

Sequencing the *Acropora millepora* transcriptome

Sylvain Forêt¹

sylvain.foret@anu.edu.au

David Hayward²

david.hayward@anu.edu.au

Eldon Ball²

eldon.ball@anu.edu.au

David Miller³

david.miller@jcu.edu.au

- ¹ Centre for Bioinformation Science, Mathematical Sciences Institute, The Australian National University, Canberra, Australia
- ² Centre for Molecular Genetics of Development, Research School of Biological Sciences, The Australian National University, Canberra, Australia
- ³ Comparative Genomics Centre, James Cook University, Townsville, Australia

Abstract

This study presents the preliminary results of an attempt to sequence the transcriptome of the coral *Acropora millepora* using 454 pyrosequencing.

Keywords: transcriptome, coral, *Acropora millepora*, 454

1 Introduction

Acropora millepora is arguably the best studied coral species of the Australian Great Barrier Reef. Besides their central ecological importance as part of the reef ecosystem, corals are of particular interest for the study of evolution and development and have played a great role in revealing the genetic complexity of the eumetazoan ancestor [1].

The genomic resources for *A. millepora*, however, are rather limited, with only around 15,000 EST sequences deposited in GeneBank. Taking advantage of the new advances in high-throughput sequencing, we are attempting to extend these resources as close as possible to the complete transcriptome. Here, we report the initial results of this effort, focusing on bioinformatics challenges and insights into metazoan evolution.

2 Method and Results

Biological material A mixture of *A. millepora* developmental stages was used to prepare the sequencing material. Sequencing was conducted at the Australian Genome Research Facility using 454 pyrosequencing. We also used a dataset produced by a concurrent project focussing on the transcriptome of *A. millepora* five days old larvae [2], also sequenced using the 454 technology. A total of 1,056,025 reads were sequenced, with a median size of 242 bp.

Assembly The existing *A. millepora* ESTs were used in addition to the 454 reads to produce an assembly with the *mira* assembler [3] in hybrid assembly mode, producing a total of 88,413 contigs. Table 1 shows the size distribution of these contigs and of the predicted transcripts of *A. millepora* closest relative with a fully sequenced genome, the sea anemone *Nematostella vectensis* [4]. This table indicates that the current sequencing effort, even in its preliminary stage, provides a considerable improvement over the existing EST contigs. The number of contigs of any size larger than 1 kb is approaching the number of predicted *N. vectensis* transcripts, but the gap increases for larger sizes. This could be caused by large transcripts being fragmented into separate contigs or to RNA decay that would affect RNA-based contigs, but not genome-based predicted transcripts.

Table 1: Size distribution of *A. millepora* contigs and *N. vectensis* predicted transcripts.

	> 500 bp	> 1 kb	> 1.5 kb	> 2 kb	> 3 kb
<i>A. millepora</i> ESTs	7,265	1,422	185	41	0
<i>A. millepora</i> ESTs + 454	27,255	8,616	3,417	1,347	192
<i>N. vectensis</i> predicted transcripts	19,100	10,573	5,940	3,361	1,224

Alternative splicing One of the main challenges in assembling a transcriptome, is that most assembly softwares perform best on linear sequences such as genomes. In a transcriptome, however, splice variants are better conceptualised as a graph of alternatively spliced exons. The assembly was thus post-processed by clustering highly similar sequences based on perfect word matches. Consensus sequences were then generated for each cluster, using the `poa` software [5] in isoform generation mode, which resulted in a total of 61,912 consensus sequences with 8,957 clusters of alternatively spliced sequences.

Homology and orthology We found that 28,277 *A. millepora* contigs have a significant hit to the NCBI RefSeq database (BLAST e-value $< 10^{-10}$). We are currently trying to group the other contigs into 1) genomic contamination, 2) UTRs or non-protein-coding genes, 3) non-conserved gene fragments or orphan genes. We also found that 19,997 of the 27,237 *N. vectensis* predicted transcripts have a hit in the *A. millepora* contigs (e-value $< 10^{-10}$). Using reciprocal best hits, 9,180 orthologous pairs were found between *N. vectensis* and *A. millepora*.

Finally, we found 467 coral sequences with a strong hit (e-value $< 10^{-20}$) in RefSeq but without any hit in the sea anemone. This implies that a number of gene losses have taken place in the *Nematostella* lineage, and that using it as the model cnidarian to reconstruct the gene set of the eumetazoan ancestor will result in an oversimplification that can be partially alleviated if the coral is taken into account.

3 Discussion

Based on the results presented here, we speculate that we are currently covering between 60% and 80% of the coral transcriptome. More sequencing is currently underway. Ultimately, this study, will not only provide knowledge on the early evolution of animal genomes, but will also provide a platform for gene expression studies in developmental and ecological contexts.

References

- [1] Ball, E.E., Hayward, D.C., Saint, R., Miller, D.J., A simple plan – cnidarians and the origins of developmental mechanisms, *Nature Reviews Genetics* 5(8):567–77, 2005.
- [2] http://www.bio.utexas.edu/research/matz_lab/matzlab/454.html
- [3] Chevreur, B., Wetter, T. and Suhai, S., Assembly Using Trace Signals and Additional Sequence Information, *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* 45–56, 1999.
- [4] Putnam N.H., Srivastava, M., Hellsten, U., *et al.*, Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization, *Science* 86(317):86–94, 2007.
- [5] Lee, C., Generating consensus sequences from partial orger multiple alignment graphs. *Bioinformatics* 19(8):999–1008, 2003.