I COBBIND
I CONGRESSO BRASILEIRO DE BIOTECNOLOGIA INDUSTRIAL
XXIV SINAFERM • XV SHEB • XV ENZITEC FLORIANÓPOLIS • 2024
Creating connections between biotechnology and industrial sustainability
August 25 to 28, 2024
Costão do Santinho Resort, Florianópolis, SC, Brazil
ENVIRONMENTAL BIOTECHNOLOGY

ASSEMBLY AND ANNOTATION OF THE SKIN TRANSCRIPTOME OF THE HUMPBACK WHALE, *Megaptera Novaeangliae* (Borowski, 1781)

Beatriz. M. Lahoz[1*], Bárbara P.H. Righetti[1], Vanessa S. Deconto[1], Karim H. Lüchmann[2], Afonso C.D. Bainy[1] & Clei E. Piazza[1]

[1] Biological Sciences/Center of Biological Sciences/Department of Biochemistry/Federal University of Santa Catarina/Laboratory of Biomarkers of Aquatic Contamination and Immunochemistry, Florianópolis, Brazil.
[2] Department of Scientific and Technological Education, Santa Catarina State University, Florianópolis, Brazil.
** Corresponding author's email address: bialahoz@gmail.com*

ABSTRACT

Biomonitoring contaminated environments stands as a critical endeavor among the current backdrop of human activities. Aquatic ecosystems, in particular, constitute the sinkhole of anthropogenic impact. This study presents a comprehensive methodology for analyzing biomarkers in the humpback whale's skin transcriptome, emphasizing its potential as an environmental sentinel. Humpback whales skin biopsies were collected in the Santos Basin and RNA was extracted. The constructed cDNA library was sequenced using the Illumina HiSeq 2500 platform with paired-end sequencing. The *de novo* assembly was performed using Trinity software, and functional annotation was conducted with BLAST+ against the SwissProt database. The sequencing generated 93,356,722 PE reads that resulted in 124,580 transcript contigs containing genes associated with biotransformation of xenobiotics, such as *CYPs*, *GSTs*, *SULTs*, *ABC-Transporters*, antioxidant enzymes, such as *GPXs* and *SODs* and nuclear receptors, such as *AhR* and *ESR*. This research underscores the importance of environmental biotechnology and ecotoxicogenomics in monitoring the impacts of anthropogenic activities on marine ecosystems. The large amount of data generated from this assembly opens doors to the discovery of new biomarkers for monitoring aquatic contamination and creates a valuable resource for research in various other fields.

Keywords: Biotechnology. Bioinformatics. Biomarkers. Marine mammals. Gene expression.

1 INTRODUCTION

Biomonitoring contaminated environments stands as a critical endeavor among the current backdrop of human activities. Aquatic ecosystems, in particular, constitute the sinkhole of anthropogenic impact, owing to their utilization in various sectors such as food production, raw material sourcing, waste disposal, and petroleum extraction.[1] Consequently, it is imperative to assess water quality using molecular biomarkers to detect chemical fluctuations and ecological stressors at early stages. Techniques like quantitative polymerase chain reaction (qPCR) are pivotal for this purpose, as they are adept at detecting gene transcripts even at minute concentrations.[2]

The transcriptome represents the dynamic gene expression repertoire within cells, tissues, or organisms at a given point in time. It comprises all RNA molecules, including messenger RNA (mRNA), essential for protein synthesis, and serves as a conduit for responding to external components such as xenobiotics.[3] Transcriptome assembly encompasses three primary strategies: reference-based assembly, *de novo* assembly, and a hybrid approach combining both methodologies. *De novo* assembly, for instance, identifies overlaps in short sequences (k-mers) from reads, culminating in the reconstruction of original transcripts. In contrast, reference-based assembly constructs the transcriptome based on the genome of the species.[4]

Ecotoxicogenomics, meanwhile, provides insights into gene expression patterns indicative of water contamination's impact on aquatic life. This interdisciplinary field not only aids in the development of novel molecular biomarkers but also in the identification of potential target genes.[5]

Marine mammals, particularly, emerge as sentinel species in terms of water contamination, owing to their extended lifespans, top predator status, and high adipose tissue content predisposing them to bioaccumulation and biomagnification of contaminants.[6] The humpback whale, *Megaptera novaeangliae*, belongs to the suborder Mysticeti, cetaceans with keratinaceous baleen plates, a filter-feeding organ that retains food while water is expelled from the side of the mouth. In this manner, the species represents an ideal monitoring candidate, since Its exposure to contaminants is not only through the trophic chain, but also through direct filtration of water, underscoring its significance in environmental surveillance efforts.[7] Hence, the present study aims to assembly and annotate the humpback whale's skin transcriptome. By doing so, it seeks to identify molecular biomarkers and target genes implicated in xenobiotic metabolism, thereby raising the capacity for effective environmental monitoring and conservation.

2 MATERIAL & METHODS

The samples were obtained from skin biopsies as part of the Cetaceans Monitoring Project, in the Santos Basin (PMC), one of the monitoring programs required by Brazil's federal environmental agency, the Institute of the Environment and Renewable Natural Resources (IBAMA), for the environmental licensing process of oil production and transport by Petrobras. Firstly, RNA was extracted from the samples following the RNeasy Fibrous Tissue Mini kit protocol (Qiagen) with a DNase I (Qiagen) treatment. Quality assessment was carried out using the NanoDrop 2000 (Thermo Fisher), visualized through gel electrophoresis and the TapeStation system with RNA Screen Tape (Agilent Technologies), and concentrations were quantified using the Qubit 2.0 fluorometer (Invitrogen). The sample with the best quality was chosen for subsequent analyses. The cDNA library was prepared following the TruSeq Stranded mRNA Library Prep kit protocol (Illumina, Inc). The cDNA quality was validated through the Bioanalyzer 2100 instument. cDNA library quantification was performed through qPCR, following the Sequencing Library qPCR Quantification kit instructions (Illumina Inc.). The sample was diluted to a concentration of 10 pM and clustered on cBot. Sequencing was carried out on the Illumina HiSeq 2500 system, with paired-ended reads PE 2 x 100 bp (100 million reads or 50 million pairs per sample). All library preparation and deep sequencing procedures were performed at the Life Sciences Core Facility (LaCTAD) from State University of Campinas (UNICAMP, Campinas, SP, Brazil). Secondly, the reads were assessed for quality using the program FastQC[8] (Version 0.11.9). Afterwards, the Trimmomatic[9] (Version 0.39) software was used to eliminate reads with low-quality scores (phred < 33) and FastQC was run again for validation.

The *de novo* assembly was performed using Trinity software (Version 2.15.1) with default settings.[10] Reads were normalized for positions with coverage above 50 bp for better performance. The abundance of transcripts was estimated using the kallisto method, available in the Trinity package and they were normalized in Transcrips Per Million (TPM). The quality of the assembly was assessed using the N50 and Ex90N50 methodologies, from Trinity. Open reading frames (ORFs) were obtained through the program TransDecoder (Version 5.0.1-3). Functional annotation was achieved using BLAST+[11] (Version 1.6.3), aligning the transcripts against the curated database SwissProt[12], with an expected value of 1e-05.

3 RESULTS & DISCUSSION

A total of 93,356,722 PE reads of 100 bp with 50% GC content were obtained using Illumina technology. None of the reads were marked with poor quality. In general, the sequences obtained excellent phred scores, all above 30, despite of the extremities, which is expected. After trimming, 42,752,625 paired-end reads were considered of high quality. In general, the sequences obtained outstanding phred scores, all around 34, with zero adaptor content.

Thereafter, the clean reads were assembled using Trinity, resulting in a total of 124,580 transcript contigs. These contigs have an average length of 1,062.88 bp and an N50 length of 2,266 bp, meaning that at least half of the total assembled bases were in contigs of 2,266 bp or longer. Additionally, the Ex90N50 statistic (N50 for the top 90% most expressed transcripts) indicated that 47,932 transcripts, which together accounted for 90% of the total expression, had lengths of at least 2,657 bp. The coding regions of the transcripts were identified through TransDecoder and only the complete ORFs were considered. The function annotation was performed combining the sequences with the curated protein database SwissProt through BLAST+.

For the annotated genes and in search for genes related to xenobiotic exposure, 8 sequences were identified as phase I biotransformation, including 1 gene of *Dimethylaniline monooxygenase* (*FMO4*) and 7 isoforms of *Cytochrome P450* (*CYP*). 16 sequences were identified as phase II biotransformation, including 13 isoforms of *Glutathione S-transferase* (*GST*) *and 3 isoforms of Sulfotransferase* (*SULT*). 19 sequences were identified as phase III biotransformation, including genes of various sub-families and members of *ABC-type oligopeptide transporter* (*ABC*). 8 sequences were identified as antioxidant enzymes, including 5 genes of *Glutathione peroxidase* (*GPX*) and 3 genes of *Extracellular superoxide dismutase* (*SOD*). Finally, 12 sequences were identified as nuclear receptors, indicating potential endocrine disruption, including 2 genes of *Retinoic acid receptor* (*RXR*), 1 gene of *Growth hormone receptor* (*GHR*)*, 2 genes of Oxysterols receptor* (*NR1*), 3 genes of *Peroxisome proliferator-activated receptor* (*PPAR*), 1 gene of *Aryl hydrocarbon receptor repressor* (*AhRR*)*, 1 gene of Aryl hydrocarbon receptor* (*AhR*) and 2 genes of *Estrogen receptor* (*ESR*).

This study allowed us to identify several biomarkers commonly associated with metabolism of xenobiotics. For example, the presence of enzymes involved in phase I, II and III biotransformation suggest a well-development detoxification system in the humpback whale, capable of handling and removing a wide array of environmental contaminants. The presence of antioxidant enzymes indicates a robust antioxidant defense system in the cetacean, essential for protecting against oxidative damage. The identification of certain nuclear receptors suggests potential pathways for endocrine disruption, possibly due to the increasing prevalence of endocrine-disrupting chemicals in marine environments. Moreover, non-destructive monitoring, using skin biopsies offers a method to study these animals without sacrificing them or relying on strand events, where carcasses are often in advance state of decomposition.[13] Furthermore, the large amount of data generated from this assembly opens doors to the discovery of new biomarkers. Its applications extend beyond aquatic contamination monitoring, reaching into fields such as genetics, immunology, evolutionary biology, cancer, neuroscience and microbiology, creating a valuable resource for future research using biopsies.

4 CONCLUSION

The identification and analysis of genes involved in xenobiotic metabolism in the humpback whale, *Megaptera novaeangliae*, underscore the species' potential as a sentinel for aquatic environmental monitoring. This study highlights the comprehensive transcriptome assembly and annotation, revealing genes associated with phase I, II, and III biotransformation processes, antioxidant enzymes, and nuclear receptors indicative of potential endocrine disruption. This research exemplifies the importance of environmental biotechnology and ecotoxicogenomics in understanding the impacts of anthropogenic activities on marine ecosystems. By elucidating the molecular mechanisms underlying xenobiotic metabolism in sentinel species like the humpback whale, we enhance our ability to assess and mitigate the effects of environmental pollutants. This work not only contributes to the development of novel biomarkers for environmental monitoring but also supports efforts in the conservation and sustainable management of marine life.

REFERENCES

[1] BORGWARDT, F. *et al*. 2019. Sci. Total Environ. 652. 1396–1408.
[2] PANTI, C. *et al*. 2011. Ecotoxicol. 20 (8). 1791–1800.
[3] SCHIRMER, K. *et al*. 2010. Anal. Bioanal. Chem. 397 (3). 917–923.
[4] MARTIN, J. A., WANG, Z. 2011. Nat. Rev. Genet. 12 (10). 671–682
[5] PIÑA, B., BARATA, C. 2011. Aqu. Toxicol. 105 (3–4). 40–49.
[6] BORRELL, A. 1993. Mar. Pollut. Bull. 26 (3). 146–151.
[7] JEFFERSON, T. A. *et al*. 2011. Cetaceans Species Accounts. *In*: Marine Mammals of the World: A Comprehensive Guide to Their Identification. 1. ed. Academic Press. California. 72
[8] ANDREWS, S. 2010. FastQC. Disponível online em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
[9] BOLGER, A.M., LOHSE, M., USADEL, B. 2014. Bioinformatics. 30. 2114-2120.
[10] HAAS, B.J *et al*. 2013. Nat. Protoc. 8. 1494-1512.
[11] CAMACHO, C. *et al*. 2009. BMC bioinformatics. 10. (421).
[12] BAIROCH, A., APWEILER, R. 2008. Nucleic Acids Res. 28. 45-48.
[13] WILEY, D. N., ASMUTIS, R. A., PITCHFORD, T. D. 1995. Fish. Bull. 93. 196–205.

ACKNOWLEDGEMENTS

We are grateful to the Cetacean Monitoring Project (PMC) for providing the samples. We thank the staff of the Life Sciences Core Facility (LaCTAD) at the State University of Campinas (UNICAMP) for their work in library preparation and deep sequencing procedures. We also appreciate the support from LABCAI/UFSC and the funding provided by Petrobras for this project.