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

Alexandre Cormier, Komlan Avia, Lieven Sterck, Thomas Derrien, Valentin Wucher, Gwendoline Andres, Misharl Monsoor, Olivier Godfroy, Agnieszka Lipinska, Marie-Mathilde Perrineau, et al.

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, Ghent, Belgium, [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, 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, Pretoria, South Africa.

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

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