Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Transcriptomic sequence dataset of a potential new model species for studying biomineralization *Onchidoris muricata* (Nudibranchia, Gastropoda, Mollusca)

Ekaterina D. Nikitenko[a,*], Ilya E. Borisenko[b], Andrey N. Anisenko[c], Elena V. Vortsepneva[a]

[a] *Department of Invertebrate Zoology, Lomonosov Moscow State University, Leninskye Gory 1/12, Moscow 119234, Russian Federation*
[b] *Embryology Department, Saint-Petersburg State University, University Embankment, 7/9, St. Petersburg 199034, Russian Federation*
[c] *Chemistry Department, Lomonosov Moscow State University, Leninskye Gory 1/10, Moscow 119234, Russian Federation*

ARTICLE INFO

Article history:
Received 19 March 2024
Revised 6 May 2024
Accepted 7 May 2024
Available online 11 May 2024

Dataset link: Transcriptome of the nudibranch mollusk Onchidoris muricata (Doridina, Gastropoda) at different stages of early ontogeny with special refer to spiculogenesis (Original data)
Dataset link: Omu-transcriptome (Original data)

Keywords:
RNA-seq
Mollusca
Gene expression
Transcriptome annotation
Biomineralization
Illumina sequencing

ABSTRACT

Onchidoris muricata is a widespread shell-less species of nudibranch molluscs, which has unique for Gastropoda skeletal elements – subepidermal calcite spicules. The general and fine morphology of the spicules, as well as their maturation process in ontogenesis, have been studied in detail by authors. The uniqueness of spicules lies in their intracellular formation and location under the ectodermal epithelium, which is more typical for deuterostomes. We present *O. muricata* as a potentially new model species for studying calcification of intracellular protein structure. A total of 96 individuals were collected in the Kandalaksha Bay of the White Sea, both manually and by scuba diving. All individuals were divided into three groups based on morphological characteristics such as specimens' size, spicule condition etc. This division suggests the existence of three stages in postembryonic ontogenesis of *O. muricata* reflecting the maturation of the spicule complex. Total RNA samples were isolated from three size groups of molluscs in three biological replicates.

* Corresponding author.
E-mail address: nikitenkocatia@yandex.ru (E.D. Nikitenko).

https://doi.org/10.1016/j.dib.2024.110526

Libraries were prepared from the polyadenylated RNA fraction and sequenced at NovaSeq6000 (Illumina), yielding a total of 112.8 Gb of 150 bp paired-end reads, corresponding to almost 1,000-fold coverage of the transcriptome. Representative transcriptome assembled *de novo* with Trinity. In addition to obtaining the transcriptome sequences of *O. muricata*, differential expression analysis was also performed for these three size groups. This allows us to trace the dynamics of molecular and biological processes during the life of a mollusc. The obtained data can then be used as a reference transcriptome for closely related species, to study specific expressed genes, to identify various unique sequences, including protein-coding ones, to understand biological processes, including biomineralization and much more.

Specifications Table

Subject	Biological sciences
Specific subject area	Zoology, transcriptomics
Type of data	Raw data: Sequence reads obtained after Illumina sequencing of RNA samples, FASTQ files Analyzed data: differential expression analysis, Table, Image, Graph, Figure.
Data collection	The collection was carried out by *Onchidoris muricata* (O.F. Müller, 1776) at a depth of 10–15 m in August 2022. In total, 96 individuals were collected. All individuals were divided into three groups based on morphological characteristics: group I from 100 µm to 1 mm, group II from 1 to 4 mm, and group III from 4 to 12 mm. The sample contained 10–11 individuals of the first group, 9–10 individuals of the second group and 6–7 individuals of the third group. 3–5 samples were collected for each size group. Total RNA was extracted from the first group of samples. From the group II and III, total RNA was extracted from the integument with spicules. This RNA was used for RNA sequencing.
Data source location	The material was collected in the vicinity of White Sea Biological Station of Lomonosov Moscow State University (WSBS LMSU) in the Kandalaksha Bay 66°33’17” N, 33°06’02” E.
Data accessibility	NCBI SRA Database, BioProject: PRJNA1033054 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1033054] GitHub [https://github.com/mezoderma/Omu-transcriptome]

1. Value of the Data

- Our data uncovers a range of downstream analyses, including assembly, annotation, differential expression of genes of the nudibranch mollusc Onchidoris muricata at different stages of ontogenesis.
- The transcriptome assembly and in silico translated proteome we obtained can be used in comparative-evolutionary studies as a resource for finding orthologs and comparing them to other taxa. It can also be used for phylogenetic analyses based on protein-coding genes.
- The quantitative data on transcript abundance that we obtained for molluscs of different ages allow us to assess changes in gene expression at these stages. Since the stages are associated with the development of the spicule network, these data may help in the study of biomineralization and the search for common features of this process in other invertebrates.
- The data analysis presented in the study was performed using exclusively open access tools, ensuring reproducibility and applicability to other Eucaryota transcriptomes.
- All data from this project is publicly available via the NCBI databases.

2. Background

Nudibranchia is a group of marine heterobranch gastropods widespread over the oceans and characterized by soft bodies without shells [1]. Nevertheless, species belonging to order Doridina possess skeletal structures called spicules located beneath the epithelium [2,3]. The subepidermal location and intracellular development of spicules are unique for Mollusca [4]. This work presents transcriptomes obtained by Illumina sequencing of RNA samples derived from the nudibranch molluscs *Onchidoris muricata* at various stages of ontogeny. *Onchidoris muricata* is a potentially new model object among invertebrates for studying the process of biomineralization is example of spicules.

The annotation of the obtained transcriptomic sequences of molluscs with spicules at different stages of maturation, as well as their comparative analysis, allows not only an expansion of our understanding of the biomineralization process, but also an assessment of the dynamics of mineralization in relation to other biological processes in ontogeny. This data serves as a fundamental basis for further study of the molecular mechanisms of calcification, cellular signaling, and search for protein sequences involved in the mineralization process.

3. Data Description

Total RNA samples were isolated from three size groups of molluscs in three biological replicates. Libraries were prepared from the polyadenylated RNA fraction and sequenced to yield a total of 112.8 Gb reads, corresponding to nearly 1000-fold coverage of the transcriptome (see below). After decontamination and collapsing sequences with greater than 95 % similarity, the assembly contained 213,764 genes and 278,617 transcripts in Trinity terms [5]. A peptide longer than 100 amino acid residues was encoded by 52,754 transcripts in 38,236 genes. Of all protein-coding transcripts, 57 % (30,084) of the peptides had a BLASTP hit against the Uniprot/Swiss-Prot database with a threshold of 1e-5, and 56 % (29,495) of the peptides had at least one domain from the Pfam database. About 37 % of peptides (19,736) had neither a BLAST hit nor a domain from Pfam. It is likely that these proteins are lineage specific.

An assembly with a nearly complete BUSCO set was obtained (Fig. 1A). As a result of decontamination, we can observe a significant reduction in duplicated genes, which are often an indication of the presence of RNA from multiple organisms. Decontaminated assembly deposed in Github repository together with counts data https://github.com/mezoderma/Omu-transcriptome.

When filtering out transcripts with expression levels lower than 1 TPM, transcripts of 156,549 genes remain. The total length of the transcriptome when counting only one isoform (with the highest expression level) per gene was 99.85 Mb.

By mapping reads to the assembly, we obtained transcript expression values. Supplementary Table S1 contains these cross-sample normalized values (TMM, in transcript per million units). Principal component analysis after regularized-logarithm transformation showed that sample 1 from group II clustered with samples from group I. Also, sample 2 from group I clusters with group II. The remaining samples of the three groups cluster together (Fig. 1B). We removed outliers from the subsequent analysis (Omu_size2_rep1 and Omu_size1_rep2).

Differential expression was analyzed with DESeq2 package [6] for R by calculating the logarithm of the change in transcript abundance and adjusted p-value. Only those transcripts were used in the analysis that had a representation higher than 1 count per million in at least two samples in a group. Transcripts whose expression changed more than 2-fold and p-value was less than 0.05 were considered differentially expressed. The groups were compared in pairs: group I vs group II (Fig. 1C), group I vs group III (Fig. 1D), group II vs group III (Fig. 1E). When comparing groups I and II, group I up-regulated expression of 1990 transcripts and group II up-regulated expression of 1608 transcripts. When comparing groups II and III, group II up-regulated expression of 781 transcripts, group III up-regulated expression of 739 transcripts.

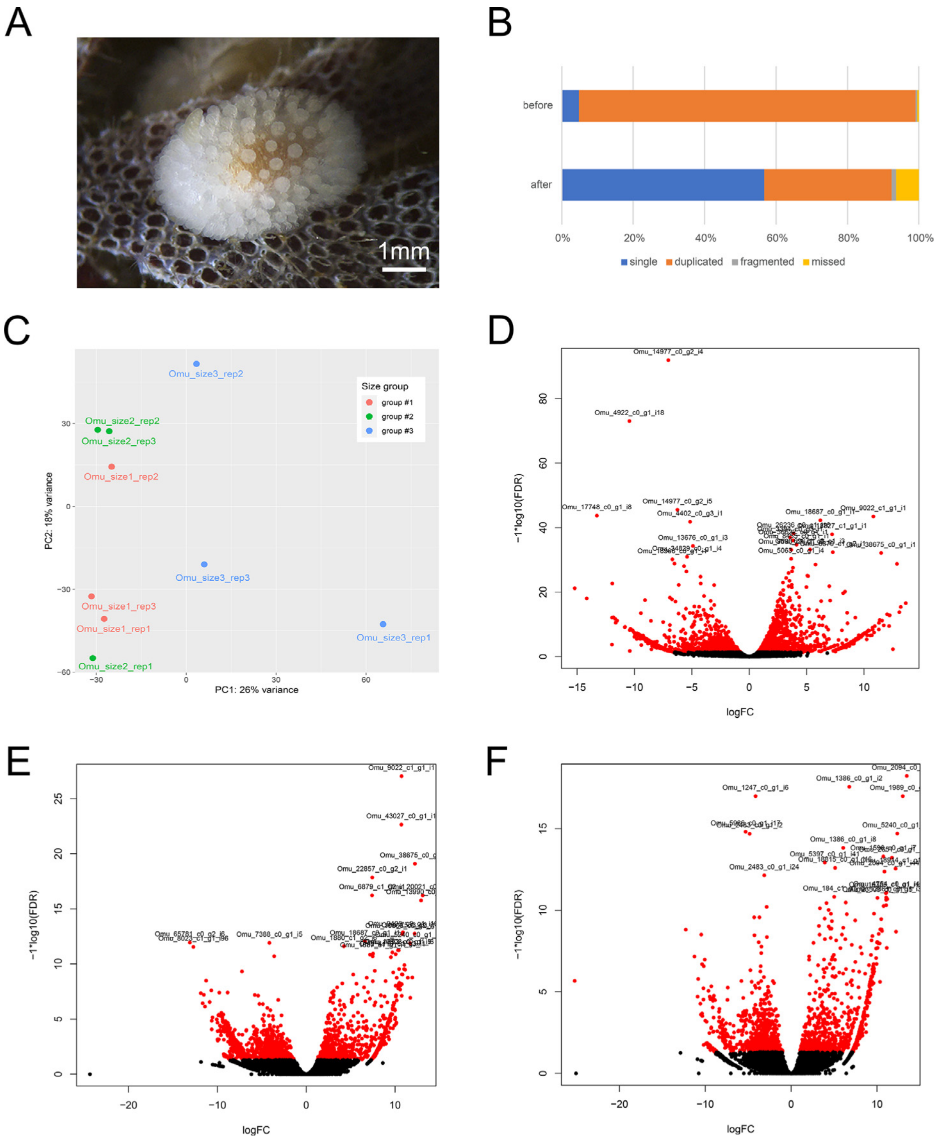

Fig. 1. Transcriptome completeness and differential expression in *Onchidoris muricata*. (A) *O. muricata* in natural environment. (B) BUSCO scores before and after decontamination of assembly. (C) Expression levels in three size groups were analyzed using the principal component analysis (PCA) method. The 500 genes with most variable expression were analyzed. (D) Differential transcript expression in groups I and II. Each dot on the graph represents a transcript; black dots have p-value > 0.05, red dots have p-value < 0.05. Red dots to the right of zero on the abscissa axis are transcripts with log fold change (L2FC) greater than one in group I (i.e., up-regulated in this group), and to the left of zero are transcripts with L2FC > 1 in group II. On the abscissa axis is L2FC, on the ordinate axis is the decimal logarithm of p-value. (E) Differential transcript expression in groups I and III. (F) Differential transcript expression in groups II and III.

4. Experimental Design, Materials and Methods

4.1. Collection of material

The study focused on *Onchidoris muricata* (O.F. Müller, 1776), with its spicule complex well described. The material was collected from the low tide zone of the Eremeevsky rapids (66°33′17″N, 33°06′02″E), using scuba diving near the WSBS (White Sea Biological Station) of Lomonosov Moscow State University the Kanadalksha Bay of the White Sea, at depths of 10 to 15 m in August. A total of 96 individuals were collected. All individuals were divided into three groups based on morphological characteristics: group I from 100 µm to 1 mm, group II from 1 to 4 mm, and group III from 4 to 12 mm. The sample contained 10–11 individuals of the first group, 9–10 individuals of the second group and 6–7 individuals of the third group. 3–5 samples were collected for each size group.

This division of the molluscs correlates with their size and the formation of the spicule complexes [4]. Afterwards, in all the molluscs except for the group I, the visceral organ complex was removed. This is because in the group I it is not yet developed, and in large molluscs it would distort the picture of gene expression in the integument. The integument in the group II and III, as well as in the group I molluscs were totally placed in an IntactRNA solution (cat # BC031, Evrogen), for 1 hour at room temperature on a shaker (110 rpm). After that, the solution was replaced with a similar fresh one. Then, the objects were frozen at −80 °C.

4.2. RNA extraction

After thawing, tissues were separated from IntactRNA by centrifugation for 5 min at 10,000 g and homogenized in ExtractRNA reagent (cat #BC032, Evrogen) using a Potter glass homogenizer. The homogenate was centrifuged at +4 °C and 15,000 g for 15 min, after which RNA was isolated from the supernatant according to the manufacturer’s protocol. RNA pellet was dissolved in nuclease-free water. The extracted RNAs were analyzed with 1 % agarose gel electrophoresis in TAE buffer with ethidium bromide staining, and on-chip electrophoresis (Bioanalyzer 2100, Agilent).

4.3. Illumina library preparation and sequencing

The cDNA libraries were prepared from the polyadenylated RNA fraction using TruSeq2 chemistry, and sequenced on an Illumina NovaSeq6000 instrument with 150 bp long paired-end reads and a yield of at least 30 million reads. Raw reads from instrument were converted to FASTQ format with bcl2fastq2 software (Illumina).

4.4. Transcriptome assembly and annotation

Raw paired-end reads were preprocessed as follows. Kmer-based read corrections was performed with Rcorrector [7] v1.0.6 followed by removing read pairs where at least one read has been flagged by rCorrector as containing an erroneous kmer, and where it was not possible to computationally correct the errors. Then quality trimming with TrimGalore v0.6.10 was implemented with flags “–length 36 -q 5 –stringency 1 -e 0.1″ to remove low-quality and too short reads. The resulting set of reads was mapped using bowtie2 [8] v2.5.2 to the ribosomal RNA database, and only reads that did not match the database were used for assembly. The database was composed of sequences of large and small subunit rRNAs SILVA 138.1 (LSUParc и SSUParc). The assembly was performed *de novo* using Trinity [5] v2.12.0 with default functions. The resulting assembled transcripts were collapsed to eliminate polymorphisms using cd-hit-est [9] v4.8.1

Table 1
RNA-seq samples, read metrics, and public SRA accessions.

Sample name	Sequencer	Animal size group	Number of raw reads (M)	Number of quality trimmed reads (M)	Uniquely mapped reads,%	SRA accession
Omu_size1_rep1	Illumina NovaSeq6000 PE 150	Group #1	40,60	38,50	92,54	SRR26553392
Omu_size1_rep2	Illumina NovaSeq6000 PE 150	Group #1	46,60	44,30	75,06	SRR26553391
Omu_size1_rep3	Illumina NovaSeq6000 PE 150	Group #1	40,00	38,00	95,44	SRR26553390
Omu_size2_rep1	Illumina NovaSeq6000 PE 150	Group #2	44,60	42,20	91,70	SRR26553389
Omu_size2_rep2	Illumina NovaSeq6000 PE 150	Group #2	42,50	40,10	93,28	SRR26553388
Omu_size2_rep3	Illumina NovaSeq6000 PE 150	Group #2	42,10	39,50	93,77	SRR26553387
Omu_size3_rep1	Illumina NovaSeq6000 PE 150	Group #3	36,30	34,90	85,70	SRR26553386
Omu_size3_rep2	Illumina NovaSeq6000 PE 150	Group #3	40,80	38,40	91,77	SRR26553385
Omu_size3_rep3	Illumina NovaSeq6000 PE 150	Group #3	42,20	39,60	87,96	SRR26553384

and an identity threshold of 0.95. The transcriptome was decontaminated using MCSC [10] for phyla Mollusca with a clustering level of $n = 4$. Decontaminated assembly used for transcripts quantification. Open reading frames (ORFs) and peptide sequences of at least 100 amino acids in length were predicted using TransDecoder. Trinotate [11] v4.0.2 pipeline including BLASTP and BLASTX using diamond [12] v2.0.15.153 with an e-value cutoff of 1e-5 against UniProtKB/Swiss-Prot database, signalp 6 [13], tmhmm 2.0 [14] and hmmscan 3.0 [15] against Pfam database were used for annotation. The obtained data are summarized in the Table 1.

4.5. Identification and analysis of differentially expressed genes

To identify differentially expressed genes, raw reads were aligned to the assembly using bowtie2 v2.5.2 and abundance estimation of each contig was calculated using RSEM [16] v1.3.3. Differential expression analysis was performed in the DESeq2 [6] package. Plots and heatmaps were generated using the pheatmap package. Transcripts were recognized as differentially expressed if the expression level between groups changed more than 2-fold ($\text{Log2FC} > 1$) and the adjusted p-value was less than 0.05. Gene Ontology (GO) enrichment analysis of differentially expressed transcripts between groups was implemented by the R package GOseq [17]. GO terms with a Benjamini and Hochberg (BH)-corrected p-value of less than 0.05 were considered significantly enriched.

4.6. Quality control

FastQC was used to assess sequencing quality before and after adapter trimming and filtering for quality and length. The quality of the assembly was checked by mapping the source reads to the assembly as well as by BUSCO [18].

Limitations

Not applicable.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Transcriptome of the nudibranch mollusk Onchidoris muricata (Doridina, Gastropoda) at different stages of early ontogeny with special refer to spiculogenesis (Original data) (NCBI).

Omu-transcriptome (Original data) (GitHub).

CRediT Author Statement

Ekaterina D. Nikitenko: Investigation, Data curation; **Ilya E. Borisenko:** Investigation, Methodology, Formal analysis, Software; **Andrey N. Anisenko:** Investigation, Validation, Formal analysis, Software; **Elena V. Vortsepneva:** Investigation, Data curation, Validation, Supervision.

Acknowledgements

The authors would like to express their gratitude to Kocot K. and Yap-Chiongco M., who provided inspiration and assistance at the beginning of this challenging journey. We are deeply grateful to Kremnyov S.V., Kozin V.V., Lavrov A.I. for their valuable advice and direct assistance in our work. The authors express their gratitude to Semenov A.A., Chava A.I., Ozerov D.A., Zakharov G.D., Petrova M.A., Prudkovsky A.A. who assisted in collecting material, as well as all our colleges from the White Sea Biological Station and colleagues from St. Petersburg University. Research was carried out using computational resources provided by Resource Center "Computer Center of SPbU". Support for research within the Interdisciplinary Scientific and Educational Schools of Moscow University No. 23-Ш04-29.

Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2024.110526.

References

[1] H. Wägele, A. Klussmann-Kolb, Opisthobranchia (Mollusca, Gastropoda)–more than just slimy slugs. Shell reduction and its implications on defence and foraging, Front. Zool. 2 (2005) 1–18, doi:10.1186/1742-9994-2-3.

[2] S.J. Foale, R.C. Willan, Scanning and transmission electron microscope study of specialized mantle structures in dorid nudibranchs (Gastropoda: Opisthobranchia: Anthobranchia), Mar. Biol. 95 (1987) 547–557, doi:10.1007/BF00393098.

[3] B.K. Penney, K.R. Ehresmann, K.J. Jordan, G. Rufo, Micro-computed tomography of spicule networks in three genera of dorid sea-slugs (Gastropoda: Nudipleura: Doridina) shows patterns of phylogenetic significance, Act. Zool. 101 (1) (2020) 5–23, doi:10.1111/azo.12266.

[4] E. Nikitenko, A. Ereskovsky, E. Vortsepneva, Ontogenetic dynamics of the subepidermal spicule complex in Nudibranchia (Gastropoda): the case of *Onchidoris muricata*, Zool 144 (2021) 125886, doi:10.1016/j.zool.2020.125886.

[5] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity, Nat. Protoc. 8 (8) (2013) 1494–1512, doi:10.1038/nprot.2013.084.

[6] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 1–21, doi:10.1186/s13059-014-0550-8.

[7] L. Song, L. Florea, Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads, Gigascience 4 (2015) s13742-015, doi:10.1186/s13742-015-0089-y.

[8] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359, doi:10.1038/nmeth.1923.

[9] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152, doi:10.1093/bioinformatics/bts565.

[10] J. Lafond-Lapalme, M.-O. Duceppe, S. Wang, P. Moffett, B. Mimee, A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm, Bioinformatics 33 (2017) 1293–1300, doi:10.1093/bioinformatics/btw793.

[11] D.M. Bryant, K. Johnson, T. DiTommaso, T. Tickle, M.B. Couger, D. Payzin-Dogru, T.J. Lee, N.D. Leigh, T.-H. Kuo, F.G. Davis, J. Bateman, S. Bryant, A.R. Guzikowski, S.L. Tsai, S. Coyne, W.W. Ye, R.M. Freeman, L. Peshkin, C.J. Tabin, A. Regev, B.J. Haas, J.L. Whited, A tissue-mapped Axolotl De Novo transcriptome enables identification of limb regeneration factors, Cell Rep. 18 (2017) 762–776, doi:10.1016/j.celrep.2016.12.063.

[12] B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND, Nat. Methods 18 (2021) 366–368, doi:10.1038/s41592-021-01101-x.

[13] F. Teufel, J.J. Almagro Armenteros, A.R. Johansen, M.H. Gíslason, S.I. Pihl, K.D. Tsirigos, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 6.0 predicts all five types of signal peptides using protein language models, Nat. Biotechnol. 40 (2022) 1023–1025, doi:10.1038/s41587-021-01156-3.

[14] S. Möller, M.D. Croning, R. Apweiler, Evaluation of methods for the prediction of membrane spanning regions, Bioinformatics 17 (2001) 646–653, doi:10.1093/bioinformatics/17.7.646.

[15] S.R. Eddy, Accelerated profile HMM searches, PLOS Comput. Biol. 7 (2011) e1002195, doi:10.1371/journal.pcbi.1002195.

[16] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinform. 12 (2011) 1–16, doi:10.1186/1471-2105-12-323.

[17] M.D. Young, M.J. Wakefield, G.K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: accounting for selection bias, Genome Biol. 11 (2010) 1–12, doi:10.1186/gb-2010-11-2-r14.

[18] M. Manni, M.R. Berkeley, M. Seppey, F.A. Simão, E.M. Zdobnov, BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, Mol. Biol. Evol. 38 (2021) 4647–4654, doi:10.1093/molbev/msab199.