Comparative transcriptomic in selected species of marine animals

mgr Magdalena Małachowicz

Transcriptomics applies to the study of the complete collection of transcripts (RNA molecules synthesized by transcription of DNA) present in a specific cell or tissue, at a certain moment. The transcriptome consists of protein-coding messenger RNA (mRNA) and non-coding RNA (ncRNA), essential in proper functioning of cells, including microRNA (miRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and other ncRNAs. In addition to the genome, the transcriptome depends on the type of tissue examined, the stage of development, physiological and environmental conditions. Transcriptome analysis provides information on the gene expression, function and regulation of genes, which help the understanding of the functioning of specific tissues and the linkage between genotype and phenotype. This helps to clarify the role of molecular mechanisms underlying genetic diversity and organismal response to environmental conditions.

Currently, two techniques are used for transcriptome analysis: hybridization of labelled cDNA/cRNA to cDNA/oligonucleotide probes fixed onto microarray and sequencing. Due to several limitations of microarray technologies (signal saturation for transcripts with a high expression, background noise caused by nonspecific hybridization and dependence on the genomic data), recently, the use of the sequencing method for transcriptome analysis has been increasing. This technique is particularly useful in non-model organisms, where full genome data is still not available and genomic resources are limited. Besides gene expression, transcriptome sequencing enables detection of known transcripts/isoforms. It provides information needed for the genome annotation such as intron/exon boundaries. It also enables discovery of novel transcripts, noncoding RNAs and alternative splicing variants, resulting from post-transcriptional modification of mRNA. It offers advantages in prediction of the protein structure and function with no prior knowledge about the genes. Transcriptome sequencing is also an effective way to obtain a large number of molecular markers such as polymorphisms of short sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs), within coding regions, determining functional genetic variation.

Rapid evolution of transcritpomics over the past decade has been driven by the development of next generation sequencing (NGS). NGS refers to massively parallel sequencing technologies, which can produce thousands to billions of reads in one run. NGS methods can be divided into two main categories: second and third generation sequencing (SGS and TGS). The first group include PCR-based technologies, such as Roche/454 pyrosequencing (2005), Illumina sequencing (2007), Ion Torrent (2010) all based on sequencing-by-synthesis (SBS) and AB SOLiD (2007) - sequencing-by-ligation (SBL). The second group are techniques capable of sequencing single DNA molecules (single-molecule sequencing, SMS) such as HeliScope (2008), PacBio RS SMRT system (2010) and sequencer using nanopore technology - Oxford Nanopore MinION (2014). The major advance offered by NGS, also known as high-throughput sequencing is the ability to produce an enormous volume of data cheaply in a short time. This advance has revolutionized molecular research on organisms, including marine species. The development of high-throughput techniques and the generation of large amounts of data including transcriptomic, has led to the combination of biological sciences such as molecular biology, genomics, transcriptomics with mathematics and informatics, leading to the creation of an interdisciplinary field of science - bioinformatics. Analysis of the sequencing results is constituted by different steps, from RNA extraction, assembling, through functional classification and transcript annotation, to biological insight (Figure 1). For this purpose, numerous bioinformatic programs have been developed, based on different algorithms and models. The diversity of the available software makes it possible to customize analysis protocols to a specific goal.

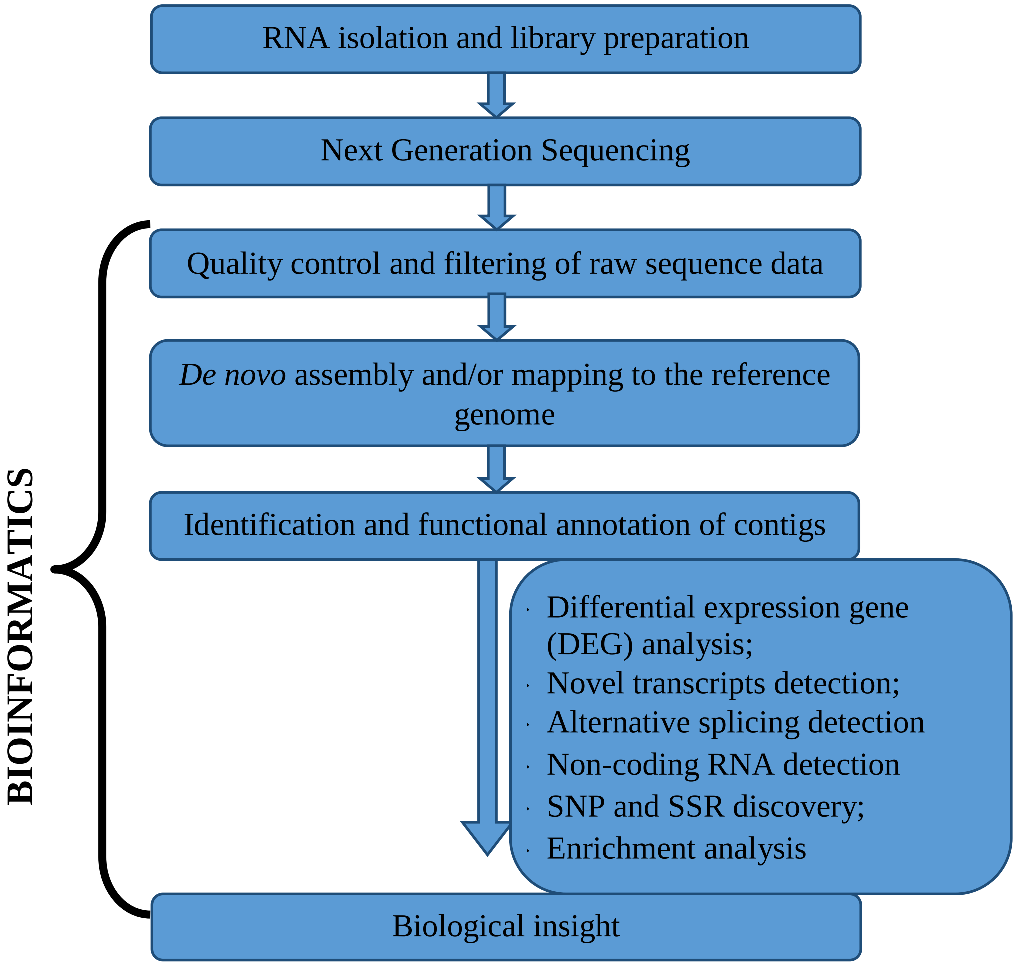
RNA isolation and library preparation
Next Generation Sequencing
Quality control and filtering of raw sequence data
De novo assembly and/or mapping to the reference genome
Identification and functional annotation of contigs
Differential expression gene (DEG) analysis;
Novel transcripts detection;
Alternative splicing detection
Non-coding RNA detection
SNP and SSR discovery;
Enrichment analysis
Biological insight
BIOINFORMATICS

Figure 1. Next-generation sequencing (NGS) workflow from RNA extraction to biological insight.

Transcriptomics coupled with next generation sequencing, bioinformatics and biostatistics provides new opportunities to address biological issues such as: physiological processes, organismal development, genetic structure and diversity of the population, adaptation to environmental conditions, evolution, phylogenetics and immunology. Thanks to widespread application, it has become the main tool for evolutionary biology. Biodiversity through the transcriptome analysis is a valuable source of information in understanding how global climate changes, anthropogenic activities and pathogen diseases affect organisms. Understanding these issues contributes to the effective management of ecologically and economically important marine species.

The primary aim of my PhD thesis was to describe the genetic diversity of selected marine species which plays a crucial role in meeting global food demand that is fish and bivalves of significant economic and ecological importance, including: Atlantic cod (*Gadus morhua*) from the Baltic Sea, sea trout (*Salmo trutta* m. *trutta*), mussels from the genus *Mytilus* (*Mytilus edulis*, *Mytilus galloprovincialis*, *Mytilus trossulus* and *Mytilus chilensis*). For this purpose I conducted transcriptome analyses using next generation sequencing method. Tissues of key importance in contact and in response to environmental conditions were selected (gills - response to salinity, skin - first line of defense of the organism, mantle - biomineralization). In all parts of my work, after RNA isolation, samples were processed using 454 pyrosequencing technique (Roche GS-FLX). Subsequently, the raw reads were pre-processed by removing adapters and low quality sequences with CLC Genomics Workbench software (v.7.5.5, CLC Bio, Qiagen, Aarhus, Denmark). The quality reads were assembled into contigs using de-Bruijn graphs in CLC Genomics Workbench [1, 2, 3]. Basic Local Alignment Search Tool (BLAST) tool was used for contigs searches against public databases e.g. the National Center of Biotechnology Information (NCBI). Next, functional classifications (gene ontology, GO, Blast2GO software) were conducted and contigs were assigned to the three main ontology categories: molecular function (MF), biological process (BP) and cellular component (CC). Using appropriate bioinformatics tools i.e. KEGG Automatic Annotation Server (KAAS), I conducted KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analyses. The results were compared using available databases, from different taxa and tissues (comparative transcriptomics) [1, 2, 3].

The first chapter of my thesis presents a transcriptome comparison between the Atlantic cod (*Gadus morhua* L.) from the Atlantic Ocean and the Baltic Sea [1]. The Atlantic cod is one of the most ecologically and economically important marine fish species in the North Atlantic Ocean. In the Baltic Sea it lives at low salinity (7-12 PSU) but requires well oxygenated water and salinity over 14 PSU for successful spawning, this limits areas suitable for reproduction. Using pyrosequencing technique 962,516 reads representing 379 Mbp of the Baltic cod transcriptome were obtained. Data was assembled into 14,029 contigs representing 13,585 genes, of which 75.39% had an ontology definition, with a dominant molecular function category. In addition, these contigs were compared with the Atlantic Ocean cod reference transcriptome (using BLASTn), from Ensembl database. Despite a high similarity between transcriptomes, 202 Baltic cod contigs (170 genes) were significantly distinct from Atlantic cod. Among these genes 100% were protein coding sequences and 68.82% of them had an ontological category. The identified differences could have been caused by geographic origin and environment conditions e.g. salinity [1]. Further analysis of transcriptome revealed the potential role of alternative splicing in the process of adaptation to altered salinity [4] (this publication is not a part of the presented doctoral dissertation).

In the second part of my work I compared the transcriptome of sea trout (*Salmo trutta* m. *trutta*) skin with other tissues (sequences obtained from literature), indicating the expression of numerous genes putatively involved in the immune response and mucus secretion, including mucins [2]. The sea trout is an anadromous form of brown trout (*Salmo trutta*) with important commercial and ecological value in Europe. The trout population decreased as a result of the construction of numerous partitions and dams on the rivers, which hindered or prevented the migration of spawners up the rivers for reproduction, as well as due to pollution, environmental degradation and diseases. Salmonids differ in their susceptibility to microorganisms due to varied skin morphology, which is the first barrier against pathogens and gene expression patterns. Thus, I characterized the skin transcriptome of sea trout using NGS. As a result of sequencing a total of 1,348,306 filtered reads were obtained and assembled into 75,970 contigs. Of these contigs 48.57% were identified using BLAST tool searches against protein databases e.g. NCBI. KEGG pathway and gene ontology analyses revealed that 13.40% and 34.57% of the annotated transcripts, respectively, represent a variety of biological processes and functions. Among the identified KEGG Orthology categories, the best represented were signal transduction (23.28%) and immune system (8.82%), with a variety of genes involved in immune pathways, implying the differentiation of immune responses in the sea trout skin. The GO enrichment analysis revealed many genes involved in stress reaction and immune response against pathogens in the skin of trout in comparison with other tissues. In addition, over 85% of homologous contigs showed similarity >95%, and 9.57% were new transcripts, potentially specific for the skin tissue. Additionally, I identified and characterized 8 types of mucins – glycoproteins, the main constituents of skin mucus, and 140 genes involved in mucin biosynthesis. Moreover, 1,119 potential simple sequence repeats were identified [2].

The third chapter of the thesis is devoted to comparative mantle transcriptome analysis within the genus *Mytilus* and to identification of putative genes involved in the biomineralization and pigmentation processes [3]. Mussels are marine bivalve species occurring in coastal waters in both the Northern and Southern Hemispheres, with significant ecological and commercial value. Mussels are the subject of a wide biogeographic research being conducted in the Laboratory of Genetics of Marine Organisms in the Institute of Oceanology PAS using the SNP genotyping. Due to their prevalence and genetic diversity, mussels are a good model for adaptation studies to environmental conditions. In addition, due to their filter-feeding abilities they also play an important role as biosensors for pollution. In mollusks, the shell secreted by mantle tissue during the biomineralization process forms the first barrier against predators and mechanical damage. Changing environmental conditions such as rising CO_2, which causes seawater pH changes, affects the shell strength, thus the ability to protect the soft body. To improve the understanding of the above issues, I characterized and compared the mantle transcriptomes of four mussel taxa from different geographic locations (*M. edulis* – the North Sea, *M. galloprovincialis* – the Mediterranean Sea and Tasmania, *M. chilensis* - Chile, *M. trossulus* – Vancouver), as representatives of "pure" taxa, using next generation sequencing. A total of 743,773 filtered reads were assembled into 20,982 contigs. Of these, 46.57%, 37.28% and 17.53% were identified using NCBI NR, GO and KEGG databases. I identified a total of 1,292 contigs potentially involved in biomineralization and melanogenesis and selected and analyzed sequences involved in response to ocean acidification, such as: carbonic anhydrase, chitinase and tyrosinase. I also identified potential SSRs (483) and SNPs (1,497), which provide resource for molecular markers studies associating biomineralization genes with different condition and taxa. For transcriptome comparisons, I conducted a GO enrichment analysis (Blast2GO, Fisher exact test) and identified orthologs (OrthoMCL software) using available databases. The GO enrichment analysis revealed pH and thermal response in *M. edulis* from the North Sea and *M. galloprovincialis* from the Mediterranean Sea. Phylogenetic analysis between the six *Mytilus* taxa using transcriptome data revealed *M. californianus* and *M. coruscus* to be genetically more distant (they formed a separate clad) from the other taxa: *M. edulis*, *M. chilensis*, *M. galloprovincialis* and *M. trossulus* [3].

Transcriptomics might be used for characterization of evolutionary relationships of the genes-of-interest between species through phylogenetic analysis. For this purpose, I used transcriptomic data obtained from commercially important marine species (Baltic cod, sea trout and mussels) for carbonic anhydrases (CAs) analysis. The CA superfamily consisting of three major families: α-CA (metazoans, bacteria, fungi), β-CA (plants) and γ-CA (archaea and bacteria) is considered as a class of the metalloenzyme, which catalyses the reversible hydration of carbon dioxide (CO_2) to form bicarbonate ion and proton ($CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow HCO_3^- + H^+$). In mammals the α-CA family is characterized by 13 different enzymatically active isoforms and 3 CA-related proteins which appear to lack CA activity (CARPs). CA plays a crucial role in ion regulation, CO_2 excretion, acid-base balance and biomineralization in fish and shellfish. The characterization of isoforms in terms of interspecific diversity, activity and tissue distribution in marine organisms is required due to rising atmospheric carbon dioxide emissions connected with both ocean acidification and global warming. To understand the evolutionary relationship and comparison of eumetazoa α-CA I compared contigs identified as CA obtained from presented studies and proteins available at the NCBI (Figure 2).

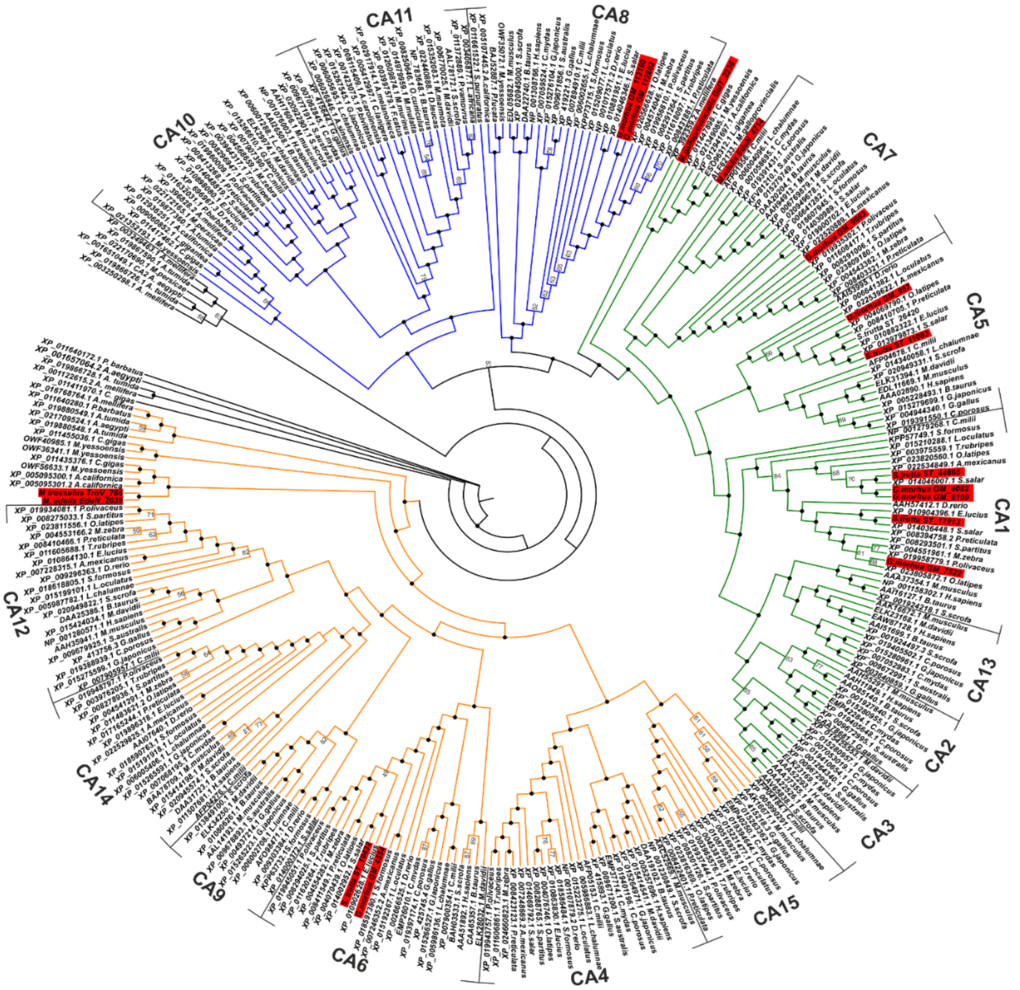

Figure 2. Phylogenetic tree of α-CAs from various eumetazoa taxa. Color scheme: red – sequences obtained from presented studies; blue - CARPs; green – intracellular CAs; orange – extracellular CAs.

Incomplete, without domains and/or highly divergent CA sequences, were excluded from the analysis. Homologous sequences (a total of 319 proteins) were aligned using multiple alignment program MUSCLE and trimmed using TrimAl with a 50% threshold. The phylogenetic tree was inferred using IQ-TREE – the Whelan and Goldman (WAG) +R9 model. Sequences obtained from the studies presented were classified mostly as an intracellular CAs, CA-related proteins and extracellular CA6 (Figure 2). The phylogenetic analysis revealed division into two major clusters: intracellular (CA1, CA2, CA3, CA5, CA7, CA13) and extracellular CAs (CA4, CA6, CA9. CA12, CA14, CA15; Figure 2). The results suggest that there is a lack of CA3, CA11 and CA13 isoforms in fish and bivalves, however this might be caused by the lack of CA sequences in database. Moreover, CARPs showed higher similarity to intracellular CA and it seems that they are well conserved through all species suggesting that they may play important biological role.

The original reports which have become the basis for my PhD thesis, had broadened knowledge of the genetic diversity and biodiversity of several important marine species including cod from the Baltic Sea, sea trout and mussels of the genus *Mytilus*. The further results obtained expand sequence databases for the above-mentioned species and provide valuable resources for further evolutionary, genomic and phylogenetic research. All findings included in my dissertation show that comparative transcriptomic analyses between closely related species from different geographic locations and between tissues, through next generation sequencing is an effective method to discover genetic diversity and its functional meaning and further, gives the opportunity to understand how organisms adapt to different environmental conditions. Thus transcriptomic analysis can be a useful approach for effective species management. One main implication of this study is to establish transcriptomic databases for key marine organisms. Very important in the analysis of sequencing results is the publication of the obtained data. Sequences generated as a part of the presented publications were deposited in NCBI databases: Sequence Read Archive (SRA) accession numbers: PRJNA273805 [1], PRJNA323793 [2] and PRJNA419475 [3] and GenBank: KY328727-KY328740 [2], MG827120–MG827134 [3]. To demonstrate the usefulness and versatile application of established transcriptomes, phylogenetic analysis of carbonic anhydrases in animals was conducted using obtained sequences. The publications presented

results from the realization of grants to Professor R. Wenne by the National Science Centre, numbers: 2011/01/M/NZ9/07207 and 2011/01/B/NZ9/04352 and the Ministry of Science and Higher Education, number: 397/N-cGRASP/2009/0.2015.

References:

1. **Malachowicz M**, Kijewska A, Wenne R (2015) Transcriptome analysis of gill tissue of Atlantic cod (*Gadus morhua* L.) from the Baltic Sea. Marine Genomics, 23: 37-40.
2. **Malachowicz M**, Wenne R, Burzynski A (2017) *De novo* assembly of the sea trout (*Salmo trutta* m. *trutta*) skin transcriptome to identify putative genes involved in the immune response and epidermal mucus secretion PLOS ONE, 12(2): e0172282.
3. **Malachowicz M**, Wenne R (2019) Mantle transcriptome sequencing of *Mytilus* spp. and identification of putative biomineralization genes. PeerJ, 6:e6245. Doi: 10.7717/peerj.6245
4. Kijewska A, **Malachowicz M**, Wenne R (2018) Alternatively spliced variants in Atlantic cod (*Gadus morhua*) support response to variable salinity environment. Scientific Reports, 8: 11607. doi: 10.1038/s41598-018-29723-w (this publication is not a part of presented doctoral dissertation).