Marine Genomics 29 (2016) 31–34

Contents lists available at ScienceDirect

Marine Genomics

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

Genomics/technical resources

De novo transcriptome sequencing of a non-model polychaete species

E. Cannarsa *, G. Zampicinini, O. Friard, A. Santovito, P. Cervella

Department of Life Sciences and Systems Biology, Università di Torino, Turin, Italy

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form 26 May 2016
Accepted 7 June 2016
Available online 20 June 2016

Keywords:
Pyrosequencing
De novo assembly
Ophryotrocha diadema
Transcriptome

ABSTRACT

Transcriptome sequencing is a useful method for studying gene-sequences data, especially in non-model organisms whose genomic sequences are yet to be determined. Indeed, even without any genome reference, transcripts can be assembled *de novo* to produce a genome-scale transcription map. Here we describe a *de novo* transcriptome assembly for the polychaete worm *Ophryotrocha diadema*, generated from 454-sequencing (Roche GS-FLX Titanium Sequencing). We describe the sequencing, the assembly and functional annotation of EST sequences, and the level of transcriptome coverage provided by our sequence data. The sequences we assembled represent a wide depiction of expressed genes from throughout the lifespan of *O. diadema*. We found gene sequences to be part of a broad range of functions (*i.e.* biological and metabolic processes, immune system, reproductive processes, development and growth, signalling and response to stimuli) and we therefore built the first transcriptomic reference for *Ophryotrocha* polychaete worms. These results encourage us to focus our investigation on many aspects of *O. diadema* and to extend our research to co-generic species.

1. Introduction

Over the past fifteen years, the next-generation sequencing technology has arisen as an innovative approach for high-throughput sequence determination and has radically improved the efficiency and speed of gene discovery (Wang et al., 2010). This has also significantly speeded up and increased the sensitivity of gene-expression profiling, improving the study of spatio-temporal data on gene expression in cells and tissues (Macagno et al., 2010). As a result, next generation sequencing approach has now been increasingly implemented in many different research fields such as biomedicine (Neuss et al., 2011), ecology (Neave et al., 2012) and behavioural ecology (Peterson et al., 2012), a state which is expected also to boost collaborative and comparative genomics studies (Blow, 2009).

In this scenario, transcriptome sequencing is more useful than the genome sequencing method for studying gene-sequences data in eukaryotes, because most of their genomes consist of non-coding DNA. EST sequences lack introns and intragenic regions that render analysis and interpretation of data more difficult (Bouck and Vision, 2007). According to this, transcriptome analysis can be particularly appropriate for species that have a large and repetitive genome, whose sequencing and assembling could be costly and challenging (Cahais et al., 2012). ESTs thus have high functional information content, and usually correspond to genes with known or predicted functions (Bouck and Vision, 2007; Andersen and Lubberstedt, 2003). Hence, transcriptome sequencing enables functional genomic studies based on global gene expression, single nucleotide polymorphism surveys (SNP), quantitative trait loci studies (QTL), genomic scans of diversity and so on (Bouck and Vision, 2007; Nagaraj et al., 2007).

Despite this evident potential, evolutionary and population genomic research has long been restricted to a small number of taxonomic groups. Thus, we still have a limited number of published complete genomic sequences that are available mainly for model organisms (Cahais et al., 2012).

Here, we utilize 454-pyrosequencing of cDNA to characterize the transcriptome of *Ophryotrocha diadema* (Annelida: Polychaeta). The aim of our study was to gather broad-spectrum information on the genes expressed in this species for developing a genomic resource to support further research in the *Ophryotrocha* genus.

Polychaetes of the genus *Ophryotrocha* are small worms living among the interstitial fauna in nutrient-rich and polluted habitats, such as harbours and lagoons. Owing to easy rearing in the laboratory, short generation times and relative simple experimental manipulations, several of the about 40 species currently described are model organisms for studying different biological aspects, including reproductive biology, sexual selection and the evolution of reproductive strategies (Cannarsa et al., 2015). Indeed, they display a variety of reproductive strategies, ranging from gonochorism – including labile gender expression – to sequential and simultaneous hermaphroditism. The ancestral reproductive mode in still not identified either by ecological or phylogenetic studies (Thornhill et al., 2009). There are crucial questions waiting for answer concerning the genetic system underlying sex determination and the evolutionary processes leading to the adoption of a particular reproductive system in a given species.

* Corresponding author at: Department of Life Sciences and Systems Biology, Università di Torino, Via Accademia Albertina 13, 10123 Turin, Italy.
E-mail address: elio.cannarsa@live.it (E. Cannarsa).

http://dx.doi.org/10.1016/j.margen.2016.06.001
1874-7787/

Table 1
Transcriptome and environmental feature.

Item name	Definition
Transcriptome analysis	
investigation type	Eukaria
Species	*Ophryotrocha diadema* (Annelida: Polychaeta)
Project name	*O. diadema* transcriptome
Geographic location	34°3.1338′ N 118°14.6208′ W
Geographic location (country and/or sea, region)	Los Angeles, CA, USA
Collection date	November the 15th, 2013
Environment (biome)	Marine — subtidal area
Environment (feature)	Harbour, lagoons, nutrient-rich seawater
Environment (material)	Laboratory strains maintained in seawater
Temp	21 °C
Salinity	34 g/l
Transciptome assembly data	
Sequencing method	Pyrosequencing
Sequencing technology	Roche 454 GS XLR70 Titanium
Assembly method	*De novo*
Assembly name	*O. diadema* transcriptome
Assembly	MIRA 4
Finishing_strategy	Draft
Annot_source	Blast2GO
Transcriptome coverage	7.8×.

Pyrosequencing produced approximately 400 Mbp of sequence data in the form of 663,996 reads ranging from 50 to 1779 bp length, with a mean length of 600 bp and mean Phred score of 31.

Large numbers of EST sequences are available from some species of polychaetes. These include *Capitella teleta*, *Hediste diversicolor* and *Platynereis dumerilii*, mainly used as bio-indicators of disturbed marine habitats (Berthet et al., 2003; Neave et al., 2012; García-Alonso et al., 2014). Nevertheless, few genomic data exist in public databases (*e.g.*, NCBI), for *Ophryotrocha* species (*e.g.* 597 sequences from *Ophryotrocha puerilis* are deposited in the EST section of the GenBank; Cervella et al., unpublished).

Here we report the results of the creation of a transcriptome resource for the polychaete worm *Ophryotrocha diadema* by a *de novo* GS-FLX Titanium Sequencing (Roche 454-Sequencing). Although the species was first described >30 years ago (Åkesson, 1976) and was intensively studied since the beginning of the 70′s, no genetic and biochemical data have been published yet. This paucity exists mainly because species of the genus *Ophryotrocha* are usually utilized in behavioural ecology studies focusing on plasticity in sex allocation and/or ecology (Cannarsa et al., 2015; Meconcelli et al., 2015; Lorenzi et al., 2015).

EST collections for *O. diadema* will contribute to the development of molecular markers for co-generic species, facilitating comparative genomics and the study of adaptive variation across the genus, in various topics from ecology to ethology. To this aim, cDNA collections from a large pool of individuals (from 4 segments larvae to 25 segments adults) were used to sample the largest number of expressed genes.

We describe the sequencing, the assembly and functional annotation of EST sequences, and the level of transcriptome coverage provided by our sequence data.

2. Data description

All experiments were carried out using laboratory populations. Specimens were reared in glass bowls, at 20 °C, with a 12 h light/dark photoperiod and fed *ad libitum* with chopped spinach. Approximately, 500 worms were reared in 200 ml of artificial sea water at a salinity of 35‰. From this population, 200 individuals were used for mRNA extraction. Those individuals were placed into bowls with clean water and starved for two days to have their gut empty of remnant food. Analysing different stages of development (see Åkesson, 1976) increases the coverage of the transcriptome. 65 adult hermaphrodite individuals (20 segments), 65 juvenile individuals (≤12 segments) and 70 larvae (≤4 segments) were pooled into 1.5 ml micro-centrifuge tube. After an addition of 500 μl of TRIZOL (Invitrogen) to every tube and lysates were homogenized. Total RNA extraction was accomplished according to the manufacturer's recommendations. After DNase (Sigma-Aldrich) treatment, the sample was purified with the RNeasy mini kit (Qiagen) in accordance with the manufacturer's instructions. Total RNA was quantified using the spectrophotometer Nanodrop ND-100. RNA integrity (RIN = 9) was determined using the Bioanalyzer 2100 (Agilent Technologies). cDNA library construction from the Poly-A RNA and the consequent sequencing were performed on one plate of Roche 454 GS XLR70 Titanium platform (Table 1).

Assuming that a similar number of cDNAs occurs in *O. diadema* as in *Caenorhabditis elegans* (about 30,000 with an average transcript length of c. 1600 bp — see the transcript list on http://www.ncbi.nlm.nih.gov/genome/41), our transcriptome coverage was estimated at 7.8×.

Before the assembly, 454 primers were trimmed from all the reads. Sequences with average quality lower than 20 were not taken into consideration as they may carry a high percentage of non-recognized bases (*i.e.* N instead of A/C/G/T). We performed different assemblies utilizing different values of "minimum relative score alignment". This parameter describes the minimum percentage of matching between two reads to be considered for assembly. Predictably, the number and the length of assembled contigs increased with decreasing match percentage parameter stringency (data not shown). We finally set a minimum relative score alignment of 70, between the suggested values of a minimum of 55 and a maximum of 80 (Chevreux, 2014). The assembling of the sequenced reads was performed by using MIRA assembler, version 4.0rc2 (Chevreux et al., 1999, Chevreux, 2007). We filtered out reads shorter than 200 bp and with a quality value lower than 20. Moreover, we considered alignments that had "quality score alignment values" higher than 70. Hence, 609,049 reads were assembled into 44,554 contigs, which have a mean length of 1069 bp. Contig length ranged from 200 bp to 11,915 bp. As expected, contig length increased with the number of reads assembled into them. Moreover, 62,535 reads resulted as singletons with a mean length of 592 bp. Their length ranged from 200 bp to 1779 bp.

106,570 assembled sequences with a mean length of 892 bp were deposited in the BioProject section of NCBI database. The assembling was considered for the consequent BLAST and annotation stage.

Since there is neither a genome reference for *O. diadema*, nor a genomic or a transcriptomic one for closely related species to compare with, we first used a BLASTX algorithm with an *E*-value threshold of 10^{-3} and

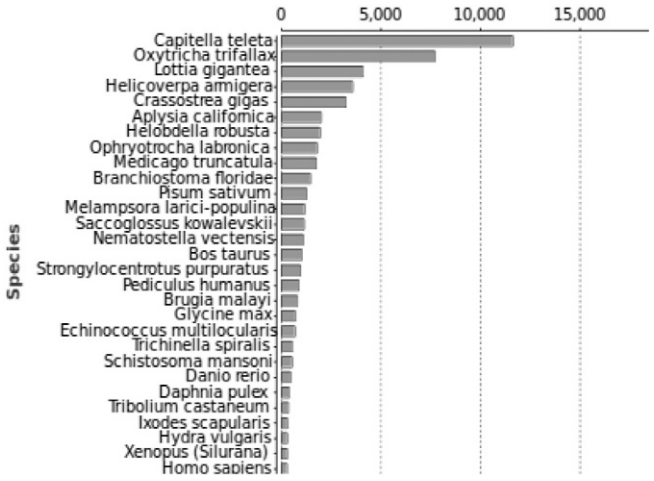

Fig. 1. Figure shows the top-hit species distribution. Assembled sequences had similarities with known proteins found in studied species.