [DOI: https://doi.org/10.1016/j.cell.2024.10.049]

2. Josué Barrera-Redondo†, Agnieszka P. Lipinska†, Pengfei Liu, **Erica Dinatale**, Guillaume Cossard, Kenny Bogaert, Masakazu Hoshino, Rory J. Craig,

Komlan Avia, Goncalo Leiria, Elena Avdievich, Daniel Liesner, Rémy Luthringer, Olivier Godfroy, Svenja Heesch, Zofia Nehr, Loraine Brillet-Guéguen, Akira F. Peters, Galice Hoarau, Gareth Pearson, Jean-Marc Aury, Patrick Wincker, France Denoeud, J. Mark Cock, Fabian B. Haas, Susana M. Coelho. 2025. **Origin and evolutionary trajectories of brown algal sex chromosomes.** *Nature Ecology and Evolution*, 9, 2127–2144. [DOI: https://doi.org/10.1038/s41559-025-02838-w]

3. **Erica Dinatale**, Rory J. Craig†, Claudia Martinho†, Hajk Georg Drost, Susana M. Coelho. 2025. **Characterization of the transposable element landscape shaping the Ectocarpus genome.** *Genome Biology,* 26:320. [DOI: https://doi.org/10.1186/s13059-025-03742-z]

List of abbreviations

AGO = Argonaute protein
H3K9me1/2/3 = Mono-, di-, or trimethylation of lysine 9 on histone H3
H3K27me3 = Trimethylation of lysine 27 on histone H3
H3K79me2 = Dimethylation of lysine 79 on histone H3
H4K20me3 = Trimethylation of lysine 20 on histone H4
HGT = Horizontal Gene Transfer
Ks = Synonymous substitution rate
PAR = Pseudo-Autosomal Region
piRNA = PIWI-interacting RNA
PTGS = Post-Transcriptional Gene Silencing
SDR = Sex Determining Region
siRNA = Small interfering RNA
sRNA = Small RNA
TE = Transposable Element
TGS = Transcriptional Gene Silencing
TRG = Taxonomically Restricted Gene
5mC = 5-methylcytosine

1. Background

1.1 Brown algae in a box

Brown algae (Phaeophyceae) are one of the most complex multicellular eukaryotic lineages and emerged approximately 450 million years ago (Mya) during the Great Ordovician Biodiversification Event (GOBE). They belong to the Stramenopiles supergroup, which diverged from plants, animals and fungi over a billion years ago (Denoeud et al. 2024). The study of brown algae has a long history; for instance, the zygotes of Fucales served as models for understanding cell polarization and asymmetric cell division since the 1950s (Jaffe 1958). However, the filamentous alga *Ectocarpus* sp. has emerged as a general model organism for this lineage, supported by the publication of its 214Mbp genome assembly in 2010 (Cock et al. 2010).

Brown algae exhibit a remarkable diversity of morphologies and developmental patterns, ranging from simple uniseriate branched filaments in *Ectocarpus* to complex parenchymatous thalli with multiple cell types and tissue differentiation found in the Fucales and Laminariales (Coelho and Cock 2020; Denoeud et al. 2024). The level of diversity is mirrored in their sexual systems and reproductive strategies. The ancestral state for the lineage is a haplodiplontic life cycle characterized by dioicy, involving an alternation between multicellular haploid gametophyte and diploid sporophyte generations (Coelho and Cock 2020). *Ectocarpus sp.* follows this haplodiplontic cycle and the two generations are morphologically similar (Ahmed et al. 2014). In the gametophyte, sex is determined by the presence of a U (female) or V (male) sex chromosome. These sex chromosomes feature a non-recombining sex-determining region (SDR) that is typically gene-poor and highly repetitive, housing a core set of genes involved in reproduction. The SDR is flanked by pseudo-autosomal regions (PARs) which exhibit intermediate characteristics between the SDR and the autosomes (Ahmed et al. 2014; Luthringer et al. 2015; Luthringer et al. 2024). Many brown algae, including *Ectocarpus*, are able to bypass syngamy and develop a partheno-sporophyte through parthenogenesis (Coelho and Cock 2020).

1.2 An “interesting type of chromosomal behavior”

The scientific Odyssey of transposable elements

Transposable elements (TEs) are repetitive and highly diverse DNA sequences that can mobilize and insert themselves within the genome of a host organism and are transmitted vertically through generations (Wells and Feschotte 2020). Their discovery was made in maize plants thanks to Barbara McClintock, while studying the instability of some kernel mutants she had genetically isolated. Starting in summer 1944, she observed a rupture occurring at a specific locus, named *Dissociation* (*Ds*), on chromosome 9 during the kernel's development, and resulting in variegated pigmentation. She further showed that the breakage of *Ds* was controlled by the presence of the *Activator* (*Ac*) locus, the introduction of which would cause *Ds* to mobilize (i.e. transpose) to another chromosomal position (McClintock 1950). She presented this and other evidence of DNA transposition at the 1951 Cold Spring Harbor Symposium, where they were met with skepticism by the scientific community. Scientists considered this phenomenon a peculiarity of maize, whereas McClintock already believed it to be the same one described in other organisms. Today, we know that TEs are ubiquitous across the tree of life and their proportion in the genome of a host is highly variable even among organisms of similar evolutionary complexity (Feschotte and Pritham 2007; Wells and Feschotte 2020; Hayward and Gilbert 2022). Only a few years later, in 1956, McClintock wrote ‘Control elements are normal components of the chromosomal complement and are responsible for controlling, in a differentiated manner, the timing and type of activity of individual genes’ (Mcclintock 1956): she was aware that she had discovered a mechanism with a functional impact on the genome.

Due to their abundance in host genomes, TEs were among the first DNA sequences sequenced when molecular biology blossomed, and in the 1980s transposition was recognized as a mutagenic force able to shape the genome. However, their presence and activity in the genome was not considered functional, but simply due to their ability to replicate independently of the rest of the genome as parasitic and selfish genetic elements. Their propagation was not thought to bring any benefit to the organism, but at most to have catastrophic consequences on its fitness (Doolittle and Sapienza 1980; Orgel and Crick 1980).

Cold Spring Harbor Laboratory Archives: Barbara McClintock is the only woman to be awarded an unshared Nobel Prize in Physiology or Medicine.

Large-scale genome sequencing revealed that TEs are widespread in genomes and that many are immobile remnants of once-active transposons, having accumulated mutations over time. This led to the concept of transposons as expendable 'junk' DNA, with larger genomes presumed to contain larger graveyards of non-functional sequences. It was only with the advent of functional genomics that it became clear that TEs contribute to the transcriptome and the regulatory apparatus of the genome, and that they can show specificity in terms of developmental stages and tissues (Feschotte 2023).

Today, all these definitions coexist in a more complex and nuanced view of transposable elements. Rather than being merely parasitic, they interact with the host genome in ways that span the spectrum from harmful to symbiotic. As research continues to uncover their roles in the physiology, development, and evolution of species, it is becoming clear that Barbara McClintock's intuition about their significance was ahead of its time.

Transposable elements propagation and evolutionary success

The impact of TEs on genome biology and evolution is fundamentally driven by their capacity to transpose. Transposition allows TEs to increase in copy number, and this proliferative potential is the measure of their success (Finnegan 1989). The diversity of transposition mechanisms is the basis of the traditional classification of eukaryotic TEs into two major classes, introduced by David Finnegan in 1989: retrotransposons (class I) and DNA transposons (class II).

TE diversity extends beyond this dichotomy. Each class is subdivided into subclasses (or orders) based on their enzymatic machinery, then into superfamilies defined by phylogenetic relationships, and finally into families (Wells and Feschotte 2020). TE families are defined by the "80-80-80 rule" (Wicker et al. 2007) and are represented by consensus sequences, approximations of the ancestral element that gave rise to

the family. Consensus sequences are the key for annotating individual TEs within the host genome.

80-80-80 rule: insertions belong to the same family if they share at least 80% sequence identity over more than 80% of their length and are longer than 80 bp.

Another important distinction is between autonomous and nonautonomous elements. Autonomous TEs encode the enzymatic machinery required for their own mobilization, whereas nonautonomous elements transpose by diverting the machinery of their autonomous relatives, despite lacking coding capacity (Wells and Feschotte 2020).

Retrotransposons replicate via reverse transcription of an RNA intermediate, producing a cDNA copy that integrates into the genome while the original remains intact; consequently, they are also known as copy-and-paste elements. They include:

RETROTRANSPOSON TYPE	KEY PROTEINS AND FUNCTIONS	REPLICATION AND INTEGRATION MECHANISM	STRUCTURAL FEATURES	EVOLUTIONARY INSIGHTS	EVOLUTIONARY RELATIONSHIPS
LTR ELEMENTS	GAG (CAPSID), POL (PROTEASE, RT, RNASE H, IN)	REVERSE TRANSCRIPTION USING TRNA PRIMER, INTEGRATION VIA IN	LONG TERMINAL REPEATS, POLYCISTRONIC RNA	RELATED TO RETROVIRUSES, RETROVIRAL ORIGIN	NO HOMOLOGS IN BACTERIA/ARCHAEA, EMERGED AFTER EUKARYOTES
NON-LTR ELEMENTS	ORF1 (RNA BINDING, OPTIONAL), ORF2 (ENDONUCLEASE, RT)	TARGET-PRIMED REVERSE TRANSCRIPTION (TPRT), OFTEN 5' TRUNCATION	NO LTRS, INTERNAL POL II PROMOTER	DIVERSE EVOLUTIONARY LINEAGES, SIMPLEST STRUCTURE	RELATED TO BACTERIAL GROUP II INTRONS, DISTINCT IN EUKARYOTES
YR ELEMENTS	TYROSINE RECOMBINASE (YR), RT	REVERSE TRANSCRIPTION WITH CIRCULAR CDNA, INTEGRATION VIA YR	TERMINAL REPEATS (INVERTED OR DIRECT)	RELATED TO LTR ELEMENTS, DISTINCT DUE TO YR	COMMON IN BACTERIA AND ARCHAEA, DISTINCT FROM EUKARYOTIC YR TES
PENELOPE ELEMENTS	GIY-YIG ENDONUCLEASE, RT	LIKELY TPRT, SIMILAR TO NON-LTR ELEMENTS	PSEUDO-LTRS, UNIQUE GIY-YIG DOMAIN	EARLY DIVERGENCE, RELATED TO TELOMERASE, COMMON IN ANIMALS	NO CLEAR BACTERIAL/ARCHAEAL HOMOLOGS, EMERGED IN EUKARYOTES

Table 1 Summary of the key features of retrotransposon subclasses.

DNA transposons, with some exceptions, mobilize via a cut-and-paste or replication-based mechanism, in which the TE sequence is excised and moved to a new genomic location. They include:

DNA TRANSPOSON TYPE	KEY PROTEINS AND FUNCTIONS	REPLICATION AND INTEGRATION MECHANISM	STRUCTURAL FEATURES	EVOLUTIONARY INSIGHTS	EVOLUTIONARY RELATIONSHIPS
DDE TRANSPOSONS	DDE TRANSPOSASE (CATALYTIC TRIAD D, D, E)	CUT-AND-PASTE TRANSPOSITION WITH EXCISION AND REINTEGRATION	SHORT TERMINAL INVERTED REPEATS (TIRS)	MOST DIVERSE AND WIDESPREAD, ANCIENT GENE FAMILY	RELATED TO BACTERIAL INSERTION SEQUENCES (IS256, IS630)
CRYPTONS	TYROSINE RECOMBINASE (YR)	NON-REPLICATIVE CUT-AND-PASTE, SIMILAR TO DDE	SHORT TIRS, SINGLE ORF	RARE IN EUKARYOTES, STRUCTURALLY SIMILAR TO INSERTION SEQUENCES	COMMON IN BACTERIA AND ARCHAEA, BUT DISTINCT FROM EUKARYOTIC YR TES
HELITRONS	REP/HEL PROTEIN (HUH ENDONUCLEASE + HELICASE)	ROLLING-CIRCLE (PEEL-AND-PASTE) REPLICATION	NO TIRS, CIRCULAR DSDNA INTERMEDIATE	ABUNDANT IN EUKARYOTES	POSSIBLY DERIVED FROM VIRUSES OR PLASMIDS, DISTINCT FROM BACTERIAL IS91
MAVERICKS/ POLINTONS	PPOLB (TYPE B DNA POLYMERASE), DDE INTEGRASE	SELF-SYNTHESIZING DNA REPLICATION AND INTEGRATION	LARGE (15-20 KB), LONG TIRS, MULTIPLE GENES	RELATED TO DSDNA VIRUSES, POSSIBLE VIROPHAGE ORIGIN	SHARE SIMILARITIES WITH DSDNA VIRUSES, POTENTIAL ENDOGENOUS VIROPHAGES

Table 2 Summary of the key features of DNA transposon subclasses.

TE classification highlights the striking breadth of their genetic structures and enzymatic repertoires, to the point that even families within the same superfamily can behave very differently. This functional and structural diversity reflects the rapid evolutionary turnover of TE sequences, and consequently, the effects of TEs on their hosts are as varied as the TEs themselves.

TEs inhabit virtually all eukaryotic genomes, with only a few known exceptions, and in many species, they constitute the majority of genomic DNA, reaching up to ~85–90% (Hufford et al. 2021; Osmanski et al. 2023). Their broad distribution across multiple eukaryotic supergroups aligns with phylogenomic analyses of conserved TE protein domains, which indicate that the major TE subclasses were already present early in eukaryotic evolution. At the same time, even closely related species can differ

dramatically in both the abundance and diversity of their TEs, underscoring the rapid and lineage-specific dynamics of TE evolution (Wells and Feschotte 2020).

The fate and long-term success of any TE family in a genome depend on several interacting forces: the rate of transposition, the probability that new insertions reach fixation, and the efficiency with which genomes delete or erode TE-derived DNA. Insertion-site preferences further modulate these outcomes (Sultana et al. 2017). Some TEs target genomic “safe havens,” whereas others insert near promoters, within heterochromatin, or into other TEs, with these patterns strongly influencing retention and impact (Xie et al. 2001). Importantly, TE propagation is not limited to vertical inheritance. Horizontal transposon transfer (HTT) occurs across all major TE groups and is especially common among certain DNA transposons, providing an additional means for TEs to invade new genomes (Peccoud et al. 2017; Zhang et al. 2020).

Altogether, the diverse strategies of replication, integration, and dispersal employed by TEs underpin their pervasive presence across eukaryotes and their enduring evolutionary success.

The ongoing discovery of transposable elements

Despite their evolutionary success and pervasive presence across the eukaryotic tree of life, many transposable elements remain undiscovered. TE annotation typically begins with automated pipelines, including increasingly sophisticated tools now incorporating deep learning, that can identify a substantial portion of the repetitive content in a genome. These approaches provide a useful first overview of TE composition and allow for broad comparisons across species, but they show limitations when addressing more detailed biological questions. As previously discussed, TE sequences evolve rapidly, and even their most conserved protein domains can diverge to the point that confident alignment becomes difficult (Wells and Feschotte 2020). As a result, automated pipelines frequently assign a large fraction of repeats to “unclassified” categories.

This large fraction of unclassified TEs does not simply reflect annotation failure. A closer inspection of these sequences reveals highly diverged representatives of known

TE families, whereas in others it uncovers genuinely novel and previously overlooked elements. The continuous release of new and taxonomically diverse genomes and the growing interest in TE biology are expanding this space for discovery, revealing new types of TEs and unusual architectures (Craig 2023; Urquhart et al. 2024). Manual curation of TE sequences remains thus necessary to reveal and capture this diversity. This is particularly true for brown algae, where current annotations contain a high proportion of unclassified repeats and TE repertoires remain far from comprehensively described (Denoeud et al. 2024).

The continued expansion of curated TE protein databases is expected to improve the performance of annotation pipelines, enabling more accurate classification of divergent elements and facilitating the discovery of novel TE families.

1.3 Transposable elements influence the host genome

In line with the ideas introduced by McClintock (McClintock 1984), the evolutionary success of TEs and their pervasive presence in eukaryotic genomes come with important functional consequences. TE mobilization shapes genome architecture, affects gene regulation, and ultimately introduces phenotypic variation within populations, contributing to adaptation.

Gene expression may (or may not) be affected by transposable elements

Many TEs carry their own regulatory sequences (e.g., promoters, enhancers, splice sites, or polyadenylation signals), and insertions can therefore activate, repress, or otherwise modulate the expression of nearby genes (Feschotte and Pritham 2007). Over evolutionary timescales, TE-derived sequences have become integrated into host regulatory networks: more than 25% of cis-regulatory elements in humans and maize originate from TEs (Pehrsson et al. 2019). Even degenerated, immobile TEs can influence gene expression. Their remnants may still contain promoters, splice sites, termination signals, or transcription factor binding motifs that interfere with existing expression patterns or generate new ones (Sammarco et al. 2022). In addition, chromatin marks associated with silenced TEs can spread into neighboring genes, altering their transcription (Hirsch and Springer 2017).

Although insertions within coding regions typically cause loss-of-function alleles, many genomes tolerate high numbers of intronic TE insertions without affecting gene expression (Hirsch and Springer 2017).

Transposable elements contribute to sex determination and phenotypic variation

TE insertions can disrupt coding sequences, rewire regulatory circuits, or modify chromosomal structure, often yielding visible and heritable traits. Several classic genetic mutations have been traced back to TE activity. For example, Mendel's wrinkled peas result from a TE-like insertion in the PsSBEI gene, causing a loss of starch-branching enzyme function and altering seed texture (Feng et al. 2025). Similarly, in Drosophila, TEs also influence specific behavioral phenotypes related to reproduction: while null mutants of the white (w) gene exhibit a failure to copulate successfully, the introduction of a mini-white (mw+) gene (a version carried on transposons/transgenes) rescues this defect (Xiao et al. 2017).

TEs also influence major developmental processes, including the evolution of sex determination systems and sex chromosome structure. One well characterized example of TE-driven sex determination evolution is found in the medaka fish *Oryzias latipes*. The master male sex-determining gene, dmrt1bY, originated from an autosomal duplication (dmrt1a) and its regulatory capacity for male-specific expression was established through the co-option of TE-derived sequences in its upstream region: binding sites essential for its function were acquired from a non-autonomous P element (Izanagi) and a subsequently inserted LINE/Rex1 retroelement (Dechaud et al. 2019).

1.3 The host genome regulates transposable elements activity

As just described, TE mobilization can be both deleterious and beneficial, and is inherently mutagenic. Controlling when and where TEs are expressed is therefore essential for genome stability and organismal survival. Accordingly, eukaryotes have evolved elaborate mechanisms to prevent or control TE proliferation. These

mechanisms frequently involve repressive epigenetic modifications and small RNA (sRNA) molecules (18–30 nucleotides long) with high sequence complementarity to TE-derived sequences, which can function either independently or in concert. Broadly, TE repression operates through two major routes: transcriptional gene silencing (TGS) and post-transcriptional gene silencing (PTGS).

Transcriptional gene silencing

Transcriptional silencing prevents TEs from being transcribed by establishing repressive chromatin states at TE loci. These epigenetic modifications reduce DNA accessibility to RNA polymerase and commonly include DNA methylation (especially 5' methylcystosine; 5mC) and histone post-translational modifications. In plants and animals, canonical repressive histone marks associated with TE silencing include mono-, di- and trimethylation of histone H3 lysine 9 (H3K9me1/2/3) and trimethylation of histone H3 lysine 27 (H3K27me3) (Igolkina et al. 2019; Millán-Zambrano et al. 2022; Hisanaga et al. 2023).

Brown algae show negligible levels of 5mC, including at TE loci (Cock et al. 2010; Bourdareau et al. 2021), and lack classical repressive histone marks discussed above. Instead, the histone modifications H3K79me2 and H4K20me3 correlate with reduced transcript abundance and with repeated elements (Bourdareau et al. 2021; Gueno et al. 2022). Notably, 88.6% of H3K79me2 domains >5 kbp contain at least one 400-bp TE, suggesting a possible TE-repressive role for this mark (Bourdareau et al. 2021). The absence of 5mC is not unprecedented among eukaryotes, as species such as *D. melanogaster* and *C. elegans* also lack canonical cytosine methylation and rely primarily on histone-based and small RNA–mediated mechanisms for TE repression (see Post-transcriptional silencing). These observations point to a broader and still underexplored diversity of transcriptional gene silencing strategies employed by eukaryotic genomes to control transposable elements, and position brown algae as a valuable model system for investigating and expanding our understanding of alternative chromatin-based regulatory mechanisms.

Post-transcriptional silencing

Post-transcriptional silencing (PTGS) targets TE RNA molecules after transcription, preventing their translation or reverse transcription. PTGS typically involves sRNA-guided recognition and degradation of TE transcripts. In flowering plants, sRNAs participate in the establishment of repressive chromatin at TE loci (see Transcriptional gene silencing) and also direct cleavage of TE mRNAs (Tang et al. 2022). In animals, TE transcripts are targeted by small interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs), whereas in ciliates small RNAs guide the programmed elimination of TE-derived sequences from the macronucleus (Allen and Nowacki 2020; Fischer 2024). sRNAs function through Argonaute (AGO) proteins, which bind a single guide strand and mediate sequence-specific repression. In many organisms, AGO family expansion enables specialization of silencing pathways, with distinct sRNA size classes preferentially sorted into different AGO proteins (Baulcombe 2004; Ghildiyal and Zamore 2009; Kim et al. 2009).

In *Ectocarpus*, however, the small RNA machinery appears markedly less diversified. The genome encodes a single AGO protein (AGO1), which predominantly associates with 21-nt sRNAs (Tarver et al. 2015; Bukhanets et al. 2026). These sRNAs are enriched over transposable elements (Cock et al. 2010), yet their uniform size distribution contrasts with the clear size-based pathway differentiation observed in plants and animals. The presence of a single AGO suggests limited functional partitioning of small RNA pathways and complicates mechanistic inference based solely on sRNA length. Although 21-nt sRNAs are commonly associated with PTGS in other systems, in *Ectocarpus* they may also contribute to transcriptional silencing. This absence of canonical size-based pathway specialization underscores the divergence of small RNA regulatory architectures in brown algae and highlights their potential to reveal alternative mechanisms of TE control.

2. Objectives of this thesis and expected outcome

Transposable elements (TEs) populate the genome of every organism, and it became clear how their presence is not only linked to their parasitic behavior, but has been co-opted by the host genomes to play roles in the development, in the physiology and in the evolution of the species. Despite the technical difficulties in annotating TEs in the genomes, they have been described in many model organisms and it has been possible to address their importance in species such as *Arabidopsis thaliana*, *C. elegans*, *Drosophila* and even humans. Despite the growing research around the topic, their diversity, functional impact on the genome and regulation remains understudied, particularly outside well-established model organisms such as plants and animals, where technical difficulties make TE annotation and detection even harder. My research addresses this gap by focusing on TEs in brown algae, a group of complex eukaryotic organisms that have evolved independently for over a billion years and exhibit unique developmental and genomic features, adding one piece to the unfinished puzzle.

The primary aim of my doctoral research was to comprehensively characterize the TE landscape in brown algae and to investigate their regulation and genomic impact in the model species *Ectocarpus*. To address this aim, I first performed large-scale TE annotation across eleven brown algal genomes, defining the composition and distribution of TE superfamilies and contributing to the characterization of TE landscapes in U/V sex chromosomes. This comparative framework allowed me to place TE diversity in brown algae into an evolutionary context and to assess how TE composition varies across species and chromosome types.

Building on this broader perspective, I then focused on *Ectocarpus* to generate a high-quality, manually curated TE reference library. This resource, made publicly available through RepDB, establishes a reliable annotation framework for closely related species and improves the accuracy of TE detection in brown algal genomes. Using this refined annotation, I investigated the behaviour and regulation of TEs within the *Ectocarpus* genome. Specifically, I examined the interaction between TEs and small RNAs (sRNAs), their association with the histone modification H3K79me2, and the potential combined contribution of these pathways to TE repression. In addition, I

assessed the impact of TE insertions on nearby gene expression to better understand how TE activity may influence genome function and stability.

By addressing these aims, my research is expected to establish a robust TE annotation framework for brown algae, facilitating comparative genomic studies within this group, and to enhance our understanding of the regulatory mechanisms that maintain genome integrity in *Ectocarpus* despite the presence of active TEs. This work will contribute to broadening our perspective on TE diversity and regulatory mechanisms across eukaryotes, by illustrating how TEs influence genome architecture and stability in a lineage evolutionarily distinct from the classic dichotomy plant-animal.

3. Contributions

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

France Denoeud, Olivier Godfroy, Corinne Cruaud, Svenja Heesch, Zofia Nehr, […], **Erica Dinatale**, […], Susana M. Coelho, Patrick Wincker, Jean-Marc Aury, J. Mark Cock. 2024. *Cell,* 187, 6943–6965.
[DOI: https://doi.org/10.1016/j.cell.2024.10.049]

Synopsis

This comprehensive comparative genomics study provides insights across the entire evolutionary history of the brown algal lineage, which emerged approximately 450 million years ago during the Great Ordovician Biodiversification Event (GOBE). The initial emergence was marked by an intense period of accelerated genome evolution, characterized by a high rate of gene birth - nearly 30% of brown algal genes could not be traced to ancestors outside the Phaeophyceae - alongside pervasive structural modification through enhanced protein domain rearrangement and increased Horizontal Gene Transfer (HGT), primarily from bacteria. This genomic activity resulted in the acquisition of key metabolic and signaling pathways essential for multicellularity and survival in dynamic coastal environments. A second major phase began around 220 Mya, driving the diversification of the main brown algal orders, which was closely linked to phenotypic divergence (e.g., complex parenchymatous thalli and life cycle shifts). This diversification was primarily facilitated by gene family amplifications and developmental innovations like the acquisition of the intercalary meristem (~190 Mya), critical for the emergence of large, perennial, forest-forming species (kelps). The study also observed that species with diploid life cycles showed increased substitution rates, consistent with the phenotypic masking of deleterious alleles. The study also revealed a pervasive, ancient influence of integrated mobile elements: insertions of large DNA viruses (Phaeovirus family, Nucleocytoviricota) are widespread, found in 67 of 69 analyzed brown algal genomes, and these elements have acted as reservoirs for viral HGT, contributing several classes of membrane-localized histidine kinases to the host genome. Finally, recent genomic analysis within

the genus *Ectocarpus* highlighted ongoing evolutionary processes involving extensive gene flow, hybridization, and introgression between species.

Own contribution

I analyzed the results obtained via REPET, an automated TE annotation pipeline, for ten brown algae genomes, and the resulting TE sequences were used as a reference to mask all other genomes presented in the study with RepeatMasker. I then proceeded to quantify TE coverage among different species. This effort highlighted the correlation between genome size and TE content in brown algae and provided the first comprehensive overview of TE diversity and distribution in the group, ranging from approximately 30% in *Ectocarpus sp*. to approximately 70% in *Fucus* species.

3.2 Origin and evolutionary trajectories of brown algal sex chromosomes

Josué Barrera-Redondo†, Agnieszka P. Lipinska†, Pengfei Liu, **Erica Dinatale**, Guillaume Cossard, Kenny Bogaert, Masakazu Hoshino, Rory J. Craig, Komlan Avia, Goncalo Leiria, Elena Avdievich, Daniel Liesner, Rémy Luthringer, Olivier Godfroy, Svenja Heesch, Zofia Nehr, Loraine Brillet-Guéguen, Akira F. Peters, Galice Hoarau, Gareth Pearson, Jean-Marc Aury, Patrick Wincker, France Denoeud, J. Mark Cock, Fabian B. Haas, Susana M. Coelho. 2025. *Nature Ecology and Evolution*, 9, 2127–2144.

[DOI: https://doi.org/10.1038/s41559-025-02838-w]

†Shared authorship

Synopsis

The comparative genomic analysis carried in this study on nine brown algal species, unveiled the complex evolutionary history of their U/V sex chromosomes, which originated between 450 and 224 million years ago following the suppression of recombination in a region containing the pivotal male-determining locus, MIN. Despite this ancient origin, seven ancestral genes within the sex-determining region (SDR) exhibit remarkable conservation. The evolution of the SDRs was highly dynamic, involving lineage-specific expansions driven by nested inversions and the engulfment of pseudo-autosomal regions (PARs), achieved through gene gains via autosomal translocation and *de novo* birth rather than gene loss. Moreover, the genomic architecture of the sex chromosomes is defined by repetitive elements and TEs; all V chromosomes are significantly repeat-rich and gene-poor compared to autosomes, with SDR size differences strongly correlating with repeat content. TEs conspicuously accumulated in the SDRs following recombination suppression, possibly causing further rearrangements through TE-mediated inversions. The specific types of TEs enriched varied across lineages: “unclassified” TEs were enriched in the PARs and SDRs of the Ectocarpales, while species that underwent genome expansion (*e.g.*, *Undaria pinnatifida*) predominantly accumulated Long Terminal Repeat (LTR) elements. This heterochromatic, TE-rich context led to significantly higher interspecies synonymous substitution rates (Ks) in the U/V chromosomes compared to autosomes, which likely promoted the unexpected enrichment of Taxonomically Restricted Genes

(TRGs) in these sex-linked regions. Finally, the research traced the demise of the U/V system through two distinct transitions: the convergent evolution to monoicy (co-sexuality) occurred when ancestral males acquired U-specific genes via translocation, and the transition to a diploid/dioecious system (as seen in *Fucus serratus*) led to the erosion of the characteristic high repeat density and TRG enrichment, although ancestral male-biased SDR genes, including MIN, retained functional regulatory roles in the new system.

Own contribution

I identified and annotated repetitive elements across the analyzed genomes, including nine brown algal species and one outgroup. The general annotation of TEs was performed using RepeatMasker and RepeatModeler2. To investigate the structural characteristics of sex chromosomes compared to autosomes, I quantified the density of TEs and other repeats across distinct genomic compartments (SDR, PAR, and autosomes). This analysis extended beyond overall repeat density to examine the taxonomic composition of the accumulating TEs in these compartments.

3.3 Characterization of the transposable element landscape shaping the Ectocarpus genome

Erica Dinatale, Rory J. Craig†, Claudia Martinho†, Hajk Georg Drost, Susana M. Coelho. 2025. *Genome Biology,* 26:320.
[DOI: https://doi.org/10.1186/s13059-025-03742-z
†Shared authorship

Synopsis

This study utilized a high-quality, chromosome-level assembly of the model brown alga *Ectocarpus* to thoroughly characterize its mobilome and the regulatory mechanisms controlling TEs. TEs constitute 28% of the Ectocarpus genome, which is at the lower end of the known brown algal range, and are notably dominated by evolutionarily young elements lacking high divergence from consensus sequences, suggesting recent or possibly ongoing activity. DNA transposons represent the most abundant and diverse TE subclass, spanning 7.9% of the genome and including several giant elements, such as those from the KDZ and EnSpm superfamilies. TEs are significantly enriched on the U and V sex chromosomes and the PAR region of the V chromosome is enriched compared to the autosomes. This pattern is hypothesized to be driven by local hopping of abundant DNA transposons originating from the non-recombining SDR. The study also analyzed TE localization relative to genes, finding that intronic TEs, which occupy 20.1% of intronic sequences, are consistently shorter and more degraded than intergenic copies, indicating selective pressure against large insertions in gene bodies. Finally, the research uncovered a complex, multilayered system of TE silencing operating in the absence of canonical epigenetic pathways like DNA methylation and H3K9me/H3K27me histone marks. Intact and potentially active TEs are preferentially associated with small RNAs (sRNAs), and the histone modification H3K79me2, which associates with transcriptional repression, particularly when covering a high proportion of the TE body. Crucially, 73.6% of intact TEs associated with H3K79me2 also exhibit sRNA accumulation, suggesting an intricate interaction between these two silencing mechanisms.

Own contribution

I annotated the most recent version of the *Ectocarpus* genome and characterized in detail its TE landscape, including the composition of the TE population and the distribution of TEs along the genome. I further examined how TE content varies between autosomes and sex chromosomes, with a specific focus on male and female SDRs. In addition, I quantified TE coverage across coding and non-coding regions of the genome and identified two distinct TE populations emerging from these analyses, which I described both quantitatively and qualitatively. To investigate potential regulatory interactions, I integrated sRNA libraries and ChIP-seq datasets and mapped these data onto TE copies to assess whether specific small RNA or chromatin signatures are associated with TE regulation. Finally, I analyzed these patterns across the *Ectocarpus* life cycle to understand how TE regulation may change during development.

4. Discussion

4.1 The diverse mobilome of brown algae

The brown algal mobilome acts as a primary driver of genomic structural diversity and serves as the principal determinant of the wide variation in genome size observed across the lineage **(paper 3.1)**. In larger genomes, such as those found in the Fucales and Laminariales, this expansion is primarily driven by the accumulation of Long Terminal Repeats (LTRs), while the more compact genome of *Ectocarpus* is dominated by DNA transposons **(paper 3.2 and 3.3)**.

The TE landscape of the model species *Ectocarpus* is remarkably diverse, containing nearly 340 TE families from 19 superfamilies, and is characterized by a significant skew toward younger elements. This suggests both (i) recent or ongoing transposition activity and (ii) strong selective pressure to maintain genome compactness by purging older, inactive elements. This active state is evidenced by the fact that over 50% of copies show less than 5% divergence from their consensus and supported by findings that approximately 98% of autonomous TE families are expressed across different life stages **(paper 3.3)**.

The predominance of DNA transposons in *Ectocarpus* became evident after extensive manual curation of the TE library, which substantially reduced the proportion of unclassified elements and revealed that most belonged to DNA transposon lineages. Among these elements, several families (e.g. KDZ, EnSpm, Sola-1) correspond to “giant” TEs (15-30 kb). While such large elements usually pose a high risk of lethal ectopic recombination (Arkhipova and Yushenova 2019), the dominance of the haploid phase in the *Ectocapus* life cycle may reduce the frequency of homologous recombination between chromosomes, providing a genomic context that allows the persistence of these complex elements. In addition, manual curation of TE proteins revealed that Sola-1 elements feature a ParB-like DNA-binding domain together with a DNA methyltransferase domain, suggesting they may have the capacity to epigenetically modify themselves **(paper 3.3)**.

The annotation and manual curation of TEs in *Ectocarpus* also expanded the known diversity of Gypsy LTR retrotransposons, uncovering complex structural events such as integrase-chromatin reader fusion and the acquisition of NADAR and Tudor domains. Notably, the Tudor domains identified in *Ectocapus* represent the first reported TE-encoded Tudor domain. More broadly, these findings highlight TE's capabilities as histone-reading or nucleic acid metabolism, serving as a reminder that TE encoded proteins are not confined to transposition: TEs possess a suite of accessory proteins responsible for their survival in the host genome and are often species-specific **(paper 3.3)**.

Overall, the brown algal mobilome is both rich and diverse. While large-scale comparative efforts, as in **paper 3.1**, have established broad evolutionary patterns, the reliance on automatic annotation pipelines - necessary for the analysis of numerous genomes - has left a substantial fraction of elements poorly characterized. This dataset nevertheless provides a valuable foundation for future work, where manual curation or improved automated classifiers can be applied strategically. The high-confidence TE library generated for *Ectocarpus* in **paper 3.3** already enhances annotation accuracy and comparative analyses across brown algae. However, the rapid evolution of TE proteins and the frequent acquisition of accessory domains indicate that much functional diversity remains to be explored. Characterizing these domains will be essential to understand the full spectrum of TE–host interactions and the broader contributions of transposable elements to genome function in eukaryotes. The currently unclassified TE fraction in brown algae therefore likely represents a reservoir of novel element types and molecular innovations.

4.2 TE-enrichment on the sex chromosome

Transposable elements are not randomly distributed in brown algae genomes, but instead accumulate in specific chromosomal environments and appear to have contributed to the structural and evolutionary diversification of brown algal sex chromosomes **(paper 3.2 and 3.3)**. Comparative analyses across species show that the non-recombining sex-determining region (SDR) is consistently enriched in repetitive sequences, and that differences in SDR size are largely explained by variation in TE content, indicating that TE accumulation has been a primary driver of

SDR expansion over evolutionary time **(paper 3.2)**. Moreover, repeat enrichment is not restricted to the SDR itself. Pseudoautosomal regions (PARs), despite recombining, are also significantly more TE-rich than autosomes **(paper 3.2 and 3.3)**. This pattern is unexpected, if compared to other haploid mating-type systems such as *Microbotryum* (Duhamel et al. 2023), where recombining regions do not show similar repeat accumulation.

The detailed analysis of *Ectocarpus* in **paper 3.3** provides mechanistic insight into this distribution. PARs and SDR display a higher relative abundance of cut-and-paste DNA transposons compared to copy-and-paste elements, with specific DNA transposon superfamilies (e.g., Tc1-Mariner, KDZ and EnSpm) significantly enriched in these compartments. This distribution is consistent with a local hopping model (Muñoz-López and García-Pérez 2010; Skipper et al. 2013), whereby DNA transposons accumulating in the recombination-suppressed SDR act as local sources of secondary insertions into adjacent PARs, thereby elevating TE density independently of recombination suppression alone.

Furthermore, the female SDR is approximately twice the size of the male SDR and contains more abundant and younger TEs, a pattern attributed to recent LTR retrotransposon expansions, particularly of the Gypsy superfamily **(paper 3.3)**. These differences likely reflect lineage-specific insertion histories coupled with differential fixation under suppressed recombination rather than pervasive sex-specific TE activation.

Beyond quantitative enrichment, TE-rich SDRs also display signatures of accelerated molecular evolution. Comparative analyses reveal elevated synonymous substitution rates (Ks) in SDR-linked regions relative to autosomes, consistent with reduced recombination and long-term differentiation of U and V chromosomes **(paper 3.2)**. In parallel, SDRs are enriched in taxonomically restricted genes (TRGs), pointing to ongoing lineage-specific gene innovation within these compartments **(paper 3.2)**. While the precise mechanisms remain to be fully resolved, the repeat-rich context of SDRs is compatible with a contribution of TEs to structural rearrangements and potentially to the emergence or diversification of novel genes.

It is interesting to notice that these genomic signatures can persist even after the functional loss of separate sexes. The monoicous species *D. dudresnayi*, which no longer maintains a functional U/V system, retains the characteristic TE-rich architecture of the ancestral sex chromosome on its U/V-homologue, including high repeat density, reduced coding density, and enrichment in TRGs **(paper 3.2)**. This long-term persistence underscores that TE accumulation is not merely a transient consequence of recombination suppression, but might represent a central force shaping the structural differentiation and evolutionary trajectory of brown algal sex chromosomes.

4.3 Intronic TEs and their relationship to genic sequence

Building on the comparative and chromosomal analyses above, we shift our focus to a finer scale, exploring how transposable elements are distributed within genes and how their insertions shape local genomic architecture in *Ectocarpus* **(paper 3.3)**.

Two major TE populations structure the genome: intergenic elements, which account for 60.8% of total TE bases, and intronic elements, representing 34.1% **(paper 3.3)**. This partitioning occurs in the context of an intron-rich genome, where introns are substantially longer than those of many model organisms with comparable genome sizes and contain exonic splice enhancers similar to those described in mammals (Wu et al. 2013). *Ectocarpus* likely maintains efficient splicing across extended intronic regions.

Despite the high density of intronic TE insertions, there is no strong relationship between intronic TE content and gene expression **(paper 3.3)**, indicating that *Ectocarpus* tolerates a moderately elevated intronic TE load without major transcriptional disruption. However, intronic TEs display clear signatures of selection: they are shorter, more degraded, and less frequently intact than intergenic TEs, and are predominantly composed of intrinsically short families such as LINEs, LINE-dependent non-autonomous elements, and non-autonomous DNA transposons. In contrast, the longest TE families, including LTR retrotransposons, are preferentially found within *Ectocarpus*-specific genes (e.g evolutionarily young genes), which may experience weaker purifying selection compared to evolutionarily ancient genes **(paper 3.3)**.

A similar selective gradient is evident in intergenic space. Insertions adjacent to genes tend to be shorter and more degraded, and most genes lack intact TEs within 2 kb of their boundaries. Consequently, long and potentially active elements, particularly LTR retrotransposons and autonomous DNA transposons, are largely concentrated in gene-poor repetitive islands that are not associated with centromeric regions **(paper 3.3)**. Similar TE-rich genomic compartments have been described in other eukaryotes and are often linked to rapid evolutionary turnover and structural polymorphism (Schrader et al. 2014; Torres et al. 2021; Fouché et al. 2022). The scarcity of intact TEs near genes may also explain the absence of a detectable relationship between gene expression and distance to the nearest TE, as would be expected if TE-associated repressive chromatin modifications exerted substantial cis effects (Uzunović et al. 2019). Future population-level analyses assessing the impact of rare TE insertions on gene expression will be necessary to resolve this question.

4.4 Regulation and repression of transposable elements *in Ectocarpus*

A particularly intriguing aspect of TE biology in brown algae concerns their regulation, which I explored specifically in *Ectocarpus* **(paper 3.3)**. The effective control of a diverse and transcriptionally active TE landscape occurs in the absence of detectable cytosine methylation, a characteristic shared across the brown algal lineage. Evidence from other eukaryotic systems indicates that TE repression can rely on alternative mechanisms (see 1.3), notably sRNAs and chromatin modifications. In *Ectocarpus*, TE control appears to depend on a multilayered regulatory architecture centered on these components.

sRNAs in *Ectocarpus* are enriched over intact TEs, indicating preferential targeting of potentially active elements. However, this association is not uniform: TE families within the same superfamily can differ markedly in their level of sRNA targeting **(paper 3.3)**. Such heterogeneity suggests that sRNAs act on specific TE subsets rather than on TE type. In addition, sRNAs exhibit consistent and localized positional patterns along TE bodies, supporting a functional role in TE repression rather than random accumulation **(paper 3.3)**. Because 21-nt sRNAs predominate and other canonical sRNA size classes are absent, their precise mode of action remains unresolved without deeper wet lab experiemtns. They may contribute to PTGS, transcriptional

repression, or context-dependent combinations of both mechanisms. Notably, no clear correlation is observed between sRNA abundance and TE gene expression **(paper 3.3)**, arguing against a simple one-to-one relationship between sRNA levels and transcript degradation.

In parallel, approximately one-third of intact TEs are associated with the histone modification H3K79me2 **(paper 3.3)**, a mark linked to reduced transcript abundance in *Ectocarpus* (Bourdareau et al. 2021). H3K79me2 enrichment typically spans the TE body and shows higher coverage over internal coding regions compared to terminal regions. Moreover, it correlates with decreased TE expression when coverage extends across more than 75% of the TE copy **(paper 3.3)**, consistent with a role in transcriptional repression. Importantly, more than 70% of intact TEs associated with H3K79me2 also accumulate sRNAs, yet many sRNA-associated TEs lack the histone mark **(paper 3.3)**. This pattern indicates that the two pathways are not strictly dependent on one another. Rather than forming a single linear cascade, sRNAs and H3K79me2 appear to define overlapping but partially independent layers of TE control.

Across development, the association between sRNAs and TEs remains remarkably stable. Fewer than 2% of TE copies exhibit differential sRNA accumulation between life stages, and the limited variation observed occurs predominantly over non-intact elements. These differences are not strongly biased toward either increased or decreased sRNA levels, indicating that the overall sRNA landscape is maintained throughout the life cycle **(paper 3.3)**. Given that any vegetative gametophyte cell can generate reproductive structures, sustained targeting of intact and potentially active TEs likely contributes to preserving genome integrity during the transition from gametophyte to gamete.

Taken together, these findings reveal a rich and dynamic *Ectocarpus* mobilome embedded within a complex, multilayered regulatory landscape shaped by small RNAs and chromatin modifications. Despite the absence of cytosine methylation and other canonical epigenetic silencing pathways characteristic of animals and plants, effective TE control is maintained. Genome stability in *Ectocarpus* appears to rely on the coordinated yet partially independent interplay between sRNAs and H3K79me2, underscoring the evolutionary flexibility of TE silencing strategies. These observations

not only expand the understanding of TE regulation in a non-model eukaryotic lineage, but also highlight how alternative epigenetic architectures can emerge to ensure genome integrity in the face of a diverse and transcriptionally active mobilome.

5. Conclusion

Transposable elements are major drivers of genome evolution, yet their diversity, organization, and regulation remain poorly characterized outside a handful of model systems. This thesis set out to provide a comprehensive and integrative view of the mobilome in brown algae, a lineage that evolved independently of plants and animals and lacks canonical TE-silencing mechanisms such as cytosine DNA methylation, with a particular focus on the model species *Ectocarpus*.

By combining high-confidence TE annotation, comparative genomics, and regulatory analyses, this work reveals a structurally diverse and evolutionarily dynamic mobilome that contributes substantially to genome architecture. TE expansion emerges as a key determinant of genome size variation across brown algae and plays a central role in shaping chromosomal regions such as sex chromosomes. At the gene scale, the coexistence of high intronic TE density with selective depletion of intact elements near genes reflects a balance between genome plasticity and purifying selection. Most notably, despite the absence of classical epigenetic repression pathways, *Ectocarpus* maintains effective control over a dense and active TE population through a multilayered system involving small RNAs and H3K79me2. The stability of this regulatory landscape across developmental stages suggests that sustained TE surveillance is essential for preserving genome integrity throughout the life cycle.

Beyond advancing our understanding of brown algal genome biology, these findings broaden the conceptual framework of TE regulation in eukaryotes. They demonstrate that effective genome defense can be achieved through alternative epigenetic strategies and highlight the evolutionary flexibility of silencing systems. By establishing a robust TE annotation framework and uncovering lineage-specific regulatory features, this thesis provides both a resource and a foundation for future comparative and functional studies.

In the long term, integrating TE dynamics into broader models of genome evolution will be essential for understanding how complex multicellular lineages balance innovation with stability. Brown algae offer a powerful system to explore this balance, and the insights gained here lay the groundwork for dissecting how mobilomes shape, and are shaped by, the evolutionary trajectories of emerging eukaryotic clades.

Bibliography

Ahmed S, Cock JM, Pessia E, Luthringer R, Cormier A, Robuchon M, Sterck L, Peters AF, Dittami SM, Corre E, et al. 2014. A Haploid System of Sex Determination in the Brown Alga Ectocarpus sp. *Curr. Biol.* 24:1945–1957.

Allen SE, Nowacki M. 2020. Roles of Noncoding RNAs in Ciliate Genome Architecture. *J. Mol. Biol.* 432:4186–4198.

Arkhipova IR, Yushenova IA. 2019. Giant Transposons in Eukaryotes: Is Bigger Better? *Genome Biol. Evol.* 11:906–918.

Barrera-Redondo J, Lipinska AP, Liu P, Dinatale E, Cossard G, Bogaert K, Hoshino M, Craig RJ, Avia K, Leiria G, et al. 2025. Origin and evolutionary trajectories of brown algal sex chromosomes. *Nat. Ecol. Evol.* 9:2127–2144.

Baulcombe D. 2004. RNA silencing in plants. *Nature* 431:356–363.

Bourdareau S, Tirichine L, Lombard B, Loew D, Scornet D, Wu Y, Coelho SM, Cock JM. 2021. Histone modifications during the life cycle of the brown alga Ectocarpus. *Genome Biol.* 22:12.

Bukhanets V, Batista RA, Haas FB, Luthringer R, Kushkush J, Zheng M, Hipp K, Alva V, Martinho C, Coelho SM. 2026. Germline fate determination by a single ARGONAUTE protein in *Ectocarpus*. *Proc. Natl. Acad. Sci.* 123:e2518712123.

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury J-M, Badger JH, et al. 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–621.

Coelho SM, Cock JM. 2020. Brown Algal Model Organisms. *Annu. Rev. Genet.* 54:71–92.

Craig RJ. 2023. Replitrons: A major group of eukaryotic transposons encoding HUH endonuclease. *Proc. Natl. Acad. Sci.* 120:e2301424120.

Dechaud C, Volff J-N, Schartl M, Naville M. 2019. Sex and the TEs: transposable elements in sexual development and function in animals. *Mob. DNA* 10:42.

Denoeud F, Godfroy O, Cruaud C, Heesch S, Nehr Z, Tadrent N, Couloux A, Brillet-Guéguen L, Delage L, Mckeown D, et al. 2024. Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems. *Cell* 187:6943-6965.e39.

Dinatale E, Craig RJ, Martinho C, Drost H-G, Coelho SM. 2025. Characterization of the transposable element landscape shaping the Ectocarpus genome. *Genome Biol.* 26:320.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601–603.

Duhamel M, Hood ME, Rodríguez de la Vega RC, Giraud T. 2023. Dynamics of transposable element accumulation in the non-recombining regions of mating-type chromosomes in anther-smut fungi. *Nat. Commun.* 14:5692.

Feng C, Chen B, Hofer J, Shi Y, Jiang M, Song B, Cheng H, Lu L, Wang L, Howard A, et al. 2025. Genomic and genetic insights into Mendel's pea genes. *Nature* 642:980–989.

Feschotte C. 2023. Transposable elements: McClintock's legacy revisited. *Nat. Rev. Genet.*:1–4.

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* 41:331–368.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5:103–107.

Fischer SEJ. 2024. Activity and Silencing of Transposable Elements in C. elegans. *DNA* 4:129–140.

Fouché S, Oggenfuss U, Chanclud E, Croll D. 2022. A devil's bargain with transposable elements in plant pathogens. *Trends Genet. TIG* 38:222–230.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10:94–108.

Gueno J, Borg M, Bourdareau S, Cossard G, Godfroy O, Lipinska A, Tirichine L, Cock JM, Coelho SM. 2022. Chromatin landscape associated with sexual differentiation in a UV sex determination system. *Nucleic Acids Res.* 50:3307–3322.

Hayward A, Gilbert C. 2022. Transposable elements. *Curr. Biol.* 32:R904–R909.

Hirsch CD, Springer NM. 2017. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* 1860:157–165.

Hisanaga T, Romani F, Wu S, Kowar T, Wu Y, Lintermann R, Fridrich A, Cho CH, Chaumier T, Jamge B, et al. 2023. The Polycomb repressive complex 2 deposits H3K27me3 and represses transposable elements in a broad range of eukaryotes. *Curr. Biol.* 33:4367-4380.e9.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373:655–662.

Igolkina AA, Zinkevich A, Karandasheva KO, Popov AA, Selifanova MV, Nikolaeva D, Tkachev V, Penzar D, Nikitin DM, Buzdin A. 2019. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. *Cells* 8:1034.

Jaffe LF. 1958. Tropistic responses of zygotes of the Fucaceae to polarized light. *Exp. Cell Res.* 15:282–299.

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 10:126–139.

Liu P, Vigneau J, Craig RJ, Barrera-Redondo J, Avdievich E, Martinho C, Borg M, Haas FB, Liu C, Coelho SM. 2024. 3D chromatin maps of a brown alga reveal U/V sex chromosome spatial organization. *Nat. Commun.* 15:9590.

Luthringer R, Lipinska AP, Roze D, Cormier A, Macaisne N, Peters AF, Cock JM, Coelho SM. 2015. The Pseudoautosomal Regions of the U/V Sex Chromosomes of the Brown Alga Ectocarpus Exhibit Unusual Features. *Mol. Biol. Evol.* 32:2973–2985.

Luthringer R, Raphalen M, Guerra C, Colin S, Martinho C, Zheng M, Hoshino M, Badis Y, Lipinska AP, Haas FB, et al. 2024. Repeated co-option of HMG-box genes for sex determination in brown algae and animals. *Science* 383:eadk5466.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* 36:344–355.

Mcclintock B. 1956. Intranuclear systems controlling gene action and mutation. *Brookhaven Symp. Biol.*:58–74.

McClintock B. 1984. The Significance of Responses of the Genome to Challenge. *Science* 226:792–801.

Millán-Zambrano G, Burton A, Bannister AJ, Schneider R. 2022. Histone post-translational modifications — cause and consequence of genome function. *Nat. Rev. Genet.* 23:563–580.

Muñoz-López M, García-Pérez JL. 2010. DNA transposons: nature and applications in genomics. *Curr. Genomics* 11:115–128.

Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–607.

Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C, et al. 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* 380:eabn1430.

Peccoud J, Loiseau V, Cordaux R, Gilbert C. 2017. Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci. U. S. A.* 114:4721–4726.

Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. 2019. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* 10:5640.

Sammarco I, Münzbergová Z, Latzel V. 2022. DNA Methylation Can Mediate Local Adaptation and Response to Climate Change in the Clonal Plant Fragaria vesca: Evidence From a European-Scale Reciprocal Transplant Experiment. *Front. Plant Sci.* [Internet] 13. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.827166/full

Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, Weichselgartner T, Kemena C, Stökl J, Schultner E, et al. 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.* 5:5495.

Skipper KA, Andersen PR, Sharma N, Mikkelsen JG. 2013. DNA transposon-based gene vehicles - scenes from an evolutionary drive. *J. Biomed. Sci.* 20:92.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* 18:292–308.

Tang Y, Yan X, Gu C, Yuan X. 2022. Biogenesis, Trafficking, and Function of Small RNAs in Plants. *Front. Plant Sci.* 13:825477.

Tarver JE, Cormier A, Pinzón N, Taylor RS, Carré W, Strittmatter M, Seitz H, Coelho SM, Cock JM. 2015. microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga Ectocarpus. *Nucleic Acids Res.* 43:6384–6398.

Torres DE, Thomma BPHJ, Seidl MF. 2021. Transposable Elements Contribute to Genome Dynamics and Gene Expression Variation in the Fungal Plant Pathogen Verticillium dahliae. *Genome Biol. Evol.* 13:evab135.

Urquhart A, Vogan AA, Gluck-Thaler E. 2024. *Starships*: a new frontier for fungal biology. *Trends Genet.* 40:1060–1073.

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. 2019. Transposable Elements Are Important Contributors to Standing Variation in Gene Expression in Capsella Grandiflora. *Mol. Biol. Evol.* 36:1734–1745.

Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* 54:539–561.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8:973–982.

Wu X, Tronholm A, Cáceres EF, Tovar-Corona JM, Chen L, Urrutia AO, Hurst LD. 2013. Evidence for Deep Phylogenetic Conservation of Exonic Splice-Related Constraints: Splice-Related Skews at Exonic Ends in the Brown Alga Ectocarpus Are Common and Resemble Those Seen in Humans. *Genome Biol. Evol.* 5:1731–1745.

Xiao C, Qiu S, Robertson RM. 2017. The white gene controls copulation success in Drosophila melanogaster. *Sci. Rep.* 7:7712.

Xie W, Gai X, Zhu Y, Zappulla DC, Sternglanz R, Voytas DF. 2001. Targeting of the Yeast Ty5 Retrotransposon to Silent Chromatin Is Mediated by Interactions between Integrase and Sir4p. *Mol. Cell. Biol.* 21:6606–6614.

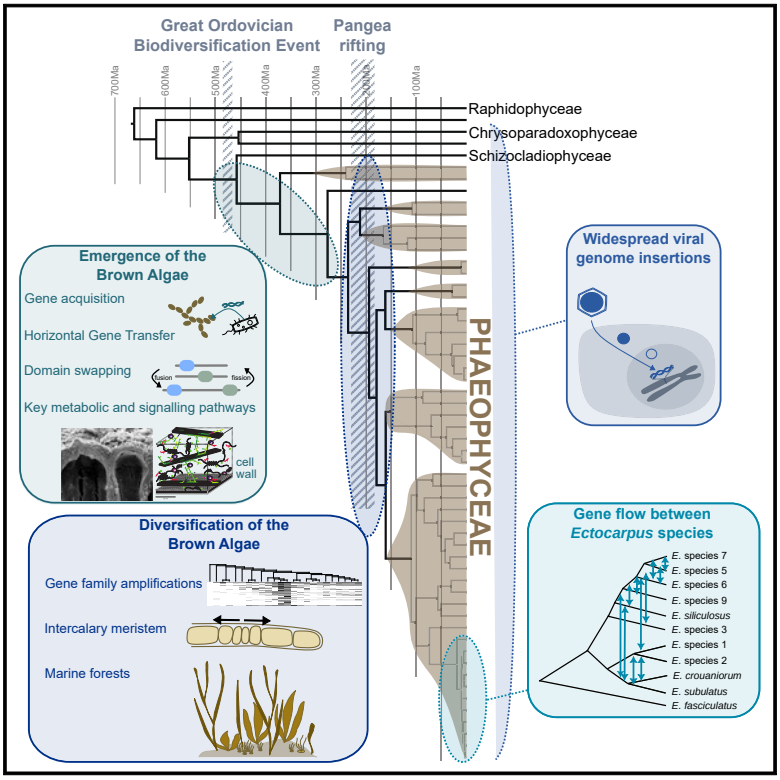
Zhang H-H, Peccoud J, Xu M-R-X, Zhang X-G, Gilbert C. 2020. Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.* 11:1362.

Appendix

Cell

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

Graphical abstract



Highlights

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

Authors

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

Correspondence

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

In brief

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

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

CellPress

Cell

CellPress
OPEN ACCESS

Article

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

France Denoeud,[1,60] Olivier Godfroy,[2,60] Corinne Cruaud,[3,61] Svenja Heesch,[4,51,61] Zofia Nehr,[4,61] Nachida Tadrent,[1,52,61] Arnaud Couloux,[1,61] Loraine Brillet-Guéguen,[5,6,61] Ludovic Delage,[7,61] Dean Mckeown,[6,61] Taizo Motomura,[8,62] Duncan Sussfeld,[1,9,62] Xiao Fan,[10,11,62] Lisa Mazéas,[2,62] Nicolas Terrapon,[12,13,62] Josué Barrera-Redondo,[14,62] Romy Petroll,[14,62] Lauric Reynes,[15,62] Seok-Wan Choi,[16,62] Jihoon Jo,[16,62] Kavitha Uthanumallian,[17,62] Kenny Bogaert,[18,53,62] Céline Duc,[19,62] Pélagie Ratchinski,[4,62] Agnieszka Lipinska,[4,14,62] Benjamin Noel,[1,62] Eleanor A. Murphy,[20,21,62] Martin Lohr,[22,62] Ananya Khatei,[23,54,62] Pauline Hamon-Giraud,[24,62] Christophe Vieira,[25,62] Komlan Avia,[26,62] Svea Sanja Akerfors,[22] Shingo Akita,[27] Yacine Badis,[4] Tristan Barbeyron,[2] Arnaud Belcour,[24,55] Wahiba Berrabah,[1] Samuel Blanquart,[24] Ahlem Bouguerba-Collin,[2] Trevor Bringloe,[28] Rose Ann Cattolico,[29] Alexandre Cormier,[30] Helena Cruz de Carvalho,[31,32] Romain Dallet,[6] Olivier De Clerck,[18]

(Author list continued on next page)

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, Evry 91057, France
[2]Sorbonne Université, CNRS, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[3]Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry 91057, France
[4]Sorbonne Université, CNRS, Algal Genetics Group, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[5]CNRS, UMR 8227, Laboratory of Integrative Biology of Marine Models, Sorbonne Université, Station Biologique de Roscoff, Roscoff, France
[6]CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Roscoff, France
[7]Sorbonne Université, CNRS, UMR 8227, ABIE Team, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[8]Muroran Marine Station, Hokkaido University, Muroran, Japan
[9]Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205, Sorbonne Université, CNRS, Museum, Paris, France
[10]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China
[11]Laboratory for Marine Fisheries Science and Food Production Processes, Laoshan Laboratory, Qingdao, Shandong 266237, China
[12]Aix Marseille University, CNRS, UMR 7257 AFMB, Marseille, France
[13]INRAE, USC 1408 AFMB, Marseille, France
[14]Department of Algal Development and Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany
[15]IRL 3614, UMR 7144, DISEEM, CNRS, Sorbonne Université, Station Biologique de Roscoff, Roscoff 29688, France
[16]Department of Biological Sciences, Sungkyunkwan University, Suwon 16419, Republic of Korea
[17]University of Melbourne, Parkville, VIC, Australia
[18]Phycology Research Group, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium
[19]Nantes Université, CNRS, US2B, UMR 6286, 44000 Nantes, France
[20]University of Bristol, Bristol, UK
[21]Marine Biological Association, Plymouth, UK
[22]Johannes Gutenberg University, Mainz, Germany
[23]Algal and Microbial Biotechnology Division, Nord University, Bodø, Norway

(Affiliations continued on next page)

SUMMARY

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

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Ahmed Debit,[31] Erwan Denis,[1] Christophe Destombe,[15] Erica Dinatale,[14] Simon Dittami,[7] Elodie Drula,[12,13] Sylvain Faugeron,[33] Jeanne Got,[24] Louis Graf,[16] Agnès Groisillier,[19] Marie-Laure Guillemin,[15,34,35] Lars Harms,[36] William John Hatchett,[37] Bernard Henrissat,[38] Galice Hoarau,[37] Chloé Jollivet,[2] Alexander Jueterbock,[23] Ehsan Kayal,[6,56] Andrew H. Knoll,[39] Kazuhiro Kogame,[40] Arthur Le Bars,[6,41] Catherine Leblanc,[7] Line Le Gall,[9] Ronja Ley,[22] Xi Liu,[6] Steven T. LoDuca,[42] Pascal Jean Lopez,[43] Philippe Lopez,[9] Eric Manirakiza,[19] Karine Massau,[6] Stéphane Mauger,[15,57] Laetitia Mest,[4,58] Gurvan Michel,[2] Catia Monteiro,[7] Chikako Nagasato,[8] Delphine Nègre,[6,59] Eric Pelletier,[1] Naomi Phillips,[44] Philippe Potin,[7] Stefan A. Rensing,[45] Ellyn Rousselot,[19] Sylvie Rousvoal,[7] Declan Schroeder,[46] Delphine Scornet,[4] Anne Siegel,[24] Leila Tirichine,[19] Thierry Tonon,[47] Klaus Valentin,[36] Heroen Verbruggen,[28] Florian Weinberger,[48] Glen Wheeler,[21] Hiroshi Kawai,[49,63,*] Akira F. Peters,[50,63,*] Hwan Su Yoon,[16,63,*] Cécile Hervé,[2,63,*] Naihao Ye,[10,11,63,*] Eric Bapteste,[9,63,*] Myriam Valero,[15,63,*] Gabriel V. Markov,[7,63,*] Erwan Corre,[6,63,*] Susana M. Coelho,[14,63,*] Patrick Wincker,[1,63,*] Jean-Marc Aury,[1,63,*] and J. Mark Cock[4,64,*]

[24]University of Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France
[25]Research Institute for Basic Sciences, Jeju National University, Jeju 63243, Republic of Korea
[26]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France
[27]Faculty of Fisheries Sciences, Hokkaido University, Minato-cho 3-1-1, Hakodate, Hokkaido 041-8611, Japan
[28]University of Melbourne, Parkville, VIC, Australia
[29]University of Washington, Seattle, WA, USA
[30]Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, 29280 Plouzané, France
[31]Institut de Biologie de l'ENS (IBENS), Département de Biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[32]Université Paris Est-Créteil (UPEC), Faculté des Sciences et Technologie, 61, Avenue du Général De Gaulle, 94000 Créteil, France
[33]Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile
[34]Núcleo Milenio MASH, Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
[35]Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile
[36]Alfred Wegener Institute (AWI), Bremenhaven, Germany
[37]Nord University, Bodø, Norway
[38]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs Lyngby, Denmark
[39]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[40]Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan
[41]CNRS, Institut Français de Bioinformatique, IFB-core, Évry, France
[42]Department of Geography and Geology, Eastern Michigan University, Ypsilanti, MI 48197, USA
[43]Centre National de la Recherche Scientifique, UMR BOREA MNHN/CNRS-8067/SU/IRD/Université de Caen Normandie/Université des Antilles, Plouzané, France
[44]Biology Department, Arcadia University, Glenside, PA, USA
[45]University of Freiburg, Freiburg im Breisgau, Germany
[46]University of Minnesota, St. Paul, MN, USA
[47]Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, UK
[48]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
[49]Kobe University Research Center for Inland Seas, Kobe, Japan
[50]Bezhin Rosko, 29250 Santec, France
[51]Present address: Applied Ecology & Phycology, Institute for Biosciences, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany
[52]Present address: Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université de Tours, Tours 37200, France
[53]Present address: Department of Algal Development and Evolution, Max Planck Institute for Biology, Tübingen 72076, Germany
[54]Present address: ICAR-Directorate of Coldwater Fisheries Research, Bhimtal, India
[55]Present address: Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
[56]Present address: Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA
[57]Present address: CNRS, La Rochelle Université, UMR7266, Littoral Environnement et Sociétés, La Rochelle, France

(Affiliations continued on next page)

INTRODUCTION

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

CellPress
OPEN ACCESS

[58]Present address: Vegenov, Saint Pol de Léon, France
[59]Present address: Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, Nantes, France
[60]These authors contributed equally
[61]These authors contributed equally
[62]These authors contributed equally
[63]Senior author
[64]Lead contact
*Correspondence: kawai@kobe-u.ac.jp (H.K.), akirapeters@gmail.com (A.F.P.), hsyoon2011@skku.edu (H.S.Y.), cherve@sb-roscoff.fr (C.H.), yenh@ysfri.ac.cn (N.Y.), epbapteste@gmail.com (E.B.), valero@sb-roscoff.fr (M.V.), gabriel.markov@sb-roscoff.fr (G.V.M.), corre@sb-roscoff.fr (E.C.), susana.coelho@tuebingen.mpg.de (S.M.C.), pwincker@genoscope.cns.fr (P.W.), jmaury@genoscope.cns.fr (J.-M.A.), cock@sb-roscoff.fr (J.M.C.)
https://doi.org/10.1016/j.cell.2024.10.049

established as key components of extensive coastal ecosystems. These seaweed-based ecosystems provide high value Earth-system-scale services, including the sequestration of several megatons of carbon per year globally, comparable to values reported for terrestrial forests,[7] but this important role of seaweed ecosystems is threatened by climate-related declines in seaweed populations worldwide.[8] However, appropriate conservation measures, coupled with the development of seaweed mariculture as a highly sustainable and low impact approach to food and biomass production, could potentially reverse this trend, allowing seaweeds to play a significant role in mitigating the effects of climate change.[9] To attain this objective, it will be necessary to address important gaps in our knowledge of the biology and evolutionary history of the brown algal lineage. For example, these seaweeds remain poorly described in terms of genome sequencing due, in part, to difficulties with extracting nucleic acids. The Phaeoexplorer project (https://phaeoexplorer.sb-roscoff.fr/) has generated a large dataset of genome sequences, spanning all the major orders of the Phaeophyceae.[10] This extensive genomic dataset has been analyzed here to study the origin and evolution of key genomic features during the emergence and diversification of this important group of marine organisms.

RESULTS

In-depth sequencing of brown algal genomes

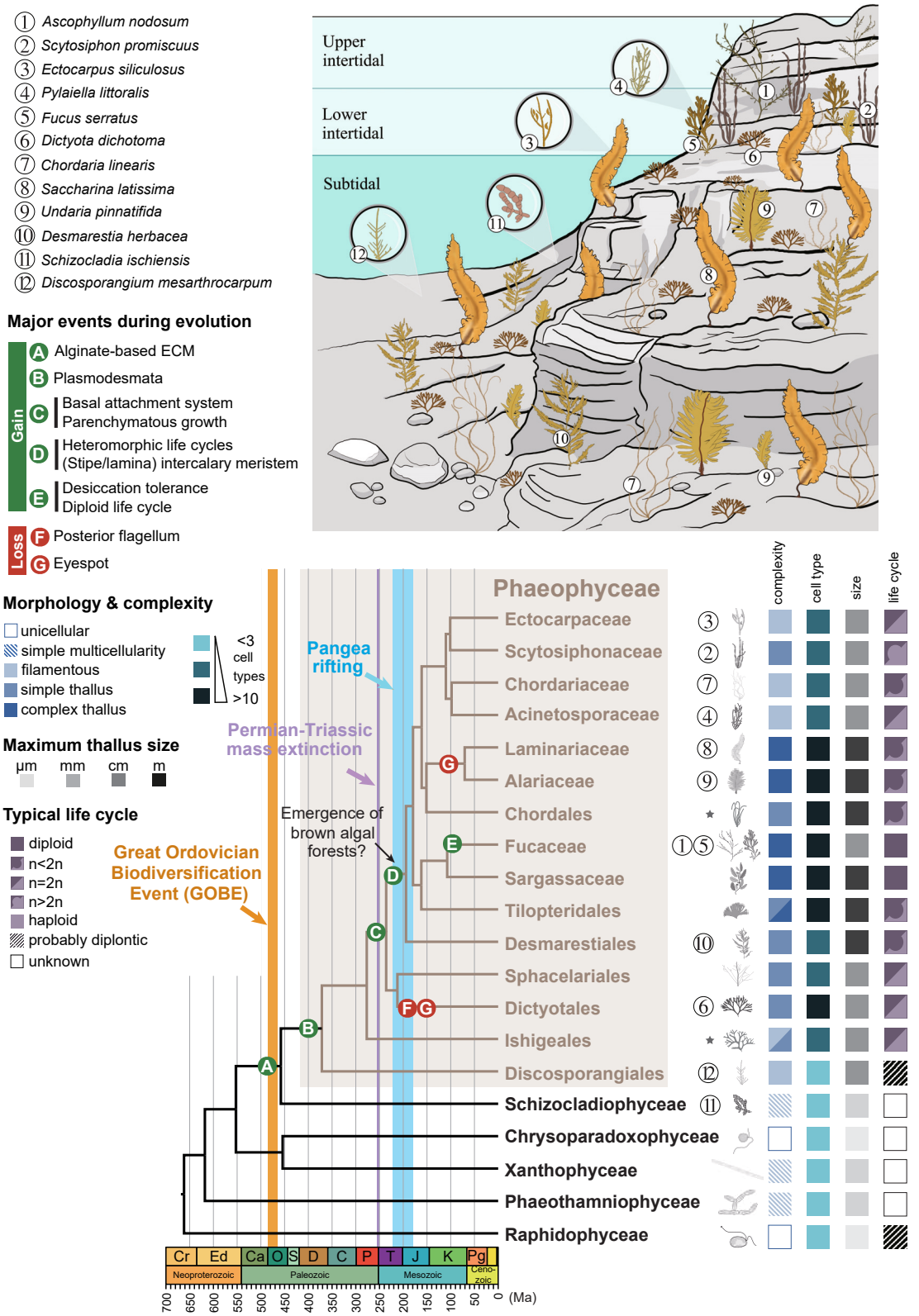
Until now, good quality genome assemblies have been obtained for only five brown algal species,[11–15] together with about 46 draft genome assemblies.[16–20] Here, we report work that has significantly expanded the genomic data available by sequencing and assembling 17 good quality genomes using long-read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

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

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

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

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



(legend on next page)

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

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

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

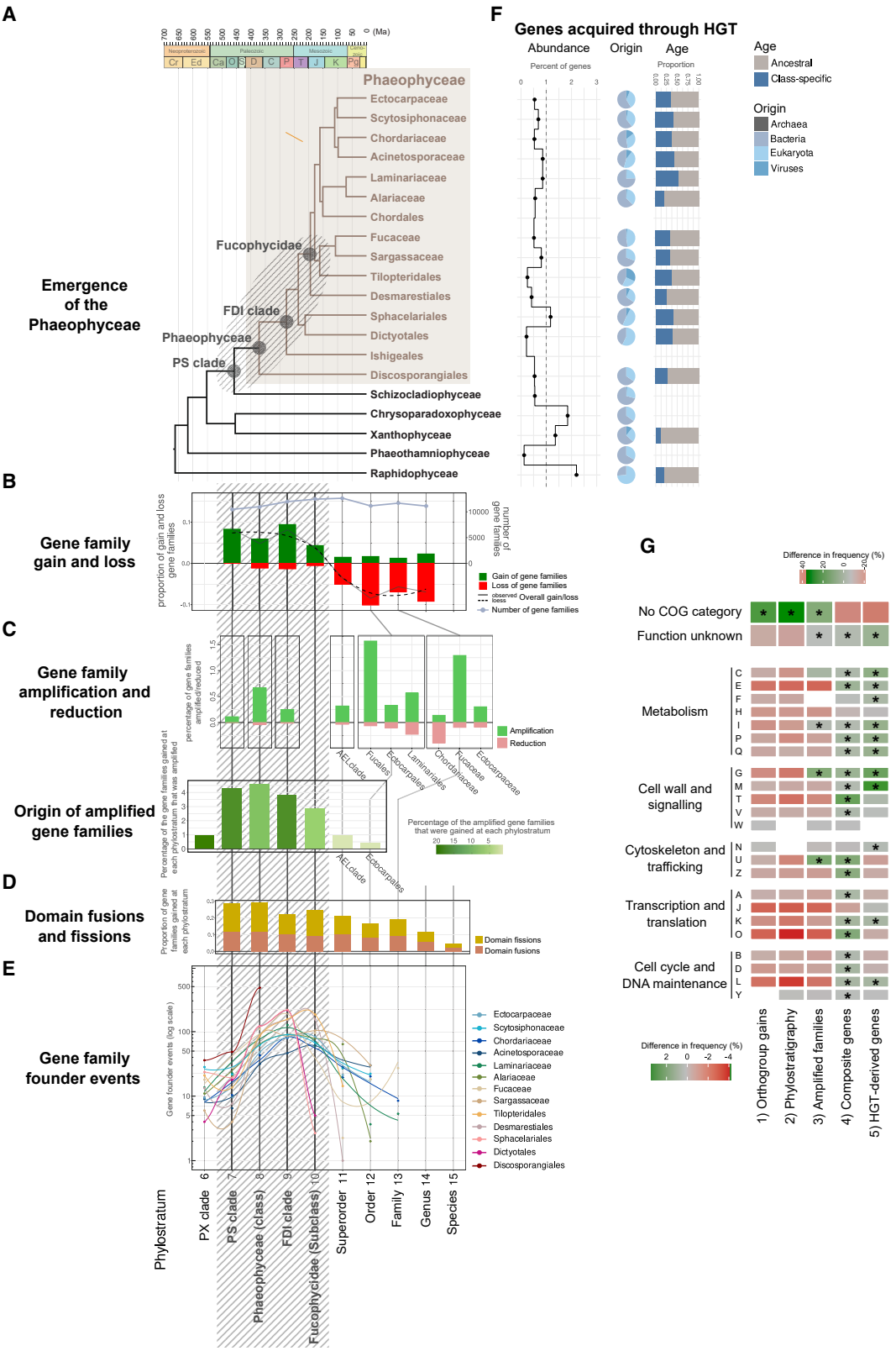
Acquisition of key metabolic and signaling pathways during the emergence of the Phaeophyceae

The success of the brown algae as an evolutionary lineage has been attributed, at least in part, to the acquisition of several key metabolic pathways, particularly those associated with cell-wall biosynthesis, and both halogen and phlorotannin metabolism.[3–5] Large complements of carbohydrate-active enzyme (CAZYme) genes (237 genes on average) were found in all brown algal orders and in their sister taxon *S. ischiensis*, but this class of gene was less abundant in the more distantly related unicellular alga *H. akashiwo* (Figures 3A, S5A, and S5B; Tables S4A and S4B). The evolutionary history of carbohydrate metabolism gene families was investigated by combining information from the genome-wide analyses of gene gain/loss, HGT and gene family amplification (Figure 3B). This analysis indicated that several key genes and gene families (mannuronan C5 epimerase [ManC5-E] and polysaccharide lyase 41 [PL41]) were acquired by the common ancestor of brown algae and *S. ischiensis*, with strong evidence in some cases that this occurred via HGT (PL41). Moreover, marked amplifications were detected for several families (AA15, ManC5-E, GH114, GT23, and PL41), indicating that both gain and amplification of gene families played important roles in the emergence of the brown algal carbohydrate metabolism gene set. Alginate is a major component of brown algal cell walls, and it plays an important role in conferring resistance to the biomechanical effect of wave action.[3] It is therefore interesting that ManC5-E, an enzyme whose action modulates the rigidity of the alginate polymer, appears to have been acquired very early (Figures 3B and 3C). The acquisition of ManC5-E, together with other alginate pathway enzymes such as PL41 (Figures 3A, 3B, and 3D), was probably an important evolutionary step, enabling the emergence of large, resilient substrate-anchored multicellular organisms in the highly dynamic and stressful coastal environment (Figure 1).

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

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

The upper panel indicates approximate positions in the intertidal of key species whose genomes have been sequenced by the Phaeoexplorer project. The lower panel illustrates the diversity of brown algae (maximal values for each taxa) and indicates a number of key evolutionary events that occurred during the emergence of the Phaeophyceae. Some lineages may have secondarily lost a characteristic after its acquisition. Note that members of the genus *Ishige* (Ishigeaceae) also exhibit desiccation tolerance (not shown). ECM, extracellular matrix; asterisk (*), these orders were not analyzed in this study; Cr, Cryogenian; Ed, Ediacaran; Ca, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene.
See also Figures S1, S2, S4, and S5.



(legend on next page)

genes probably diverged from an intermediate-type ancestral gene similar to the *S. ischiensis* genes early during Phaeophyceae evolution. It is likely that the initial acquisitions of algal- and bacterial-type vHPOs represented independent events although the presence of probable vestiges of bacterial-type vHPO genes in *S. ischiensis* means that it is not possible to rule out acquisition of both types of vHPO through a single event.

Gene gain may not, however, have been the proximal factor responsible for all the key metabolic innovations that occurred in the emerging brown algal lineage. Phlorotannins are characteristic brown algal polyphenolic compounds that occur in all Phaeophyceae species, with the exception of some members of the Sargassaceae. Phlorotannins are derived from phloroglucinol and brown algae possess three classes of type III polyketide synthase, two of which (PKS1 and PKS2) were acquired prior to the emergence of the Phaeophyceae and the third (PKS3) evolving much later within the Ectocarpales (Figures 3A and 4A; Table S4E). Interestingly, PKS1 proteins from different brown algal species have been shown to have different activities leading to the production of distinct metabolites,[32–34] indicating that the acquisition of novel functions by this class of enzymes may have played an important role in the emergence of the brown algal capacity to produce phlorotannins. Moreover, many stramenopile PKS type III genes encode proteins with signal peptides or signal anchors (Table S4E). For the brown algae, this feature is consistent with the cellular production site of phlorotannins and the observed transport of these compounds by physodes, secretory vesicles characteristic of brown algae.[35] Cross-linking of phlorotannins, embedded within other brown algal cell-wall compounds such as alginates, has been demonstrated *in vitro* through the action of vHPOs[36–38] and indirectly suggested by *in vivo* observations colocalizing vHPOs with physode fusions at the cell periphery.[39,40] Consequently, vHPOs are good candidates for the enzymes that cross-link phlorotannins and other compounds, perhaps even for the formation of covalent bonds between phloroglucinol monomers and oligomers, which could occur via activation of aromatic rings through halogenation. These observations suggest that the acquisition of vHPOs by the common ancestor of brown algae and *S. ischiensis*, together perhaps with modifications of the existing PKS enzymes, triggered the emergence of new metabolic pathways leading to the production of the phlorotannin molecules characteristic of the Phaeophyceae lineage.

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

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

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

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

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

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

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

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

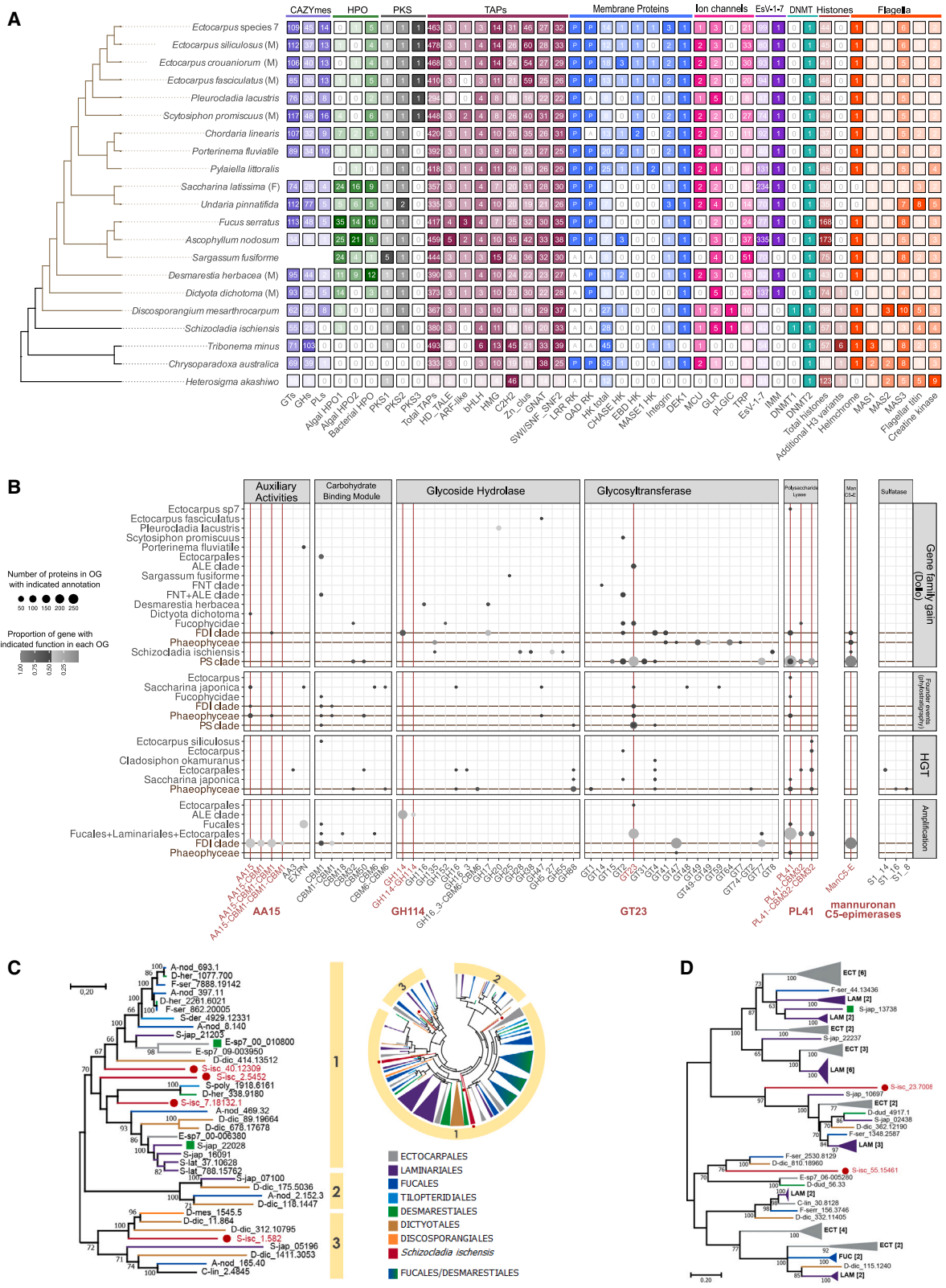
(D) Composite gene analysis. Proportions of gene families showing domain fusion (orange) or domain fission (yellow) at different age strata.

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

(F) Horizontal-gene-transfer-derived genes in orthologous groups and across species. The black trace represents the percentage of genes resulting from HGT events per species. Pie charts summarize the predicted origins (donor taxa) of the HGT genes. The right-hand bar graph indicates the proportions of ancestral (i.e., acquired before the root of the phylogenetic class, in gray) and class-specific (i.e., acquired within the phylogenetic class, in blue) HGT genes.

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

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



(legend on next page)

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

We found indications that the diversification of life cycles, in some cases linked with the emergence of large, complex body architectures, impacted genome evolution through population genetic effects. Most brown algae have haploid-diploid life cycles involving alternation between sporophyte and gametophyte generations, the only exception being the Fucales, which have diploid life cycles. The theoretical advantages of different types of life cycle have been discussed in detail,[48] and one proposed advantage of a life cycle with a haploid phase is that this allows effective purifying counter-selection of deleterious alleles. When the brown algae with haploid-diploid life cycles were compared with species from the Fucales, increased rates of both synonymous and non-synonymous mutation rates were detected in the latter, consistent with the hypothesis that deleterious alleles are phenotypically masked in species where most genes function in a diploid context (Figure S5E). Comparison of non-synonymous substitution rates (dN) for genes in brown algae with different levels of morphological complexity, ranging from simple filamentous thalli though parenchymatous to morphologically complex, indicated significantly lower values of dN for filamentous species (Figure S5E). This observation suggests that the emergence of larger, more complex brown algae may have resulted in reduced effective population sizes and consequently weaker counter-selection of non-synonymous substitutions.[24]

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

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

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

Figure 3. Gene family evolution during the emergence of the brown algal lineage and a focus on carbohydrate metabolism

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

(B) Overview of information from the orthogroup Dollo analysis, the phylostratigraphy analysis, the horizontal gene transfer analysis and the gene family amplification analysis for a selection of cell-wall active protein (CWAP) families. Dots represents functional family/orthogroup couples, with the size being proportional to the number of proteins annotated in the orthogroup (OG), and the color representing the proportion of the functional annotation that falls into this OG. Phaeophyceae plus Schizocladiophyceae (PS) and FDI clade, identified as gene innovation stages, are highlighted in brown. Functional categories with interesting evolutionary histories are highlighted in red.

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

(D) Phylogenetic tree of the polysaccharide lyase 41 (PL41) family. Green squares, biochemically characterized proteins. Brown algal sequences are color-coded in relation to their taxonomy, as indicated in (C). Schizocladiaphyceae sequences are shown in red and with a red circle.

P, present; A, absent; CAZYmes, carbohydrate-active enzymes; HPO, vanadium haloperoxidase; PKS, type III polyketide synthase; TAPs, transcription-associated proteins; EsV-1-7, EsV-1-7 domain proteins; DNMT, DNA methyltransferase; GTs, glycosyltransferases; GHs, glycoside hydrolases; ARF, auxin response factor-related; bHLH, basic-helix-loop-helix; HMG, high mobility group; Zn-clus, zinc cluster; C2H2, C2H2 zinc finger; GNAT, Gcn5-related N-acetyltransferase; SNF2, sucrose nonfermenting 2; LRR, leucine-rich repeat; QAD, β-propeller domain; RK, membrane-localized receptor kinase; HK, histidine kinase; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor 1 domain; DEK1, defective kernel1; MCU, mitochondrial calcium uniporter; GLR, glutamate receptor; pLGIC, pentameric ligand-gated ion channel; TRP, transient receptor potential channel; IMM, IMMEDIATE UPRIGHT; H3, histone H3; MAS, mastigoneme proteins; AA, auxiliary activity; ECT, Ectocarpales; LAM, Laminariales; FUC, Fucales; DES, Desmarestiales.

See also Figure S5.

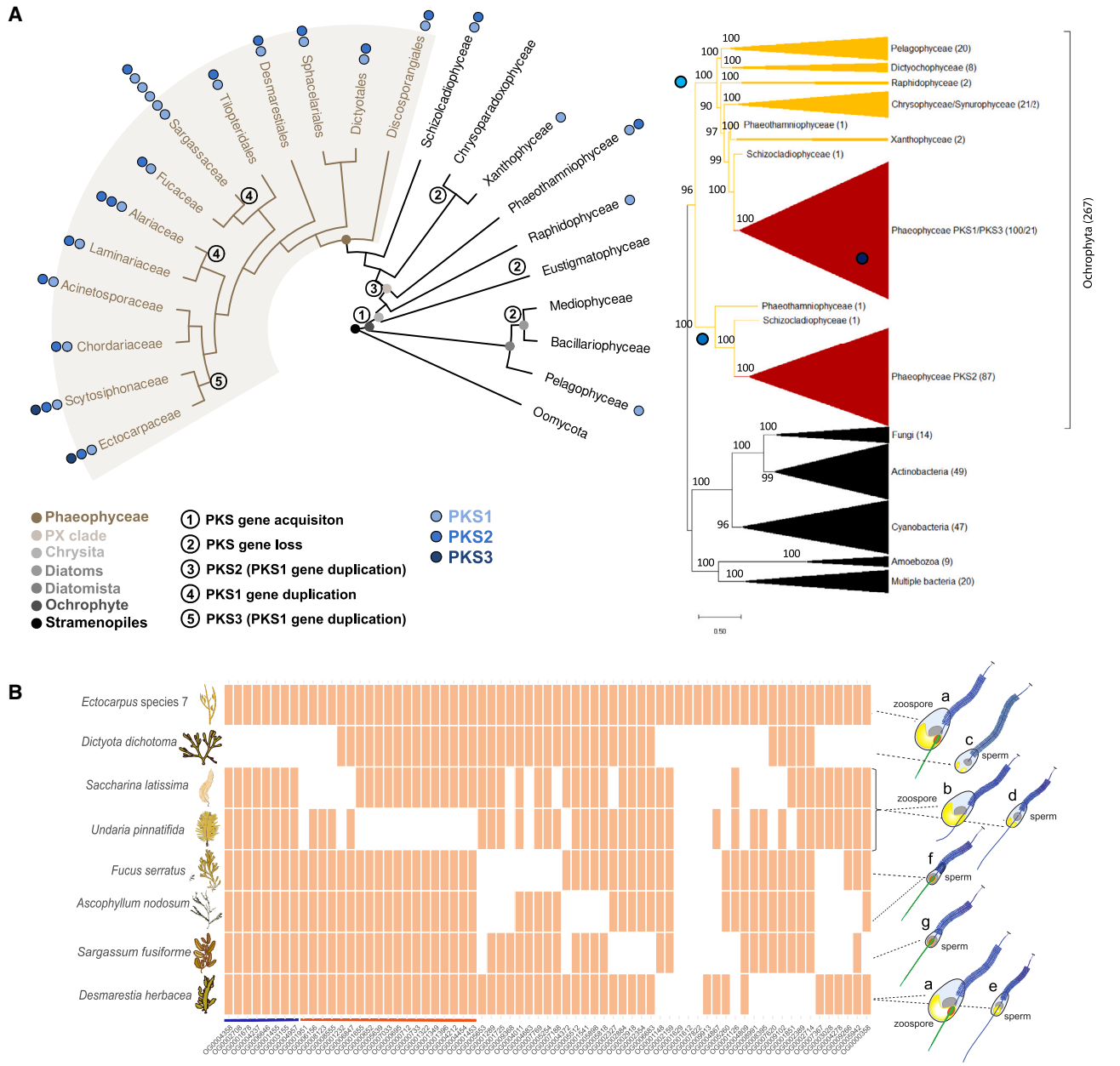


Figure 4. Evolution of key gene families during the emergence of the brown algal lineage

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

(B) Loss of orthogroups corresponding to flagellar proteome components[41] in eight brown algal species from five orders. For the zoid drawings: gray, nucleus; yellow, chloroplast; blue, anterior flagella with mastigonemes; red, eyespot. The posterior flagellum is shown either in green to indicate the presence of green autofluorescence correlated with the presence of the eyespot or in blue in species without an eyespot. Bars below the heatmap indicate gene losses associated with loss of just the eyespot (orange) or of the entire posterior flagellum (blue).

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

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

forests of today, however, seem to have emerged only relatively recently, during the mid-Neogene, following the expansion of cooler water shelf environments.[54,55]

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

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

One of the proposed roles of halogenated molecules in brown algae is in biotic defense[4] and, clearly, an effective defense system would have been an important prerequisite for the emergence of the large, perennial organisms that constitute marine forests. Additional immunity-related families[61] that expanded during the diversification of the brown algae include five orthogroups that contain either GTPases with a central Ras of complex proteins/C-terminal of Roc domain tandem (ROCO GTPases) or nucleotide-binding adaptor shared by apoptotic protease-activating factor 1, R proteins, and CED-4 tetratricopeptide repeat (NB-ARC-TPR) genes (Table S3).

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

Recent evolutionary events within the genus *Ectocarpus*

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

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

Brown algal genomes contain large amounts of inserted viral sequences

A particularly striking result of this study was the identification of extensive amounts of integrated DNA sequence corresponding to large DNA viruses of the *Phaeovirus* family (Figure 6A; Table S5), which integrate into brown algal genomes as part of their lysogenic life cycles.[65] Analysis of 72 genomes in the Phaeoexplorer and associated public genome dataset identified a total of 792 viral regions (VRs) of *Nucleocytoviricota* (NCV) origin in 743 contigs, with a combined length of 32.3 Mbp.

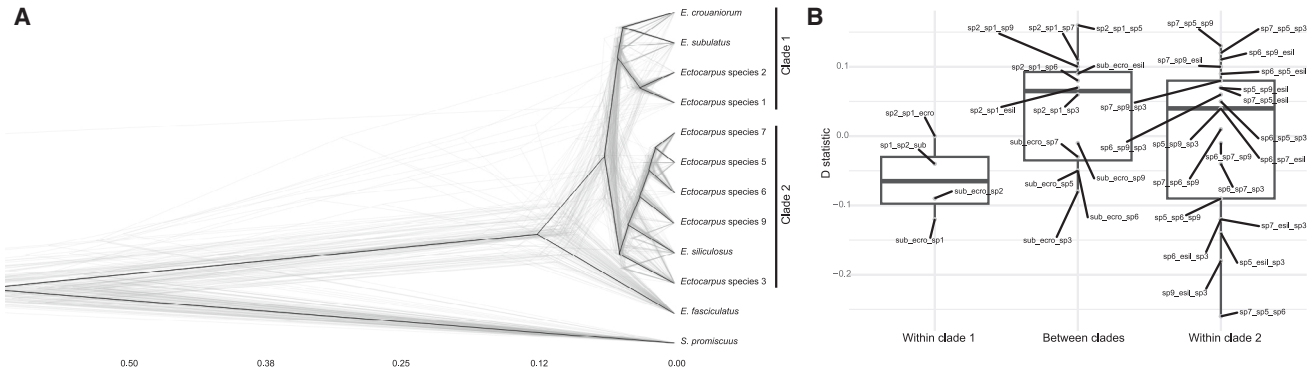


Figure 5. Evidence for gene flow within the genus *Ectocarpus*

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

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

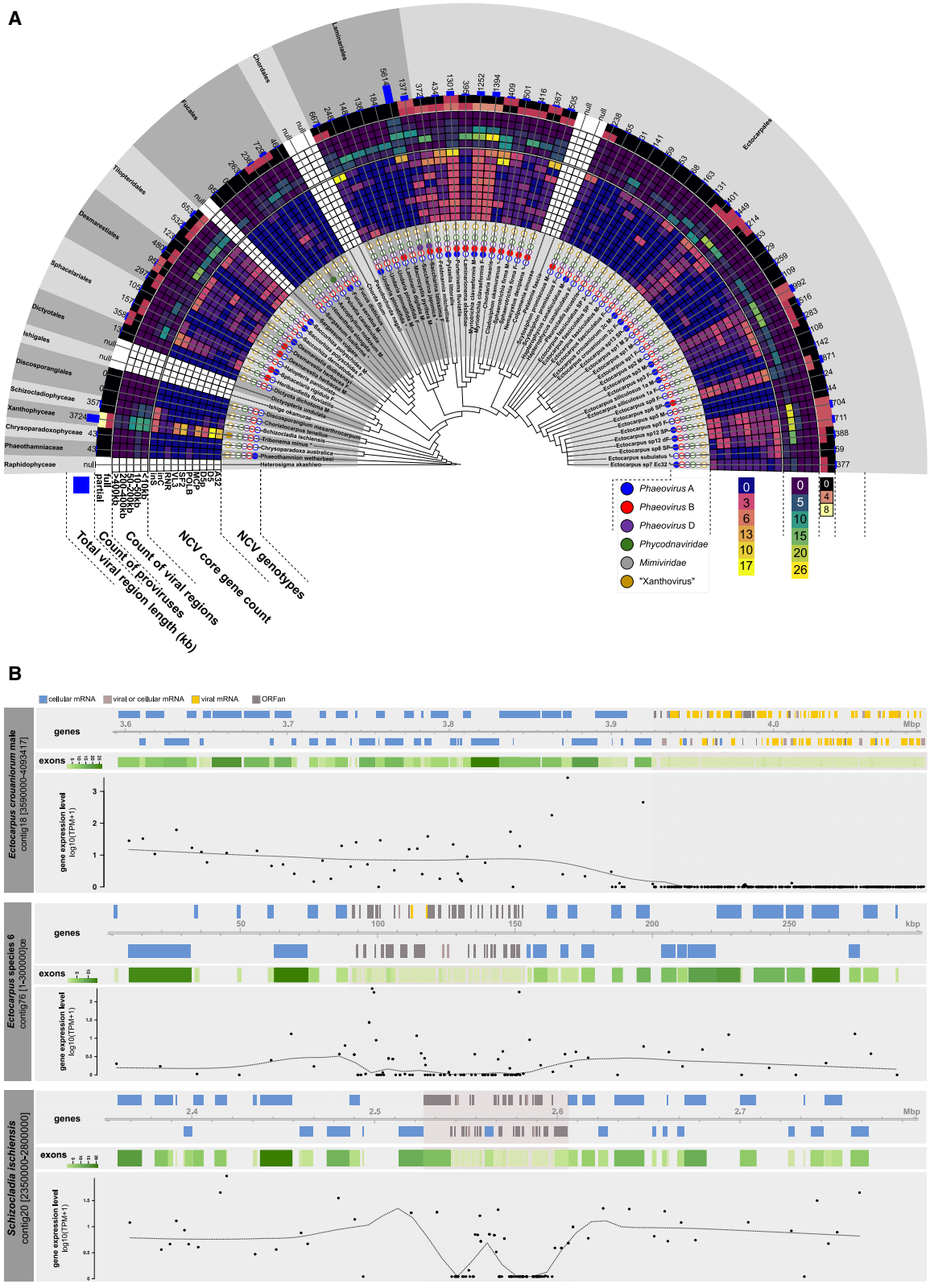
See also Figure S2.

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

On average, each of the 72 analyzed genomes contained 469 kbp of VR (with a maximum of 5,614 kbp) and only two genomes contained no VRs (both from the Discosporangiales). There were a number of outlier genomes that contained more than 1 Mbp of VRs (*T. minus*, *S. latissima*, *S. japonica*, *P. fluviatile*, and *Myriotrichia clavaeformis* male and female). At least one partial provirus (a VR possessing several key NCV marker genes) was present in 39 genomes, 29 of which had at least one full provirus with a complete set of seven key NCV marker genes (Figure 6A; Table S5). In addition to the previously known infections in Ectocarpales[65] and Laminariales,[66] integrated NCV proviruses were found in all Phaeophyceae orders screened, except the Discosporangiales and Dictyotales, and were also detected in *T. minus* (Xanthophyceae). Moreover, NCV marker gene composition indicated that multiple integrated proviruses were present in 16 genomes from multiple Phaeophyceae orders (Ectocarpales, Desmarestiales, Sphacelariales, Tilopteridales, and Laminariales), and the Xanthophyceae (Figure 6A; Table S5). Phylogenetic analysis of the major capsid protein (MCP) and DNA polymerase genes indicated that the majority of the integrated NCVs belonged to the genus *Phaeovirus*, the sole viral group known to infect brown algae (Figures S6A and S6B). However, this analysis also revealed integrated sequences corresponding to other viral groups. Viral sequences in *T. minus* belonged to a putative novel genus closely related to *Phaeovirus*, for which we propose the name *Xanthovirus*. Finally, mimiviridae-related VRs were identified in *S. latissima* and *Pelvetia canaliculata*, but since they are partial proviruses and do not appear to possess integrase genes, they may have originated from ancient endogenization events, similar to those described in chlorophytes.[67]

The identification of integrated NCVs across almost all brown algal orders and in closely related outgroup taxa suggests that the lysogenic life cycle strategy of phaeoviruses is ancient and that giant viral genomes have been integrating into the genomes of brown algae throughout the latters' evolutionary history. This conclusion was supported by the phylostratigraphic analysis, which detected the appearance of many novel virus-related genes dating back to the origin of the Phaeophyceae (Figure S3A). Marked differences were detected in total VR size and NCV marker gene presence across the brown algal genome set, and large differences were even detected between strains from the same genus (between 24 and 992 kbp of VR in different *Ectocarpus* spp. for example; Figure 6A; Table S5). These differences indicate dynamic changes in VR content over evolutionary time, presumably due, at least in part, to differences in rates of viral genome integration, a process that can involve multiple, separate insertion events,[68] and rates of VR loss due to meiotic segregation.[69] In addition, the abundant presence of partial proviruses and NCV fragments in brown algal genomes indicates that inserted VRs can degenerate and fragment, probably also leading to VR loss over time. The identification of large-scale viral genome insertion events over such a long timescale (at least 450 mya[1]) suggests that NCVs may have had a major impact on the evolution of brown algal genomes throughout the emergence of the lineage.

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



(legend on next page)

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

DISCUSSION

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

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

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

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

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

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

Figure 6. Inserted viral regions in brown algal genomes

(A) Annotated phylogeny summarizing key statistics of the presence of *Nucleocytoviricota* (NCV) sequences in the genomes of brown algae and closely related taxa. Eight genomes sourced from public databases are labeled with an asterisk. Outer layers around the tree are as follows (1) NCV genotypes in each genome, (2) NCV core gene count indicates the number of copies of each viral core gene (A32, A32 packaging ATPase; D5/D5p, D5 helicase/primase; MCP, major capsid protein; POLB, DNA polymerase B; SF2, superfamily 2 helicase; VL3, very late transcription factor 3; RNR, ribonucleotide reductase; inC, integrase recombinase; inS, integrase resolvase), (3) count of viral regions is the number of viral regions within each size range category as indicated, (4) count of proviruses is the estimated number of complete or partial integrated viral genomes in a genome, (5) total viral region length is the sum of the lengths in kbp of all viral regions within a genome. The outermost layer indicates the taxonomic class or order of the host clades.

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

See also Figure S6.

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

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

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

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

Data and code availability

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

ACKNOWLEDGMENTS

This work was supported by the France Génomique National Infrastructure project Phaeoexplorer (ANR-10-INBS-09), the European Research Council project Sexsea (638240), the Investissements d'Avenir project Idealg (ANR-10-BTBR-04-01), the European BG-01 BlueGrowth H2020 project Genialg (727892), Laoshan Laboratory grants (LSKJ202203801, LSKJ202203204), the Taishan Scholars Program and China Agriculture Research System (CARS-50), the CNRS international research network DABMA (00022), the ANR projects Epicycle (ANR-19-CE20-0028-01), BrownSugar (ANR-20-CE44-0011), HaloGene (ANR-22-CE20-0025), Seabioz (ANR-20-CE43-0013) and BrownLincs (ANR-23-CE20-0048-01), the National Research Foundation of Korea (2022R1A2B5B03002312, 2022R1A5A1031361) and Ministry of Oceans and Fisheries (KIMST-20180430) granted to H.S.Y., the projects Connect Talent EpiAlg Région Pays de la Loire-Nantes Métropole and Etoiles Montantes M-EpiCC Région Pays de la Loire, the MITI-funded project Algometabionte, Dr. Karl Feldbausch-Stiftung, the CNRS, and Sorbonne University. We are grateful to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), which is part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing both help and computing and storage resources.

AUTHOR CONTRIBUTIONS

Software, L.B.-G., R.D., A.L.B., X.L., D.N., and E.C.; formal analysis, F.D., O.G., L.D., D.M., T.M., D. Sussfeld, X.F., L. Mazéas, N. Terrapon, J.B.-R., R.P., L.R., S.-W.C., J.J., K.U., K.B., C. Duc, P.R., A.L., E.A.M., M.L., A.K., P.H.-G., C.V., S.S.A., S.A., K.A., Y.B., T. Barbeyron, A.B., S.B., A.B.-C., A. Cormier, H.C.d.C., A.D., E. Dinatale, S.D., E. Drula, J.G., L.G., A.G., M.-L.G., L.H., B.H., A.J., E.K., A.H.K., C.L., L.L.G., R.L., S.T.L., P.J.L., E.M., S.M., G.M., C.N., S.A.R., E.R., D. Schroeder, A.S., L.T., T.T., K.V., H.V., G.W., H.K., A.F.P., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.-M.A., and J.M.C.; investigation, C.C., S.H., Z.N., N. Tadrent, A. Couloux, B.N., W.B., E. Denis, C.J., L. Mest, S.R., and D. Scornet; resources, O.G., A. Couloux, L.B.-G., T. Bringloe, R.A.C., C. Destombe, S.F., W.J.H., G.H., K.K., A.L.B., K.M., C.M., N.P., P.P., S.R., D. Scornet, H.V., F.W., H.K., A.F.P., M.V., E.C., and J.-M.A.; data curation, O.G., C.C., A. Couloux, L.B.-G., J.-M.A., and J.M.C.; writing – original draft, F.D., O.G., J.-M.A., and J.M.C.; writing – review and editing, all authors; visualization, F.D., O.G., L.B.-G., L.D., D.M., T.M., D. Sussfeld, X.F., L. Mazéas, J.B.-R., R.P., L.R., K.U., K.B., P.R., A.L., M.L., P.H.-G., C.V., A.D., A.L.B., H.K., E.C., and J.M.C.; supervision, O.G., N. Terrapon, M.L., R.A.C., H.C.d.C., O.D.C., S.D., G.H., A.J., C.B., E.P., P.L., S.A.R., A.S., L.T., H.V., G.W., H.K., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.-M.A., and J.M.C.; project administration, F.D., E.B., M.V., G.V.M., E.C., S.M.C., P.W., J.-M.A., and J.M.C.; funding acquisition, H.C.d.C., C.B., P.P., C.H., N.Y., S.M.C., and J.M.C.; conceptualization, J.M.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

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

- *Ectocarpus* species 1 strain Ec sil Puy CHCH Z9 G3m
- *Ectocarpus* species 1 strain Ec03
- *Ectocarpus* species 12 strain Ec fas CH92 Nie 2f
- *Ectocarpus* species 12 strain Ec fas CH92 Nie 3m
- *Ectocarpus* species 13 strain EcNAP12-S#4-19m
- *Ectocarpus* species 2 strain Ec06
- *Ectocarpus* species 3 strain Ec10
- *Ectocarpus* species 3 strain Ec11
- *Ectocarpus* species 5 strain Ec13
- *Ectocarpus* species 5 strain Ec12
- *Ectocarpus* species 6 strain EcLAC-371f
- *Ectocarpus* species 7 strain Ec32
- *Ectocarpus* species 8 strain EcLAC-412m
- *Ectocarpus* species 9 strain EcSCA-722f
- *Ectocarpus subulatus* strain Bft15b
- *Feldmannia mitchelliae* strain KU-2106 Giff mitch BNC GA
- *Fucus distichus*
- *Fucus serratus*
- *Fucus serratus*
- *Halopteris paniculata* strain Hal grac a UBK
- *Hapterophycus canaliculatus* strain Oshoro5f
- *Hapterophycus canaliculatus* strain Oshoro7m
- *Hapterophycus canaliculatus* strain Oshoro 3F x 9M
- *Hapterophycus canaliculatus* strain Oshoro 4F x 9M
- *Hapterophycus canaliculatus* strain Oshoro 6F x 6M
- *Heribaudiella fluviatilis* strain SAG. 13.90
- *Heterosigma akashiwo* strain CCMP452
- *Himanthalia elongata*
- *Laminaria digitata* strain LdigPH10-18mv
- *Laminarionema elsbetiae* strain ELsaHSoW15
- *Macrocystis pyrifera* strain P11A1
- *Macrocystis pyrifera* strain P11B4
- *Myriotrichia clavaeformis* strain Myr cla04
- *Myriotrichia clavaeformis* strain Myr cla05
- *Myriotrichia clavaeformis* strain Myr cla12
- *Pelvetia canaliculata*
- *Phaeothamnion wetherbeei* strain SAG 119.79
- *Pleurocardia lacustris* strain SAG 25.93
- *Porterinema fluviatile* strain SAG 2381
- *Pylaiella littoralis* strain U1.48
- *Pylaiella littoralis* strain F24
- *Saccharina japonica* strain Ja
- *Saccharina latissima* strain SLPER63f7
- *Saccorhiza dermatodea* strain SderLü1190fm
- *Saccorhiza polyschides* strain SpolBR94f
- *Saccorhiza polyschides* strain SpolBR94m
- *Saccorhiza polyschides*
- *Sargassum fusiforme*
- *Schizocladia ischiensis* strain KU-0333
- *Scytosiphon promiscuus* strain 000310-Muroran-5-female
- *Scytosiphon promiscuus* strain Ot110409-Otamoi-16-male
- *Scytosiphon promiscuus* strain SXS107
- *Sphacelaria rigidula* strain Sph rig Cal Mo 4-1-68b
- *Sphacelaria rigidula* strain Sph rig Cal Mo 4-1-G3b
- *Sphacelaria rigidula* strain Sph rig Cal Mo SP
- *Sphaerotrichia firma* strain ET2f
- *Sphaerotrichia firma* strain Sfir13m
- *Tribonema minus* strain UTEX B 3156
- *Undaria pinnatifida* strain Kr2015

- METHOD DETAILS
  - Biological material
  - DNA extraction
  - Illumina library preparation and sequencing
  - Oxford Nanopore library preparation and sequencing
  - RNA extraction, Illumina RNA-seq library preparation and sequencing
  - Assembly strategies
  - Assembly decontamination
  - Transcriptome assembly
  - *De novo* transcriptomes
  - Detection and masking of repeated sequences and transposons
  - Gene prediction
  - Annotation decontamination
  - Analyses aimed at deducing functional characteristics of predicted proteins
  - Detection of tandemly duplicated genes
  - Relative orientation of adjacent genes and lengths of intergenic regions
  - Detection of long non-coding RNAs
  - Intron conservation
  - Phylogenomic tree of the Phaeophyceae
  - Bayesian divergence time estimation for the brown algae
  - Detection of orthologous groups
  - Dollo analysis of orthogroup gain and loss
  - Phylostratigraphy analysis
  - Detection of gene family amplifications
  - Composite genes
  - Horizontal gene transfer (HGT)
  - Gene codon usage, functional annotation and expression
  - Comparative analysis of gene sets identified by genome-wide analyses of evolutionary history
  - Detection of viral genome insertions and viral regions in algal genomes
  - Phylogenetic analysis of viral genes
  - Metabolic networks
  - CAZymes
  - Sulfatases
  - Haloperoxidases
  - Ion channels
  - Membrane-localised proteins
  - Transcription-associated proteins
  - EsV-1-7 domain proteins
  - Histones
  - DNA methyltransferases
  - Spliceosome
  - Flagella proteins
  - Detection of *Porterinema fluviatile* genes differentially expressed in freshwater and seawater
  - Identification of genes with generation-biased expression patterns
  - Life cycle and thallus architecture
  - Assembly and analysis of organellar genomes
  - Analysis of *Ectocarpus* genome synteny
  - Analysis of *Ectocarpus* gene evolution
  - Phylogenetic analysis of *Ectocarpus* species
  - *Ectocarpus* introgression analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

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

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

REFERENCES

1. Choi, S.-W., Graf, L., Choi, J.W., Jo, J., Boo, G.H., Kawai, H., Choi, C.G., Xiao, S., Knoll, A.H., Andersen, R.A., et al. (2024). Ordovician origin and subsequent diversification of the brown algae. Curr. Biol. *34*, 740–754.e4. https://doi.org/10.1016/j.cub.2023.12.069.

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

3. Mazéas, L., Yonamine, R., Barbeyron, T., Henrissat, B., Drula, E., Terrapon, N., Nagasato, C., and Hervé, C. (2023). Assembly and synthesis of the extracellular matrix in brown algae. Semin. Cell Dev. Biol. *134*, 112–124. https://doi.org/10.1016/j.semcdb.2022.03.005.

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

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

6. Cock, J.M., Godfroy, O., Macaisne, N., Peters, A.F., and Coelho, S.M. (2014). Evolution and regulation of complex life cycles: a brown algal perspective. Curr. Opin. Plant Biol. *17*, 1–6. https://doi.org/10.1016/j.pbi.2013.09.004.

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

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

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

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

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

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

13. Graf, L., Shin, Y., Yang, J.H., Choi, J.W., Hwang, I.K., Nelson, W., Bhattacharya, D., Viard, F., and Yoon, H.S. (2021). A genome-wide investigation of the effect of farming and human-mediated introduction on the ubiquitous seaweed *Undaria pinnatifida*. Nat. Ecol. Evol. *5*, 360–368. https://doi.org/10.1038/s41559-020-01378-9.

14. Wang, S., Lin, L., Shi, Y., Qian, W., Li, N., Yan, X., Zou, H., and Wu, M. (2020). First draft genome assembly of the seaweed *Sargassum fusiforme*. Front. Genet. *11*, 590065. https://doi.org/10.3389/fgene.2020.590065.

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

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

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

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

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

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

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

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

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

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

25. Vosseberg, J., Stolker, D., von der Dunk, S.H.A., and Snel, B. (2023). Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome. Mol. Biol. Evol. *40*, msad011. https://doi.org/10.1093/molbev/msad011.

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

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

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

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

30. Radlow, M., Czjzek, M., Jeudy, A., Dabin, J., Delage, L., Leblanc, C., and Hartung, J. (2018). X-ray diffraction and density functional theory provide insight into vanadate binding to homohexameric bromoperoxidase II and the mechanism of bromide oxidation. ACS Chem. Biol. *13*, 1243–1259. https://doi.org/10.1021/acschembio.8b00041.

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

32. Meslet-Cladière, L., Delage, L., Leroux, C.J.J., Goulitquer, S., Leblanc, C., Creis, E., Gall, E.A., Stiger-Pouvreau, V., Czjzek, M., and Potin, P. (2013). Structure/function analysis of a Type III polyketide synthase in the brown alga *Ectocarpus siliculosus* reveals a biochemical pathway in phlorotannin monomer biosynthesis. Plant Cell *25*, 3089–3103. https://doi.org/10.1105/tpc.113.111336.

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

34. Zhao, D.-S., Hu, Z.-W., Dong, L.-L., Wan, X.-J., Wang, S., Li, N., Wang, Y., Li, S.-M., Zou, H.-X., and Yan, X. (2021). A Type III polyketide synthase (SfuPKS1) isolated from the edible seaweed *Sargassum fusiforme* exhibits broad substrate and catalysis specificity. J. Agric. Food Chem. *69*, 14643–14649. https://doi.org/10.1021/acs.jafc.1c05868.

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

36. Salgado, L.T., Cinelli, L.P., Viana, N.B., Tomazetto de Carvalho, R., De Souza Mourão, P.A., Teixeira, V.L., Farina, M., and Filho, A.G.M.A. (2009). A vanadium bromoperoxidase catalyzes the formation of high-molecular-weight complexes between brown algal phenolic substances and alginates(1). J. Phycol. *45*, 193–202. https://doi.org/10.1111/j.1529-8817.2008.00642.x.

37. Berglin, M., Delage, L., Potin, P., Vilter, H., and Elwing, H. (2004). Enzymatic cross-linking of a phenolic polymer extracted from the marine alga *Fucus serratus*. Biomacromolecules *5*, 2376–2383. https://doi.org/10.1021/bm0496864.

38. Bitton, R., Berglin, M., Elwing, H., Colin, C., Delage, L., Potin, P., and Bianco-Peled, H. (2007). The influence of halide-mediated oxidation on algae-born adhesives. Macromol. Biosci. *7*, 1280–1289. https://doi.org/10.1002/mabi.200700099.

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

40. Salgado, L.T., Tomazetto, R., Cinelli, L.P., Farina, M., and Amado Filho, G.M. (2007). The influence of brown algae alginates on phenolic compounds capability of ultraviolet radiation absorption in vitro. Braz. J. Oceanogr. *55*, 145–154. https://doi.org/10.1590/S1679-87592007000200007.

41. Fu, G., Nagasato, C., Oka, S., Cock, J.M., and Motomura, T. (2014). Proteomics analysis of heterogeneous flagella in brown algae (stramenopiles). Protist *165*, 662–675. https://doi.org/10.1016/j.protis.2014.07.007.

42. Kloareg, B., Badis, Y., Cock, J.M., and Michel, G. (2021). Role and evolution of the extracellular matrix in the acquisition of complex multicellularity in eukaryotes: A macroalgal perspective. Genes *12*, 1059. https://doi.org/10.3390/genes12071059.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

60. Colin, C., Leblanc, C., Michel, G., Wagner, E., Leize-Wagner, E., Van Dorsselaer, A., and Potin, P. (2005). Vanadium-dependent iodoperoxidases in *Laminaria digitata*, a novel biochemical function diverging from brown algal bromoperoxidases. J. Biol. Inorg. Chem. *10*, 156–166. https://doi.org/10.1007/s00775-005-0626-8.

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

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

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

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

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

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

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

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

69. Bräutigam, M., Klein, M., Knippers, R., and Müller, D.G. (1995). Inheritance and meiotic elimination of a virus genome in the host *Ectocarpus siliculosus* (Phaeophyceae). J. Phycol. *31*, 823–827. https://doi.org/10.1111/j.0022-3646.1995.00823.x.

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

71. Kabbara, S., Hérivaux, A., Dugé de Bernonville, T., Courdavault, V., Clastre, M., Gastebois, A., Osman, M., Hamze, M., Cock, J.M., Schaap, P., et al. (2019). Diversity and evolution of sensor histidine kinases in eukaryotes. Genome Biol. Evol. *11*, 86–108. https://doi.org/10.1093/gbe/evy213.

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

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

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

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

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

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

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

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

80. Shen, Y., Motomura, T., Ichihara, K., Matsuda, Y., Yoshimura, K., Kosugi, C., and Nagasato, C. (2023). Application of CRISPR-Cas9 genome editing by microinjection of gametophytes of *Saccharina japonica* (Laminariales, Phaeophyceae). J. Appl. Phycol. *35*, 1431–1441. https://doi.org/10.1007/s10811-023-02940-1.

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

82. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

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

84. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

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

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

87. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

109. Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C.D., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M. (2016). Gmove a Tool for Eukaryotic Gene Predictions Using Various Evidences. F1000Research *5*. https://doi.org/10.7490/f1000research.1111735.1.

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

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

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

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

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

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

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

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

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

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

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

121. Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2018). CompositeSearch: A generalized network approach for composite gene families detection. Mol. Biol. Evol. *35*, 252–255. https://doi.org/10.1093/molbev/msx283.

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

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

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

125. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

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

127. Aylward, F.O., and Moniruzzaman, M. (2021). ViralRecall-A flexible command-line tool for the detection of giant virus signatures in 'Omic data. Viruses *13*, 150. https://doi.org/10.3390/v13020150.

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

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

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

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

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

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

133. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
134. Petroll, R., Schreiber, M., Finke, H., Cock, J.M., Gould, S.B., and Rensing, S.A. (2021). Signatures of transcription factor evolution and the secondary gain of red algae complexity. Genes *12*, 1055. https://doi.org/10.3390/genes12071055.
135. Petroll, R., Varshney, D., Hiltemann, S., Finke, H., Schreiber, M., de Vries, J., and Rensing, S.A. (2024). Enhanced sensitivity of TAPscan v4 enables comprehensive analysis of streptophyte transcription factor evolution. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.13.602682.
136. Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Res. *49*, W216–W227. https://doi.org/10.1093/nar/gkab225.
137. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.
138. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680. https://doi.org/10.1093/nar/22.22.4673.
139. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.
140. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.
141. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
142. Andrews, S. (2016). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
143. Krueger, F. (2015). Trim Galore!: A Wrapper around Cutadapt and FastQC to Consistently Apply Adapter and Quality Trimming to FastQ Files, with Extra Functionality for RRBS Data (Babraham Institute).
144. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.
145. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
146. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.
147. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.
148. Wallau, G.L., Capy, P., Loreto, E., Le Rouzic, A., and Hua-Van, A. (2016). VHICA, a new method to discriminate between vertical and horizontal transposon transfer: application to the mariner family within *Drosophila*. Mol. Biol. Evol. *33*, 1094–1109. https://doi.org/10.1093/molbev/msv341.
149. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18. https://doi.org/10.1093/nar/gkw955.
150. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
151. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. *45*, W6–W11. https://doi.org/10.1093/nar/gkx391.
152. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16. https://doi.org/10.1093/nar/gkh152.
153. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.
154. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.
155. Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics *33*, 2197–2198. https://doi.org/10.1093/bioinformatics/btx144.
156. Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics *20*, 3643–3646. https://doi.org/10.1093/bioinformatics/bth397.
157. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *10*, e1003537. https://doi.org/10.1371/journal.pcbi.1003537.
158. Douglas, J., Jiménez-Silva, C.L., and Bouckaert, R. (2022). StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent. Syst. Biol. *71*, 901–916. https://doi.org/10.1093/sysbio/syac010.
159. Bouckaert, R.R., and Drummond, A.J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC Evol. Biol. *17*, 42. https://doi.org/10.1186/s12862-017-0890-6.
160. Kloepper, T.H., and Huson, D.H. (2008). Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol. Biol. *8*, 22. https://doi.org/10.1186/1471-2148-8-22.
161. Gschloessl, B., Guermeur, Y., and Cock, J. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinf. *9*, 393.
162. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
163. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.
164. Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics *36*, 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.
165. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.
166. Wickham, H., Chang, W., and Wickham, M.H. (2016). Package 'ggplot2.' Create Elegant Data Visualisations Using the Grammar of Graphics, version 2, pp. 1–189.
167. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.
168. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic,

prokaryotic, and viral genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

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

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

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

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

173. Barbeyron, T., Brillet-Guéguen, L., Carré, W., Carrière, C., Caron, C., Czjzek, M., Hoebeke, M., and Michel, G. (2016). Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. PLoS One *11*, e0164846. https://doi.org/10.1371/journal.pone.0164846.

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

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

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

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

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

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

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

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

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

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

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

185. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. (2018). A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere *3*, e00069-18. https://doi.org/10.1128/mSphereDirect.00069-18.

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

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

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

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

190. Starr, R.C., and Zeikus, J.A. (1993). UTEX-The culture collection of algae at the University of Texas at Austin 1993 list of cultures. J. Phycol. *29*, 1–106. https://doi.org/10.1111/j.0022-3646.1993.00001.x.

191. Dittami, S.M., Gravot, A., Goulitquer, S., Rousvoal, S., Peters, A.F., Bouchereau, A., Boyen, C., and Tonon, T. (2012). Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to low salinities in *Ectocarpus* (brown algae). Plant J. *71*, 366–377. https://doi.org/10.1111/j.1365-313X.2012.04982.x.

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

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

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

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

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

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

198. Forcelloni, S., and Giansanti, A. (2020). Evolutionary forces and codon bias in different flavors of intrinsic disorder in the human proteome. J. Mol. Evol. *88*, 164–178. https://doi.org/10.1007/s00239-019-09921-4.

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

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

201. Steffen, R., Ogoniak, L., Grundmann, N., Pawluchin, A., Soehnlein, O., and Schmitz, J. (2022). paPAML: an improved computational tool to explore selection pressure on protein-coding sequences. Genes *13*, 1090. https://doi.org/10.3390/genes13061090.

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

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

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

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Descriptions of all sequenced samples have been deposited in the EBI/ENA database	This study.	EBI/ENA project PRJEB76691
Critical commercial assays		
OmniPrep Genomic DNA Purification Kit	G Biosciences, St. Louis, MO, USA	N/A
Nucleospin Plant II midi DNA Extraction Kit	Macherey-Nagel, Düren, Germany	N/A
NEBNext DNA Modules Products	New England Biolabs, Ipswich, MA, USA	N/A
NEBNext Sample Reagent Set	New England Biolabs, Ipswich, MA, USA	N/A
Ampure XP	Beckmann Coulter Genomics, Danvers, MA, USA	N/A
Kapa Hifi Hotstart NGS library Amplification kit	Roche, Basel, Switzerland	N/A
Short Read Eliminator Kit	Pacific Biosciences, Menlo Park, CA, USA	N/A
1D Genomic DNA by Ligation	Oxford Nanopore Technologies Ltd, Oxford, UK	SQK-LSK109, SQK-LSK108 or SQK-LSK110
Qiagen RNeasy kit or the Macherey Nagel RNAplus kit	Macherey-Nagel, Düren, Germany	N/A
TruSeq Stranded mRNA Sample Prep	Illumina	N/A
NEBNext Ultra II Directional RNA Library Prep for Illumina	New England BioLabs	N/A
Deposited data		
The sequence data generated by this project is described in Table S1.	This study.	EBI/ENA: PRJEB76691
CNRS Research Data dataset "Data for Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems"	This study.	CNRS Research Data: https://doi.org/10.57745/9U1J85
Experimental models: Organisms/strains		
The strains used for genome and transcriptome sequencing are listed in Table S1A.	Culture collection references are provided where relevant.	See strain names and culture collection accessions for identifiers.
Software and algorithms		
MEGAHIT version 1.1.1	Li et al.[82]	RRID:SCR_018551 https://github.com/voutcn/megahit
MetaGene version 2008.8.19	Noguchi et al.[83]	http://metagene.cb.k.u-tokyo.ac.jp/
BLAST	Altschul et al.[84]	RRID:SCR_004870 http://blast.ncbi.nlm.nih.gov/Blast.cgi
Burrows-Wheeler Aligner	Li and Durbin[85]	RRID:SCR_010910 http://bio-bwa.sourceforge.net/
Bowtie2 version 2.3.5.1	Langmead and Salzberg[86]	RRID:SCR_016368 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SPAdes assembler version 3.8.1	Bankevich et al.[87]	RRID:SCR_000131 https://cab.spbu.ru/software/spades/
filtlong	Wick, R.	RRID:SCR_024020 https://github.com/rrwick/Filtlong
Smartdenovo	Liu et al.[88]	RRID:SCR_017622 https://github.com/ruanjue/smartdenovo

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Redbean	Ruan and Li[89]	N/A
Flye	Kolmogorov et al.[90]	RRID:SCR_017016 https://github.com/fenderglass/Flye
Necat	Chen et al.[91]	https://github.com/xiaochuanle/necat
Racon	Vaser et al.[92]	RRID:SCR_017642 https://github.com/isovic/racon
Hapo-G	Aury et al.[93]	https://www.genoscope.cns.fr/hapog/
Metabat 2	Kang et al.[94]	RRID:SCR_019134 https://bitbucket.org/berkeleylab/metabat/src/master/
SortMeRNA	Kopylova et al.[95]	RRID:SCR_014402 http://bioinfo.lifl.fr/RNA/sortmerna/
Velvet version 1.2.07	Zerbino and Birney[96]	RRID:SCR_010755 http://www.molecularevolution.org/software/genomics/velvet
Oases version 0.2.08	Schulz et al.[97]	RRID:SCR_011896 http://www.ebi.ac.uk/~zerbino/oases/
TransDecoder	Haas, B.J.	RRID:SCR_017647 https://github.com/TransDecoder/TransDecoder
CDDsearch	Marchler-Bauer et al.[98]	N/A
Trimmomatic version 0.38 and version 0.39	Bolger et al.[99]	RRID:SCR_011848 http://www.usadellab.org/cms/index.php?page=trimmomatic
Trinity version version 2.6.5	Grabherr et al.[100]	RRID:SCR_013048 http://trinityrnaseq.sourceforge.net/
rnaSPAdes version version 3.13.1	Bushmanova et al.[101]	RRID:SCR_016992 http://cab.spbu.ru/software/rnaspades/
RepeatMasker version 4.1.0	Smit et al.[102]	RRID:SCR_012954 http://repeatmasker.org/
Tandem repeats finder	Benson et al.[103]	RRID:SCR_022193 https://github.com/Benson-Genomics-Lab/TRF
REPET	Flutre et al.[104]	N/A
BLAT	Kent[105]	RRID:SCR_011919 http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Genewise	Birney et al.[106]	RRID:SCR_015054 http://www.ebi.ac.uk/Tools/psa/genewise/
DIAMOND version 0.9.30	Buchfink et al.[107]	RRID:SCR_009457 http://www.nitrc.org/projects/diamond/
Est2Genome	Mott[108]	https://galaxy-iuc.github.io/emboss-5.0-docs/est2genome.html
Gmove	Dubarry et al.[109]	RRID:SCR_019132 http://www.genoscope.cns.fr/gmove
votingLNC	Debit, A.	https://gitlab.com/a.debit/votinglnc
AliView version 1.26	Larsson[110]	RRID:SCR_002780 https://github.com/AliView
RAxML version 8.2.	Stamatakis[111]	RRID:SCR_006086 https://github.com/stamatak/standard-RAxML
Tracer version 1.7.2	Rambaut et al.[112]	RRID:SCR_019121 https://bioweb.pasteur.fr/packages/pack@Tracer@v1.6
OrthoFinder version 2.5.2	Emms and Kelly[113]	RRID:SCR_017118 https://github.com/davidemms/OrthoFinder
Count version 9.1106	Csűös[114]	https://www.iro.umontreal.ca/~csuros/gene_content/count.html
MUSCLE version 3.8.1551	Edgar[115]	RRID:SCR_011812 http://www.ebi.ac.uk/Tools/msa/muscle/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
OD-Seq version 1.0	Jehl et al.[116]	https://bioconductor.org/packages/release/bioc/manuals/odseq/man/odseq.pdf
HMMER3 package versions 3.1b1 and 3.3.2	Mistry et al.[117]	RRID:SCR_005305 http://hmmer.janelia.org/
GenEra	Barrera-Redondo et al.[118]	N/A
MCL	Enright et al.[119]	RRID:SCR_024109 https://micans.org/mcl/
Foldseek	Kempen et al.[120]	https://search.foldseek.com/search
CleanBlastp	Pathmanathan et al.[121]	N/A
SEED	Overbeek et al.[122]	RRID:SCR_002129 http://www.theseed.org/wiki/Home_of_the_SEED
IPR2GO	Paysan-Lafosse et al.[123]	http://www.ebi.ac.uk/interpro/search/sequence-search
eggNOG	Huerta-Cepas et al.[124]	RRID:SCR_002456 http://eggnog.embl.de
eggNOG-mapper	Cantalapiedra et al.[125]	RRID:SCR_021165 http://eggnog-mapper.embl.de
Spearman's rank correlation analysis tool version 1.1.23-r7	P. Wessa, Free Statistics Software, Office for Research Development and Education	https://www.wessa.net/
Prodigal version 2.6.3	Hyatt et al.[126]	RRID:SCR_011936 https://github.com/hyattpd/Prodigal
ViralRecall version 2.0	Aylward et al.[127]	https://github.com/faylward/viralrecall
esl-translate version 0.48	Rivas, E.	https://github.com/EddyRivasLab/easel/blob/master/miniapps/esl-translate.man.in
bedtools version 2.29.2	Quinlan and Hall[128]	RRID:SCR_006646 https://github.com/arq5x/bedtools2
MMseqs cluster version 13.45111	Hauser et al.[129]	RRID:SCR_008184 https://github.com/eturro/mmseq#mmseq-transcript-and-gene-level-expression-analysis-using-multi-mapping-rna-seq-reads
MAFFT v7	Katoh and Standley[130]	RRID:SCR_011811 http://mafft.cbrc.jp/alignment/server/
MEGA	Tamura et al.[131]	RRID:SCR_023017 https://www.megasoftware.net/
NGphylogeny platform	Lemoine et al.[132]	https://ngphylogeny.fr/.
TrimAl	Capella-Gutiérrez et al.[133]	RRID:SCR_017334 http://trimal.cgenomics.org/
TAPscan version 4	Petroll et al.[134,135]	https://plantcode.cup.uni-freiburg.de/tapscan/
Expasy web translator	Duvaud et al.[136]	RRID:SCR_024703 https://web.expasy.org/translate/
Geneious versions 11.0.5 and 11.1.5	Geneious	RRID:SCR_010519 http://www.geneious.com/
Interproscan 94.0	Jones et al.[137]	RRID:SCR_005829 http://www.ebi.ac.uk/Tools/pfa/iprscan/
Clustal 2.1	Thompson et al.[138]	RRID:SCR_001591 http://www.ebi.ac.uk/Tools/msa/clustalo/
Gblocks	Castresana[139]	RRID:SCR_015945 http://molevol.cmima.csic.es/castresana/Gblocks_server.html
Kallisto version 0.44.0.	Bray et al.[140]	RRID:SCR_016582 https://pachterlab.github.io/kallisto/about
Deseq2	Love et al.[141]	RRID:SCR_015687 https://bioconductor.org/packages/release/bioc/html/DESeq2.html

(*Continued on next page*)

Continued

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

(*Continued on next page*)

Continued

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

EXPERIMENTAL MODEL DETAILS

Ascophyllum nodosum

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

Chordaria linearis strain ClinC8C

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

Choristocarpus tenellus strain KU-1152

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

Chrysoparadoxa australica strain CS-1217

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

Cladosiphon okamuranus strain S-strain

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

Desmarestia dudresnayi strain DdudBR16

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

Sex: monoicous
Maintenance: Maintained in culture

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sphaerotrichia firma strain Sfir13m

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

Tribonema minus strain UTEX B 3156

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

Undaria pinnatifida strain Kr2015

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

METHOD DETAILS

Biological material

Sequencing brown algal genomes has been hampered by the significant challenges involved, including inherent problems with growing brown algae, the presence of molecules that interfere with sequencing reactions and complex associations with microbial symbionts. To address these problems, cultured, unialgal filamentous gametophyte material was used whenever possible (i.e. for species with haploid-diploid life cycles) and the extraction methodology was adapted for each species.

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

All strains except those belonging to the Fucales were grown under laboratory conditions. The latter cannot be maintained long-term in the laboratory so field material was harvested for extractions. The haploid gametophyte generation was grown in culture for species with characterised haploid-diploid life cycles, with the exception of *Ectocarpus* strains, for which haploid partheno-sporophytes or diploid sporophytes were cultivated. All cultures were grown either in 140 mm diameter Petri dishes or in 2–10 L bottles, the latter aerated by bubbling with sterile air. Most cultures were grown in Provasoli-enriched[104] natural seawater (PES medium) under fluorescent white light (10–30 μM photons/m2•s) at 13°C (or at 10°C for *Hapterophycus canaliculatus* and *Chordaria linearis* or 20°C for *Sphacelaria rigidula*, *Dictyota dichotoma*, *Schizocladia ischiensis* and *Chrysoparadoxa Australica*). Exceptions included the freshwater species *Pleurocladia lacustris*, *Porterinema fluviatile* and *Heribaudiella fluviatilis*, which were grown in natural seawater that had been diluted to 5% with distilled water (i.e., 95% distilled water / 5% seawater) before addition of ES medium (http://sagdb.uni-goettingen.de/culture_medi a/01%20Basal%20Medium.pdf) micronutrients (at 20°C for *P. lacustris*) and *Phaeothamnion wetherbeei*, which was grown in MIEB12 (medium 7 in Letunic et al.[177]). Whole thallus was extracted for all species except the Fucales, where either dissected meristematic regions or released male gametes were extracted. Tissue samples were frozen in liquid nitrogen and stored at -80°C before extraction.

DNA extraction

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

Illumina library preparation and sequencing

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

Oxford Nanopore library preparation and sequencing

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

RNA extraction, Illumina RNA-seq library preparation and sequencing

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

Assembly strategies

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

Short-read-based genome assembly

When sequencing was performed exclusively using short reads, reads corresponding to bacterial contaminants were filtered out early in the assembly process because, typically, the initial datasets were too large to run assemblers like SPAdes. To remove bacterial contaminants, an assembly based on the initial illumina dataset was first generated for each strain using a fast and non-greedy algorithm, MEGAHIT[82] version 1.1.1 with the parameters –k-min 101 –k-max 131 –k-step 10. Assigning taxonomy is easier when working with contigs than with reads. Contigs exceeding 500 bp in each preliminary assembly underwent taxonomic classification based on gene models predicted using the *ab initio* software MetaGene[83] version 2008.8.19 with default parameters and then aligning proteins against UniprotKB using BLASTp (e-value $<10e^{-4}$). A superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 50% of their genes assigned to Bacteria and with at least one gene every 10 kbp were classified as bacterial sequences. For each strain, the initial Illumina sequencing reads were aligned against the corresponding bacterial sequences using latest version of the Burrows-Wheeler Aligner[85] (BWA) with default parameters and mapped short-reads were labelled as contaminants, and assembled for the purpose of obtaining more contiguous contigs. These bacterial contigs were then used to build a contaminant sequence database. Finally, the clean subset of reads was obtained by aligning the whole Illumina dataset against this strain-specific bacterial contig database, using Bowtie2[86] version 2.2.9 with default parameters. A final assembly was then generated for each strain using the contaminant-free read datasets and the SPAdes[87] assembler version 3.8.1 with the parameters -k 21,57,71,99,127 -m 2000 –only-assembler –careful. Genome assemblies based only on short-reads were more fragmented (N50 ranged from 3 kbp to 31 kbp) than assemblies that used long reads but the sizes of the former were consistent with expectations.

Long-read-based genome assemblies

A subset of the strains produced DNA of both adequate quality and quantity, enabling successful long-read sequencing. In these cases, long reads were assembled directly and the detection of possible bacterial contigs was carried out after the assembly step. To produce long-read-based genome assemblies we generated three samples of reads i) all reads, ii) 30X coverage of the longest reads and iii) 30X coverage of the filtlong (https://github.com/rrwick/Filtlong) highest-score reads. The three samples were used as input data for four different assemblers, Smartdenovo,[88] Redbean,[89] Flye[90] and Necat.[91] Based on the cumulative size and contiguity, we selected the best assembly for each strain. This assembly was then polished three times using Racon[92] with nanopore reads, and twice with Hapo-G[93] and Illumina PCR-free reads.

Assembly decontamination

Contigs from the short- and long-read genome assemblies were inspected for potential bacterial sequences. This process was carried out using a combination of several analysis and tools: GC composition, read coverage, Metabat 2 (for tetramer composition and clustering)[94] and Metagene (for gene prediction and taxonomic identification, as described previously). Contigs were manually removed based on their characteristics.

Transcriptome assembly

Ribosomal-RNA-like reads were detected using SortMeRNA[95] and filtered out. The Illumina RNA-seq short reads from each strain were assembled using Velvet[96] version 1.2.07 and Oases[97] version 0.2.08 with kmer sizes of 61, 63 and 65 bp. BUSCO[168] analysis (v5, eukaryota_odb10) was then performed on the three resulting assemblies for each strain in order to select the best assembly, i.e. the most complete at the gene level. Reads were mapped back to the contigs with BWA-mem, and only consistent paired-end reads were retained. Uncovered regions were detected and used to identify chimeric contigs. In addition, open reading frames (ORF) and domains were identified using TransDecoder (Haas, B.J., https://github.com/TransDecoder/TransDecoder) and CDDsearch,[98]

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

De novo transcriptomes

The RNA-seq data was also used to generate *de novo* transcriptomes. For each strain, all the RNA-seq data available was cleaned to remove poor quality sequence and adapter sequences using Trimmomatic[99] v0.39 prior to being assembled using either Trinity[100] version v2.6.5 or rnaSPAdes[101] version v3.13.1. The strandness and Kmer-length parameters of the assemblers were adjusted to take into account RNA-seq read characteristics. The *de novo* transcriptomes represented an alternative source to identify and characterise genes if they were not detected in the genome assemblies. The *de novo* transcriptomes are available from the CNRS Research Data dataset (https://doi.org/10.57745/9U1J85) and from the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/).

Detection and masking of repeated sequences and transposons

Prior to gene annotation, each genome assembly was masked based on the repeat library from *Ectocarpus* species 7 (formerly *Ectocarpus siliculosus*)[11] and using RepBase with RepeatMasker[102] version v.4.1.0, default parameters. Tandem repeats finder (TRF)[103] was also used to mask tandem repeat duplications. In addition, transposons were annotated in ten species using REPET[104] and the transposons detected were used as a reference to mask all genomes with RepeatMasker[102] version v4.1.0, default parameters.

Gene prediction

For each strain, gene prediction was performed using both homologous proteins and RNA-seq data. Proteins from *Ectocarpus* species 7 (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2)[178] and UniRef90 (https://www.uniprot.org/uniref/) were aligned against each genome assembly. First, BLAT[105] with default parameters was used to quickly localise putative genes corresponding to the *Ectocarpus* species 7 proteins. The best match and matches with a score $\geq$ 90% of the best match score were retained. Second, the alignments were refined using Genewise[106] with default parameters, which is more precise for intron/exon boundary detection. Alignments were retained if more than 80% of the length of the protein was aligned to the genome. To detect conserved proteins and allow detection of horizontal gene transfer, UniRef90 proteins (without *E. siliculosus* sequences) were aligned with DIAMOND[107] (v0.9.30 with parameters –evalue 0.001 –more-sensitive) to genomic regions lacking alignments with an *Ectocarpus* species 7 protein. Only the five best matches per locus were retained, based on their bitscore. Selected proteins from UniRef90 were aligned to the whole genome using Genewise as described previously, and alignments with at least 50% of the aligned protein length were retained. The assembled transcriptome for each strain was aligned to the strain's genome assembly using BLAT[105] with default parameters. For each transcript, the best match was selected based on the alignment score, with an identity greater or equal to 90%. Selected alignments were refined using Est2Genome[108] in order to precisely detect intron boundaries. Alignments were retained if more than 80% of the length of the transcript was aligned to the genome with a minimal identity of 95%. Finally, the protein homologies and transcript mapping were integrated using a combiner called Gmove.[109] This tool can find coding sequences (CDSs) based on genome-located evidence without any calibration step. Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for open reading frames (ORFs) consistent with the protein evidence. Translated proteins of predicted genes were then aligned against NR prot (release 19/02/2019) and the *Ectocarpus* species 7 version v2 proteome[178] (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2) using DIAMOND BLASTp with parameters –evalue 10-5 –more-sensitive –unal 0. All predicted genes with significant matches (the smallest protein had to be aligned for at least 50% of its length) were retained. In addition to these genes, we also retained genes with CDS size greater than 300 bp and with a coding ratio (CDS size / mRNA size) greater or equal to 0.5.

Annotation decontamination

After predicting the genes, an additional analysis was carried out to detect bacterial sequences. If a contig did not contain any genes, it was analysed with MetaGene and the predicted proteins added to the gene catalogue for the purpose of detecting bacterial sequences. Proteins generated from predicted genes (Gmove plus MetaGene) were then aligned against UniprotKB using BLASTp (e-value < $10e^{-4}$) and superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 80% of their genes assigned to bacteria, Archaea or viruses were classified as bacterial sequences and removed from the final assembly file. Genes belonging to these contigs were also removed from the final gene catalogue. Finally, completeness of each predicted gene catalogue was assessed using BUSCO[168] (v5.0.0; eukaryota_odb10).

In addition, the quality of the annotations was assessed by comparing the length of coding regions in pairs of orthologous proteins (best reciprocal hits) between each genome and *Ectocarpus* species 7, which was used as a reference because its high-quality annotation has been extensively curated.[178] The correlation between orthologous CDS lengths was higher for genomes sequenced with long reads than for genomes only sequenced with short reads (Figure S1B). This difference was probably principally due to a higher proportion of underestimated protein lengths in the latter (Table S1B) which likely corresponded to fragmented genes. The qualities of Ectocarpales genome annotations were very high (BUSCO and length of predicted CDS) even when the genomes were sequenced

using only short reads, probably because their phylogenetic proximity to *Ectocarpus* species 7 facilitated the building of good quality gene models.

Analyses aimed at deducing functional characteristics of predicted proteins

Several different analyses of the predicted proteomes of each species were carried out to provide information about the cellular functions of the encoded proteins. These included eggNOG-mapper[125] analyses (v2.1.8 or v2.0.1, with emapperDB v5.0.2 or v4.5.1) to provide multiple functional annotations (Gene Ontology, Kyoto Encyclopedia of genes and genomes, Clusters of Orthologous Genes, Pfam), Interproscan[137] analyses (versions v5.55-88.0, v5.51-85.0 or v5.36-75.0) to detect functional domains, Hectar[161] (v1.3) predictions of protein subcellular localisation and various DIAMOND[107] (v2.0.15 vs UniRef90 2022_03, with parameter "evalue" set to $10e^{-5}$) sequence similarity searches aimed at identifying homologous proteins with functional annotations.

Detection of tandemly duplicated genes

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

Relative orientation of adjacent genes and lengths of intergenic regions

For each species, the proportion of pairs of adjacent genes localized on opposite strands was compared to the expected proportion of 0.5 using a binomial test (with p=0.5). The p-values are shown in Table S1B (p-values of <0.05 correspond to cases where the proportion is significantly higher than 0.5).

The lengths of intergenic regions between pairs of adjacent genes located on opposite strands (i.e. divergently or convergently transcribed) were compared with the lengths of intergenic regions between genes located on the same strand (i.e. transcribed in the same direction). Contingency tables were constructed for each species using a threshold of 1000 bp for the intergenic length and the number of intergenic regions in each of four categories were counted: 1) same strand genes, intergenic <1000 bp, 2) opposite strand genes, intergenic <1000 bp, 3) same strand genes, intergenic $\geq$1000 bp, 4) opposite strand genes, intergenic $\geq$1000 bp. Fisher exact tests were applied to the contingency tables (alternative hypothesis: true odds ratio is greater than 1). The p-values are shown in Table S1. When p-values are <0.05, short intergenic lengths are significantly associated with pairs of genes on opposite strands. All calculations were performed with R[162] (version 4.3.0).

Detection of long non-coding RNAs

Transcriptome data for 11 species (Table S1F), including nine brown algal strains and two outgroup taxa, was analysed to identify lncRNAs. Any transcripts with invalid nucleotide DNA symbols were discarded and sequences shorter than 200 nucleotides were removed to avoid the detection of small RNA transcripts. The transcriptome sequences in Fasta format were analysed with votingLNC (https://gitlab.com/a.debit/votinglnc) to detect lncRNA transcripts and assign a confidence level for each transcript. A similar approach was used to detect lncRNAs in the lncPlankton database.[179] VotingLNC is a meta-classifier combining the predictions of the ten most commonly used coding potential tools. Based on a majority voting ensemble procedure, the meta-tool assigns the final coding potential class to a transcript as the class label predicted most frequently by the ten classification models included in the ensemble. Alongside the majority voting class, a reliability score was calculated for each transcript. A cut-off non-coding reliability score of $p > 0.5$ was chosen to treat a transcript as lncRNA and to decrease false-positive identification. The set of transcripts predicted as lncRNA by the majority-voting procedure and having an ORF(s) encoding peptide(s) with length $\geq$ 100aa were discarded. lncRNA transcripts that had significant matches in either the Pfam[174] (hmmscan e-value < 0.001) or SwissProt (BLASTp e-value < $1e^{-5}$ and similarity $\geq$ 90%) databases were removed from the dataset. Transcript length, GC content, and the length of the longest ORF were compared between lncRNAs and protein-coding RNAs. The comparison was carried out using a Wilcoxon test. R version V.4.1.2 was used for all the analyses and ggplot2[166] (version 3.4.0) for plotting.

Intron conservation

Intron positions were compared in a set of single copy genes that are conserved across all the Phaeophyceae and the outgroup species. The analysis focused on the 21 reference genomes (Table S1F) and on orthogroups that occurred exactly once in at least 20 of the 21 genomes, allowing the gene to be absent from only one of the 21 genomes. In addition, orthogroups were discarded if more than three copies had been annotated in the other Phaeophyceae genomes. These filters produced a set of 235 conserved (ancestral) orthogroups. Multiple alignments were carried out for each orthogroup using MUSCLE[115] version 3.8.1551 with default parameters and conserved blocks were identified with Gblocks[139] version 0.91b with the parameters -p=t -s=n -b5=a -b2=[nsp] -b1=[nsp] -b3=6, where "nsp" is equal to 90% of the number of proteins aligned. A shell script was then used to compare intron positions in the alignments. For each intron in the multiple sequence alignment, we obtained a corresponding conservation profile listing which species contains an intron at that position. The profiles obtained for the 949 introns that are in conserved blocks of the multiple alignments are

shown in Figure S4B. Both phase and length of ancestral introns (e.g. that were conserved in most Phaeophyceae and at least two sister clades) were compared to the phase and length of *Ectocarpus* species 7 introns as a reference. The same approach was used to compare intron positions across 11 *Ectocarpus* species, with *Scytosiphon promiscuus* as an outgroup, by selecting 831 conserved monocopy orthogroups. The number of introns per gene in brown algae and in closely-related outgroup species were compared using a contingency test (Table S1C).

Phylogenomic tree of the Phaeophyceae

To provide a phylogenetic framework for the analyses of the Phaeoexplorer genome dataset, the 41-species phylogenomic tree reported by Akita et al.[180] was updated by adding 15 additional species using the same methodology. Briefly, for the additional species, amino acid sequences were recovered for the 32 single-copy orthologous genes used to construct the published tree and these were aligned manually with the existing sequences using the alignment software AliView[110] v.1.26. The aligned sequences of the final 56 species were concatenated and maximum likelihood analysis was carried out with 10,000 rapid bootstraps using RAxML[111] v.8.2.9 and the gamma model. The best-fit evolutionary model for each gene was determined using AIC.

Bayesian divergence time estimation for the brown algae

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

Detection of orthologous groups

Predicted proteins from the 60 strains sequenced in Phaeoexplorer complemented with 16 public proteomes covering the Ochrophytina subphylum and the terrestrial oomycetes were clustered using OrthoFinder[113] v2.5.2 with default parameters. This generated 56,340 orthogroups that contained 90.1% of the proteins (1,415,341 of the 1,571,648). Seventy-one of the 76 strains had more than 75% of their proteins in an orthogroup shared with at least one other strain. The orthogroups contain between 2 and 6,220 proteins with a mean of 25.1 proteins and a median of three.

Dollo analysis of orthogroup gain and loss

An analysis of evolutionary events of gene family gain and loss was carried out on a selection of strains covering the brown algal phylogeny and sister groups as distant as the Raphidophyceae under the Dollo parsimony law using orthogroups as proxies for gene families. To limit possible problems due to the fragmentation of predicted proteins in some assemblies, we selected 24,410 orthogroups present in at least one of 17 strains that had both good quality genome assembly and good quality gene predictions. Dollo parsimony analysis was then run using Count[114] version v9.1106 based on a cladogram of a subset of 24 species representative of the Phaeoexplorer project and excluding all public outgroups more distant than *Heterosigma akashiwo*. The cladogram was based on the topology of the brown algae phylogenetic tree published by Akita et al.[182]

Phylostratigraphy analysis

GenEra[118] was used to estimate gene family founder events for each genome assembly by running DIAMOND[107] in ultra-sensitive mode against the Phaeoexplorer protein dataset and the NCBI non-redundant database. All sequence matches with e-values $< 10^{-5}$ were treated as being homologous with the query genes in the target genomes. The NCBI taxonomy was used as an initial template to infer the evolutionary relationships of each query gene with their matches in the sequence database but taxonomic assignments within the PX clade and Phaeophyceae were then modified to reflect the evolutionary relationships that were inferred in the maximum likelihood tree. Gene families were predicted based on a clustering analysis of the query proteins against themselves using an e-value cutoff of 10^{-5} in DIAMOND and an inflation parameter of 1.5 with MCL.[119] Estimated evolutionary distances were extracted for each pair of species from the maximum likelihood species tree (substitutions/site) to calculate homology detection failure probabilities.[183] Taxonomic sampling of the species tree enabled homology detection failure tests to be carried out within the PX clade. Gene families whose ages could not be explained by homology detection failure were analysed by inspecting the functional and domain annotations for *Ectocarpus* species.[7,179] Structural alignments were performed using Foldseek[120] against the AlphaFold protein structure database.[170]

Detection of gene family amplifications

A binomial test with a parameter of 17/21 was carried out to detect gene families (OGs) that had significantly expanded in 17 Phaeophyceae reference genomes compared with four closely-related outgroup species (*Schizocladia ischiensis*, *Tribonema minus*, *Chrysoparadoxa australica* and *Heterosigma akashiwo*; Table S1F). Expanded gene families deviated significantly from the expected proportion (17/21 under the null hypothesis where there are equal gene numbers in all species). Benjamini–Hochberg FDR correction for multiple testing was then applied and 233 candidate OGs with corrected p-values of < 0.001 were retained. All calculations were performed with R (version 4.1.0).

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

1) A protein consensus was first deduced for each orthogroup. Protein sequences representative of all lineages were extracted and aligned using MUSCLE[115] version 3.8.1551 with default parameters and the multiple alignments were filtered using OD-Seq[116] version 1.0 to remove outlier sequences, with parameter –score set to 1.5. The consensus sequences were then extracted from the multiple alignments of non-outlier sequences using hmmemit in the HMMER3[117] package version 3.1b1 with default parameters.
2) In order to estimate gene family copy number independently of the assembly and annotation processes, short read sequences for each genome were mapped onto the orthogroup consensus sequences using DIAMOND.[107] Unique matches were retained for each read and depth of coverage was calculated for each consensus orthogroup. The depth obtained for each orthogroup was normalised for each species by dividing by the depth obtained on a set of conserved single-copy genes, so that the final value obtained was representative of the gene copy number. Then, for each candidate amplified orthogroup, the average depth for the 17 Phaeophyceae species and the average depth for the four outgroup species was calculated and OGs where the depth for outgroups was more than half the depth for the Phaeophyceae were discarded. We retained 227 out of 233 orthogroups after this step.
3) Finally, functional annotations were used to remove orthogroups that were likely to correspond to transposable elements. A final list of 180 OGs was retained (Table S3).

The amplified gene families were manually categorised into functional classes based on the output of automatic functional annotation programs (InterProScan,[137] EggNOG,[124] nr BLASTp) and an amplification profile was assigned to each orthogroup by identifying the taxonomic group where the amplification of the family was most marked (Table S3).

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

Composite genes

The amino-acid sequences of all 530,598 genes present in the selected genomes were compared in an all-against-all pairwise alignment using DIAMOND BLASTp[107] version 2.0.11; "very-sensitive" mode; e-value threshold $1e^{-5}$. This raw alignment was then filtered using CleanBlastp, from the CompositeSearch suite,[121] to remove sequence alignments with under 30% residue identity and produce the final sequence similarity network. CompositeSearch was then used on this network to identify putative composite gene families among the orthologous groups (OGs) previously computed by OrthoFinder.[113] Composite OGs containing two or more genes and having non-overlapping regions aligned to their component OGs were retained for further analysis, while singleton composite OGs and composites with overlapping component regions were discarded. A phylogeny-based approach,[184] which uses information from extant genomes to apply a Dollo parsimony model in Count,[114] was used to reconstruct the evolutionary events

(domain fusions and fissions) that led to structural rearrangements of composite genes, allowing them to be labelled as fusion or fission events (or as complex events when sequentiality could not be clearly deduced).

Horizontal gene transfer (HGT)

Dataset and experimental approach

Uneven data collection across taxa can impact HGT identification. The phylogeny-based HGT screening approach used here requires the establishment of a comprehensive and taxonomically diverse reference dataset. The analysis focused on the Phaeoexplorer genomes using a background database called REFAL and an automated bioinformatics tool called RoutineTree, which screens for HGTs using phylogenetics. The background database was built using a starting database, GNM1157, which includes a diverse set of 17,250,679 protein sequences from 1157 genomes spanning various prokaryotic and eukaryotic lineages (540 bacteria, 45 archaea, 431 Opisthokonta, 15 Rhodophyta, 83 Viridiplantae, and 43 genomes from CRASH lineages). Data from NCBI RefSeq (updated as of May 2020) and MMETSP were integrated into GNM1157 to form the background database REFAL. To enhance data quality and reduce redundancy, CD-HIT version 4.5.4 was used to remove highly similar sequences (with sequence identity $\geq$90%) within each taxonomic order. This curation process resulted in a protein database consisting of 39.9 million sequences, representing over 7,786 taxa and providing comprehensive coverage across the diverse branches of the tree of life. To obtain the best assembled genome within a genus, the latest version was selected if multiple versions were available. In addition, the dataset was expanded by searching for genomes in other repositories such as the Joint Genome Institute. Special attention was paid to achieving balanced representation of the Rhodophyta and Viridiplantae, which are particularly crucial for HGT analysis within the Chromalveolate group. To accomplish this, protein data from six red algal transcriptomes sourced from MMETSP was added. The HGT search was applied to 72 Stramenopile genomes, including 45 newly sequenced and 27 public genomes.

Phylogenetic Tree Reconstruction

The pipeline for constructing phylogenetic trees splits fasta files into individual sequence files and then carries out a search for homologous sequences, followed by multiple sequence alignment and tree-building. Nested positions within the trees were identified using artificial intelligence and hU and hBL methods were used for HGT verification. Instead of using all available sequences, sequences with the best BLAST hit scores from each kingdom, phylum, and class were used for tree construction to expedite tree-building and enhance clarity. Each gene, regardless of whether it was a copy or not, was used as a query for tree construction. To improve precision, four different methods were used for tree building: neighbour-joining, maximum parsimony, maximum likelihood and Bayesian. As a result, each node within a tree was associated with four support values. To create single-gene phylogenetic trees, a BLASTp[84] search was carried out against the background database, employing an e-value cutoff of $1e^{-05}$. For each query, the top 1,000 significant matches were sorted by bit-score in descending order as the default criterion. Matching sequences were then retrieved from the database, with a constraint of no more than three sequences per genus and no more than 12 sequences per phylum. To further refine the selection, significant matches with a query-subject alignment length of at least 120 amino acids were re-sorted based on query-subject identity in descending order. A second set of homologous sequences was then retrieved from the database following the same procedure. These two sets of homologous sequences, along with the query, were merged and aligned using MUSCLE[115] version 3.8.31 with default settings. The resulting alignments, trimmed to a minimum length of 50 amino acids using TrimAl[133] version 1.2 in automated mode (-automated1), were used to construct phylogenetic trees with FastTree version 2.1.7, with the 'WAG + CAT' model and four rounds of minimum-evolution SPR moves (-spr 4) along with exhaustive ML nearest-neighbour interchanges (-mlacc 2 -slownni). Branch supports were estimated using the Shimodaira-Hasegawa (SH)-test.

Inferring HGT based on tree topology

Phylogenetic trees were examined to identify specific topologies where Phaeoexplorer query sequences were nested among other sequences, defined as a situation where two or more monophyletic clades consist of both queries and prokaryotic sequences, supported by distinct nodes within the tree. These monophyletic clades are considered to group together if they share the same set of prokaryotic sequences but differ in sequences from optional taxa. Singletons for both the donor and receptor genes were excluded to minimise contamination and recent HGT interference. To retain only robustly supported nested positions, positions were required to be multiply supported, with a minimum of $\geq$0.70 for the SH-test and aBayes-test support from at least two Phaeoexplorer receptor nodes and three donor supporting nodes. Furthermore, queries that displayed significantly different amino acid compositions ($P < 0.05$) compared to the remaining sequences in the alignment were discarded. Queries from the CRASH category that nested among sequences from other kingdoms (supported by >70% UFBoot at one or more supporting nodes) were retained.

Enhancing accuracy and establishing the timing of HGTs

To enhance accuracy, a minimum requirement was imposed for all supporting nodes and for strongly supported nodes that indicate query-donor monophyly. To determine the timing of HGT events, temporal information, primarily derived from the timetree database, was incorporated into each node. We assigned the "smallest boundary" role to pinpoint the most recent common ancestor at the time of the HGT event. Essentially, if all descendants of a given query protein sequence can be traced back to the initial HGT event, a common ancestral node can be identified whose occurrence time can be inferred using a molecular clock approach based on archaeological and fossil evidence. The taxonomy boundaries of HGT descendants were determined by identifying the smallest ancestor shared by both the donor and receptor taxa from the monophyletic clades within the tree. By considering the emergence

times of both taxa, the timing of the transfer of genes from earlier taxa to later taxa can be determined, as the reverse scenario is not considered plausible.

Verification of HGTs

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

Further validation was based on the following concepts:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gene codon usage, functional annotation and expression

Indices of codon usage and GC content were calculated using Codonw 1.4.4 (http://codonw.sourceforge.net). Gene functions were assigned by searching against the Gene Ontology (GO) database using blast2GO (ref blast2GO 08) and the KEGG database using blastKOALA (http://www.kegg.jp/blastkoala/) with default parameters. The full gene sets of each species were set as the background for KEGG and GO enrichment analyses by applying Student's t-test (p-value cutoff = 0.01). HGTs were also analysed with SEED (http://www.theseed.org/wiki/Home_of_the_SEED), IPR2GO (http://www.ebi.ac.uk/interpro/search/sequence-search), eggNOG[124] (http://eggnogdb.embl.de/#/app/home) and Pfam.[175] For each species, the differences between mean gene expression levels for HGTs and non-HGT genes with common GO terms were accessed using Student's t-test. Go terms with less than five genes in either gene category were ignored. The differences in expression dispersal (coefficient of variation: standard deviation across genes or samples / mean value) and expression specificity (frequencies of a gene to be detected as unexpressed, defined as transcripts per kilobase million (TPM) = 2, in any condition) were accessed in a similar manner. Given the variable experimental conditions associated with different transcriptome data for each species, gene expression values for a gene were used indiscriminately regardless of the conditions. Correlation tests between the codon adaptation index (CAI) and gene expression were carried out using the Spearman's rank correlation analysis tool (P. Wessa, Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7, https://www.wessa.net/).

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

Genes identified as belonging to orthogroups that were predicted to be gained at specific nodes of the phylogenetic tree based on the Dollo parsimony analysis, to belong to either significantly amplified gene families (binomial analysis) or to belong to gene families that have significantly changed in size over evolutionary time (CAFE5 analysis), to correspond to founder events (Phylostratigraphy analysis), to have been remodelled (composite gene analysis) or to have been derived from an HGT (HGT analysis) were extracted from the output of each of these analysis and aggregated in a single datatable. Correspondences were established manually between phylogenetic tree nodes and phylostrata and this information was integrated into the datatable. Counting and calculations of the

frequency of events at specific time points were carried out using *ad hoc* R scripts (R version 4.4.1) and the tidyverse[167] package (version 2.0.0). Graphs were generated using the ggplot2[166] package (version 3.5.1). For each gene, a COG functional category was retrieved from the eggNOG mapper output and the COG enrichment analysis was carried out in R using the clusterProfiler[165] package (version 4.6.2) by comparing each set of gene families with the full set of gene families.

Detection of viral genome insertions and viral regions in algal genomes

To reduce the dataset size for analysis, 64 Phaeoexplorer and eight public genomes were initially filtered to retain only contigs that were more than 10 kbp in length. Gene prediction was then carried out on all contigs using Prodigal[126] (V2.6.3, settings: default, meta) and the resulting proteins were used as queries against the NCVOG[171] and VOGDB[185] databases using hmmscan (HMMER 3.3.2 with default settings). The contigs detected by hmmscan were then filtered to retain only sequences with at least one match to either viral database at a defined e-value cutoff ($1e^{-20}$ for NCVOG, and $1e^{-80}$ for VOGDB). The resulting positive 4,951 contigs were then analysed using ViralRecall[127] version 2.0 with settings -w 50 -g 1 -b -f -m 2 using the built-in Nucleocytoviricota (NCV) database GVOG and a window size of 50 kbp. To ensure that viral genes were not missed because they had not been annotated by Prodigal, six-frame translations of the contigs were generated using esl-translate (version 0.48 with default settings), and the resulting proteins queried against the same databases used by ViralRecall using hmmsearch (HMMER 3.3.2, settings: -E 1e-10). The ViralRecall results were then parsed using an in-house workflow. Six-frame translations were removed from the results if they overlapped (even partially) with any Prodigal gene prediction, as identified using bedtools[128] (v2.29.2; intersect). Likewise, overlapping six-frame translations and gene predictions with the same NCVOG match were removed to reduce redundancy. Based on the distance between query sequences with the same GVOG hit, queries were flagged as frame-shifted (less than 100 bp gap), intron-containing (100-5,000 bp gap) or mono-exonic (greater than 5,000 bp gap). All queries were also checked for overlaps with multi-exonic genes that had been annotated by the Phaeoexplorer gene prediction procedure (using Gmove[109]), and flagged if they did. All queries were then filtered to retain only those that matched a set of key NCV marker genes, identified by NCVOG code (A32, D5 helicase, D5 DNA primase, MCP, DNA polymerase B, SFII and VLTF3) or some Phaeovirus integrase genes (integrase recombinase, integrase resolvase and RNR). The marker gene proteins were clustered with the protein sequences of NCVOGs using MMseqs cluster[129] (version 13.45111 with settings –min-seq-id 0.3 -c 0.8). Finally, the parsed results of the NCV marker gene set identified by the ViralRecall screen were manually curated, retaining only those queries with varying combinations of the following properties: placement within a viral region as identified by ViralRecall, similar hmmsearch results (score and e-value) and gene length to that of known NCV genes, not part of a multi-exonic gene, lack of Pfam HMM matches to cellular domains sharing homology to the marker gene (specific to certain marker genes), and clustered with an NCVOG in the MMseqs analysis. We noted that the median number of viral regions found in genomes assembled with long reads was very similar to that for genomes assembled with short reads (9 and 10, respectively). The marker gene content of the viral regions was manually assessed to estimate the number of complete or partial inserted viruses in each genome. VRs were considered to be complete proviruses if they contained all seven of the key NCV marker genes listed above. VRs were classed as partial proviruses if they only contained a subset of the seven key NCV marker genes, the presence of the MCP and DNA polymerase B genes being particularly strong indicators of a partial provirus.

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

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

Phylogenetic analysis of viral genes

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

Metabolic networks

Genome-scale metabolic networks were reconstructed using AuCoMe[46] version 0.5.1 using the MetaCyc[187] version 26 database. A first dataset, consisting of the 60 species listed in Table S1F (column "Metabolic networks") plus two public diatom genomes

already used in the initial AuCoMe study (*Fragilariopsis cylindrus* and *Fistulifera solaris*) was processed to build the largest possible database (phaeogem) for exploratory comparisons (https://gem-aureme.genouest.org/phaeogem/). Then, a second comparison was performed on all long-read species plus outgroups. Based on Multidimensional-scaling (MDS) analyses, the most divergent long-read species (*Choristocarpus tenellus*, *Laminaria digitata*, *Phaeothamnion wetherbeei* and the public genome of *Sargassum fusiforme*) were excluded to construct a 16 species dataset, balancing assembly quality and phylogenetic coverage (https://gem-aureme.genouest.org/16bestgem/). MDS plots were build using the vegan package, version 2.6-4 (https://github.com/vegandevs/vegan) with R 4.1.2,[162] using Jaccard distances. A third stricter dataset (fwgem), enriched in high-quality long-read Ectocarpales, was built to address questions related to freshwater adaptation (https://gem-aureme.genouest.org/fwgem/). A set of reactions that were overrepresented in brown algae compared to the outgroup was created by taking reactions present in 100% of brown algae and less than 70% of outgroups. Reactions corresponding to genes lost in freshwater species were also extracted. These reaction sets were extracted from all the networks using the Aucomana library (https://github.com/PaulineGHG/aucomana). Online wikis (phaeogem, 16bestgem and fwgem) were generated using AuReMe.[188]

CAZymes

CAZyme genes were identified based on shared homology with biochemically characterised proteins, either individually or as hidden Markov model (HMM) profiles. For phylogenetic analyses, proteins were aligned using MAFFT[130] with the iterative refinement method and the scoring matrix Blosum62. The alignments were manually refined and trees were constructed using the maximum likelihood approach. Alignment reliability was tested by a bootstrap analysis using 100 resamplings of the dataset. Only bootstrap values above 60% are shown. The phylogenetic trees were displayed with MEGA.[131] The annotated genes are listed in Table S4B with accession numbers.

Sulfatases

The sulfatases encoded by each brown algal genome were identified and assigned to their respective family and subfamily using the SulfAtlas database[173,174] (https://sulfatlas.sb-roscoff.fr/). Each predicted proteome was first submitted to the SulfAtlas HHM server (https://sulfatlas.sb-roscoff.fr/sulfatlashmm/), which allows rapid identification of sulfatase candidates and (sub)family assignment using hidden Markov model profiles for each SulfAtlas (sub)family. Each sulfatase candidate sequence was then used as a query in a BLASTp[84] search against the SulfAtlas database (https://blast.sb-roscoff.fr/sulfatlas/). Sequences with at least 50% identity with sulfatases from marine bacteria or other marine microorganisms were considered to be contaminants. Below this threshold, additional examination of the predicted gene structure and genomic context of the candidate sequence was undertaken to identify possible horizontal gene transfers.

Haloperoxidases

vHPO genes were identified based on sequence homology and active site conservation. Maximum likelihood phylogenetic analyses were carried out using the NGphylogeny platform at https://ngphylogeny.fr/. MAFFT was used to align vHPO sequences and alignments were automatically curated with TrimAl,[133] leading to the selection of 444 informative positions from the initial 1450 positions for the algal-type vHPOs and 402 informative positions from the initial 1078 positions for the bacterial-type vHPOs. Maximum likelihood trees were constructed using FastTree with the WAG+G gene model and 1000 bootstrap replicates. Maximum likelihood Newick files were formatted as circular representations using iTOL. Only bootstrap values between 0.7 and 1 were conserved. The lists of annotated vHPO genes are in Tables S4C and S4D.

Ion channels

A search was carried out for 12 classes of ion channel in the predicted proteomes of the 21 Phaeoexplorer reference genomes plus those of two diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. Predicted proteomes were screened using BLASTp[84] and query sequences from *Ectocarpus* species 7 and seven other species from diverse eukaryotic taxa.

Membrane-localised proteins

Membrane protein family genes were identified either by carrying out BLASTp[84] searches of the predicted Phaeoexplorer proteomes using *Ectocarpus* species 7 sequences as queries or by recovering orthogroups containing the relevant *Ectocarpus* species 7 sequences as members. The BLASTp approach was used for DEK1-like calpains, fasciclins, tetraspanins, CHASE, ethylene-binding-domain-like and MASE1 domain histidine kinases whereas the orthogroup approach was used to recover other members of the histidine kinase family. Both approaches were used to search for integrins and transmembrane receptor kinases. For integrins the two methods detected exactly the same set of proteins. For receptor kinases the BLASTp and orthogroup analyses detected 99.3% and 98.3% of the 269 genes, respectively. For these analyses, either the whole genome dataset was analysed or only the set of 21 reference genomes (Table S1F), depending on the size of the gene family.

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

Transcription-associated proteins

TAPscan v4[135] was used to analyse the transcription-associated protein (TAP) complements of 21 species. TAPscan[134] is a comprehensive tool for annotating TAPs based on the detection of highly conserved protein domains using HMM profiles with specific thresholds and coverage cut-offs. Following detection, specialised rules are applied to assign protein sequences to TAP families based on the detected domains. TAPscan v4 can assign proteins to 138 different TAP (sub)families with high accuracy.

EsV-1-7 domain proteins

EsV-1-7 domain proteins were identified in the 31 brown algal and sister taxa genomes (Table S1F) by recovering the members of all orthogroups (with the exception of OG0000001, which is a very large OG that consisting principally of transposon sequences) that either contained one or more of a curated set of 101 EsV-1-7 domain proteins[62] for *Ectocarpus* species 7 or contained an EsV-1-7 domain protein based on a match to the Pfam EsV-1-7 motif PF19114. The recovered proteins were screened manually for the presence of at least one EsV-1-7 domain and a total of 2018 were finally identified as members of the EsV-1-7 family.

To identify orthologues of the EsV-1-7 protein IMMEDIATE UPRIGHT[62] (IMM), BLASTp searches of 25 brown algal and four sister taxa predicted proteomes were carried out with the amino-terminal domain of the IMM protein minus the five EsV-1-7 repeats as this domain is unique to IMM. Proteins were retained as IMM orthologues if they were more similar to IMM than to the most closely-related protein in *Ectocarpus* species 7, Ec-17_002150.

Histones

Histone protein sequences were analysed in *Ascophyllum nodosum*, *Chordaria linearis*, *Chrysoparadoxa australica*, *Desmarestia herbacea*, *Dictyota dichotoma*, *Discosporangium mesarthrocarpum*, *Ectocarpus crouaniorum*, *Ectocarpus fasciculatus*, *Ectocarpus siliculosus*, *Fucus serratus*, *Heterosigma akashiwo*, *Pleurocladia lacustris*, *Porterinema fluviatile*, *Pylaiella littoralis*, *Saccharina latissima*, *Sargassum fusiform*, *Schizocladia ischiensis*, *Scytosiphon promiscuus*, *Sphacelaria rigidula*, *Tribonema minus* and *Undaria pinnatifida* using BLASTp against the complete predicted proteomes (https://blast.sb-roscoff.fr/phaeoexplorer/) with the histone protein sequences from the diatom *Phaeodactylum tricornutum* as queries. The genes and transcripts coding for the identified histones were then retrieved from the genomes and predicted transcripts using BLAST (https://blast.sb-roscoff.fr/phaeoexplorer/). The proteins encoded by the identified genes and transcripts were predicted with the Expasy web translator (https://web.expasy.org/translate/). In order to identify truncated proteins or incorrect start codons, the following constraints were applied: H2A proteins must start with the SGKGKGGR sequence, H2B with AKTP, canonical H3.1 and variants H3.3 with ARTKQT and H4 with SGRGKGGKGLGKGG. For the linker histone H1, protein sequences had to be lysine-rich and sequences with incorrect start codons were determined by alignments of all identified H1 proteins. For proteins with incorrect start codons, the region upstream of the correct start codon was removed. For truncated proteins, *i.e.* proteins whose transcripts lacked either the start (no methionine) or stop codons, the protein sequence was completed based on alignment with the corresponding genomic region using the Geneious 11.0.5 software. When the sequence could not be completed, a BLAST was performed against the Phaeoexplorer *de novo* transcriptomes (https://blast.sb-roscoff.fr/phaeoexplorer_denovo/) when this data was available (this was not possible for the public genomes *T. minus*, *U. pinnatifida* and *S. fusiforme*). Based on the nomenclature established by,[189] H3 histones were classified as follows: canonical H3.1 proteins harbour AT residues at positions 31-32 while histone variants H3.3 harbour TA residues, H3 proteins with other residues at positions 31-32 were named H3.4 and so on. CenH3 variants of H3 were identified by analysis with Panther 17.0 (www.pantherdb.org/tools/sequenceSearchForm.jsp?) and/or Interproscan[137] 94.0 (www.ebi.ac.uk/interpro/search/sequence/).

Species abbreviations used in histone phylogenetic trees are: Atr, *Amborella trichopoda*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Di, *Dictyostellium discoideum*; Dr, *Danio rerio*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Pp, *Physarum polycephalum*; Ppa *Physcomitrium patens*; Sc, *Saccharomyces cerevisiae*; Tm, *Tetrahymena thermophila*; Zm, *Zea mays*; Mp, *Marchantia polymorpha* subsp. *Ruderalis*; Bd, *Brachypodium distachyon*; Ccr, *Chondrus crispus*; Gs, *Galdieria sulphuraria*; Cm, *Cyanidioschyzon merolae*; Cr, *Chlamydomonas reinhardtii*; Ol, *Ostreococcus luciminarinus*; Ot, *Ostreococcus tauri*; To, Thalassiosira oceanica; Pt, *Phaeodactylum tricornutum*; An, *Ascophyllum nodosum*; Cl, *Chordaria linearis*; Ca, *Chrysoparadoxa australica*; Dh, *Desmarestia herbacea*; Ddi, *Dictyota dichotoma*; Dme, *Discosporangium mesarthrocarpum*; Ec, *Ectocarpus crouaniorum*; Ef, *Ectocarpus fasciculatus*; Es, *Ectocarpus siliculosus*; Fse, *Fucus serratus*; Ha, *Heterosigma akashiwo*; Pla, *Pleurocladia lacustris*; Pf, *Porterinema fluviatile*; Pli, *Pylaiella littoralis*; Sl, *Saccharina latissima*; Sf, *Sargassum fusiform*; Si, *Schizocladia ischiensis*; Sp, *Scytosiphon promiscuus*; Sri, *Sphacelaria rigidula*; Tm, *Tribonema minus*; Up, *Undaria pinnatifida*.

DNA methyltransferases

Searches were carried out for methyltransferases and demethylases in the predicted proteomes of 20 of the high quality brown algal reference genome assemblies (based on Nanopore long-read sequence) plus the sister taxa *Chrysoparadoxa australica* and *Schizocladia ischiensis* using BLASTp (Table S1F). A methyltransferase reference database was constructed by recovering sequences from NCBI, ENSEMBL and UniProtKB. Methyltransferase sequences were recovered for stramenopiles such as *Nannochloropsis gaditana*, the diatom *Phaeodactylum tricornutum*, the oomycete *Phytophthora infestans* and for species from more distant lineages including *Arabidopsis thaliana*, *Homo sapiens* and the fungus *Neurospora crassa*. The proteomes of the selected brown algal strains were then queried against this database using BLASTp and matches with an e-value of < 0.001, a bit score > 70, a maximum gap of 5 and percentage identity of >30% were retained. The retained matches were screened against the NCBI, UniProt and

SwissProt databases to identify and remove contaminating bacterial or viral proteins. Methyltransferase domains were detected in the retained matches using the Simple Modular Architecture Research Tool (SMART)[177] domain architecture analysis and InterPro searches (https://www.ebi.ac.uk/interpro/). Sequences with methyltransferase domains were retained for further analysis. Validated brown algal methyltransferases were aligned with reference methyltransferases using Clustal[138] 2.1.

Spliceosome

Components of the Major Spliceosome were identified using a reference set of 147 human components (https://www.genenames.org/data/genegroup/#!/group/1518), excluding the five small nuclear RNAs (snRNAs). Including isoforms, this query set consisted of 626 proteins. These proteins were used to screen the predicted proteomes of 54 genomes (Table S1F) using BLASTp and matches were retained if they had an e-value of at most $1e^{-40}$ and coverage >30%. Searches were also carried out for components of LSm and Sm complexes which have roles as scaffolds in the formation of ribonucleoprotein particles (RNPs), in the maturation of mRNAs (including splicing, such as the cytoplasmic complex LSm1-7, LSm2-8 which is part of the core U6 snRNP and other complexes important for the formation of the 3' ends of histone transcripts), in the assembly of P-Bodies and in the maintenance of telomeres.

Flagella proteins

A previous proteomic analysis of anterior and posterior flagella of the brown alga *Colpomenia bullosa* identified a total of 592 proteins across the two proteomes.[41] Here the *Ectocarpus* species 7 orthologues of 70 of these proteins that had been detected with a very high level of confidence were used to identify the corresponding orthogroups and the presence or absence of these orthogroups was scored for seven representative species (Table S1F).

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

Six independent cultures of *Porterinema fluviatile* were cultivated for four weeks in 140 mm Petri dishes with Provasoli-enriched culture medium,[190] which was renewed every two weeks. For three Petri dishes, the culture medium was based on autoclaved natural seawater (high salinity treatment), for the other three Petri Dishes natural seawater was diluted 1:19 vol/vol with distilled water (low salinity treatment). Cultures were harvested with 40 μm nylon sieves, dried with a paper towel, and immediately frozen in liquid nitrogen. RNA extraction library construction and sequencing were carried out as described in section "RNA extraction, Illumina RNA-seq library preparation and sequencing". RNA-seq reads were cleaned with Trimmomatic[99] V0.38 and then mapped to the *P. fluviatile* genome using Kallisto[140] version 0.44.0. Differentially expressed genes were identified using the DESeq2 package[141] included in Bioconductor version 3.11, considering genes with an adjusted $p < 0.05$ and a $\log_2$ fold-change > 1 as differentially expressed. To compare the differentially expressed genes in *P. fluviatile* with an equivalent set previously identified for *Ectocarpus subulatus* in a microarray experiment using nearly identical growth conditions,[191] orthologues in the two species were detected using Orthofinder version 2.3.3. Of the 10,066 shared orthogroups, 6,606 had microarray expression data for *E. subulatus*. This information was used to classify differentially expressed genes for the two species as either shared orthologues or as lineage-specific.

Identification of genes with generation-biased expression patterns

RNA-seq data (two to five replicates per condition) was recovered for gametophyte and sporophyte generations of ten species (Table S1F). Data quality was assessed with FastQC[142] version 0.11.9 and sequences were then trimmed with Trim Galore version 0.6.5 with the parameters –length 50, - quality 24, –stringency 6, –max_n 3. The cleaned reads were mapped onto the corresponding genome for each species using HISAT2 version 2.1.0 with default options. Counting was carried out with featureCounts[145] from the subread package (version 2.0.1) on CDS features grouped by Parent. Transcript Per Kilobase Million (TPM) tables were generated for all conditions and differentially expressed genes were detected using DESeq2[141] version 1.30.1. Genes were classified into six categories based on the differential expression analysis and the TPM values: gametophyte-biased, mean TPM $\geq$1 in gametophyte and sporophyte, $\log_2$(fold change) $\geq$1, adjusted *p*-value <0.05; sporophyte-biased: mean TPM $\geq$1 in gametophyte and sporophyte, $\log_2$(fold change) $\leq$-1, adjusted *p*-value <0.05; gametophyte-specific, mean TPM <1 in sporophyte and $\geq$1 in gametophyte, $\log_2$(fold change) $\geq$1, adjusted *p*-value <0.05; sporophyte-specific, mean TPM <1 in sporophyte and $\geq$1 in gametophyte, $\log_2$(-fold change) $\leq$-1, adjusted p-value <0.05; unbiased genes: mean gametophyte and sporophyte TPMs $\geq$1, $\log_2$(fold change) <1 or >-1 and/or adjusted p-value $\geq$0.05; unexpressed genes, mean gametophyte and sporophyte TPM <1.

Life cycle and thallus architecture

Genome dataset and traits

To study the impact of body architecture, the brown algae were divided into three categories: 22 filamentous species, eight simple parenchymatous species and 13 species with elaborate thalli (Table S1F). For the life-cycle-based assessment, the groups were: 30 haploid-diploid species and six diploid species (Table S1F). Body architecture information was available for 43 species, and life cycle information was available for 36 species; species without body plan or life cycle information were not used in subsequent analyses. Two approaches were used to estimate selection intensity across the phylogeny, (i) a model-based method, and (ii) by evaluating codon usage bias and nucleotide composition. Two evolutionary models were used, one based on architecture and the other based on life cycle. For model-based methods the phylogeny was categorised based on the above traits, and selection intensity parameters were estimated using PAML[146] version 4.9i. Rate estimates were obtained for non-synonymous substitutions (dN), synonymous

substitutions (dS) and omega (dN/dS) for the multiple sequence alignments of all genes within each orthogroup using the variable-ratio model of CODEML from PAML, which allows different omegas for different branch categories. The traits were assigned to the branches of the phylogeny using ancestral state estimation by stochastic mapping with the phytools R package.[147,162]

Evolutionary models to study impacts of body architecture

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

Evolutionary models to study the impacts of life cycle

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

Selection of intensity parameters

Omega (dN/dS) provides an estimate of the ratio of substitutions at sites under selection compared to neutral sites, and is generally used to infer the strength of purifying selection. Omega needs to be interpreted with caution because not all synonymous sites are neutral[192] and also synonymous substitutions are often underestimated due to saturation of synonymous sites, which might in turn impact the omega ratios.[193] Omega values lower than one indicate substitutions are less frequent at sites under selection compared to neutral sites and are characteristic of highly conserved genes or genes evolving under strong purifying selection. As we used primarily low copy number genes in this study, the analysed genes were expected to evolve under strong purifying selection, with omega values much lower than one. Using omega for near neutral studies is challenging because near neutral sites are determined by effective population size, that is to say, sites under mild selection constraint in larger populations can behave as neutral sites in smaller populations. It is therefore difficult to infer the amount of mutation from relative values of omega. In order to obtain better insight into selection intensity, mutation accumulation was not only investigated using rates of synonymous (dS) and non-synonymous (dN) substitutions but also by estimating codon bias and nucleotide composition. Codon usage bias was used, in addition to omega, to infer selection intensity across species as the former reflects selection efficacy at synonymous sites.[194–196] We inferred codon usage bias by estimating the effective number of codons (ENC) for each species using the enc method from the VHICA package.[148,182] The effective number of codons (ENC) quantifies the extent of deviation of codon usage of a gene from equal usage of synonymous codons. For the standard genetic code, ENC values range from 20 (where a single codon is used per amino acid implying strong codon usage bias) to 61 (implies that all synonymous codons are equally used for each amino acid[197]). Low ENC indicates constrained use of codons, which potentially highlights stronger codon bias due to stronger selection at synonymous sites. As nucleotide composition can also influence codon bias, we calculated the overall GC composition, GC at the third codon position (GC3) and the theoretical expected ENC (EENC) based on GC3 using local R scripts. The lower the observed ENC (OENC, estimated from the gene sequence) relative to EENC, the stronger the influence of selection due to translation on codon usage. This was studied by estimating the difference (DENC = EENC - OENC) between the expected ENC and the observed ENC.[198] Positive DENC indicates a role for selection constraints on codon usage in addition to the influence of nucleotide composition. DENC values of zero or less indicate that codon bias is entirely driven by nucleotide composition. DENC values were used to study the influence of translation selection and nucleotide composition on codon usage bias.

Assembly and analysis of organellar genomes

Plastid and mitochondrial genomes were assembled *de novo* using NOVOPlasty[149] v3.7 and *rbcL* and *cox1* nucleotide sequences as seeds. Assembled genomes were checked by aligning reads using Bowtie2[86] v2.3.5.1 and processed with SAMtools[150] v1.5. Annotation of protein-coding genes was performed with GeSeq[151] v2.03. Annotation of tRNAs, tmRNAs and rRNAs was performed with ARAGORN[152] v1.2.38.

Maximum-likelihood (ML) phylogenetic trees were constructed using 92 plastid genomes (11 non-brown outgroup sequences) and 89 mitochondrial genomes (seven non-brown outgroup sequences). The conserved coding-region amino acid sequences of 139 plastid genes (31,159 amino acids) and 35 mitochondrial genes (7,461 amino acids) were used to construct these phylogenetic trees. The sequence for each gene was aligned individually using MAFFT[130] v7 (–maxiterate 1000) and then concatenated. Alignment partitions were assigned based on genes. Each of the aligned gene sequences was trimmed with trimAl[133] v1.2 (-automated1). ML phylogenetic trees were constructed with IQ-TREE 2.[163] The protein substitution models in each gene partition were selected using ModelFinder.[153] Statistical support for tree branches was assessed with 1,000 replicates of ultrafast bootstrap (UFBoot2).[154]

Analysis of *Ectocarpus* genome synteny

Global genome synteny analysis was performed using SynMap[155] on the CoGe platform (https://genomevolution.org/coge/) with the following genomes: *Ectocarpus crouaniorum* male, *Ectocarpus fasciculatus* male, *Ectocarpus siliculosus* male, *Ectocarpus* species 7 male and *Ectocarpus subulatus*. SynMap identifies syntenic regions between two or more genomes using a combination of sequence similarity and collinearity algorithms. Last[199] was used as the BLAST algorithm and syntenic gene pairs were identified using DAGChainer[156] with settings "Relative Gene Order", -D = 20, -A =5. Neighbouring syntenic blocks were merged into larger blocks. Substitution rates between the synthetic CDS pairs were calculated using CodeML,[146] which was also implemented in SynMap, CoGe. In detail, protein sequences were aligned using the Needleman-Wunsch algorithm implemented in nwalign (https://pypi.org/project/nwalign/) and then translated back to aligned codons. CodeML was run five times for each alignment using the default parameters and the lowest dS was retained, with the upper cutoff for dS values set at 2. *Ectocarpus* genes were grouped according to their age based on the phylostratigraphic analysis and by chromosomal location based on their chromosome position in *Ectocarpus* species 7. All plots and statistical analysis were carried out in R version v.4.3.1. Local synteny analysis was based on orthologous genes as identified by Orthofinder.

Analysis of *Ectocarpus* gene evolution

Protein sequence alignments were used to remove gaps with trimAl[133] and then translated back to DNA with backtranseq.[200] Only DNA fasta files with a minimum of 70 bp were retained (831 single-copy orthologs). PhyML trees were built with Geneious v11.1.5 (https://www.geneious.com). Maximum likelihood analysis was carried out to detect site specific, branch-site specific and branch specific positive selection as well as sites under negative selection, using PAML.[201]

Phylogenetic analysis of *Ectocarpus* species

Phylogenetic analysis was carried out for 11 *Ectocarpus* species plus *Scytosiphon promiscuus* as an outgroup (Table S1F). Of the 933 single-copy orthogroups identified for these 12 species, 261 high-confidence alignments were retained for gene tree and species tree inferences following the removal of low-quality alignments using BMGE.[202] Bayesian inference of the phylogeny of the *Ectocarpus* species complex was performed using BEAST[157] v2.7. The analysis was conducted under the multi-species coalescent (MSC) model, implemented in StarBEAST3[158] v1.1.7. The MSC model coestimates gene trees and the species tree within a multispecies coalescent framework, enabling the assessment of incongruences among genes with respect to the species tree. To account for substitution model uncertainty, bModelTest[159] was employed to average over a set of substitution models for each alignment. StarBEAST3 was run under both the Yule model and the strict clock model. A total of 300,000,000 Markov Chain Monte Carlo (MCMC) generations were conducted, with tree states stored every 50,000 iterations. Posterior tree samples were combined, discarding the initial 10% burn-in, using LogCombiner v2.4.7. A maximum clade credibility tree was generated using TreeAnnotator[157] v2.4.7.

Ectocarpus introgression analysis

To distinguish introgression from shared ancestry, D estimates (i.e. ABBA-BABA tests) were generated from 36 four-taxon combinations[203]: four to test the level of introgression within clade 1 (i.e. *E. subulatus*, *E. crouaniorum*, *Ectocarpus* species 1, *Ectocarpus* species 2), 20 to test the level of introgression within clade 2 (i.e. *Ectocarpus* species 6, *Ectocarpus* species 7, *Ectocarpus* species 5, *Ectocarpus* species 9, *E. siliculosus*, *Ectocarpus* species 3) and 12 to test the level of introgression between these two clades. Tests were designed using a four-taxon fixed phylogeny (((P1,P2)P3)O), where P1 and P2 are closely related species from the same clade, P3 is a more divergent species that may have experienced admixture with one or both of the (P1,P2) taxa, and an out-group (O). *E. fasciculatus* was used as the out-group taxon for all ABBA-BABA tests. Details about how P1, P2 and P3 taxa were selected for each test are given in Table S6. Previous results of species tree inference were used to inform subsequent ABBA–BABA tests and to define the (((P1,P2)P3)O) phylogenies. ABBAs are sites at which the derived allele (called B) is shared between the taxa P2 and P3, whereas P1 carries the ancestral allele (called A), as defined by the outgroup while BABAs are sites at which the derived allele is shared between P1 and P3, whereas P2 carries the ancestral allele. Under incomplete lineage sorting, conflicting ABBA and BABA patterns should occur in equal frequencies, resulting in a D statistic equal to zero. Historical gene flow between P2 and P3 causes an excess of ABBA, generating positive values of D. Historical gene flow between P1 and P3 causes an excess of BABA, generating negative values of D. Patterson's D-statistic was calculated for the concatenated alignments of the 261 orthologroups. Significance was detected using a block-jackknifing approach,[203–205] with a block size of 5 kbp. For the jackknife procedure, one block of adjacent sites was removed n times. A Z-score was finally obtained by dividing the value of the D statistic by the standard error over n sequences of 5 kbp. The ParimonySplits network was reconstructed for the genus *Ectocarpus* using SplitsTree 4[160] (version 4.14.6) with 1000 bootstrap replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

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

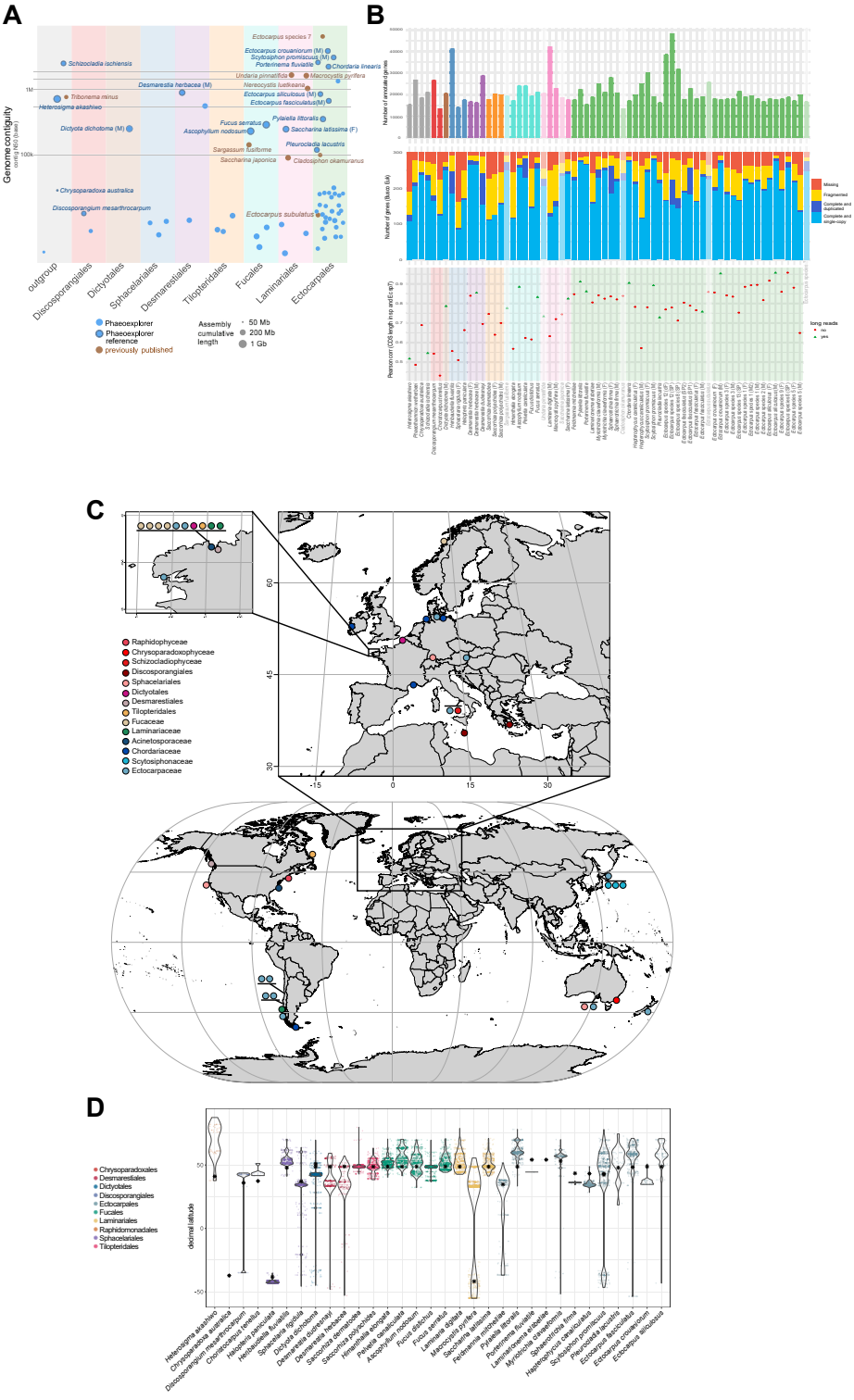
ADDITIONAL RESOURCES

The Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr) provides access to all the annotated genome assemblies described in this study as downloadable files. The output files from the Orthofinder,[113] Interproscan,[137] Hectar[161] and eggNOG-mapper[125] analyses, together with the results of the various DIAMOND[107] sequence similarity analyses (see section "Analyses aimed at deducing functional characteristics of predicted proteins"), can also be downloaded. In addition, the site provides genome browser interfaces for the genomes and multiple additional tools and resources including BLAST interfaces for genomes, proteomes and *de novo* transcriptomes, various experimental protocols, an AskOmics genomic data query interface (PhaeoAskOmics), an RShiny-based transcriptomic aggregator for the model brown alga *Ectocarpus* species 7 strain Ec32, a link to genome-wide metabolic networks for the Phaeoexplorer species and a list of project-related publications.

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

CellPress
OPEN ACCESS

Supplemental figures



(legend on next page)

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

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

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

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

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

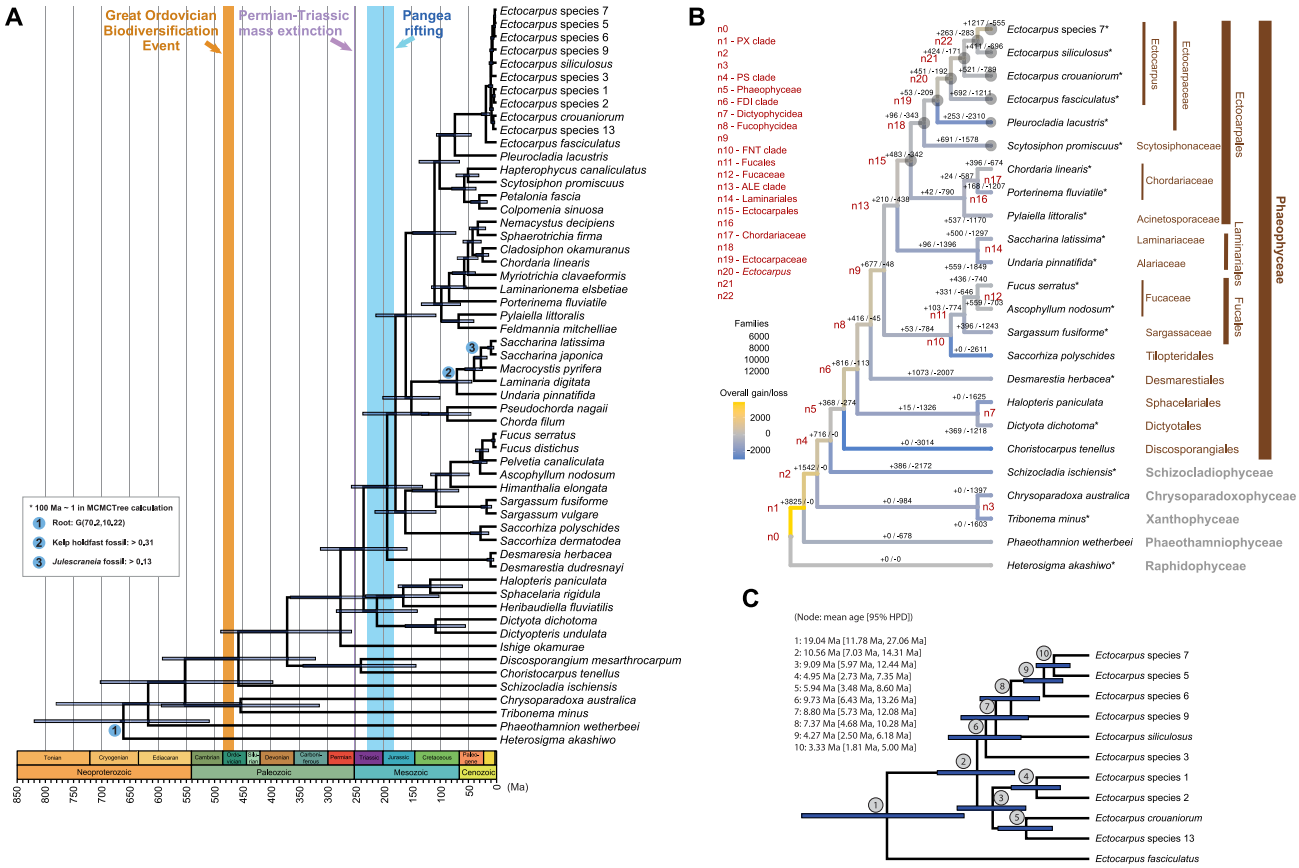
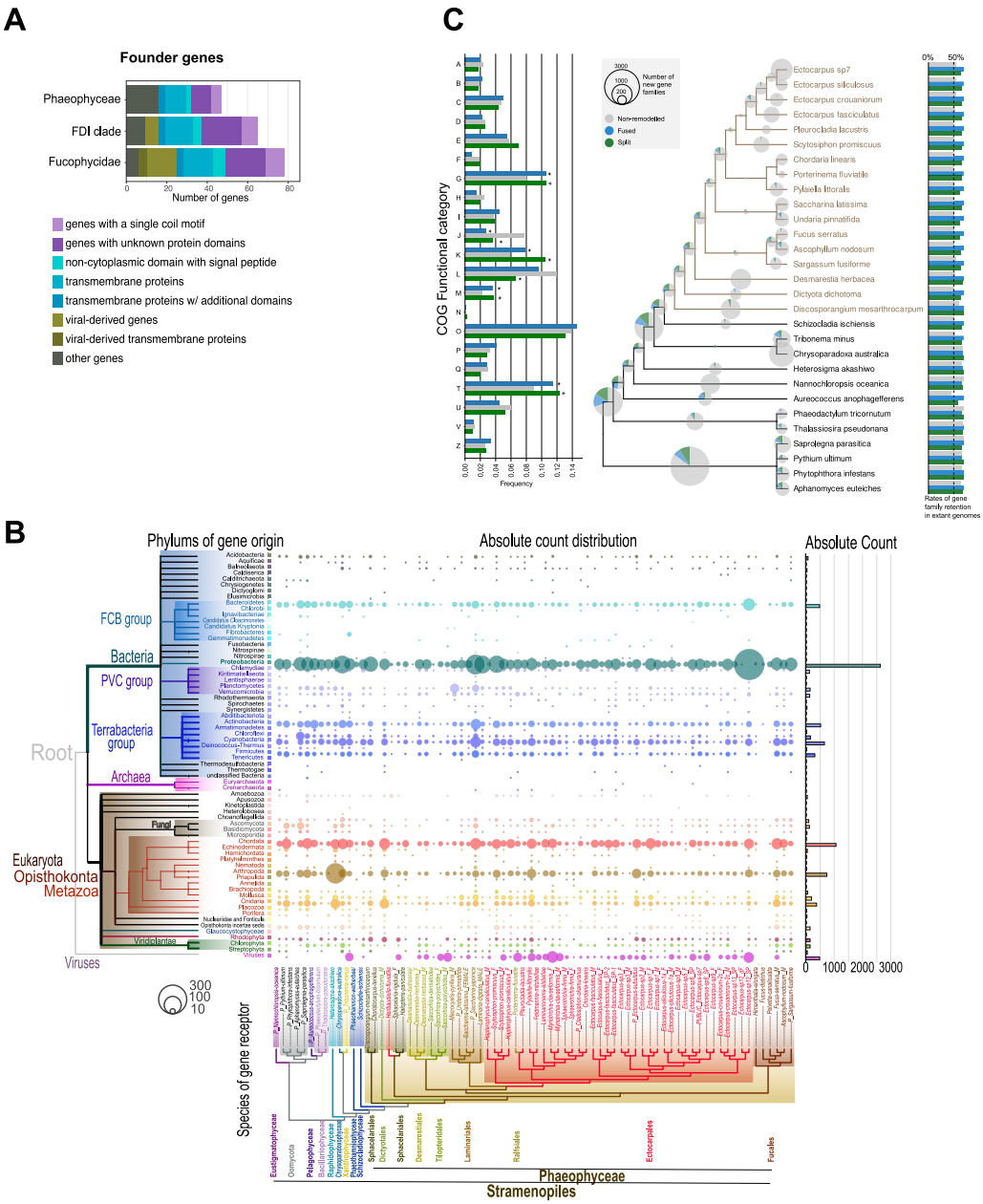


Figure S2. Fossil-calibrated time trees and Dollo logic analysis of orthogroups, related to Figures 1 and 5

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

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

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



(legend on next page)

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

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

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

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

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

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

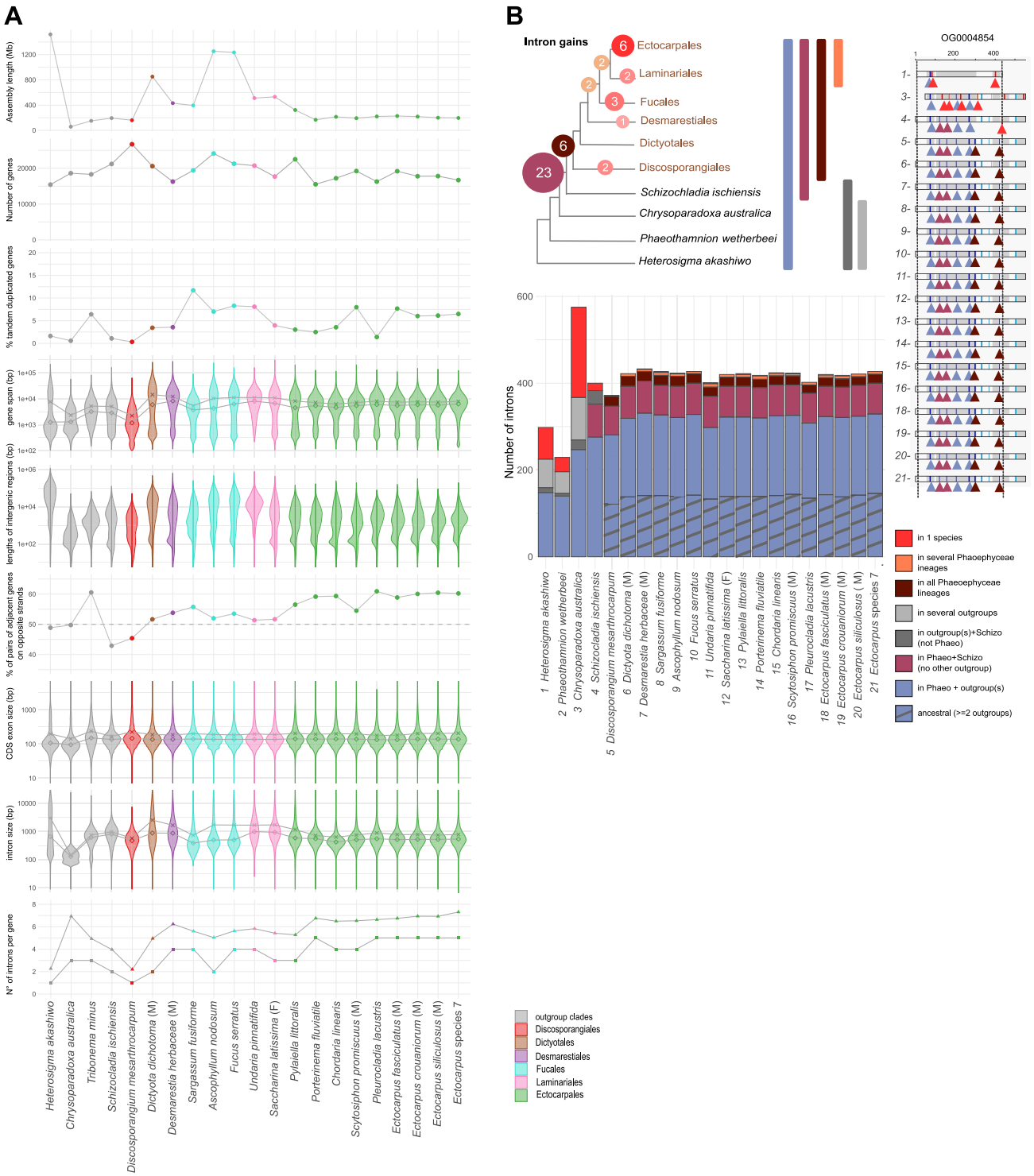
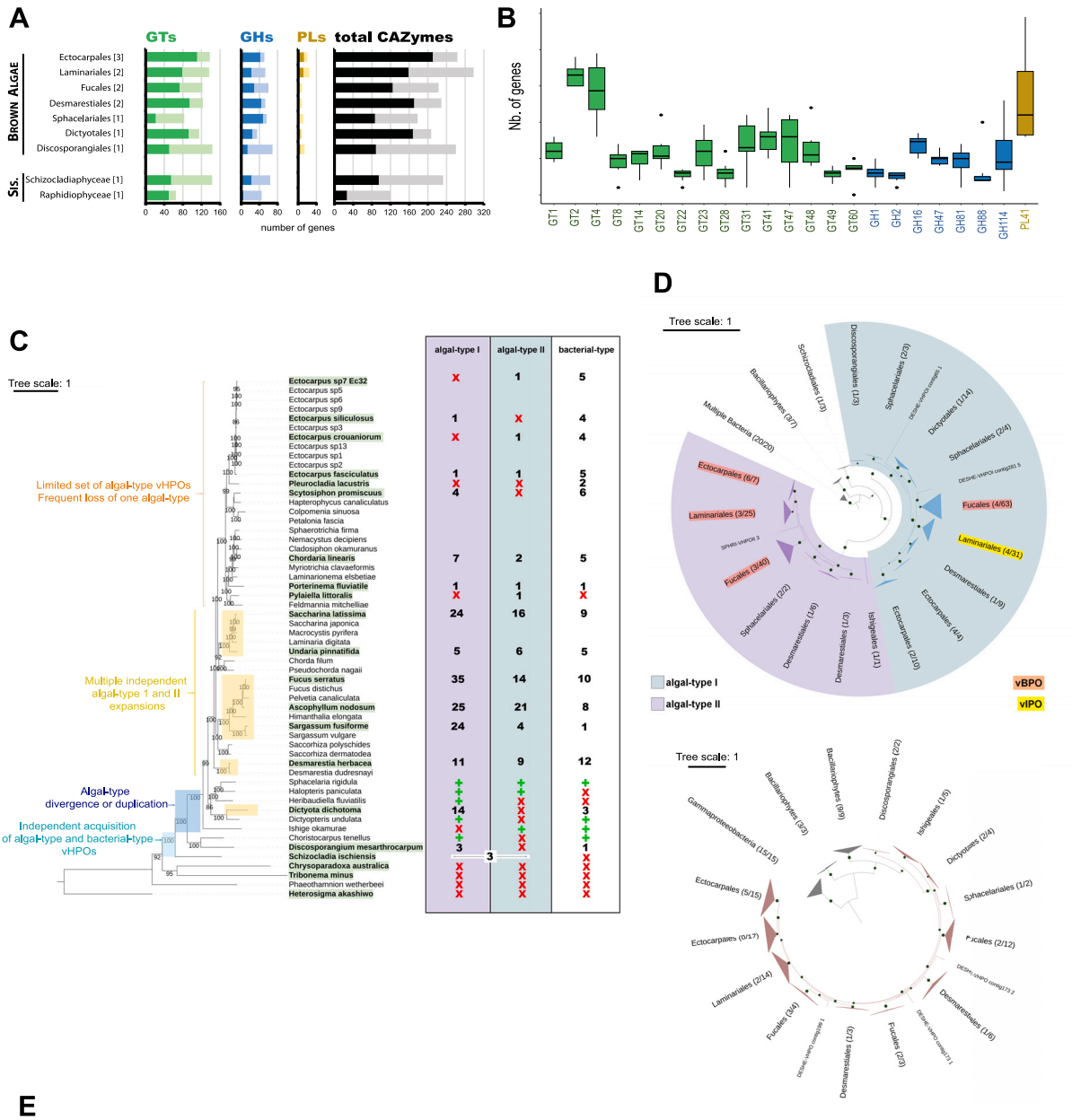


Figure S4. Structural features of the predicted genes in the Phaeoexplorer genomes, related to Figure 1

(A) Genome and gene statistics for the 21 reference genomes (Figure S1A; Table S1B). Violin plots display size distributions. Cross or triangle, mean; diamond or square, median. For intergenic regions, half violins on the left correspond to intergenic distances between adjacent genes on opposite strands, and half violins on the right to intergenic distances between adjacent genes on the same strand.

(B) Intron acquisition during the emergence of the brown algal lineage (left). The analysis examined 949 introns in 235 conserved orthologs. Colors correspond to the species distributions of introns. Hatched blue indicates introns that are shared with at least two outgroups (ancestral introns). The right panel shows an example of an intron conservation profile for the orthogroup OG0004854. Color code and species numbering as for the left. F, female; M, male.



(legend on next page)

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

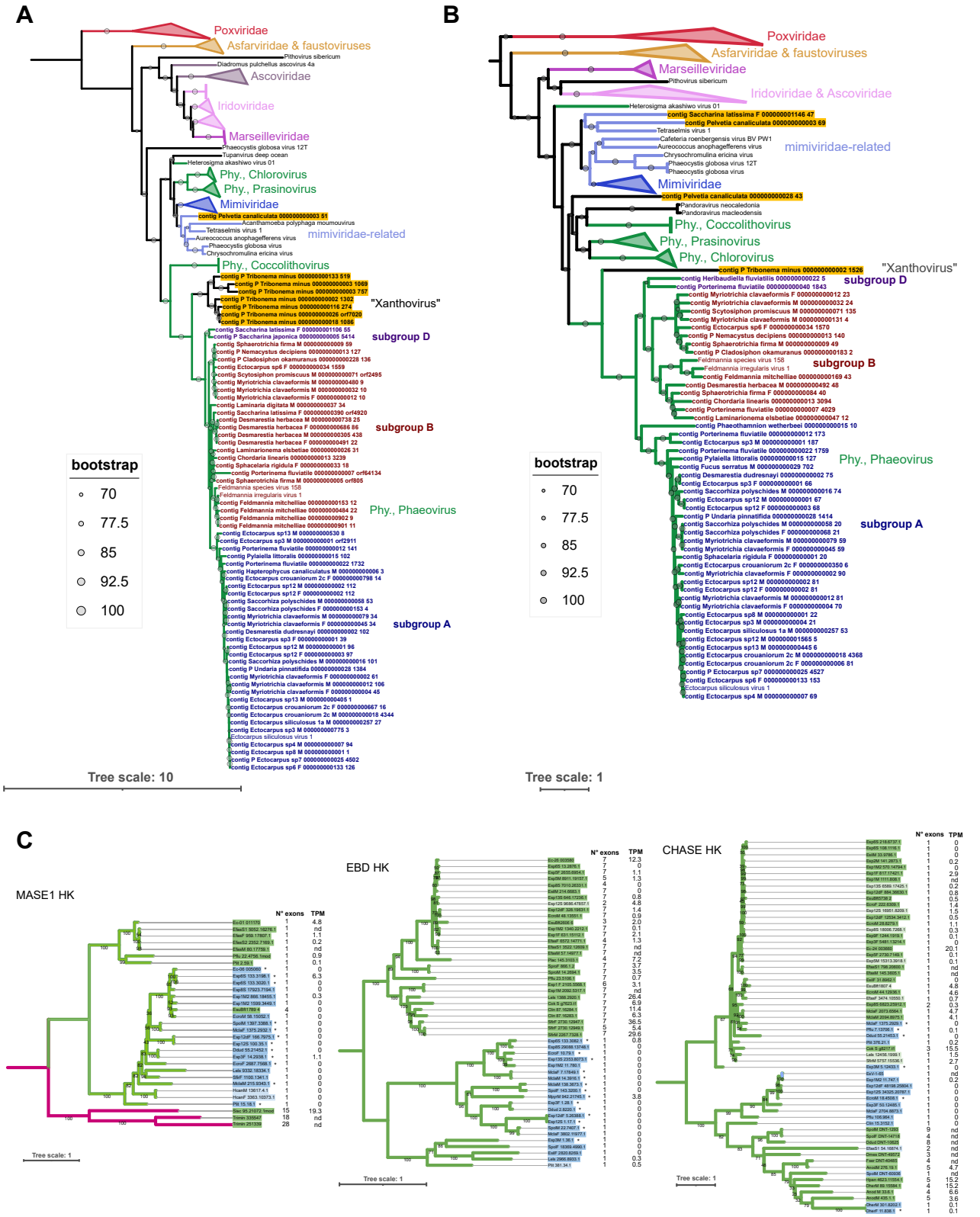
(A) Counts of numbers of genes predicted to encode glycosyltransferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs), and all CAZymes (GTs, GHs, PLs, carbohydrate esterases [CEs], auxiliary activities [AAs], carbohydrate-binding modules [CBMs]), showing numbers for both full-length proteins (dark colors) and fragments (light colors). The data are averaged by order with the number of species analyzed per order in brackets.

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

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

(D) Maximum-likelihood phylogenetic trees for 259 algal-type vHPOs (left) and for bacterial-type vHPOs (right). Algal-type I vHPOs are colored in blue and algal-type II vHPOs in violet. The clades that have structurally or biochemically characterized enzymes are highlighted in red for vBPOs and in yellow for vIPOs. Strongly supported representative branches have been collapsed. The clade names correspond either to taxa or to individual gene names. For the bacterial vHPOs, the FastTree reconstruction tool with 1,000 bootstraps was drawn as a circular representation taking the gammaproteobacterial group as the starting point to arbitrary root the tree. Brown algal branches are colored in brown. Green dots indicate bootstrap values of between 0.7 and 1.0 (1,000 replicates).

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



(legend on next page)

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

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

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

nature ecology & evolution

Article

https://doi.org/10.1038/s41559-025-02838-w

Origin and evolutionary trajectories of brown algal sex chromosomes

Received: 22 September 2024

Accepted: 22 July 2025

Published online: 25 August 2025

Check for updates

Josué Barrera-Redondo[1,9,11], Agnieszka P. Lipinska[1,2,11], Pengfei Liu[1], Erica Dinatale[1], Guillaume Cossard[1], Kenny Bogaert[1], Masakazu Hoshino[1,10], Rory J. Craig[1], Komlan Avia[3], Goncalo Leiria[1], Elena Avdievich[1], Daniel Liesner[1], Rémy Luthringer[1], Olivier Godfroy[2], Svenja Heesch[2], Zofia Nehr[2], Loraine Brillet-Guéguen[2,4], Akira F. Peters[5], Galice Hoarau[6], Gareth Pearson[7], Jean-Marc Aury[8], Patrick Wincker[8], France Denoeud[8,12], J. Mark Cock[2,12], Fabian B. Haas[1,12] & Susana M. Coelho[1,12]

Research on the biology and evolution of sex chromosomes has primarily focused on diploid XX/XY and ZW/ZZ systems. In contrast, the rise, evolution and demise of U/V systems has remained an enigma. Here we analyse genomes of nine brown algal species with different sexual systems to determine the history of their sex determination. U/V sex chromosomes emerged between 450 and 224 million years ago, when a region containing the pivotal male-determinant *MIN* ceased recombining. Seven ancestral genes within the sex-determining region show remarkable conservation over this vast evolutionary time, although nested inversions caused expansions of the sex locus, independently in each lineage. We evaluate whether these expansions are associated with increased morphological complexity and sexual differentiation, and show that taxonomically restricted genes evolve unexpectedly often in U and V chromosomes. We also investigate two situations in which U/V-linked regions have changed. First, we demonstrate that convergent evolution of two monoicous species occurred by ancestral males acquiring U-specific genes. Second, the *Fucus* dioecious system involves new sex-determining gene(s), acting upstream of formerly V-specific genes during development. Both situations have led to the demise of U and V chromosomes and erosion of their specific genomic characteristics.

The mechanisms controlling the development of male or female identities, or co-sexuality, when individuals express both sex functions, vary widely across different organisms[1,2]. In species with separate sexes, sex chromosomes may be present, carrying a sex-determining region (SDR)[3] that encodes factors directing sex identity and which often does not undergo recombination in the heterogametic sex (XY or ZW)[4] of diploid species (dioecious), or in the diploid stage of haploid-dominant (dioicous) species. Sex chromosomes have independently evolved from autosomes multiple times and may be subject to specific evolutionary forces, including differential selection between sexes, asymmetrical expression of deleterious mutations and hemizygosity, meiotic silencing and dosage compensation[3].

Research on the biology and evolution of sex chromosomes has primarily focused on diploid XX/XY and ZW/ZZ systems in mammals, birds, fish, *Drosophila* and diploid plants[4,5]. U/V haploid sex-determination systems, such as those of bryophytes and algae[6,7], have been less explored. In U/V systems, sex is not determined at fertilization but during meiosis, when haploid spores inherit either a

A full list of affiliations appears at the end of the paper. e-mail: susana.coelho@tuebingen.mpg.de

U chromosome, and will develop into a female gametophyte, or a V chromosome, controlling male gametophyte formation[8]. These fundamental inheritance differences between U/V and XX/XY or ZW/ZZ systems have broad evolutionary and genomic implications[9,10]. However, so far, only the U/V systems of the brown alga *Ectocarpus* and the U/Vs of four distantly related bryophyte taxa[11–14] have been fully sequenced and assembled into chromosomes. While these studies helped understand the genomic structure of bryophyte U- and V-linked regions, the species involved diverged ~500 Ma (million years ago) and do not share homologous U/V chromosomes[15]. As a result, we still lack a broad comparative view across multiple homologous U/V systems that would inform a reconstruction of their evolutionary history. Brown algae represent exceptional models for studying sex chromosome evolution because they display diverse reproductive systems, life cycles and sex chromosome systems in a single lineage[16]. Their ancestral state probably involved separate sexes[16], suggesting that their sex chromosomes could share a common origin. Here we study the origin, evolution and demise of U/V sex chromosomes in the brown algae.

Results

The origin of brown algal sex chromosomes

We focused on species covering the phylogenetic, morphological and reproductive diversity of the brown algal clade[17] and their closest extant outgroup, *Schizocladia ischiensis*[17,18]. We substantially improved the brown algal genome datasets available[18–20] to reach chromosome or near-chromosome-level genome assemblies (Extended Data Table 1 and Supplementary Table 1). This revealed that brown algae have 27–33 chromosomes and largely conserved macrosynteny (Fig. 1a).

We identified the female (U) and male (V) sex-determining regions (SDRs) in the dioicous species using a combination of bioinformatic and experimental approaches (see 'Discovery of the U/V sex determination regions' in Methods; Supplementary Figs. 1–5). All U/V species share the same, albeit highly rearranged, ancestral sex chromosome, showing remarkable stability despite the large evolutionary time (Fig. 1b,c and Extended Data Table 1). The recombination suppression event leading to the birth of U/V sex chromosomes occurred after the split of *S. ischiensis* and *Dictyota dichotoma*, ~450–224 Ma[21] (Fig. 1a). The male-determining gene *MIN*[22] is the only V-specific gene consistently present in all V-SDRs of the dioicous species. We note that one dioecious (*Fucus serratus*) and two monoicous (haploid, co-sexual *Chordaria linearis* and *Desmarestia dudresnayi*) species lack U/V sex chromosomes but still retain *MIN* on a chromosome homologous to the ancestral U/V ('U/V-homologue' hereafter). The outgroup *S. ischiensis* has low synteny with the brown algae and exhibits putative fusion-with-mixing events[23] (Fig. 1a and Extended Data Fig. 1).

We next examined the U/V-SDRs by comparing male and female genome assemblies (Methods). The SDRs contain a small number of genes overall (between 18 and 52), and compared with the pseudo-autosomal regions (PARs) (between 229 and 904), with considerable variation in gene content and size across species, the smallest being found in the Ectocarpales (*Ectocarpus* sp. 7, *Ectocarpus crouaniorum*, *Scytosiphon promiscuus*; Fig. 1b,c, Extended Data Fig. 2a and Extended Data Table 1). SDR size differences across species are strongly correlated with the number of genes (R^2 = 0.97; Extended Data Fig. 2b) and the repeat content (R^2 = 0.99; Extended Data Fig. 2c) inside these regions. Many genes located in the PARs of the Ectocarpales are within the V-SDRs of *Undaria pinnatifida*, *Desmarestia herbacea* and *D. dichotoma*, indicating that the SDR boundaries have changed across species. The boundary differences coincide with extensive structural rearrangements, particularly inversions, even among closely related taxa (Fig. 1b,c and Extended Data Fig. 3). Note that the centromere in the V chromosome of *Ectocarpus* is found within the SDR[19], so we cannot exclude that a centromere-related suppression of recombination may have preceded the inversion events found on the SDR[24].

Together, our results indicate that the brown algal U/V sex chromosomes evolved between 450–224 Ma, via suppressed recombination in a genomic region that contained *MIN* (henceforth male-determining locus). The presence of *MIN* in distantly related lineages could push the age of the U/V chromosomes further back in time, but more evidence would be required to establish that dioicy existed in these organisms.

The evolution of the SDRs involved boundary expansions and gene gains

The brown algal U- and V-SDRs carry homologous genes (gametologue pairs), indicating descent from a common ancestral region (Supplementary Table 2). *Ectocarpus* sp. 7 and *D. herbacea* show similar ratios of gametologues and U- or V-specific genes (16/14 and 11/7 gametologue/sex-specific genes in *Ectocarpus* and *D. herbacea*, respectively; Supplementary Table 2). Only ten genes share SDR orthologues between both species, while the rest were mostly acquired independently in the SDR of each species, with one gene that was retained as a gametologue pair in *Ectocarpus* sp. 7 but lost both copies in *D. herbacea* (Supplementary Table 2). Five gametologue pairs conserved both copies in the two species, while another three gametologue pairs lost either the male or the female copy in *D. herbacea* (Supplementary Table 2). In addition, *MIN* and a U-specific gene are also conserved between species (Supplementary Table 2). Although the total number of U/V-SDR genes differs between *Ectocarpus* sp. 7 (18 genes) and *D. herbacea* (30 genes), each species shows an equal number of gametologues and sex-specific genes in its U- and V-SDRs (Supplementary Table 2). This intraspecies symmetry supports the idea that the U and V chromosomes may have undergone parallel evolutionary changes within each lineage[10,25,26]. The V-SDR of *D. herbacea* contains 20 additional genes that belong to endogenous viral elements, which are common across brown algal genomes[18].

Diploid sex chromosome in animals and plants exhibit evolutionary strata representing different recombination suppression events over time. Strata are identified by analysing synonymous substitutions (K_s) between male/female gametologue pairs[27] whose locations in fully X or Z-linked regions are known. However, detecting evolutionary strata in U/V systems is difficult because neither of these fully sex-linked regions recombines and gene movements and chromosome rearrangements disrupt collinearity of both chromosomes between species[25,28,29]. Moreover, in the absence of a recombining outgroup (which does not exist in brown algae), the ancestral gene order cannot be reliably inferred. In both *Ectocarpus* sp. 7 and *D. herbacea*, the V- and U-SDR rearrangements differ by inversions (Fig. 2a,b and Extended Data Fig. 4), consistent with the idea that inversions may lead to suppressed recombination between sex chromosomes. An analysis of gametologue pair divergence revealed saturated levels of K_s values (Fig. 2b,c and Supplementary Table 3), further limiting the inference of evolutionary strata across brown algal SDRs. Nonetheless, the gametologue K_s values are broadly consistent between orthologues in *Ectocarpus* sp. *7* and *D. herbacea*, where shared SDR gametologues between species have higher K_s values and probably spent more evolutionary time diverging than the gametologues that are not shared between species (Supplementary Table 3). Furthermore, the location of gametologues with the lowest K_s values in the U-SDR of *D. herbacea*, relative to the PAR genes in *Ectocarpus* sp. 7, suggests that inversions involving the entire U-SDR and adjacent PAR segments probably contributed to the expansion of the U/V-SDR boundaries in *D. herbacea*, in a process we term 'engulfment' (Extended Data Fig. 4). The expansion of the SDR boundaries in *D. herbacea* led to the engulfment of a region containing four genes in the PAR1 of *Ectocarpus* sp. 7, and a second region with 13 genes located on the PAR2 (Fig. 2d, Supplementary Table 4 and Extended Data Fig. 4). Twelve of these engulfed genes into the SDR of *D. herbacea* were retained as gametologues. These observations support a scenario where expansions in the SDR boundaries of brown algae

https://doi.org/10.1038/s41559-025-02838-w

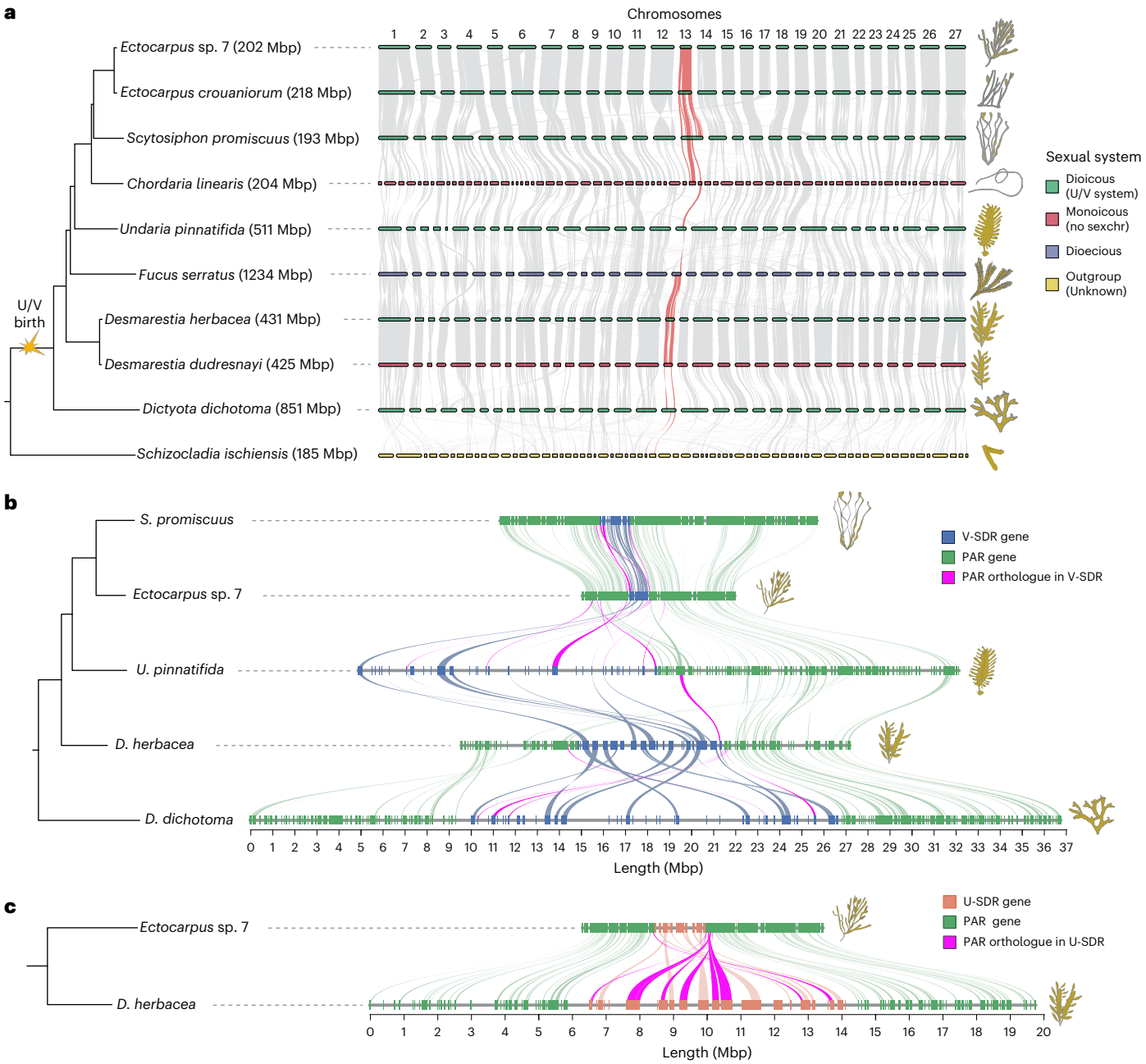


Fig. 1 | Origins of U/V sex chromosomes in brown algae. a, Macrosynteny plot comparing genomes of six dioicous (green), two monoicous (red), one dioecious (blue) and one outgroup species (yellow). The chromosomes were originally numbered by their physical size in the *Ectocarpus* v2 genome[141]. Note that the dioecious species *F. serratus* has a fully diploid life cycle (without gametophytes[7]). Syntenic blocks of the V sex chromosome are highlighted in red, with the emergence of U/V chromosomes shown in the phylogeny. Genome sizes are indicated in brackets. **b**, Microsynteny plot of V chromosomes in five dioicous species, highlighting the male sex-determining regions (blue) and the PARs (green). The PAR genes whose orthologues are found within the SDR of other species are highlighted in purple. **c**, Microsynteny plot of U chromosomes in two dioicous species, highlighting the female sex-determining regions (peach) and the PARs (green). The PAR genes whose orthologues are found within the SDR of other species are highlighted in purple. Note that the genome assemblies for *C. linearis* and *S. ischiensis* are not chromosome level, leading to a high number of contigs.

occur through nested inversions. Two chromatin-related transcription factors in the *Ectocarpus* PARs were independently incorporated into the SDRs of four other dioicous species (Supplementary Tables 4 and 5).

The observation of greater V-SDR gene content in early diverging lineages (such as *D. dichotoma*) than in the later-diverging Ectocarpales (Extended Data Table 1) could reflect either gene loss in the V-SDRs of Ectocarpales or independent gene gains in the V-SDRs of each lineage (as predicted in ref. 10), from an ancestral state with low V-SDR gene content that is retained in Ectocarpales. To distinguish between these possibilities, we reconstructed the ancestral SDR gene content (Supplementary Table 5), focusing on the V chromosome, as the genomic data are of better quality (Supplementary Table 1), and assuming parallel U/V-SDR evolution[10,25,26] as seen in *Ectocarpus* sp. 7 and *D. herbacea* (Fig. 2d and Supplementary Table 2). This analysis revealed that brown algal V-SDR evolution occurred via lineage-specific gene gains rather than gene loss in the Ectocarpales (Fig. 2e,f). Gene gains were caused by a combination of three processes: expansions of the SDR boundaries into the PARs, translocation of autosomal genes into the SDR and lineage-specific gene birth events within the SDR (Fig. 2e and Supplementary Table 5). Consistently, ancestral

V-SDR genes were associated with higher gametologue K_s values, while independently acquired gametologues in *D. herbacea* had lower K_s values (Supplementary Table 2 and Fig. 2c).

The seven genes in the ancestral V-SDR (Fig. 2f and Extended Data Fig. 2d) include the male-determinant *MIN*[22] and six V gametologues of genes that are also carried on the U chromosome (Fig. 2f and Supplementary Table 6). As predicted by early models of U/V-SDR evolution[10], all seven genes are probably related to sex determination processes. Gametologue pairs include putative transmembrane proteins that may play a role in gamete recognition[30], STE20 serine/threonine kinase gametologues probably involved in pheromone pathways[31], and a casein kinase, a MEMO-like domain protein and a GTPase-activating protein which may act in signal transduction (Supplementary Table 6). All of these genes are gametologue pairs in *Ectocarpus* sp. 7, but in *D. herbacea* the casein kinase was lost from the U-SDR and the putative transmembrane receptor was lost in both sexes. We noticed that these ancestral V-SDR genes remain in the U/V-homologue of the species that have lost their U/V system (Fig. 2f and Supplementary Table 6), emphasizing their importance for pathways in sex organ development even in the absence of sex chromosomes.

The V-SDR size appears to be associated with the level of sexual dimorphism (Fig. 2e and Extended Data Fig. 2), but the small sample size is insufficient for formal statistical analysis. Species with low sexual dimorphism (anisogamous) retained the ancestral V-SDR genes with very few gene gains, further suggesting that they may represent the V-SDR ancestral state. Although the number of SDR changes is small, oogamous species each independently gained diverse V-SDR genes, and one gene (ATP-dependent RNA helicase) was convergently acquired in all (OG0003211 in Extended Data Fig. 2d). All the detected autosomal translocations into the V-SDRs of *Ectocarpus* sp. 7 and *D. herbacea* (Fig. 2e) also involve sex-specific genes (Supplementary Tables 2 and 5), consistent with a model where sexual antagonism in autosomal loci may be solved by gaining sex linkage[32]. In contrast, we found no correlation between autosomal sex-biased gene (SBG) expression and sexual dimorphism level (false discovery rate (FDR)-corrected $P > 0.01$; Supplementary Table 7), supporting previous studies[33] (Extended Data Fig. 5a). However, we observed an enrichment of male-biased genes on the PARs in all species (chi-square test $P < 0.01$) except *D. dichotoma* (Extended Data Fig. 5b).

Most U/V-SDR genes were prominently expressed in fertile haploid gametophytes, consistent with gene preservation via haploid purifying selection (Supplementary Table 8). Gametologues had typically higher expression levels than sex-specific genes (present in only one of the SDRs) (Wilcoxon test, $P = 0.00075$ in *D. herbacea*; $P = 0.08843$ in *Ectocarpus* sp. 7) (Fig. 2d). A comparative analysis in fertile gametophytes between SDR genes and their autosomal counterparts in other species showed that newly acquired genes on the SDR had similar expression levels to their autosomal counterparts (Extended Data Fig. 6), suggesting either a co-option of autosomal biological activity into male-specific functions in the V-SDR or the general importance of these genes for gametophyte development. Examining expression levels across multiple tissues in *Ectocarpus* sp. 7 revealed that activity of U/V-SDR genes is not confined to fertile gametophytes (Fig. 2d). Therefore, the SDRs contain not only genes involved in sex determination and gametophyte fertility but also genes playing a broader role in development.

Altogether, our analyses illustrate how brown algal U/V-SDRs undergo structural changes, evolving mainly by lineage-specific gene gains associated with increasing levels of sexual dimorphism. We identified a set of conservatively sex-linked genes in dioicous brown algae, suggesting their role in sex determination and/or differentiation, along with genes potentially involved in other developmental pathways.

Structural features and evolutionary dynamics of brown algal U/V sex chromosomes

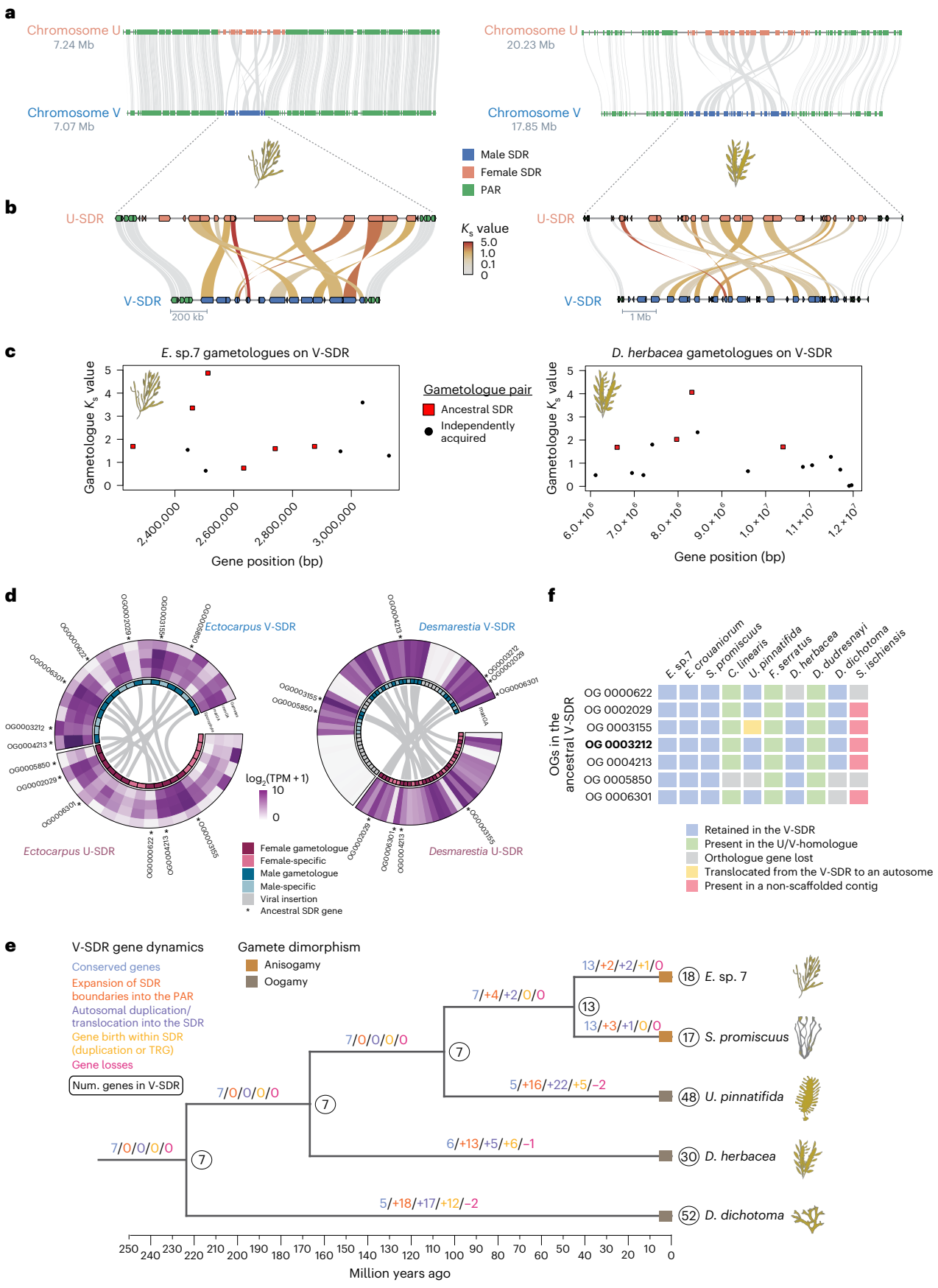
We next examined the structural features that differentiate the entire U/V sex chromosomes (V-SDR and PARs) from the rest of the genome (Fig. 3a and Supplementary Fig. 12a). As expected for non-recombining regions[25,34], all V sex chromosomes are repeat rich and gene poor (Wilcoxon rank-sum test, FDR-corrected $P < 0.01$; Fig. 3a, Extended Data Fig. 7 and Supplementary Tables 9–11). V-SDRs have significantly higher repeat density than the PARs or the autosomes (permutation test, FDR-corrected $P < 0.001$; Extended Data Fig. 8). This low gene density is not influenced by the presence of centromeres within the SDRs, as the coding density in the *Ectocarpus* sp. 7 V centromere (3.51%) is slightly higher than in the rest of the V-SDR (2.85%), presumably due to the small size of the centromere (153 kbp)[19]. The PARs were also significantly enriched in repeats when compared with the autosomes, although less so than the V-SDRs (permutation test, FDR-corrected $P < 0.001$; Extended Data Fig. 8). Among repetitive elements, 'unclassified' transposable elements (TEs) were enriched in the PARs and SDRs of the Ectocarpales (permutation test, FDR-corrected $P < 0.01$), while the V-SDRs of species that underwent genome expansion (for example, *U. pinnatifida*, *D. herbacea*, *D. dichotoma*) predominantly accumulated long terminal repeat (LTR) elements (Supplementary Fig. 6 and Supplementary Table 9).

Moreover, sex chromosomes had fewer orthologues conserved between species compared with the autosomes (chi-square test, $P < 10^{-4}$, Supplementary Table 12), possibly reflecting increased numbers of taxonomically restricted genes (TRGs; that is, genes that are not detectable outside of a defined taxonomic group). Phylostratigraphy analyses[35,36] confirmed an enrichment of TRGs in the sex chromosomes of all dioicous species (Wilcoxon rank-sum test, FDR-corrected $P < 0.01$; Fig. 3b and Supplementary Tables 13 and 14). TRG enrichment was localized in the PARs of the Ectocarpales, but this pattern extended to the entire sex chromosome, including the SDRs, in species with larger V-SDRs (permutation test, FDR-corrected $P < 0.001$; Fig. 3b). Importantly, sex chromosomes have statistically younger TRGs than the last common ancestor of the five dioicous species (same Order or broader taxonomic groups), indicating that TRG enrichment arose independently in each species (Pearson standardized residuals >2.4; Supplementary Figs. 7–11).

We previously proposed a theoretical model where generation-antagonistic selection may favour the retention of young sporophyte-beneficial loci in the PARs of *Ectocarpus* sp. 7 U/Vs[37]. Consistent with

Fig. 2 | Lineage-specific U/V-SDR expansion from an ancestral SDR and its association with sexual dimorphism. a, Microsynteny plot between the U and V chromosomes of *Ectocarpus* sp. 7 and *D. herbacea*. **b**, Synteny between the U and V gametologues within the V-SDRs of both species, coloured by synonymous substitutions per site (K_s). **c**, Identification of ancestral SDR gametologues (red squares) and independently acquired gametologues (black dots) with respect to their gametologue K_s values and their position in the V-SDR. **d**, Circos plot linking gametologue pairs in each species with expression levels of all SDR genes ($\log_2$(TPM + 1)) across different life stages in *Ectocarpus* sp.7 and mature gametophytes (matGA) in *D. herbacea*. Gametologues are highlighted in dark colours, sex-specific genes are highlighted in light colours, viral insertions are marked in grey and asterisks denote conserved SDR genes (also in **f**). **e**, The ancestral state reconstruction of V-SDR gene content across brown algae, showing the expected number of genes in the SDR (white circles), gene retention (blue numbers), gene gain through expansion of the SDR boundaries (orange), gene gain through autosomal translocation (purple), gene birth event inside the SDR (yellow) and gene loss (red) along with changes in gamete dimorphism[16]. **f**, Schematic of the seven ancestral V-SDR gene orthogroups (OGs), with genomic locations marked: retained in the V-SDR (blue), found in the U/V-homologue of non-dioicous species (green), translocated from the V-SDR to an autosome (yellow), present in a non-scaffolded contig (red), and lost (grey). Bold: *MIN*. See Supplementary Tables 5 and 6.

https://doi.org/10.1038/s41559-025-02838-w



https://doi.org/10.1038/s41559-025-02838-w

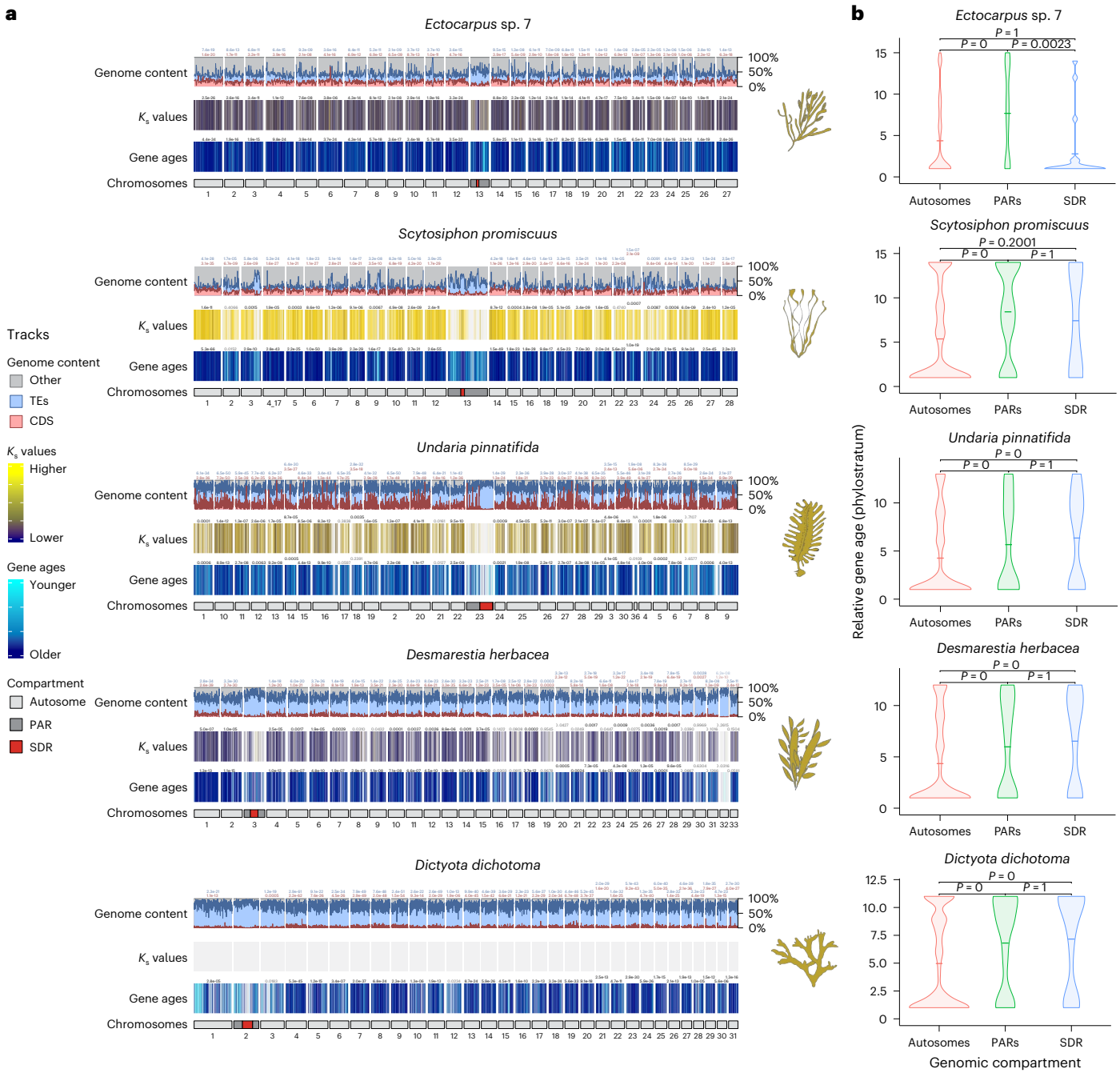


Fig. 3 | The U/V sex chromosomes are enriched in taxonomically restricted genes. a, Karyoplots for five dioicous species showing the following features from bottom to top: chromosome compartments (autosomes, PARs and SDR), relative gene ages, interspecies K_s values, and proportion of coding (CDS, red) and repeat (TEs, blue) density. Statistically significant differences for each feature between each autosome and the V chromosome are depicted on top of the track for that autosome (FDR-corrected two-sided Wilcoxon rank-sum test; values indicated with solid colours when $P < 0.01$ for the tested hypothesis). The precise range of gene age categories and interspecies K_s values for each species can be found in Supplementary Figs. 7–11. **b**, Violin plots for five dioicous species showing the relative gene age ranks (higher ranks equate to younger ages) of the TRGs across chromosome compartments (autosomes, PARs and SDR). Statistically significant differences in mean values of gene ages (centre line) were assessed using FDR-corrected two-sided permutation tests.

this model, sporophyte-biased genes are indeed enriched in the sex chromosomes of *Ectocarpus* sp. 7 and *U. pinnatifida* (fold change >2, adjusted $P < 0.05$), but less so in *D. dichotoma* and *S. promiscuus* (Supplementary Table 15). Moreover, we explored additional mechanisms underlying TRG emergence by estimating interspecies K_s values between orthologues in closely related species (Supplementary Table 16) and comparing these values between chromosomes and genomic compartments (V-SDR, PAR, autosomes). If synonymous mutations behave neutrally[38,39], then interspecies K_s can be used as a proxy for mutation rates[40,41]. Consistently, we found higher interspecies K_s values in the V sex chromosomes compared with autosomes across all dioicous species (Wilcoxon rank-sum test, FDR-corrected $P < 0.01$; Fig. 3a and Supplementary Tables 16 and 17), suggesting higher mutation rates relative to autosomes. Higher interspecies K_s values are also localized in the PARs, mirroring the pattern observed with the TRGs (Supplementary Figs. 7–10). Therefore, the enrichment of TRGs in the U and V is associated with both enrichment of sporophyte-biased genes and higher synonymous substitution rates.

To test the generality of the pattern of TRG enrichment on U and V chromosomes, we applied the same approach in other organisms with haploid sex determination, the plants *Ceratodon purpureum*, *Sphagnum angustifolium*, *Marchantia polymorpha*[11,12,42] and the fungus *Cryptococcus neoformans*[43]. We observed a clear enrichment of TRGs in the V chromosomes of *C. purpureum* and *S. angustifolium* (Wilcoxon rank-sum test, FDR-corrected $P < 0.01$; Supplementary Figs. 12 and 13 and Supplementary Tables 13 and 14), but not in the U/V chromosomes of *M. polymorpha* or the mating-type chromosome of *C. neoformans* (Supplementary Figs. 14 and 15 and Supplementary Tables 13 and 14).

Fate of U/V sex chromosomes following loss of dioicy

We studied the evolutionary trajectory of brown algal genomes after the loss of the U/V system, by exploring two independent transitions to monoicy in *C. linearis* and *D. dudresnayi* that undergo sexual reproduction and develop male and female gametangia[33]. Most genes in the 'ex'-sex chromosomes (U/V-homologues) of both monoicous species are male derived, indicating that monoicy emerged from a male background (Fig. 4a,b). The U/V-homologue of *C. linearis* contains several rearrangements spanning the regions that are homologous to the PAR and SDR (SDR-homologue), with 11 V-SDR-derived and 2 U-SDR-derived orthologues located within the SDR-homologue, with an additional V-SDR-derived gene that was translocated elsewhere in the U/V-homologue (Fig. 4a and Supplementary Table 18). Likewise, *D. dudresnayi* underwent at least two inversion events within the SDR-homologue after splitting from *D. herbacea* (Fig. 4b), containing 20 V-SDR-derived genes and 4 U-SDR-derived genes (Supplementary Table 19).

Both monoicous species retained mostly male and a few female copies for most of the U/V-SDR-derived gametologues (91% in *C. linearis* and 100% in *D. dudresnayi*), whereas several U- and V-specific orthologues were lost in these species (10 and 17 sex-specific genes in *C. linearis* and *D. dudresnayi*, respectively). Of these lost orthologues, 60% and 43% present closely related autosomal paralogues in *C. linearis* and *D. dudresnayi*, respectively (Supplementary Tables 18 and 19), although it is unclear whether the expression of these autosomal paralogues is compensating the activity of the lost genes. The only three U-SDR-derived orthologues in *C. linearis* are flanked by PAR orthologues translocated at the end of the V-SDR-derived region, suggesting that the U/V-homologue (contig 12) of *C. linearis* acquired its U-SDR-derived genes through two translocations (Extended Data Fig. 9). Three U-SDR-derived genes in *D. dudresnayi* are dispersed across the V-SDR-derived region of the U/V-homologue, suggesting independent U-SDR translocations into the V-SDR, while the fourth U-SDR-derived gene was translocated to an autosome (Extended Data Fig. 9).

The seven ancestral V-SDR genes are transcriptionally active ($\log_2(\text{TPM} + 1) > 2$) during reproductive stages of both monoicous species (Supplementary Table 20), emphasizing their role in reproduction despite the absence of a U/V system, particularly *MIN*[22] which is retained in both species. While most U-SDR-derived genes are absent in monoicous species, a single intracellular cholesterol transporter gene was convergently preserved in both monoicous genomes (Supplementary Table 20) and actively expressed during fertility in both *Ectocarpus* sp. 7 and *D. herbacea* (Supplementary Table 8).

The U/V-homologue of *D. dudresnayi* retains some vestiges of its past as a U/V chromosome, such as low coding density, high repeat density and enrichment of TRGs (Wilcoxon rank-sum test, FDR-corrected $P < 0.01$), although we found non-significant differences in interspecies K_s values across the genome (Fig. 4c, Extended Data Fig. 7, Supplementary Fig. 16 and Supplementary Tables 9–14, 16 and 17).

Finally, we examined the transition from haploid to diploid sex determination, which has remained unstudied in eukaryotes. Although the ancestral state for the brown algae is a U/V sexual system, the Fucales recently transitioned to a diploid life cycle[44], with many species, such as *F. serratus*[45], exhibiting diploid separate sexes (dioecy)[46]. Dioecy probably evolved from monoecy (both sexes in the same diploid individual) in the last common ancestor of the *Fucus* genus[47] (25–5 Ma[18]), consistent with a young sex chromosome in *F. serratus*. Our extensive bioinformatic analysis and PCR sex-linkage testing for candidate genes such as *MIN* (Methods) failed to identify sex-linked sequences in *F. serratus* (Supplementary Fig. 17), suggesting that the SDR is small and undifferentiated. However, male *F. serratus* conserves the *MIN* gene and all the ancestral V-SDR genes in its U/V-homologue (Figs. 2f and 5a). Importantly, although none of the U/V-SDR-derived genes are sex-linked in *F. serratus*, *MIN* and four other ancestral V-SDR genes are exclusively expressed in males (fully silenced in females) (fold change (FC) > 2, $p_{adj} < 0.05$; Fig. 5a and Supplementary Table 21). This pattern is consistent in three other Fucales species (Supplementary Table 22). Therefore, the ancestral V-SDR genes probably still play roles in male sex determination or differentiation pathways.

Contrary to the observations in *D. dudresnayi*, the U/V-homologue of *F. serratus* lacks the TRG enrichment pattern and all the other distinctive features of the U/V chromosomes (Wilcoxon rank-sum test, FDR-corrected $P > 0.01$). Thus, this 'ex'-sex chromosome has lost all the evolutionary vestiges of its past as a U/V chromosome (Fig. 5b, Extended Data Fig. 7, Supplementary Fig. 18 and Supplementary Tables 9–14, 16 and 17).

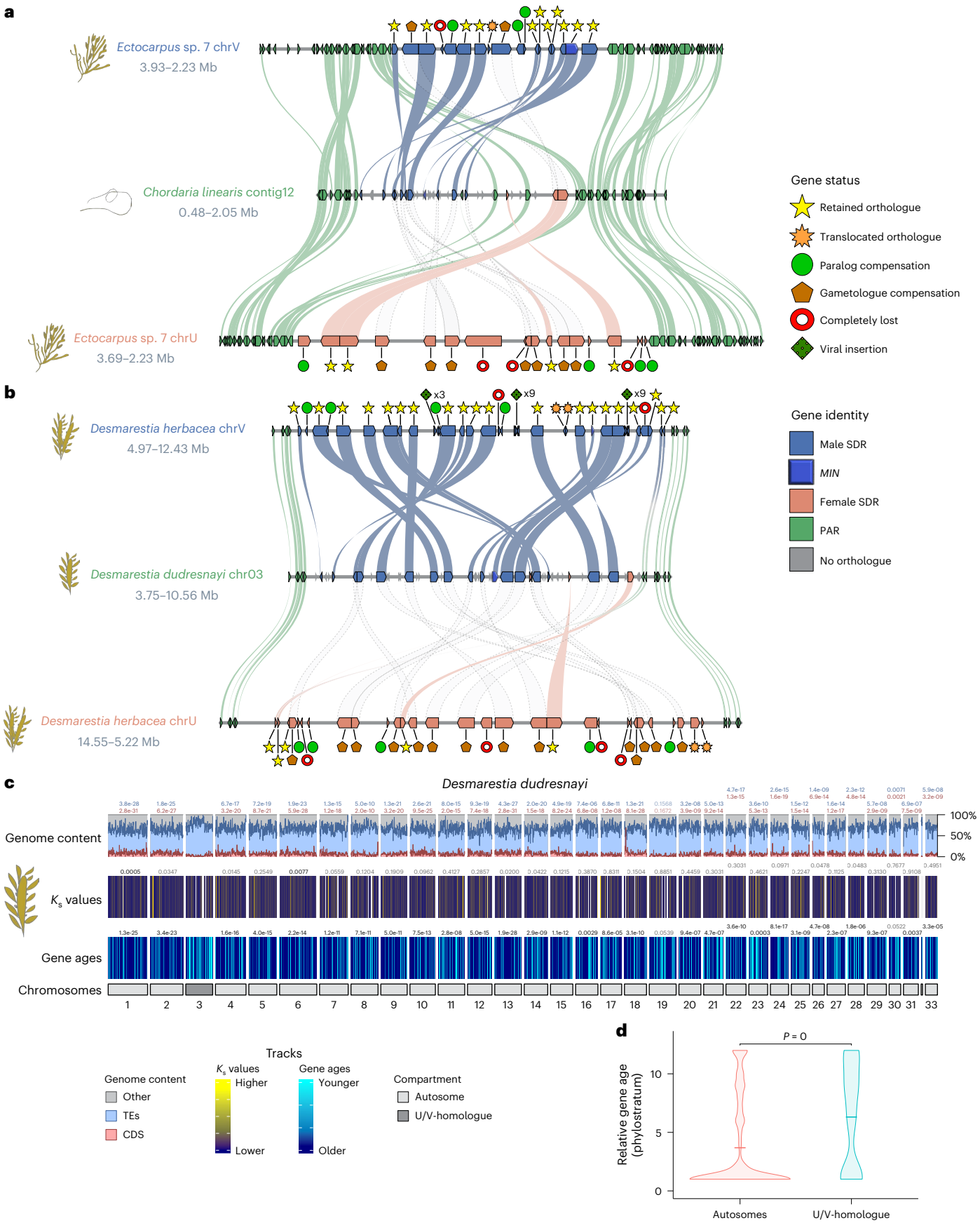
Discussion

Here we characterized the evolutionary trajectory of brown algal sex chromosomes (Fig. 6). Brown algal sex chromosomes date back 450–244 Ma[21], at the origin of brown algae. We propose that the male-determining gene *MIN* underlies the birth of this U/V system[22]. The ancestral cassette with seven V-SDR genes suggests a very early evolution of these genes into a non-recombining locus during the evolution of the U/V-SDRs. The ancestral V-SDR genes probably contribute to reproduction, but may also be involved in broader developmental functions. Despite their old age, brown algal U/V chromosomes retain large PARs bordering the SDR, unlike haploid systems in non-vascular plants that mostly lack detectable PARs[11–14].

Brown algal genomes have a high degree of synteny conservation. The U/V-SDR is however prone to accumulate structural rearrangements, including inversions that may have caused the initial recombination suppression event in proto-sex chromosomes and the later expansion of the U/V-SDRs into the PARs. Similar to other haploid systems[48,49], TEs conspicuously accumulated in the SDRs following recombination suppression[50], possibly causing further rearrangements through TE-mediated inversions[51].

Fig. 4 | Fate of sex chromosomes during transitions from dioicy to co-sexuality (monoicy). a, Comparison of the U/V-homologue in *C. linearis* against the U and V chromosomes of *Ectocarpus* sp. 7. **b**, Comparison of the U/V-homologue in *D. dudresnayi* against the U and V chromosomes of *D. herbacea*. The colour code represents the identity of the genes alongside the chromosomes, while the shapes represent the evolutionary fate of each SDR gene in the monoicous genome. The matching shades between the SDRs and the U/V-homologue are either colour coded by their ancestral background or they appear as transparent dotted shades if the gametologue of the other sex was retained. **c**, Karyoplot of *D. dudresnayi* showing the following features from bottom to top: chromosome compartment (autosomes and U/V-homologue), relative gene ages, interspecies K_s values, proportion of coding (CDS, red) and repeat (TEs, blue) density. Statistically significant differences for each feature between each autosome and the U/V-homologue are depicted on top of the track for that autosome (FDR-corrected two-sided Wilcoxon rank-sum test; values indicated with solid colours when $P < 0.01$ for the tested hypothesis). **d**, Violin plot showing the relative gene age ranks (higher ranks equate to younger ages) of the TRGs between the autosomes and the U/V-homologue of *D. dudresnayi*. Statistically significant difference in mean values of gene ages (centre line) was assessed using an FDR-corrected two-sided permutation test.

https://doi.org/10.1038/s41559-025-02838-w



https://doi.org/10.1038/s41559-025-02838-w

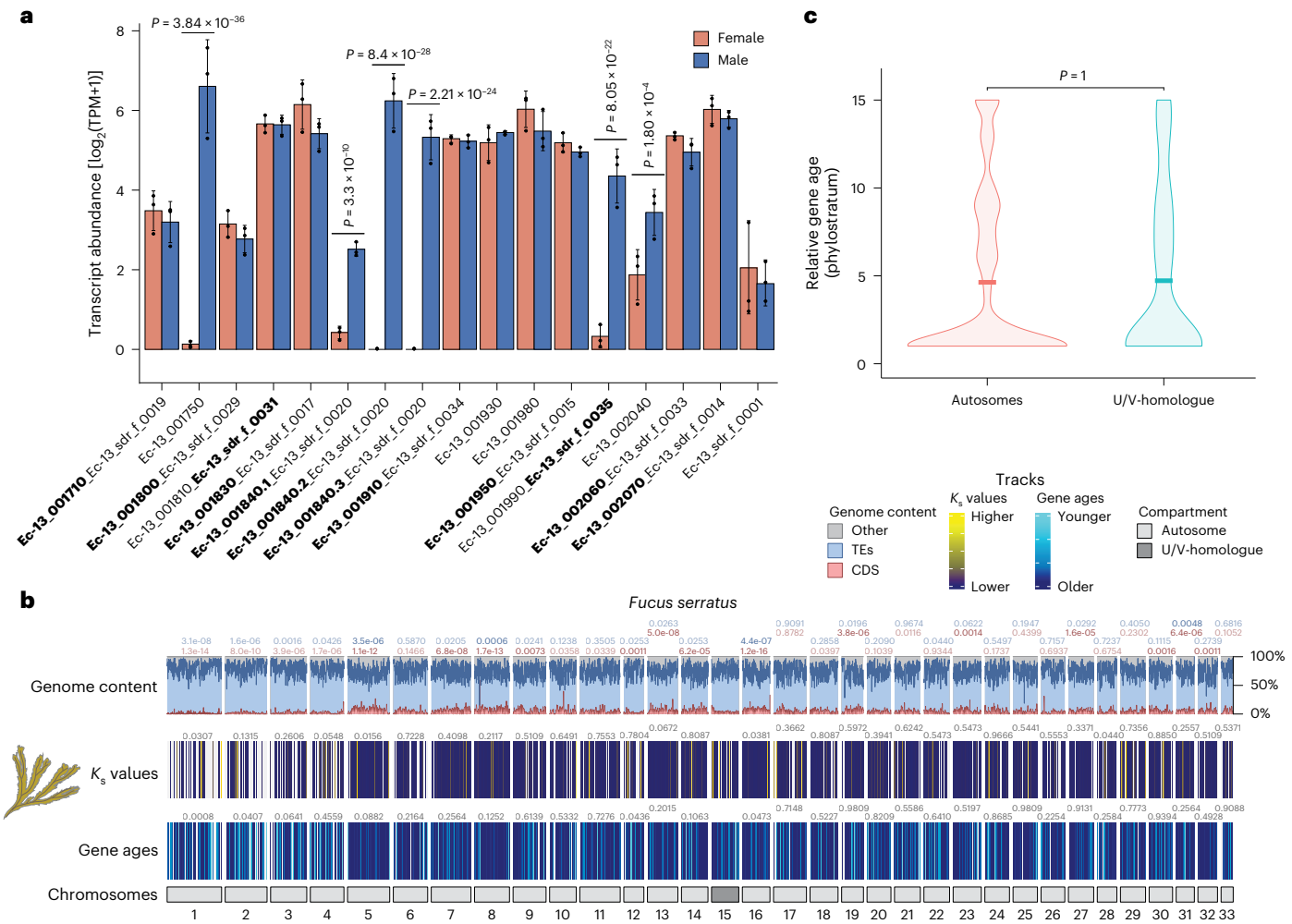


Fig. 5 | Transition from haploid to diploid sex determination. a, Expression of ancestral U/V-SDR genes in the diplontic species *F. serratus*. Gene expression of mature algae (using 3 males and 3 females, see Methods) is given as $\log_2(\text{TPM} + 1)$ and bars represent standard deviation of the mean. *P* values shown on the plot are derived from differential expression analysis performed using DESeq2, which applies a two-sided Wald test with Benjamini–Hochberg correction for multiple testing. Bold text represents whether the gene in *F. serratus* corresponds to an ancestral male or the female gametologue. **b**, Karyoplot of *F. serratus* showing the following features from bottom to top: chromosome compartment (autosomes and U/V-homologue), relative gene ages, interspecies K_s values (between 0.00079 and 6.838, with an average value of 0.148), proportion of coding (CDS, red) and repeat (TEs, blue) density. Statistically significant differences for each feature between each autosome and the U/V-homologue are depicted on top of the track for that autosome (FDR-corrected two-sided Wilcoxon rank-sum test; values indicated with solid colours when $P < 0.01$ for the tested hypothesis). **c**, Violin plot showing the relative gene age ranks (higher ranks equate to younger ages) of the TRGs between the autosomes and the U/V-homologue of *F. serratus*. The mean values of gene ages (centre line) are not significantly different (FDR-corrected two-sided permutation test).

Models of XX/XY and ZW/ZZ sex chromosome evolution suggest that sexually antagonistic selection may lead to SDR expansions, making former PAR genes fully sex linked and gaining new genes[52,53]. We show that the three brown algal species with greater sexual dimorphism in gametes (oogamy) indeed have enlarged U/V-SDR gene contents, although statistical testing is not possible given the small sample size. Oogamy is proposed to be ancestral in the brown algae[16,21], but this trait seems to be highly labile and our analyses suggest multiple independent transitions to oogamy from a less dimorphic ancestor with a small ancestral SDR, accompanied by SDR gene gain. Anisogamous species only experienced few gene gains in their SDRs, as predicted in ref. 10. Similar to *M. polymorpha*, the relatively simple gene content of the U/V-SDRs in the brown algae could regulate an autosomal effector gene network controlling sexual dimorphism, including differences in somatic development between male and female gametophytes[54]. Accordingly, we observed substantial sex-biased expression of autosomal genes between mature male and female gametophytes carrying sex organs. This contrasts with *C. purpureus*, where sex chromosomes carry thousands of genes and few sex-biased autosomal genes[11]. Consistent with other studies[33,55], we found no correlation between levels of sex-biased expression and sexual dimorphism in brown algae. Sexual dimorphism in this lineage may be controlled by a relatively small subset of genes, and most sex-biased genes may influence gametophyte physiology or vegetative development.

In diploid sexual systems, recombination suppression between sex chromosomes explains Y and W degeneration[56] (but see refs. 57–59). Differences in the evolution of U/V systems are predicted, since sex determination occurs in the haploid stage, in which deleterious mutations are more efficiently removed by selection than in diploid systems[29,60] and some of the predicted differences have been documented in bryophytes[29]. In brown algae, as in diploid systems, the SDRs lost recombination, leading to TE accumulation and consequent reduction in gene density. However, unlike the degeneration documented in some XX/XY and ZW/ZZ systems, our ancestral state reconstruction revealed more gene gains than losses or gene movements out of the V-SDR during evolution of brown algal lineages.

https://doi.org/10.1038/s41559-025-02838-w

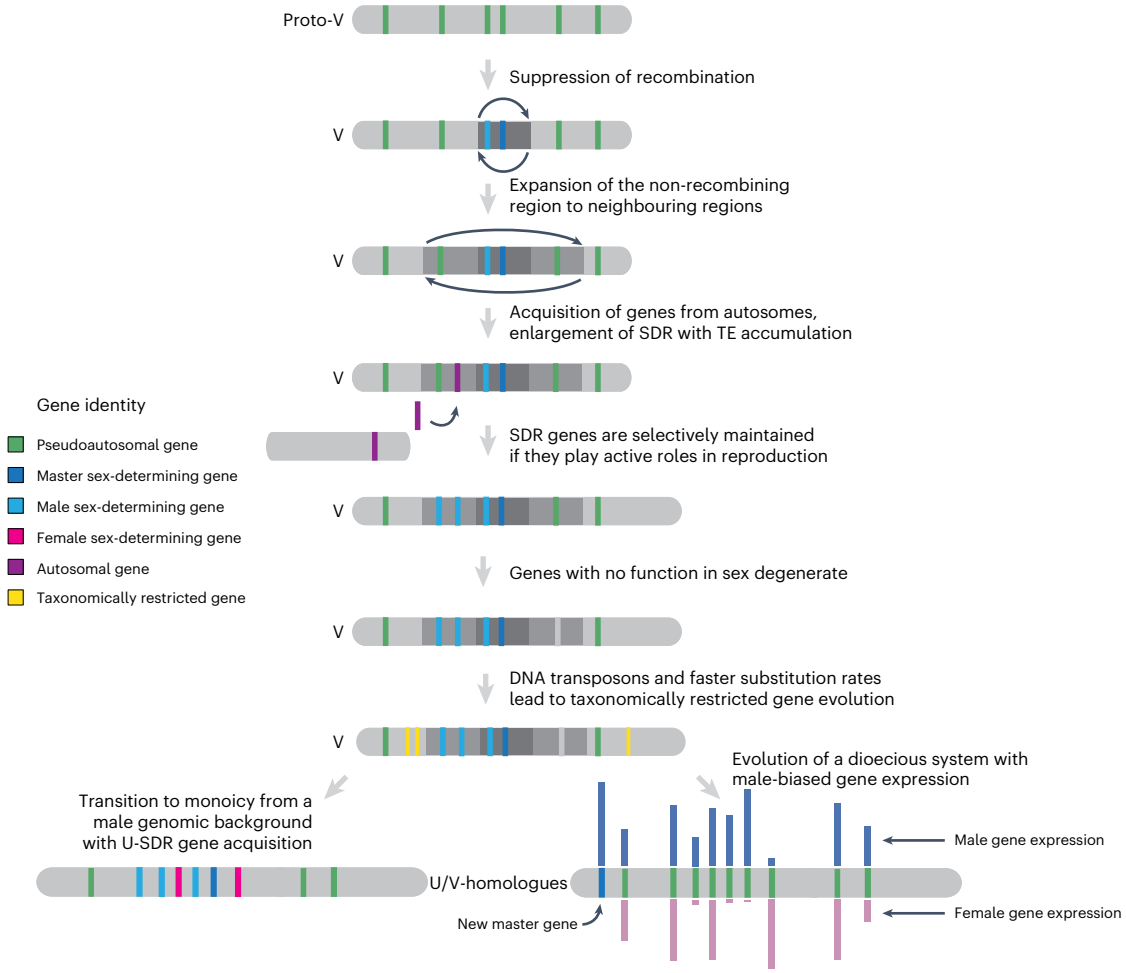


Fig. 6 | Hypothetical model for U/V sex chromosome evolution. U/V sex chromosomes arose from an ancestral autosome, via suppression of recombination that probably occurred via an inversion. The SDR boundaries expanded into neighbouring PAR via inversions, but also by recruitment of genes from autosomes; expansion occurred in a lineage-specific fashion, concomitant with increased sexual dimorphism of the different species. SDR genes are maintained within the SDR if they have roles in sex, whereas genes with no role in sex are lost. Faster substitution rates, probably a consequence of the heterochromatic context of the sex chromosome, may promote the rise of taxonomically restricted genes, which are selectively maintained on the sex chromosome if they have advantages to the sporophyte generation. In species that switch to a diploid life cycle, the U/V system disappears, but the genes that are in the V-specific region retain roles in sex, although they are no longer masters. Transition from U/V separate sexes to co-sexuality (monoicy) occurred when a male haploid individual acquired female-specific genes via translocations. During the demise of the U/V sex chromosomes, their structural and evolutionary footprints slowly disappear.

Unclassified repeats accumulate in the Ectocarpales sex chromosomes, including the PARs, but not in species with larger genomes, where LTR retroelements become dominant. DNA transposons are over-represented among unclassified repeats[61], and they often insert near the progenitor locus in a process called local hopping[62]. We propose that the U/V-SDR may thus act as a source of DNA transposons that hop to the PARs, thus increasing their repeat density, whereas increased colonization of LTR elements obscures this pattern in larger genomes.

Brown algal U/V chromosomes display an excess of TRGs. What mechanisms underlie this pattern? The SDR and the PARs of U/V chromosomes are enriched in heterochromatin[63], involved in repressing TEs[64], and heterochromatic regions tend to have higher mutation rates due to reduced access of the DNA repair machinery during replication[65]. Accordingly, we consistently observe higher interspecies K_s values in U/Vs, particularly in the PARs, since these regions recombine between sexes and thus have a higher rate of neutral fixation than the SDR, which experiences the Hill–Robertson effect that reduces fixation probabilities of neutral mutations due to its linkage with other sites under selection[66]. We speculate that this feature could facilitate the evolution of TRGs. Alternatively, the high density of DNA transposons within the U/V could also promote the co-option of their regulatory motifs and enable de novo transcript birth, as seen in *Drosophila*[67]. Note that the pattern could be reinforced through generation-antagonistic selection[37], but DNA transposons and higher mutation rates may be sufficient to initiate this pattern in species lacking sporophyte-biased gene expression. Importantly, the TRG enrichment is unique to U/V systems and gradually disappears when these systems are lost. This pattern extends beyond brown algae to other eukaryotes with U/V systems, such as *C. purpureus*[11,68] and *S. angustifolium*[12]. However, we could not detect enrichment of TRGs on the small U/V sex chromosomes of *M. polymorpha*[42,69], or on the mating-type chromosome of *C. neoformans*[43]. Mosses such as *C. purpureus* and *S. angustifolium* display relatively more complex sporophyte body plans than liverworts such as *M. polymorpha*[70], which could underlie stronger generation-antagonistic selection[37]. Unlike brown algae, most of the sex chromosomes in bryophytes are sex linked, with minimal space for PARs[29]. In this context, the sex chromosome of *C. purpureus* expanded its SDR very recently in evolutionary time through fusions of autosomes with earlier established U and V chromosomes[71], retaining its evolutionary footprints as PARs,

while the sex-linked region in *M. polymorpha* is much older[69], which could also limit the formation of TRGs that are predominantly observed in the PARs of the brown algae. The pattern thus appears to be specific to U/V systems where chromosomal degeneration is mild and linked to haploid–diploid life cycles where the sporophyte stage is sufficiently complex, highlighting a key role for generation selection[37]. Therefore, our study hints at a unique interplay between complex life cycles, heterochromatic landscape, DNA transposons and higher mutation rates that may lead to TRG enrichment in U/V chromosomes, and this process is pervasive across distant, independently evolved eukaryotic kingdoms.

Monoicous brown algae have transcriptomic profiles resembling ancestral females[33]. However, our results show that monoicy arose at least twice from a male ancestor that acquired female genes. The male pathway requires *MIN*[22,25,72], which could have facilitated the evolution of monoicy from males, as also seen in the green lineage[73,74]. We note the presence of U-SDR-derived gene(s) in all monoicous species, particularly a cholesterol transporter gene that is found in all U-SDRs, suggesting that the U-SDR contains femaleness-promoting factor(s), consistent with reports in kelps[75]. We thus propose that monoicy evolved via translocation events adding essential femaleness-promoting genes to a male genetic background. Since the resulting monoicous individuals were capable of producing both male and female reproductive structures, individuals with U chromosomes were no longer essential for sexual reproduction, ultimately leading to the loss of the U-SDR in monoicous species. Although the combination of key female and male genes is essential for this evolutionary transition, the retention of a sex chromosome is not. For example, in *Volvox africanus*, monoicy required the retention of female SDR-like regions, while most male SDR genes were lost except for a multicopy array of the male-determining gene *MID*[76].

The evolution of a dioecious system in *F. serratus* is associated with an irreversible transition to diploidy in Fucales[16,47]. The U/V to dioecy transition has remained elusive[2,47], but our data in brown algae imply that it involved an intermediate monoicous stage, supporting previous predictions from ancestral state reconstruction analyses[16]. A small, undifferentiated Y-specific region consistent with a young XY system may explain why the sex chromosome in *F. serratus* was undetectable. Nonetheless, all ancestral V-SDR genes are found in the U/V-homologue of *F. serratus*, several showing a male-biased gene expression across Fucales species, particularly *MIN*[22]. Our findings imply that *MIN* and possibly other ancestral V-SDR genes are still involved in male differentiation, but shifted downwards in the sex determination cascade. These results thus support and extend the 'bottom-up' hypothesis of sex determination, where downstream components of sex differentiation are conserved across taxa, and new master sex regulators can replace older ones[77].

Methods

Biological material

Scytosiphon promiscuus, Dictyota dichotoma, *Undaria pinnatifida* and *Desmarestia dudresnayi* haploid gametophytes were cultivated under laboratory conditions as in ref. 78. We cultivated the gametophytes at 14 °C with a photoperiod of 12:12 h light:dark at an irradiance of 25 µmol photons m^{-2} s^{-1}. The media consisted of filtered natural seawater (NSW), which was autoclaved and enriched with half-strength Provasoli nutrient solution (Provasoli-enriched seawater; PES)[78]. We grew the first biomass in 140 mm Petri dishes and the gametophytes were later transferred to a 1 l flask with gentle aeration. The gametophytes were fragmented once a month and the media were changed every 2 weeks to promote biomass production. Before freezing, gametophytes were treated with antibiotics for 3 days with gentle agitation and under the same culture conditions. The first day, gametophytes were treated with a mix of streptomycin (2 g l^{-1} of PES), penicillin G (0.5 g l^{-1} of PES) and chloramphenicol (0.1 g l^{-1} of PES); the next day with ampicilin (1 g l^{-1} of PES), and on the last day with kanamycin (1 g l^{-1} of PES). Between each day of treatment and before freezing, gametophytes were rinsed with 500 ml of NSW to remove traces of antibiotic.

Samples of fucoid algae sexual and vegetative tissue were collected in the intertidal zone during low tides in June 2012 from Viana do Castelo (*F. vesiculosus*, *A. nodosum*) and Caminha (Rio Minho; *F. ceranoides*), northern Portugal. Sexual phenotypes were verified in the field by sectioning and observing receptacles under a field microscope. Tissue samples were flash frozen in liquid nitrogen on the shore and transported to the laboratory in a cryoshipper, after which they were lyophilised and stored dry at room temperature on silica crystals (see Supplementary Table 1 for a list of strains used in this study).

DNA and RNA extraction and sequencing

Genomic DNA was isolated from algal tissue (~100 mg) by grinding into fine powder under liquid nitrogen and subsequent cell lysis in 500 µl of Genomic Lysis Buffer (OMNIPREP for plant kit) for 1 h at 60 °C. The lysate was cleaned up with 200 µl of chloroform and DNA was precipitated in ethanol. The DNA pellet was digested in CF buffer (Macherey–Nagel) for 45 min at 65 °C and purified using NucleoBond AXG20 Mini columns according to the user manual (Macherey–Nagel). Final high molecular weight genomic DNA was quantified (Qubit), analysed for purity (Nanodrop) and checked for size distribution (Femto Pulse System) before preparing the sequencing libraries. We sequenced the libraries using an Oxford Nanopore Technologies (ONT) MinION Mk1B. We prepared the ONT libraries using an SQK-LSK110 library preparation kit for R9.4.1 flow cells and an SQK-LSK114 library preparation kit for R10.4.1 flow cells. Two libraries were sequenced for *D. dudresnayi* on R9.4.1 flow cells and a third library was sequenced on an R10.4.1 flow cell.

RNA was isolated from mature gametophytes of *U. pinnatifida* and *S. promiscuus* following modified procedure of the Qiagen RNAeasy kit, and the TruSeq RNA Library Prep Kit v.2 was used to sequence the transcriptomes in an Illumina NextSeq 2000 platform (150 bp, PE reads). Extraction of total RNA from fucoid algae (*F. vesiculosus*, *A. nodosum* and *F. ceranoides*) was performed following ref. 79 and RNA libraries were sequenced on an Illumina HiSeq 2000 machine (100 bp, PE reads).

Genome assembly and annotation

High-quality, chromosome-level assemblies of brown algal genomes have been notoriously difficult to obtain due to technical challenges in extracting nucleic acids. Whole-genome assemblies and annotations of *S. promiscuus* male, *D. dichotoma* male, *D. herbacea* male and female, *E. crouanorium* male, *C. linearis*, *S. ischiensis* and *F. serratus* male were obtained from ref. 18. We also downloaded the genome of *Ectocarpus* sp. 7 (ref. 19) and the male genome of *U. pinnatifida*[20], which were already assembled at a chromosome level. For *D. dudresnayi*, we performed genome sequencing, de novo genome assembly and ab initio gene annotation. Base calling was done using ONT Guppy[80] with the configuration files 'dna_r9.4.1_450bps_sup.cfg' and 'dna_r10.4.1_e8.2_400bps_sup.cfg' and the options '–trim_adapters –trim_primers', yielding 17.4 Gbp of data in 2,871,152 reads. We merged all the reads and analysed them using Kraken (v.2.1.2)[81] and the bacteria database (August 2022) to remove potential contaminant sequences. All data classified as bacterial reads by Kraken were screened using blastN (v.2.13.0+)[82] (-evalue 0.001 -num_alignments 20) against the NCBI genbank bacterial database (downloaded November 2023). The blastN output was visualized in MEGAN (v.6.23.4)[83], and all reads that were declared as bacterial were extracted and removed from further analyses. We obtained 1,908,772 decontaminated reads with an average length of 5.1 Kbp (9.8 Gbp of data, 20× coverage), which were deposited on the NCBI Sequence Read Archive (Supplementary Table 1).

The decontaminated reads were assembled de novo using flye (v.2.9.1-b1780)[84] with the options '–nano-raw -g 450 m -t 28 -i 3 –scaffold'. The draft assembly consisted of 1,032 contigs with a total size of 425 Mbp, an N50 of 4.6 Mbp and an L50 of 29 contigs. We used TransposonPSI (http://transposonpsi.sourceforge.net/) to predict

the TEs and RepeatScout (v.1.0.6)[85] to predict the simple repeats in the genome assembly. Both predictions were combined to soft mask the repetitive content in the genome assembly using bedtools maskfasta (v.2.27.1)[86]. We mapped the RNA-seq data of *D. dudresnayi* from the PhaeoExplorer database[18] to the soft-masked genome assembly using STAR (v.2.7.1a)[87]. We used BRAKER v.2.1.6 alongside the RNA-seq data[88] to predict the protein-coding genes in the soft-masked genome assembly.

Hi-C library preparation and sequencing for chromosome-level assemblies

We generated Hi-C libraries for three male genomes (*S. promiscuus*, *D. herbacea* and *D. dichotoma*) and two female genomes (*Ectocarpus* sp. 7 and *D. herbacea*). Fresh algal tissue was cross-linked for 20 min at room temperature in a solution of 2% formaldehyde with filtered NSW and then transferred into a 400 mM glycine solution with filtered NSW for 5 min to quench the formaldehyde. The samples were then stored at −80 °C until use. The Hi-C libraries were prepared as follows. The samples were de-frosted in 1 ml of 1× *Dpn*II buffer with protease inhibitors (Roche cOmplete), transferred to Precellys VK05 lysis tubes (Bertin) and disrupted using the Precellys apparatus with five grinding cycles of 30 s at 7,800 r.p.m., followed by 20 s pauses. SDS was added to the lysate at 0.5% final concentration and samples were incubated for 10 min at 62 °C, followed by the addition of Triton X-100 to a final concentration of 1% and 10 min of incubation at 37 °C under gentle shaking. We added 500 U of *Dpn*II to 4.6 ml of the digestion mixture and incubated the samples for 2 h at 37 °C under gentle shaking (180 r.p.m. in an inclined rack to prevent sedimentation), followed by the addition of another 500 U of *Dpn*II and an overnight incubation under the same conditions. The digested samples were centrifuged at 4 °C for 20 min at 16,000 × *g*. The supernatant was discarded and the pellet was incubated for biotinylation at 37 °C for 1 h under a constant shaking (300 r.p.m.) in a 500 ml biotinylation mix with a concentration of 1× ligation buffer, 0.09 mM dATP-dGTP-dTTP, 0.03 mM biotin-14-dCTP and 0.64 U ml^{-1} Klenow fragments. After biotinylation, the samples were incubated for 3 h at room temperature in a 1.2 ml ligation reaction with a concentration of 1× ligation buffer, 100 mg ml^{-1} BSA, 1 mM ATP and 0.4 U ml^{-1} T4 DNA Ligase. The samples were then incubated overnight at 65 °C after adding 20 µl 0.5 M EDTA, 80 µl 10% SDS and 1.6 mg Proteinase K. DNA was extracted with 1 volume of phenol/cholorform/isoamyl alcohol (24:24:1), followed by 30 s of vortexing at top speed and a 5-min centrifugation at top speed. We precipitated the DNA by adding 1/10 volume of 3 M NaAC pH 5 and two volumes of 100% cold ethanol, followed by a 30-min incubation at −80 °C and a 20-min centrifugation at 14,000 × *g* and 4 °C. The DNA pellet was washed with 1 ml 70% ethanol, then dried at 37 °C for 10 min and resuspended in 100 µl 1× TE buffer with 1 mg ml^{-1} RNase. DNA was sheared to 250–500 bp fragments using a Covaris S220 ultrasonicator, purified with AMPure beads (0.6×) (Beckman) and eluted in 20 µl 10 mM Tris pH 8.0. Biotinylated but not ligated DNA fragments were first removed by T4 DNA polymerase treatment (final concentration, 300 U per pellet; NEB), and the biotin-labelled fragments were selectively captured by Dynabeads MyOne Streptavidin C1 beads (Invitrogen). The libraries were prepared using the NEB Ultra II library preparation system and sequenced on the NextSeq 2000 Illumina platform (2 × 150 bp) (Supplementary Table 1).

We scaffolded the genomes from ref. 18 into chromosome-level assemblies using the Hi-C data. We filtered the low-quality Hi-C reads using Trimmomatic (v.0.39)[89] (ILLUMINACLIP:2:30:10 LEADING:25 TRAILING:25 SLIDINGWINDOW:4:15 MINLEN:75 AVGQUAL:28). We mapped the Hi-C reads against each genome assembly using BWA-mem v.0.7.17-r1188d in the Juicer v.1.6 pipeline[90] to generate a contact map, which was then fed to 3D-DNA v190716 (ref. 91) to scaffold the genomes into chromosomes. The obtained scaffolds were manually inspected against the contact maps to solve the limits of each chromosome using Juicebox (v.1.11.08)[92]. The PhaeoExplorer gene annotations[18] were lifted into the new assemblies using Liftoff (v.1.6.1)[93], while the annotation of TEs was performed using RepeatModeler2 (ref. 94). We scaffolded the genomes of *E. crouaniorum* and *D. dudresnayi* into chromosomes using a reference-guided assembly with RagTag (v.2.0.1)[95] against the chromosome-level assemblies of *Ectocarpus* sp. 7 and *D. herbacea*, respectively. All genes within the SDRs in the brown algal species studied (see below) were manually curated to exclude any TE-related genes from the annotation.

Discovery of the U/V sex determination regions

Male sex-determining regions (V-SDR) in *S. promiscuus*, *U. pinnatifida*, *D. herbacea* and *D. dichotoma*, as well as female sex-determining region (U-SDR) in *D. herbacea* were analysed following two complementary methods: (1) a *k*-mer-based YGS approach, originally designed to detected Y-linked sequences in heterogametic systems, developed in ref. 96 and (2) genomic coverage analysis, designed to identify sex-linked regions through differences in read depth between male and female individuals[97]. These methods are well suited for organisms with divergent sex chromosomes, such as brown algae, where U and V haplotypes have diverged over extended evolutionary time.

The YGS method principle is to identify male or female sex-linked scaffolds by comparing *k*-mer frequencies between reference genome assembly and *k*-mers generated from DNA-seq reads of the opposite sex. Regions in the male reference genome that contain *k*-mers that are absent in female reads will indicate candidate male SDR sequences; similarly, female genomic scaffolds with low coverage in male *k*-mers will denote female SDR region. For each species, 15-base-pair *k*-mer sequences were generated separately from male and female Illumina reads (see Supplementary Table 1 for data accession numbers) using Jellyfish v.2.3.0 count (-m 15 -s 10 G -C –quality-start=33 –min-quality=20) and converted to fasta format with Jellyfish dump (–lower-count=5)[98]. Next, non-overlapping 500-kb sliding windows (*Desmarestia*, *Dictyota* and *Undaria*) or 200-kb sliding windows (*Scytosiphon*) of the reference chromosome genomes (from the sex whose SDR was to be identified) were created using seqkit (v.2.3.1)[99] and used as input for the YGS.pl script[96] together with the fasta *k*-mer files produced in the previous step. Each window was then analysed to calculate the proportion of *k*-mers in the reference window that are not present in the opposite-sex *k*-mer database. Genomic windows with a minimum of ≥50% of unmatched single-copy *k*-mers were then retained as candidate male or female SDR sequences. Because the borders of the SDRs cannot be precisely defined at the single-nucleotide level with the available data, we focused on genes within these regions and defined the SDR boundaries on the basis of the flanking genes located at the transition to pseudoautosomal regions (PARs).

Candidate SDR regions identified by YGS were further validated by analysing sex-specific differences in read coverage. In detail, the short Illumina reads coming from males and females of each investigated species were trimmed with Trimmomatic[89] (see above) and mapped to the reference genome for which the SDR was to be studied, using HISAT2 (ref. 100) (default settings). Bam files produced by HISAT2 were used as input for Mosdepth[101] to calculate coverage in 10-kb windows along the genome sequence (-m -n -b 10000 –fast-mode -Q 30). Read mapping depth in genomic windows was normalized by the genome-wide mean for each sex, and the coverage in genomic intervals was then compared between males and females. Because V-SDR-linked sequences are present only in males, we expect them to have similar read coverage as autosomal regions in males, but little or no coverage in females (and conversely for the U-SDR sequences in *D. herbacea*). The comparison focused on regions within male reference genomes where the coverage in males fell within the range of 75–125% of the genome average, while the coverage in females remained below 50% of the genome average.

Both coverage and *k*-mer analysis identified identical genomic regions, providing high-confidence candidate SDRs (Extended Data Table 1 and Supplementary Figs. 1–5). In *D. herbacea*, where both male and female chromosome-level genome assemblies were available,

we directly compared U and V chromosomes to further confirm the SDR borders by analysing the collinearity of pseudoautosomal regions flanking the SDRs. The SDR scaffolds for all studied species were further validated by PCR amplification (Supplementary Table 1) using 4 males and 4 females.

Genetic mapping and search for the sex chromosome in *F. serratus*

Three different sets of materials were used in this study: (1) 12 male and 12 female field samples, hereafter denoted the 24-individual natural population; (2) 157 sporophyte progeny population derived from a cross between one male sample and one female sample collected from the field and (3) 3 male and 3 female samples collected from the field for whole-genome sequencing. The 157-progeny population and 24-individual natural population were genotyped using the double digest RAD sequencing approach (ddRAD-seq). Briefly, individual genomic DNA was digested with the restricted enzymes PstI and HhaI to obtain fragments that were size selected between 400 and 800 bp before sequencing on in Illumina HiSeq 2500 platform (paired-end 2 × 125 bp). See ref. 102 for detailed protocol of the ddRAD-seq.

We performed whole-genome sequencing on an Illumina HiSeq 2500 system (2 × 150 bp paired-end) for the 3 male and 3 female samples. For ddRAD-seq data, raw reads were cleaned and trimmed with Trimmomatic as above and mapped to the draft genome of *F. serratus* male. For the progeny population, genotypes were called from the obtained bam files using the Stacks pipeline (v.2.5)[103]. The obtained vcf files were filtered with VCFtools (v.0.1.16)[104] and bcftools[105] (max missing per locus: 30%, max missing per sample: 40%, max mean coverage:30, minQG:20).

The filtered vcf file of the progeny population was used to construct a genetic map with Lep-MAP3 (ref. 106). Briefly, the ParentCall2 module was used to call parental genotypes, the SeparateChromosomes2 module was used to split the markers into linkage groups and the OrderMarkers2 module was used to order the markers within each linkage group using 30 iterations per group and finally computing genetic distances. Phased data were converted to informative genotypes with the script map2genotypes.awk.

We used different approaches to identify the SDR in *F. serratus*.

Coverage analysis. We combined whole-genome sequence data from the 3 males and 3 females alongside the ddRAD-seq data of the 24-individual natural population, mapping both datasets to the *F. serratus* male genome assembly using bwa-mem[107]. Coverage analyses was done in several ways:

- Using SATC (sex assignment through coverage)[108], a method that uses sequencing depth distribution across scaffolds to jointly identify: (1) male and female individuals and (2) sex-linked scaffolds. This identification was achieved by projecting the scaffold depths into a low-dimensional space using principal component analysis and subsequent Gaussian mixture clustering. Male and female whole-genome sequences were used for this analysis.
- Using the method SexChrCov described in ref. 109 with the 24-individual natural population.
- Using the method DifCover[110] which identifies regions in a reference genome for which the read coverage of one sample is significantly different from the read coverage of another sample when aligned to a common reference genome. The 24-individual natural population was used for this analysis.
- Using soap.coverage (v.2.7.9)[111] to calculate the coverage (number of times each site was sequenced divided by the total number of sequenced sites) of each scaffold in each sample. For each scaffold, the male to female (M:F) fold change coverage was calculated as $\log_2$(average male coverage) − $\log_2$(average female coverage). The 24-individual natural population was used for this analysis.

Fixation index (F_{ST}) and sex-biased heterozygosity. This approach has been previously used to find sex-linked genomic regions in several studies[112,113]. Using the 24-individual natural population, F_{ST} was calculated using vcftools[104]. Sex-biased heterozygosity was defined as the $\log_{10}$ of the male heterozygosity:female heterozygosity ratio, where heterozygosity was measured as the fraction of sites that are heterozygous. This ratio is expected to be zero for autosomal scaffolds and elevated on young sex scaffolds due to excess heterozygosity in males.

Identification of eventual female scaffolds that failed to map to the male reference genome. Vcftools and bedtools were used to extract female regions that did not map to the reference genome consistently in the 3 resequenced female samples.

All candidate contigs were tested by PCR in 4 males and 4 females.

Synteny analyses, K_s analysis and transitions to co-sexuality

Whole-genome synteny comparisons were performed for each pair of chromosome-level assemblies using MCscan (v.1.2.14)[114], both between different species, between sex chromosomes in the same species and between monoicous species and their closest relatives with U/V chromosomes. The putative gametologues between sex chromosomes that were predicted with MCscan were reassessed using OrthoFinder (v.2.5.4)[115] and best reciprocal DIAMOND (v.2.1.8.162)[116] hits.

We calculated the number of synonymous substitutions per synonymous site (K_s) for each pair of male and female gametologues as a proxy to assess the relative time at which both genes diverged from each other. The amino acid sequences of each pair of gametologues were aligned with MAFFT (v.7.520)[117] and subsequently aligned into codons using pal2nal (v.14)[118]. The gametologue K_s values were calculated using the model in ref. 119 as implemented in KaKs_calculator (v.2.0)[120].

We evaluated the male or female identity of the genes in the monoicous species whose orthologues were found within the SDR in their closest dioicous relatives. For this, we compared the results obtained with MCscan[114] against the orthogroup prediction performed with OrthoFinder[115], with best reciprocal DIAMOND[116] hits and by calculating gene trees for each orthogroup using an amino acid alignment with MAFFT[117] and gene tree reconstructions using FastTree (v.2.1.11)[121].

Ancestral reconstruction of the male SDR

The brown algal phylogeny was obtained from ref. 18. The species tree is based on 32 single-copy nuclear genes whose protein sequences were aligned manually using AliView[122], and whose best-fit substitution models were assessed independently using the Akaike information criterion. The tree was generated using a maximum likelihood approach implemented in RAxML bootstraps and the gamma model. Every node in the phylogeny has 99–100% bootstrap support values. Divergence times were subsequently calculated using MCMCtree[123] and three calibration points. The MCMC chains were run for 1.5 million generations and the first 200,000 MCMC chains were discarded as burn-in.

We searched for orthologue genes within the V-SDR of five species (*Ectocarpus* sp. 7, *S. promiscuus*, *U. pinnatifida*, *D. herbacea* and *D. dichotoma*) in our OrthoFinder results. For each V-SDR gene, we coded its orthologue in the other species as 'present' (1) if it is also sex linked in the V-SDR, whereas it was coded as 'absent' (0) if the orthologue resides in the PARs, in an autosome or if there is no detectable orthologue in that species. Once we generated this presence/absence matrix with the evolutionary relationship of the genes within the V-SDR (Supplementary Table 5), we used it as the input file for the software Count (v.10.04)[124] to estimate the ancestral content of the V-SDR throughout a phylogeny and determine the most likely scenario of V-SDR evolution in the brown algae. We employed posterior probabilities under a phylogenetic birth-and-death model with independent gain and loss rates across each branch in the phylogeny. We modelled the independent gain and loss rates through 10 gamma categories and performed 1,000 optimization rounds with a convergence threshold

on the likelihood >0.1 to find the best fitting model for the data. The branch lengths in the tree that were used for the ancestral state reconstruction were retrieved from the molecular clock analysis performed in ref. 16. We distinguished between conserved V-SDR genes that are ancestral and parallel acquisitions of the same gene in the V-SDR by analysing gene trees between male and female genomes, in addition to female transcriptome assemblies of *D. dichotoma* and *U. pinnatifida*. Sequence alignments were done using MAFFT[117] with default settings and uploaded to the https://www.phylogeny.fr/ platform. Alignments were further curated using Gblocks (v.0.91b)[125] (min. seq. for flank pos.: 85%, max. contig. nonconserved pos.: 8, min. block length: 10). Trees were produced using PhyML (v.3.1l)[126] with the default model and visualized in TreeDyn (v.198.3)[127]. The approximate likelihood-ratio test was chosen as the statistical test for branch support. We inferred the function of the ancestral V-SDR genes through the annotation of genes in *Ectocarpus* sp. 7 belonging to that orthogroup. The most likely acquisition mechanism of each SDR gene in each species was assessed on the basis of the position of each orthologue in the other species (pseudoautosomic, autosomic or missing; Supplementary Table 5).

Genomic content across chromosomes

We used closely related genome assemblies available in the PhaeoExplorer database[18] to assess the depletion of orthologues in the sex chromosome. We predicted one-to-one orthologues using OrthoFinder[115] between the following species pairs: *Ectocarpus* sp. 7 with *Ectocarpus siliculosus*, *S. promiscuus* with *C. linearis*, *U. pinnatifida* with *Saccharina japonica*, *F. serratus* with *Fucus distichus*, *D. herbacea* with *D. dudresnayi*, and *D. dichotoma* with *Halopteris paniculata* (Supplementary Table 16). We calculated the expected number of detectable orthologues for each chromosome and compared it against the observed number of detected orthologues using chi-square tests. We performed Benjamini–Hochberg corrections to the *P* values of the chi-square tests to control the FDR in the analysis[128].

GenEra[36] was used by running DIAMOND in ultra-sensitive mode[116] against the NCBI NR database and all the PhaeoExplorer proteins[18] to perform a phylostratigraphic analysis (*e*-value threshold of 10^{-5}) and calculate the relative ages of each gene in each genome (Supplementary Table 13). Phylostratigraphy is a genetic statistical method developed to date the putative origin of all the genes contained in the genome of a target species by detecting homologues across species at different evolutionary distances (all the way from species within the same genus to species from different domains of life). Finding the most distant homologues of each gene can link them to their founder events (that is, the first instance where a gene homologue is found in the history of that lineage), allowing us to then determine their relative ages, coded as the taxonomic group where that gene is detected[35,36,129]. The gene age categories outside of the brown algae and *S. ischiensis* were based on the taxonomic classification of each species within the NCBI Taxonomy database[130], while the gene ages within the brown algae were manually assessed to reflect the evolutionary relationships obtained in the PhaeoExplorer maximum likelihood tree[18]. We performed Wilcoxon rank-sum tests in R (v.4.3.1)[131] to assess non-random differences in gene age distributions between pairs of chromosomes (Supplementary Table 14). We performed Benjamini–Hochberg corrections to the *P* values of the Wilcoxon rank-sum tests to control the FDR in the analysis[128]. The gene ages responsible for these differences were found by evaluating the standardized residuals using mosaic plots.

We used the interspecies K_s values between pairs of species as a proxy for neutral mutation rates across six of the seven chromosome-level assemblies by using the most closely related genome assemblies available in the PhaeoExplorer database[18]. We used the same set of one-to-one orthologues detected between species pairs as for the orthologue-depletion test (Supplementary Table 16). However, the evolutionary distance between *D. dichotoma* and *H. paniculata* prevented us from calculating reliable interspecies K_s values for this species since synonymous substitutions reached the point of saturation. The amino acid sequences of each pair of orthologues were aligned with MAFFT[117] and subsequently aligned into codons using pal2nal[118]. The interspecies K_s values were calculated using the model in ref. 119 as implemented in KaKs_calculator (v.2.0)[120]. We also evaluated the difference in interspecies K_s values between the autosomes and the sex chromosomes through FDR-corrected Wilcoxon rank-sum tests (Supplementary Table 17). We calculated the protein-coding density, the density of TEs and the taxonomic identity of these TEs within 100-kb non-overlaping windows across each chromosome using bedtools[86] (Supplementary Table 9). The differences in protein-coding space and repeat content between the autosomes and the sex chromosomes were also tested using FDR-corrected Wilcoxon rank-sum tests (Supplementary Tables 10 and 11). The differences in repeat density, percentage of unclassified repeats and gene ages across genomic compartments (SDR, PARs and autosomes) were tested using FDR-corrected permutation tests with 10,000 permutations. All genomic features were plotted using karyoploteR (v.1.20.3)[132].

Gene expression analysis

We used kallisto (v.0.44.0)[133] to calculate gene expression levels using 31-base-pair-long *k*-mers and 1,000 bootstraps. Transcript abundances were then summed within genes using the tximport v.3.19 package[134] to obtain the expression level for each gene in transcripts per million (TPM). Differential expression analysis was done in the DESeq2 v.3.19 package[135] in R v.4.3.1, applying FC ≥ 2 and $p_{adj} < 0.05$ cut-offs. Sex-biased gene expression analysis in *Ectocarpus* sp. 7, *S. promiscuus*, *U. pinnatifida*, *D. herbacea* and *D. dichotoma* was performed between mature male and female gametophytes (gametophytes bearing reproductive structures). To discover genes with sporophyte-biased expression in *Ectocarpus* sp.7, *S. promiscuus*, *U.pinnatifida* and *D. dichotoma*, we first calculated the differential expression between male gametophytes and sporophytes, as well as female gametophytes and sporophytes. Genes that showed significant sporophyte-biased expression (FC ≥ 2, $p_{adj} < 0.05$) in both comparisons were considered sporophyte biased.

A total of 314.2 M RNA-seq reads from *F. vesiculosus* male, female and vegetative tissue were assembled de novo with rnaSPAdes[136] using *k*-mer values of 33 and 49. Assembly quality was assessed by (pseudo) mapping reads back onto the resulting assembly and retaining 'good' contigs as defined using TransRate (v.1.0.3)[137] with default settings. The resulting 159,108 contigs were aligned with BLASTx[82] against a database of Stramenopile proteins, and those with top hits against brown algae (Phaeophyceae) were retained as the final curated reference transcriptome (36,394 contigs, N50 = 1,770 bp). Transcript expression levels were determined by mapping the reads from all samples against the reference transcriptome using Bowtie2 (ref. 138) and the RSEM-EBSeq (v.1.3.3)[139] pipeline, and relative expression values were recorded as TPM. All samples used in the gene expression analysis can be found in Supplementary Table 1.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The accession numbers and download links for all the chromosome-level genome assemblies and annotations that were generated and used in this study are available in Supplementary Table 1 and in the Edmond Repository[140] at https://doi.org/10.17617/3.OOWB2Y. The raw sequence reads for the Oxford Nanopore data, Hi-C libraries and RNA-seq libraries are available in the Sequence Read Archive under BioProject accession number PRJNA1059008. All genome assemblies and annotations are also accessible through the Phaeoexplorer database (https://phaeoexplorer.sb-roscoff.fr/) for comparative genomics

analyses. This paper does not report original data. Further information and requests for resources and reagents should be directed to and will be fulfilled by S.M.C. (susana.coelho@tuebingen.mpg.de).

Code availability

This paper does not report original code.

References

1. Bachtrog, D. et al. Sex determination: why so many ways of doing it? *PLoS Biol.* **12**, e1001899 (2014).
2. Beukeboom, L. W. & Perrin, N. *The Evolution of Sex Determination* (Oxford Univ. Press, 2014).
3. Vicoso, B. Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nat. Ecol. Evol.* **3**, 1632–1641 (2019).
4. Mignerot, L. & Coelho, S. M. The origin and evolution of the sexes: novel insights from a distant eukaryotic linage. *C. R. Biol.* **339**, 252–257 (2016).
5. Ma, W. & Rovatsos, M. Sex chromosome evolution: the remarkable diversity in the evolutionary rates and mechanisms. *J. Evol. Biol.* **35**, 1581–1588 (2022).
6. Umen, J. & Coelho, S. Algal sex determination and the evolution of anisogamy. *Annu. Rev. Microbiol.* **73**, 267–291 (2019).
7. Coelho, S. M. & Umen, J. Switching it up: algal insights into sexual transitions. *Plant Reprod.* **34**, 287–296 (2021).
8. Avia, K. et al. Genetic diversity in the UV sex chromosomes of the brown alga *Ectocarpus*. *Genes* **9**, 286 (2018).
9. Coelho, S. M., Mignerot, L. & Cock, J. M. Origin and evolution of sex-determination systems in the brown algae. *New Phytol.* **222**, 1751–1756 (2019).
10. Bull, J. J. Sex chromosomes in haploid dioecy: a unique contrast to Muller's theory for diploid dioecy. *Am. Nat.* **112**, 245–250 (1978).
11. Carey, S. B. et al. Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. *Sci. Adv.* **7**, eabn2488 (2021).
12. Healey, A. L. et al. Newly identified sex chromosomes in the *Sphagnum* (peat moss) genome alter carbon sequestration and ecosystem dynamics. *Nat. Plants* **9**, 238–254 (2023).
13. Bowman, J. L. et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304.e15 (2017).
14. Silva, A. T. et al. To dry perchance to live: insights from the genome of the desiccation-tolerant biocrust moss *Syntrichia caninervis*. *Plant J.* **105**, 1339–1356 (2021).
15. Bechteler, J. et al. Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *Am. J. Bot.* **110**, e16249 (2023).
16. Heesch, S. et al. Evolution of life cycles and reproductive traits: insights from the brown algae. *J. Evol. Biol.* **34**, 992–1009 (2021).
17. Bringloe, T. T. et al. Phylogeny and evolution of the brown algae. *Crit. Rev. Plant Sci.* **39**, 281–321 (2020).
18. Denoeud, F. et al. Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems. *Cell* **187**, 6943–6965.e39 (2024).
19. Liu, P. et al. 3D chromatin maps of a brown alga reveal U/V sex chromosome spatial organization. *Nat. Commun.* **15**, 9590 (2024).
20. Shan, T. et al. First genome of the brown alga *Undaria pinnatifida*: chromosome-level assembly using PacBio and Hi-C technologies. *Front. Genet.* **11**, 140 (2020).
21. Choi, S.-W. et al. Ordovician origin and subsequent diversification of the brown algae. *Curr. Biol.* https://doi.org/10.1016/j.cub.2023.12.069 (2024).
22. Luthringer, R. et al. Repeated co-option of HMG-box genes for sex determination in brown algae and animals. *Science* **383**, eadk5466 (2024).
23. Simakov, O. et al. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**, eabi5884 (2022).
24. Sun, Y., Svedberg, J., Hiltunen, M., Corcoran, P. & Johannesson, H. Large-scale suppression of recombination predates genomic rearrangements in *Neurospora tetrasperma*. *Nat. Commun.* **8**, 1140 (2017).
25. Ahmed, S. et al. A haploid system of sex determination in the brown alga *Ectocarpus* sp. *Curr. Biol.* **24**, 1945–1957 (2014).
26. Charlesworth, D. Plant contributions to our understanding of sex chromosome evolution. *New Phytol.* **208**, 52–65 (2015).
27. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
28. Lipinska, A. P. et al. Multiple gene movements into and out of haploid sex chromosomes. *Genome Biol.* **18**, 104 (2017).
29. Charlesworth, D. The mysterious sex chromosomes of haploid plants. *Heredity* **129**, 17–21 (2022).
30. Bolwell, G. P., Callow, J. A., Callow, M. E. & Evans, L. V. Fertilization in brown algae: II. evidence for lectin-sensitive complementary receptors involved in gamete recognition in *Fucus serratus*. *J. Cell Sci.* **36**, 19–30 (1979).
31. Zhao, Z.-S., Leung, T., Manser, E. & Lim, L. Pheromone signalling in *Saccharomyces cerevisiae* requires the small GTP-binding protein Cdc42p and its activator *CDC24*. *Mol. Cell. Biol.* **15**, 5246–5257 (1995).
32. Ponnikas, S., Sigeman, H., Abbott, J. K. & Hansson, B. Why do sex chromosomes stop recombining? *Trends Genet.* **34**, 492–503 (2018).
33. Cossard, G. G. et al. Selection drives convergent gene expression changes during transitions to co-sexuality in haploid sexual systems. *Nat. Ecol. Evol.* **6**, 579–589 (2022).
34. Bachtrog, D. Accumulation of Spock and Worf, two novel non-LTR retrotransposons, on the neo-Y chromosome of *Drosophila miranda*. *Mol. Biol. Evol.* **20**, 173–181 (2003).
35. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
36. Barrera-Redondo, J., Lotharukpong, J. S., Drost, H.-G. & Coelho, S. M. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biol.* **24**, 54 (2023).
37. Luthringer, R. et al. The pseudoautosomal regions of the U/V sex chromosomes of the brown alga *Ectocarpus* exhibit unusual features. *Mol. Biol. Evol.* **32**, 2973–2985 (2015).
38. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
39. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
40. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
41. Keightley, P. D. & Eyre-Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331–333 (2000).
42. Montgomery, S. A. et al. Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr. Biol.* **30**, 573–588.e7 (2020).
43. Loftus, B. J. et al. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324 (2005).
44. Clayton, M. N. Isogamy and a fucalean type of life history in the Antarctic brown alga *Ascoseira mirabilis* (Ascoseirales, Phaeophyta). *Bot. Mar.* **30**, 447–454 (1987).
45. Coyer, J. A., Peters, A. F., Hoarau, G., Stam, W. T. & Olsen, J. L. Hybridization of the marine seaweeds, *Fucus serratus and Fucus evanescens* (Heterokontophyta: Phaeophyceae) in a 100-year-old zone of secondary contact. *Proc. R. Soc. B* **269**, 1829–1834 (2002).

46. Hatchett, W. J. et al. Evolutionary dynamics of sex-biased gene expression in a young XY system: insights from the brown alga genus *Fucus*. *New Phytol.* **238**, 422–437 (2023).
47. Cánovas, F. G., Mota, C. F., Serrão, E. A. & Pearson, G. A. Driving south: a multi-gene phylogeny of the brown algal family Fucaceae reveals relationships and recent drivers of a marine radiation. *BMC Evol. Biol.* **11**, 371 (2011).
48. Lee, S. C., Ni, M., Li, W., Shertz, C. & Heitman, J. The evolution of sex: a perspective from the fungal kingdom. *Microbiol. Mol. Biol. Rev.* **74**, 298–340 (2010).
49. Duhamel, M., Hood, M. E., Rodríguez de la Vega, R. C. & Giraud, T. Dynamics of transposable element accumulation in the non-recombining regions of mating-type chromosomes in anther-smut fungi. *Nat. Commun.* **14**, 5692 (2023).
50. Dolgin, E. S. & Charlesworth, B. The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* **178**, 2169–2177 (2008).
51. Gray, Y. H. M. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16**, 461–468 (2000).
52. Charlesworth, D. Evolution of recombination rates between sex chromosomes. *Phil. Trans. R. Soc. B* **372**, 20160456 (2017).
53. Jordan, C. Y. & Charlesworth, D. The potential for sexually antagonistic polymorphism in different genome regions. *Evolution* **66**, 505–516 (2012).
54. Luthringer, R. et al. Sexual dimorphism in the brown algae. *Perspect. Phycol.* **1**, 11–25 (2014).
55. Scharmann, M., Rebelo, A. G. & Pannell, J. R. High rates of evolution preceded shifts to sex-biased gene expression in *Leucadendron*, the most sexually dimorphic angiosperms. *Elife* **10**, e67485 (2021).
56. Charlesworth, B. Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl Acad. Sci. USA* **75**, 5618–5622 (1978).
57. Lenormand, T., Fyon, F., Sun, E. & Roze, D. Sex chromosome degeneration by regulatory evolution. *Curr. Biol.* **30**, 3001–3006. e5 (2020).
58. Jeffries, D. L., Gerchen, J. F., Scharmann, M. & Pannell, J. R. A neutral model for the loss of recombination on sex chromosomes. *Phil. Trans. R. Soc. B* **376**, 20200096 (2021).
59. Jay, P., Tezenas, E., Véber, A. & Giraud, T. Sheltering of deleterious mutations explains the stepwise extension of recombination suppression on sex chromosomes and other supergenes. *PLoS Biol.* **20**, e3001698 (2022).
60. Coelho, S. M., Gueno, J., Lipinska, A. P., Cock, J. M. & Umen, J. G. UV chromosomes and haploid sexual systems. *Trends Plant Sci.* **23**, 794–807 (2018).
61. Peona, V. et al. Teaching transposon classification as a means to crowd source the curation of repeat annotation—a tardigrade perspective. *Mob. DNA* **15**, 10 (2024).
62. Ivics, Z. & Izsvák, Z. The expanding universe of transposon technologies for gene and cell engineering. *Mob. DNA* **1**, 25 (2010).
63. Gueno, J. et al. Chromatin landscape associated with sexual differentiation in a UV sex determination system. *Nucleic Acids Res.* **50**, 3307–3322 (2022).
64. Bourdareau, S. et al. Histone modifications during the life cycle of the brown alga *Ectocarpus*. *Genome Biol.* **22**, 12 (2021).
65. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
66. Charlesworth, B. The effects of deleterious mutations on evolution at linked sites. *Genetics* **190**, 5–22 (2012).
67. Lebherz, M. K., Fouks, B., Schmidt, J., Bornberg-Bauer, E. & Grandchamp, A. DNA transposons favor de novo transcript emergence through enrichment of transcription factor binding motifs. *Genome Biol. Evol.* https://doi.org/10.1093/gbe/evae134 (2024).
68. McDaniel, S. F. Divergent outcomes of genetic conflict on the UV sex chromosomes of *Marchantia polymorpha* and *Ceratodon purpureus*. *Curr. Opin. Genet. Dev.* **83**, 102129 (2023).
69. Iwasaki, M. et al. Identification of the sex-determining factor in the liverwort *Marchantia polymorpha* reveals unique evolution of sex chromosomes in a haploid system. *Curr. Biol.* **31**, 5522–5532.e7 (2021).
70. Ligrone, R., Duckett, J. G. & Renzaglia, K. S. The origin of the sporophyte shoot in land plants: a bryological perspective. *Ann. Bot.* **110**, 935–941 (2012).
71. McDaniel, S. F., Neubig, K. M., Payton, A. C., Quatrano, R. S. & Cove, D. J. Recent gene-capture on the UV sex chromosomes of the moss *Ceratodon purpureus*. *Evolution* https://doi.org/10.1111/evo.12165 (2013).
72. Vigneau, J. et al. Interactions between U and V sex chromosomes during the life cycle of *Ectocarpus*. *Development* **151**, dev202677 (2024).
73. Singh, S., Davies, K. M., Chagné, D. & Bowman, J. L. The fate of sex chromosomes during the evolution of monoicy from dioicy in liverworts. *Curr. Biol.* **33**, 3597–3609.e3 (2023).
74. Takahashi, K. et al. Reorganization of the ancestral sex-determining regions during the evolution of trioecy in *Pleodorina starrii*. *Commun. Biol.* **6**, 590 (2023).
75. Liesner, D. et al. Developmental pathways underlying sexual differentiation in the U/V sex chromosome system of giant kelp. *Dev. Cell* https://doi.org/10.1016/j.devcel.2024.12.022 (2025).
76. Yamamoto, K. et al. Three genomes in the algal genus Volvox reveal the fate of a haploid sex-determining region after a transition to homothallism. *Proc. Natl Acad. Sci. USA* **118**, e2100712118 (2021).
77. Herpin, A. & Schartl, M. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO Rep.* **16**, 1260–1274 (2015).
78. Starr, R. C. & Zeikus, J. A. UTEX—the culture collection of algae at the University of Texas at Austin 1993 list of cultures. *J. Phycol.* **29**, 1–106 (1993).
79. Pearson, G., Lago-Leston, A., Valente, M. & Serrão, E. Simple and rapid RNA extraction from freeze-dried tissue of brown algae and seagrasses. *Eur. J. Phycol.* **41**, 97–104 (2006).
80. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
81. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
82. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
83. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
84. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
85. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
86. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
87. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
88. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).

89. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
90. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
91. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
92. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
93. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
94. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
95. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
96. Carvalho, A. B. & Clark, A. G. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* **23**, 1894–1907 (2013).
97. Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S. & Bachtrog, D. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biol.* **11**, e1001643 (2013).
98. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
99. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
100. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
101. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
102. Avia, K. et al. High-density genetic map and identification of QTLs for responses to temperature and salinity stresses in the model brown alga *Ectocarpus*. *Sci. Rep.* **7**, 43241 (2017).
103. Rochette, N. C., Rivera-Colón, A. G. & Catchen, J. M. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* **28**, 4737–4754 (2019).
104. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
105. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
106. Rastas, P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33**, 3726–3732 (2017).
107. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
108. Nursyifa, C., Brüniche-Olsen, A., Garcia-Erill, G., Heller, R. & Albrechtsen, A. Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data. *Mol. Ecol. Resour.* **22**, 458–467 (2022).
109. Malde, K., Skern, R. & Glover, K. A. Using sequencing coverage statistics to identify sex chromosomes in minke whales. Preprint at https://doi.org/10.48550/arXiv.1902.06654 (2019).
110. Smith, J. J. et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat. Genet.* **50**, 270–277 (2018).
111. Darolti, I. et al. Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proc. Natl Acad. Sci. USA* **116**, 19031–19036 (2019).
112. Brelsford, A., Lavanchy, G., Sermier, R., Rausch, A. & Perrin, N. Identifying homomorphic sex chromosomes from wild-caught adults with limited genomic resources. *Mol. Ecol. Resour.* **17**, 752–759 (2017).
113. Toups, M. A., Rodrigues, N., Perrin, N. & Kirkpatrick, M. A reciprocal translocation radically reshapes sex-linked inheritance in the common frog. *Mol. Ecol.* **28**, 1877–1889 (2019).
114. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
115. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
116. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
117. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
118. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, 609–612 (2006).
119. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
120. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).
121. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
122. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
123. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
124. Csűös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
125. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
126. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
127. Chevenet, F., Brun, C., Bañuls, A.-L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
128. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
129. Lotharukpong, J. S. et al. A transcriptomic hourglass in brown algae. *Nature* **635**, 129–135 (2024).
130. Schoch, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
131. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021); https://www.r-project.org/
132. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
133. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

134. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
135. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
136. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-seq data. *Gigascience* **8**, giz100 (2019).
137. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
138. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
139. Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).
140. Haas, F. B. Several brown algal chromosome level assemblies, V1. *Edmond* https://doi.org/10.17617/3.OOWB2Y (2025).
141. Cormier, A. et al. 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 (2017).

Acknowledgements

This work was supported by the MPG, the CNRS, Sorbonne University, the ERC (grant no. 864038 and 638240 to S.M.C.), the France Génomique National infrastructure project Phaeoexplorer (ANR-10-INBS-09), the JSPS Overseas Research Fellowships (to M.H.), the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A), the Investissements d'Avenir project Idealg (ANR-10-BTBR-04-01), the European BG-01 Blue Growth H2020 project Genialg (727892) and the ANR project Epicycle (ANR-19-CE20-0028-01). S.M.C. was supported by the Moore Foundation (GBMF11489) and the Bettencourt-Schueller Foundation. J.B.-R. was supported by a Humboldt Research Fellowship for postdoctoral researchers from the Alexander von Humboldt Foundation. We thank the members of the Phaeoexplorer consortium, in particular C. Jolivet, L. Mest and D. Scornet for assistance with algae cultures, C. Cruaud for help with sequencing libraries preparation, E. Corre and A. Le Bars for support with the Phaeoexplorer database and A. Couloux for the genome assemblies and annotations. We also thank the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and Biogenouest network, for providing computing and storage resources.

Author contributions

J.B.-R. and A.P.L. conducted investigation (equal) and formal analysis (equal), designed the methodology (equal), performed visualization (equal), wrote the original draft (equal), and reviewed and edited the manuscript (equal). P.L. conducted investigation (supporting) and formal analysis (supporting). E.D., G.C., O.G., K.B., M.H., R.J.C., K.A., G.L., E.A., D.L., R.L., O.G., S.H., Z.N., L.B.-G. and A.F.P. conducted investigation (supporting). G.H., J.-M.A,. G.P., P.W., F.D. and J.M.C. performed data curation (supporting) and data acquisition (supporting). F.B.H. conducted investigation (supporting), designed the methodology (equal), and performed data curation (equal) and formal analysis (supporting). S.M.C. conceptualized the project (lead), acquired funding (lead), administered and supervised the roject (lead), designed the methodology (equal), performed visualization (supporting), wrote the original draft (equal), and reviewed and edited the manuscript (lead).

Funding

Open access funding provided by Max Planck Society.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-025-02838-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-025-02838-w.

Correspondence and requests for materials should be addressed to Susana M. Coelho.

Peer review information *Nature Ecology & Evolution* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

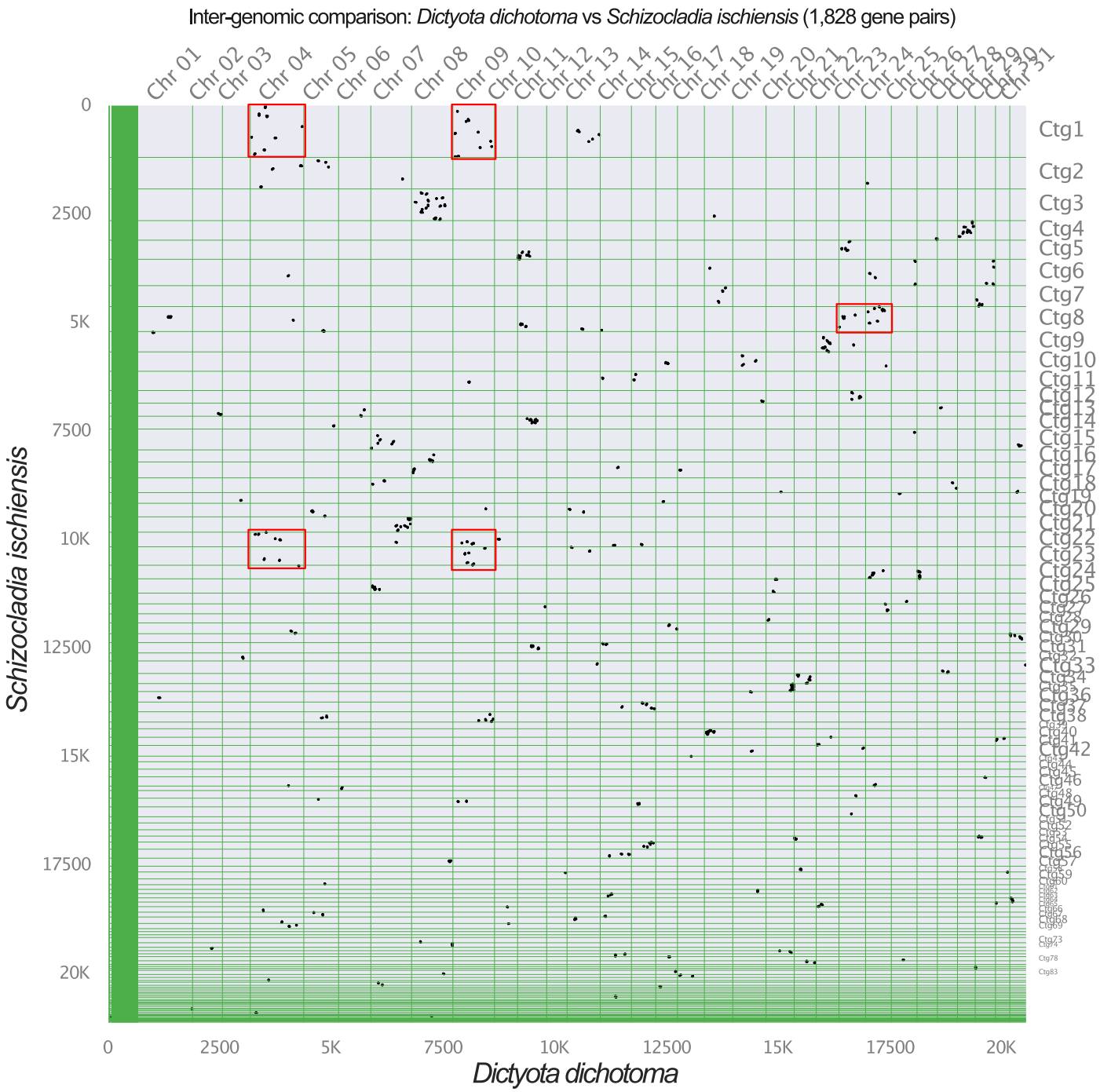
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

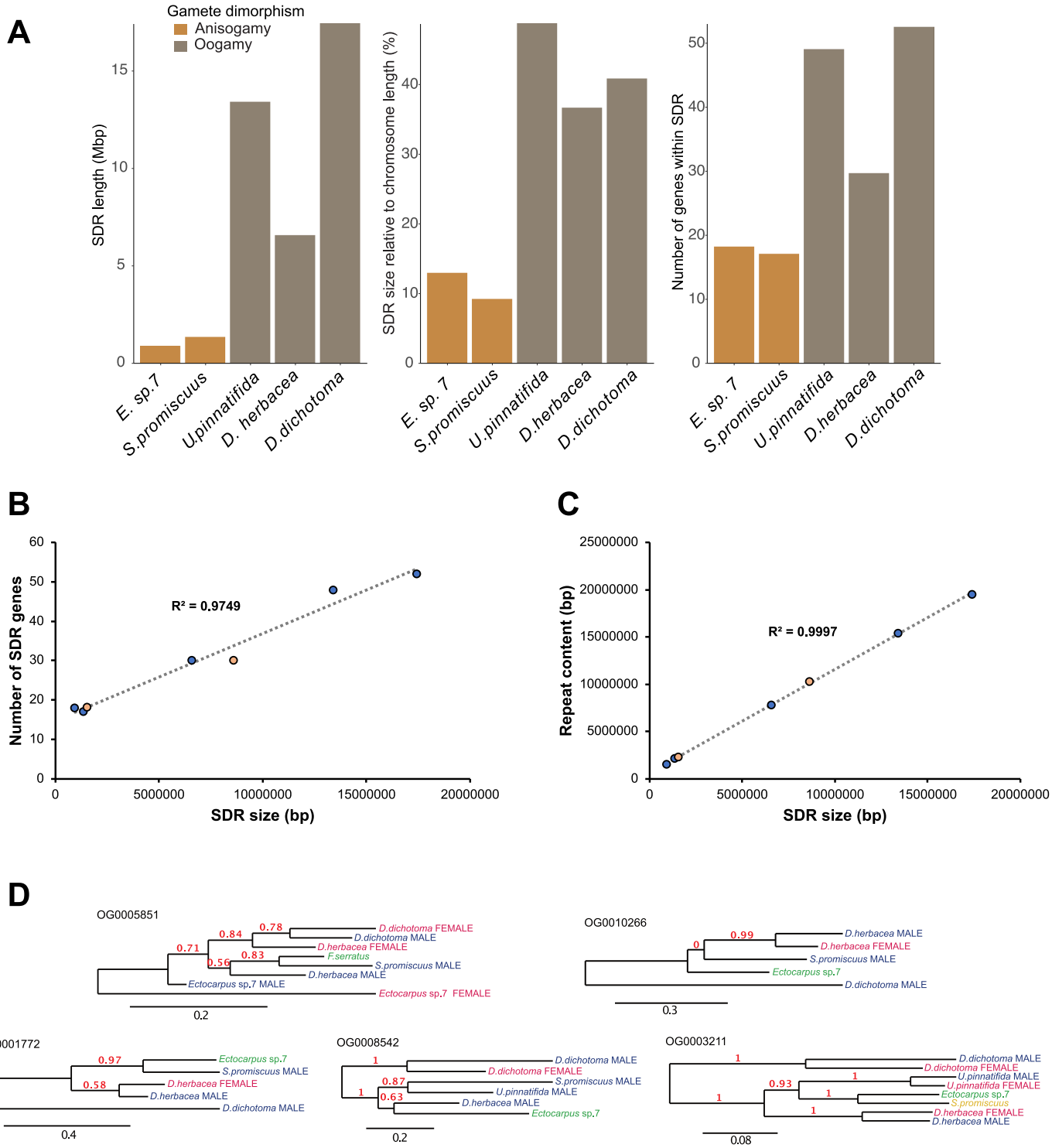
[1]Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, Tübingen, Germany. [2]Sorbonne Université, CNRS, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France. [3]INRAE, Université de Strasbourg, UMR SVQV, Colmar, France. [4]CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Roscoff, France. [5]Bezhin Rosko, Santeg, France. [6]Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway. [7]Universidade do Algarve, UALG·Centro de Ciências do Mar (CCMAR), Montenegro, Portugal. [8]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [9]Present address: Departamento de Biotecnología y Bioquímica, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Unidad Irapuato, Irapuato, Mexico. [10]Present address: Research Center for Inland Seas, Kobe University, Nadaku, Kobe, Japan. [11]These authors contributed equally: Josué Barrera-Redondo, Agnieszka P. Lipinska. [12]These authors jointly supervised this work: France Denoeud, J. Mark Cock, Fabian B. Haas, Susana M. Coelho.
✉e-mail: susana.coelho@tuebingen.mpg.de

https://doi.org/10.1038/s41559-025-02838-w



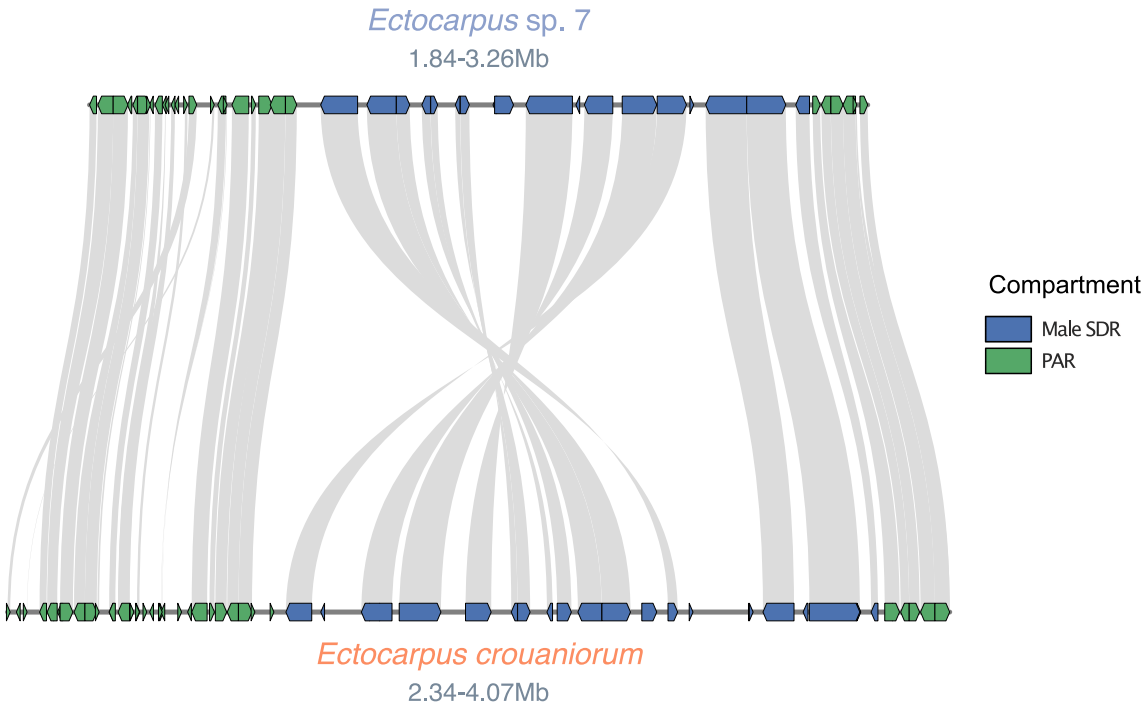
Extended Data Fig. 1 | Macrosynteny plot between *Schizocladia ischiensis* and *Dictyota dichotoma* using 1,828 orthologs. We highlight two fusion-with-mixing events (red squares) between chromosomes 4 and 9, and between chromosomes 23 and 24 in *D. dichotoma*.

https://doi.org/10.1038/s41559-025-02838-w



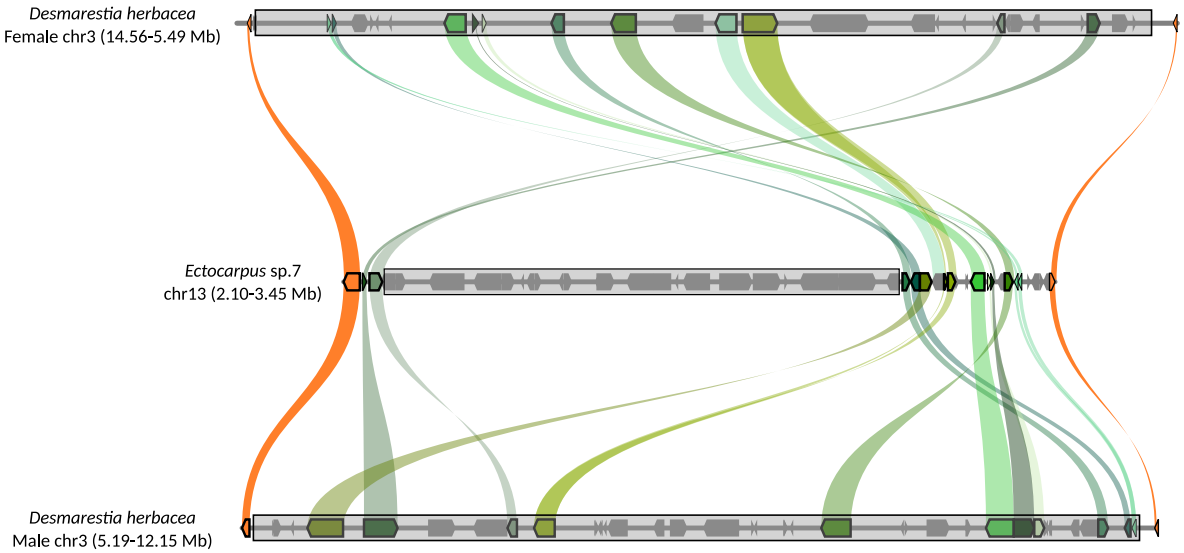
Extended Data Fig. 2 | SDR size differences and detection of independently-acquired V-SDR genes across species. (**a**) Differences in the size of the male SDR between brown algal species based on the total sequence length, the relative size of the SDR compared to the length of the V chromosome and the number of protein-coding genes retained within the SDR. The bars are colored accord-ing to the level of gamete dimorphism in each species (based on 16). (**b**) Correlation between the V (blue) and U (pink) SDR sizes and the SDR gene content across species. (**c**) Correlation between the V (blue) and U (pink) SDR sizes and the SDR repeat content across species. (**d**) Gene trees showing the independent acquisition of SDR gametologues across species that were previously interpreted as part of the ancestral male SDR genes.

https://doi.org/10.1038/s41559-025-02838-w



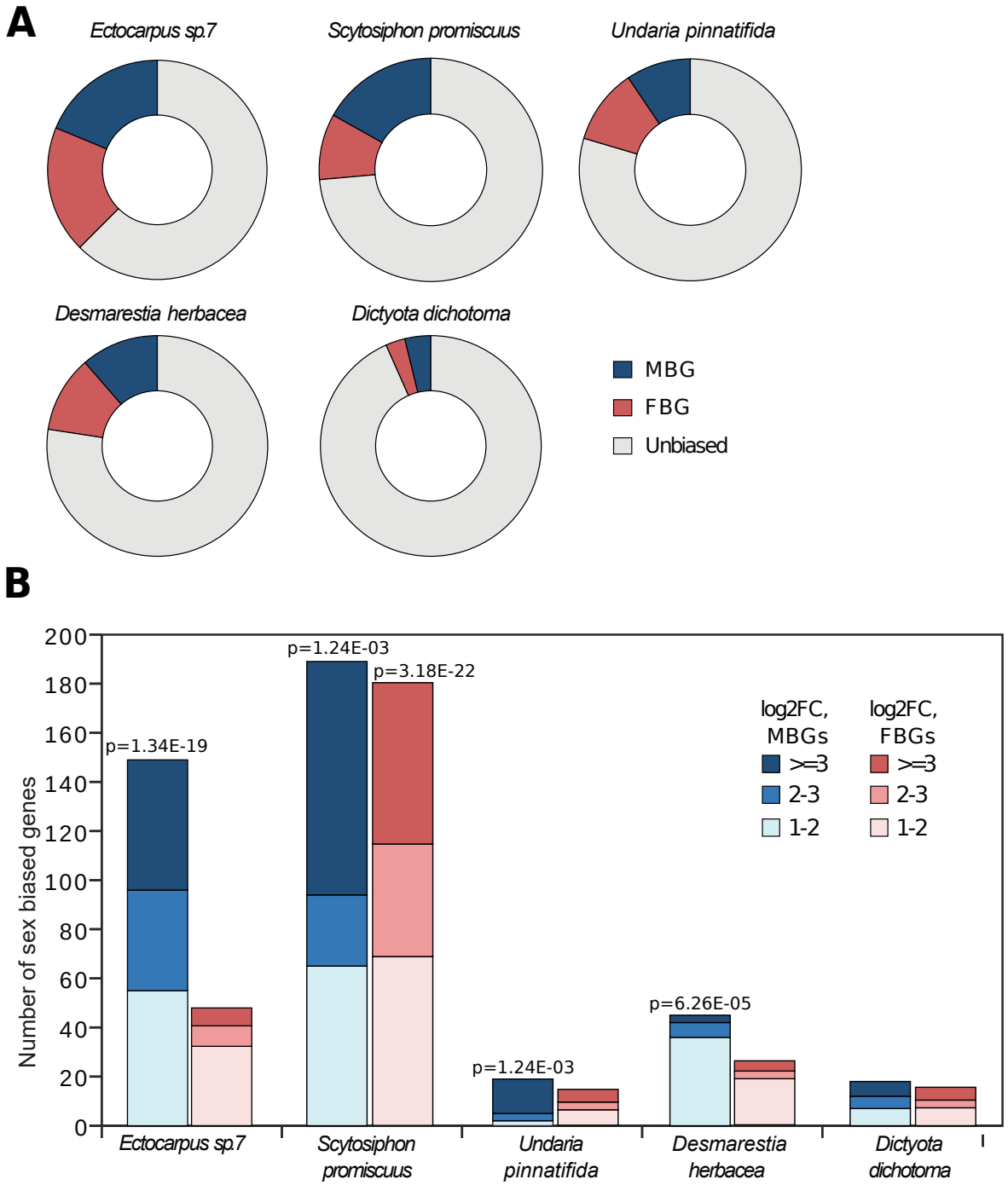
Extended Data Fig. 3 | Male SDR synteny between *Ectocarpus* sp. 7 and *Ectocarpus crouaniorum*. One of the species underwent a recent inversion event within the SDR. The arrows in the boxes represent the orientation of each gene within the chromosome.

https://doi.org/10.1038/s41559-025-02838-w



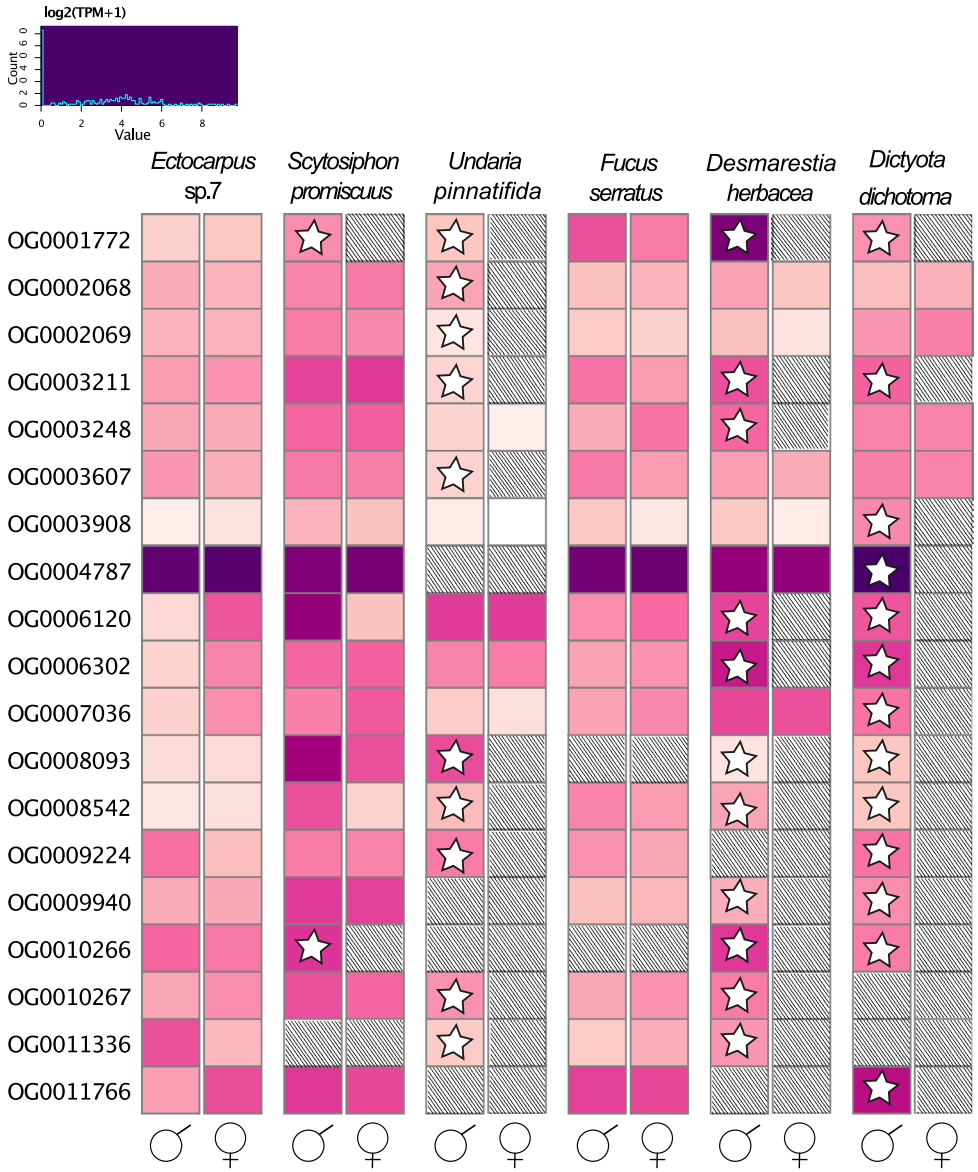
Extended Data Fig. 4 | Synteny analysis plot illustrating the expansion of the *Desmarestia herbacea* U- and V-sex-determining regions (SDRs) into the surrounding pseudoautosomal region (PAR). The *Ectocarpus* sp. 7 V-chromosome is shown in the middle as a reference, with the SDR regions outlined by grey boxes. Green lines trace syntenic relationships between *Ectocarpus* PAR genes and the recently-acquired *Desmarestia* SDR genes, with each gene pair represented in a distinct shade of green. This demonstrates that nearly all PAR genes from *Ectocarpus*, which have entered the expanded SDR in *Desmarestia*, are retained as gametologues. Orange lines highlight the PAR boundary genes in *Desmarestia*, which remain within the PAR of *Ectocarpus*.

https://doi.org/10.1038/s41559-025-02838-w



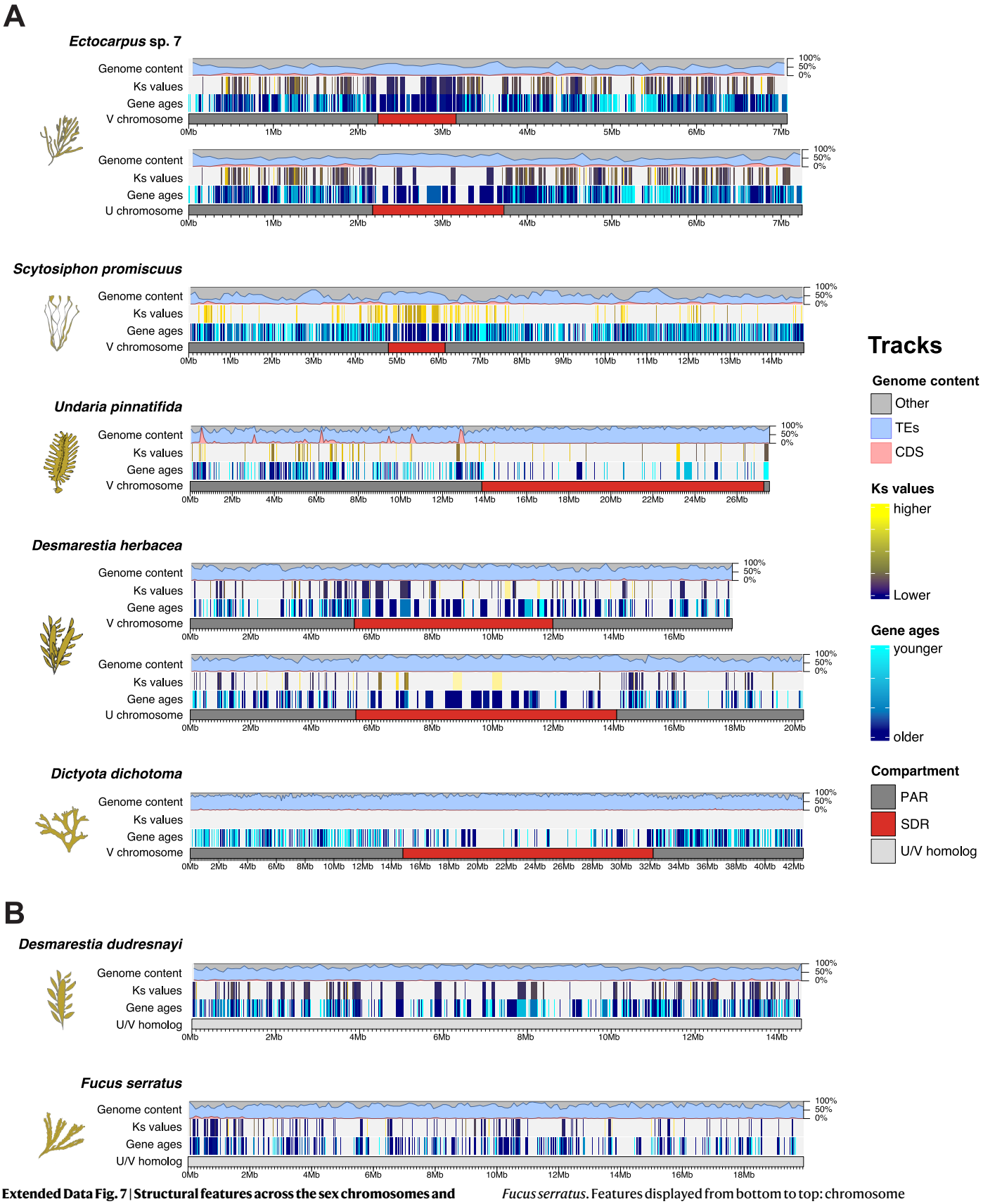
Extended Data Fig. 5 | Sex-biased gene expression per dioicous species. **a**) Proportion of sex biased genes in each of the five dioicous species. MBG: male-biased genes; FBG: female biased genes. (**b**) Number of sex-biased genes in the pseudoautosomal regions of sex chromosomes (U-V-SDRs excluded), male-biased genes are shown in blue and female-biased genes in red. Exact p-values above the bars mark significant enrichment of the sex-biased genes on the PAR (Chi-square test).

https://doi.org/10.1038/s41559-025-02838-w



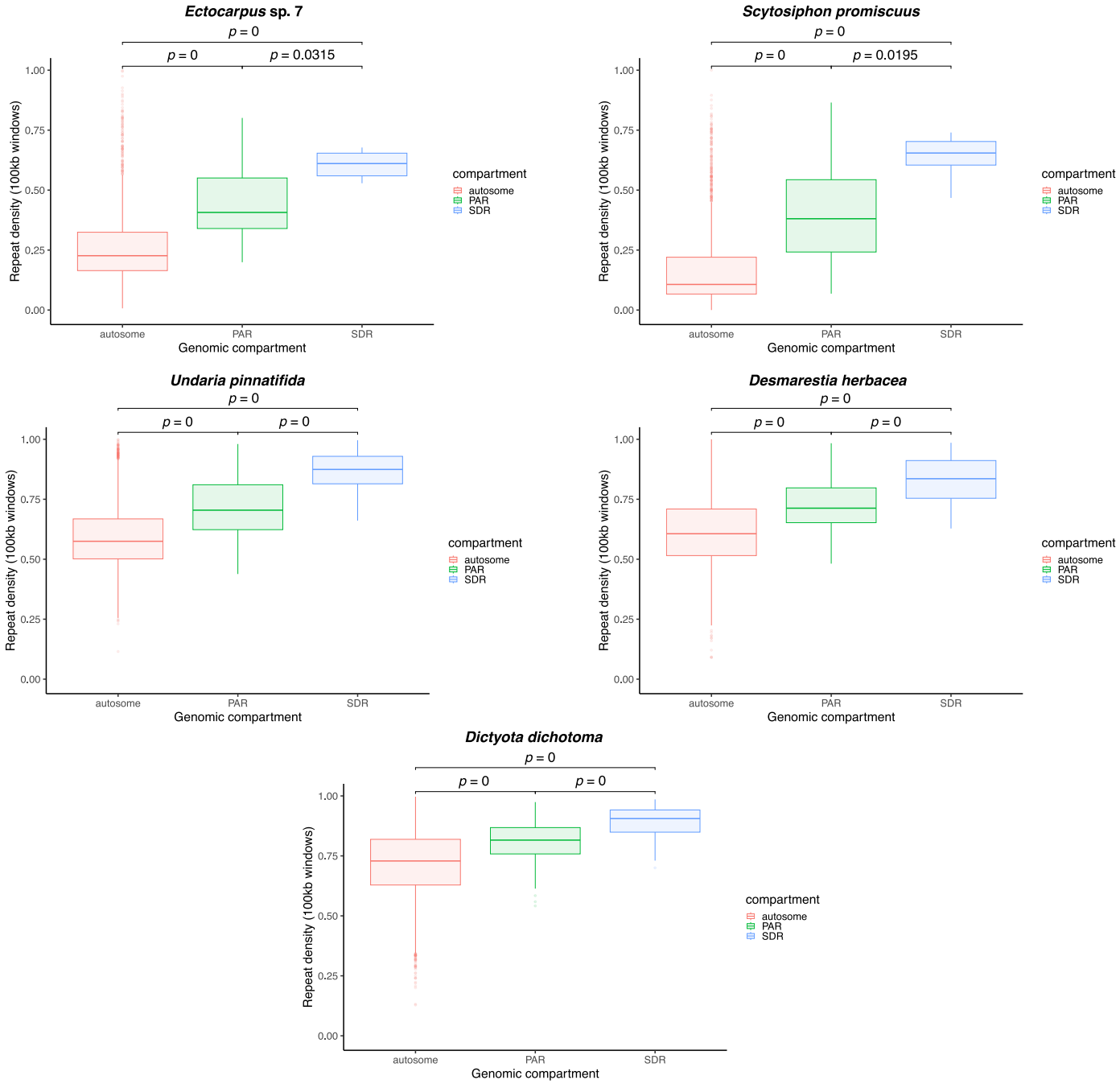
Extended Data Fig. 6 | Expression of genes (log2(TPM + 1)) that entered the SDR independently in different species. Expression is measured in mature male and female gametophytes, hashing marks missing orthologues, stars inside the cells indicate that the gene is inside the male non-recombining region (V-SDR). Orthogroups containing orthologues in less than three species or with multicopy genes were excluded from this analysis. M: male; F: female.

https://doi.org/10.1038/s41559-025-02838-w



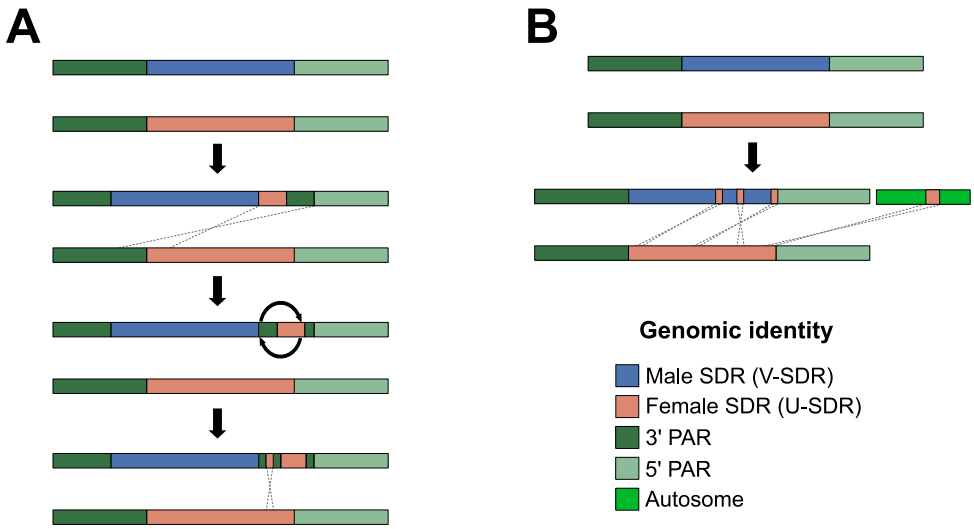
Extended Data Fig. 7 | Structural features across the sex chromosomes and U/V-homologues of brown algae. (**a**) V and U chromosomes of *Ectocarpus* sp. 7, *Scytosiphon promiscuus*, *Undaria pinnatifida*, *Desmarestia herbacea* and *Dictyota dichotoma*. (**b**) U/V-homologues of *Desmarestia dudresnayi* and *Fucus serratus*. Features displayed from bottom to top: chromosome compartments (PARs, SDR, U/V-homologue); relative gene ages, inter-species *Ks* values, and proportion of coding (CDS, red) and repeat (TEs, blue) density.

https://doi.org/10.1038/s41559-025-02838-w



Extended Data Fig. 8 | Accumulation of repetitive elements in the V-SDRs and PARs of five dioicous species. Statistically significant differences in median values of repeat density (center line) were assessed using FDR-corrected permutation tests.

https://doi.org/10.1038/s41559-025-02838-w



Extended Data Fig. 9 | Proposed scenarios for the transition from dioicy to monoicy in *Chordaria linearis* and *Desmarestia dudresnayi*. (**a**) The ancestor of *Chordaria linearis* likely underwent an initial translocation event from the U chromosome to the V chromosome, inserting part of the U-SDR and a piece of the 3′ PAR towards the 5′ end of the V-SDR potentially through an ectopic recombination event. A subsequent inversion within this translocation spread the 3′PAR genes to both sides of the U-SDR insertion. Finally, a second translocation led to the insertion of an additional piece of the U-SDR within the 3′ PAR translocation. (**b**) The ancestor of *Desmarestia dudresnayi* underwent three translocations of U-SDR genes into the V-SDR. Additionally, a fourth translocation event happened between the U-SDR and an autosome (chr_04).

https://doi.org/10.1038/s41559-025-02838-w

Extended Data Table 1 | General characteristics of the genomes and sex chromosomes in representative brown algal species and an outgroup (*Schizocladia ischiensis*)

Sexual system	Dioicous (V)						Dioicous (U)		Dioecious	Monoicous		Unknown
Species	*E. sp. 7*	*E.crou*	*S.prom*	*U. pinn*	*D. herb*	*D. dich*	*E. sp. 7*	*D. herb*	*F. serr*	*C. lin*	*D. dud*	*S. isch*
Genome size (Mbp)	200.170	218.478	193.199	511.280	430.876	851.153	197.371	484.711	1 091.539	214.613	425.034	194.512
Chrom-level scaffolds	27	27	27	30	32	31	27	32	33	NA	32	NA
Total n. scaffolds	33	175	451	114	935	1467	46	1756	3,805	217	823	130
N50 (bp)	6 893 608	7 597 912	6 781 292	16 510 065	13 135 153	26 426 000	6 914 137	12 441 740	17 863 197	2 249 057	13 227 088	2 524 267
Sex chrom. or U/V-homolog	chr_13	chr_13	chr_13	HiC_scaffold_23	chr_03	chr_02	chr_13 (Ec25_SDR_F)	chr_03	LG15	C-linearis_contig12	chr_03	NA
Sex chrom. or U/V-homolog size (bp)	7 072 209	9 658 235	14 770 496	27 543 478	17 928 803	42 672 188	7 248 464	20 286 227	19 924 529	3 613 932	14 536 050	NA
SDR length (bp)	923 344	1 084 112	1 363 928	13 409 094	6 568 004	17 412 526	1 551 053	8 626 479	NA	NA	NA	NA
N. genes in SDR	18	18	17	48	30 (+ 20 viral-derived genes)	52	18	30	NA	NA	NA	NA
N. genes PAR	421	519	904	306	229	451	421	229	NA	NA	NA	NA
Gamy	anisogamous	anisogamous	anisogamous	oogamous	oogamous	oogamous	anisogamous	oogamous	NA	NA	NA	NA
CDS density SDR	2.96%	NA	2.31%	0.51%	0.99%	0.28%	1.75%	0.81%	NA	NA	NA	NA
CDS density PAR or U/V-homolog	8.50%	NA	6.69%	7.35%	1.98%	1.70%	8.90%	1.98%	2.07%	NA	2.83%	NA
CDS density autosomes	14.97%	NA	15.94%	13.51%	5.17%	2.95%	14.97%	5.17%	2.28%	NA	9.27%	NA
Repeat density SDR	60.64%	NA	64.29%	86.94%	83.82%	89.19%	68.88%	84.53%	NA	NA	NA	NA
Repeat density PAR or U/V homolog	44.31%	NA	40.19%	72.51%	72.92%	80.88%	41.46%	71.76%	74.26%	NA	74.68%	NA
Repeat density autosomes	30.12%	NA	18.92%	59.24%	61.53%	72.10%	29.38%	60.97%	75.16%	NA	57.63%	NA
Genome source	Liu et al.[18]	This study	This study	Shan et al.[19]	This study	This study	Liu et al.[18]	This study	This study	Denoeud et al.[17]	This study	Denoeud et al.[17]

Note: Note that *Fucus serratus* (F. serr) has a chromosome (LG15) that is homologous to the ancestral U/V sex chromosome because it contains several genes present in the ancestral U/V-SDR. However, none of these genes are sex-linked in *F. serratus* (see text for details). Note that the dioecious species *F. serratus* has male and female (diploid) sexes and no gametophyte generation, that is, an animal-like life cycle. The CDS and repeat densities of the U/V-homologues were placed in the same rows as the PAR values of the dioicous species.

nature portfolio

Corresponding author(s): Susana Coelho

Last updated by author(s): 22/06/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	details are provided in the method section of the manuscript
Data analysis	details on data analysis are provided in the methods section of the manuscript

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data and accession codes are provided in the supplemental tables of the manuscript. There are no restrictions on data availability

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We identified the sex chromosomes and sex determining regions of several brown algal species represenitng the full phylogeny and level of morphological complexity across the lineage.
Research sample	brown algal lines
Sampling strategy	samples for each brown algal species used were grown in the laboratory in separate petri dishes as haploids. Sampeles were obtained from RCC (culture collection of Roscoff) and all details of codes and origin are provided in the manuscript.
Data collection	na
Timing and spatial scale	na
Data exclusions	no data was excluded
Reproducibility	for expression analysis the experiments were performed at least in triplicate
Randomization	samples were grown in the culture room in a randomized fashion.
Blinding	n.a

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

Genome Biology

RESEARCH

Open Access

Characterization of the transposable element landscape shaping the *Ectocarpus* genome

Erica Dinatale[1], Rory J. Craig[1,5†], Claudia Martinho[1,2†], Hajk-Georg Drost[3,4*] and Susana M. Coelho[1*]

†Rory J. Craig and Claudia Martinho jointly supervised E.D.

*Correspondence: hdrost001@dundee.ac.uk; susana.coelho@tuebingen.mpg.de

[1] Department of Algal Development and Evolution, Max Planck Institute for Biology, Tübingen, Germany
[2] Present Address: School of Life Sciences, Division of Plant Sciences, University of Dundee, James Hutton Institute, Invergowrie, Dundee, Scotland, UK DD25DA
[3] Computational Biology Group, Max Planck Institute for Biology, Tübingen, Germany
[4] Digital Biology Group, Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK
[5] School of BioSciences, University of Melbourne, Parkville, Australia VIC 3010

Abstract

Background: Comprising up to 90% of eukaryotic genomes, transposable elements (TEs) are mobile genetic units that play fundamental roles in evolution. Brown algae, one of the most complex multicellular eukaryotic groups that evolved independently from plants, fungi, and animals, are particularly underexplored in their transposon biology, especially when studied in a developmental context.

Results: Here, we explore the TE landscape of the model brown alga *Ectocarpus*, using a high-quality genome assembly complemented by extensive manual curation. TEs account for 28% of the genome, with a predominance of evolutionarily young elements. DNA transposons represent the most abundant and diverse TE subclass. Notably, TEs are significantly enriched along the sex chromosomes, a pattern potentially driven by local transposition events from the non-recombining sex-determining region into the pseudoautosomal regions. The genome harbors a high density of intronic TEs, which show minimal impact on host gene expression; however, intronic TEs tend to be shorter and more degraded than intergenic copies, suggesting selective pressures on their retention in the genome. Intact and potentially active TEs are preferentially associated with small RNAs and the histone modification H3K79me2, with over 70% of H3K79me2-marked intact TEs also enriched in small RNAs. This stable association indicates tight and sustained silencing of intact TEs throughout the life cycle of Ectocarpus.

Conclusions: Our study highlights the genetic diversity of the *Ectocarpus* mobilome and presents a complex, multilayered landscape of TE regulation mechanisms which involves small RNAs and chromatin modifications in the absence of an epigenetic silencing machinery that would be comparable to animals or plants.

Background

Brown algae are one of the most complex multicellular lineages on Earth, only distantly related to animals, land plants, and fungi, from which they diverged over 1 billion years ago [1, 2]. They have colonized various marine habitats and evolved diverse morphologies,

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

developmental patterns, life cycles, and sex determination systems [3, 4]. *Ectocarpus* is a filamentous brown alga that has emerged as a model organism representing this group of eukaryotes [5, 6]. *Ectocarpus* undergoes a haploid-diploid life cycle, alternating between haploid gametophyte, in which sex is determined by the presence of a male (V) or female (U) sex chromosome, and diploid sporophyte generations. Both generations are multicellular and relatively simple in terms of development [7]. Although significant progress has been made in understanding development, sex determination, and gene regulation in the *Ectocarpus* genome (reviewed in [7, 8]), relatively little is known about mobile genetic elements in this species. This lack of knowledge extends across brown algae, hindering our understanding of how mobile elements have influenced the evolution of this major multicellular lineage.

Transposable elements (TEs) are mobile genetic units that display substantial mechanistic diversity and are near-ubiquitous across the tree of life [9]. TEs can copy or excise their sequence and insert at new genomic locations, with this replicative nature frequently leading to their accumulation in genomes. Both the proportion and diversity of TEs in a host genome can vary dramatically, even between closely related species [10, 11]. TE abundance also differs between and within chromosomes, and the genomic distribution of different TEs is influenced by multiple forces, including the strength of selection against TE insertions, the local recombination landscape, and TE insertion biases [12, 13]. In brown algae, estimated repeat contents range from ~ 20% in *Ectocarpus* to 70% (e.g., the Fucales and Dictyotales), and variation in the abundance of repeats largely explains differences in genome size [2]. TEs are enriched on the U and V sex chromosomes, especially in the sex-determining region (SDR) [4, 14]. Yet, in contrast to many animal, plant and fungal systems, most TE families in the *Ectocarpus* genome, and consequently across brown algae, remain unclassified and largely uncharacterized.

TEs have long been considered selfish genetic elements and threats to genome stability [15, 16], or as benign "junk" DNA that accumulates in genomes as degraded and inactive copies [9, 17]. However, in a return to the foundational hypotheses of transposon biology [18], it is now clear that TEs play many fundamental roles in evolution. For example, TE insertions can induce phenotypic change and adaptation, and TE co-option acts as a major driving force of gene regulatory element evolution [10, 19]. Nevertheless, TEs are believed to be inherently mutagenic and eukaryotes have evolved elaborate mechanisms to prevent or control the proliferation of TEs.

Broadly, TE silencing mechanisms include transcriptional and post-transcriptional inhibition of the TE sequences and genes required for their replication and mobilization. These mechanisms frequently involve repressive epigenetic modifications and small RNA (sRNA) molecules with high sequence complementarity to a TE target region, which can function either independently or in concert. In flowering plants, sRNAs are involved in the establishment of repressive DNA and chromatin modifications at TE loci, and also direct small interfering RNA-mediated cleavage of TE mRNA [20, 21]. In animals, TE transcripts can also be the target of small interfering RNAs (siRNAs), while PIWI-interacting RNAs (piRNAs) repress both TE transcripts and genomic loci [22]. In ciliates, sRNAs guide the removal of TEs from the macronucleus [23]. Repressive modifications include DNA methylation (especially 5' methylcytosine; 5mC) and histone post-translational modifications that reduce chromatin accessibility. In particular, in plants,

animals and some ciliate species, trimethylation of histone H3 lysine 27 (H3K27me3) and mono-, di- and trimethylation of histone H3 lysine 9 (H3K9me1/2/3) are associated with transcriptional suppression and TE silencing [24–26].

How TEs are silenced in brown algae is largely unknown. *Ectocarpus* displays negligible levels of 5mC, both genome-wide and on TEs [27]. Furthermore, classical repressive histone marks, including H3K27me3 and H3K9me1/2/3, are absent in the species. The histone marks H3K79me2 and H4K20me3 have been associated with decreased transcript abundance [28–30] and repeats [31]. In particular, 88.6% of H3K79me2 regions of > 5 kbp included a ≥ 400-bp repeated element, suggesting a potential role for this mark in TE control [31]. Furthermore, *Ectocarpus* sRNAs show an association with TEs [27]. sRNAs exhibit a predominance for 21-nt species, including microRNA families [32].

Here, we combine manual TE curation with a recent chromosome-level genome assembly [33] to comprehensively characterize the *Ectocarpus* mobilome and investigate TE regulation across its life cycle. We show that the majority of TEs in the *Ectocarpus* genome are evolutionarily young, with diverse DNA transposons representing the most abundant subclass and contributing significantly to repeat enrichment on the sex chromosome. Using integrated ChIP-seq, sRNA-seq, and RNA-seq datasets, we further demonstrate that intact and potentially active TEs are preferentially associated with putative silencing mechanisms, suggesting a tightly regulated control of TE activity throughout development.

Results

The Ectocarpus genome features a high diversity of recently active TEs

To characterize the genomic landscape and diversity of TEs in *Ectocarpus* sp. 7 (hereafter *Ectocarpus*), we used the latest high-quality, near-gapless genome assembly, which is based on a haploid male individual (strain Ec32 [33],) and spans 200.2 Mb at telomere-to-telomere level for many chromosomes [33]. This highly contiguous reference genome provides the opportunity to characterize TE dynamics at unprecedented resolution in a brown alga.

An existing *Ectocarpus* TE library features 897 consensus sequences [27] and was generated using the automated repeat annotation tool REPET [34]. Following a stringent post-processing and quality control procedure for these 897 sequences (see Methods), we detected high levels of redundancy, mis-classification, and lack of classification (i.e., unknown repeats). Repeat classification is particularly challenging in taxonomically diverse species, since the process largely relies on homology to TEs and TE proteins from species with curated TE libraries [35]. To overcome these limitations, we produced a de novo TE library complemented by a manual curation procedure, primarily targeting the most abundant elements (i.e., TE families contributing at least ~ 100 kb to the genome) (Methods). This annotation and curation effort yielded 165 verified consensus sequences, each of which served as a representative for a single autonomous or nonautonomous TE family (or subfamily, if relevant) (Supplementary Material 1 and 2). For the remaining low abundance TE models from the REPET library, we performed sequence clustering analysis to remove redundancy and performed additional re-classification via homology comparisons to the manually curated models (see Methods). The final library was produced by combining the manual TE models with the non-redundant REPET

models, yielding 574 consensus sequences, corresponding to 347 families (Table 1, see Supplementary Material 3 for detailed annotation notes for TE families).

The *Ectocarpus* genome harbors considerable TE diversity. Most of the 347 TE families correspond to the major subclasses of eukaryotic TEs, which can broadly be grouped as either class I (moving via an RNA intermediate) or class II (a DNA intermediate) [36]. The *Ectocarpus* genome features all class I retrotransposon subclasses: LTR retrotransposons, long interspersed nuclear elements (LINEs), tyrosine recombinase retrotransposons (DIRS), and *Penelope*-like elements (PLEs) [9, 37]. For Class II, our curation revealed substantial diversity in DD(E/D) DNA transposons (i.e., canonical *cut*-and-*paste* DNA transposons, hereafter simply DNA transposons), in addition to Helitron elements that move via a *copy*-and-*paste* rolling circle mechanism [38]. We identified individual TE copies in the *Ectocarpus* genome by passing our library to RepeatMasker, and also identified satellite and microsatellite arrays using Tandem Repeats Finder. Overall, repeats constitute 31% of the *Ectocarpus* genome, the vast majority of which are TEs (28% of the genome, i.e., 55.2 Mb). DNA transposons account for 7.9%, LTR retrotransposons 6.3%, and LINEs 4.9% (Fig. 1A, Table 1). The TE landscape is largely comprised of copies with high sequence similarity within a given family, with 53% of the total genomic TE content attributable to TE copies exhibiting < 5% divergence from their

Table 1 Summary statistics of TE superfamilies in the *Ectocarpus* genome

Category	Families	Copies	Intact copies	Genome coverage (%)
LINE/*CRE*	2	4821	30	0.42%
LINE/*RTE*	3	5010	188	0.97%
LINE/*RTE-X*	7	16,418	4100	3.41%
LINE/Unknown	3	641	16	0.14%
LTR/*Copia*	35	5023	561	3.00%
LTR/*Gypsy*	32	4221	196	2.78%
LTR/Unknown	26	1144	198	0.47%
DIRS/*Ngaro*	15	3893	689	2.45%
DIRS/*PAT-like*	5	842	63	0.58%
PLE/*Chlamys*	3	4782	27	0.43%
DNA/Unknown	51	9109	1308	1.70%
DNA/*Dada*	1	15	1	0.01%
DNA/*Harbinger*	15	2251	916	0.93%
DNA/*KDZ*	8	6403	148	1.80%
DNA/*Mariner-Tc1*	17	11,912	3007	1.43%
DNA/*MuDR*	1	56	42	0.13%
DNA/*PiggyBac*	4	1588	289	0.48%
DNA/*Sola-1*	1	1989	2	0.17%
DNA/*Sola-2*	1	48	14	0.04%
DNA/*hAT*	4	128	13	0.05%
DNA/*EnSpm*	5	6366	161	1.11%
Helitron	13	4688	522	1.10%
Unknown	94	22,127	4037	3.93%
Tandem repeats	/	/	/	3.35%
Total	393	113,475	16,528	30.91%

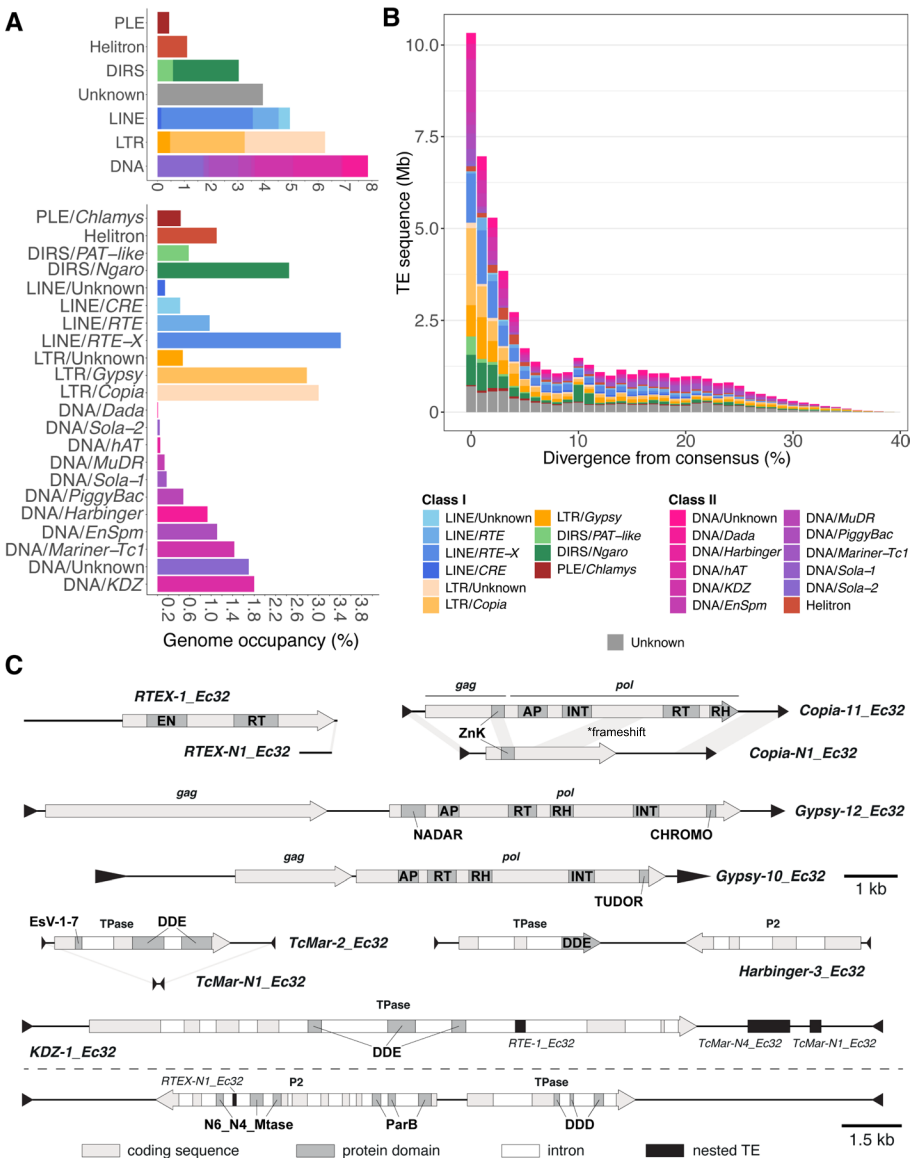


Fig. 1 Genomic occupancy and landscape of 113,475 annotated TEs (574 consensus sequences) in *Ectocarpus*. **A** Percentage of the genome spanned by different subclasses and superfamilies of TEs. DNA refers to DD(E/D) transposons. **B** TE landscape showing cumulative genomic TE content relative to divergence from TE family consensus sequence (a proxy for TE age). **C** Schematics of *Ectocarpus* TEs showing terminal motifs, genes, and protein domains. Models are to scale except for the terminal inverted repeats of DNA transposons which are enlarged for visualization. Domain abbreviations are: EN = endonuclease, RT = reverse transcriptase, INT = integrase, RH = RNaseH, ZnK = Gag zinc knuckle, AP = aspartic protease, DDE = DD(E/D)-type integrase

family consensus sequence (Fig. 1B). This broad lack of degradation suggests that most TE families in the *Ectocarpus* genome may have been recently active, and that many TE copies may have retained the competency to transpose.

At the level of individual families, the most abundant TEs are LINEs and non-autonomous elements that putatively rely on autonomous LINEs for their transposition (Supplementary Material 4). Specifically, multiple families from the *RTE-X*, *RTE*, and *CRE*

clades are each present in several thousand individual copies (Supplementary Material 4). The most abundant single family, present in more than 6000 copies and totaling 2.48 Mb of the genome, is a ~ 600 bp non-autonomous element (*RTEX-N1_Ec32*) that exhibits sequence conservation at its 3' end to autonomous *RTE-X* elements (*RTEX-1* and *RTEX-2_Ec32*, Fig. 1C). It is possible that *RTEX-N1_Ec32* is a SINE (short interspersed nuclear element); however, we were unable to determine the origin of its 5' region. For simplicity, we group this family and all other similar families with LINEs, since some classification systems require SINEs to have RNA polymerase III promoters [39].

Although less abundant as individual families, LTR retrotransposons exhibit a greater family-level diversity than LINEs and collectively constitute more sequence in the *Ectocarpus* genome (Fig. 1A, B). All *Ectocarpus* LTR retrotransposons belong to the *Metaviridae* (i.e., *Ty3/Gypsy*, 32 families) and *Pseudoviridae* (i.e., *Ty1/Copia*, 35 families) groups. As previously reported in *Ectocarpus* [27] and some plant genomes [40], we found several *Copia* families that carry a single ORF with at least partial homology to the *gag* ORF of fully autonomous elements (e.g., *Copia-N1_Ec32* and *Copia-11_Ec32*, Fig. 1C). *Gypsy* families from the *Ectocarpus* genome display considerable structural diversity, and phylogenetic analysis revealed four distinct lineages distributed among the diversity of eukaryotic *Gypsy* elements (Suplementary Material 5: Fig. S1). Notably, families from all four lineages encode polyproteins featuring chromatin reader domains that are associated with chromatin recognition, remodeling, and targeted insertion. Families from three of the lineages feature putative C-terminal chromodomains fused to the integrase domain of Pol (e.g., *Gypsy-12_Ec32*, Fig. 1C). One of these lineages corresponds to *ECR-1*, which we previously described as an *Ectocarpus* centromere-specific element [33]. Interestingly, the *Ectocarpus* chromodomain-containing *Gypsy* lineages are not phylogenetically related to chromoviruses (the major group of chromodomain-containing *Gypsy* elements [41]) nor the Chronos chromodomain-containing *Gypsy* elements of oomycetes [42] (Supplementary Material 7: Fig. S1). The final lineage corresponds to an element carrying a C-terminal Tudor domain (*Gypsy-10_Ec32*, Fig. 1C), which together with chromodomains form part of a larger family of histone readers [43]. To our knowledge, this is the first reported TE that carries a Tudor domain. We recovered a Tudor domain-containing homolog of *Gypsy-10_Ec32* in the yellow-green alga *Tribonema minus*, suggesting acquisition of this domain prior to the origin of brown algae more than 450 mya [44]. Collectively, our results support multiple independent acquisitions of chromatin reader domains by *Gypsy* elements in stramenopiles, potentially indicating unique interactions between retroelements and the host chromatin-landscape in this evolutionary lineage.

Furthermore, the families from one of the four *Gypsy* lineages all feature a highly conserved YbiA-family NADAR domain at the N-terminal region of Pol (e.g. *Gypsy-12_Ec32*, Fig. 1C; Supplementary Material 7: Fig. S2). YbiA-family NADAR (NAD and ADP-ribose) domain proteins are present in prokaryotes, eukaryotes, and viruses and play diverse roles in enzymatic pathways involving NAD (nicotinamide adenine dinucleotide) and ADP-ribose, including processing of nucleic acids and proteins [45]. Excluding *Ectocarpus*, blastp queries of the *Gypsy* NADAR domains almost exclusively retrieved bacterial hits, suggesting possible acquisition of the domain by horizontal

transfer (Supplementary Material 8: Table S1). Querying the Repbase database of TE proteins recovered divergent hits to the ORF1 proteins from several families of *Tx1* LINEs from the Pacific Oyster *Crassostrea gigas*, which also appear to constitute YbiA-family NADAR domains (Supplementary Material 7: Fig. S2). Although their role in retrotransposition is unclear, the presence of YbiA-family NADAR domains in RNA viruses has been hypothesized to function in RNA processing [45] and it may be the case that the NADAR domains are involved in processing the RNA intermediates of transposition in these *Ectocarpus* and *C. gigas* elements.

The remaining retrotransposons belong to the *Ngaro* and *PAT-like* groups of DIRS elements, and the *Chlamys* group of PLEs. *Ngaro* elements are particularly diverse (15 families) and abundant (2.45% of the genome) (Fig. 1A). PLEs have not previously been reported in TSAR (telonemids-stramenopiles-alveolates-rhizaria) [46], and their discovery in *Ectocarpus* extends their known range to a new eukaryotic supergroup.

DNA transposons are both the most abundant and most diverse subclass of TEs in *Ectocarpus*, with at least 10 superfamilies present (Fig. 1A, B). *Harbinger* (15 families) and *Mariner-Tc1* (17 families) are the most diverse DNA transposons at the family level, and *Mariner-Tc1* DNA transposons collectively contribute almost 12,000 genomic copies (Table 1), most of which are non-autonomous (i.e., miniature inverted-repeat transposable elements; e.g., the 217 bp *TcMar-N1_Ec32*, which relies on the 4.4 kb autonomous *TcMar-2_Ec32* for its transposition and exhibits 27-fold higher copy number; Fig. 1C, Supplementary Material 4). The transposase of *TcMar-2_Ec32* has acquired an N-terminal EsV-1–7 domain (Fig. 1C), which may be an uncharacterized class of zinc finger domain that has proliferated in the genomes of brown algae and are otherwise mostly known from *Nucleocytoviricota* giant viruses [47]. Although all of the autonomous *Harbinger* elements encode two tail-to-tail genes, they can be divided to two groups: elements that target ANT trinucleotides (where N is any nucleotide) and cause corresponding ANT target site duplications (TSDs; previously reported for the *Pangu* subgroup [48] and those that match the classic TNA insertion/target site duplication pattern of *Harbinger* (e.g., *Harbinger-3_Ec32*, Fig. 1C).

A remarkable feature of the *Ectocarpus* mobilome is the presence of several giant DNA transposons from the superfamilies *KDZ*, *EnSpm*, and *Sola-1*, which generally span more than 10 kb, but can exceed 30 kb. In the case of *KDZ* and *EnSpm*, the combination of their considerable lengths and high copy numbers results in these families comprising substantial proportions of the genome (1.8% and 1.1%, respectively; Fig. 1A, Table 1). Most giant transposon families feature nested insertions of unrelated TEs (e.g., *KDZ-1_Ec32*, Fig. 1C). These insertions are generally fixed among multiple copies, implying that these giant elements have continued to transpose following insertion by other TEs. Many families carry additional transcribed genes, and frequently the transposase-encoding gene is predicted to be pseudogenized, implying mobilization in *trans* via copies that can produce an active transposase. Indeed, several giant families only carry genes that have no homology to transposases and are presumably entirely non-autonomous (Supplementary Material 3), analagous to the Pack-MULE transposons of plants [49]. While most of the accessory genes are of unknown function, we identified one intriguing gene transcribed on the opposite strand to the transposase in the 21.5 kb *Sola1-1_Ec32* family (Fig. 1C). The predicted product of this gene is a 1104 aa protein featuring two domains

typically associated with prokaryotes: a ParB-like DNA binding domain and a DNA methyltransferase domain predicted to produce either N4-methylcytosine or N6-methyladenine. Similar proteins have been reported in green and red algal transposons, and they may represent a novel class of transposon accessory proteins that are capable of binding and epigenetically modifying their own transposon sequence [50, 51].

Chromosomal distribution of TEs reveals repeat enrichment on the sex chromosomes

The *Ectocarpus* v5 haploid reference genome features 27 chromosomes ranging from 4.5 to 11 Mb, including the male sex chromosome (i.e., chromosome 13 or V) that harbors the male-specific sex determining region (SDR) [33]. Although not analyzed in the above section, Liu et al. [33] also produced a gapless assembly of the female SDR (chromosome U) via sequencing of a haploid female gametophyte, which is available as part of the v5 assembly. The SDRs of the sex chromosome harbor loci involved in male (V) and female (U) sex determination [52] and do not undergo recombination [4, 14, 53]. Consequently, the SDRs are more repetitive, less gene dense, and have larger introns and lower GC content relative to the autosomes [4, 14]. The pseudoautosomal regions (PARs) flanking the SDR exhibit intermediate values to those of the SDR and autosomes for the aforementioned genomic features, despite showing a recombination rate similar to the autosomes [14, 53, 54]. To explore the evolutionary dynamics of TE accumulation on autosomes and the sex chromosomes, we analyzed the repeat content across the different genomic compartments: autosomes, PARs, and male and female sex determining regions (V and U-SDR).

Overall, the distribution of TEs is not uniform, with total TE proportions of 22% on the autosomes, 39% on the PARs, 51% on the male SDR, and 65% on the female SDR (Supplementary Material 8: Table S2). The entire male sex chromosome harbors a significantly higher fraction of TE sequence compared to the autosomes (Kruskal–Wallis p-value $< 2.2e-16$; Dunn's test with Bonferroni correction shows p-value < 0.05 for all comparisons, except for $p = 0.0597$ in the comparison with chromosome 22; Fig. 2A, Supplementary Material 7: Fig. S3A), as previously reported [14]. Considering the SDR and PARs of the male sex chromosome separately, TEs are enriched in the SDR compared to the PARs (Fig. 2B, Wilcoxon rank sum test with continuity correction p-value $= 1.60e-16$), and the PARs are enriched compared to the autosomes (Wilcoxon rank sum test with continuity correction p-value $< 2.2e-16$).

As introduced above, the difference in genomic TE proportions between the PARs and autosomes is not expected given their similar recombination landscapes. Since DNA transposons are abundant in the *Ectocarpus* genome, we asked whether the mechanistic properties of DNA transposons could be responsible for the TE enrichment on the PARs. Unlike copy-and-paste elements, cut-and-paste DNA transposons exhibit biased insertion patterns termed "local hopping," in which new insertions occur more frequently in *cis*-linked sites in the vicinity of the donor locus [55, 56]. We hypothesized that DNA transposons that have accumulated in the recombination-suppressed SDR could act as a source of local hopping events, which may have resulted in an accumulation of DNA transposons on the nearby PARs, relative to the autosomes. To test this hypothesis, we scaled the abundance of DNA transposons relative to copy-and-paste TEs (retrotransposons and Helitrons) in different genomic regions. On the autosomes, we find a copy:cut

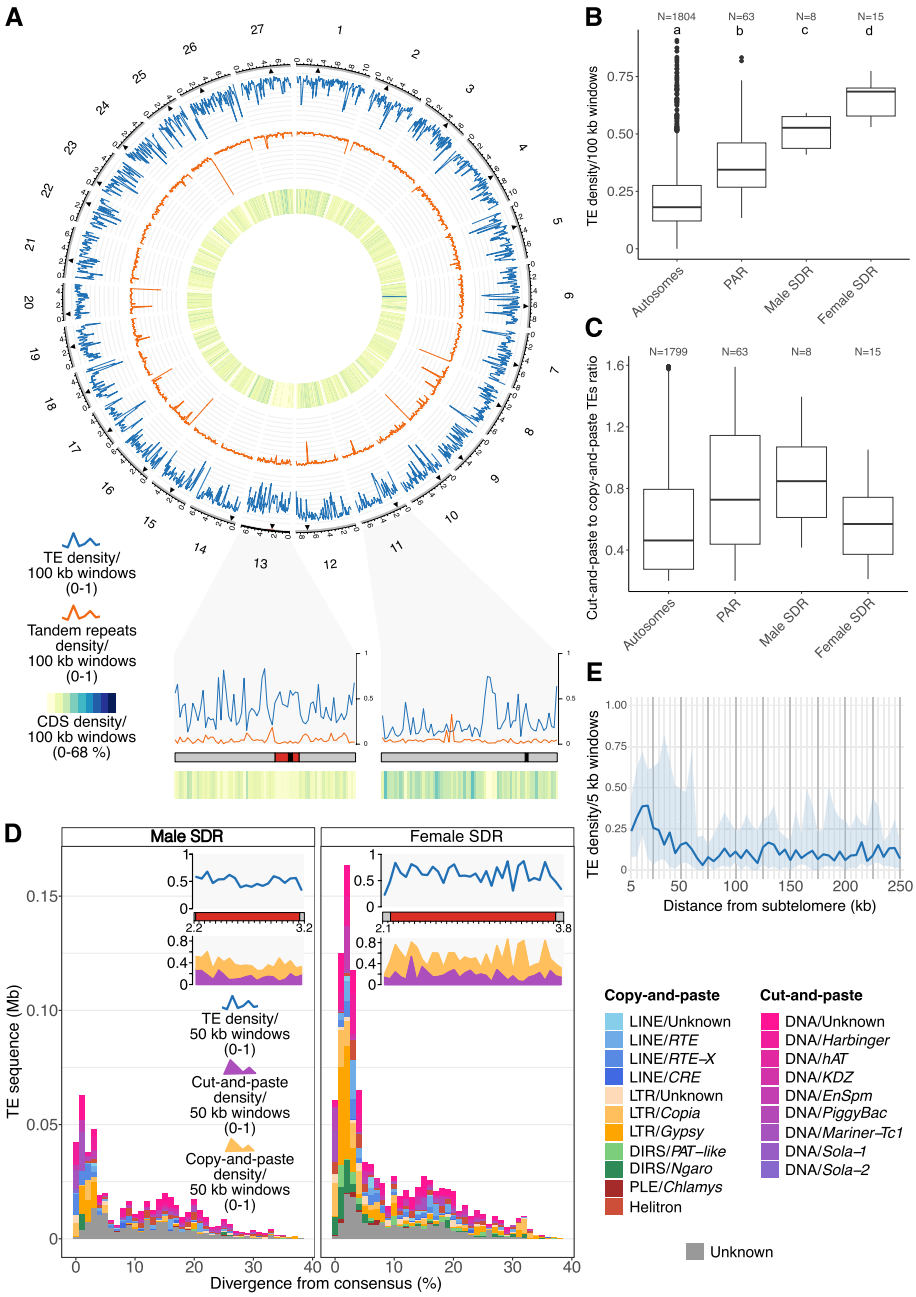


Fig. 2 **A** Circos plot representation of the *Ectocarpus* genome. Light gray outer blocks represent chromosomes, with the sex determining region (SDR) highlighted in red on the male sex chromosome (chromosome 13). Black triangles represent the centromeric regions. Tracks represent the TE and tandem repeat densities and the inner heatmap shows CDS density. Chromosome 11 and 13 are expanded as linear representations to contrast an autosome to the sex chromosome. **B** TE density in 100-kb windows across different genomic compartments. **C** Ratio of cut-and-paste to copy-and-paste TEs in 100-kb windows. **D** TE landscapes and TE densities of the male and female SDRs. Repeat densities were calculated in 50-kb windows. **E** TE density across 5-kb non-overlapping windows extending towards the centromere from the subtelomeric satellites; median density (blue line), first and third quantiles (light blue area)

abundance ratio of 0.48 (Fig. 2C, Supplementary Material 8: Table S2). In contrast, the PARs of the male sex chromosome exhibit a significantly higher ratio of DNA transposons to copy-and-paste elements (0.71) relative to the autosomes (Wilcoxon rank

sum test with continuity correction p-value = 2.60e − 04), as does the male SDR (0.79) (Fig. 2C). When examining TE superfamily densities per megabase, we observed that only a subset of DNA transposon superfamilies exhibit consistent enrichment on the sex chromosomes relative to the autosomes (Supplementary Material 5, Supplementary Material 7: Fig. S3B, S3C). Namely, *Tc1-Mariner*, in addition to the giant *KDZ* and *EnSpm* DNA transposons, were enriched ($\log_2$FC > 1) in both the SDR and PAR. Taken together, these observations support the possibility that local hopping of DNA transposons, and particularly giant elements, from the SDR has contributed to the elevated DNA TE content observed in the adjacent PARs.

We next examined the male and female SDRs in more detail (Fig. 2D; Supplementary Material 8: Table S3). The female SDR is approximately twice the size of the male SDR (Supplementary Material 8: Table S2) and harbors a significantly higher abundance of TEs, with 64.5% TE content compared to 50.9% in the male SDR (Wilcoxon rank sum test with continuity correction p-value = 1.42e − 03) (Fig. 2B, D). Surprisingly, the ratio of cut-and-paste to copy-and-paste elements on the female SDR is 0.52, not significantly higher (Wilcoxon rank sum test with continuity correction p-value = 1) than the autosomes (Fig. 2C). This observation can be explained by an increased abundance of LTR retrotransposons specifically in the female SDR; LTR retrotransposons comprise 16 and 8% of the female and male SDR, respectively, and *Gypsy* elements comprise almost 10% of the female SDR. While both SDRs share similar distributions of TE divergence values for DNA transposons, the female SDR features a higher abundance of low-divergence LTR retrotransposons (Fig. 2D). TEs on the female SDR are thus significantly younger than the ones on the male SDR (Wilcoxon rank sum test with continuity correction p-value = 8.698e − 06), consistent with a recent expansion of LTR retrotransposons that has specifically impacted the female SDR.

We also investigated the TE distribution along autosomes. The median TE density per autosomal 100-kb window is 18.9%, although there are many outlier genomic windows that exhibit substantially higher TE densities (Fig. 2B). Indeed, the autosomal landscape is punctuated by discrete regions that are TE-rich and gene-poor; for example, a ~ 300-kb region on chromosome 6 has a TE density of 64.7% and a coding sequence (CDS) density of only 6.4% (Fig. 2A). These TE-rich islands are independent of centromeres, which were recently characterized as short regions featuring the *ECR-1* centromere-specific *Gypsy* element [33]. The centromeres span on average only 39 kb, and this highly localized enrichment does not substantially contribute to variation in TE density at a chromosome-wide scale (Fig. 2A). Conversely, we did find an enrichment of TEs towards the chromosome termini. The *Ectocarpus* subtelomeric regions are defined by the presence of a specific satellite array [33], and highly elevated abundances of tandem repeats can be observed at several chromosome termini (Fig. 2A). Immediately upstream of these subtelomere-specific satellites, we observed a localized increase in TE density extending ~ 70 kb away from the subtelomere (Fig. 2E).

Finally, although the *Ectocarpus* v5 genome is near-complete, the assembly of a small number of chromosomes did not reach the subtelomere or telomere, and there are five unplaced scaffolds (13.75 Mb in total) that likely correspond to these incomplete chromosomal termini [33]. The scaffolds are massively repetitive (as much as 70% TE sequence, Supplementary Material 8: Table S4), and although they span only 7% of the

genome, the scaffolds constitute 18% of the total genomic TE content. These gene-poor, TE-rich regions likely correspond to large repetitive islands and putatively represent extreme cases of the TE-enrichment of chromosomal termini reported above.

Distribution and characteristics of TE insertions relative to genic sequence

The consequences of transposon insertions on the host genome often depends on the type of the element and on its insertion site [57]. In particular, TEs can impact the regulation of nearby genes via multiple complex interactions [10]. Therefore, we examined the local distance-distribution of TEs relative to genes focusing on the 27 assembled chromosomes representing the haploid male genome.

Intergenic regions contain the highest density of TEs at 52.9%, followed by 20.1% of intronic sequence, 7.0% of 5' UTRs, 8.8% of 3' UTRs, and 2.4% of CDS (Fig. 3A, Supplementary Material 8: Table S5). Conversely, tandem repeats exhibit a relatively even distribution across the genome (Fig. 3A, Supplementary Material 8: Table S5). The *Ectocarpus* genome is generally highly compact with regard to intergenic space, with a median intergenic length of 1302 bp on the autosomes, rising to 2853 bp on the male sex chromosome (Fig. 3B). To explore the difference between intergenic sequence immediately adjacent to genes and that far from genes, we divided intergenic regions into gene "proximal" (sequences within 500 bp of a gene) and "distal" (sequences > 500 bp from a gene) categories. E This 500 bp threshold from both flanks of neighboring genes captures the median intergenic tract length. Proximal intergenic regions exhibited a substantially lower TE density (17.8%) relative to distal intergenic sequences (59.6%) (Fig. 3A). Thus, genes generally do not harbor a high density of TEs in their immediate flanking sequence, and the high overall density of intergenic TEs can largely be explained by longer repeat-rich regions that are generally distant from genes.

The *Ectocarpus* genome, and indeed most brown algal genomes [2], has a very high intronic content (~ 46% of the genome) [58]. Furthermore, the *Ectocarpus* introns are far larger than model species with similarly compact genomes, where intron lengths are generally tightly distributed with a peak of 60–110 bp (e.g., *Arabidopsis thaliana* and *Drosophila melanogaster*) [59]. The *Ectocarpus* introns have a median length of 524 bp on the autosomes and 765 bp on the sex chromosome (Fig. 3B). These intron features, combined with the high percentage of intronic TEs, prompted us to investigate the relative characteristics of intronic and intergenic TEs. TEs inserted within introns are shorter (Kruskal–Wallis rank sum test p-value < 2.2e − 16) and more divergent from their consensus sequences (Kruskal–Wallis rank sum test p-value < 2.2e − 16) compared to intergenic TEs (Supplementary Material 8: Table S6). These differences are consistent for both intact and non-intact elements (Fig. 3C, D), where intact elements are defined as copies that span > 80% of the consensus length and are < 20% divergent from the consensus (i.e., copies that may retain the capacity to transpose).

The intergenic and intronic populations of TEs also differ in terms of their TE compositions. Intergenic proximal and intronic regions are predominantly populated by DNA transposons and LINEs, whereas LTR retrotransposons contribute the majority of TE sequence to intergenic distal regions (Fig. 3E, Supplementary Material 8: Table S7). DNA transposons consistently exhibit the highest copy numbers relative to the other subclasses, and autonomous elements dominate the total abundance in intergenic

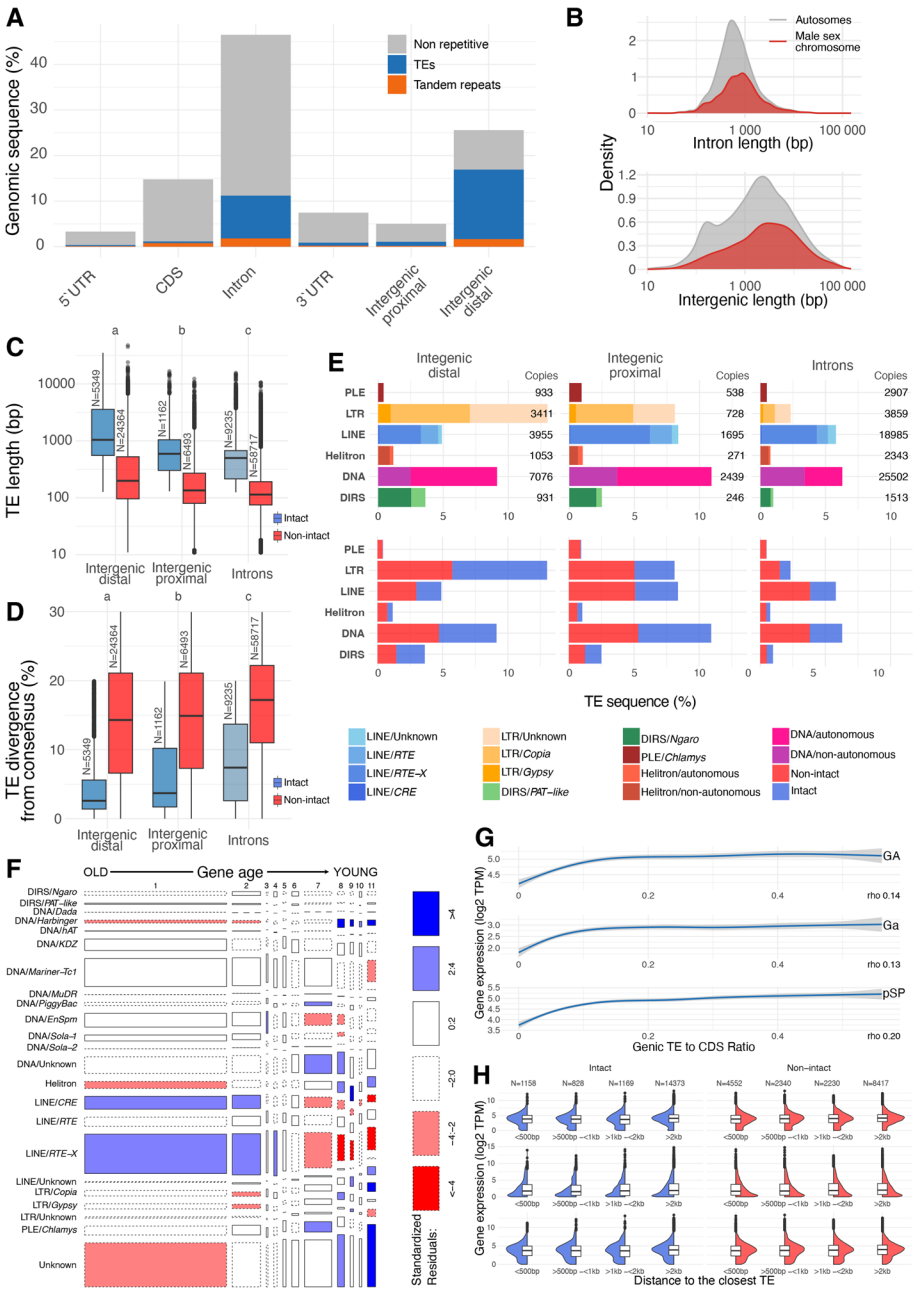


Fig. 3 **A** Density of TE bases by genomic site class. The blue and orange bars represent the proportion of sequence occupied by TEs and tandem repeats, respectively; gray bars represent the proportion of non-repetitive sequence. **B** Density plots showing the length distribution of introns (upper panel) and intergenic regions (lower panel) in autosomes and on the male sex chromosome. **C** Intact (blue) and non-intact (red) TE lengths by site class. **D** Divergence from consensus for intact (blue) and non-intact (red) TE copies; TEs with divergences > 30% were excluded. **E** Density of different TE classifications by site class. The bars are colored by order and superfamily in the upper panel (or autonomous/nonautonomous status) and by intact status in the lower panel. Note that the densities differ from those in panel **A** since an entire TE copy was considered "proximal" if any of its sequence was within 500 bp of a gene. **F** Mosaic plot showing the relationship between intronic TEs of different superfamilies and the age of the genes in which they are inserted; the gene age is shown as numbers from 1 to 11 where 1 are the most evolutionary ancient genes and 11 are *Ectocarpus*-specific genes. **G** Expression of genes relative to the ratio of CDS to total genic TE content, for different life stages. GA: gametophyte, Ga: gametes, pSp (parthenosporophyte). **H** Gene expression relative to the distance to the closest intact or non-intact TE

sequence but not introns, where the abundance of autonomous and non-autonomous DNA transposons is approximately equal (Fig. 3E; Supplementary Material 8: Table S7). Across all subclasses, the proportion of intact elements is generally highest in intergenic distal sequences, followed by intergenic proximal sequences, and introns (Fig. 3E). These results suggest that the shorter length of intronic TEs can be explained by both a higher density of shorter insertions (non-autonomous DNA transposons, LINEs that are frequently 5' truncated and include very short abundant non-autonomous families, see Fig. 1C) and a greater abundance of older, degraded TE copies. Conversely, LTR elements, which insert as full-length copies and have lengths ranging between 3.9 and 14.5 kb in *Ectocarpus*, dominate the intergenic distal regions together with autonomous DNA transposons, which include the giant families discussed above.

Brown algal genomes feature genes that span an extensive range of evolutionary ages, with many genes taxonomically restricted to the entire clade or to specific brown algal lineages [2, 4, 60]. We next tested whether there was any difference in the properties of intronic TE insertions among genes of different evolutionary ages. We found that LINEs tend to be enriched within evolutionarily old genes (phylostratum rank 1), assessed with genomic phylostratigraphy [60], whereas *Harbinger* DNA transposons and LTR retrotransposons are over-represented in younger genes that emerged after the evolution of the Ectocarpales (phylostratum rank 8) (Fig. 3F). Notably, elements that can be very large, specifically LTR retrotransposons and Helitrons, are over-represented within *Ectocarpus*-specific genes (phylostratum rank 11) (Fig. 3F). Together with the above observations that intronic TEs are generally shorter and more degraded relative to intergenic TEs, these results suggest that selection may play a major role in shaping the composition of intronic TE sequences, with longer TEs only tolerated in evolutionarily young genes that may be evolving under weaker selection.

Finally, we explored the effects of TE insertions on gene expression. Since intronic TEs are ubiquitous in *Ectocarpus* (81% of genes feature at least one genic insertion), we tested whether the abundance of intronic TEs is associated with gene expression levels. To quantify intronic TE abundance per gene, we calculated the ratio of genic TE nucleotides to CDS nucleotides. We found a very weak positive monotonic correlation between intronic TE abundance and gene expression (rho 0.14 in the gametophyte (GA) stage, rho 0.13 in the gametes, rho 0.20 in the parthenosporophytes (pSP). While the p-values are highly significant (p-value < 2.2e−16), this is likely due to the large sample size, and the relationship between TE content and gene expression is rather weak (Fig. 3G), suggesting that genes can be highly expressed despite containing substantial intronic TE densities. The absence of any major effect of intronic TE content on transcript abundance was consistent across three distinct life stages, the GA, gametes, and pSP. Considering the distance to the nearest intergenic TE, we found no effect of the vicinity of either intact or non-intact TEs on gene expression (Fig. 3H). Furthermore, 82% of genes had no intact TE within 2 kb, although 52% of genes did have at least one non-intact TE in these regions (Fig. 3H). This result suggests that proximal TE insertions are selected against in the *Ectocarpus* genome and is consistent with the presence of large gene-poor TE-rich islands that harbor mostly intact TEs.

Small RNAs and non-canonical histone modifications are associated with potentially active TEs

Both small RNAs and H3K79me2 have been shown to associate with TEs in Ectocarpus [27, 31]. We therefore asked whether these associations play a role in the regulation of TEs. We performed sRNA-seq across three different stages of the life cycle, namely the vegetative and fertile gametophyte and gametes of the Ectocarpus sp. 7 male reference strain Ec32 (Fig. 4A). Consistent with [27] and [32], we recovered a peak in sRNA abundance at 21 nt in all three life stages (Fig. 4B) and a strong uracil (U) bias at the 5' end (Supplementary Material 7: Fig. S4A), consistent with the presence of a single Argonaute (AGO) protein encoded in the genome. Across all three developmental stages, the abundance of sRNAs that mapped to TEs was

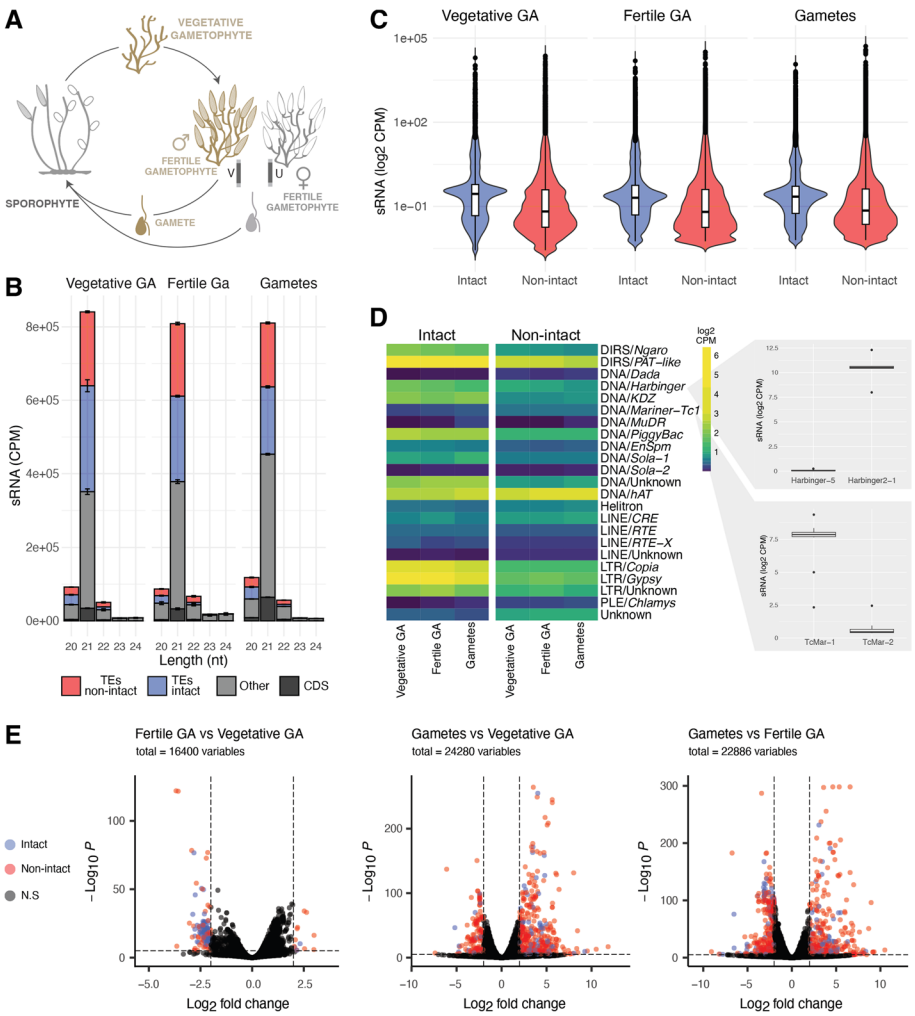


Fig. 4 **A** Life cycle of a male *Ectocarpus*. The colored life stages (vegetative gametophyte, fertile gametophyte, gametes) represent those sampled for sRNA sequencing in this study. **B** Size distribution and genomic feature annotation of sRNAs. Error bars represent SD of the mean of the biological replicates. **C** sRNA accumulation (CPM) on intact (blue) and non-intact (red) TE copies across the life stages. **D** sRNA accumulation (CPM) by TE superfamily. The zoomout panels show sRNA accumulation (CPM) differences between specific families from the *Tc1-Mariner* and *Harbinger* superfamilies. **E** Volcano plots illustrating TEs with differential sRNA expression throughout the *Ectocarpus* life cycle. Intact TEs are represented by blue dots, and non-intact TEs by red dots

substantially higher than those mapping to CDS (49,647.51 CPM average versus 10,640.45, Fig. 4B). A substantial proportion of sRNAs also mapped to other regions of the genome, which likely includes miRNA loci [32]. We found that CDS-mapping sRNA reads of 24 nt or longer were almost entirely in the same sense as the associated mRNA (Supplementary Material 7: Fig. S4B, S4C), consistent with degradation products. Conversely, on both TEs and CDS, sRNAs < 24 nt exhibited relatively equal proportions of sense and antisense mapping, and only 1% (0.4% in gametes) of TE-associated reads were 24 nt or longer (Supplementary Material 7: Fig. S4D, S4E, Supplementary Material 8: Table S8). We therefore focused on sRNAs of 20–24 nt for our analyses.

sRNAs overlap with 28.1% of TEs across all developmental stages (CPM > 0, Supplementary Material 6). Noticeably, the percentage of TEs associated with sRNA in all life stages increases to 64.7% when considering intact TEs exclusively and falls to 21.8% for non-intact TEs (Supplementary Material 8: Table S9). Furthermore, sRNA accumulation is significantly higher on intact TEs compared to non-intact TEs (Wilcoxon rank sum test with continuity correction p-value < 2.2e−16) (Fig. 4C). This enrichment towards intact TEs holds for all developmental stages and across TE superfamilies (Fig. 4C, D). Therefore, sRNAs in *Ectocarpus* are preferentially associated with intact, potentially active, transposons.

For both intact and non-intact TEs, sRNA accumulation varies substantially across different TE categories (Fig. 4D). For example, sRNA CPM values are consistently higher for LTR retrotransposons relative to LINEs; however, due to the challenges of sRNA normalization in terms of TE size, the result should be interpreted with caution. For other subclasses, we observed considerable differences in sRNA accumulation between superfamilies, e.g., *Ngaro* DIRS elements exhibit lower sRNA CPMs (1.81 log2CPM) than *PAT*-like DIRS (6.38 log2CPM), as do *Harbinger* DNA transposons (1.67 log2CPM) relative to *Mariner-Tc1* DNA transposons (0.33 log2CPM). Most strikingly, we found that specific families within the same superfamily could differ drastically in their association with sRNAs. For example, within the *Mariner-Tc1* superfamily, the autonomous family *TcMar-1_Ec32* exhibits 12-fold higher sRNA log2CPMs than *TcMar-2_Ec32* (Fig. 4D), and two *Harbinger* families differ similarly (Fig. 4D). The association between sRNAs and TEs thus appears to be dynamic and clearly non-random, suggesting a role for sRNA in the regulation of specific TE families.

To further assess the dynamics of TE regulation across the life cycle of *Ectocarpus*, we performed differential sRNA accumulation analysis between the three sampled life stages (Fig. 4A). Most TE copies (> 98%) exhibited no significant differential sRNA accumulation between the life stages, suggesting that TE regulation via sRNAs is stably maintained. However, 1.2 and 1.4% of the total TE copies showed differential sRNA accumulation between vegetative gametophytes and gametes, and fertile gametophyte and gametes, respectively (Fig. 4E, Supplementary Material 7: Fig. S5A, S5B; Supplementary Material 8: Table S10, S11). The changes in sRNA association with TEs during this transition were not strongly biased towards either upregulation or downregulation, suggesting a general reconfiguration of TE-associated sRNA accumulation in gametes (Supplementary Material 8: Table S10, S11). Non-intact TEs were significantly more likely to exhibit differential sRNA accumulation compared to

intact elements in both gametophyte-to-gamete comparisons (Supplementary Material 8: Table S12; chi-square test, $p < 0.00001$). This pattern may reflect the importance of maintaining silencing mechanisms for intact TEs in the new generation.

A small subset of 166 TE copies showed differential sRNA accumulation at the transition to fertility (i.e., between vegetative and fertile gametophytes); however, we noticed that this transition featured substantially more downregulated TEs, a significant proportion of which were intact elements (61.2% intact, relative to less than 30% in other transitions) (Fig. 4D, Supplementary Material 7: Fig. S5A). Moreover, 92.5% of these downregulated intact TEs belong to a single family of non-autonomous *Harbinger* DNA transposons (*Harbinger-N4_Ec32*, Supplementary Material 8: Table S13). These *Harbinger* elements are not associated with any significant changes in sRNA accumulation between any other life stages (Supplementary Material 7: Fig. S5B), potentially suggesting that the transition to fertility is associated with a highly specific shift in sRNA abundance in this family. However, since sRNAs almost invariably multimap to different copies of the same TE family, we cannot rule out an alternative explanation where a single copy of *Harbinger-N4_Ec32* may be inserted at a genomic locus that undergoes differential sRNA regulation.

We next assessed the relationship between H3K79me2 and TEs. We identified H3K79me2 peaks in the *Ectocarpus* v5 genome (Methods), and we found that 30.2% of intact TEs overlap with an H3K79me2 peak, relative to 23.1% of non-intact TEs. Furthermore, H3K79me2 signal was consistently higher within the TE body relative to the flanking regions, with a more pronounced increase in signal for intact TEs (Fig. 5A). Among TE superfamilies, *Harbinger* DNA transposons exhibited the most substantial enrichment in TE body H3K79me2 signal (Fig. 5B, Supplementary Material 7: Fig. S6A). Note that many TEs are in close proximity to other TEs, which reduces the relative signal of TE body to flanking sequence.

To explore the potential interplay between sRNA and H3K79me2 in TE regulation, we mapped both ChIP-Seq and sRNA reads to TE consensus sequences (Methods). Specifically, we split autonomous TEs into their left and right termini and genic sequence; for retrotransposons, we considered CDS, for DNA transposons exons and introns. We found that DIRS ($p = 0.024$), DNA transposons ($p = 0.023$), and LTR retrotransposons ($p = 5.9 \times 10^{-7}$) exhibited significant differences in H3K79me2 coverage across annotated features. In particular, DIRS and LTR retrotransposons showed significantly higher coverage over internal coding regions compared to terminal regions, especially the 5′ end (Fig. 5C, Supplementary Material 8: Table S14). For DNA transposons, although the overall *p*-value suggested a trend towards greater coverage over internal features (exons, introns), none of the pairwise comparisons remained significant after multiple testing correction, indicating a weaker or more variable pattern. Other TE groups did not show statistically significant differences. Nonetheless, individual TE families often exhibited distinct coverage profiles, with internal regions frequently showing elevated H3K79me2 coverage.

These observations led us to ask whether H3K79me2 preferentially targets TE-encoded genes. Supporting this idea, we found that autonomous TEs, which typically contain intact coding sequences, exhibited significantly higher H3K79me2 coverage than non-autonomous TEs ($p = 0.034$; Supplementary Material 8: Table S15). Together,

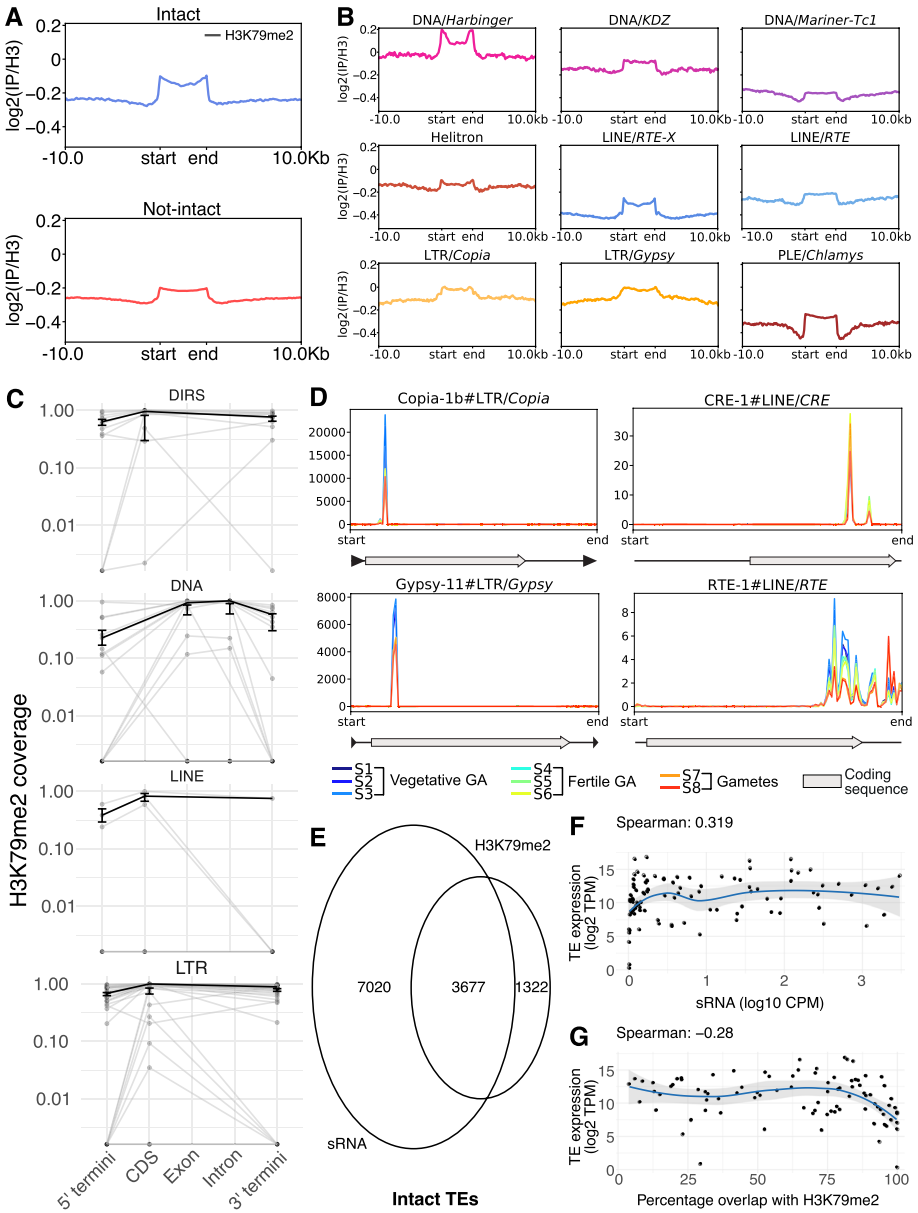


Fig. 5 **A** H3K79me2 (normalized on H3) accumulation profiles on intact and non-intact elements. Start and end indicate the TE body, with 10-kb flanking regions. **B** H3K79me2 (normalized on H3) accumulation profiles on TE superfamilies. Start and end indicate the TE body, with 10-kb flanking regions. **C** H3K79me2 coverage over TE features. Each gray dot represents the mean coverage value, across 8 biological replicates, for each TE family consensus sequence. Gray lines connect dots belonging to the same TE family. The black line represents the median coverage value for each TE feature, with SD, within the TE subgroup represented. **D** sRNA reads distribution over TE body. Each line represents a biological replicate. **E** Venn diagram showing the intersect between intact TE copies associated with sRNA and H3K79me2. **F** Scatter plot and fitted curve (Loess regression) showing the correlation between sRNA abundance (CPM) on intact TE copies and TE gene expression. Each point represents a TE family. **G** Scatter plot and fitted curve (Loess regression) showing the correlation between the average proportion of intact TE body overlapped by H3K79me2 peaks and the TE gene expression

these results indicate that H3K79me2 may play a role in the transcriptional repression of TE genes.

Similarly, sRNA accumulation profiles were highly variable across TE sequences. Within superfamilies, sRNA peaks were not significantly associated with specific annotated TE features (Fig. 5D, Supplementary Material 7: Fig. S6B, S6C). However, sRNAs did not accumulate uniformly along the TE body, but tended to concentrate in localized regions, consistently across life stages. For example, both LTR/*Copia* and LTR/*Gypsy* elements accumulate sRNA preferentially towards the 5' end of the TE body; on the contrary LINEs typically exhibit sRNA accumulation towards the 3' end (Fig. 5D, Supplementary Material 7: Fig. S6B).

Considering sRNA and H3K79me2 collectively, only 8.6% of TE copies are associated with both sRNA mapping and a H3K79me2 peak (Supplementary Material 7: Fig. S7A). This number increases to 22.2% of intact TEs, and strikingly, 73.6% of intact TEs that intersect an H3K79me2 peak also exhibit sRNA mapping (Fig. 5E). Thus, most intact TEs associated with H3K79me2 are also associated with sRNAs, whereas the majority of intact TEs associated with sRNAs are not associated with H3K79me2.

We next investigated the differences between the sRNA population associated with intact TEs that intersect an H3K79me2 peak and the set associated with TEs that do not intersect with an H3K79me2 peak. The length distribution of sRNA in the two populations confirms the peak at 21nt (Supplementary Material 7: Fig. S7B, S7C), with only small differences in strand preference: the 21nt sRNA associated with TEs that intersect an H3K79me2 peak are equally sense and antisense, with antisense sRNA becoming more abundant among 23 and 24 nt sRNA (Supplementary Material 7: Fig. S7B). In contrast, sRNAs associated with TEs that do not overlap with the histone mark present a higher abundance of antisense 21 nt sRNA (Supplementary Material 7: Fig. S7C). This observation potentially highlights the importance of post-transcriptional gene silencing (PTGS) in TEs that lack H3K79me2. Alternatively, it may suggest that sense sRNAs contribute to the deposition of H3K79me2. Furthermore, we observed no difference in 5′ nucleotide bias between the two sRNA populations; both show a strong preference for uracil (U) at the 5′ end (Supplementary Material 7: Fig. S7D, S7E; Supplementary Material 8: Table S16). Regarding the distribution of intact TEs relative to genes, the majority of both intergenic and intronic TE copies lack H3K79me2 marks (59 and 72%, respectively; Supplementary Material 8: Table S17). Nonetheless, intact intergenic TEs are more likely to associate with an H3K79me2 peak than intronic copies (chi-square p-value < 0.00001). These results suggest that H3K79me2-marked TEs tend to be located further from genes, potentially reflecting a mechanism to prevent unintended silencing of nearby genes due to the broad distribution of H3K79me2 peaks. Alternatively, the lower proportion of marked TEs within introns may reflect the reduced presence of autonomous TEs in these regions, as H3K79me2 appears to preferentially target TEs encoding transposition-related proteins.

Finally, we explored the relationships between TE gene expression across the same life stages, and sRNA and H3K79me2 association. Restricting the analysis to manually curated TE families that encode genes, the nine most highly expressed TE families are all *Copia* LTR retrotransposons (Supplementary Material 7: Fig. S7C, Supplementary

Material 8: Table S18), which dominate the most expressed TE families across all life stages (Supplementary Material 7: Fig. S7D, Supplementary Material 8: Table S18). In total, 97.9% of these gene-encoding families exhibited expression of at least 1 TPM (transcripts per million) consistently across life stages, suggesting broad transcriptional activity for intact TEs in *Ectocarpus*. We found a weak positive correlation (Spearman's rho = 0.319) between TE gene expression and sRNA abundance on intact copies, suggesting that TEs associated with high sRNA abundance can still be highly expressed (Fig. 5F). Conversely, we observed a weak negative correlation (Spearman's rho = -0.28) between TE gene expression and the average percentage of intact TEs overlapped by H3K79me2 peaks (Fig. 5G). In particular, the curve applied to highlight the trend between variables (Loess Regression) shows a pronounced decline in TE expression when more than 75% of the TE body intersects with H3K79me2 peaks (Fig. 5G). Therefore, coverage of a high proportion of the TE body by H3K79me2 is associated with reduced transcriptional activity.

Discussion

The diverse mobilome of *Ectocarpus* and TE-enrichment on the sex chromosome

Despite occupying only 28% of its genome, *Ectocarpus* features a high diversity of TEs, harboring nearly 340 families from six subclasses and 19 superfamilies. Furthermore, more than half of the cumulative TE sequence in the genome belongs to copies that exhibit less than 5% divergence from their family consensus sequence, suggesting recent or potentially ongoing transposition. This is supported by our observation that ~ 98% of curated autonomous TE families are expressed consistently across life stages, potentially leading to the production of transposon proteins. Population genomics or experimental data will be required to demonstrate active transposition in *Ectocarpus*. The skew of the TE landscape towards younger elements also implies that the *Ectocarpus* genome is under selection to maintain compactness, with inactive and degrading TEs being purged from the genome. Although detailed analyses of TE landscapes and TE diversity have not been performed for other brown algae, their greater genome sizes and repeat densities [2] suggest that selection against TE accumulation may be weaker in several species, many of which are morphologically more complex and would be expected to have lower effective population sizes.

The substantial diversity of *Ectocarpus* TEs highlights the benefits of studying taxonomically diverse genomes. For example, our annotation of *Gypsy* LTRs increases the number and diversity of known fusions between integrases and chromatin-reader domains, and revealed acquisitions of NADAR domains by diverse retrotransposons. Whether these retrotransposons may facilitate this diversification process through extrachromosomal recombination as suggested for other species [61] will be an exciting area for future research. Notably, DNA transposons are the most abundant and diverse TE subclass in *Ectocarpus*, spanning almost 8% of the genome. This proportion is likely to be an underestimation, as most of the remaining unknown TEs are expected to be DNA elements. Automated classification of DNA transposons is highly challenging [35]; for autonomous families, nucleotide-to-protein homology searches are complicated by the abundance of introns in transposase genes and low sequence conservation of DD(E/D)

transposases, whereas nonautonomous families are often very short and lack obvious structural characteristics. Although DNA transposons undergo cut-and-paste transposition, they can increase in copy number during certain transposition events [9, 62], and the relative abundance of retrotransposons and DNA transposons differs substantially between eukaryotes. For example, although TEs comprise approximately the same genomic space, in *C. elegans* DNA transposons account for more than 70% of the total TE content [63] whereas in *A. thaliana* this figure is ~ 27% [64] and in *D. melanogaster* only ~ 7% [65]. DNA transposons are also the dominant subclass in many, but not all, fish genomes, whereas they are absent from most mammalian genomes [66]. Furthermore, TE horizontal transfer events in vertebrates overwhelmingly occur among fish [66]. Like many fish, brown algae undergo external fertilization in aquatic environments where opportunities for horizontal transfer may be increased, and it would be interesting to address the evolutionary origin of brown algal DNA transposons in future work.

An additional contributing factor to the abundance of DNA transposons in the *Ectocarpus* genome is the presence of several giant elements, specifically from the *KDZ*, *EnSpm*, and *Sola-1* superfamilies, which have maximum lengths ranging from 15.2 kb (*EnSpm*) to 30.3 kb (*KDZ*). *KDZ* elements encode complex transposases that can exceed 1300 amino acids [67], frequently carry accessory genes in fungi [68], and are among the largest DNA transposons reported. Active transposition of a 31.7 kb *KDZ* element was observed in the green alga *Chlamydomonas reinhardtii* [69], and giant *EnSpm* elements carrying multiple accessory genes have been reported in red algae [51]. The evolution of giant TEs is thought to be limited due to the risk of ectopic recombination occurring between copies, which would be strongly selected against by the host [70]. *Ectocarpus*, along with many other taxa associated with giant TEs (fungi, green and red algae), spends a substantial portion of its life cycle in the haploid stage [7]. Although DNA repair mechanisms in *Ectocarpus* remain uncharacterized, a potentially low frequency of homologous recombination in the haploid stage could help explain the evolution of giant transposable elements in this species.

Consistent with previous studies, we found that TEs are enriched on the *Ectocarpus* sex chromosomes [4, 53]. Notably, this enrichment is not restricted to the non-recombining SDR but also extends into the PARs, despite PARs recombining at rates comparable to autosomes [4, 53]. This pattern contrasts with observations in other organisms with haploid chromosomal sex or mating type determination; for example, *Microbotryum* fungi show no TE enrichment in their recombining PARs [71]. We hypothesize that the high abundance of DNA transposons in the *Ectocarpus* genome may underlie this pattern. In this model, DNA transposons would initially accumulate in the non-recombining sex-determining region (SDR), along with other repetitive elements. Due to the known tendency of DNA transposons to insert near their donor copies (a phenomenon known as "local hopping"; [55]), the SDR could act as a source of transposition events, leading to increased insertion rates in the adjacent PARs. Over time, the resulting accumulation of DNA transposons in the PARs could create a reservoir of relatively neutral sequence, facilitating the secondary accumulation of other TEs that might otherwise be subject to stronger purifying selection. Although this hypothesis is difficult to test directly, we do observe a significant enrichment of DNA transposons (relative to copy-and-paste elements) in the *Ectocarpus* PARs. In contrast, *Microbotryum* genomes

contain relatively few DNA transposons and are instead dominated by LTR retrotransposons [71], consistent with the absence of TE enrichment in their recombining PARs.

The *Ectocarpus* female SDR contains a distinct set of TEs compared to the male SDR. Similar patterns of sex-chromosome TE differences have been reported in other species and attributed to sex specific activity of certain TEs or differences in epigenetic silencing between sexes. For example, in *Silene*, the Y chromosome shows different TE content compared to autosomes [72] and in the dioecious plant *Rumex acetosa* the X and Y chromosomes exhibit differing TE distributions [73]. In animals, sexually dimorphic TE expression in the germline has been linked to elevated TE activity in males, which undergo more mitotic divisions, a trend consistent with the positive correlation between TE activity and mitotic rate [74]. However, *Ectocarpus* has minimal morphological and developmental differences between male and female gametophytes. Thus, sex-specific silencing or expression is unlikely to account for the observed TE differences between male and female SDRs. A more plausible explanation is that, due to high population polymorphism, the female lineage sampled in this study randomly acquired a different set of TEs in the SDR, which then became fixed due to the lack of recombination in this region. Population genomics approaches will be required to test whether differences between the *Ectocarpus* male and female SDRs hold across the species.

Ectocarpus has high intronic TE density, but intact TEs are primarily located far from genes

Ectocarpus exhibits two major TE populations within the genome: intergenic TEs (60.8% of TE bases) and intronic TEs (34.1%). Introns comprise a major fraction of the *Ectocarpus* genome and are substantially longer than introns in many model organisms with comparable genome sizes. Indeed, *Ectocarpus* introns feature exonic-splice enhancers similar to those of mammals, which may be considered an adaptation to the intron-richness of the genome [58]. We found no strong relationship between intronic TE content and gene expression, suggesting that *Ectocarpus* can tolerate a moderately high intronic TE content with little impact on gene expression, similar to several plant species [75]. Nonetheless, we do find evidence for selection against TEs in introns; intronic TEs are shorter, more degraded, and less frequently intact relative to intergenic TEs, and dominated by TE families that are naturally shorter (e.g., LINEs and LINE-dependent nonautonomous elements, nonautonomous DNA transposons). Indeed, insertions of the longest TE families, LTRs and Helitrons, are only preferentially found in *Ectocarpus*-specific genes, which may be evolving under much weaker selection relative to evolutionarily ancient genes.

Intergenic TEs in *Ectocarpus* show distinct characteristics depending on their proximity to genes. TE insertions adjacent to genes are more frequently shorter and more degraded, and most genes do not feature an intact TE of any length within 2 kb. Hence, the greatest abundance of TEs in the *Ectocarpus* genome, and especially longer elements such as LTR retrotransposons and autonomous DNA transposons, are concentrated in gene-poor repetitive islands that are not associated with the centromeres. Many genomes feature similar TE-rich islands, for example in some fungi [76] and insects [77], and these regions can be associated with rapid evolution and extensive structural polymorphism at the population level [78]. The paucity of intact TEs near to genes may also explain the absence of any obvious relationship between gene expression and distance to

the nearest TE, which, for example, may be expected if repressive histone marks associated with the TE were to influence expression of the neighboring gene. Assessment of the effect of rare TE insertions on gene expression across an *Ectocarpus* population would be one avenue to address this question [79].

Regulation and repression of transposable elements in *Ectocarpus*

Host silencing and regulation of TEs is highly complex and frequently involves sRNAs, DNA, and histone modifications, as well as crosstalk between these mechanisms [80, 81]. In *Ectocarpus*, which notably lacks cytosine methylation, the mechanisms underlying effective TEs control remain largely unknown, and this is particularly striking given the high density, diversity, and transcriptional activity of intact and potentially active elements.

Our results reveal a clear enrichment for sRNAs towards intact TEs. Furthermore, sRNAs association is not uniform across different TEs; TE families within the same superfamily (e.g., *Tc1-Mariner* DNA transposons) can exhibit markedly different sRNA targeting. These findings suggest that sRNAs do not simply target recent, potentially active TE insertions, but rather a specific subset of elements. We also show that sRNAs show consistent, localized patterns along TE bodies, supporting the hypothesis of a functional role in TE silencing. 21-nt sRNAs are typically linked to post-transcriptional gene silencing (PTGS) in other systems such as plants, where sRNAs of different lengths are associated with distinct silencing pathways [82]. In *Ectocarpus*, given the lack of other characteristic sRNA classes and a consistent association between sRNA and specific TE features, it remains unclear whether these sRNAs mediate PTGS, participate in transcriptional silencing, or play multiple roles depending on the context.

Approximately a third of intact TEs are also associated with H3K79me2, which has been linked with decreased transcript abundance in *Ectocarpus* [29]. H3K79me2 peaks are typically localized within the TE gene body, suggesting a role in transcriptional repression. Notably, most intact TEs that are associated with H3K79me2 enrichment show sRNA accumulation, whereas most intact TEs that lack H3K79me2 are still associated with sRNAs. This could imply that there is some independence between these putative silencing mechanisms, or that sRNAs play multiple roles in TE control. Indeed, while we did find a weak repressive effect of H3K79me2 on TE gene expression when the mark is present across most of the TE body (> 75% of TE copy), there was no clear relationship between sRNA abundance and TE gene expression. sRNAs are known to play multiple roles in TE silencing [83]. Beyond their well-characterized involvement in post-transcriptional gene silencing (PTGS) through the cleavage of target mRNAs [84, 85], sRNAs are also implicated in transcriptional silencing pathways. In animals lacking cytosine methylation, such as *C. elegans*, sRNAs have been demonstrated to directly initiate the deposition of repressive histone marks on TEs [86, 87]. Similarly, in plants, sRNAs are involved in pathways that facilitate the establishment of repressive histone modifications, such as H3K9me2 at specific TE loci and developmental stages including *Arabidopsis* embryos [88]. Although the interaction between histone PTMs and sRNAs remains elusive in brown algae, the observation that more than 70% of intact H3K79me2-associated TEs also associate with sRNAs suggests that a pathway similar to that observed in other organisms may exist. However, given the prevalence of

TEs associated with sRNAs but lacking H3K79me2, it is unlikely that sRNAs function exclusively to drive histone modifications. Disentangling these possibilities will require experimental tools currently lacking in *Ectocarpus*. For instance, degradome sequencing could test whether sRNAs direct mRNA cleavage, while loss-of-function mutants for AGO1 or histone methyltransferases could determine the functional consequences of sRNA and H3K79me2 loss.

The association between sRNA and TEs remains largely stable throughout the life cycle of *Ectocarpus*, with only a small fraction of TE copies (< 2%) showing differential sRNA association during the developmental transitions studied. Despite this overall stability, some differences in sRNA accumulation are observed particularly between the gametophyte stage (both vegetative and fertile) and the gametes. These differences are pronounced over non-intact TEs, possibly reflecting their reduced silencing priority or a role in reproductive reprogramming as documented in other organisms [81, 89]. The observed differences are not biased towards increased or decreased sRNA accumulation, as some TE copies exhibit higher sRNA association in gametes compared to the gametophyte stage, while others show the opposite trend.

Our sRNA sampling spanned the developmental transition from immature to fertile gametophytes. In *Ectocarpus*, any vegetative gametophyte cell has the potential to develop into reproductive structures where gametes are formed. Therefore, a stable sRNA landscape, particularly targeting potentially active TEs, may play a critical role in safeguarding genome integrity during this sensitive developmental window.

Conclusions

Our study reveals a rich genetic diversity of the *Ectocarpus* mobilome and uncovers a complex, multilayered regulatory landscape shaped by small RNAs and chromatin modifications (Fig. 6). These mechanisms operate in the absence of canonical epigenetic silencing pathways found in animals and plants, suggesting the presence of unique, lineage-specific mechanisms for transposable element control in *Ectocarpus* that remain

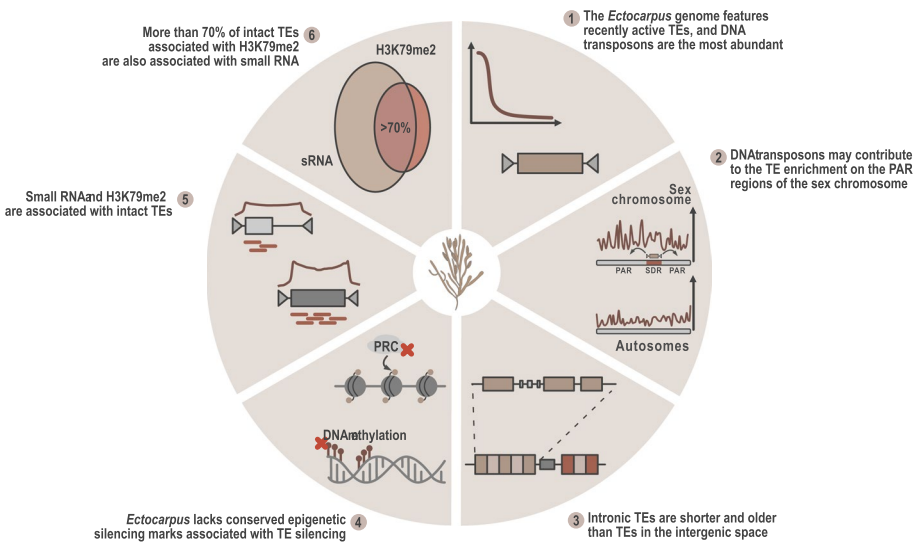


Fig. 6 Schematic illustration of the key findings on the *Ectocarpus* mobilome

to be uncovered. This work provides new insights into TE regulation in non-model eukaryotes and expands our understanding of epigenetic innovation across lineages.

Methods

De novo annotation and manual curation of transposable elements

Manual curation of the most abundant *Ectocarpus* TE families was performed following the methodology of [90]. Briefly, preliminary repeat models were generated by running RepeatModeler v2.0.5 [91] on the v5 reference genome using the "LTRStruct" model [92, 93]. The abundance of each automated repeat model in the genome was then estimated by passing the output of RepeatModeler to RepeatMasker v4.0.9 (https://www.repeatmasker.org/). Repeat models that contributed more than 100 kb to the genome were then selected for curation. Individual copies of each repeat were retrieved from the genome using blastn v2.9.0 ("-evalue 10e-10") [94] extended to include flanking sequence of 2 kb, and aligned using mafft v7.453 ("E-INS-i" setting) [95]. Alignments were then viewed manually and full-length transposons were identified where possible by identifying regions where sequence from multiple independent aligned copies transitioned into unalignable flanking sequence (with iteration of the previous steps performed if the initial extension and alignment did not reach both TE termini). Consensus sequences corresponding to individual full-length families were then produced using the AdvancedConsensusMaker tool (hiv.lanl.gov). For transposons carrying genes, putative protein sequences were predicted by searching for the longest ORFs (intron-less genes) or by using RNA-seq alignments (see below) to curate gene models (intron-containing genes) in the IGV genome browser v2.15.4 [96]. A small number of lower abundance elements were also subjected to manual curation, including several autonomous DNA transposons (to aid classification of associated and more abundant nonautonomous elements) and all families that had previously been manually curated for *Ectocarpus* and made available at the Repbase repository (version 29.02) [97]. Classification was performed based on established characteristics of TE superfamilies, including homology to curated transposases, terminal structures and motifs (LTRs, terminal inverted repeats, etc.), and TSDs if present. Protein domains were identified using InterProScan [98] or HHpred [99] available online at EMBL-EBI or the MPI Bioinformatics Toolkit [100], and all non-canonical domains were manually inspected by visualizing multiple sequence alignment (e.g., chromatin reader and NADAR domains of *Gypsy* elements, see Supplementary Material 7: Fig. S1, S2).

A phylogeny of *Gypsy* families was produced by combining the 13 available Pol protein sequences of the manually curated elements with the dataset of 85 *Gypsy* reverse transcriptase domain sequences used for phylogenetic analysis by [42]. Sequences were aligned using mafft (L-INS-i parameters), manually trimmed to reduce the alignment to only the reverse transcriptase domain, and automatically trimmed to remove gap-rich alignment columns with trimAl v1.4.rev22 ("gappyout" model) [101]. The resulting alignment was passed to IQ-TREE v2.3.0 [102] which was run using the parameters "-m MFP -bb 1000" to perform model selection [103] and ultrafast bootstrapping [104].

The manually curated elements were incorporated to the existing TE library published in Cock et al. (2010) (available at: https://urgi.versailles.inra.fr/Data/Trans

posable-elements/Ectocarpus), containing 897 consensus sequences obtained running the REPET pipeline [34] on the v3 assembly of the genome. Before merging, the nomenclature of the consensus sequences was standardized (e.g., instances of "EsGypsy" were replaced with "LTR/Gypsy," and "EsCopia" with "LTR/Copia") and uniqueness was ensured by appending a unique identifier based on an incremental counter to each sequence header. To address misclassification, the unified library was clustered using CD-HIT-EST v4.8.1 [105] with 80% identity and length thresholds (-c 0.8 -aS 0.8 -d 0 -G 0 -g 1 -b 500 -M 8000). Following clustering, REPET consensus sequences were reclassified based on their association with manually curated elements to maintain consistency within the library.

A known issue with REPET is the misclassification of tandem repeats as nonautonomous LTRs [106]. Tandem repeats in the v5 genome were identified using Tandem Repeat Finder (TRF) v4.09.1 [107] using the parameters "2 5 7 80 10 50 2000 -d -ngs", and resulting coordinates were subsequently intersected with the results of a preliminary RepeatMasker run (see below). For each TE family, the percentage of its cumulative length that overlapped with tandem repeats identified by TRF was calculated, and any family with more than 60% overlap was reclassified as a tandem repeat in the library.

Redundancy within the library was addressed through a two-step process. First, CD-HIT-EST v4.8.1 [105] was executed again with a 95% similarity threshold (-c 0.95 -aS 0.95 -d 0 -G 0 -g 1 -b 500 -M 8000). The clustering results were then used to identify and remove redundant sequences. To perform a final round of reclassification, remaining unclassified elements were used as queries in a blastx search with -evalue 0.001 against TE proteins from Repbase, supplemented with novel *Ectocarpus* TE proteins annotated as part of the curation process (Supplementary Material 2).

Analysis of transposon distribution along the Ectocarpus genome

TEs were mapped on the *Ectocarpus* v5 genome [33] using RepeatMasker version open-4.0.9 (Smit et al. 2013–2015; http://www.repeatmasker.org) installed with rm-blast v2.9.0. The following parameters were set: -a -s -gff -xsmall -cutoff 230. The RepeatMasker output files were processed using the Perl script onecodetofindthemall.pl [108] with the parameters — unknown and — strict, which reconstructs fragmented repeats and full-length LTR retrotransposons, excluding tandem repeats from the analyses. After OneCodeToFindThemAll, the number of copies decreased from 233,743 detected through the annotation step with RepeatMasker to 131,728.

The output files of OneCodeToFindThemAll were concatenated in a single file, then processed through an in-house script to generate BED and TSV files providing coordinates of each TE copy. TE copies were defined as "intact" if they exhibited less than 20% divergence from their consensus sequence and spanned more than 80% of the consensus sequence length.

Tandem repeats were identified and annotated by combining the output of RepeatMasker (simple repeats and consensus sequences identified as tandem repeats, see above) with the output of TRF. When TE and tandem repeat annotations overlapped, TEs were given precedence. Computationally reproducible analysis scripts can be found at (https://github.com/edinatale/Ectocarpus-TEs.git).

The genome was divided into 100-kb non-overlapping windows and the density of TEs, tandem repeats and CDS was calculated for each window with coverageBED. To visualize the distribution of TEs, tandem repeats and CDS specifically in the male and female SDRs, 50-kb windows were used given the small size of these genomic regions. The densities in each window were plotted with Circos [109] and KaryoploteR [110]. TE density per 100-kb window was used to compare sex chromosome versus autosomes and between PAR, SDR, and autosomes using respectively the Kruskal–Wallis test followed by Dunn's test with Bonferroni correction, and Wilcoxon rank-sum tests.

To visualize the distribution of TEs adjacent to subtelomeres, each chromosome was split to 5-kb windows where the first window begins immediately adjacent to the subtelomeric satellite repeats. For each window the median, first and third quantiles were calculated among all chromosomes. The sex chromosome was excluded from this analysis because one subtelomeric region has not been defined.

TE location relative to genes

The *Ectocarpus* annotation was first checked against repeat coordinates to avoid including any genes carried by TEs in analyses. An in-house script was used to calculate the intersect between TE coordinates and CDS per transcript, and any transcripts exhibiting > 70% intersect with TEs were removed.

TE densities were calculated by site class by dividing the genome into 5' and 3' UTRs, CDS, introns, and intergenic region. Moreover, intergenic sequence was divided to gene proximal sequences (within 500 bp of a gene) or gene distal sequences. The densities calculated in Fig. 3A were obtained using BEDTools coverageBed [111]. When considering individual TE copies, a copy was defined as intergenic proximal if > 1% of its sequence fell within 500 bp of a gene. TE copies assigned as intergenic proximal, intergenic distal, and intronic were compared based on their length and their percentage of divergence from their consensus sequence, and statistical analyses were conducted using Kruskal–Wallis and Dunn's tests.

To test the relationship between intronic TE content and gene expression, the ratio of the total genic TE content to CDS length was calculated. Mono-exonic genes were not considered in this analysis, because of their overall low expression in the *Ectocarpus* genome [27]. The evolutionary ages of *Ectocarpus* genes were estimated using GenEra [60] using default parameters.

The distance of each gene to the nearest TE was calculated using BEDTools closest [111], with intronic TEs excluded. For each gene, a maximum distance of 2 kb from the nearest TE was considered, given that the typical intergenic space is approximately 1 kb. Genes were grouped based on the distance to the nearest TE: within 500 bp, between 500 bp and 1 kb, between 1 and 2 kb, and greater than 2 kb. These groups were subsequently used to assess the relationship between the proximity of the nearest TE and the expression levels (log2 TPM) of the corresponding genes based on previous gene expression data [29, 112].

Algae strains and growth conditions

All biological material manipulations were carried out under sterile conditions. *Ectocarpus sp7* strain Ec32 male parthenosporophytes were grown at 14 °C for 3 weeks in sterile

and filtered natural seawater in 12 h/12 h dark/light cycles. When reaching sexual maturity, unilocular sporangia were manually isolated and incubated in a moist chamber for 24–48 h in high light conditions as described in [113] until spore release. Gametophytes were incubated for approximately 2 weeks until collected (immature gametophytes) and then for 4–5 weeks at 14 °C—until production of plurilocular gametangia was confirmed under microscope (fertile stage). Gametes were span at 10,000 *g* for 5 min; supernatant was removed carefully without disturbing the gamete pellet which was flash frozen in Liquid Nitrogen and stored at − 80 °C until RNA extraction was performed. Approximately, 1500 gametophytes (150 Petri dishes with 10 gametophytes each) were used for gamete release per biological replicate. Immature and fertile gametophyte tissue was collected from 5 Petri dish with 10 gametophytes each at the same time of development. Immature gametophytes were grown in the same conditions but harvested 3 weeks before the gametangia development and lack of plurilocular sporangia was confirmed under the light microscope. Three independent biological replicates were collected for each sample, except for the gametes in which 2 biological replicates were collected.

sRNA library preparation and short-read sequencing

sRNA libraries were prepared starting with total RNA isolated from 50–100 mg of tissue, except for the "gamete" sample in which it was not possible to estimate weight. Instead, all gametes harvested from one gamete release from approximately 1500 gametophytes were used. Total RNA was isolated as previously described [30]. Briefly, gametophytes were grounded to a fine powder in liquid nitrogen in 1.5 mL Eppendorf tubes using a micropestle. Gametes were directly resuspended in the extraction buffer by up and down pipetting as well as vertexing. The algae powder was further homogenized in 500 µl of cetlytrimethylammonium bromide (CTAB)-based extraction buffer [100 mM Tris–HCl pH 8, 1.4 M NaCl, 20 mM EDTA pH 8, 2% Plant RNA Isolation Aid (PVP, Invitrogen AM9690), 2% CTAB and 1% β-mercaptoethanol by vortexing and incubating at 65 °C until all samples were processed (5–20 min). RNA was extracted by mixing the homogenate with 1:1 volume of chloroform/isoamylalcohol (24:1). Supernatant was collected after centrifugation at 10,000 *g* for 15 min at 4 °C and the chloroform/isoamylalcohol extraction step was repeated. RNA was precipitated with 1:1 (v/v) isopropanol at − 20 °C overnight. Samples were centrifuged at 20,000 *g* for 30 min at 4 °C and pellets washed twice with 1 mL of 70% ethanol and resuspended in nuclease free water. DNase treatment was performed using a TURBO DNase Kit for 20 min, according to the manufacturer's instructions (Thermo Fisher Scientific, AM1907). Total RNA was purified with the RNA clean and concentrator kit (Zymo Research, R1013) following the manufacturer's instructions with the following modifications: columns were washed twice with 400 µl of RNA Prep buffer and four times with 700 µl of RNA Wash Buffer. RNA was eluted in 12 µL of nuclease-free water. sRNA libraries were prepared as described in [114] starting from 1 µg of total RNA. In brief, NEBNext multiplex small RNA library prep kit (New England Biolabs) was used according to the manufacturer's instructions. Libraries were indexed during the PCR step with 15 cycles and size-selected using BluePippin 3% agarose cassettes

(Sage Science). Equimolecular pooled libraries were sequenced with Illumina NextSeq2000 on a 1×50 bp single-end reads. The sRNA sequencing data can be accessed in the SRA Knowledge Base with the accession number PRJNA1197354.

sRNA-sequencing analysis

The small RNA reads generated (Supplementary Material 8: Table S19) in triplicate for the vegetative (S1, S2, S3) and fertile (S4, S5, S6) gametophytes, and in duplicate for gamete samples (S7, S8), were processed using snakemake_sRNAseq (https://github.com/seb-mueller/snakemake_sRNAseq), an automated Snakemake [115] pipeline for small RNA-sequencing data analysis. Read quality was assessed with FastQC (v0.11.7, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptor sequences were removed from raw reads using cutadapt (v4.2; [116] with quality trimming (-q 20) and length filtering flags.

Two independent mapping analyses were performed. In the first step, reads < 18 or > 40 bp were discarded (length filtering -m 18 -M 40), and the remaining sequences were aligned to the Ectocarpus sp. 7 reference genome [33] and the TE consensus sequences. In the second step, only reads between 20 and 24 bp (length filtering -m 20 -M 24) were selected and mapped to the same reference genome. Mapping was performed with bowtie version 1 [117], using the parameters "multi" or "unique", -v 0, -m 1, optimized for high-accuracy alignment of short reads [118]. Both uniquely and multi-mapped reads were retained for downstream analysis. The configuration files used for this analysis are available at https://doi.org/10.17617/3.4SCUJN.

Small RNA abundance was quantified as counts per million (CPM) over individual TE copies using the featureCounts function from the R package Subread [119], with the options countMultiMappingReads = TRUE, allowMultiOverlap = TRUE and fracOverlap = 1. Differential accumulation of sRNAs over TEs was assessed with DESeq2 [120] using raw count data. Raw read counts were filtered to retain only TE loci with a minimum of 10 reads across all samples, and shrinkage was applied using the apeglm method. Comparisons were conducted separately for the following groups: vegetative gametophyte vs. fertile gametophyte, vegetative gametophyte vs. gametes, and fertile gametophyte vs. gametes.

Chip-seq mapping and analysis

FastQ files for the ChIP-seq data (H3K79me2) from *Ectocarpus* male gametophyte were downloaded from the SRA database [29]. Data were mapped to the V5 reference genome and the TE consensus annotated genes using the nf-core/chipseq v2.0.0 [121] with BWA as the aligner, allowing multi-mapping ("keep_multi_map": true). To normalize the H3K79me2 ChIP-seq signal against the H3 control, bamCompare from the DeepTools suite [122] was used with a bin size of 10 bp (–bs 10) to calculate log2(IP/H3) ratios (–operation log2). The resulting normalized signal files (bigWig format) were used to compute signal intensities over annotated TEs using computeMatrix in scale-regions mode. Parameters included -bs 10, -b 10,000, -a 10,000, –regionBodyLength 1000. The flanking regions of 10 kb were used to ensure inclusion of non-TE sequence, thereby reducing noise in the analysis, as TEs are often located in close

proximity to each other. Profiles showing the distribution of normalized ChIP-seq signal over TEs were generated using plotProfile, with the –perGroup option to create separate profiles for each TE superfamily.

Broad peaks of H3K79me2 enrichment were identified using MACS2 [123] with the –broad option, and using the corresponding H3 ChIP-seq data as the control. Peak files were intersected with TE annotations using BEDTools [111] to calculate the proportion of each TE copy overlapping with H3K79me2-enriched regions.

TE expression

RNA-seq data from the wildtype male *Ectocarpus*, encompassing the vegetative gametophyte, fertile gametophyte, and gamete stages, was obtained from [124] and mapped to the *Ectocarpus* V5 reference genome using the nf-core/rnaseq v3.12.0 [121] with default parameters. TE expression levels were quantified using TEspeX [125], and transcript abundance for each TE family was calculated in transcripts per million (TPM), normalized by the length of the consensus sequence representing each family.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03742-z.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Supplementary Material 7: Figures S1-S7.

Supplementary Material 8: Tables S1-S19.

Acknowledgements

This work was supported by the Max Planck Society and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). HGD thanks Detlef Weigel for support and sponsorship. We thank Adrian Contreras, Andrea Movilli, Josué Barrera-Redondo, Jaruwatana Sodai Lotharukpong for valuable discussions. ED thanks the International Max Planck Research School 'From Molecules to Organisms'.

Peer review information

Leandro Quadrana and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

SMC and HGD conceived and led this study. SMC provided resources. CM performed sRNA experiments. ED, RJC, CM analysed the data. RJC and CM provided technical supervision. ED wrote the original draft of the manuscript. SMC, RJC, HGD, CM edited the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Max Planck Society, the European Research Council (grant 864038 to SMC), Marie Skłodowska-Curie grant agreement 101109906 to RJC, and the Moore Foundation (SMC). HGD was supported by a Royal Society Wolfson Fellowship (RSWF\R1\241004).

Data availability

The transposable element consensus sequences are available on Figshare (https://doi.org/10.6084/m9.figshare.29832329). The sRNA sequencing data can be accessed in the SRA Knowledge Base (https://www.ncbi.nlm.nih.gov/sra) with the accession number PRJNA1197354, which also corresponds to the associated BioProject https://www.ncbi.nlm.nih.gov/bioproject/1197354. Supplementary Material and scripts are available in GitHub at https://github.com/edinatale/Ectocarpus-TEs.

Declarations

Ethical approval
Not applicable to this study.

Competing interests
The authors declare no competing interests.

Received: 18 December 2024 Accepted: 12 August 2025
Published online: 29 September 2025

References

1. Bringloe TT, Starko S, Wade RM, Vieira C, Kawai H, Clerck OD, et al. Phylogeny and evolution of the brown algae. Crit Rev Plant Sci. 2020;39:281–321.
2. Denoeud F, Godfroy O, Cruaud C, Heesch S, Nehr Z, Tadrent N, et al. Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems. Cell. 2024;S0092867424012728.
3. Coelho SM, Gueno J, Lipinska AP, Cock JM, Umen JG. UV chromosomes and haploid sexual systems. Trends Plant Sci. 2018;23:794–807.
4. Barrera-Redondo J, Lipinska AP, Liu P, Dinatale E, Cossard G, Bogaert K, et al. Origin and evolutionary trajectories of brown algal sex chromosomes [Internet]. Evolutionary Biology; 2024 Jan. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.01.15.575685
5. Coelho SM. The brown seaweed Ectocarpus. Nat Methods. 2024;21:363–4.
6. Coelho SM, Umen J. Switching it up: algal insights into sexual transitions. Plant Reprod. 2021;34:287–96.
7. Coelho SM, Peters AF, Müller D, Cock JM. Ectocarpus: an evo-devo model for the brown algae. EvoDevo. 2020;11:19.
8. Cock J, Godfroy O, Strittmatter M, Scornet D, Uji T, Farnham G, et al. Emergence of *Ectocarpus* as a model system to study the evolution of complex multicellularity in the brown algae. 2015. p. 153–62.
9. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet. 2020;54:539–61.
10. Feschotte C, Pritham EJ. DNA Transposons and the Evolution of Eukaryotic Genomes. Annu Rev Genet. 2007;41:331–68.
11. Hayward A, Gilbert C. Transposable elements. Curr Biol. 2022;32:R904–9.
12. Bartolomé C, Charlesworth B. Evolution of amino-acid sequences and codon usage on the Drosophila miranda neo-sex chromosomes. Genetics. 2006;174:2033–44.
13. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 2008;18:359–69.
14. Ahmed S, Cock JM, Pessia E, Luthringer R, Cormier A, Robuchon M, et al. A haploid system of sex determination in the brown alga *Ectocarpus* sp. Curr Biol. 2014;24:1945–57.
15. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284:601–3.
16. Hurst GDD, Werren JH. The role of selfish genetic elements in eukaryotic evolution. Nat Rev Genet. 2001;2:597–606.
17. Biémont C. A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. Genetics. 2010;186:1085–93.
18. McClintock B. The Significance of Responses of the Genome to Challenge. Science. 1984;226:792–801.
19. Feschotte C. Transposable elements: McClintock's legacy revisited. Nat Rev Genet. 2023;24:797–800.
20. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet. 2014;15:394–408.
21. Sigman MJ, Slotkin RK. The First Rule of Plant Transposable Element Silencing: Location, Location. Location Plant Cell. 2016;28:304–13.
22. Onishi R, Yamanaka S, Siomi MC. piRNA- and siRNA-mediated transcriptional repression in Drosophila, mice, and yeast: new insights and biodiversity. EMBO Rep. 2021;22: e53062.
23. Allen SE, Nowacki M. Roles of Noncoding RNAs in Ciliate Genome Architecture. J Mol Biol. 2020;432:4186–98.
24. Igolkina AA, Zinkevich A, Karandasheva KO, Popov AA, Selifanova MV, Nikolaeva D, et al. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. Cells. 2019;8:1034.
25. Millán-Zambrano G, Burton A, Bannister AJ, Schneider R. Histone post-translational modifications — cause and consequence of genome function. Nat Rev Genet. 2022;23:563–80.
26. Hisanaga T, Romani F, Wu S, Kowar T, Wu Y, Lintermann R, et al. The Polycomb repressive complex 2 deposits H3K27me3 and represses transposable elements in a broad range of eukaryotes. Curr Biol. 2023;33:4367–4380.e9.
27. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, et al. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature. 2010;465:617–21.
28. Bourdareau S, Tirichine L, Lombard B, Loew D, Scornet D, Coelho SM, et al. Histone modifications during the life cycle of the brown alga *Ectocarpus*. bioRxiv. 2020;2020.03.09.980763.
29. Gueno J, Borg M, Bourdareau S, Cossard G, Godfroy O, Lipinska A, et al. Chromatin landscape associated with sexual differentiation in a UV sex determination system. Nucleic Acids Res. 2022;50:3307–22.
30. Vigneau J, Martinho C, Godfroy O, Zheng M, Haas FB, Borg M, et al. Interactions between U and V sex chromosomes during the life cycle of *Ectocarpus*. Development. 2024;dev.202677.
31. Bourdareau S, Tirichine L, Lombard B, Loew D, Scornet D, Wu Y, et al. Histone modifications during the life cycle of the brown alga Ectocarpus. Genome Biol. 2021;22:12.

32. Tarver JE, Cormier A, Pinzón N, Taylor RS, Carré W, Strittmatter M, et al. microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga *Ectocarpus*. Nucleic Acids Res. 2015;43:6384–98.
33. Liu P, Vigneau J, Craig RJ, Barrera-Redondo J, Avdievich E, Martinho C, et al. 3D chromatin maps of a brown alga reveal U/V sex chromosome spatial organization. Nat Commun. 2024;15:9590.
34. Quesneville H, Nouaud D, Anxolabehere D. Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. J Mol Evol. 2003;57:S50–9.
35. Peona V, Martelossi J, Almojil D, Bocharkina J, Brännström I, Brown M, et al. Teaching transposon classification as a means to crowd source the curation of repeat annotation – a tardigrade perspective. Mob DNA. 2024;15:10.
36. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5:103–7.
37. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
38. Thomas J, Pritham EJ. *Helitrons* , the Eukaryotic Rolling-circle Transposable Elements. Chandler M, Craig N, editors. Microbiol Spectr. 2015;3:3.4.03.
39. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2019;94:233–52.
40. Chaparro C, Gayraud T, De Souza RF, Domingues DS, Akaffou S, Laforga Vanzela AL, et al. Terminal-Repeat Retrotransposons with GAG Domain in Plant Genomes: A New Testimony on the Complex World of Transposable Elements. Genome Biol Evol. 2015;7:493–504.
41. Gorinšek B, Gubenšek F, Kordiš D. Evolutionary Genomics of Chromoviruses in Eukaryotes. Mol Biol Evol. 2004;21:781–98.
42. Ustyantsev K, Blinov A, Smyshlyaev G. Convergence of retrotransposons in oomycetes and plants. Mob DNA. 2017;8:4.
43. Lu Y, Tan F, Zhao Y, Zhou S, Chen X, Hu Y, et al. A Chromodomain-Helicase-DNA-Binding Factor Functions in Chromatin Modification and Gene Regulation. Plant Physiol. 2020;183:1035–46.
44. Choi S-W, Graf L, Choi JW, Jo J, Boo GH, Kawai H, et al. Ordovician origin and subsequent diversification of the brown algae. Curr Biol. 2024;34:740–754.e4.
45. De Souza RF, Aravind L. Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. Mol BioSyst. 2012;8:1661.
46. Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. An Ancient Clade of *Penelope* -Like Retroelements with Permuted Domains Is Present in the Green Lineage and Protists, and Dominates Many Invertebrate Genomes. Townsend J, editor. Molecular Biology and Evolution. 2021;38:5005–20.
47. Macaisne N, Liu F, Scornet D, Peters AF, Lipinska A, Perrineau M-M, et al. The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development. 2017;144:409.
48. Han M-J, Xiong C-L, Zhang H-B, Zhang M-Q, Zhang H-H, Zhang Z. The diversification of PHIS transposon superfamily in eukaryotes. Mob DNA. 2015;6:12.
49. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
50. Iyer LM, Abhiman S, Aravind L. Natural History of Eukaryotic DNA Methylation Systems. Progress in Molecular Biology and Translational Science [Internet]. Elsevier; 2011 [cited 2025 May 24]. p. 25–104. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123876850000020
51. Petroll R, West JA, Ogden M, McGinley O, Craig RJ, Coelho SM, et al. The expanded Bostrychia moritziana genome unveils evolution in the most diverse and complex order of red algae. Current Biology. 2025;S0960982225005081.
52. Luthringer R, Raphalen M, Guerra C, Colin S, Martinho C, Zheng M, et al. Repeated co-option of HMG-box genes for sex determination in brown algae and animals. Science. 2024;383:eadk5466.
53. Luthringer R, Lipinska AP, Roze D, Cormier A, Macaisne N, Peters AF, et al. The pseudoautosomal regions of the U/V sex chromosomes of the brown alga *Ectocarpus* exhibit unusual features. Mol Biol Evol. 2015;32:2973–85.
54. Avia K, Lipinska AP, Mignerot L, Montecinos AE, Jamy M, Ahmed S, et al. Genetic diversity in the UV sex chromosomes of the brown alga *Ectocarpus*. Genes (Basel). 2018;9.
55. Munoz-Lopez M, Garcia-Perez J. DNA Transposons: Nature and Applications in Genomics. CG. 2010;11:115–28.
56. Skipper KA, Andersen PR, Sharma N, Mikkelsen JG. DNA transposon-based gene vehicles - scenes from an evolutionary drive. J Biomed Sci. 2013;20:92.
57. Kazazian HH. Mobile Elements: Drivers of Genome Evolution. Science. 2004;303:1626–32.
58. Wu X, Tronholm A, Cáceres EF, Tovar-Corona JM, Chen L, Urrutia AO, et al. Evidence for Deep Phylogenetic Conservation of Exonic Splice-Related Constraints: Splice-Related Skews at Exonic Ends in the Brown Alga Ectocarpus Are Common and Resemble Those Seen in Humans. Genome Biol Evol. 2013;5:1731–45.
59. Craig RJ, Hasan AR, Ness RW, Keightley PD. Comparative genomics of *Chlamydomonas*. Plant Cell. 2021;33:1016–41.
60. Barrera-Redondo J, Lotharukpong JS, Drost H-G, Coelho SM. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol. 2023;24:54.
61. Drost H-G, Sanchez DH. Becoming a Selfish Clan: Recombination Associated to Reverse-Transcription in LTR Retrotransposons. Eyre-Walker A, editor. Genome Biology and Evolution. 2019;11:3382–92.
62. Fricker AD, Peters JE. Vulnerabilities on the Lagging-Strand Template: Opportunities for Mobile Elements. Annu Rev Genet. 2014;48:167–86.
63. Wells JN, Chang N-C, McCormick J, Coleman C, Ramos N, Jin B, et al. Transposable elements drive the evolution of metazoan zinc finger genes. Genome Res. 2023;33:1325–39.
64. Quesneville H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mob DNA. 2020;11:28.

65. Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in Drosophila. Mobile. DNA. 2020;11:23.
66. Zhang H-H, Peccoud J, Xu M-R-X, Zhang X-G, Gilbert C. Horizontal transfer and evolution of transposable elements in vertebrates. Nat Commun. 2020;11:1362.
67. Bohne A, Zhou Q, Darras A, Schmidt C, Schartl M, Galiana-Arnoux D, et al. Zisupton–A Novel Superfamily of DNA Transposable Elements Recently Active in Fish. Mol Biol Evol. 2012;29:631–45.
68. Iyer LM, Zhang D, De Souza RF, Pukkila PJ, Rao A, Aravind L. Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. Proc Natl Acad Sci USA. 2014;111:1676–83.
69. López-Cortegano E, Craig RJ, Chebib J, Balogun EJ, Keightley PD. Rates and spectra of de novo structural mutations in *Chlamydomonas reinhardtii*. Genome Res. 2023;33:45–60.
70. Arkhipova IR, Yushenova IA. Giant Transposons in Eukaryotes: Is Bigger Better? Angert E, editor. Genome Biol Evol. 2019;11:906–18.
71. Duhamel M, Hood ME, Rodríguez De La Vega RC, Giraud T. Dynamics of transposable element accumulation in the non-recombining regions of mating-type chromosomes in anther-smut fungi. Nat Commun. 2023;14:5692.
72. Kubat Z, Zluvova J, Vogel I, Kovacova V, Cermak T, Cegan R, et al. Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. New Phytol. 2014;202:662–78.
73. Steflova P, Tokan V, Vogel I, Lexa M, Macas J, Novak P, et al. Contrasting Patterns of Transposable Element and Satellite Distribution on Sex Chromosomes (XY1Y2) in the Dioecious Plant Rumex acetosa. Genome Biol Evol. 2013;5:769–82.
74. Pasquesi GIM, Perry BW, Vandewege MW, Ruggiero RP, Schield DR, Castoe TA. Vertebrate Lineages Exhibit Diverse Patterns of Transposable Element Regulation and Expression across Tissues. Genome Biol Evol. 2020;12:506–21.
75. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2017;1860:157–65.
76. Torres DE, Thomma BPHJ, Seidl MF. Transposable Elements Contribute to Genome Dynamics and Gene Expression Variation in the Fungal Plant Pathogen *Verticillium dahliae*. Yi S, editor. Genome Biology and Evolution. 2021;13:evab135.
77. Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun. 2014;5:5495.
78. Fouché S, Oggenfuss U, Chanclud E, Croll D. A devil's bargain with transposable elements in plant pathogens. Trends Genet. 2022;38:222–30.
79. Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. Transposable Elements Are Important Contributors to Standing Variation in Gene Expression in Capsella Grandiflora. Parsch J, editor. Molecular Biology and Evolution. 2019;36:1734–45.
80. Liu P, Cuerda-Gil D, Shahid S, Slotkin RK. The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. Annu Rev Genet. 2022;56:63–87.
81. Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, et al. Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. Cell. 2009;136:461–72.
82. Erdmann RM, Picard CL. RNA-directed DNA Methylation. PLoS Genet. 2020;16: e1009034.
83. Cheng J, Martinez G. Enjoy the silence: Canonical and non-canonical RNA silencing activity during plant sexual reproduction. Curr Opin Plant Biol. 2024;82: 102654.
84. Tang Y, Yan X, Gu C, Yuan X. Biogenesis, Trafficking, and Function of Small RNAs in Plants. Front Plant Sci. 2022;13: 825477.
85. Fischer SEJ. Activity and Silencing of Transposable Elements in C. elegans. DNA. 2024;4:129–40.
86. Mao H, Zhu C, Zong D, Weng C, Yang X, Huang H, et al. The Nrde Pathway Mediates Small-RNA-Directed Histone H3 Lysine 27 Trimethylation in Caenorhabditis elegans. Curr Biol. 2015;25:2398–403.
87. Buckley BA, Burkhart KB, Gu SG, Spracklin G, Kershner A, Fritz H, et al. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. Nature. 2012;489:447–51.
88. Parent J-S, Cahn J, Herridge RP, Grimanelli D, Martienssen RA. Small RNAs guide histone methylation in *Arabidopsis* embryos. Genes Dev. 2021;35:841–6.
89. Yang F, Xi R. Silencing transposable elements in the Drosophila germline. Cell Mol Life Sci. 2017;74:435–48.
90. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022;13:7.
91. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
92. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.
93. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. 2018;176:1410–22.
94. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
95. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.
96. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
97. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
98. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005;33:W116–120.
99. Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics. 2020;72: e108.

100. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430:2237–43.
101. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
102. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015;32:268–74.
103. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
104. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Mol Biol Evol. 2013;30:1188–95.
105. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
106. Rodriguez F, Arkhipova IR. An Overview of Best Practices for Transposable Element Identification, Classification, and Annotation in Eukaryotic Genomes. In: Branco MR, De Mendoza Soler A, editors. Transposable Elements [Internet]. New York, NY: Springer US; 2023 [cited 2024 Nov 23]. p. 1–23. Available from: https://link.springer.com/10.1007/978-1-0716-2883-6_1
107. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
108. Bailly-Bechet M, Haudry A, Lerat E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA. 2014;5:13.
109. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
110. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Hancock J, editor. Bioinformatics. 2017;33:3088–90.
111. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
112. Lipinska, Cormier A, Luthringer R, Peters AF, Corre E, Gachon CMM, et al. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *Ectocarpus*. Mol Biol Evol. 2015;32:1581–97.
113. Coelho SM, Scornet D, Rousvoal S, Peters NT, Dartevelle L, Peters AF, et al. How to cultivate *Ectocarpus*. Cold Spring Harb Protoc. 2012;2012:258–61.
114. Martinho C, Wang Z, Ghigi A, Buddle S, Barbour F, Yarur A, et al. CHROMOMETHYLTRANSFERASE3/KRYPTONITE maintains the *sulfurea* paramutation in *Solanum lycopersicum*. Proc Natl Acad Sci USA. 2022;119: e2112240119.
115. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.
116. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. 2011;10–2.
117. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
118. Zytnicki M, Gaspin C. srnaMapper: an optimal mapping tool for sRNA-Seq reads. BMC Bioinformatics. 2022;23:495.
119. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
120. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
121. Harshil Patel, Jose Espinosa-Carrasco, Chuan Wang, Phil Ewels, nf-core bot, Tiago Chedraoui Silva, et al. nf-core/chipseq: nf-core/chipseq v2.1.0 - Platinum Willow Sparrow [Internet]. Zenodo; 2024 [cited 2025 May 24]. Available from: https://zenodo.org/doi/10.5281/zenodo.3240506
122. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–191.
123. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
124. Lotharukpong JS, Zheng M, Luthringer R, Liesner D, Drost H-G, Coelho SM. A transcriptomic hourglass in brown algae. Nature. 2024;635:129–35.
125. Ansaloni F, Gualandi N, Esposito M, Gustincich S, Sanges R. TEspeX: consensus-specific quantification of transposable element expression preventing biases from exonized fragments. Mathelier A, editor. Bioinformatics. 2022;38:4430–3.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.