J. Phycol. **54,** 775–787 (2018)

DOI: 10.1111/jpy.12765

ORGANELLAR GENOMICS: A USEFUL TOOL TO STUDY EVOLUTIONARY RELATIONSHIPS AND MOLECULAR EVOLUTION IN GRACILARIACEAE (RHODOPHYTA)[1]

Cintia Iha

Department of Botany, Institute of Biosciences, University of São Paulo, R Matão 277, São Paulo SP 05508-090, Brazil
School of BioSciences, University of Melbourne, Melbourne Victoria 3010, Australia

Christopher J. Grassa

Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge Massachusetts 02138, USA

Goia de M. Lyra

Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge Massachusetts 02138, USA
Laboratório de Algas Marinhas, Instituto de Biologia, Universidade Federal da Bahia, Rua Barão de Jeremoabo, s/n, Salvador Bahia 40170-115, Brazil

Charles C. Davis

Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge Massachusetts 02138, USA

Heroen Verbruggen[2]

School of BioSciences, University of Melbourne, Melbourne Victoria 3010, Australia

and Mariana C. Oliveira[2]

Department of Botany, Institute of Biosciences, University of São Paulo, R Matão 277, São Paulo SP 05508-090, Brazil

Gracilariaceae has a worldwide distribution including numerous economically important species. We applied high-throughput sequencing to obtain organellar genomes (mitochondria and chloroplast) from 10 species of Gracilariaceae and, combined with published genomes, to infer phylogenies and compare genome architecture among species representing main lineages. We obtained similar topologies between chloroplast and mitochondrial genomes phylogenies. However, the chloroplast phylogeny was better resolved with full support. In this phylogeny, *Melanthalia intermedia* is sister to a monophyletic clade including *Gracilaria* and *Gracilariopsis*, which were both resolved as monophyletic genera. Mitochondrial and chloroplast genomes were highly conserved in gene synteny, and variation mainly occurred in regions where insertions of plasmid-derived sequences (PDS) were found. In mitochondrial genomes, PDS insertions were observed in two regions where the transcription direction changes: between the genes *cob* and *trn*L, and *trn*A and *trn*N. In chloroplast genomes, PDS insertions were in different positions, but generally found between *psd*D and *rrs* genes. Gracilariaceae is a good model system to study the impact of PDS in genome evolution due to the frequent presence of these insertions in organellar genomes. Furthermore, the bacterial *leu*C/*leu*D operon was found in chloroplast genomes of *Gracilaria tenuistipitata*, *G. chilensis*, and *M. intermedia*, and in extrachromosomal plasmid of *G. vermiculophylla*. Phylogenetic trees show two different origins of *leu*C/*leu*D: genes found in chloroplast and plasmid were placed with proteobacteria, and genes encoded in the nucleus were close to Viridiplantae and cyanobacteria.

Key index words: **chloroplast; gene synteny; genome architecture; mitochondria; phylogenomics; plasmid-derived sequences; plasmids**

***Abbreviations*: aa, amino acid; CDS, coding sequence; FSO, homologous freestanding ORF; *G.*, *Gracilaria*; *Gp.*, *Gracilariopsis*; LCB, locally collinear blocks; ML, maximum likelihood; ORF, open-reading frame; PDS, plasmid-derived sequence**

The use of high-throughput sequencing techniques has led to an increase in studies using

[1]Received 29 October 2017. Accepted 29 June 2018. First Published Online 10 July 2018. Published Online 12 September 2018, Wiley Online Library (wileyonlinelibrary.com).
[2]Author for correspondence: e-mail mcdolive@ib.usp.br
Editorial Responsibility: K. Müller (Associate Editor)

complete organellar genomes to infer phylogenetic relationships in Rhodophyta, first for the study of higher taxonomic levels (Janouškovec et al. 2013, Yang et al. 2015, Lee et al. 2016a) and more recently for order and families (Costa et al. 2016, Díaz-Tapia et al. 2017). Those data also facilitated the investigation of other aspects of red algal genomes, including gene synteny and horizontal gene transfers (Yang et al. 2015, Lee et al. 2016a,b), and, also to find plasmids (Lee et al. 2016b). Plasmids are extrachromosomal DNA molecules, autonomously replicating, generally circular, AT-rich, double-stranded and may have active open read frames (Goff and Coleman 1990, Moon and Goff 1997, Lee et al. 2016b). They are recognized as mobile elements and can behave as transposable elements (Harrison and Brockhurst 2012, Lee et al. 2016b). For example, sequences derived from Rhodophyta plasmids (e.g., Goff and Coleman 1990, Villemur 1990a,b, Moon and Goff 1997) were observed in the chloroplast genomes in Rhodophyta (Lee et al. 2016b).

The family of red algae Gracilariaceae includes 230 species in six genera (Lyra et al. 2015, Guiry and Guiry 2017) of which *Congracilaria* and *Gracilariophila*, are parasitic; *Gracilaria*, *Gracilariopsis*, *Melanthalia* and *Curdiea* are benthic and free-living. The family is largely pantropical but also occurs in temperate and boreal regions. Species of the family are commercially important not only for the production of agar (Zemke-White and Ohno 1999) but also for numerous pharmacological applications, which possess antiviral, anti-inflammatory, and antihypertensive properties (Smit 2004).

At the time of the writing of this paper, six chloroplasts and nine mitochondrial genomes of Gracilariaceae species are available in GenBank. Most of these are from economically important crop species, e.g., *Gracilariopsis chorda* (Yang et al. 2014), *Gracilariopsis lemaneiformis* (Zhang et al. 2012, 2016), *Gracilaria firma* (Ng et al. 2017), and *Gracilaria tenuistipitata* var. *liui*, which was the first complete chloroplast genome published of Florideophyceae (Hagopian et al. 2004). Mitochondrial genomes were also investigated in the parasitic taxon *Gracilariophila oryzoides* and its host *Gracilariopsis andersonii* (Hancock et al. 2010).

Furthermore, the family is highly interesting for investigations of genome evolution and the influence of plasmid-derived sequences (PDSs), as most of the published chloroplast genomes present these insertions (Lee et al. 2016b, Ng et al. 2017). Moreover, horizontal gene transfers from a bacterial plasmid, which contained the genes *leu*C and *leu*D (henceforth *leu*C/*leu*D), were also reported in the chloroplast genomes of *Gracilaria tenuistipitata* var. *liui* (Hagopian et al. 2004, Janouškovec et al. 2013) and *Gracilaria chilensis* (Lee et al. 2016b). However, relatively little attention has been paid to examining genome evolution in Gracilariaceae, and phylogenetic analyses based on a large data set from organellar genomes are not yet available for the family in the literature.

Our study aims to (i) characterize new mitochondrial and chloroplast genomes of Gracilariaceae, expanding sampling among diverse lineages within the family, (ii) evaluate the evolutionary history of the family using phylogenomics, and (iii) compare the genomic architecture of each of the organelles, including an investigation of PDSs.

MATERIAL AND METHODS

Taxon sampling and culturing. We sequenced whole-genome from ten Gracilariaceae species: *Gracilaria caudata*, *G. ferox*, *G. gracilis*, *G. rangiferina*, *G. tenuistipitata*, *G. vermiculophylla*, *Gracilariopsis longissima*, *Gp. mclachlanii*, *Gp. tenuifrons*, and *Melanthalia intermedia* (Table S1 in the Supporting Information). Except for *Gracilaria caudata*, *G. vermiculophylla*, and *M. intermedia*, samples were obtained from the LAM-USP Germplasm Bank (Costa et al. 2012). *Gracilaria caudata* has been maintained in laboratory culture but has no Germplasm Bank voucher (Table S1). These samples were cultivated in vitro for ~1 month. Apical fragments obtained from the Germplasm bank were transferred to Erlenmeyer flasks 150 mL with 50 mL of modified von Stosch enriched seawater solution (Ursi and Plastino 2001). Cultures were maintained under 150 μmol · photons · m^{-2} · s^{-1} photosynthetically active radiation provided by 40 W daylight fluorescent tubes on a 14 h light/10 h dark cycle and aerated for 30 min · h^{-1}. We renewed the medium weekly when we rinsed the algae in the freshwater and MilliQ water to clean possible contaminants. Twenty-four hours before DNA extraction, we added 50 mg · L^{-1} of penicillin and 25 mg · L^{-1} of streptomycin to the 50 mL of modified von Stosch medium to eliminate possible bacterial contamination; plus 2 mL · L^{-1} of Micostatin to eliminate possible fungi contamination; and 1–5 mL · L^{-1} of germanium dioxide to eliminated possible diatom contamination. Silica dry material of *G. vermiculophylla* and *M. intermedia* were provided by S. Frederiqc's laboratory collection.

Molecular techniques and high-throughput sequencing. We extracted total genomic DNA from samples (~200–800 mg of biomass) using the method described in Faugeron et al. 2001. To prepare the DNA libraries, except for *Gracilaria vermiculophylla*, *Gracilariopsis tenuifrons*, and *Melanthalia intermedia*, we used TruSeq DNA PCR-Free LT Library preparation kit (Illumina, San Diego, CA, USA), following the manufacturer's instruction. Before the DNA library preparation, we diluted the genomic DNA to a final volume of 55 μL at 40 ng · μL^{-1}, measured with the Qubit® 3.0 Fluorometer using the Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific Inc, Waltham, MA, USA). DNA was sheared using Covaris S-series (Woburn, MA, USA) with the following parameters: 175 W of incident power, 5% duty factor, 200 cycles, intensity level 2, for 40 s. We used the Kapa Hyperplus Library Preparation Kit to prepare the DNA library of *Gracilaria vermiculophylla* and *M. intermedia*, diluting genomic DNA to a final volume of 35 μL at a minimum of 20 ng of total DNA input, with enzymatic fragmentation (8 min). To prepare the *Gracilariopsis tenuifrons* DNA library, we diluted the genomic DNA to a final volume of 50 μL at 100 ng · μL^{-1} and sonicated under the following parameters: 175 W of incident power, 10% duty factor, 200 cycles, intensity level 2, for 50 s. For this species, we constructed the DNA library using NEBNext DNA Library Prep Master Mix and NEBNext Multiplex oligos (New England Biolabs, Ipswich, MA, USA), following

the manufacturer's protocol. We quantified all the libraries with real-time PCR using the Roche Light 480 II Cycler with the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Kapa Biosystems, Wilmington, NC, USA). Except for *Gracilariopsis tenuifrons*, *Gracilaria vermiculophylla*, and *M. intermedia*, all the prepared DNA libraries were diluted in ddH_2O to a final concentration of 2 nM and pooled together. We sequenced the pool using Illumina HiSeq 2500 at the Human Genome and Stem Cell Research Center (HUG-CELL) at the University of São Paulo (IB-USP, São Paulo, Brazil). We implemented the Rapid Run mode using the HiSeq Rapid PE Cluster v2 and HiSeq Rapid SBS (500 cycles) kits to generate read lengths of 250 base pairs (bp). *Gracilariopsis tenuifrons* was sequenced in a paired-end mode using 200 cycles on an Illumina HiScanSQ at the Laboratório Multiusuários Centralizado (ESALQ-USP, Piracicaba, Brazil) to generate reads of 100 bp length. *Gracilaria vermiculophylla* and *M. intermedia* were sequenced in a paired-end mode using 300 cycles on an Illumina NextSeq 550.

Genome assembly and annotation. Raw Illumina read quality was analyzed in the FastQC program (Andrews 2011). We used Trimmomatic v0.036 (Bolger et al. 2014) to clean and trim low-quality reads and bases. We normalized the raw reads to 100× coverage using BBnorm (Bushnell 2014) for the DNA libraries with exceptionally high coverage. We assembled the genomes with SPAdes (Nurk et al. 2013) setting different k-mer sizes (i.e., 22, 33, 55, 77, 99, and 127). Mitochondrial and chloroplast assembled contigs were identified using BLASTn (Altschul et al. 1990) searches against a custom-built database containing published organellar genomes from Gracilariaceae. These contigs were imported to Geneious 9.1.8 (Biomatters, Auckland, New Zealand) and original reads were mapped using Bowtie 2 (Langmead and Salzberg 2012) to confirm the assembly. Circularity was verified by changing the start and end position and remapping the original reads to the assemblies. For *Gracilariopsis tenuifrons*, we could not complete the circular genome and we used MyCC (Lin and Liao 2016) to identify mitochondrial and chloroplast contigs. Both mitochondrial and chloroplast genomes were annotated using MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) to find coding sequences (CDS) and ARAGORN (http://130.235.46.10/ARAGORN/) and RNAweasel (http://megasun.bch.umontreal.ca/cgi-bin/RNAweasel/RNAweaselInterface.pl) to find RNA and intron sequences. Large and Small ribosomal RNA were checked using the SILVA rRNA database (Quast et al. 2013). We also performed manual inspections and corrections looking for open-reading frames (ORFs) using the ORF finder plugin available in Geneious 9.1.8. Doubtful ORFs were verified using BLASTx. We reannotated the organellar genomes obtained from GenBank when genes were missing or differed from the same organellar genomes

TABLE 1. Homologous ORFs found in plasmids of Gracilariaceae species

Freestanding ORF	Organism	Plasmid	ORF	Length (bp)
FSO1	*Gracilaria chilensis*	Gch3937	ORF1	1,218
	Gracilaria chilensis	Gch7220	ORF1	813
	Gracilaria chilensis	Gch7220	ORF4	1,218
	Gracilaria chilensis	Gch7220	ORF6	243
	Gracilaria chilensis	Gch7220	ORF7	153
	Gracilaria chilensis	GC2	Unique ORF	1,236
	Gracilaria ferox	Gfe3115	ORF2	1,254
	Gracilaria robusta	Gro4970	ORF1	1,236
	Gracilaria robusta	Gro4059	ORF2347	573
	Gracilaria vermiculophylla	Gve7464	ORF1	1,248
	Gracilaria vermiculophylla	Gve4548	ORF2	1,242
	Gracilariopsis lemaneiformis	Gle4293	ORF1	1,257
FSO2	*Gracilaria ferox*	Gfe3115	ORF3	456
	Gracilaria robusta	Gro4970	ORF5	585
	Gracilaria vermiculophylla	Gve7464	ORF4	534
FSO3	*Gracilaria chilensis*	Gch3937	ORF4	348
	Gracilaria ferox	Gfe3115	ORF4	351
	Gracilaria robusta	Gro4970	ORF6	405
FSO4	*Gracilaria chilensis*	Gch3937	ORF2	543
	Gracilaria chilensis	Gch7220	ORF5	594
	Gracilaria vermiculophylla	Gve7464	ORF2	204
FSO5	*Gracilaria vermiculophylla*	Gve4548	ORF1	519
	Gracilariopsis lemaneiformis	Gle4293	ORF2	495
FSO6	*Gracilaria vermiculophylla*	Gve4548	ORF3	600
	Gracilariopsis lemaneiformis	Gle4293	ORF5	564
FSO7	*Gracilaria vermiculophylla*	Gve4548	ORF4	687
	Gracilariopsis lemaneiformis	Gle4293	ORF4	543
FSO8	*Gracilaria vermiculophylla*	Gve4548	ORF5	804
	Gracilariopsis lemaneiformis	Gle4293	ORF3	789
FSO9	*Gracilaria robusta*	Gro4059	ORF1478	465
FSO10	*Gracilariopsis lemaneiformis*	GL3	ORF2	435
FSO11	*Gracilaria robusta*	Gro4970	ORF3	414
FSO12	*Gracilaria chilensis*	Gch7220	ORF2	213
	Gracilaria chilensis	Gch3937	ORF3	213
FSO13	*Gracilaria chilensis*	Gch7220	ORF8	348
	Gracilaria chilensis	Gch7220	ORF9	213
	Gracilaria chilensis	Gch7220	ORF10	159
	Gracilaria chilensis	Gch7220	ORF16	330
	Gracilariopsis lemaneiformis	GL3	ORF1	381

from different species. Revised annotations of published genomes are found in Table S2 in the Supporting Information.

We searched for plasmid sequences in the assembled contigs from all sequenced species using BLASTn (-task blastn) against published Rhodophyta plasmids (Table S3 in the Supporting Information). Resulting contigs from BLASTn were imported to Geneious 9.1.8 and original reads were mapped using Bowtie 2 to confirm the assembly. Circularity was verified by changing the start and end position and remapping to the original reads. ORFs were identified using the ORF finder plugin and verified using BLASTx on the web. We also compared all homologous plasmid genes of known Gracilariaceae plasmids. The homologous freestanding ORFs were classified in different categories called FSO (Table 1).

We searched for PDS in all complete organellar genomes, including the genomes downloaded from Genbank, using 10 published red algae plasmids available in Genbank (Table S3). The PDSs were searched using tBLASTx with 52 plasmid CDS queries against our database of Gracilariaceae organellar genomes. We also used BLASTn with complete plasmid sequences against genomes database (-task blastn). For both searches, we considered only hits with *e*-value less than $1.0e^{-20}$.

Phylogenetic and genomic analyses. For each organelle, we constructed a concatenated sequence matrix using amino acid (aa) data. Both matrices were assembled using CDS that were present in at least four species. We used 25 CDS for the mitochondrial matrix and 196 CDS for the chloroplast matrix (Table S4 in the Supporting Information). The matrices were constructed using genes from either mitochondrial or chloroplast genomes from Gracilariaceae available in GenBank plus the species newly sequenced in this study (Table 2). In total, our mitochondrial and chloroplast matrices included 19 and 15 samples, respectively. For all phylogenetic reconstructions, we used *Chondrus crispus*, *Gelidium elegans*, *Gracilaria vagum*, and *Rhodymenia pseudopalmata* as outgroups (Leblanc et al. 1995, Janouškovec et al. 2013, Kim et al. 2014, Yang et al. 2014, 2015, Lee et al. 2016a,b). The aa matrices were aligned using MAFFT version 7 (Katoh and Standley 2013) as implemented in Geneious 9.1.8. We performed maximum likelihood (ML) analyses using IQ-TREE 1.5.4 (Nguyen et al. 2015) with 1,000 replicates of ultrafast bootstrap (Minh et al. 2013). We performed partitioned analyses (Chernomor et al. 2016) using PartitionFinder (Lanfear et al. 2012) available in IQ-TREE 1.5.4 (-m TESTMERGE) to select a model for each gene for the aa matrix.

We standardized the start position of all linear genomes to assess synteny. Mitochondrial genomes were linearized at the Asparagine tRNA gene (*trn*N); chloroplast genomes were linearized at the small subunit ribosomal RNA gene (*rrs*). The start positions were chosen because they are the beginning of large conserved regions in the genomes. Both organellar genomes were aligned using a progressive Mauve algorithm (Darling et al. 2010) in Geneious 9.1.8 applying the full alignment option, automatically calculated seed weights and automatically calculated minimum locally collinear blocks (LCB) to score and compare genetic synteny between species within genomes.

We searched for *leu*C and *leu*D genes in the assembled contigs in all samples sequenced. We used as our query the *Gracilaria tenuistipitata* and *Gracilaria chilensis* chloroplast-encoded genes and *Chondrus crispus* nuclear-encoded genes (*leu*C = XP_005715507; *leu*D = XP_005718118). Other *leu*C and *leu*D sequences were obtained from GenBank using closest homologs from BLASTp searches upon excluding samples from environment samples and multispecies, which usually are not well identified. Protein sequences were aligned using MAFFT version 7 as implemented in Geneious 9.1.8. Maximum likelihood inferences were completed in IQ-TREE 1.5.4 using "-m TEST" parameter to choose the best model with ProtTest and 1,000 replicates of ultrafast bootstrap.

RESULTS

Sequencing and assembly. All samples sequenced using the Illumina HiSeq 2500 had extraordinarily high coverage (>5,000×) of both organellar genomes. Because excessive coverage can affect assembly quality (Lonardi et al. 2015), read coverage was normalized to 100× prior to assembly. *Gracilariopsis tenuifrons*, in contrast, had an acceptable, but low coverage assembly (20.81 on average for the mitochondrial and 43.62 on average for the chloroplast genome). Thus, it was not possible to determine complete organellar chromosomes for this species (Table S1).

We completed the assembly and annotations of both organellar genomes without gaps for *Gracilaria caudata*, *G. ferox*, *G. gracilis*, *G. rangiferina*, *G. vermiculophylla*, *Gracilariopsis longissima*, *Gp. mclachlanii*, and *Melanthalia intermedia*. For *G. tenuistipitata*, we completed only the mitochondrial genome because the chloroplast genome had been previous published for the same strain (Hagopian et al. 2004). For *Gracilariopsis tenuifrons*, we extracted 19 mitochondrial genes (GenBank accession number MH396176–MH396194) and 147 chloroplast genes (GenBank accession number MH396029–MH396175) from 6 mitochondrial and 56 chloroplast contigs, respectively.

Genome structure. All general characteristics of the organellar genomes of Gracilariaceae species are described in Table 2. The nine complete mitochondrial genomes we sequenced were similar in GC content and length. GC content was 28.1% on average (min = 25.4% in *Gracilaria rangiferina*, max = 29.2% in *Gracilaria gracilis*); length was 26,006 bp on average. *Gracilaria gracilis* had the shortest (25,865 bp) and *Gracilariopsis longissima* had the longest genome (26,744 bp). Mitochondrial genomes were also similar in gene content (Table 2), comprising 25 protein-coding genes, three rRNAs (*rrl*, *rrs*, *rrn*5), and 24 tRNAs. The exceptions were *Gracilariopsis longissima*, *G. chorda*, and *G. lemaneiformis* with 26 protein-coding genes and *Gracilaria ferox* with 23 tRNAs and *Melanthalia intermedia* with 20 tRNAs. (Table 2, Table S5 in the Supporting Information). All mitochondrial genomes had a Group II intron (ranging from 416 bp in *Gracilaria rangiferina* to 499 bp in *Gracilaria chilensis*) in the *trn*I tRNA gene.

All *Gracilaria* and *Gracilariopsis* complete chloroplast genomes sequenced in this study were also similar in GC content and length. However, *M. intermedia* chloroplast genome is larger and had higher GC content (Table 2). GC content was 28.8% on average (min = 27.3% in *Gracilaria longissima*, max = 31.8% in *Melanthalia intermedia*); the length was 186.048 bp on average. *Gracilaria rangiferina* had the shortest (178,841 bp) and *Melanthalia intermedia* the longest genome (218,416 bp). Chloroplast genomes of *Gracilaria* and *Gracilariopsis* species were also similar

TABLE 2. General information of mitochondrial and chloroplast genomes of Gracilariaceae

Species	RefSeq	INSDC	Size (kb)	GC%	Protein	rRNA	tRNA	Other RNA	Intron	Publish
Mitochondrion										
Gracilaria caudata	–	MH396017	26,030	28.8	25	3	24	–	1	This study
Gracilaria changii	NC_034681.1	KX980031	25,729	27.7	25	3	24	–	1	Song et al. (2017)
Gracilaria chilensis	NC_026831.1	KP728466	26,898	27.6	25	3	26	–	1	Lee et al. (2015)
Gracilaria ferox	–	MH396018	25,588	27.4	25	3	23	–	1	This study
Gracilaria gracilis	–	MH396019	25,865	29.2	25	3	24	–	1	This study
Gracilaria rangiferina	–	MH396020	25,908	25.4	25	3	24	–	1	This study
Gracilaria salicornia	NC_023784.1	KF824534	25,272	28.4	25	3	21	–	1	Campbell et al. (2014)
Gracilaria salicornia	–	KT373903	25,915	28.6	25	3	24	–	1	Song et al. (2016)
Gracilaria tenuistipitata	–	MH396021	25,880	27.1	25	3	24	–	1	This study
Gracilaria vermiculophylla	NC_027064.1	KJ526627	25,973	28.1	25	3	22	–	1	Chi, S., Qian, H., Zhang, L., Lv, H., Li, T.-Y., Liu, C. & Liu,T. unpub. data
Gracilaria vermiculophylla	–	MH396022	26,149	28.4	25	3	24	–	1	This study
Gracilariophila oryzoides	NC_014771.1	KX687879	25,184	28.1	25	3	21	–	1	Salomaki and Lane (2016)
Gracilariopsis andersonii	NC_014772.1	KX687878	27,011	28	25	3	21	–	1	Salomaki and Lane (2016)
Gracilariopsis chorda	NC_023251.1	KC875851	26,534	27.6	26	3	24	–	1	Yang et al. (2014)
Gracilariopsis lemaneiformis	–	JQ071938	25,883	27.5	26	3	21	–	1	Zhang et al. (2012)
Gracilariopsis longissima	–	MH396023	26,744	28.4	26	3	24	–	1	This study
Gracilariopsis mclachlanii	–	MH396024	26,012	28.5	25	3	24	–	1	This study
Gracilariopsis tenuifrons	–	–	≥19,883	28.8	≥19	≥1	≥14	–	1	This study
Melanthalia intermedia	–	MH396025	25,348	29.3	25	3	20	–	1	This study
Chloroplast										
Gracilaria caudata	–	MH396009	182,933	28.8	207	3	30	2	1	This study
Gracilaria chilensis	NC_029860.1		185,637	29.3	205	3	30	2	1	Lee et al. (2016b)
Gracilaria ferox	–	MH396010	180,255	28.8	205	3	30	2	1	This study
Gracilaria firma	NC_033877.1	KX601051	187,001	28.1	216	3	30	2	1	Ng et al. (2017)
Gracilaria gracilis	–	MH396011	180,807	29.1	207	3	30	2	1	This study
Gracilaria rangiferina	–	MH396012	178,841	27.7	203	3	30	2	1	This study
Gracilaria salicornia	NC_023785.1	KF861575	179,757	28.8	204	3	30	2	1	Campbell et al. (2014)
Gracilaria tenuistipitata var. liui	NC_006137.1	AY673996	183,883	29.2	206	3	30	2	1	Hagopian et al. (2004)
Gracilaria vermiculophylla	–	MH396013	180,254	29.5	204	3	30	2	1	This study
Gracilariopsis chorda	NC_031149.1	KX284722	182,459	27.4	203	3	30	2	1	Lee et al. (2016a)
Gracilariopsis lemaneiformis	NC_029644.1	KU179794	182,505	27.4	208	3	30	2	1	Zhang et al. (2016)
Gracilariopsis longissima	–	MH396014	181,896	27.3	204	3	30	2	1	This study
Gracilariopsis mclachlanii	–	MH396015	184,980	27.8	205	3	30	2	1	This study
Gracilariopsis tenuifrons	–	–	≥144,929	28.4	≥147	≥0	≥24	≥1	1	This study
Melanthalia intermedia	–	MH396016	218,416	31.8	224	3	30	2	1	This study

in gene content comprising 203–207 protein genes, while the *M. intermedia* chloroplast had 224 protein genes. All chloroplast genomes had 3 rRNA (*rrl*, *rrs*, *rrn*5), 30 tRNA, 1 tmRNA, and 1 npDNA (*rnp*B; Table 2, Table S6 in the Supporting Information). All chloroplast genomes possessed a Group II intron in the *trn*Me tRNA gene (ranging from 1,998 bp in *Gracilariopsis longissima* to 2,111 bp in *Gracilariopsis lemaneiformis*).

Plasmids. From the assembled contig data, we found three circular plasmid sequences, one in *Gracilaria ferox* (Gfe3115) and two in *Gracilaria*

vermiculophylla (Gve4548 and Gve7464; Fig. 1, Table S3). The Gfe3115 plasmid was 3,115 bp (82× average coverage with normalized reads to 100×) and contained four ORFs: FSO1, FSO2, FSO3, and ORF1, of which ORF1 had no similarity to other sequences in GenBank. The Gve4548 plasmid was 4,548 bp (157.7× average coverage) and had FSO1, FSO5, FSO6, FSO7, and FSO8 homologous ORFs, which presented similarity in sequence and synteny with the *Gracilariopsis lemaneiformis* plasmid Gle4293 (Fig. 1). The Gve7464 plasmid was 7,464 bp (154.2× average coverage) with six ORFs including *leu*C and *leu*D genes, FSO1, FSO2, FSO4, and ORF2, of which ORF2 had no similarity to other sequences in GenBank.

Phylogeny. Mitochondrial and chloroplast ML trees were largely congruent (Figs. 2A and 3A). When only species for which both chloroplast and mitochondrial genomes are available are used for phylogenomic analyses, chloroplast and mitochondrial trees are identical (data not shown). Gracilariaceae is monophyletic, *M. intermedia* is sister to a monophyletic clade containing both monophyletic genera *Gracilaria* and *Gracilariopsis*. Comparing the trees of both organellar genomes, the mitochondrial tree presented lower support in some nodes, including the position of *Gracilaria rangiferina* as a sister taxon to the other *Gracilaria* species in the tree (83%), *G. vermiculophylla* grouped with the *G. tenuistipitata*/*G. chilensis* clade (81%), *G. ferox* grouped with *G. salicornia* (84%), and *Gracilariopsis mclachlanii* grouped with the *Gracilariopsis lemaneiformis*/*Gp. chorda* clade (81%). On the other hand, the chloroplast tree had full support for all those nodes. We used these ML trees as references for inferring the evolution of Gracilariaceae organellar genomes.

Genomic comparison and PDSs of mitochondria. Mitochondrial genomes are highly conserved especially in two main regions. The first one between *trn*N and *cob* genes (blue LCB in Fig. 2A) and the other between *trn*L and *trn*A genes (purple LCB in Fig. 2A). Regions that were specific for each genome were generally caused by PDS insertion. We investigated the occurrence of PDS in the newly

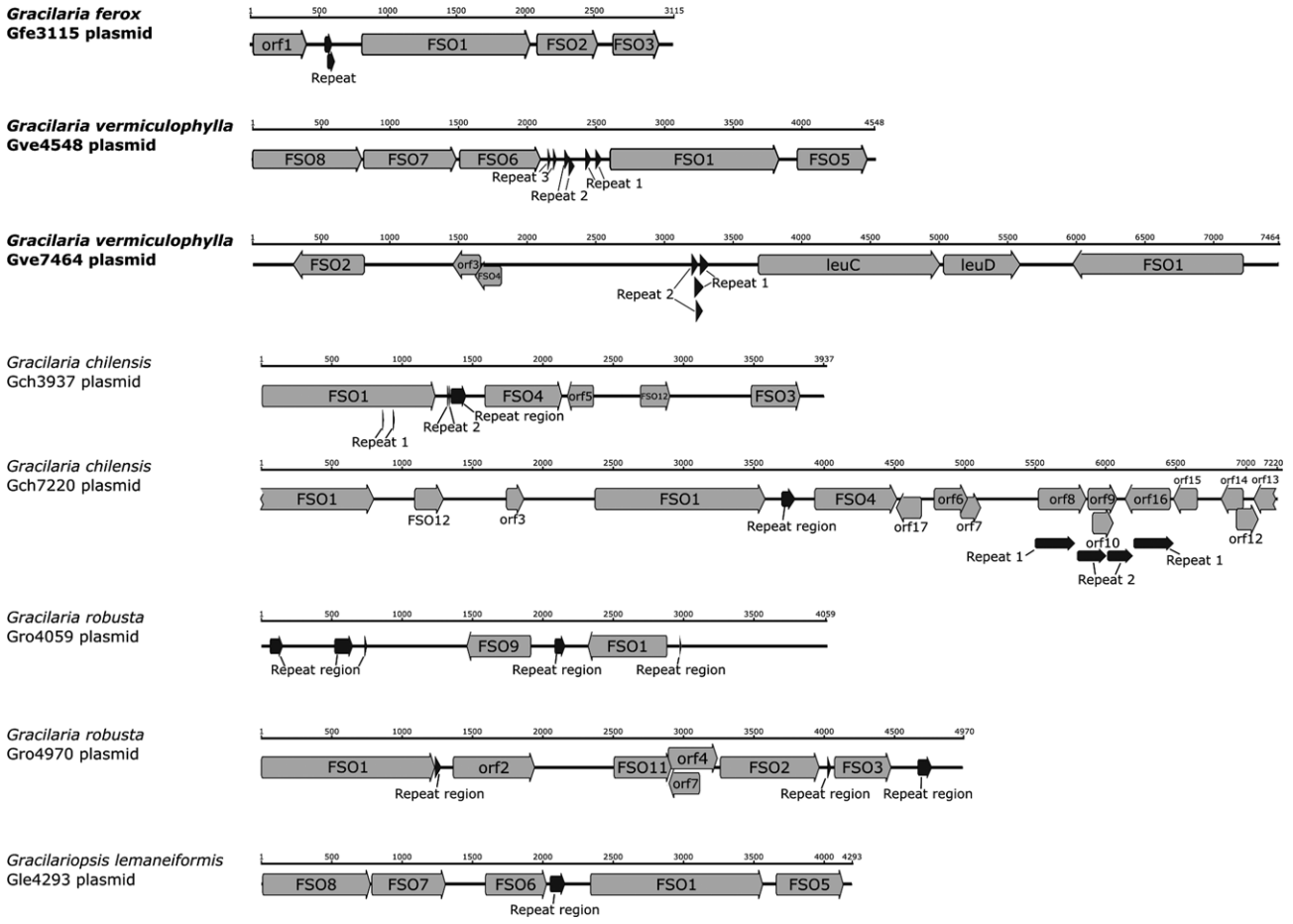

FIG. 1. Linear schematic representation of circular Gracilariaceae plasmids. Plasmids with bold names were sequenced in this study, others were obtained from GenBank (accession numbers showed in Table S3). Light gray arrows are ORFs. Black arrows are repeated regions. Regions identified as "Repeat 1, 2, and 3" are entirely repeated sequences in different positions of the genome. Regions identified as "Repeat region" are repetitions in tandem. Scale line in base pairs.

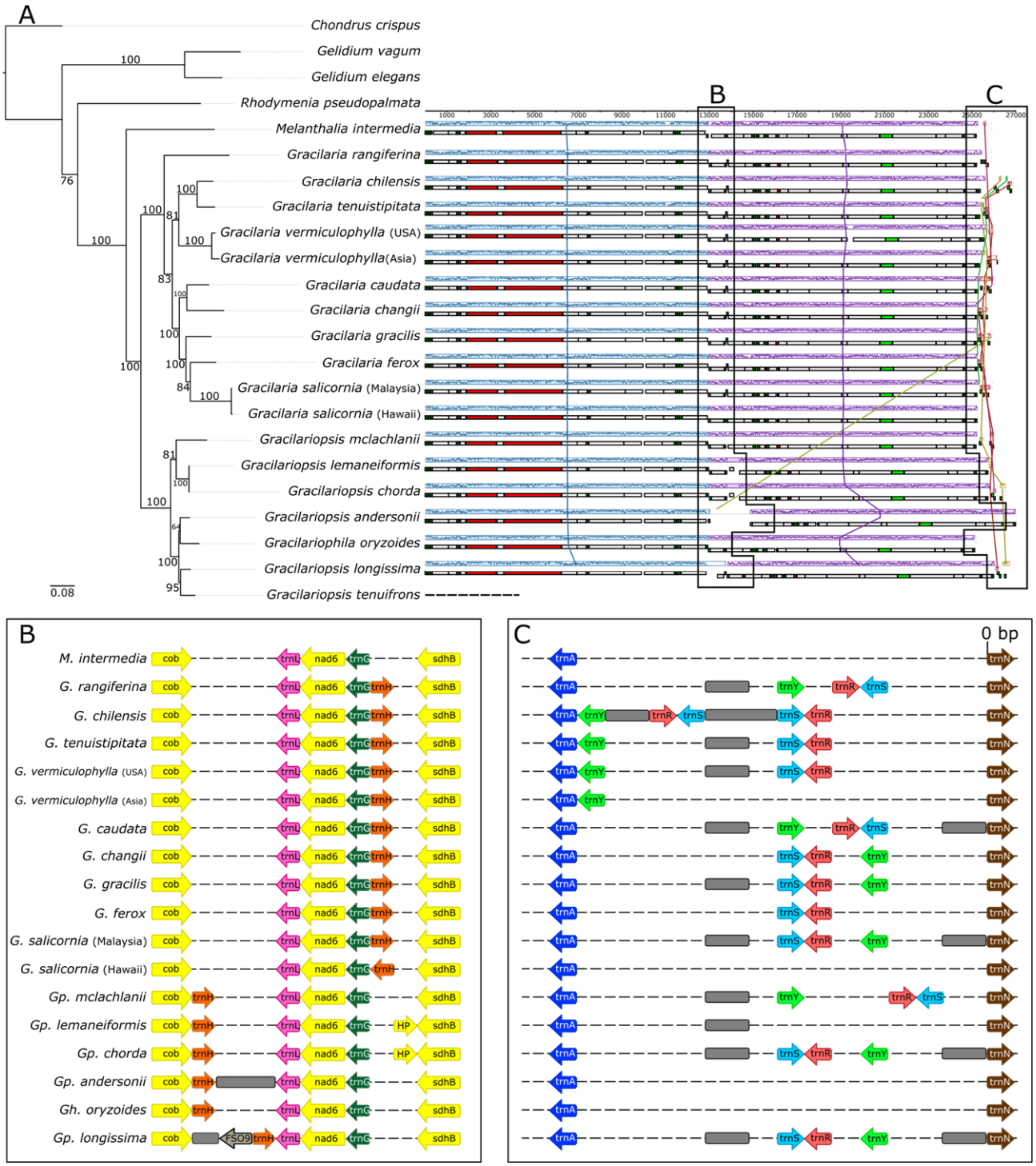

FIG. 2. Mitochondrial genome phylogeny and genomic comparison of Gracilariaceae. (A) Left, maximum likelihood phylogenetic tree based on an amino acid matrix with 25 coding genes sequences. Right, mitochondrial progressive MAUVE alignment. Each linear genome corresponds to the name in the phylogenetic tree. There is no assembled genome of *Gracilariopsis tenuifrons* in the MAUVE alignment. Horizontal axis above refers to genome length in base pairs (bp). Synteny between genomes is represented by locally collinear blocks (LCB). LCB above the line is the forward strand; below the line is the reverse strand. The same color line of LCB along the alignment indicate similarity region among the genomes. Gene maps are shown below the LCBs, protein-coding genes in white, rRNA genes in red, tRNA genes in dark green, and intron in light green. Position of boxes above (forward) and below (reverse) line also refers to gene orientation. (B and C) Loci marked in black line in A in detail. Protein-coding genes are represented in arrow box in yellow, plasmid-derived sequence in gray, and each *trn* is marked in different colors. Arrow boxes do not represent gene length and the scheme represents only the synteny and not alignment.

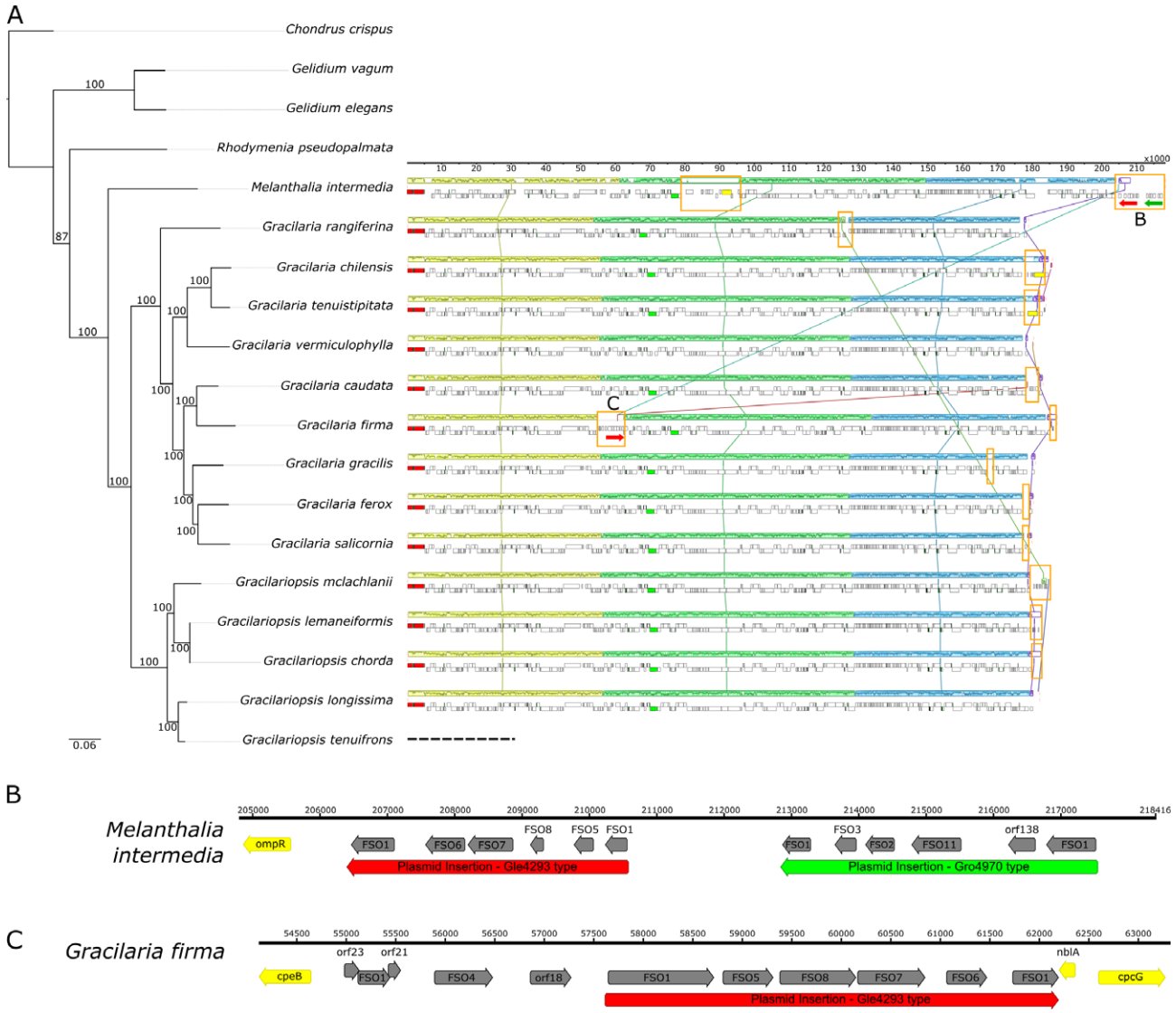

FIG. 3. Chloroplast genome phylogeny and genomic comparison of Gracilariaceae. (A) Left, maximum likelihood phylogenetic tree based on an amino acid matrix with 196 coding genes sequences. Right, mitochondrial progressive MAUVE alignment. Each linear genome corresponds to the name in the phylogenetic tree. There is no assembled genome of *Gracilariopsis tenuifrons* in the MAUVE alignment. Horizontal axis above refers to genome length in base pair (bp). Synteny between genomes is represented by locally collinear blocks (LCB). LCB above the line is the forward strand, below the line is the reverse strand. The same color line of LCB along the alignment indicate similarity region among the genomes. Gene maps are shown below the LCBs, protein-coding genes in white, rRNA genes in red, tRNA genes in black, intron in light green, and *leu*C/*leu*D operon in yellow (showed in *Melanthalia* sp., *Gracilaria chilensis*, and *Gracilaria tenuistipitata*). Position of boxes above (forward) and below (reverse) line also refers to gene orientation. Continuous orange line boxes represent plasmid-derived sequences. Red arrows indicate plasmid insertion similar to the Gle4293 plasmid. Green arrows indicate plasmid insertion similar to the Gro4970 plasmid. (B and C) Loci with plasmid insertions marked in A in detail. Protein-coding genes are represented in arrow box in yellow, including plasmid-derived ORFs (FSO).

genomes and in the previously published genomes. These insertions occurred mainly in two regions of the genome where the transcription direction changes (black line boxes in Fig. 2A).

All PDS insertions found in the mitochondrial genomes (Fig. 2, B and C) were similar (*e*-value between $1.0e^{-24}$ and $1.0e^{-46}$) to the plasmids previously described for *Gracilaria robusta* (Goff and Coleman 1988), identified as Gro4059 (Fig. 1; Table S3). *Gracilariopsis andersonii* had the largest mitochondrial genome (Table 2) due to the ~2,000 bp PDS insertion between the genes *trn*H and *trn*L. *Gracilariopsis longissima* had a PDS located between *cob* and *trn*H (Fig. 2B). In *Gp. longissima*, one of the plasmid-derived CDS is homologous to FSO9 (Fig. 2B; Table 1) that is also found in *G. robusta* Gro4059 plasmid (Fig. 1). This gene was the only plasmid-derived CDS found in the mitochondrial genomes. The other region where PDSs were found was between *trn*A and *trn*N genes (Fig. 2C). We did not find PDS insertions in the mitochondrial genomes of *Gracilaria vermiculphylla* from Asia,

G. changii, *G. ferox*, *G. salicornia* from Hawaii, *Gracilariophila oryzoides*, and *Melanthalia intermedia*. All the PDS indels occurred between tRNA genes, with one exception (*Gracilariopsis longissima* PDS between *cob* and *trn*H; Fig. 2).

The gene *trn*H is in different positions in *Gracilaria* and *Gracilariopsis*, and it is absent in *Melanthalia intermedia* (Fig. 2B). In *Gracilariopsis*, it is found between *cob* and *trn*L. In *Gracilaria*, it is found between *trn*G and *sdh*B. This gene is in the other strand in *G. salicornia* from Hawaii when compared to the other genomes (Fig. 2B). *Gracilariopsis chorda* and *G. lemaneiformis* had a hypothetical protein gene between *trn*G and *sdh*B (Fig. 2B) that did not match with any known gene in GenBank. The region between *trn*A and *trn*N is highly variable (Fig. 2, A and C). For example, *Gracilaria rangiferina*, *G. caudata*, and *Gracilariopsis mclachlanii* had an inversion in the region including the genes *trn*Y, *trn*R, and *trn*S (Fig. 2C). The *trn*Y gene is translocated in *Gracilaria vermiculophylla*, *G. chilensis*, and *G. tenuistipitata* (Fig. 2C). *Gracilaria chilensis* had a duplication of *trn*R and *trn*S genes close to an extra PDS insertion. *Gracilaria salicornia* from Hawaii, *Gracilariopsis lemaneiformis*, *Gracilariophila oryzoides*, *Gracilariopsis andersonii*, and *Melanthalia intermedia* missed the *trn*Y, *trn*R, and *trn*S tRNA genes. *Gracilaria vermiculophylla* from Asia missed *trn*R and *trn*S tRNA genes, while *G. ferox* missed *trn*Y gene. It is interesting that *G. salicornia* from Hawaii had only 38 bp between *trn*A and *trn*N, while *G. salicornia* from Malaysia had 682 bp which contained three tRNA genes and a PDS (Fig. 2C). *Gracilaria vermiculophylla* from the United States had 806 bp between *trn*A and *trn*N, which contain two tRNA genes, while *G. vermiculophylla* from Asia had 420 bp without tRNA genes.

Genomic comparison and PDSs of chloroplast. As well as mitochondrial genomes of red algae, chloroplast genomes are very conserved in synteny and genome content, presenting three main conserved regions. The first one was between *rrs* and *cpe*B genes (yellow LCB in Fig. 3A). The second was between *nbl*A and *dna*K (green LCB in Fig. 3A). The last one was between *rpl*3 and *psb*D (blue LCB in Fig. 3A). Species-specific features of chloroplast genomes are generally caused by PDS insertion (Fig. 3). They are mainly found between *psb*D and *rrs* genes. Despite the overall conservation, some differences were found. The gene *pet*P was present only in *Gracilariopsis* and *Melanthalia intermedia*. *Gracilaria salicornia*, *G. rangiferina*, *Gracilariopsis mclachlanii*, and *M. intermedia* missed *ycf*23, while *Gracilaria caudata* and *G. gracilis* had a putative pseudogene. *Gracilaria chilensis* and *M. intermedia* missed *pbs*A and *G. firma* had a small CDS, probably a pseudogene. *Melanthalia intermedia* had a pseudogene of *pgm*A. In the *trn*Me-intron, all *Gracilariopsis*, *Gracilaria firma*, and *M. intermedia* had a CDS (*mat*), while the other *Gracilaria* species had two CDSs (Table S6).

Melanthalia intermedia had at least nine species-specific features (white spaces in LCBs in Fig. 3A). Two of them were larger and had ORFs homologous to plasmids which indicated that they are PDS insertions (orange line block in Fig. 3A). The first one is between ORF684 and *ycf*58 and had the *leu*C and *leu*D operon (yellow rectangle in Fig. 3A) and had also homologous to plasmid ORFs as FSO1, FSO4, and FSO10 (Fig. S1 in the Supporting Information, Table 1). The second one is between *omp*R and *rrs* and had two complete plasmid insertions (Fig. 3A, B) similar to Gle4293 and Gro4970 (Fig. 1). Both complete plasmid insertions are flanked by the homologous FSO1 (Fig. 3B). *Gracilaria rangiferina* had a unique PDS insertion between *dna*K and *rpl*3, which contains two genes homologous to FSO1 and FSO6 (orange line block in Fig. 3, Fig. S1, Table 1). *Gracilaria chilensis* and *G. tenuistipitata* had an insertion including *leu*C/*leu*D genes (yellow rectangle in Fig. 3A) between *psd*D and *omp*R genes. *Gracilaria chilensis* PDS had two small ORFs similar to FSO1 and FSO8 between *psd*D gene and the bacterial plasmid insertion *leu*D (Fig. S1). *Gracilaria tenuistipitata* PDS had two small ORFs matched FSO1 between *omp*R and *rrs* genes (Fig. S1). *Gracilaria caudata* had a PDS insertion with three CDSs between *psb*D and *omp*R genes similar to FSO4, FSO11, FSO1, and an ORF with no similarity found in GenBank (orf142). *Gracilaria firma* had a unique, large PDS insertion region with about 7,000 bp and 11 ORFs. In this large PDS, we found a complete plasmid insertion similar to the Gle4293 plasmid (Fig. 3, A and C). This species also had another PDS with two small ORFs homologous to FSO4 between *omp*R and *rrs* genes (Fig. S1). *Gracilaria gracilis* had a small PDS insertion between *rpl*11 and *moe*B genes without ORFs. *Gracilaria ferox* had a PDS insertion between *psb*D and *omp*R genes without ORFs. *Gracilaria salicornia* also had a PDS insertion between *psb*D and *omp*R and had a gene homologous to FSO1 (Fig. S1). *Gracilariopsis mclachlanii* had a larger PDS insertion ca 5,000 bp with fragments of five CDSs similar to FSO1, FSO8, FSO7, FSO6, and FSO1 between *omp*R and *rrs* (Fig. S1). *Gracilariopsis lemaneiformis* and *Gp. chorda* had a PDS insertion between *omp*R and *rrs*. In *Gp. lemaneiformis*, the PDS had four small CDS homologous to FSO7, FSO6, and FSO1 (Fig. S1) and there is no ORFs in PDS insertion in *Gp. chorda*. We did not find any PDS insertions in the *Gp. longissima* and *G. vermiculophylla* chloroplast genome.

leuC/leuD operon. As we found the *leu*C/*leu*D genes in the chloroplast genomes of *Melanthalia intermedia*, *Gracilaria tenuistipitata*, *G. chilensis* (Fig. S1), and in *Gracilaria vermiculophylla* plasmid Gve7464 (Fig. 1), we searched for homologous genes in all assembled contigs of the other species. We found *leu*C/*leu*D encoded in nuclear contigs in all species except for *G. chilensis*, *G. tenuistipitata*, and *G. vermiculophylla*, while in *M. intermedia* we found only

pseudogenes. These species are exactly the species that present these genes in the chloroplast or plasmid. The *M. intermedia leu*D nucleus-encoded gene was so divergent that we could not align it in the matrix used to reconstruct the phylogeny. ML phylogenies inferred from each of those two genes are not congruent (Figs. S2 and S3 in the Supporting Information). In *leu*C ML phylogeny, all genes encoded in chloroplast and plasmids (*Gracilaria vermiculophylla*) grouped with Proteobacteria clade (Fig. S2). However, in *leu*D ML phylogeny, those genes grouped with one proteobacteria (*Xenorhabdus budapestensis*), while all other proteobacteria formed a separate, nonrelated clade (Fig. S3). All *leu*C/*leu*D nucleus-encoded genes grouped in a clade with Viridiplantae. The only exception was the unicellular red algae *Cyanidioschyzon merolae*, which was placed as sister of the clade that comprises cyanobacteria and Viridiplantae. Gracilariaceae nucleus-encoded genes grouped with other Rhodophyta taxa, which also have these genes in the nuclear genome.

DISCUSSION

Organellar phylogenomics of Gracilariaceae. We completed mitochondrial and chloroplast genomes of nine species of Gracilariaceae. Although we could not obtain a circularized organellar genome of *Gracilariopsis tenuifrons*, we obtained most of the genes to reliably infer the phylogeny (Tables S5 and S6). Our phylogenies are the first using larger gene content matrices (25 genes for the mitochondrial and 196 genes for the chloroplast) for Gracilariaceae and had very high bootstrap support. Phylogenetic analysis of mitochondria and chloroplast genes showed similar topology. However, the chloroplast phylogeny was better resolved (full support in all nodes; Figs. 2A and 3A). Chloroplast genomes have been well established as a promising resource to resolve red algal phylogenies, due to its highly conserved nature and adequate phylogenetic signal (Janouškovec et al. 2013, Costa et al. 2016, Lee et al. 2016a).

Although our taxonomic sampling was not extensive (19 species for mitochondrial and 15 species for chloroplast), our sampling represents the main clades of Gracilariaceae (Gurgel and Fredericq 2004, Lyra et al. 2015). Along these lines, generic circumscription within Gracilariaceae has been controversial, especially whether *Hydropuntia* is distinct from *Gracilaria* (Bellorin et al. 2002, Gurgel and Fredericq 2004, Lyra et al. 2015). Furthermore, a yet unnamed genus proposed for the clade represented by *Gracilaria chilensis*, *G. tenuistipitata*, and *G. vermiculophylla* has also been suggested (Gurgel and Fredericq 2004, Lyra et al. 2015). The largest and most well resolved phylogeny of Gracilariaceae reduced *Hydropuntia* and the unnamed genus to *Gracilaria*, forming a monophyletic genus, based on three molecular markers (*rbc*L, UPA, and COI-5P; Lyra et al. 2015). Previous members of *Hydropundia*, including *Gracilaria rangiferina*, *G. caudata*, *G. changii*, and *G. firma* (Abbott et al. 1991, Gurgel and Fredericq 2004, Lyra et al. 2015, Ng et al. 2017) did not form a monophyletic clade in any of our topologies (Figs. 2A and 3A), which corroborates previous findings (Lyra et al. 2015). On the other hand, *Gracilaria chilensis*, *G. tenuistipitata*, and *Gracilaria vermiculophylla* form a monophyletic group, but within *Gracilaria* (sensu Lyra et al. 2015). Thus, the transfer of these three species to a new genus would render *Gracilaria* non-monophyletic. Furthermore, our analyses show the genera *Gracilaria* and *Gracilariopsis* are monophyletic, with full support. In addition, the monophyly of the latter two genera is supported by the loss of the plastid gene *pet*P (in *Gracilaria*) and the position of the gene *trn*H in the mitochondrial genome (Fig. 2B). In *Gracilaria*, this gene is between *trn*G and *sdh*B and in *Gracilariopsis* this gene is between *cob* and *trn*L genes.

Genome architecture and the influence of plasmids in the evolution of mitochondrial genomes in Gracilariaceae. Our newly sequenced genomes are consistent with those known for Gracilariaceae and other Florideophyceae. Together, they demonstrate a conserved gene content and synteny among these groups (Figs. 2A and 3A, Tables S5 and S6; Campbell et al. 2014, Yang et al. 2015, Lee et al. 2016a). The 18 complete mitochondrial genomes (6 for *Gracilariopsis*, 11 for *Gracilaria*, and 1 for *Melanthalia*) are highly conserved except in two regions that include variations in tRNA genes and inserted PDS (Fig. 2A, boxes B and C). There is a rearrangement of *trn*H between *Gracilariopsis* and *Gracilaria* (Fig. 1B) that was previously reported by Lee et al. (2015); this gene is otherwise absent in *M. intermedia*. However, this rearrangement occurred several times in Florideophyceae (Yang et al. 2015). The locus with *trn*L+*nad*6+*trn*G genes (Fig. 2B) is highly conserved in Florideophyceae, except in Hildenbrandiophycidae, and it is located between two highly divergent loci (Yang et al. 2015). The region between *trn*A and *trn*N (Fig. 2C) is also highly variable in terms of tRNA content in Florideophyceae, except in Nemaliophycidae and Hildenbrandiophycidae (Yang et al. 2015). This variation does not characterize different clades in the phylogenies of these groups (Yang et al. 2015), and can also be observed within species, as in *Gracilaria salicornia* and *G. vermiculophylla* (Fig. 2C). This loci in the mitochondrial genome can be highly variable and could thus be an interesting tool for population genetics, biogeographic, and phylogeographic studies (Song et al. 2016). For example, the mitochondrial genomes of *Gracilaria salicornia* specimens from Hawaii and Malaysia have 96.9% nucleotides identity. However, if the 682 bp sequence between *trn*A and *trn*N—which include three tRNA genes

and two PDSs (Fig. 2C)—from the Malaysian specimen is excluded, the identity is 99.4%. *Gracilaria salicornia* is an invasive species in Hawaii, and its type locality is in the Philippines (Fig. 2, B and C).

The hotspots of PDS insertions in mitochondrial genomes are exactly where the origin and terminus of replication may occur, where strand direction changes (Fig. 2). These PDSs were previously reported in *Gracilaria chilensis*, *Gracilariopsis chorda*, and *Gracilariopsis lemaneiformis* (Zhang et al. 2012, Yang et al. 2014, Lee et al. 2015). Moreover, all the PDS insertions found in mitochondria were similar to *Gracilaria robusta* plasmid Gro4059 (Fig. 1), which suggests that similar plasmids could have been inserted in the mitochondria. *Gracilariopsis longissima* is the only species in which a plasmid homologous ORF (FSO9) was found in the mitochondrial genome (Fig. 2B). The intron in the *trn*I gene was found in all sequenced mitochondrial genomes of Florideophyceae (Table S2), suggesting that this intron was gained in the florideophycean ancestor (Yang et al. 2015).

Genome architecture and the influence of plasmids in the evolution of chloroplast genomes in Gracilariaceae. To infer chloroplast genome evolution in Gracilariaceae, we compared gene synteny from 14 complete genomes available (Fig. 3), four *Gracilariopsis*, nine *Gracilaria*, and one *Melanthalia*. As it was observed for the mitochondrial genomes, chloroplast genome organization is highly similar, except for regions with PDS insertions (Fig. 3). PDS insertions are common in Rhodophyta (Lee et al. 2016b) and were found in almost all Gracilariaceae chloroplasts except in *Gracilaria vermiculophylla* and *Gracilariopsis longissima* (Fig. 3A). Most PDS insertions have ORFs, which are homologous to ORFs found in previously published plasmids (Table 1) and they are positioned between the *psd*D-*omp*R-*rrs* locus (Fig. 3A). This region is the hotspot for plasmid insertion in chloroplast genomes in Gracilariaceae (Lee et al. 2016b, Ng et al. 2017), but the correlation with the origin and terminus of replication is not as clear as in the mitochondrial genome.

Melanthalia intermedia possesses two whole plasmid insertions similar in homologous ORFs and synteny to the Gle4293 and Gro4970 plasmids in the chloroplast genomes (Fig. 3B); *Gracilaria firma* possesses a whole plasmid insertion similar to the Gle4293 plasmid (Fig. 3C; Ng et al. 2017). The fact that all these insertions are flanked by FSO1 (Fig. 3, B and C), which is a protein belonging to the family DUF1368 with unknown function and specific to red algae (Lee et al. 2016b, Ng et al. 2017), could mean that this homologous gene may be associated with the mobility of the complete plasmid or plasmid regions. FSO1 is the most common and widespread gene found in plasmids and chloroplast genomes of red algae (Lee et al. 2016b); the origin of this plasmid-derived gene is unknown (Lee et al. 2016b). The presence of red algal plasmid regions in the organellar genome has been reported in Rhodophyta with or without the co-occurrence of extrachromosomal plasmids (Ng et al. 2017). We found extrachromosomal circular plasmids in *Gracilaria ferox* and *Gracilaria vermiculophylla*. The plasmid in *G. ferox* has FSO1, FSO2, FSO3, and an ORF with no similarity in GenBank protein database (Fig. 1). The PDS insertion in the chloroplast of *G. ferox* had three degenerated ORFs that are different from the plasmid: FSO5, FSO8, and FSO6, which are ORFs found in the Gle4293 plasmid. In *G. vermiculophylla*, there is no plasmid-derived ORFs in organellar genomes. The Gve4548 plasmid shares a strong similarity in ORF content and synteny with Gle4293 (Fig. 1), which seems to be the most common plasmid type found in Gracilariaceae.

The other plasmid of *Gracilaria vermiculophylla*, Gve7464, contains two genes of the proteobacterial operons related to leucine biosynthesis *leu*C and *leu*D genes (Fig. 1) and this is the first record of these genes in a eukaryotic plasmid. The confirmation that this plasmid is not a bacterial plasmid contaminant is that it contains Rhodophyta exclusive plasmid ORFs: FSO1, FSO2, and FSO4 (Fig. 1, Table 1). Three other species have *leu*C/*leu*D genes in their chloroplast genome: *G. tenuistipitata* var. *liui*, *G. chilensis*, and *Melanthalia intermedia* (Hagopian et al. 2004, Janouškovec et al. 2013, Lee et al. 2016b). This operon is potentially functional and it was probably originated from horizontal gene transfers involving protobacterial donors (Janouškovec et al. 2013). The ML phylogenies of *leu*C and *leu*D (Figs. S2 and S3) using encoded-chloroplast/plasmid genes and nucleus-encoded genes clearly show that there are two origins for those genes in Rhodophyta. It is interesting that the species that have chloroplast/plasmid-encoded genes do not seem to have a functional nucleus-encoded gene. *Melanthalia intermedia* had pseudogenes. The fact that there are more copies of chloroplast and plasmid genomes than nucleus genomes may lead to a prevalence of the chloroplast/plasmid-encoded genes in these species and the nucleus-encoded genes may be lost in the nuclear genome.

The origin and function of most red algae plasmid genes are unknown and difficult to determine, because, so far, the only match observed for those genes is with organellar genomes and plasmids (Lee et al. 2016b). Beside this, the copy number and the position of the homologous plasmid-derived ORFs in the organellar genomes are inconsistent in Rhodophyta phylogenies (Lee et al. 2016b, Ng et al. 2017). It is possible that the plasmids may have been integrated into the organellar genomes randomly during the evolutionary history of the organisms and this incorporation into the maternally inherited organelles may have rendered their fixation in a population (Lee et al. 2016b, Ng et al. 2017). However, most plasmid-derived genes in organellar genomes seem to be pseudogenes

because they used to be smaller and highly divergent compared with the genes in the plasmids. Therefore, apparently, the plasmid genes lost their function subsequent to being inserted in the organellar genomes. The only exception is the *leu*C and *leu*D genes, which seem to be functional. Plasmids are analogous to transposable elements with regard to mobility and can contribute to the gain or loss of genes in the genomes (Lee et al. 2016b, Ng et al. 2017). It is not possible to know whether these genes were inserted in the chloroplast genome from a Rhodophyta plasmid or if the plasmid that contained these genes was disintegrated from the chloroplast genome.

Genomic architecture is useful to understand the nature and source of change that occurred during the evolution of organisms. Organellar genomes are mostly highly conserved in Rhodophyta, and also have variable regions originated by plasmid horizontal gene transfer, which can be useful for comparisons at species or population levels. Plasmids have been recognized as mobile elements, but their origin and detailed process of horizontal gene transfer into organellar genomes of red algae remain unclear (Lee et al. 2016b). Moreover, Gracilariaceae is a good model group for studying the impact of PDS in genome evolution due to the significant presence of these sequences in organellar genomes.

We are grateful to R. Petti for technical support in the cultivation of Gracilariaceae specimens and to A.G.C. Martins for bioinformatics support. We thank E. Plastino and L. Ayres-Ostrock for collecting the *Gracilaria caudata* specimen; and Suzanne Fredericq and Frederico Gurgel for providing *Gracilaria vermiculophylla* and *M. intermedia* samples. We thank the Core Facility for Scientific Research – University of São Paulo (CEFAP-USP/GENIAL) and its staffs, S.I.S. Vançan, T.A. Souza and J. Gaiarsa, for support during DNA library preparation and informatics. We also thank the Human Genome and Stem Cell Research Center – University of São Paulo (HUG-CELL) for Illumina sequencing. CI is grateful to CNPq (152939/2014-8) and CAPES (88881.134422/2016-01) for scholarships. GML is thankful for funding from FAPESB (TO INT0001/2016). MCO is thankful for funding from CNPq (301491/2013-5; 406351/2016-3), and Biota-FAPESP (2013/11833-3). We thank SPRINT bilateral grant between FAPESP (2015/50078-1) and the University of Melbourne.

Abbott, I. A., Junfu, Z. & Bangmei, X. 1991. *Gracilaria mixta*, sp. nov. and other Western Pacific species of the genus (Rhodophyta: Gracilariaceae). *Pacific Sci.* 45:12–27.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10.

Andrews, S. 2011. FASTQC: A Quality Control Tool for High Throughput Sequence Data. Babraham Institute, Cambridge, UK. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (last accessed 23 October 2017).

Bellorin, A. M., Oliveira, M. C. & Oliveira, E. C. 2002. Phylogeny and systematics of the marine algal family Gracilariaceae (Gracilariales, Rhodophyta) based on small subunit rDNA and ITS sequences of Atlantic and Pacific species. *J. Phycol.* 38:551–63.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–20.

Bushnell, B. 2014. BBTools software package. Available at: https://jgi.doe.gov/data-and-tools/bbtools (last accessed 23 October 2017).

Campbell, M. A., Presting, G., Bennett, M. S. & Sherwood, A. R. 2014. Highly conserved organellar genomes in the Gracilariales as inferred using new data from the Hawaiian invasive alga *Gracilaria salicornia* (Rhodophyta). *Phycologia* 53:109–16.

Chernomor, O., von Haeseler, A. & Minh, B. Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65:997–1008.

Costa, J. F., Lin, S., Macaya, E. C., Fernández-garcía, C. & Verbruggen, H. 2016. Chloroplast genomes as a tool to resolve red algal phylogenies: a case study in the Nemaliales. *BMC Evol. Biol.* 16:205.

Costa, E. S., Plastino, E. M., Petti, R., Oliveira, E. C. & Oliveira, M. C. 2012. The Gracilariaceae Germplasm Bank of the University of São Paulo, Brazil–A DNA barcoding approach. *J. Appl. Phycol.* 24:1643–53.

Darling, A. E., Mau, B. & Perna, N. T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147.

Díaz-Tapia, P., Maggs, C. A., West, J. A. & Verbruggen, H. 2017. Analysis of chloroplast genomes and a supermatrix inform reclassification of the Rhodomelaceae (Rhodophyta). *J. Phycol.* 53:920–37.

Faugeron, S., Valero, M., Destombe, C., Martínez, E. A. & Correa, J. A. 2001. Hierarchical spatial structure and discriminant analysis of genetic diversity in the red alga *Mazzaella laminarioides* (Gigartinales, Rhodophyta). *J. Phycol.* 37:705–16.

Goff, L. J. & Coleman, A. W. 1988. The use of plastid DNA restriction endonuclease patterns in delineating red algal species and populations. *J. Phycol.* 24:357–68.

Goff, L. J. & Coleman, A. W. 1990. Red algal plasmids. *Curr. Genet.* 18:557–65.

Guiry, M. D. & Guiry, G. M. 2017. Algaebase. Available at: http://www.algaebase.org/ (last accessed 23 October 2017).

Gurgel, C. F. D. & Fredericq, S. 2004. Systematics of the Gracilariaceae (Gracilariales, Rhodophyta): a critical assessment based on *rbc*L sequence analyses. *J. Phycol.* 40:138–59.

Hagopian, J. C., Reis, M., Kitajima, J. P., Bhattacharya, D. & Oliveira, M. C. 2004. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J. Mol. Evol.* 59:464–77.

Hancock, L., Goff, L. & Lane, C. 2010. Red algae lose key mitochondrial genes in response to becoming parasitic. *Genome Biol. Evol.* 2:897–910.

Harrison, E. & Brockhurst, M. A. 2012. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* 20:262–7.

Janoušković, J., Liu, S. L., Martone, P. T., Carré, W., Leblanc, C., Collén, J. & Keeling, P. J. 2013. Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PLoS ONE* 8:e59001.

Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–80.

Kim, K. M., Park, J. H., Bhattacharya, D. & Yoon, H. S. 2014. Applications of next-generation sequencing to unravelling the evolutionary history of algae. *Int. J. Syst. Evol. Microbiol.* 64:333–45.

Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29:1695–701.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth.* 9:357–9.

Leblanc, C., Boyen, C., Richard, O., Bonnard, G., Grienenberger, J. & Kloareg, B. 1995. Complete Sequence of the Mitochondrial DNA of the Rhodophyte Chondrus crispus (Gigartinales). Gene Content and Genome Organization. *J. Mol. Biol.* 250:484–95.

Lee, J. M., Boo, S. M., Mansilla, A. & Yoon, H. S. 2015. Unique repeat and plasmid sequences in the mitochondrial genome of *Gracilaria chilensis* (Gracilariales, Rhodophyta). *Phycologia* 54:20–3.

Lee, J., Cho, C. H., Park, S. I., Choi, J. W., Song, H. S., West, J. A., Bhattacharya, D. & Yoon, H. S. 2016a. Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. *BMC Biol.* 14:75.

Lee, J., Kim, K. M., Yang, E. C., Miller, K. A., Boo, S. M., Bhattacharya, D. & Yoon, H. S. 2016b. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. *Sci. Rep.* 6:23744.

Lin, H. H. & Liao, Y. C. 2016. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* 6:24175.

Lonardi, S., Mirebrahim, H., Wanamaker, S., Alpert, M., Ciardo, G., Duma, D. & Close, T. J. 2015. When less is more: "slicing" sequencing data improves read decoding accuracy and de novo assembly quality. *Bioinformatics* 31:2972–80.

Lyra, G. M., Costa, E. S., Jesus, P. B., Matos, J. C. G., Caires, T. A., Oliveira, M. C., Oliveira, E. C., Xi, Z., Nunes, J. M. C. & Davis, C. C. 2015. Phylogeny of Gracilariaceae (Rhodophyta): evidence from plastid and mitochondrial nucleotide sequences. *J. Phycol.* 51:356–66.

Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30:1188–95.

Moon, D. A. & Goff, L. J. 1997. Molecular characterization of two large DNA plasmids in the red alga *Prophyra pulchra*. *Curr. Genet.* 32:132–8.

Ng, P. K., Lin, S. M., Lim, P. E., Liu, L. C., Chen, C. M. & Pai, T. W. 2017. Complete chloroplast genome of *Gracilaria firma* (Gracilariaceae, Rhodophyta), with discussion on the use of chloroplast phylogenomics in the subclass Rhodymeniophycidae. *BMC Genom.* 18:40.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32:268–74.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A. et al. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. *In* Deng, M., Jiang, R., Sun, F. & Zhang, X. [Eds.] *Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 158–70.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41:590–6.

Salomaki, E. D. & Lane, C. E. Red Algal Mitochondrial Genomes Are More Complete than Previously Reported. *Genome Biol. Evol.* 9:48–63.

Smit, A. J. 2004. Medicinal and pharmaceutical uses of seaweed natural products: a review. *J. Appl. Phycol.* 16:245–62.

Song, S. L., Yong, H. S., Lim, P. E., Ng, P. K. & Phang, S. M. 2016. Complete mitochondrial genome, genetic diversity and molecular phylogeny of *Gracilaria salicornia* (Rhodophyta: Gracilariaceae). *Phycologia* 55:371–7.

Song, S. L., Yong, H. S., Lim, P. E. & Phang, S. M. 2017. Complete mitochondrial genome of Gracilaria changii (Rhodophyta: Gracilariaceae). *J. Appl. Phycol.* 29:2129–34.

Ursi, S. & Plastino, E. M. 2001. Crescimento in vitro de linhagens de coloração vermelha e verde clara de *Gracilaria* sp. (Gracilariales, Rhodophyta) em dois meios de cultura: análise de diferentes estádios reprodutivos. *Rev. Bras. Bot.* 24:587–94.

Villemur, R. 1990a. The DNA sequence and structural organization of the GC2 plasmid from the red alga *Gracilaria chilensis*. *Plant Mol. Biol.* 15:237–43.

Villemur, R. 1990b. Circular plasmid DNAs from the red alga *Gracilaria chilensis*. *Curr. Genet.* 18:251–7.

Yang, E. C., Kim, K. M., Kim, S. Y., Lee, J., Boo, G. H., Lee, J. H., Nelson, W. et al. 2015. Highly conserved mitochondrial genomes among multicellular red algae of the Florideophyceae. *Genome Biol. Evol.* 7:2394–406.

Yang, E. C., Kim, K. M., Kim, S. Y. & Yoon, H. S. 2014. Complete mitochondrial genome of agar-producing red alga *Gracilariopsis chorda* (Gracilariales). *Mitochondrial DNA* 25:339–41.

Zemke-White, W. L. & Ohno, M. 1999. World seaweed utilization: an end-of-century summary. *J. Appl. Phycol.* 11:369–76.

Zhang, Y., Guo, Y., Li, T., Chen, C., Shen, K. & Hsiao, C. 2016. The complete chloroplast genome of *Gracilariopsis lemaneiformis*, an important economic red alga of the family Gracilariaceae. *Mitochondrial DNA Part B: Resources* 1:2–3.

Zhang, L., Wang, X., Qian, H., Chi, S., Liu, C. & Liu, T. 2012. Complete sequences of the mitochondrial DNA of the wild *Gracilariopsis lemaneiformis* and two mutagenic cultivated breeds (Gracilariaceae, Rhodophyta). *PLoS ONE* 7:e40241.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Figure S1. Gene synteny of the plasmid-derived sequence regions found in chloroplast genomes. These regions are identified in orange line box in Figure 3A.

Figure S2. Maximum likelihood phylogeny of *leu*C. Numbers at nodes indicate bootstrap replicates. Bootstrap support values are shown only above 50%.

Figure S3. Maximum likelihood phylogeny of *leu*D. Numbers at nodes indicate bootstrap replicates. Bootstrap support values are shown only above 50%.

Table S1. Genome sequencing statistics of species analyzed in this study. All specimen herbaria vouchers are deposited at SPF herbarium at University of São Paulo, São Paulo, Brazil.

Table S2. Notes about annotations done in published genomes in Genbank. "c" indicated reverse strand.

Table S3. General information of plasmid sequences used in present study.

Table S4. List of Gracilariaceae protein-coding genes used in the ML tree search.

Table S5. Mitochondrial genes from species used in present study. Gene present is marked with color and "+" or name. Gene absent is marked in white. Genes as generally arranged by different color. Line indicated same locus.

Table S6. Chloroplast genes from species used in present study. Gene present is marked with color and "+" or name. Gene absent is marked in white. Genes as generally arranged by different color. Line indicated same locus, except to unique ORFs. Asterisk (*) indicated pseudogene.