Sequencing of Coral Genomes Obtained From the Environment

Dexter Jiunn Herng Lee, Christopher Lok Yung Voo#
Biotechnology Research Institute, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, MALAYSIA
Corresponding author. E-Mail: cvooly@ums.edu.my; Tel: +6088-320000 ext 5602; Fax: +6088-435324

ABSTRACT Cultured *Acropora gemmifera* and *A. tenuis* obtained from the environment were sequenced and the sequenced obtained were binned. The corals were obtained from a coral farm located in the waters near the shores of Semporna, Sabah. The absence of symbionts commonly associated with corals, as well as the presence of reads with no hits, suggest that a more comprehensive and exhaustive database should be used instead of the current database. The presence of contaminants and the variability of the symbionts may vary depending on the location and environment from which the corals were obtained.

KEYWORDS: DNA sequencing; environmental samples; corals; binning; contaminants

Full Article - *Environmental biotechnology*
Received 30 August 2017 **Online** 28 November 2017
© Transactions on Science and Technology 2017

INTRODUCTION

Corals are marine invertebrates from the Cnidaria phylum. Typically found between the tropics of Cancer and Capricorn as well as at varying depths, coral reefs serve as habitats for a wide variety of marine lifeforms. However, corals are under threat from pollution, climate change, ocean acidification, pollution, destructive fishing practices and overfishing as outlined in the United Nation's Resolution 2/12 in 2016 (UNEP/EA.2/Res.12). The nature of the coral's habitat makes it less accessible as compared to terrestrial organisms and as such, it is a rarely tapped resource of which the corals and organisms associated with them may produce novel biocompounds with potential commercial applications.

The first coral genome (*Acropora digitifera*) was sequenced by the Marine Genomics Unit (MGU) of the Okinawa Institute of Science and Technology (OIST) in 2011 (Shinzato *et al.*, 2011). Both the assembled genome (along with the raw sequencing data) and assembled transcriptome were made available on the National Center for Biotechnology Information (NCBI) and MGU websites. A hybrid sequencing approach was used to sequence the *A. digitifera* genome, utilising reads generated from the Illumina Genome Analyzer (paired-end and mate-pair libraries) and 454 GS-FLX (paired-end and single end libraries) sequencers. *De novo* assembly of the sequencing data resulted in a genome size of approximately 419 Mbp. The DNA for sequencing was obtained from the sperm of the *A. digitifera*. The coral was a part of larger colony, and was maintained in an aquarium after collection. Other coral genomes (*A. millepora* and *A. muricata*) were sequenced by other institutions, but the assembled genome and raw sequencing data are not publicly available (as of December 2016).

Culturing of the coral in a relatively controlled environment and harvesting its sperm for DNA sequencing would shift greater abundance of coral DNA with relatively lesser amounts of non-coral DNA. However, by obtaining the adult colony structure of corals straight from the environment and extracting its DNA, one would expect the presence of non-coral DNA (from symbionts and other organisms) and this would, depending on the variability and DNA abundance of non-coral organisms present on or within the coral, would, increase assembly complexity and sequencing costs (to achieve the required sequencing depth) respectively. This would also consume greater time and

resources (computing and storage) needed prior to genome assembly. As such, binning and classifying of sequencing reads are required so as to enable one to gauge the amount of reads needed for subsequent sequencing runs to reach the sufficient coverage (determined based on the initial relative DNA abundance of coral and non-coral). The binning process would also potentially reduce the time and resources needed for the *de novo* assembly process (since identified non-coral sequences can be removed) as compared to assembling the reads without prior binning. This paper will attempt to elucidate the distribution of DNA sequences across multiple bins representing the extraction and sequencing the DNA of corals (from two *Acropora* species) obtained straight from the environment.

METHODOLOGY

Sample acquisition

The corals were obtained from a coral farm located off the coast of Semporna, Sabah. The corals were morphologically identified as *A. gemmifera and A. tenuis*. The live, cultured corals were shipped to Universiti Malaysia Sabah via overnight surface shipping, individually packed with seawater, activated carbon beads and air, all placed in a polystyrene box with ice. It was noted that some of the corals were dead on arrival. Nevertheless, the corals were cut into smaller fragments (small enough to fit into a 50-mL falcon tube) and were snap frozen using liquid nitrogen. The snap frozen coral fragments were placed into labelled 50-mL falcon tubes and were stored in -80 ˚C prior to DNA extraction for a minimum of 1 week.

DNA extraction

The DNA was extracted using a combination conventional and kit based methods. A guanidium thiocyanate based lysis buffer (4 M guanidium thiocyanate, 2 % beta-mercaptoethanol, 0.1 M Tris-HCl, 0.05 M EDTA, pH 8.0) was used, followed by extraction using phenol chloroform. Precipitation of nucleic acids was done using 2 M NaCl (final concentration) and 2 volumes of 99.8 % ethanol. The resulting pellet was washed twice with 70 % ethanol solution, after which the partially dried pellet was dissolved in the Digestion Solution of the MagJET Genomic DNA Kit (Thermo Scientific, Cat. K2721). The dissolved pellet was processed as per the kit's instructions, with the proteinase K digestion step omitted.

DNA quality and quantity assessment

The extracted DNA was assessed using the Nanodrop 2000 spectrophotometer (Thermo Scientific), Qubit 2.0 fluorometer using the High Sensitivity dsDNA assay (Invitrogen) and 1 % agarose gel electrophoresis prior to sequencing using the Illumina HiSeq 3000 sequencer.

DNA sequencing and sequencing reads quality control

The extracted DNA from the two coral species were outsourced for library preparation (270 bp and 800 bp libraries, PCR-free) and sequencing (PE150). The raw data sequencing data was subjected to trimming using Trimmomatic (Bolger *et al.*, 2014), with reads having a minimum base quality of Q30 and 100 bp length being kept. The quality of the reads (pre and post quality control) was assessed using FastQC (Andrews, 2010).

Database construction for binning

Kraken (Wood and Salzberg, 2014) was used to bin and classify the sequencing reads. A custom Kraken database, comprising of genomic sequences from the NCBI reference sequence (RefSeq) database (O'Leary *et al.*, 2016) was constructed. The sequences were divided into 17 parts, with each

part having a size of slightly below 40 GB. Database construction was done on all 17 parts, with the kmer database reduced to a final size of 40 GB for each part.

Binning of sequencing reads

The sequencing reads were binning using Kraken running on a node with 32 threads and 128 GB of RAM. The results from all 17 parts were merged using custom PERL scripts, after which the kraken report for the merged results were generated and visualized using Krona (Ondov *et al.*, 2011).

RESULT AND DISCUSSION

Table 1 and Figure 1 show the number of reads and bases before (raw) and after quality control.

Table 1. Number and percentage of reads before and after quality control (QC). Percentage figures were rounded up to two decimal points. SE: single-end, PE: paired-end, bp: base pairs.
NOTE: Value obtained by adding both reads of the pair.

Species	*A. gemmifera*	Percentage (%)	*A. tenuis*	Percentage (%)
Raw reads count	497,932,398	100.00	499,498,698	100.00
Raw data (Gbp)	74.69	100.00	74.92	100.00
Reads post-QC (PE)[1]	469,820,084	94.35	469,111,010	93.92
Reads post-QC (SE)	3,843,015	0.77	5,019,965	1.01
Reads discarded	24,269,299	4.87	25,367,723	5.08
Post-QC bases PE (Gbp)[1]	68.81	92.13	68.58	91.53
Post-QC bases SE (Mbp)	553.87	0.74	722.11	0.96
Discarded bases (Gbp)	5.33	7.13	5.63	7.51

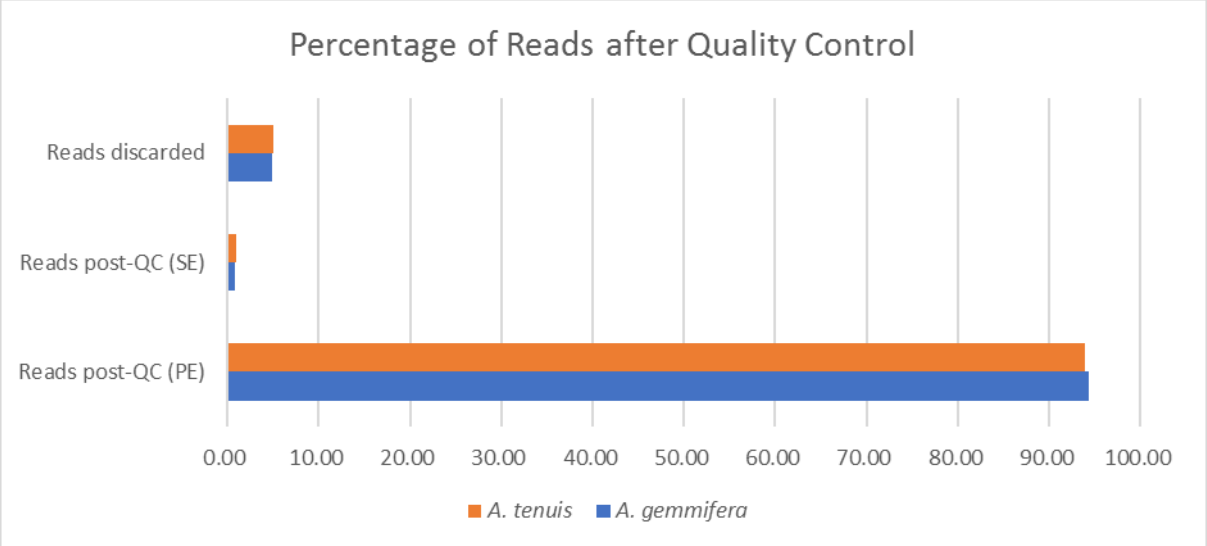

Figure 1. Percentage of reads after quality control

The binning of the sequencing reads (after quality control) was done only on the paired-end data, given that the single-end reads passing quality control was under 2 %. Given the large

differences in the number of reads between paired-end and singled-end reads, the binning results of the paired-end reads will reflect that of the single-end reads if binning was to be done on the single-end reads. Figure 2 and 3 show the merged binning results for reads from *A. gemmifera* and *A. tenuis*, while Table 2 shows the organisms whose reads are more or equal than 0.1 % of the total number of reads.

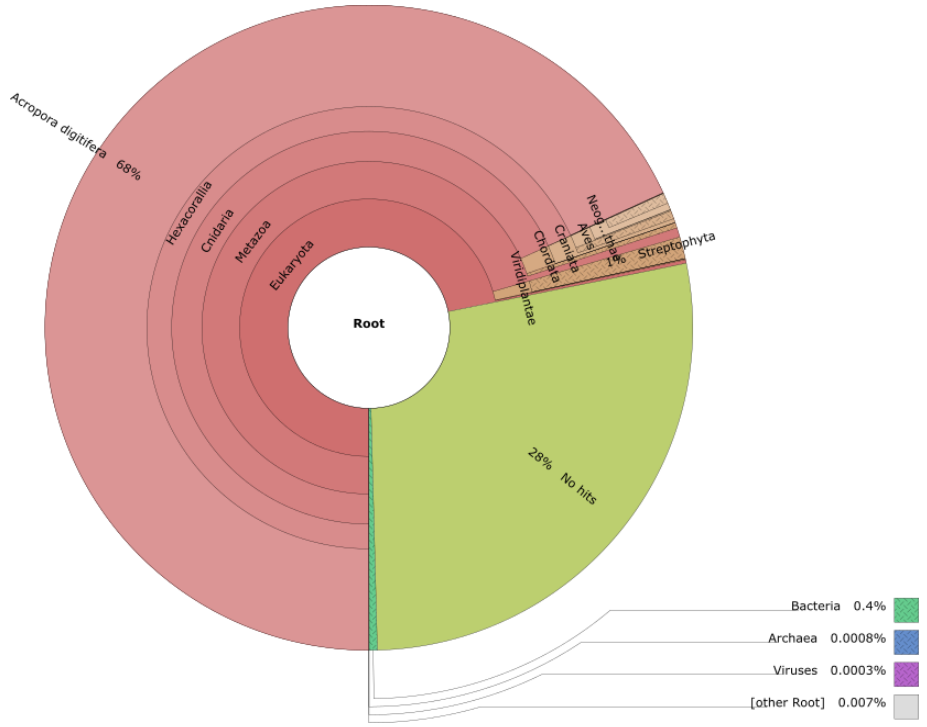

Figure 2. Visualisation of the binning results of *A. gemmifera* reads.

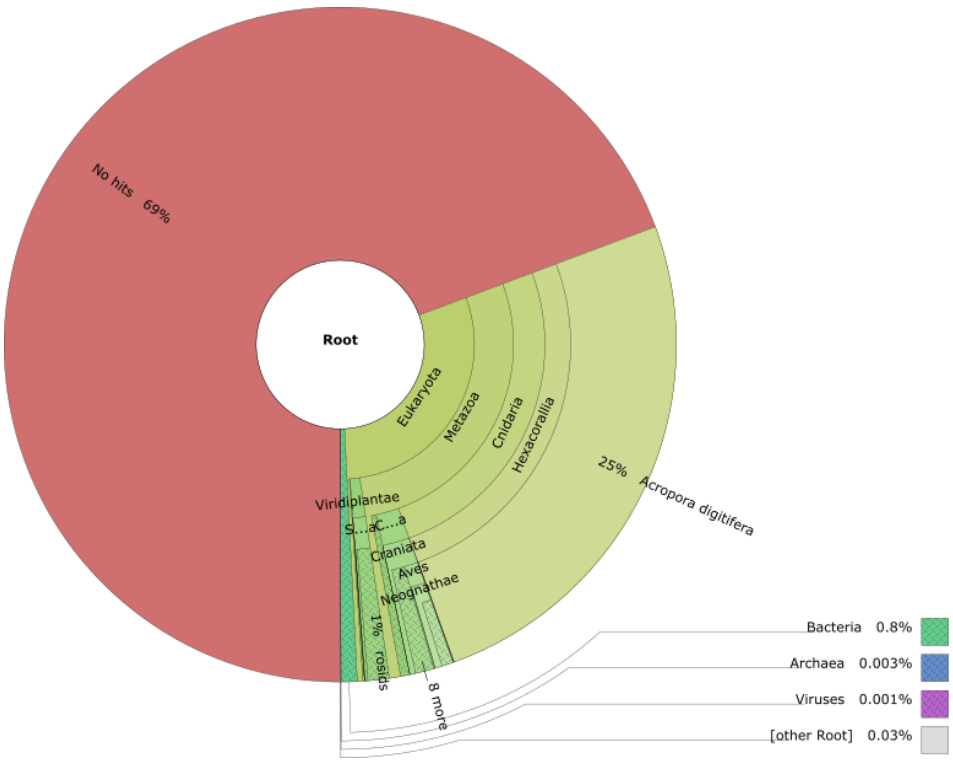

Figure 3. Visualisation of binning results of *A. tenuis* reads.

Table 2. Organisms with reads more than 0.1 % of total reads.

Dataset	*A. gemmifera*	*A. tenuis*
A. digitifera (species)	68 %	25 %
Neognathae (subclass)	0.9 %	0.9 %
Brassica napus (species)	0.5 %	0.8 %
Brachypodium distachyon (species)	0.2 %	< 0.1 %
Sorghum bicolor (species)	0.2 %	< 0.1 %
Serratia (genus)	0.2 %	< 0.1 %
Medicago truncatula (species)	< 0.1 %	0.1 %
No hits	28 %	69 %

A large portion of the classified reads in both datasets were identified as reads with close similarity to *A. digitifera*. This is expected as both coral species are of the same genus as *A. digitifera*. However, there is a stark difference in the number of reads identified as *A. digitifera* between both datasets (68 % against 25 %). There are two factors (or a contribution from both) that may lead to this, one, *A. gemmifera* is more closely related to *A. digitifera* than *A. tenuis* is to *A. digitifera*, and two, *A. tenuis* has far lesser coral DNA distribution (relative to other DNA) in the coral structure that was processed for DNA extraction. Construction of phylogenetic trees using common phylogenetic markers (18s rRNA, cytochrome oxidase I, and cytochrome b) was inconclusive as the bootstrap support for the nodes were too low (data and figures not shown).

The presence of reads from terrestrial plants and birds suggest that the coral farm is located somewhere near the coast, possibly with moderate bird traffic and relatively low human related activities. The presence of reads from members of the *Serratia* genus (majority of which were from *S. ureilytica*) may indicate a possible candidate for symbiotic relationship with the said coral species (Determination as a strict symbiotant would require more datasets from different locations). Although unicellular algae are usually associated with corals, none of the binned reads were from the *Symbiodinium* genus. This might be due to the sequences not being present in the NCBI RefSeq genome database.

Given the nature of the library preparation (PCR-free) and the amount of data generated, it is suspected that some of the reads under the "no hits" bin may very well belong to the corals, as well as possibly other organisms. The binning method used in this case is only as good as the database from which the reads are compared against. As such, reads with no hits may belong to organisms which are yet to be sequenced, or are not present in the NCBI RefSeq genome database. It is hypothesised that number of reads with no hits might be reduced if full kmer databases were used instead of reduced kmer databases.

CONCLUSION

Sequencing reads from *A. gemmifera* and *A. tenuis* obtained from the environment were binned. Based on the organisms present, it is possible that the corals were cultured near the coastal area with relatively low human related activities. The presence of reads with no hits, as well as the absence of reads from the *Symbiodinium* genus illustrated one of the shortcomings of reference-based binning with available databases. However, the binning process did shift a significant portion and

abundance of coral-like sequences aiding subsequent assembly of the coral genome while reducing misassemblies due to leak-over of genomes from non-coral species.

ACKNOWLEDGEMENTS

The results obtained in this study were part and parcel of the research on corals funded by the Sabah State Government through Sabah Biodiversity Centre and Protein Technologies Limited (United Kingdom) under the Universiti Malaysia Sabah (UMS) grant code GL0110.

REFERENCES

[1] Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[2] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–20. https://doi.org/10.1093/bioinformatics/btu170

[3] Environment Assembly resolution 2/12, *Sustainable coral reefs management*, UNEP/EA.2/Res.12 (4 August 2016), available from https://documents-dds-ny.un.org/doc/UNDOC/GEN/K16/072/34/pdf/K1607234.pdf

[4] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell C. M., Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey K. M., Murphy M. R., O'Neill K, Pujar S, Rangwala S. H., Rausch D., Riddick L. D., Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy T. D. & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

[5] Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

[6] Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., Fujie, M., Fujiwara, M., Fujiyama, A., Miller, David J. & Satoh, N. (2011). Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature*, **476**(7360), 320–3.

[7] Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**(3), R46.