Analysis of sexual differentiation in the brown alga *Ectocarpus* by RNA-seq

Alexandre CORMIER[1], Susana COELHO[1], Mark COCK[1] and Erwan CORRE[2]

[1] Génétique des Algues, UMR7139 CNRS, Place Georges Teissier, 29680, Roscoff, France acormier@sb-roscoff.fr, coelho@sb-roscoff.fr, cock@sb-roscoff.fr, [2] ABIMS, FR2424 CNRS-UPMC, Place Georges Teissier, 29680, Roscoff, France corre@sb-roscoff.fr

Abstract: We analyze in this study the differential expression of the gametophytic male and female transcriptome in the brown algae *Ectocarpus siliculosus*. Complementary assembly approaches (reference genome vs. de novo) have been used and a set of statistical tools to analyze the differential expression. Results will be discussed. **Key-words:** RNAseq, differential expression, *de novo* and reference transcriptome assembly.

Introduction

Ectocarpus represent an extremely interesting group for the study of sex evolution. Male and female gametophytes present morphological differences and *Ectocarpus* gametes are morphologically identical (isogamy). The female produces a pheromone and the male gamete tracks the pheromone to achieve fertilization (physiological and behavioural anisogamy). We are using an RNAseq approach (Illumina) to explore transcriptome differences between male and female *Ectocarpus* gametophytes to explore the sexual differentiation.

We uses two approaches in parallels : an assembly with the male reference genome and a *de novo* assembly. Both approaches were used to not be limited to the unique use of the male genome and thus be able to find female specific transcripts. Results of both assemblies and differentially expressed genes identified in both approaches will be compared.

Material / data avaible

▪ References data

Male genome [1] — Male annotations

▪ Sequencing

Male and female fertile gametophyte (isogenic lines) → RNA-seq (Illumina) → 2 replicats for the female / 2 replicats for the male

▪ Cleaning

The raw data were cleaned with FASTX toolkit to increase the quality of the reads used in assemblies

Raw data → Trimming (Min quality : 25, Min size : 60pb) → Filtering (Min quality : 20, Min % of base with this value : 75) → Cleaned data

Data type	Raw	Cleaned
Reads number per replicat	26 000 000	23 000 000

Methodolgy

▪ Assembly and abundance estimation

❖ **With reference**

First, cleaned reads of each replicats were aligned against the reference genome with *TopHat* [2,3]. After, transcripts were assembled and a isoforms detection was performed with *Cufflinks* [2,3]. Finally, assemblies were merged with *Cuffmerge* to allow results comparison.

❖ ***de novo***

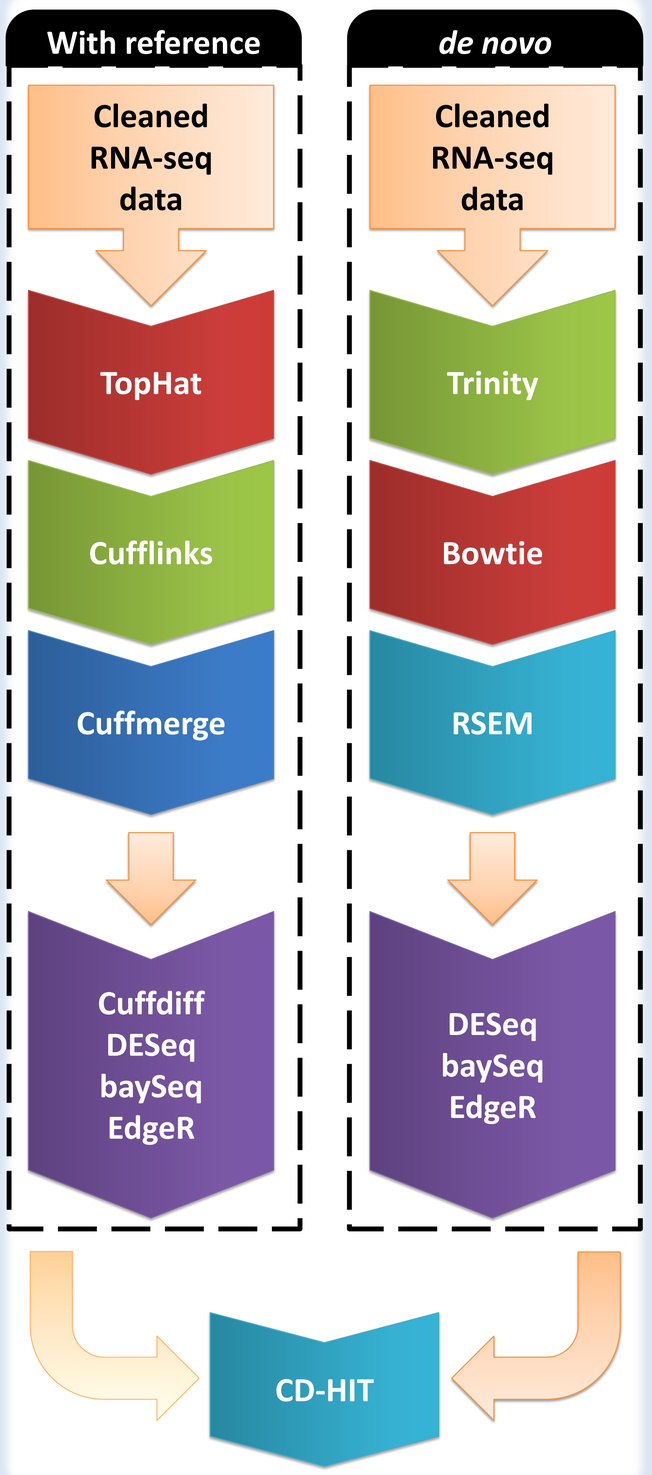
First, a single assembly **of all reads** was performed with *Trinity* [4]. After, reads were aligned for each replicates against transcripts with *Bowtie*, and the abundance estimation was performed with *RSEM*.

▪ Differential expression

The differential expression analysis is performed by using 3 R packages (*DESeq*, *baySeq* and *EdgeR*). The analysis tool provides by *Cufflinks* - *Cuffdiff* - is added to assembly with reference genome. Only the transcripts identified as differentially expressed by all tools are kept for further analysis.

▪ Clustering

Transcripts identified in the two approaches are compared by clustering transcripts sequences with *CD-HIT-EST-2D*



Results

▪ Computational time

All calculations were performed on a cluster with 24 Intel Xeon cores @2,53GHz. *TopHat*, *Cufflinks* and *Cuffmerge* were used 8 cores, 10 cores for *Trinity* and 1 core for *Bowtie* and *RSEM*.

	With reference			*de novo*		
Software	TopHat	Cufflinks	Cuffmerge	Trinity	Bowtie	RSEM
Time per run (number of run)	3h (x4)	3h (x4)	2h30 (x1)	24h (x1)	2h (x4)	2h (x4)

▪ Assemblies

The *de novo* assembly generates a larger number and shorter transcripts, shorter than the assembly with reference genome. *De novo* transcripts have a lower number of exons.

Assembly	With reference	*de novo*
Transcritps (isoforms included)	41 045	82 518
Mean size (pb)	2 480	816
Exon per transcript	9	3

▪ Differential expression

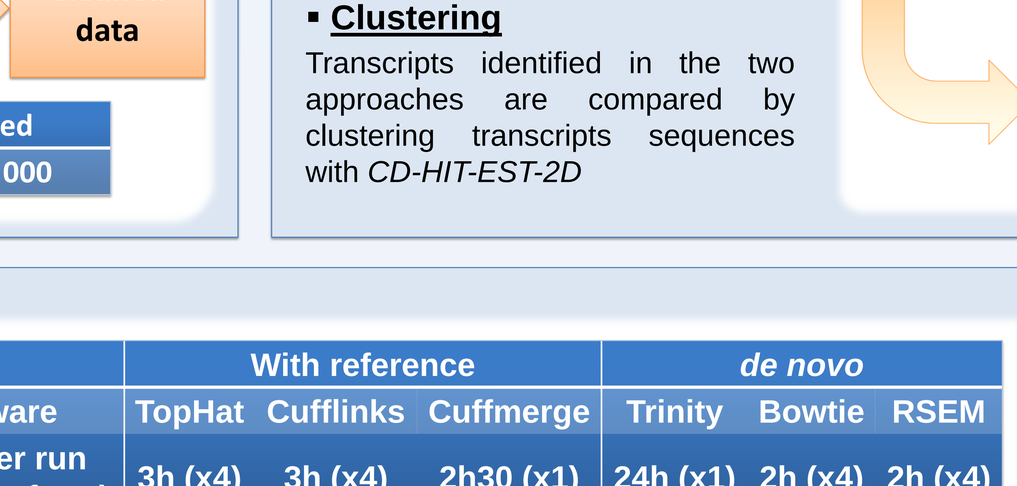
❖ For assembly "with reference", the old version of *Cuffdiff* gives a number of transcripts differentially expressed much higher than R packages. In the new version of *Cuffdiff*, number of transcripts differentially expressed is the same than for R packages.

❖For both approaches, statistical software has identified a similar number of differentially expressed transcripts

▪ Transcripts comparison

❖ After clustering between differentially expressed transcripts in both approaches with *CD-HIT-EST-2D*, only 24% off the 345 transcripts identified in *de novo* assembly are common with the 276 transcripts identified in “with reference” assemblies.

❖ For common transcripts, results for differential expression are identical for both approaches



Venn diagram: with reference results

Venn diagram: *de novo* results

Ongoing

▪ Bioinformatics

Checking new softwares releases : algorithms evolution and better performances

Structural comparison between differentially expressed transcripts

▪ Biology

Functional annotation of differentially expressed transcripts was realized with Blast2GO (data not show)

Validation of sex biased genes by qRT-PCR

Comparing our results with results of differential expression for gametes : isolate specific transcripts of gametophytes

[1] Cock, J. M. *et al.* The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
[2] Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols***7**, 562–578 (2012).
[3] Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech***28**, 511–515 (2010).
[4] Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **advance online publication**, (2011).