Marine Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Identification and expression analysis of long noncoding RNAs in embryogenesis and larval metamorphosis of *Ciona savignyi*

Jiankai Wei[a], Bo Dong[a,b,c,*]

[a] Ministry of Education Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China
[b] Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China
[c] Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao 266003, China

ARTICLE INFO

Keywords:
Ciona
lncRNAs
RNA-Seq
Embryogenesis
Larval metamorphosis

ABSTRACT

Long noncoding RNAs (lncRNAs) play important roles in diverse developmental and pathological processes through chromatin reprogramming, cis regulation and posttranscriptional modification. They have been extensively studied in both vertebrates and invertebrates. However, the information of lncRNAs in urochordate is still lacking. In this study, we used the RNA-Seq data from three developmental stages (18, 21 and 42 hours post fertilization, hpf) of embryos and larvae in *Ciona savignyi* to identify candidate lncRNAs and analyze their expression profiles. A total of 29,944 unigenes were predicted as lncRNAs, five of which had hits with lncRNAs in NONCODE database. The acquired lncRNAs had an average length of 466 nt. The peaks of length, GC content and minimum free energy of the lncRNAs were significantly lower than that of the message RNAs (mRNAs). The average expression levels of lncRNAs were also lower than those of mRNAs. Among the three developmental stages, highly expressed lncRNAs concentrated in 18 hpf embryos. While, for those lncRNAs specifically up-regulated in 21 hpf embryos, their co-expressed mRNAs were enriched in GO terms of membrane, indicating these lncRNAs are involved in the regulation of luminal membrane biogenesis, and extracellular matrix secretion through membrane localized proteins during *Ciona* notochord tubulogenesis. The lncRNAs in 42 hpf larvae were distinct from those in 18 and 21 hpf embryos. This result is associated with the fact that swimming larvae are transiting into metamorphic juveniles at this stage, indicating lncRNAs are involved in the regulation of larval metamorphosis. Overall, our study identified a large number of lncRNAs in *C. savignyi* and revealed their expression characteristics and dynamics during *Ciona* embryogenesis and larval metamorphosis. The results will help to further understand the function of lncRNAs in chordate development and the evolution of lncRNAs.

1. Introduction

Long non-coding RNAs (lncRNAs) are typically defined as RNA molecules that are at least 200 nucleotides (nt) in length and do not display potential to encode proteins (Ulitsky, 2016). They are generally transcribed in a regulated manner by Pol II, and often process similarly to mRNAs: they are generally capped, spliced, and polyadenylated (Ulitsky and Bartel, 2013). There is broad consensus that lncRNAs are transcribed and processed similarly to mRNAs, and play crucial functions independent of translation (Rutenberg-Schoenberg et al., 2016). In particular, lncRNAs have been implicated in developmental and pathological processes through reprogramming of chromatin, cis regulation at enhancers, and post-transcriptional regulation of mRNA processing (Iyer et al., 2015).

Without a doubt, the advent of sensitive, high-throughput genomic technologies such as microarrays and next-generation sequencing (NGS) results in an unprecedented ability to detect novel transcripts, the vast majority of which seem not to be derived from annotated protein-coding genes (Kung et al., 2013). Recent studies have identified thousands of lncRNAs in human (Cabili et al., 2011; Iyer et al., 2015), mouse (Guttman et al., 2009), fruit fly (Young et al., 2012), nematode (Nam and Bartel, 2012) and zebrafish (Ulitsky et al., 2011). Although several highly conserved lncRNAs are known (Chodroff et al., 2010), lncRNAs generally have modest sequence conservation (Marques and Ponting, 2009). Only a very small fraction of lncRNAs are well characterized. The rapid evolution of lncRNAs can provide insights into their function, but the absence of lncRNA annotations in non-model organisms has precluded comparative analysis.

Ascidians are the largest class within the subphylum Tunicata (Urochordata) in the chordate phylum. Ascidians of the genus *Ciona* are widely used model organisms for chordate developmental genomics because of their similar embryonic body plan to that of vertebrates

* Corresponding author at: Ocean University of China, No. 5 Yushan Road, Qingdao 266003, China.
E-mail address: bodong@ouc.edu.cn (B. Dong).

https://doi.org/10.1016/j.margen.2018.05.001
Received 19 December 2017; Received in revised form 28 April 2018; Accepted 4 May 2018

(Stolfi and Christiaen, 2012). The genome of *Ciona intestinalis* (now is *C. robusta*) (Brunetti et al., 2015) and *Ciona savignyi* both have been sequenced (Dehal et al., 2002; Small et al., 2007). The experimental malleability and unique phylogenetic position of the sea squirt *Ciona* provide a model system to study the lncRNAs in embryogenesis and larval metamorphosis. Moreover, tunicates lack the genomic duplications, facilitating functional genomic studies by avoiding the complications of functional overlap and redundancy (Vassalli et al., 2015). A series of studies have focused on the prediction of non-coding RNAs especially microRNAs in tunicates. Computational methods were used to predict 14 microRNA gene families in *C. intestinalis* (Norden-Krichmar et al., 2007). This study reported the first collection of validated microRNAs in *C. intestinalis*. Then conserved non-coding sequences from the *C. intestinalis* genome were extracted and computationally folded to identify putative hairpin-like structures and 458 miRNA candidates were obtained (Keshavan et al., 2010). Another study reported a network analysis of mRNA and microRNA expression during appendage regeneration in *Ciona* (Spina et al., 2017). A comprehensive homology-based annotation of non-coding RNAs was also conducted in the recently sequenced genome of *Didemnum vexillum* (Velandia-Huerto et al., 2016). Only one conserved lncRNA (RMST 9) was identified due to their low levels of sequence conservation. Genome-wide survey of *Halocynthia roretzi* was also presented and 319 candidate miRNAs were identified (Wang et al., 2017). The identification of lncRNAs has already been reported in other marine species such as sea cucumber (Mu et al., 2016), oyster (Yu et al., 2016) and rainbow trout (Al-Tobasei et al., 2016; Wang et al., 2016). However, the high-throughput identification and expression analysis of lncRNAs in tunicates are still lacking.

Until recently, the most common sequencing methods used oligo (dT)-based enrichment for poly (A) RNAs, which include the vast majority of functionally characterized lncRNAs (Ulitsky, 2016). In this study, we used the RNA-Seq data from *C. savignyi* embryonic and larval libraries at different stages to identify lncRNAs and acquire their expression profiles. Our results reveal the characteristics and dynamics of lncRNAs during embryogenesis and larval metamorphosis of *C. savignyi* and will help to further understand the function of lncRNAs in chordate development and the evolution of lncRNAs.

2. Results

2.1. High throughput RNA-Seq and de novo assembly of transcriptomic data in C. savignyi

To acquire lncRNAs that play the roles in embryonic and larval development of *C. savignyi*, 18 hours post-fertilization (hpf), 21 hpf, and 42 hpf staged samples (a total of nine samples, three for each stage) were collected for RNA-Seq. At 18 hpf (St. 24) (Hotta et al., 2007), notochord cells elongated into a single line; at 21 hpf (St. 25) (Hotta et al., 2007), apical extracellular lumen formed and expanded between adjacent notochord cells; at 42 hpf, the tail of the swimming larva disappeared in the process of metamorphosis. Seven libraries including two 18 hpf samples, two 21 hpf samples, and three 42 hpf samples were successfully constructed and sequenced. After removing adaptors and trimming low quality reads, 288 million clean reads with 43.3 billion nucleotides were obtained (Table 1), which were then assembled into 147,212 transcripts and clustered into 110,279 unigenes by Trinity. The unigenes, totaling to 66 million nucleotides, with an average length of 599 nt, and N50 length of 990 nt, were then utilized to map to the genome of *C. savignyi* for subsequent lncRNA identification.

2.2. Identification of lncRNAs in C. savignyi

For a generic pipeline to identify lncRNAs from RNA-Seq data, RNA-Seq reads were either first mapped to the genome, and then assembled into transcripts, or first assembled into transcripts, and then mapped to the genome. Then multiple filtering steps were applied to remove various artifacts and protein-coding genes. For *C. savignyi*, the genome annotation was not as comprehensive as model animals. Therefore, we first assembled the transcripts and mapped the unigenes to the genome of *C. savignyi*, and then filtered the artifacts and protein-coding genes. An optimized computational pipeline (Fig. 1) was developed to identify lncRNAs and mRNAs in *C. savignyi*. The pipeline can also be used to identify lncRNAs in other non-model organisms. The reads were first assembled and clustered into unigenes, and then unigenes were filtered by length (length $>$ 200 nt) and read coverage (reads ≥ 3), and 57,580 unigenes were left. Then, they were aligned to the genome of *C. savignyi* (CSAV 2.0) by BLAT, 51,081 unigenes were aligned to *C. savignyi* genome. After that blast search was conducted with protein databases with a cutoff *E*-value of 1e-5. Among them, 16,680 unigenes found putative homologs in three main protein databases (Nr, Swissprot and Pfam). These unigenes were identified as mRNAs, while the other 34,401 unigenes were treated as raw candidate lncRNAs. 31,722 unigenes were acquired after trimming unigenes with maximum ORF length $>$ 100 aa by Getorf, 30,668 unigenes were left after trimming unigenes with cpc value $>$ -1 by cpc. Finally, after removing tRNAs, rRNAs and other ncRNAs by blast search and removing unigenes, which have hits in UTRdb database (Grillo et al., 2010) and UTR region of *C. savignyi*, 29,944 unigenes (Table 2) were acquired as candidate lncRNAs in *C. savignyi*.

Compared to other marine animals, the number of identified lncRNAs is larger than that in sea cucumber, oyster and rainbow trout. To verify our result, two additional alignment-free lncRNA annotation programs were performed. Coding-Non-Coding Index (CNCI) (Sun et al., 2013) result showed that only 55 of the 29,944 lncRNAs had coding potential (score $>$ 0). Coding Potential Assessment Tool (CPAT) (Wang et al., 2013) result showed that only 2 of 29,944 lncRNAs had coding probability (score $>$ 0.38).

We also analyzed the genomic position of these candidate lncRNAs. The results showed that among the 29,944 unigenes, 14,571 (48.66%) of them located in sequence space that do not overlap with mRNA regions, which are referred to “long intergenic noncoding RNAs (lincRNAs)”. Most of the lincRNA (85.09%) were $<$ 10 kb from the protein-coding genes (Fig. 2). For the other 15,373 (51.34%) of the candidate lncRNAs, which have overlap with mRNA regions, 7853 of them are antisense transcripts to mRNAs. 7520 of them are sense to mRNAs and encoded within the intron regions of annotated genes.

To explore the sequence conservation features of the lncRNAs

Table 1
RNA-seq samples used for lncRNA assembly and analysis.

Sample	Raw reads	Clean reads	Clean bases	Q20 (%)	Q30 (%)	GC content (%)
18 hpf sample 1	48,025,418	39,470,900	5.92G	96.82	91.30	42.32
18 hpf sample 2	46,761,988	41,675,704	6.25G	95.57	89.11	43.34
21 hpf sample 1	50,795,502	42,544,888	6.38G	96.90	91.34	45.84
21 hpf sample 2	44,062,316	40,128,044	6.02G	95.74	89.31	46.14
42 hpf sample 1	55,471,004	45,215,904	6.78G	96.72	91.01	45.15
42 hpf sample 2	47,077,378	39,647,522	5.95G	96.89	91.34	44.78
42 hpf sample 3	43,598,910	39,771,512	5.97G	95.84	89.59	44.77