THE RHODOEXPLORER PLATFORM FOR RED ALGAL GENOMICS AND WHOLE GENOME ASSEMBLIES FOR SEVERAL GRACILARIA SPECIES

Agnieszka P. Lipinska[1,3#,*], Stacy A. Krueger-Hadfield[2,#,*,!], Olivier Godfroy[3], Simon Dittami[3], Lígia Ayres-Ostrock[4,5], Guido Bonthond[6], Loraine Brillet-Guéguen[3,7], Susana Coelho[1], Erwan Corre[7], Guillaume Cossard[1], Christophe Destombe[8], Paul Epperlein[1], Sylvain Faugeron[8,9], Elizabeth Ficko-Blean[3], Jessica Beltrán[8,9], Emma Lavaut[8], Arthur Le Bars[7,10], Fabiana Marchi[4], Stéphane Mauger[8], Gurvan Michel[3], Philippe Potin[3] , Delphine Scornet[3], Erik E. Sotka[11], Florian Weinberger[12], Mariana Cabral de Oliveira[4], Marie-Laure Guillemin[8,13,14], Estela M. Plastino[4], Myriam Valero[8]

[1] Department of Algal Development and Evolution, Max Planck Institute for Biology Tubingen, Tubingen, Germany

[2] Department of Biology, University of Alabama at Birmingham, 1300 University Blvd, Birmingham, AL, 35294

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

[4] Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, Rua do Matão 277, Cidade Universitária 05508-090, São Paulo, SP, Brasil.

[5] Hortimare - Breeding & Propagating Seaweed. Altonstraat 25A 1704 CC Heerhugowaard. The Netherlands

[6] Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University Oldenburg, Schleusenstrasse 1, 26382, Wilhelmshaven, Germany

[7] CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, 29680, Roscoff, France

[8] CNRS, Sorbonne Université, Pontificia Universidad Católica de Chile, Universidad Austral de Chile, IRL 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, CS 90074, F-29688 Roscoff, France

[9] Núcleo Milenio MASH, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile

[10] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

[11] Department of Biology, College of Charleston, Charleston SC 29412

[12] GEOMAR Helmholtz-Zentrum für Ozeanforschung, Marine Ecology Division, Düsternbrooker Weg 20, 24105 Kiel, Germany

[13] Núcleo Milenio MASH, Facultad de Ciencias, Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Casilla 567, Valdivia, Chile

[14] Centro FONDAP de Investigación de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile

[#] Shared first authors

! Current address: Virginia Institute of Marine Science, Eastern Shore Laboratory, 40 Atlantic Ave, Wachapreague, VA 23480, sakh@vims.edu

* Authors for correspondence: **Agnieszka P. Lipinska**, Department of Algal Development and Evolution, Max Planck Institute for Developmental Biology, Tuebingen, Germany, +49-7071-601-1370, alipinska@tuebingen.mpg.de; **Stacy A. Krueger-Hadfield**, Department of Biology, University of Alabama at Birmingham, 1300 University Blvd, Birmingham, AL, 35294, +1-205-934-6034, sakh@uab.edu & Virginia Institute of Marine Science, Eastern Shore Laboratory, 40 Atlantic Ave, Wachapreague, VA 23480, sakh@vims.edu

ABSTRACT

Macroalgal (seaweed) genomic resources are generally lacking as compared to other eukaryotic taxa, and this is particularly true in the red algae (Rhodophyta). Understanding red algal genomes is critical to understanding eukaryotic evolution given that red algal genes are spread across eukaryotic lineages from secondary endosymbiosis and red algae diverged early in the Archaeplastids. The Gracilariales are highly diverse and widely distributed order whose species can serve as ecosystem engineers in intertidal habitats, including several notorious introduced species. The genus *Gracilaria* is cultivated worldwide, in part for its production of agar and other bioactive compounds with downstream pharmaceutical and industrial applications. This genus is also emerging as a model for algal evolutionary ecology. Here, we report new whole genome assemblies for two species (*G. chilensis* and *G. gracilis*), a draft genome assembly of *G. caudata*, and genome annotation of the previously published *G. vermiculophylla* genome. To facilitate accessibility and comparative analysis, we integrated these data in a newly created web-based portal dedicated to red algal genomics (https://rhodoexplorer.sb-roscoff.fr). These genomes will provide a resource for understanding algal biology and, more broadly, eukaryotic evolution.

KEYWORDS

evolution, ecology, omics, ploidy, Rhodophyta

SIGNIFICANCE STATEMENT

The Gracilariales are an ecologically and economically important red algal order found throughout the coastal regions of the world. Understanding the biology, ecology, and evolution

of species in this order, and that of red algae more broadly, has been hampered by the limited phylogenetic coverage of genomic resources. Here, we present whole genome assemblies and gene annotations for four *Gracilaria* species that will serve as a key resource for algal research on evolution, ecology, biotechnology and aquaculture.

INTRODUCTION

Red algae (Rhodophyta) represent a lineage of photosynthetic eukaryotes in the Archaeplastids that diverged from green algae around 1700 MYA (Yang et al. 2016). Within the Rhodophyta, the Cyanidiophyceae were the earliest to diverge approximately 1200 MYA, while the Florideophyceae diverged more recently (i.e., 412 MYA; Yang et al. 2016) and constitute the most speciose group (Graham et al. 2016). In this context, the genomic resources currently available (Table S1) represent only a fraction of the evolutionary diversity of red algae, limiting our capacity to reconstruct the evolutionary history of the unique features of this group.

The Florideophyceae have a life cycle in which haploid male and female gametophytes alternate with a diploid tetrasporophyte (Figure S1). After fertilization, the zygote develops on the female into a cystocarp, in which the zygote is mitotically copied. Male gametes (spermatia) and spores are non-motile, and the female gamete (carpogonium) is retained on the female thallus. The cystocarp was thought to have evolved in response to low fertilization success (Searles 1980), but recent work has shown that many males fertilize a single female (Engel et al. 1999, Krueger-Hadfield et al. 2015) and that animal-mediated fertilization can increase reproductive success (Lavaut et al. 2022). Many species have ‘isomorphic’ gametophytes and tetrasporophytes, which are hard to discern without the aid of molecular tools (e.g., sex-linked

92 markers, Martinez et al. 1999; Guillemin et al. 2012; or microsatellites, Krueger-Hadfield et al.
93 2016).

94 Here, we focus on four *Gracilaria*[1] species spanning roughly 170 million years of
95 evolution (Lyra et al. 2021). These species were chosen based on their evolutionary, ecological,
96 and/or economic importance. Species in the genus *Gracilaria* produce agars in their cell wall
97 (Popper et al. 2011), they can be propagated vegetatively, and serve as ecosystem engineers in
98 intertidal zone (Kain and Destombe 1995). The four taxa chosen can be divided into three clades
99 based on their molecular divergence: (i) *G. chilensis* and *G. vermiculophylla*, (ii) *G. caudata*, and
100 (iii) *G. gracilis* (Lyra et al. 2021). *Gracilaria gracilis* and *G. caudata* are evolutionarily more
101 distinct than the phylogenetic group that contains *G. chilensis* and *G. vermiculophylla*.
102 *Gracilaria chilensis* C.J. Bird et al. is an important crop along the Chilean coastline, where it has
103 been both harvested and subsequently planted after a crash in natural stands likely due to
104 overharvesting (Buschmann et al. 2001). The artificial selection for tetrasporophytes has resulted
105 in early stages of domestication (Valero et al. 2017) and loss of sexual reproduction (Guillemin
106 et al. 2008). *Gracilaria vermiculophylla* (Ohmi) Papenfuss is a successful invader in many of the
107 bays and estuaries of North America, northwestern Africa, and Europe (Krueger-Hadfield et al.
108 2017). The invasion success was likely facilitated by adaptive shifts in temperature and salinity
109 tolerance (e.g., Sotka et al. 2018) and to biofoulers (e.g., Bonthond et al. 2020), as well as the
110 ability to asexually fragment (Krueger-Hadfield et al. 2016). *Gracilaria caudata* J. Agardh can
111 form dense stands in the intertidal zone (Plastino and Oliveira 1997) and has been subjected to
112 intense harvesting pressure, leading to declines in native populations (Hayashi et al. 2014, see
113 also Ayres-Ostrock et al. 2019). For this species, we re-analyzed the genome published by

[1] There is controversy over the systematics of *Gracilaria* Greville, but for the purposes of this paper, we consider the four species as belonging to the genus *Gracilaria* (*sensu* Lyra et al. 2021, Guiry and Guiry 2022).

Flanagan et al. (2021). Finally, *Gracilaria gracilis* (Stackhouse) Steentoft, L.M. Irvine & Farnham is a long-lived species that inhabits tidepools along European coastlines. This species serves as model species to test hypotheses related to the evolution of sex (e.g., alternation of haploid and diploid phases in life cycles, Destombe et al. 1989, 1992, 1993, Hughes and Otto 1999; mating system and sexual selection, Richerd et al. 1993, Engel et al. 2002).

The availability of genomic and genetic resources for these four *Gracilaria* species should aid in our understanding of the evolutionary ecology of red algae in their dynamic environment, during invasions of new habitats, under cultivation practices, and in response to climate change. Moreover, these new resources will add to the existing genomic data and illuminate key processes in eukaryotic evolution. The Rhodoexplorer – Red Algal Genome Database currently includes the *Gracilaria* species discussed here but will include all the high-quality genomic resources available for the Rhodophyta (e.g., genomes, transcriptomes), thereby providing a unique resource for comparative analyses.

RESULTS AND DISCUSSION

Genome Assembly

Final genome assembly sizes, based on long and short read sequencing, ranged from 76 to 80 Mbp for *G. gracilis* and *G. chilensis,* respectively. In addition, we created a draft genome assembly based on the Illumina sequencing only for *G. caudata* (32 Mbp) and reassembled the genome of *G. vermiculophylla* (Flanagan et al. 2021) to a final 47 Mbp after bacterial contamination removal. The above genome assemblies were comparable to the genomes of *G. domingensis* (78 Mbp, Nakamura-Gouvea et al. 2022) and *G. changii* (36 Mbp, Ho et al. 2017). PacBio assemblies of *G. chilensis* and *G. gracilis* produced in this study (< 300 contigs per

genome) are the most contiguous red macroalgal genomes presently available in public databases, apart from *G. vermiculophylla* and *P. yezoensis* where the addition of a HiC library enabled scaffolding nearly at the chromosome level (Wang et al. 2020, Flanagan et al. 2021). Despite the differences in assembly size, BUSCO scores were similar across the long read-sequenced *G. gracilis* and *G. chilensis*, and the more fragmented *G. caudata* genome, with 81.6 to 83.6% of conserved proteins present (Eukaryota_odb10, Manni et al. 2021, Simão et al. 2015; Table 1). The re-assembled genome of *G. vermiculophylla* contained 71.8% of the conserved proteins. Given the diversity of Rhodophyta and the lack of lineage-specific databases, these results are in the expected range. A recent study estimated the presence of conserved eukaryotic genes (Eukaryota_odb10) in red algal genomes at a median level of 69% (Hanschen et al. 2020).

Red algal genomes are repeat-rich, with half or more of their genomic sequence being constituted by repetitive elements, as reported previously for *Porphyra umbilicalis* (43.9%, Brawley et al. 2017), *Pyropia yezoensis* (48%, Wang et al. 2020) and *Chondrus crispus* (73%, Collen et al. 2013). In agreement with this general trend, between 45.7-66.2% of the *Gracilaria* genomes corresponded to repetitive elements (Figure 1 and Table 1).

Gene prediction and Annotation

Gene prediction yielded a total of 8,042, 9,065 and 9,674 coding sequences for *G. chilensis*, *G. caudata* and *G.gracilis* (Table 1), which was comparable with other red macroalgal genomes, *Chondrus crispus* (9,815 genes, Collen et al. 2013) and *Gracilaria changii* genome (10,912 genes, Ho et al. 2022). In addition, we annotated the reassembled genome of *G. vermiculophylla*, which yielded fewer genes (7,048). Among these genes, 70.6-76.6% did not contain any introns, as typical for the compact genomes of red algae (Qiu et al. 2015). Most

Gracilaria genes had homologous sequences in the Uniprot database (84.2-89.7%) and were annotated with at least one INTERPRO hit (91.7-93.6%). Between 47.9% and 54.4% of genes were associated with GO annotations.

Orthofinder analyses enabled us to identify 4,666 core groups of orthologous proteins present in all four of the sequenced genomes (Figure 1B) versus 408-620 orthogroups or orphan genes specific to only one of the sequenced species (Figure 1B). Among the species-specific sequences, the rate of GO annotation was lower than for the entire dataset, ranging from 12.7% for *G. chilensis* to 18.2% for *G. caudata*. The fact that the two species *G. caudata* and *G. gracilis* share more genes between them than with the phylogenetic group of *G. chilensis* and *G. vermiculophylla* was expected due to divergence between the two clades of *Gracilaria* species. Both the annotated and the unknown species-specific genes constitute attractive targets to study their role in adaptation and speciation.

Rhodoexplorer – Red Algal Genome Database

In addition to depositing the raw reads and sequenced genome in a public repository (see Data Availability section), all four genomes were also integrated into the newly created Rhodoexplorer Red Algal Genome Database (https://rhodoexplorer.sb-roscoff.fr), hosted at the ABiMS bioinformatic platform. This platform will gradually include more red algal genomes in the future. The services provided include:

- Information about the sequenced strains, with links to external databases (NCBI, WoRMS, Algaebase)
- Assembly and annotation metrics

- Data downloads: genomic, genes and proteomic datasets, structural and functional annotations, orthology clusters, etc.
- A BLAST interface with a selection of red algal genomes, predicted and *de novo* assembled transcriptomes and proteomes.
- Visualization tools: a genome browser to visualize the predicted genes and the RNAseq data mapped on the genome and a web interface to visualize functional annotations and retrieve individual protein sequences.

MATERIALS AND METHODS

Sampling of the biological material

Adult female and male *Gracilaria* thalli, all bearing reproductive structures, used for the sequencing were collected from natural populations: *G. chilensis* in Lenca (Chile, -41.607, -72.692), *G. vermiculophylla* in Charleston, SC (USA, 32.752, -79.900), *G. caudata* in Paracuru, CE (Brazil, -3.399, -39.012), and *G. gracilis* in Cape Gris-Nez (France, 50.872, 1.584). *Gracilaria caudata* and *G. chilensis* were maintained as clonal, unialgal cultures under laboratory conditions prior to nucleic acid extractions (see *Culture conditions*). Field-collected *G. gracilis* and *G. vermiculophylla* thalli were transported to the laboratory, examined under a microscope, and cleaned of contaminants. If visible, cystocarps were excised prior to preservation of the thalli at -80°C before further use. Table S2 provides details of the *Gracilaria* species used in this study.

Culture conditions

Cultures were initiated either from lab crosses or from tetraspores released by field-collected tetrasporophytes. *Gracilaria caudata* was grown in the modified von Stosch nutrient solution (Ursi and Plastino 2001) diluted to 25% in seawater (32 psu), with weekly renewals. The algae were kept in culture chambers at 25ºC under fluorescent illumination of 70 µmol.m-2.s-1 14h photoperiod, following previously established optimal growth conditions (Yokoya and Oliveira 1992a,b). *Gracilaria chilensis* was grown in Provasoli medium (McLachlan 1973), changed weekly during the first two months and twice a week thereafter. Cultures were kept at 13°C under 40-60 µmol.m-2.s-1 of light with 12h day length.

Nucleic acid extraction, library preparation, and sequencing

Genomic DNA was extracted from mature male gametophytes using DNeasy PowerPlant Pro Kit for *G. caudata* or an in-house protocol based on Faugeron et al. (2001) for *G. chilensis* and *G. gracilis*. The concentration and purity of DNA were measured with NanoDrop and Qubit before sequencing on an Illumina HiSeq 2500 (125 bp PE reads for *G. chilensis* and *G. gracilis*; 100bp PE reads for *G. caudata*) or PacBio Sequel II with sheared gDNA large insert library (*G. gracilis* and *G. chilensis*) (Table S2).

For genome annotation, total RNA was extracted from mature thalli of male and female gametophytes of *G. chilensis* (2 males and 2 females), *G. caudata* (4 males and 4 females), and *G. gracilis* (1 male and 1 female) using the Rneasy Mini Plant Kit (Qiagen) following the manufacturer's instructions. Total RNA was extracted from *G. vermiculophylla* (4 males and 4 females) using the Macherey Nagel Nucleospin RNA for Plant kit following the manufacturer's instructions. Paired-end 150bp Illumina reads were generated with Illumina HiSeq 2500 Table S2).

Genome assembly

De novo genome assemblies for *G. gracilis* and *G. chilensis* were generated based on 203-fold and 116-fold coverage of PacBio long reads, respectively. Briefly, bacterial sequences were removed from raw data (subreads) using Blobtools v1.1.1 (Laetsch and Blaxter 2017). For each species, two independant assemblies were generated using CANU (Koren et al., 2017) and FLYE (Kolmogorov et al., 2019). Based on congruity (QUAST v.5.0.2 – Mikheenko, et al., 2018) and BUSCO score (Simão FA, et al. 2015) the best assemby was kept and polished using three iterations of RACON v.1.4.20. Finally, PacBio sequencing error were corrected using 150bp paired-end Illumina reads with PILON v.1.23 software (Walker et al. 2014). The draft genome assembly of *G. caudata*. was generated using 171-fold coverage of 150bp paired-end Illumina reads only. First, a meta-genome was produced using metaSPAdes v3.12.0 (Nurk et al. 2017) and bacterails contigs were detected using Blobtools. Reads corresponding to eukaryotic contigs were then assembled using SPAdes v3.12.0 (Bankevich et al. 2012). Quality of all *de novo* genome assemblies was assessed with QUAST and DNAseq remapping for congruity and BUSCO and RNAseq mapping for completness.

For *G. vermiculophylla*, we updated the existing chromosome-scale genome assembly (Flanagan et al. 2021) by reassembling the Illumina reads using SPAdes v 3.12.0 (Bankevich et al. 2012) and scaffolding with Hi-C libraries, following the Dovetail Genomics proprietary pipeline (Elbers et al. 2019). This process ameliorated the genome continuity (N50 increased from 2.06Mb to 2.68Mb) and completeness (BUSCO score increased from 57,6% to 65,9% of complete genes using the Eukaryota_odb10 dataset).

We used Blobtools v1.1.1 (Laetsch and Blaxter 2017) with maximal accuracy settings to validate the quality of the four *Gracilaria* genome assemblies and identify potential bacterial contaminations. In brief, DNAseq reads of each species were first mapped to their corresponding reference genome using HISAT2 v 2.2.1 (Kim et al. 2019). Next, BAM coverage files produced by HISAT2, Diamond blastx v2.0.11 (Buchfink et al. 2015, 2021) hit-file against non-redundant protein sequences archive from NCBI (-sensitive, –max-target-seqs 1, -e-value 1e − 20), and Blast v2.12.0 (Camacho et al. 2009) output against nucleotide archive from NCBI (-max_target_seqs 10 -max_hsps 1 -evalue 1e − 20) were used as input for Blobtools. Genomic scaffolds classified as bacterial or with a coverage of less than 1 (sum of coverages for each sequence across all coverage files) were removed from the assembly. Genome assembly completeness was assessed using BUSCO scores with the eukaryotic data set (Eukaryota_odb10, Simão et al. 2015, Manni et al. 2021).

Chloroplastic and mitochondrial genomes of each species were reconstructed from Illumina raw reads using NovoPlasty (Dierckxsens et al. 2016) throught the European Galaxy web portal (https://usegalaxy.eu/). Annotation of those *de novo* organellar genomes were done using the GeSeq web tool (Tillich et al. 2017 – https://chlorobox.mpimp-golm.mpg.de/geseq.html). Public sequences from *Gracilaria caudata* voucher SPF:57390 (NC_039146, NC_039139), *Gracilaria chilensis* voucher CNU050183 (KP728466, KT266788), *Gracilaria gracilis* voucher SPF:55734 (NC_039141, NC_039148) and *Gracilaria vermiculophylla* (MN853882, MH396022) were retrieved from NCBI and used as seeds and references for both assembly and annotation.

Genome annotation

Each reference genome was first masked using RepeatMasker v4.0.9 (Smit et al. 2015) with Dfam v3.0 database (Wheeler et al. 2013) and a customized repeat library produced from concatenated outputs of RepeatScout v1.0.6 (Price et al. 2005) and TransposonPSI v1.0.0 (Hass 2007-2011). Initial quality assessment of the RNA-Seq reads was performed with FastQC v0.11.9 (Andrews et al. 2010), and reads were trimmed using Trimmomatic v0.39 (TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50; Bolger et al. 2014). Clean reads were mapped to the reference genome assembly using HISAT2 v 2.2.1 (Kim et al. 2019). The resulting alignment files were used to annotate protein-coding genes with BRAKER2 v2.1.6 pipeline (Bruna et al. 2021). Functional annotation of the reference transcriptomes was performed using eggNOG-mapper (Huerta-Cepas et al. 2019, Cantalapiedra et al. 2021).

All code used for genomes assembly and annotation is available on the Gitpage dedicated to the genome database project https://abims-sbr.gitlab.io/rhodoexplorer/doc/data_process/.

Rhodoexplorer – Red Algal Genome Database

The main web portal (https://rhodoexplorer.sb-roscoff.fr) has been implemented using the Python web framework Django, with data stored in a relational database (PostgreSQL).

For each red algal species, an integrated environment of visualization tools has been deployed based on the Galaxy Genome Annotation (GGA) project (Bretaudeau et al. 2019). Each GGA environment deployed for the Rhodoexplorer genome database includes: Chado – a PostgreSQL relational database schema for storing biological data (Mungall et al. 2007); JBrowse – a web-based genome browser (Buels et al. 2016); Tripal – a Drupal-based application for creating biological websites (Sanderson et al. 2013); Elasticsearch – a distributed, free and

295 open search and analytics engine for all types of data
296 (https://www.elastic.co/products/elasticsearch); Galaxy – a browser accessible workbench for
297 scientific computing used as a data loading orchestrator for administrators (The Galaxy
298 Community 2022). To facilitate the deployment and the administration of the GGA service, a set
299 of Python tools has been developed (http://gitlab.sb-roscoff.fr/abims/e-infra/gga_load_data)
300 allowing mass deployment of Docker containers and automated data loading through Galaxy
301 with the Bioblend API (Sloggett et al 2013).
302 The BLAST interface (https://blast.sb-roscoff.fr/rhodoexplorer/) includes an
303 implementation of the BLAST algorithm using SequenceServer (Priyam et al. 2019) graphical.
304 The documentation website for navigating the platform web portal and resources
305 (https://abims-sbr.gitlab.io/rhodoexplorer/doc/) is published from a GitLab repository, with
306 Pages and MkDocs, a static site generator.
307 The entire informatic infrastructure is deployed and maintained on the ABiMS
308 Bioinformatics platform of the Roscoff Biological Station, part of the national infrastructure
309 French Bioinformatic Institute.
310
311 *SUPPLEMENTARY MATERIAL*
312 Supplementary Figure S1: Life cycle of *Gracilaria.*
313 Supplementary Table S1: Available red algal genomic resources.
314 Supplementary Table S2: Species used in this study.

315 *ACKNOWLEDGEMENTS*

316 This project was supported by start-up funds from the College of Arts and Sciences at the
317 University of Alabama at Birmingham to SAKH; ANID NCN2021-033 and FONDECYT

1221456 and 1221477 to MLG, JB and SF; the International Research Networks DEBMA "Diversity, Evolution and Biotechnology of Marine Algae" (CNRS GDRI 0803) and DABMA "Diversity, Adaptation, and Biotechnology of Marine Algae" (CNRS IRN 00022); the ERC (grant number 864038 to SMC); and the ANR project IDEALG (ANR-10-BTBR-04, "Investissements d'Avenir, Biotechnologies-Bioressources"). We are grateful to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, and the Max Planck Institute for Biology Tubingen for providing computational resources. We also wish to thank Kristy Hill-Spanik, Rosário Petti, and Vivian Viana for field and technical support.

DATA AVAILABILITY

Sequencing data has been deposited in the SRA database under BioProjects PRJNA936482, PRJNA931233, PRJNA938301, PRJNA938403. The accession numbers for the raw sequence data are provided in Table S2.

Gracilaria chilensis, Gracilaria gracilis, and *Gracilaria caudata* Whole Genome Shotgun project have been deposited at DDBJ/ENA/GenBank under the accessions JARGXX000000000, JARGSG000000000 and JASCIV000000000, respectively. *Gracilaria vermiculophylla* updated assembly has been deposited under JAHNZQ000000000.

337 *REFERENCES*

338 Ayres-Ostrock LM, et al. 2019. Dual influence of terrestrial and marine historical processes on
339 the phylogeography of the Brazilian intertidal red alga *Gracilaria caudata*. J Phycol 55: 1096-
340 1114.

341

342 Andrews. 2010. FastQC: A quality control tool for high throughput sequence data [Online].
343 Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

344

345 Bankevich A, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to
346 single-cell sequencing. J Comp Biol 19: 455-477.

347

348 Bolger Am, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence
349 data. Bioinformatics 30: 2114-2120.

350

351 Bonthond G, et al. 2020. How do microbiota associated with an invasive seaweed vary across
352 scales? Mol Ecol 29:2094–2108.

353

354 Brawley SH et al. 2017. Insights into the red algae and eukaryotic evolution from the genome of
355 Porphyra umbilicalis (Bangiophyceae, Rhodophyta). PNAS 114: E6361–E6370.

356

Bretaudeau A, et al. 2019. Galaxy genome annotation: easier genome annotation using Galaxy and GMOD tools [version 1; not peer reviewed]. F1000Research 8:1026. https://doi.org/10.7490/f1000research.1116992.1

Bruna T, et al. 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3(1):lqaa108.

Buels R, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 17:66. https://doi.org/10.1186/s13059-016-0924-1

Buchfink B, Xie C, Huson DH 2015. Fast and sensitive protein alignment using DIAMOND. Nat Meth 12: 59-60.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Meth 18: 366-368.

Buschmann AH, et al. 2001. Red algal farming in Chile: a review. Aquaculture 194: 203–220.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioninform. 10: 421.

Cantalapiedra CP, et al. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38: 5825-5829.

Challis R. 2017 Assembly-stats, Assembly statistic visualisation.
https://github.com/rjchallis/assembly-stats/tree/17.02

Collén J, et al. 2013. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. PNAS.110: 5247-5252.

Destombe C, et al. 1989. What controls haploid-diploid ratio in the red alga, *Gracilaria verrucosa*? J Evol Biol 2:317–338.

Destombe C, et al. 1992. Differences in dispersal abilities of haploid and diploid spores of *Gracilaria verrucosa* (Gracilariales, Rhodophyta). Bot Mar 35:93–8.

Destombe C, et al. 1993. Differences in response between haploid and diploid isomorphic phases of *Gracilaria verrucosa* (Rhodophyta:Gigartinales) exposed to artificial environmental conditions. Hydrobiologia 260:131–137.

Dierckxsens, N., Mardulyn, P., & Smits, G. (2016). NOVOPlasty:de novoassembly of organelle genomes from whole genome data. *Nucleic Acids Research*, gkw955.
https://doi.org/10.1093/nar/gkw955

Elbers JP, et al. 2019. Improving Illumina assemblies with Hi-C and long reads: An example with North African dromedary. Mol Ecol Res. 19: 1015-1026.

403 Engel CR, et al. 1999. Performance of non-motile male gametes in the sea: analysis of paternity
404 and fertilization success in a natural population of a red seaweed, *Gracilaria gracilis.* Proc Royal
405 Soc. B 266: 1879-1886.
406
407 Engel CR, et al. 2002. Non-random mating in controlled multiple-donor crosses in *Gracilaria gracilis*
408 (Gracilariaceae, Rhodophyta). European Journal of Phycology 37:179-190.
409
410 Faugeron S, et al. 2001. Hierarchical Spatial Structure and Discriminant Analysis of Genetic
411 Diversity in the Red Alga Mazzaella Laminarioides (gigartinales, Rhodophyta). J Phycol
412 37:705–716.
413
414 Flanagan BA, et al. 2021. Founder effects shape linkage disequilibrium and genomic diversity of
415 a partially clonal invader. Mol Ecol 30:1962-1978.
416
417 The Galaxy Community. 2022. The Galaxy platform for accessible, reproducible and
418 collaborative biomedical analyses: 2022 update, Nucleic Acids Res. 50: W345–W351.
419 doi:10.1093/nar/gkac247
420
421 Graham LE, et al. 2016. *Algae.* Madison: LJLM Press. 720 pp.
422

423 Guillemin ML, et al. 2008. Genetic variation in wild and cultivated populations of the haploid–
424 diploid red alga *Gracilaria chilensis*: how farming practices favor asexual reproduction and
425 heterozygosity. Evol 62:1500–1519.
426
427 Guillemin ML, Huanel OR, Martinez EA 2012. Characterization of genetic markers linked to sex
428 determination in the haploid-diploid red alga *Gracilaria chilensis*. J Phycol 48:365–72.
429
430 Guiry MD, Guiry GM 2022. *AlgaeBase*. World-wide electronic publication, National University
431 of Ireland, Galway. https://www.algaebase.org
432
433 Haas BJ 2007-2011. TransposonPSI. http://transposonpsi.sourceforge.net
434
435 Hanschen ER, et al. 2020. An evaluation of methodology to determine algal genome
436 completeness. Algal Res, 51:102019.
437
438 Hayashi L, et al. 2014. Cultivation of red seaweeds: a Latin American perspective. J Appl Phycol
439 26:719–727.
440
441 Ho, C.-L., Lee, W.-K., & Lim, E.-L. 2018. Unraveling the nuclear and chloroplast genomes of an
442 agar producing red macroalga, *Gracilaria changii* (Rhodophyta, Gracilariales). Genomics 110:
443 124–133.
444

445 Huerta-Cepas J, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically
446 annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acid Res 47:
447 D309-D314.
448
449 Hughes JS, Otto SP 1999. Ecology and the evolution of biphasic life cycles. Am Nat 154:306–
450 320.
451
452 Kain JM, Destombe C 1995. A review of the life history, reproduction and phenology of
453 *Gracilaria*. J Appl Phycol 7:269–281.
454
455 Kim D, et al. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-
456 genotype. Nat Biotechnol 37: 907–915.
457
458 Kolmogorov M, Yuan J, Lin Y and Pevzner P "Assembly of Long Error-Prone Reads Using
459 Repeat Graphs", Nature Biotechnology, 2019 doi:10.1038/s41587-019-0072-8
460
461 Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read
462 assembly via adaptive k-mer weighting and repeat separation. Genome Research. (2017).
463 doi:10.1101/gr.215087.116
464
465 Krueger-Hadfield SA, et al. 2015. O father where art thou? Paternity analyses in a natural
466 population of the haploid–diploid seaweed *Chondrus crispus*. Heredity 114: 185-194.
467

Krueger-Hadfield SA, et al. 2016. Invasion of novel habitats uncouples haplo-diplontic life cycles. Mol Ecol 25:3801–3816.

Krueger-Hadfield SA, et al. 2017. Genetic identification of source and likely vector of a widespread marine invader. Ecol Evol 7:4432–4447.

Lavaut E, et al. 2022. Pollinators of the sea: A discovery of animal-mediated fertilization in seaweed. Science 377:528-530.

Laetsch DR, Blaxter ML 2017. KinFin: Software for taxon-aware analysis of clustered protein sequences. G3 7: 3349-3357.

Lyra GDM, et al. 2021. Phylogenomics, divergence time estimation and trait evolution provide a new look into the Gracilariales (Rhodophyta). Mol Phylogenetics Evol 165:107294.

Martinez, EA., Destombe, C., Quillet, MC., & Valero, M. 1999. Identification of random amplified polymorphic DNA (RAPD) markers highly linked to sex determination in the red alga *Gracilaria gracilis*. Mol Ecol 8: 1533-1538.

Manni M, et al. 2021. BUSCO: Assessing genomic data quality and beyond. Current Protocols. https://currentprotocols.onlinelibrary.wiley.com/doi/full/10.1002/cpz1.323

490 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A, Versatile genome assembly
491 evaluation with QUAST-LG. *Bioinformatics* (2018) 34 (13): i142-i150.
492 doi:10.1093/bioinformatics/bty266
493
494 Mungall CJ, et al. 2007. A Chado case study: an ontology-based modular schema for
495 representing genome-associated biological information. Bioinformatics 23: i337–i346.
496 https://doi.org/10.1093/bioinformatics/btm189
497
498 McLachlan J. 1973. Growth media – marine. In: Handbook of Phycological Methods, Culture
499 Methods and Growth Measurements; ed.: Janet R. Stein; pp. 25 – 52
500
501 Nakamura-Gouvea N, et al. (2022). Insights into agar and secondary metabolite pathways from
502 the genome of the red alga *Gracilaria domingensis* (Rhodophyta, Gracilariales). J Phycol 58:
503 406–423.
504
505 Nurk S, Meleshko D, Korobeynikov A and Pevzner P. 2017. metaSPAdes: a new versatile
506 metagenomic assembler. Genome Research. 27: 824-834. 10.1101/gr.213959.116
507
508 Plastino EM, Oliveira EC 1997. *Gracilaria caudata* J. Agardh (Gracilariales, Rhodophyta) –
509 restoring an old name for a common western Atlantic alga. Phycologia 36:225–332.
510
511 Popper ZA, et al. 2011. Evolution and diversity of plant cell walls: from algae to flowering
512 plants. Ann Rev Plant Biol 62: 567-590.

513

514 Price AL, Jones NC, Pevzner PA 2005. De novo identification of repeat families in large
515 genomes. To appear in Proceedings of the 13 Annual International conference on Intelligent
516 Systems for Molecular Biology (ISMB-05). Detroit, Michigan.

517

518 Priyam A. et al. 2019. Sequenceserver: a modern graphical user interface for custom BLAST
519 databases. Mol Biol Evol 36: 2922–2924. doi: 10.1093/molbev/msz185

520

521 Qiu H, et al. 2015. Evidence of ancient genome reduction in red algae (Rhodophyta). J Phycol
522 51: 624–636.

523

524 Richerd S, et al. 1993. Variation of reproductive success in a haplo-diploid red alga, *Gracilaria*
525 *verrucosa*: effects of parental identities and crossing distance. American Journal of Botany
526 80:1379-1391

527

528 Sanderson LA, et al. 2013. Tripal v1.1: a standards-based toolkit for construction of online
529 genetic and genomic databases, Database, 2013: bat075. https://doi.org/10.1093/database/bat075

530

531 Sloggett C, Goonasekera N, Afgan E. 2013. BioBlend: automating pipeline analyses within
532 Galaxy and CloudMan. Bioinformatics 29: 1685-1686. doi:10.1093/bioinformatics/btt199

533

534 Simão FA, et al. 2015. BUSCO: assessing genome assembly and annotation completeness with
535 single-copy orthologs. Bionformatics 31: 3210-3212.
536
537 Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015
538 http://www.repeatmasker.org
539
540 Searles RB 1980. The strategy of the red algal life history. Am Nat 115:113–120.
541
542 Sotka EE, et al. 2018. Combining niche shift and population genetic analyses predicts rapid
543 phenotypic evolution during invasion. Evol Appl 11:781-793.
544
545 Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017)
546 GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* 45:
547 W6-W11
548
549 Ursi S, Plastino EM 2001. Crescimento in vitro de linhagens de coloração vermelha e verde clara
550 de *Gracilaria birdiae* (Gracilariales, Rhodophyta) em dois meios de cultura: análise de diferentes
551 estádios reprodutivos. Rev. Brasil. Bot. 24: 585-592.
552
553 Valero M et al. 2017. Perspectives on domestication research for sustainable seaweed
554 aquaculture. Perspectives in Phycology 4: 33-46.
555

556 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
557 Wortman J, Young SK, Earl AM (2014) Pilon: An Integrated Tool for Comprehensive Microbial
558 Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963.
559 doi:10.1371/journal.pone.0112963

560 Wang D, et al. 2020. Pyropia yezoensis genome reveals diverse mechanisms of carbon
561 acquisition in the intertidal environment. Nat Commun. 11:4028.

562

563 Wheeler TJ, et al. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov
564 models. Nucleic Acids Res. 41: D70–D82.
565

566 Yang EC, et al. 2016. Divergence time estimates and the evolution of major lineages in the
567 florideophyte red algae. Sci Rep., 6:21361.
568

569 Yokoya NS, Oliveira EC 1992a. Effects of salinity on the growth rate, morphology and water
570 content of some Brazilian red algae of economic importance. Cienc Mar 18:49–64.
571

572 Yokoya NS, Oliveira EC 1992b. Temperature response of economically important red algae and
573 their potential for mariculture in Brazilian waters. J Appl Phycol 4:339–345.
574
575
576
577
578
579
580

TABLES

Table 1: Assembly statistics.

	G. chilensis	*G. vermiculophylla*	*G. caudata*	*G. gracilis*
Strain	NLEC103-M9	HapMaleFtJ-2017	M-176_S67	GNS1m
Sequencing	PacBio	Illumina, HiC	Illumina	PacBio
Genome size	76.07 Mbp	44.95 Mbp	30.28 Mbp	72.49 Mbp
Contigs	138	4240	5535	279
GC contents	48.9%	49.5%	49.9%	46.6%
N50	1.56 Mbp	2.56 Mbp	20.8 Kbp	563 Kbp
L50	18	6	396	38

Repeat content	66.2%	48.3%	45.7%	60.7%
Protein-coding genes	7943	6807	8737	9460
Av. gene length	1404 bp	1751 bp	1409 bp	1643 bp
Genes w. interpro / Uniprot 90*	93.4% / 88.8%	93.6% / 89.7%	91.7% / 86.5%	92.0% / 84.2%
Genes with GO annotation	52.7%	54.4%	49.9%	47.9%
Genes with intron	23.4%	24.1%	28.6%	29.4%
Busco complete	75.3%	65.1%	73.0%	77.3%
Busco fragmented	6.3%	6.7%	8.6 %	6.3%
Busco missing	18.4%	28.2%	18.4%	16.4%

603 * e-value cutoff 1e-5

604

605 *FIGURE LEGENDS*

606 Fig. 1. A) Genome assembly metrics of *Gracilaria chilensis* (top left), *Gracilaria*
607 *vermiculophylla* (top right), *Gracilaria caudata* (bottom left) and *Gracilaria gracilis* (bottom
608 right), (Challis 2017, https://github.com/rjchallis/assembly-stats). The inner radius (red) of the
609 circular plot represents the length of the longest scaffold in the assembly and the proportion of
610 the assembly that it represents. The cumulative number of scaffolds within a given percentage of
611 the genome is plotted in light purple originating at the center of the plot. The N50 and N90

scaffold lengths are indicated by dark and light orange, respectively. Genome scaffolds are plotted in gray from the circumference and the length of segment at a given percentage indicates the cumulative percentage of the assembly that is contained within scaffolds of at least that length. The GC content is marked by the dark blue outer circle. Complete, fragmented and duplicated BUSCO genes are shown in green in the upper right corner. B) Venn diagram of shared and species-specific orthogroups and orphan genes among the four sequenced *G.* species. C) *G. chilensis* (top left), *G. vermiculophylla* (top right), *G. caudata* (bottom left) and *G. gracilis* (bottom right). Photo credit in order: M-L. Guillemin, S. Krueger-Hadfield, E. M. Plastino, C. Destombe.

Fig. S1. Life cycle of *Gracilaria.* The life cycle consists of an alternation between haploid dioecious gametophytes and a diploid tetrasporophyte. The tetrasporophyte produces meiospores through meiosis, which develop as gametophytes after release. The sex of the gametophytes is determined by haploid sex chromosomes (UV system). Spores that receive the V sex chromosome develop as male gametophytes whereas spores that carry U chromosome will produce female gametophytes. After fertilization, the zygote develops within the carposporophyte on the female gametophyte and is mitotically amplified—producing thousands of diploid carpospores that after release will give rise to tetrasporophytes.

635

636

637

638

639

640

641

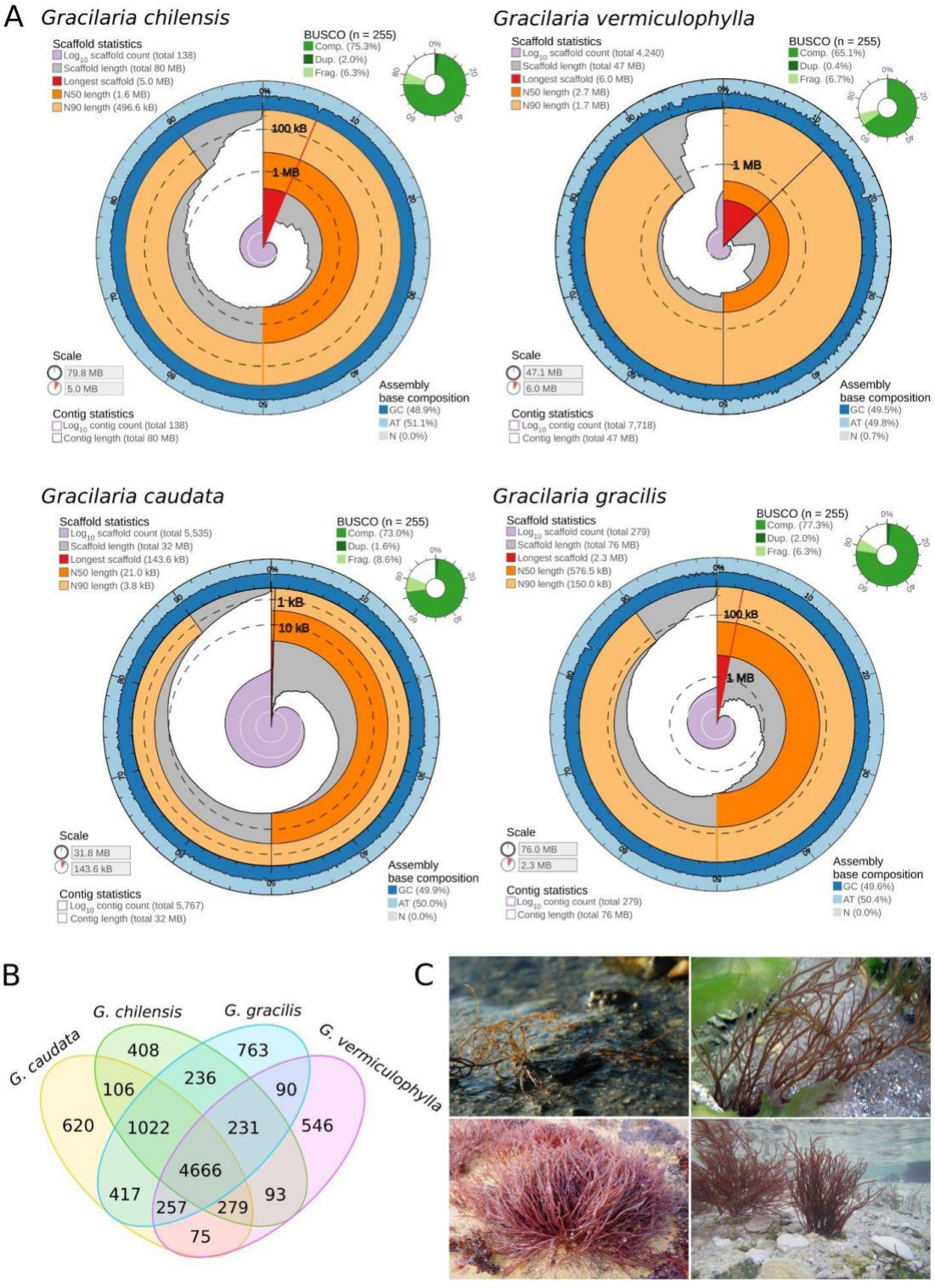
642

643

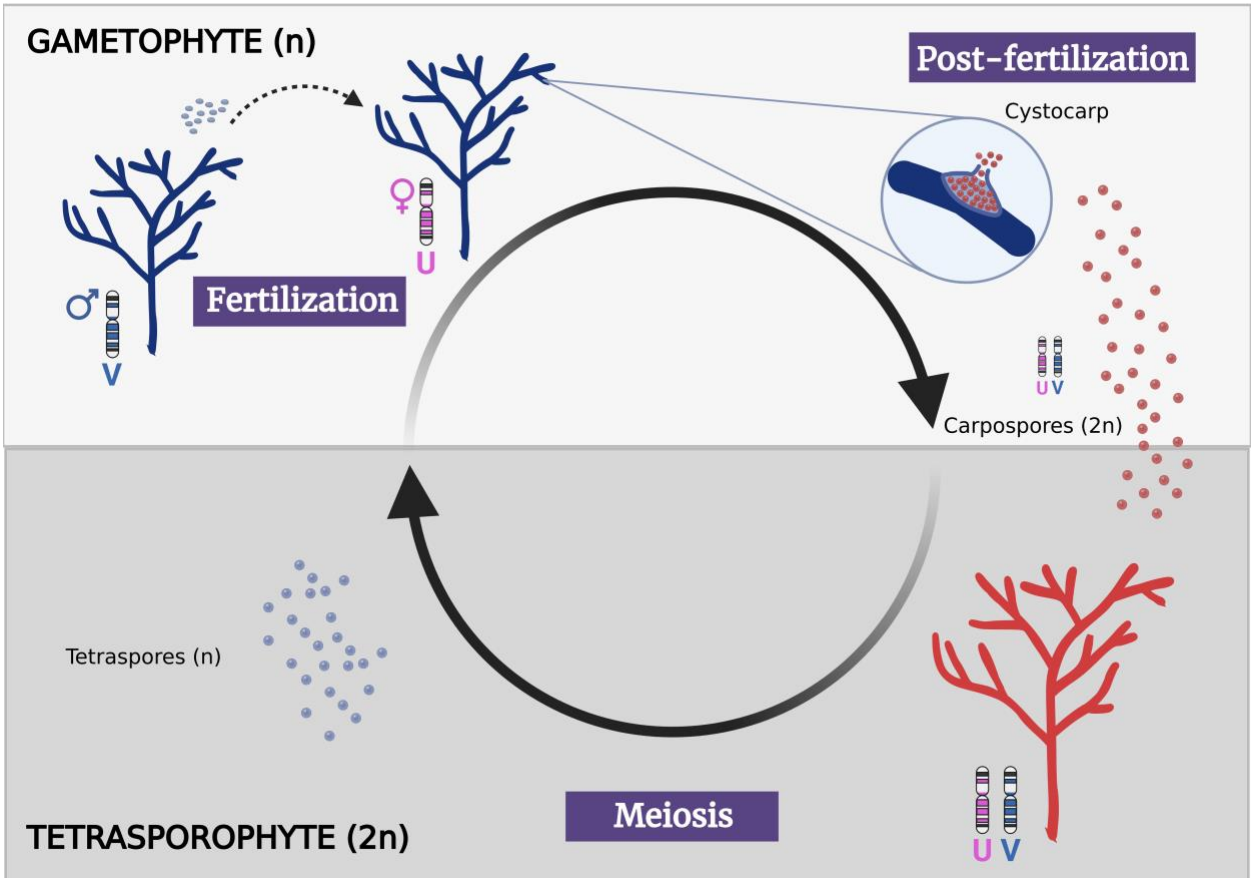
644

645

646 **Figure 1**



647



Supplementary Figure S1: Life cycle of *Gracilaria*.

Supplementary Table S1: Available red algal wholes genome sequences. M=multicellular, U=unicellular.

Species	Order	N50	U/M	Citation
Chondrus crispus	Gigartinales	250kb	M	https://doi.org/10.1073/pnas.1221259110
Galdieria sulphuraria	Cyanidiales	230kb	U	https://doi.org/10.7554/eLife.45017
Galdieria phlegrea	Cyanidiales	201kb	U	https://doi.org/10.7554/eLife.45017
Gracilaria changii	Gracilariales	17kb	M	https://doi.org/10.1016/j.ygeno.2017.09.003
Gracilaria domingensis	Gracilariales	189kb	M	https://doi.org/10.1111/jpy.13238
Gracilaria vermiculophylla	Gracilariales	2Mb	M	https://doi.org/10.1111/mec.15854
Gracilariopsis chorda	Gracilariales	220kb	M	https://doi.org/10.1093/molbev/msy081
Gracilariopsis lemaneiformis	Gracilariales	35kb	M	https://doi.org/10.1186/s12870-018-1309-2
Calliarthron tuberculosum	Corallinales	n/a	M	https://doi.org/10.1016/j.cub.2011.01.037
Porphyridium purpureum	Porphyridiales	20kb	U	https://doi.org/10.1038/ncomms2931
Porphyra umbilicalis	Bangiales	202kb	M	https://doi.org/10.1073/pnas.1703088114
Neoporphyra haitanensis	Bangiales	650kb	M	https://doi.org/10.1093/molbev/msab315
Neopyropia yezoensis	Bangiales	34Mb	M	https://doi.org/10.1038/s41467-020-17689-1
Kappaphycus alvarezii	Gigartinales	849kb	M	https://doi.org/10.1101/2020.02.15.950402
Asparagopsis taxiformis	Bonnemaisoniales	2kb	M	https://doi.org/10.1021/acschembio.0c00299

Cyanidium caldarium	Cyanidiales	13kb	U	https://www.ncbi.nlm.nih.gov/genome/7354*
Cyanidiococcus yangmingshanensis	Cyanidiales	653kb	U	https://doi.org/10.1111/jpy.13056
Cyanidioschyzon merolae	Cyanidioschyzonales	846kb	U	https://doi.org/10.1186/1741-7007-5-28 https://doi.org/10.1038/nature02398

660 * no publication associated
661 n/a data no longer accessible
662

663 **Supplementary Table S2:** Species used in this study

Species	Strain name	Isolation location	Sex	Type of data	Accession numbers
Gracilaria chilensis	NLEC103-F17	Lenca, Region of Puerto Montt, Chile (-41.607, -72.692)	Female	RNAseq	SRR23519128
	NLEC103-F17		Female	DNAseq	SRR23519124
	NLEC103-F20		Female	RNAseq	SRR23519127
	NLEC103-F20		Female	DNAseq	SRR23519123
	NLEC103-M9		Male	RNAseq	SRR23519129
	NLEC103-M9		Male	DNAseq	SRR23519122, SRR23519125
	NLEC103-M1		Male	RNAseq	SRR23519130
	NLEC103-M2		Male	DNAseq	SRR23519126
Gracilaria gracilis	GNS1m	Cap-Gris-Nez Northen France (50.872, 1.584)	Male	DNAseq	SRR23565662, SRR23565669
	GNS1m		Male	RNAseq	SRR23565661 , SRR23565660, SRR23565659, SRR23565670
	GNS1f		Female	DNAseq	SRR23565672, SRR23565663
	GNS1f		Female	RNAseq	SRR23565671, SRR23565666 , SRR23565665 , SRR23565664

Species	Strain name	Isolation location	Sex	Type of data	Accession numbers
	GNH218m		Male	DNAseq	SRR23565667
	GNH47aAf		Female	DNAseq	SRR23565668
Gracilaria caudata	172F	Pedra Rachada beach, Paracuru, Ceará, Brazil (-3.399, -39.012)	Female	DNAseq	SRR23610505
	172F		Female	RNAseq	SRR23610508
	174F		Female	DNAseq	SRR23610506
	174F		Female	RNAseq	SRR23610509
	176M		Male	DNAseq	SRR23610514
	176M		Male	RNAseq	SRR23610511
	178M		Male	DNAseq	SRR23610513
	178M		Male	RNAseq	SRR23610515
	179M		Male	RNAseq	SRR23610516
	177M		Male	RNAseq	SRR23610512
	175F		Female	RNAseq	SRR23610510
	171F		Female	RNAseq	SRR23610507
Gracilaria vermiculophylla	Gver_F	Charleston, SC, USA (32.752, -79.900)	Female	DNAseq	SRR23609120
	fjs03mal		Male	RNAseq	SRR23609119
	fjs33mal		Male	RNAseq	SRR23609118
	fjs36mal-New		Male	RNAseq	SRR23609117
	Fjs50mal-New		Male	RNAseq	SRR23609116
	fjs34fem		Female	RNAseq	SRR23609115
	fjs39fem		Female	RNAseq	SRR23609114
	fjs40fem		Female	RNAseq	SRR23609113
	fjsfem		Female	RNAseq	SRR23609112

664

665