Re-annotation, improved large-scale assembly and establishment of a catalogue of non-coding loci for the genome of the model brown alga *Ectocarpus*

Alexandre Cormier[1], Komlan Avia[1], Lieven Sterck[2,3,4], Thomas Derrien[5], Valentin Wucher[5], Gwendoline Andres[6], Misharl Monsoor[6], Olivier Godfroy[1], Agnieszka Lipinska[1], Marie-Mathilde Perrineau[1], Yves Van De Peer[2,3,4,7], Christophe Hitte[5], Erwan Corre[6], Susana M. Coelho[1], J. Mark Cock[1*]

[1]Sorbonne Université, UPMC Univ Paris 06, CNRS, Algal Genetics Group, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff, France, [2]Department of Plant Systems Biology,VIB, B-9052 Ghent, Belgium, [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9000 Ghent, Belgium, [4]Bioinformatics Institute Ghent, Technologiepark 927, 9052 Ghent, Belgium, [5]IGDR CNRS-UMR6290 – Université Rennes 1, Rennes, France, [6]Abims Platform, CNRS-UPMC, FR2424, Station Biologique de Roscoff, CS 90074, 29688 Roscoff, France, [7]Department of Genetics, Genomics Research Institute, University of Pretoria, 0028 Pretoria, South Africa.

*Author for correspondence: Tel: 33 (0)2 98 29 23 60; Email: cock@sb-roscoff.fr

Brief heading: Re-annotation of the genome of the model brown alga *Ectocarpus*

Key words: Alternative splicing, Brown algae, *Ectocarpus*, Genetic markers, Genome reannotation, Long non-coding RNAs, *Saccharina japonica*, Stramenopile

Summary

- The genome of the filamentous brown alga *Ectocarpus* was the first to be completely sequenced from within the brown algal group and has served as a key reference genome both for this lineage and for the stramenopiles.
- We present a complete structural and functional reannotation of the *Ectocarpus* genome.

- The large-scale assembly of the *Ectocarpus* genome was significantly improved and genome-wide gene re-annotation using extensive RNA-seq data improved the structure of 11,108 existing protein-coding genes and added 2,030 new loci. A genome-wide analysis of splicing isoforms identified an average of 1.6 transcripts per locus. A large number of previously undescribed non-coding genes were identified and annotated, including 717 loci that produce long non-coding RNAs. Conservation of lncRNAs between *Ectocarpus* and another brown alga, the kelp *Saccharina japonica*, suggests that at least a proportion of these loci serve a function. Finally, a large collection of SNP-based markers was developed for genetic analyses. These resources are available through an updated and improved genome database.
- This study significantly improves the utility of the *Ectocarpus* genome as a high-quality reference for the study of many important aspects of brown algal biology and as a reference for genomic analyses across the stramenopiles.

Introduction

Ectocarpus has been studied since the nineteenth century and work on this organism has provided many insights into novel aspects of brown algal biology (Müller, 1967; Charrier *et al.*, 2008). This long research history, together with several features of the organism that make it well adapted for genetic and genomic approaches (Coelho *et al.*, 2012a), led to it being proposed as a general model organism for the brown algae in 2004 (Peters *et al.*, 2004) and to the initiation of a genome sequencing project that produced a first complete genome assembly in 2010 (Cock *et al.*, 2010). The publication of the genomic sequence was followed up with the development of many additional tools and resources including a genetic map (Heesch *et al.*, 2010), gene mapping techniques, microarrays (Dittami *et al.*, 2009; Coelho *et al.*, 2011), transcriptomic data (Ahmed *et al.*, 2014; Lipinska *et al.*, 2015), proteomic techniques (Ritter *et al.*, 2008) and bioinformatics tools (Gschloessl *et al.*, 2008; Prigent *et al.*, 2014). These genomic resources are currently being exploited to further our understanding of a broad range of processes, including life cycle regulation (Coelho *et al.*, 2011), sex determination (Lipinska *et al.*, 2013, 2015; Ahmed *et al.*, 2014), development and morphology (Le Bail *et al.*, 2011), interactions with pathogens (Zambounis *et al.*, 2012) and metabolism (Meslet-Cladière *et al.*, 2013; Prigent *et al.*, 2014).

The brown algae are an important taxonomic group for several reasons; they are key primary producers in many coastal ecosystems and have a major influence on marine biodiversity and

ecology (Dayton, 1985; Steneck *et al.*, 2002; Bartsch *et al.*, 2008; Klinger, 2015; Wahl *et al.*, 2015). Brown algae also represent an important resource of considerable commercial value (Kijjoa & Sawangwong, 2004; Smit, 2004; Hughes *et al.*, 2012) and industrial exploitation of these organisms has increased markedly in recent years with the expansion of aquaculture activities, particularly in Asia (Tseng, 2001). Finally, brown algae are also of phylogenetic interest because they are very distantly related to well-studied groups such as the animals, fungi and land plants and, moreover, have evolved complex multicellularity independently of these other lineages (Cock *et al.*, 2010; Cock & Collén, 2015). Comparative analyses between brown algae and members of these other eukaryotic supergroups therefore potentially provide a means to explore deep evolutionary events of broad, general importance.

A high-quality genome resource is essential if these important features of the brown algae are to be investigated effectively. The version of the *Ectocarpus* genome that was published in 2010 (Cock *et al.*, 2010) included detailed manual annotations of many of the genes but gene structure predictions were based on a limited amount of transcriptomic data (Sanger expressed sequence tags) and the large-scale assembly of the sequence contigs only associated about 70% of the genome sequence with linkage groups. Moreover, annotation efforts had focused almost exclusively on protein-coding genes, largely ignoring the non-coding component of the genome. The study described here set out to address these shortfalls, exploiting the large amount of transcriptomic data now available and using recently developed genetic and bioinformatic approaches to improve both the assembly and annotation of the genome. A high-density, RAD-seq-based genetic map was used to anchor sequence scaffolds onto the chromosomes, considerably improving the large-scale assembly of the genome. In addition, a complete reannotation of the genome was carried out based on extensive RNA-seq data. This updated version of the genome annotation includes information about transcript isoforms and integrates non-coding loci such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs). Finally, we report additional resources including a genome-wide set of single nucleotide polymorphisms for genetic mapping and improvements to the genome database such as the addition of a JBrowse-based genome browser that allows multiple types of genome-wide data to be visualised simultaneously.

Materials and Methods

Biological material

Ectocarpus strains were cultured as described previously (Coelho *et al.*, 2012b). The male genome sequenced strain Ec32 (reference CCAP 1310/4 in the Culture Collection of Algae and Protozoa, Oban, Scotland) is a meiotic offspring of a field sporophyte, Ec17, collected in 1988 in San Juan de Marcona, Peru (Peters *et al.*, 2008). Ec722 is a UV-mutagenised descendant of Ec32. The female outcrossing line Ec568 is derived from a sporophyte collected in Arica in northern Chile (Heesch *et al.*, 2010).

RNA-seq

The analyses carried out in this study used RNA-seq data generated for biological replicate (duplicate) samples of partheno-sporophytes and of both young and mature samples for both male and female gametophytes (ten libraries in all). The production of the young (Lipinska *et al.*, 2015) and mature (Ahmed *et al.*, 2014) gametophyte RNA-seq data has been described previously. For each of the replicate partheno-sporophyte samples, total RNA was extracted and used as a template by Fasteris (CH-1228 Plan-les-Ouates, Switzerland) to synthesise cDNA using an oligo-dT primer. The cDNA libraries were sequenced with Illumina HiSeq 2000 technology to generate 100 bp single-end reads. Data quality was assessed using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and the reads were trimmed and filtered using a quality threshold of 25 (base calling) and a minimal size of 60 bp. Only reads in which more than 75% of nucleotides had a minimal quality threshold of 20 were retained. Table S1 shows the number of raw reads generated per sample and the number of reads remaining after trimming and filtering (cleaned reads). The cleaned reads were mapped to the *Ectocarpus* sp. genome (Cock *et al.*, 2010) (available at Orcae Sterck *et al.*, 2012) using Tophat2 and the Bowtie2 aligner (Kim *et al.*, 2013). More than 90% of the sequencing reads for each library mapped to the genome.

De novo assembly of the pooled RNA-seq data from the ten libraries was carried out using Trinity (Grabherr *et al.*, 2011) in normalized mode with default parameters. Weakly expressed transcripts (isoform percentage <1 and RPKM <1) were removed from the dataset. The remaining transcripts were aligned against the *Ectocarpus* reference genome (Ec32) using GenomeThreader (Gremme *et al.*, 2005) with a maximum intron length of 26,000 bp, a minimum coverage of 75% and a minimum alignment score of 90%.

Gene prediction

Gene prediction was carried out using the EuGene program (Foissac *et al.*, 2008), as described previously (Cock *et al.*, 2010). Alignments of the Trinity RNA-seq-derived transcripts against the *Ectocarpus* sp. reference genome were added to the EuGene pipeline in addition to the data used for the v1 annotation, which included splice site predictions generated by SpliceMachine (Degroeve *et al.*, 2005) and *Ectocarpus* Sanger EST data. The new set of EuGene gene structure predictions were compared with the gene structures of the v1 annotation using AEGeAn (Standage & Brendel, 2012) and a combination of automated and manual approaches was used to select the optimal gene structures. Briefly, automatic validation of new predictions was applied for genes where there were modifications to the UTRs, where additional exons were added or where there were modifications to the detailed structure of existing exons. In cases where the new model predicted exon lost, the prediction was retained only if there was 65% similarity between the reference and the new model. This threshold was reduced to 30% similarity when the reference gene had only 4 exons or less. A subset of about one hundred genes for each class was manually reviewed to validate the automatic selection of gene structures. GenomeView (Abeel *et al.*, 2012) was used to visualise RNA-seq read mapping information.

Manual annotation

The v2 annotation took into account the functional and structural annotation of 325 and 410 genes, respectively, carried out through the Orcae database (Sterck *et al.*, 2012) since the publication of the v1 annotation. Many of the structural annotations were based on the same set of RNA-seq data that was used for the genome-wide gene structure prediction but exploited transcripts that had been generated using a reference-guided approach with Tophat2 and Cufflinks2 (Trapnell *et al.*, 2010; Kim *et al.*, 2013). Tophat2 was able to map 92% of the cleaned reads to the genome sequence and 36,565 transcripts were assembled by Cufflinks2 (including multiple transcripts for some loci) using the mapping information and the initial gene models as guides.

Annotation of gene functions

Putative functions were assigned to the v2 genes based on the identification of protein domains using InterProScan, which carried out searches against all its component databases (Jones *et al.*, 2014). Gene ontology categories were assigned using Blast2GO (Conesa *et al.*,

2005). For genes where a manually assigned function was already available (3,442 genes), the InterProScan-based prediction was compared manually with the existing annotation and the most relevant annotation retained.

Detection of alternative transcripts

To detect alternative transcripts of the set of 17,418 protein-coding loci, 507,634,855 million reads of RNA-seq data corresponding to diverse tissues and life cycle stages (Table S1) were mapped to the *Ectocarpus* genome using Bowtie2 (Langmead *et al.*, 2009) and transcripts were predicted genome-wide using Stringtie (Pertea *et al.*, 2015) with default parameters, guided by the annotation file from the v2 annotation. A Stringtie prediction was made for each library based on TopHat2 mapping files. The results were merged using Cuffmerge (Trapnell *et al.*, 2010). Cuffcompare was used to assign the predicted transcripts to the reference genes. Transcripts with 3' UTRs > 9300 bp and/or 5' UTRs > 2500 bp were discarded. Only potential isoforms (class code = J, O and C) were retained. Prediction of the coding regions of the alternative transcripts was carried out using Transdecoder (Haas *et al.*, 2013). ORF predictions were filtered to retain complete coding sequences with both initiation and stop codons. The longest ORF was retained for each transcript.

A global classification and quantification of the different types of alternative splicing that generated the transcript isoforms was obtained using SplAdder (Kahles *et al.*, 2016) based on the mapping of the pooled RNA-seq data.

Detection of non-protein-coding genes

The detection of microRNA, ribosomal RNA and snoRNA loci has been described previously (Tarver *et al.*, 2015).

Ectocarpus lncRNA loci were detected using FEELnc (https://github.com/tderrien/FEELnc) with default parameters and the output transcripts of the Stringtie analysis described in the previous section. The same specificity threshold (0.97) was used for both protein-coding and non-coding transcripts to predict lncRNA loci. Transcripts overlapping annotated protein-coding genes (v2 annotation) were eliminated and a random forest approach based on ORF coverage (i.e. length of the longest ORF / length of the lncRNA transcript), transcript size and k-mer frequency was implemented to classify the remaining transcripts as mRNAs or lncRNAs. Loci with mono-exonic transcripts were eliminated to limit the inclusion of false positive loci due to read mapping ambiguity. An

arbitrary minimum size of 200 nt was applied to eliminate loci encoding small RNA transcripts. FEELnc also classifies the predicted lncRNA loci by determining 1) if they overlap (genic) or not (intergenic) with the nearest gene on the genome, designated the adjacent gene (and which can be a protein-coding gene or small-RNA-encoding locus), 2) if genic lncRNAs overlap with intron or exon regions of the adjacent gene and in which orientation, sense or antisense, and 3) how intergenic lncRNAs are orientated with respect to the adjacent gene (within 10 kbp) on the chromosome (same strand, convergent or divergent).

A similar approach was used to detect *S. japonica* lncRNA loci. For this genome, the Stringtie transcript prediction used as input for FEELnc was based on mapping of 220,551,196 million RNA-seq reads to the *S. japonica* genome (Ye *et al.*, 2015). The RNA-seq data corresponded to female gametes (127,607,414 reads, accession number SRR2064656), spores (30,552,978 reads, accession number SRR2064654), thalli grown under blue light (11,981,830 reads, accession number SRR371552) or in the dark (12,657,652 reads, accession number SRR371551), young sporophytes grown under blue (13,333,334 reads, accession number SRR496757) or white (17,181,148 reads, accession number SRR496799) light and thalli subjected to heat stress (7,236,840 reads, accession number SRR947066). Orthologous *Ectocarpus* and *S. japonica* lncRNA loci were detected by carrying out reciprocal Blastn searches (E-value < 10^{-4}). Alignments of lncRNA sequences were carried out with SIM (http://web.expasy.org/sim/) and visualised with Lalnview (Duret *et al.*, 1996).

DESeq2 with default parameters was used to detect *Ectocarpus* lncRNA and protein-coding loci that were differently expressed in sporophyte basal versus upright filaments.

Genome-wide identification of sequence variants

Genome sequence data was generated for the female outcrossing line Ec568 using Illumina HiSeq2500 technology (Fasteris, Switzerland), which produced 25,976,388,600 bp of 2x100 bp paired-end sequence. Sequence variants were detected as described previously (Godfroy *et al.*, 2015).

To determine whether sequence variants behaved as Mendelian loci, a cross between a UV-mutagenised derivative of the reference genome strain Ec32 (strain Ec722) and the female outcrossing line Ec568 (Heesch *et al.*, 2010) was used to generate a population of 180 progeny each corresponding to an independent meiotic event, segregating the two parental alleles of each variant locus. Two libraries were constructed with pools of 84 and 96 haploid,

partheno-sporophyte individuals and sequenced using Illumina HiSeq2500 technology (Fasteris, Switzerland) to generate 20,785,058,400 bp and 23,429,143,400 bp of 2x100 bp paired-end sequence, respectively. Sequence variants were detected in each dataset as described previously (Godfroy *et al.*, 2015) and VarScan was used to identify SNPs shared by the two pools of haploid individuals. For each of these SNPs the sum of the variant frequencies observed in the two pools was calculated, and only those for which this sum was between 0.8 and 1.2 were retained. VarScan compare was then used to extract the Ec568 variants from the list of Mendelian segregating SNPs.

Database curation of the v2 annotation

A Genome Browser was implemented based on Jbrowse (Buels *et al.*, 2016) using a Chado database (Mungall & Emmert, 2007). The browser integrates both v1 and v2 reference gene models, raw gene models predicted by EuGene, transcripts predicted by Cufflinks and EST and RNA-seq read data.

Accession numbers

The accession numbers for the sequence data used in this article are given in supplementary Table S1.

Results

Improved chromosome-scale assembly of the *Ectocarpus* genome

A microsatellite-based genetic map (Heesch *et al.*, 2010) was originally used to produce a large-scale assembly of the *Ectocarpus* genome consisting of 34 pseudo-chromosomes (Cock *et al.*, 2010) corresponding to the 34 linkage groups of the genetic map. The pseudo-chromosomes were constructed by concatenating sequence scaffolds based on the genetic order of sequence-anchored microsatellite markers on the genetic map (Cock *et al.*, 2010). However, due to the low density of the markers, the large-scale assembly included only 325 of the 1,561 sequence scaffolds (70.1% of the total sequence length) and, moreover, only 40 (12%) of the mapped scaffolds could be orientated relative to the chromosome (i.e. only 12% of the scaffolds contained at least two microsatellite markers which recombined relative to each other).

To improve the large-scale assembly of the *Ectocarpus* genome, we took advantage of a high-density, single nucleotide polymorphism (SNP)-based genetic map that has recently

been generated using a Restriction site associated DNA (RAD)-seq method (K. Avia, personal communication). The 3,588 SNP markers used to construct the genetic map were mapped to sequence scaffolds and the recombination information for these markers used to construct a new set of pseudo-chromosomes (Fig. 1). The new large-scale assembly represents a significant improvement because it integrates 531 of the 1,561 sequence scaffolds onto genetic linkage groups (90.5% of the total sequence length) and 49% of these scaffolds have been orientated with respect to their chromosome. Moreover, the high-density genetic map has allowed several fragmented linkage groups / pseudo-chromosomes to be fused, reducing the total number from 34 to 28. The exact number of chromosomes in *Ectocarpus* sp. strain Ec32 is not known but cytogenetic analysis of another *Ectocarpus* species, *E. siliculosus* indicated the presence of approximately 25 chromosomes (Müller, 1966, 1967).

Reannotation of gene structure based on RNA-seq data

The initial set of *Ectocarpus* gene models (referred to hereafter as the v1 annotation) was generated using EuGene (Foissac *et al.*, 2008) based on a limited amount of transcriptomic information (91,041 Sanger expressed sequence tags, ESTs; Cock *et al.*, 2010) and therefore involved a significant amount of *de novo* prediction. The v1 annotation has been gradually improved since 2010 by the addition of 325 functional and 410 structural annotations for individual genes through the Orcae database (Sterck *et al.*, 2012). This gene-by-gene approach improved the quality of the annotation of a number of selected genes but it was necessary to extend the approach to improve annotation quality across the whole genome.

A genome-wide reannotation, hereafter referred to as the v2 annotation, was therefore carried out based on the analysis of 642 million reads of RNA-seq data from ten different libraries (Ahmed *et al.*, 2014; Lipinska *et al.*, 2015 and this study; Table S1). This data was assembled into 34,551 *de novo* transcripts using Trinity (Grabherr *et al.*, 2011). GenomeThreader (Gremme *et al.*, 2005) was able to align 91% of these transcripts to the genome. Gene prediction for the v2 annotation was then carried out using EuGene and the 34,551 *de novo* transcripts, along with 83,502 Sanger ESTs and SpliceMachine (Degroeve *et al.*, 2005) splice site predictions. The 21,958 preliminary gene models generated by this prediction were then compared with the 16,256 genes of the v1 annotation (Cock *et al.*, 2010) using AEGeAn (Standage & Brendel, 2012) and a combination of automatic and manual criteria were used to evaluate the predictions and select the optimal gene model for each

294 locus. This genome-wide reannotation integrated the results of the manual gene-by-gene
295 annotation carried out since publication of the v1 annotation by preferentially retaining high
296 quality, expert functional and structural annotations.
297 The 21,958 preliminary gene predictions included 1) genes that were identical to the v1
298 prediction (10,426 genes), 2) genes that were structurally different to their v1 counterpart
299 (6,295 genes) and 3) novel loci that were not predicted by the v1 annotation (5,237 genes).
300 For the first set of genes, the v1 gene models were replaced with the RNA-seq-based models,
301 providing considerable additional information about the UTR structure of the genes (e.g. Fig.
302 2A). When the RNA-seq-based prediction differed from the v1 model, manual inspection was
303 used to select the optimal model for each locus (e.g. Fig. 2B; see Methods and Materials for
304 details). This second set of genes also included predictions which indicated that v1 annotation
305 genes needed to be fused (e.g. Fig. 2C) or split (e.g. Fig. 2D). Novel RNA-seq-based
306 predictions, not present in the v1 annotation, were filtered to remove probable false positives.
307 Predictions were retained only if 1) their transcripts had an abundance of >1 RPKM across
308 the entire (merged) set of RNA-seq data, 2) the start codon of the gene was not located in a
309 repeated region (to exclude transposon-derived ORFs; Yandell & Ence, 2012) and 3) their
310 coding region was >100 bp. After applying these filters, 2,030 of the new predictions were
311 retained and integrated into the v2 annotation.
312 Overall, the addition of these new genes and updates to the existing genes (fusing or
313 splitting existing gene models) brought the total number of genes in the v2 annotated genome
314 to 17,418 (Table 1). The transition from the v1 to the v2 version of the genome annotation
315 involved the modification of 11,108 of the v1 gene models, of which 5,336 were altered
316 within their coding regions (Table 2). Of the former, 784 involved gene fusions (to produce
317 404 genes in the v2 annotation), 19 involved splitting v1 annotation gene predictions (to
318 create 38 genes in the v2 annotation) and 123 genes were removed. The v2 annotation now
319 includes coordinates for at least one of the UTR regions for 78.7% of the 17,418 genes
320 (compared to 52.6% for the v1 annotation; Fig. 3, Table 1). The v2 annotation is publically
321 available through the ORCAE database
322 (http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi; Sterck *et al.*, 2012).
323 The *Ectocarpus* genome database was modified to take into account the large-scale
324 assembly of the sequence scaffolds. In particular, the sequentially numbered locusIDs were
325 modified to indicate sequencial position on the pseudochromosome. The correspondence
326 between the LocusIDs of the v1 and v2 annotations is given in Table S2 and is also available

as a download from the genome database (https://bioinformatics.psb.ugent.be/gdb/ectocarpusV2/).

Prediction of gene function

The final 17,418 genes of the v2 annotation were further analysed to improve the prediction of gene function by comparing protein sequences with the InterPro database using InterProScan (Jones *et al.*, 2014) and by using Blast2GO (Conesa *et al.*, 2005) to assign gene ontology (GO) categories. This process allowed functional annotations and GO categories to be assigned to 10,688 and 7,383 of the 17,418 v2 annotation genes, respectively (compared with 5,583 and 5,989, respectively, for the v1 annotation; Table 1). Of the 2,030 genes that were present in the v2 annotation but not the v1 annotation, 212 had matches in the public databases and 135 and 79 were assigned functional annotations and GO categories, respectively.

Alternative splicing

A previous search for alternative gene transcripts based on the 91,041 Sanger ESTs detected isoforms for only a small percentage (2.9%) of the *Ectocarpus* genes (Cock *et al.*, 2010). Here we carried out an updated search for alternative transcripts using the available RNA-seq data (Table S1). The analysis focused on transcript isoforms with alternatively spliced coding regions because variants of this type are more likely to have biological roles through the production of variant protein products. A total of 10,723 alternative transcripts of this type were detected genome-wide, associated with 7,362 (42.3%) of the 17,418 protein-coding genes. This corresponded to an average of 1.62 transcripts per locus.

Whilst alternative splicing of gene transcripts can potentially lead to the production of two or more protein products with different biological activities from a single genetic locus, this is not necessarily the case and alternative transcripts can also represent spliceosomal errors or correspond to variants that do not differ significantly from the principal transcript in terms of transcript functionality. To assess the extent to which alternative splicing has the potential to impact gene function in *Ectocarpus*, we used Interproscan (Jones *et al.*, 2014) to compare the domain structures of the predicted protein products of the principal and alternative gene transcripts of the 7,362 genes that exhibited alternative splicing of their coding region. This analysis indicated that, on average, each isoform lacked about 21% of the domains that were detected in the principal transcript. These marked differences between the domain structures

of the protein products of principal and alternative transcripts are likely to significantly modify the activities of the alternative protein products.

In addition to this genome-wide approach, a more detailed analysis was carried out for four genes that encoded proteins with multiple, repeated copies of small protein domains. Fig. 4 shows that the alternative transcripts of these genes encode multiple protein variants in which repeated domains are included or excluded from the protein product in different combinations, producing proteins with markedly different domain structures. Together with the genome-wide analysis described above, these analyses suggested that alternative splicing is used in *Ectocarpus* to combine protein domain modules to generate multiple protein isoforms from individual loci.

Analysis of the types of alternative splicing events that give rise to transcript isoforms in *Ectocarpus* using the program SplAdder (Kahles *et al.*, 2016) indicated that the most common event was the use of an alternative 3' acceptor site (Table 3). Intron retention events were relatively rare, representing less than 12% of the detected events.

Identification and integration of non-protein-coding genes

With the exception of tRNA loci (Cock *et al.*, 2010), the v1 annotation provided very little information about non-protein-coding genes. The v2 annotation includes considerably more information about this type of locus, in particular integrating 64 microRNA (miRNA) loci, nine ribosomal RNA loci (rRNA) and 610 of the small nucleolar RNA (snoRNA) loci recently predicted by Tarver et al. (2015). The rRNA and snoRNA loci are listed in Tables S3 and S4; information about the miRNA loci can be found in Tarver et al. (2015).

In vertebrates most snoRNAs are located in introns (Hoeppner & Poole, 2012) but this is not the case in all species and only about 30% of *Ectocarpus* snoRNAs are intronic. Work in other species has shown that the main function of snoRNAs is to direct chemical modification of other RNA molecules, particularly ribosomal RNAs (reviewed in Bratkovic & Rogelj, 2014). The two major classes of snoRNA, C/D box and H/ACA box, are principally involved in methylation and pseudouridylation of RNA molecules, respectively, but several alternative functions have been reported (Kehr *et al.*, 2014). *Ectocarpus* is predicted to have 95 C/D box and 515 H/ACA box snoRNAs. Note that the *Ectocarpus* snoRNAs were detected using ACAseeker and CDseeker and should therefore be considered predictions until their functions have been investigated experimentally.

A search of the *Ectocarpus* genome indicated that the core protein components that associate with both C/D and H/ACA box snoRNAs to form of sno-ribonucleoproteins (snoRNPs) are highly conserved in *Ectocarpus* (Table S5).

A screen was also carried out for potential long non-coding RNAs (lncRNAs) using the FEELnc lncRNA prediction pipeline (https://github.com/tderrien/FEELnc) and the RNA-seq data listed in Table S1. This analysis predicted the presence of 717 lncRNA loci in the *Ectocarpus* genome (Table S6), corresponding to a total of 952 different transcripts (1.3 isoforms per locus on average). The mean size of the lncRNA transcripts was 1,708 nucleotides and varied between 200 (the defined minimal size) and 7,988 nucleotides. The lncRNA loci were classified based on their configuration relative to the nearest protein-coding gene in the genome (referred to in the following text as the adjacent gene) and included both loci that were located entirely in an intergenic region (i.e. long intergenic non-coding RNAs or lincRNAs) and loci that overlapped with their adjacent gene (Fig. S1). About 45% of the lncRNAs were classed as lincRNAs. Expression analysis indicated that lncRNA transcripts were about eight-fold less abundant on average than those of protein-coding genes (Fig. 5). A similar difference in mean expression level has been observed in animal and land plant systems (Ulitsky & Bartel, 2013; Chekanova, 2015 and references therein). The *Ectocarpus* lincRNA loci tend to occur in regions of the genome of low gene density. The mean distance of lincRNA loci from flanking protein-coding genes is 8,654 bp, which is significantly longer (Wilcoxon test $P < 2.2e\text{-}6$) than the mean distance between protein-coding loci (4,154 bp).

To determine whether lncRNAs exhibited differential expression patterns in different tissues, we compared abundances of lncRNA transcripts in replicate samples of two different tissues of the sporophyte stage, the strongly adhering, prostrate filaments of the basal system and the upright filaments of the apical system (Peters *et al.*, 2008). DESeq2 identified 219 lncRNA loci that were differentially expressed between these two tissues, and 4,019 differentially expressed protein-coding genes (padj < 0.1 and $|\log2\text{fold-change}| \geq 1$ in both cases).

To determine the extent to which the sequences of the *Ectocarpus* lncRNAs have been conserved over evolutionary time, we carried out a search for lncRNA loci in a second brown algal genome, that of the kelp *Saccharina japonica* (Ye *et al.*, 2015). The *Ectocarpus* sp. and *S. japonica* lineages are thought to have diverged between 80 and 110 mya (Silberfeld *et al.*, 2010; Kawai *et al.*, 2015). Predicted lncRNA loci were compared between the two species rather than simply searching for sequences related to *Ectocarpus* lncRNAs in the *S. japonica*

genome as the former approach is more likely to detect *bona fide* orthologues (Ulitsky & Bartel, 2013). *S. japonica* transcripts were predicted using Stringtie (Pertea *et al.*, 2015) based on the mapping of 220,551,196 million reads of RNA-seq data (Ye *et al.*, 2015), corresponding to several different tissues, to the assembled genome sequence. Based on these data, FEELnc predicted the presence of 2,840 lncRNA loci in the *S. japonica* genome (Table S7), corresponding to a total of 3,568 different transcripts (1.3 isoforms per locus on average). The mean size of the *S. japonica* lncRNA transcripts was 2,036 nucleotides and varied between 200 (the defined minimal size) and 26,887 nucleotides. As with the *Ectocarpus* lncRNAs, the *S. japonica* lncRNAs were found to be organised in a range of configurations relative to the adjacent gene on the genome (Fig. S2). Comparison of the sets of predicted lncRNAs from *Ectocarpus* and *S. japonica* using Blastn identified 64 pairs of loci that exhibited reciprocal best Blast matches with E-values lower than 10^{-4} (Table S8). These loci are highly likely to be orthologous. Note that Blast comparisons may underestimate the extent of similarity between *Ectocarpus* and *S. japonica* lncRNAs because the program relies on the presence of short regions of high sequence conservation to seed alignments.

Comparison of pairs of orthologous lncRNAs from *Ectocarpus* and *S. japonica* (e.g. Fig. 6) indicated that they tended to contain both conserved and species-specific domains, with the latter usually being located at the ends of the RNA molecules. This suggests that there may not be strong selection pressure on the length of the lncRNA molecules nor on the precise sites of initiation and termination of the mature transcripts.

Impact of the updated large-scale assembly and gene annotation on large-scale genome features including the sex chromosome and an integrated viral genome

Linkage group 30 of the v1 assembly was recently shown to correspond to the sex chromosome in *Ectocarpus* (Ahmed *et al.*, 2014). This linkage group consisted of 20 scaffolds in the v1 assembly but has been considerably extended in the v2 assembly (chromosome 13 in Fig. 1) with the addition of a further 16 scaffolds, increasing the estimated physical length of the chromosome (cumulative scaffold length) from 4,994 to 6,933 kbp. The non-recombining sex-determining region was not affected by this update, as all the additional scaffolds are located in the pseudoautosomal regions of the chromosome. However, as we have recently described a number of unusual features of the pseudoautosomal regions (Luthringer *et al.*, 2015), we verified that these observations were

still valid for the updated version of the chromosome. This analysis confirmed that the updated pseudoautosomal regions continue to exhibit a number of structural features that are intermediate between those of the autosomes and the sex-determining region. In particular, compared with the autosomes, the updated pseudoautosomal regions still exhibit significantly reduced gene density, increased content of transposable element sequences, lower %GC content and the genes had significantly smaller and fewer exons (supplementary Fig. S3). The conclusions of the Luthringer *et al*. (2015) study therefore remain valid for the updated version of the sex chromosome.

The genome of *Ectocarpus* strain Ec32 contains an integrated copy of a large DNA virus, closely related to the *Ectocarpus* phaeovirus EsV-1 (Cock *et al.*, 2010). Microarray analysis had shown that all the viral genes were silent (Cock *et al.*, 2010) and the RNA-seq data analysed here confirmed this observation, indicating complete silencing of this region of the chromosome under all the conditions analysed (Fig. S4).

A genome-wide variant resource for genetic analysis of brown algal gene function

To create an additional genetic resource for gene mapping in *Ectocarpus*, a genome re-sequencing approach was used to identify sequence variants (single nucleotide polymorphisms, SNPs, and indels) across the entire genome. Hi-seq2500 Illumina technology was used to generate 25,976,388,600 bp of paired-end, sequence reads (121x genome coverage) for the female outcrossing line Ec568 (Heesch *et al.*, 2010). A total of 340,665 high quality sequence variants (Table S9) were identified by comparing this data with the reference genome of the male strain Ec32 (Cock *et al.*, 2010) plus the sex-determining region from the Ec32-related female strain Ec597 (Ahmed *et al.*, 2014).

To further validate the sequence variants as potential genetic markers, we used a bulked segregant approach to determine whether they behaved as Mendelian loci. Genomic DNA extracts from a population of 180 segregating progeny derived from a cross between a UV-mutagenised derivative of the reference genome strain Ec32 (strain Ec722) and the female outcrossing line Ec568 were grouped into two bulked segregant pools (84 and 96 individuals) and sequenced on an Illumina platform. Lists of SNP variants were then generated for the two bulked segregant pools and the two lists compared to identify 390,804 shared SNPs that exhibited a 1:1 segregation pattern in the progeny population and were therefore behaving as Mendelian loci. Using this data, 237,839 of the 340,665 sequence variants obtained by mapping the Ec568 DNA-seq data against the reference scaffolds (see above) were validated

as Mendelian genetic markers (Table S9). The average distance between adjacent pairs of the genetic markers identified is 823 bp, providing a high-density resource for genetic analysis in this species.

Extension and improvement of the *Ectocarpus* genome database

The v1 annotation of the *Ectocarpus* genome has been publically available on the Orcae database (Sterck *et al.*, 2012) since its publication in 2010. We have updated the database by adding the v2 annotation described in this study. In addition, a v2 annotation-based Jbrowse genome browser has been created (http://mmodev.sb-roscoff.fr/jbrowse/) to allow simultaneous visualisation of multiple types of data in a genome context. The Jbrowse genome browser allows parallel visualisation of gene models for both coding and non-coding loci, transcript predictions based on RNA-seq data, genetic markers including microsatellites and SNP markers, raw RNA-seq data for both messenger RNAs and small RNAs, Sanger EST data, micro-array data and tiling array data. The Jbrowse genome browser is complementary to the Orcae database, providing an environment for the compilation and analysis of newly generated data before information is definitively incorporated into Orcae, which is the reference database. It is possible for registered users of the Jbrowse genome browser to create private versions in order to upload unpublished and working datasets.

Discussion

The objective of the work reported here was to improve the utility of the *Ectocarpus* genome sequence as a genomic resource.

A high-density, RAD-seq-based genetic map was exploited to significantly improve the large-scale assembly of the genome. This approach allowed 90.5% of the genome sequence to be assembled into 28 pseudo-chromosomes, providing a high quality reference genome for future comparisons with other brown algal genomes focused on synteny and large-scale organisation of chromosomal regions.

In addition, extensive RNA-seq data was used to improve 11,108 existing gene models and to identify 2,030 new protein-coding genes. New data available in the public databases has allowed the functional annotation associated with the protein-coding genes to be considerably improved. Sixty-one percent of genes have now been assigned functional information, compared with 34% in the v1 annotation.

The RNA-seq data was also exploited to evaluate the extent to which protein-coding genes generate alternative transcripts. Wu et al. (2013) reported strong skews in codon usage at both the 5' and 3' ends of *Ectocarpus* exons. Based on a preliminary analysis that indicated a low level of alternative splicing compared with humans, these authors suggested that the skews might reflect strong selection to preserve exon splicing enhancers to avoid mis-splicing of gene transcripts. Our analysis, which was based on a significantly larger transcriptomic dataset, detected a frequency of alternative splicing of about 1.62 transcripts per gene on average. It is difficult to precisely evaluate whether *Ectocarpus* exhibits a particularly low level of alternative splicing compared to other model organisms based on this value because estimates for these other organisms are constantly being revised as more extensive transcriptomic datasets become available. Based on current estimates, however, the frequency of alternative splicing in *Ectocarpus* falls within the range of 1.2 to 3.4 transcripts per intron-containing gene proposed for diverse model organisms with intron-rich genomes including humans, mouse, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* (Kianianmomeni *et al.*, 2014; Chen *et al.*, 2014; Lee & Rio, 2015; Zhang *et al.*, 2015), and therefore does not appear to be exceptionally low.

As far as the types of alternative splicing events are concerned, the *Ectocarpus* genome does not show the same bias towards intron retention events that has been observed with members of the green lineage such as *Arabidopsis* or *Volvox* (Reddy *et al.*, 2013; Kianianmomeni *et al.*, 2014). Instead, use of alternative 3' acceptor sites is very common (41% of events), a bias that has not been observed in other genomes as far as we are aware. Analysis of the domain composition of predicted protein products of alternative transcripts indicated that alternative splicing is likely to contribute significantly to the complexity of the *Ectocarpus* proteome.

The initial v1 annotation focused on protein-coding genes. In this study a genome-wide search was also carried out for non-coding genes, particularly lncRNA loci. Comparison of the *Ectocarpus* lncRNAs with the lncRNA complement of the kelp *S. japonica* indicated that some of the lncRNA loci were already present in the last common ancestor of these two species and have been at least partially conserved, at the sequence level, over the period of about 80 and 110 mya (Silberfeld *et al.*, 2010; Kawai *et al.*, 2015) since the divergence of the two species. Conserved regions were often associated within the same lncRNA with regions that had no equivalent in the opposite species suggesting that brown algal lncRNAs may be organised in a modular fashion and be relatively insensitive to the presence or absence of additional lengths of sequence associated with functional modules. The catalogues of

Ectocarpus and *S. japonica* lncRNA loci are expected to serve as important reference sets for future analyses of lncRNA function in the brown algae.

A genome-wide SNP resource was also developed as part of this study. This collection of SNPs will be a valuable tool for future genetic analyses using *Ectocarpus* as a model system (Cock *et al.*, 2011; Coelho *et al.*, 2012a). All of these new and updated resources have been integrated into the *Ectocarpus* genome database, which has also been improved and extended to facilitate exploitation of the genome data and associated information.

With the integration of the new information and resources described here, the *Ectocarpus* genome represents one of the most extensively annotated genomes within the stramenopile group and, as such, will serve as an important reference genome for future genome analysis projects. Recently, the *Ectocarpus* genome provided a reference for the analysis of the larger and more complex genome of the kelp *Saccharina japonica* (Ye *et al.*, 2015) and similar comparisons are expected in the future as part of the many ongoing brown algal and stramenopile genome projects.

Acknowledgements

We thank Toshiki Uji for providing RNA-seq data, diverse members of the *Ectocarpus* Genome Consortium for manual annotation of genes through the Orcae database and an anonymous reviewer for comments that led to significant improvement of the manuscript. This work was supported by the Centre National de la Recherche Scientifique, the Agence Nationale de la Recherche (project Bi-cycle ANR-10-BLAN-1727, project Idealg ANR-10-BTBR-04-01 and project Sexseaweed ANR-12-JSV7-0008), the University Pierre et Marie Curie and the European Research Council (grant agreement 638240). A.C. was supported by a grant from the Brittany Region.

Author contributions

AC reannotated the *Ectocarpus* genome, identified and characterised alternative transcripts and prepared data for database integration, LS and YVDP created the *Ectocarpus* v2 Orcae database, KA and AC constructed the pseudochromosomes using the genetic map, TD, VW, AC and CH identified the *Ectocarpus* and *S. japonica* lncRNAs, GA and MM created the JBrowse database, OG identified and catalogued the SNP markers, AL and MMP analysed data, JMC, EC and SMC designed and coordinated the research, JMC wrote the manuscript. All authors read and approved the final manuscript.

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y. **2012**. GenomeView: a next-generation genome browser. *Nucleic Acids Res* **40**: e12.

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

Bartsch I, Wiencke C, Bischof K, Buchholz C, Buck B, Eggert A, Feuerpfeil P, Hanelt D, Jacobsen S, Karez R, *et al.* 2008. The genus *Laminaria sensu lato*: recent insights and developments. *Eur J Phycol* **43**: 1–86.

Bratkovic T, Rogelj B. **2014**. The many faces of small nucleolar RNAs. *Biochimica et biophysica acta* **1839**: 438–443.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, *et al.* 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* **17**: 66.

Charrier B, Coelho S, Le Bail A, Tonon T, Michel G, Potin P, Kloareg B, Boyen C, Peters A, Cock J. **2008**. Development and physiology of the brown alga *Ectocarpus siliculosus*: two centuries of research. *New Phytol* **177**: 319–32.

Chekanova JA. **2015**. Long non-coding RNAs and their functions in plants. *Curr Opin Plant Biol* **27**: 207–16.

Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. **2014**. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Molecular Biology and Evolution* **31**: 1402–1413.

Cock JM, Collén J. **2015**. Independent emergence of complex multicellularity in the brown and red algae. In: Ruiz-Trillo I, In: Nedelcu AM, eds. Advances in Marine Genomics. Evolutionary transitions to multicellular life. Springer Verlag, Dordrecht, Netherlands 335–361.

Cock JM, Peters AF, Coelho SM. **2011**. Brown algae. *Curr Biol* **21**: R573–5.

Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury J, Badger J, *et al.* 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617–21.

Coelho SM, Godfroy O, Arun A, Le Corguillé G, Peters AF, Cock JM. **2011**. *OUROBOROS* is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga *Ectocarpus*. *Proc Natl Acad Sci U S A* **108**: 11518–11523.

Coelho SM, Scornet D, Rousvoal S, Peters N, Dartevelle L, Peters AF, Cock JM. **2012a**. *Ectocarpus*: A model organism for the brown algae. *Cold Spring Harbor Protoc* **2012**: 193–198.

Coelho SM, Scornet D, Rousvoal S, Peters NT, Dartevelle L, Peters AF, Cock JM. **2012b**. How to cultivate *Ectocarpus*. *Cold Spring Harb Protoc* **2012**: 258–261.

628 **Conesa A, Götz S, García-Gómez J, Terol J, Talón M, Robles M**. **2005**. Blast2GO: a universal tool for
629 annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–6.

630 **Dayton P**. **1985**. Ecology of Kelp Communities. *Annu Rev Ecol Syst* **16**: 215–245.

631 **Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y**. **2005**. SpliceMachine: predicting splice
632 sites from high-dimensional local context representations. *Bioinformatics* **21**: 1332–8.

633 **Dittami S, Scornet D, Petit J, Ségurens B, Da Silva C, Corre E, Dondrup M, Glatting K, König R, Sterck**
634 **L, *et al.* 2009**. Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae)
635 reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biol*
636 **10**: R66.

637 **Duret L, Gasteiger E, Perrière G**. **1996**. LALNVIEW: a graphical viewer for pairwise sequence
638 alignments. *Computer applications in the biosciences: CABIOS* **12**: 507–510.

639 **Foissac S, Gouzy JP, Rombauts S, Mathé C, Amselem J, Sterck L, Van de Peer Y, Rouzé P, Schiex T**.
640 **2008**. Genome Annotation in Plants and Fungi: EuGene as a model platform. *Current Bioinformatics*
641 **3**: 87–97.

642 **Godfroy O, Peters AF, Coelho SM, Cock JM**. **2015**. Genome-wide comparison of ultraviolet and ethyl
643 methanesulphonate mutagenesis methods for the brown alga *Ectocarpus*. *Mar Genomics* 24: 109-
644 113.

645 **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,**
646 **Raychowdhury R, Zeng Q, *et al.* 2011**. Full-length transcriptome assembly from RNA-Seq data
647 without a reference genome. *Nat Biotechnol* **29**: 644–52.

648 **Gremme G, Brendel V, Sparks ME, Kurtz S**. **2005**. Engineering a software tool for gene structure
649 prediction in higher organisms. *Information and Software Technology* **47**: 965–978.

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

652 **Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B,**
653 **Lieber M, *et al.* 2013**. De novo transcript sequence reconstruction from RNA-seq using the Trinity
654 platform for reference generation and analysis. *Nature Protocols* **8**: 1494–1512.

655 **Heesch S, Cho GY, Peters AF, Le Corguillé G, Falentin C, Boutet G, Coëdel S, Jubin C, Samson G,**
656 **Corre E, *et al.* 2010**. A sequence-tagged genetic map for the brown alga *Ectocarpus siliculosus*
657 provides large-scale assembly of the genome sequence. *New Phytol* **188**: 42–51.

658 **Hoeppner MP, Poole AM**. **2012**. Comparative genomics of eukaryotic small nucleolar RNAs reveals
659 deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC evolutionary biology* **12**:
660 183.

661 **Hughes AD, Kelly MS, Black KD, Stanley MS**. **2012**. Biogas from Macroalgae: is it time to revisit the
662 idea? *Biotechnol Biofuels* **5**: 86.

663 **Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka**
664 **G, *et al.* 2014**. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford,*
665 *England)* **30**: 1236–1240.

Kahles A, Ong CS, Zhong Y, Rätsch G. **2016**. SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**: 1840-1847.

Kawai H, Hanyuda T, Draisma SGA, Wilce RT, Andersen RA. **2015**. Molecular phylogeny of two unusual brown algae, *Phaeostrophion irregulare* and *Platysiphon glacialis*, proposal of the Stschapoviales ord. nov. and Platysiphonaceae fam. nov., and a re-examination of divergence times for brown algal orders. *Journal of Phycology* **51**: 918–928.

Kehr S, Bartschat S, Tafer H, Stadler PF, Hertel J. **2014**. Matching of Soulmates: coevolution of snoRNAs and their targets. *Molecular Biology and Evolution* **31**: 455–467.

Kianianmomeni A, Ong CS, Rätsch G, Hallmann A. **2014**. Genome-wide analysis of alternative splicing in *Volvox carteri*. *BMC genomics* **15**: 1117.

Kijjoa A, Sawangwong P. **2004**. Drugs and Cosmetics from the Sea. *Mar Drugs* **2**: 73–82.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. **2013**. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Klinger T. **2015**. The role of seaweeds in the modern ocean. *Perspect Phycol* **2**: 31–39.

Langmead B, Trapnell C, Pop M, Salzberg SL. **2009**. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Le Bail A, Billoud B, Le Panse S, Chenivesse S, Charrier B. **2011**. ETOILE Regulates Developmental Patterning in the Filamentous Brown Alga *Ectocarpus siliculosus*. *Plant Cell* **23**: 1666–1678.

Lee Y, Rio DC. **2015**. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry* **84**: 291–323.

Lipinska A, Cormier A, Luthringer R, Peters AF, Corre E, Gachon CMM, Cock JM, Coelho SM. **2015**. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *Ectocarpus*. *Molecular Biology and Evolution* **32**: 1581–1597.

Lipinska AP, D'hondt S, Van Damme EJM, De Clerck O. **2013**. Uncovering the genetic basis for early isogamete differentiation: a case study of *Ectocarpus siliculosus*. *BMC genomics* **14**: 909.

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

Meslet-Cladière L, Delage L, Leroux CJ, Goulitquer S, Leblanc C, Creis E, Gall EA, Stiger-Pouvreau V, Czjzek M, Potin P. **2013**. Structure/Function Analysis of a Type III Polyketide Synthase in the Brown Alga *Ectocarpus siliculosus* Reveals a Biochemical Pathway in Phlorotannin Monomer Biosynthesis. *Plant Cell* **25**: 3089–103.

Müller DG. **1966**. Untersuchungen zur Entwicklungsgeschichtebder Braunalge Ectocarpus siliculosus aus Neapel. *Planta* **68**: 57–68.

Müller DG. **1967**. Generationswechsel, Kernphasenwechsel und Sexualität der Braunalge *Ectocarpus siliculosus* im Kulturversuch. *Planta* **75**: 39–54.

Mungall CJ, Emmert DB. **2007**. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**: i337–46.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. **2015**. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–5.

Peters AF, Marie D, Scornet D, Kloareg B, Cock JM. **2004**. Proposal of *Ectocarpus siliculosus* (Ectocarpales, Phaeophyceae) as a model organism for brown algal genetics and genomics. *J Phycol* **40**: 1079–1088.

Peters AF, Scornet D, Ratin M, Charrier B, Monnier A, Merrien Y, Corre E, Coelho SM, Cock JM. **2008**. Life-cycle-generation-specific developmental processes are modified in the *immediate upright* mutant of the brown alga *Ectocarpus siliculosus*. *Development* **135**: 1503–12.

Prigent S, Collet G, Dittami SM, Delage L, Ethis de Corny F, Dameron O, Eveillard D, Thiele S, Cambefort J, Boyen C, *et al.* 2014. The genome-scale metabolic network of *Ectocarpus siliculosus* (EctoGEM): a resource to study brown algal physiology and beyond. *Plant J* **80**: 367–81.

Reddy ASN, Marquez Y, Kalyna M, Barta A. **2013**. Complexity of the alternative splicing landscape in plants. *The Plant Cell* **25**: 3657–3683.

Ritter A, Goulitquer S, Salaün J, Tonon T, Correa J, Potin P. **2008**. Copper stress induces biosynthesis of octadecanoid and eicosanoid oxygenated derivatives in the brown algal kelp *Laminaria digitata*. *New Phytol* **180**: 809–21.

Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, de Reviers B, Rousseau F. **2010**. A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): Investigating the evolutionary nature of the 'brown algal crown radiation'. *Mol Phylogenet Evol* **56**: 659–74.

Smit AJ. **2004**. Medicinal and pharmaceutical uses of seaweed natural products: A review. *J Appl Phycol* **16**: 245–262.

Standage DS, Brendel VP. **2012**. ParsEval: parallel comparison and analysis of gene structure annotations. *BMC Bioinformatics* **13**: 187.

Steneck RS, Graham MH, Bourque BJ, Corbett D, Erlandson JM, Estes JA, Tegner MJ. **2002**. Kelp forest ecosystems: biodiversity, stability, resilience and future. *Environ Conserv* **29**: 436–459.

Sterck L, Billiau K, Abeel T, Rouzé P, Van de Peer Y. **2012**. ORCAE: online resource for community annotation of eukaryotes. *Nat Methods* **9**: 1041.

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

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. **2010**. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.

Tseng C. **2001**. Algal biotechnology industries and research activities in China. *J. Appl. Phycol.* **13**: 375–380.

Ulitsky I, Bartel DP. **2013**. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**: 26–46.

Wahl M, Molis M, Hobday AJ, Dudgeon S, Neumann R, Steinberg P, Campbell AH, Marzinelli E, Connell S. **2015**. The responses of brown macroalgae to environmental change from local to global scales: direct versus ecologically mediated effects. *Perspect Phycol* **2**: 11 – 29.

Wu X, Tronholm A, Caceres EF, Tovar-Corona JM, Chen L, Urrutia AO, Hurst LD. **2013**. Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga *Ectocarpus* are common and resemble those seen in humans. *Genome Biol Evol* **5**: 1731–45.

Yandell M, Ence D. **2012**. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329–42.

Ye N, Zhang X, Miao M, Fan X, Zheng Y, Xu D, Wang J, Zhou L, Wang D, Gao Y, *et al.* 2015. *Saccharina* genomes provide novel insight into kelp biology. *Nat Commun* **6**: 6986.

Zambounis A, Elias M, Sterck L, Maumus F, Gachon CM. **2012**. Highly dynamic exon shuffling in candidate pathogen receptors... What if brown algae were capable of adaptive immunity? *Mol Biol Evol* **29**: 1263–1276.

Zhang R, Calixto CPG, Tzioutziou NA, James AB, Simpson CG, Guo W, Marquez Y, Kalyna M, Patro R, Eyras E, *et al.* 2015. AtRTD - a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in *Arabidopsis thaliana*. *The New phytologist* **208**: 96–101.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Classification of *Ectocarpus* lncRNAs**.**

Fig. S2 Classification of *S. japonica* lncRNAs.

Fig. S3 Comparisons of structural characteristics of the sex-determining and pseudoautosomal regions of the sex chromosome with both a representative autosome and with all autosomes for both the v1 and v2 versions of the *Ectocarpus* genome annotation. **a** percent of sequence that is transposon, **b** number of genes per Mbp, **c** gene size, **d** coding region size, **e** percent GC, **f** percent GC3, **g** cumulative intron size, **h** number of exons.

Fig. S4 Suppressed transcription from a viral genome inserted into chromosome 6.

Table S1 *Ectocarpus* RNA-seq data used in this study. Reads were cleaned using the Fastx toolkit**.**

Table S2 Correspondences between v1 and v2 LocusIDs.

Table S3 List of the rRNA loci in the assembled *Ectocarpus* genome.

Table S4 List of predicted snoRNA loci in the *Ectocarpus* genome.

Table S5 *Ectocarpus* orthologues of core protein components of snoRNPs.

Table S6 List of predicted lncRNA loci in the *Ectocarpus* genome.

Table S7 List of predicted lncRNA loci in the *S. japonica* genome.

Table S8 Comparisons of pairs of orthologous lncRNA loci from *Ectocarpus* and *S. japonica*. Orthologous loci were detected by comparing FEELnc-predicted lncRNA loci from *Ectocarpus* and *S. japonica* using Blastn with a cut off of 10^{-4}.

Table S9 List of 341,426 sequence variants between the genome of the reference male strain Ec32 and the female outcrossing line Ec568.

Tables

Table 1 Comparison of genome-wide statistics for the v1 and v2 annotations of the *Ectocarpus* genome

	v1 annotation	v2 annotation
Genes (including UTRs)		
Number of genes	16,256	17,418
Mean gene length (bp)	6,859	7,542
Longest gene (bp)	122,137	123,931
Shortest gene (bp)	134	150
Exons		
Total number	129,875	134,690
Mean number per gene	7.3	7.96
Max number per gene	171	173
Mean length (bp)	242.2	299.8
Introns		
Total number	113,619	121,264
Mean length (bp)	703.8	739.87
Max length (bp)	25,853	36,147
UTRs		
Genes with only annotated 5' UTR	1,098	918
Genes with only annotated 3' UTR	4,766	3,056
Genes with annotated 5' and 3' UTR	2,484	9,737
Genes without any annotated UTR	7,598	3,715
Mean 5' UTR length (bp)	120.60	139.61
Mean 3' UTR length (bp)	674.74	901.66
Annotation of gene functions		
Genes with predicted functions	5,583	10,688
Genes with associated GO terms	5,989	7,383
miRNA loci	26	64
rRNA loci	n/a	5
snoRNA loci	n/a	656
lncRNA loci	n/a	717

Table 2 Overview of the modifications to the v1 annotation during the production of the v2 annotation of the *Ectocarpus* genome

	Number of genes
N° of v1 models with modified CDS region in the v2 annotation	5,336
N° of v1 models with modified CDS and/or UTR in the v2 annotation	11,108
N° of v1 models fused in the v2	784
N° of v1 models split in the v2	19
N° of v1 gene models removed	123

N° of new gene models in the v2 annotation	2,030

Table 3 Proportions of the different types of alternative splicing events that generate alternative transcripts in *Ectocarpus*

	Mean occurrence per gene	Proportions of alternative splicing events for the genome (%)
Alternative 3' acceptor site	0.481	40.95
Alternative 5' donor site	0.248	21.07
Intron retention	0.139	11.79
Single exon skipping	0.254	21.59
Skipping of multiple exons	0.054	4.58

Figures

Fig. 1 Large-scale assembly of the *Ectocarpus* scaffolds into pseudochromosomes based on a high-density, RAD-seq-based genetic map. Each bar represents one of the 28 chromosomes. Sequence scaffolds (supercontigs) are drawn to scale and identified with numbers (e.g. 207, sctg_207). Left or right pointing arrowheads indicate that the scaffolds have been orientated with respect to the chromosome (i.e. scaffolds with at least two markers separated by at least one recombination event); unorientated scaffolds are indicated with a spot. Chromosome 13 corresponds to the sex chromosome and the non-recombining sex-determining region is indicated with a bar.

Fig. 2 Representative comparisons of v1 and v2 annotation gene predictions illustrating the major types of annotation correction carried out during the transition between the two versions. Protein coding exons are in light or dark green for genome annotation versions v1 and v2, respectively, UTRs are in grey and introns are indicated by thin black lines. **a** analysis of the RNA-seq data allowed the identification of UTRs for gene Ec-27_006370. **b** v2 genes Ec-27_006520 and Ec-05_002440 have been extended and modified compared to their v1 equivalents. **c** v1 genes Esi0002_0099 and Esi0002_0101 were fused to create a single locus, Ec-01_007860. **d** v1 gene Esi0002_0311 was split to create two loci, Ec-01_006420 and Ec-01_006425. Arrows indicate gene features that were not identified or misidentified by the v1 annotation.

Fig. 3 Comparison of the degree of completeness of gene annotations in the v1 and v2 versions of the *Ectocarpus* genome annotation.

Fig. 4 Protein variants predicted to be encoded by alternative transcripts of four genes. **a** alternative products of the ROCO LRR GTPase gene Ec-06_001640 with different LRR repeat structures, **b** alternative products of the nucleotide-binding adaptor shared by the NB-ARC TPR domain containing gene Ec-25_000110 with different TPR domain contents, **c** alternative products of the Notch domain gene Ec-19_004380 with different Notch repeat structures, **d** alternative products of the Ankyrin repeat gene Ec-09_000460 with different Ankyrin repeat structures. Grey lines indicate domains shared between proteins. Roc, Ras of

complex proteins domain; DUF4782, domain of unknown function 4782; VPS9, Vacuolar Protein Sorting-associated 9 domain. The LocusID of each isoform is indicated.

Fig. 5 *Ectocarpus* lncRNA transcript abundance. On average, lncRNA transcripts are about eight-fold less abundant than those of protein coding genes. Boxes indicate interquantile range and median. Whiskers indicate the 10th and 90th percentiles. Data points outside this range are shown as individual points.

Fig. 6 Examples of lncRNA loci conserved between *Ectocarpus* and *Saccharina japonica*. lncRNA loci (in blue) are shown for each species, along with the nearest protein-coding locus on the chromosome (in red). Genes above the line, which represents the chromosome, are transcribed to the right, genes below the line to the left. Percent identities over the aligned regions of *Ectocarpus* and *S. japonica* lncRNA transcripts are indicated. Ec, *Ectocarpus*, Sj, *S. japonica.*