University of Minho
School of Engineering

Beatriz Ferreira da Silva

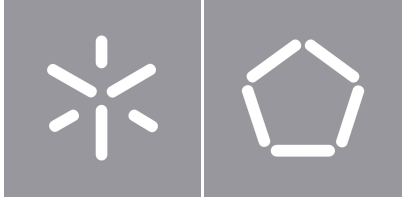
# Identify gene expression profiles in freshwater mussels under thermal stress

january 2024

Identify gene expression profiles in freshwater mussels under thermal stress

Beatriz Ferreira da Silva

UMinho | 2024



**University of Minho**
School of Engineering

Beatriz Ferreira da Silva

# Identify gene expression profiles in freshwater mussels under thermal stress

Masters Dissertation
Master's in Bioinformatics

Dissertation supervised by
**Elsa Maria Branco Froufe Andrade**
**Ronaldo Gomes de Sousa**

january 2024

## DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras de boas práticas internacionalmente aceites, no que diz respeito aos direitos de autor e direitos conexos. Assim, o presente documento pode ser utilizado nos termos previstos na licença abaixo indicada. Caso o utilizador careça de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

***Licença concedida aos utilizadores deste trabalho***



**Atribuição-NãoComercial-SemDerivações**
**CC BY-NC-ND**
https://creativecommons.org/licenses/by-nc-nd/4.0/

# AGRADECIMENTOS

Começo por agradecer aos meus orientadores, Professor Ronaldo Sousa e Doutora Elsa Froufe, pela orientação que me prestaram durante a realização deste trabalho, em particular à Doutora Elsa Froufe, pelo seu apoio, preocupação e motivação constante. Além disso, os meus sinceros agradecimentos vão para o Doutor André Santos por me ter apoiado em tudo do início ao fim do meu percurso. Fica também um especial obrigado à restante equipa do Ciimar, pelo ambiente acolhedor que proporcionaram durante esta estadia, e, em particular, ao Rui Pinto, por me ter auxiliado em todas as minhas dúvidas sobre o desenvolvimento do código.

Quero também agradecer aos meus pais, em muito, por me terem dado a possibilidade de realizar esta dissertação. Um obrigada, por sempre acreditarem em mim, mesmo quando eu própria não acreditei. Ao meu irmão, que nada percebia da minha aréa e que sabia a minha dissertação de trás para a frente, um obrigada.

Ao Diogo, com quem partilhei todas as minhas frustações e vitórias nesta jornada, um obrigado por seres um pilar e por sempre acreditares em mim.

Como um refúgio extra nos momentos de maior stress foram todos os meus amigos, que sempre me deram forças e ajudaram a descontrair quando assim necessitava. Não me poderei esquecer também das minhas grandes amigas do secundário, Andreia Taveira, Ana Mota e Beatriz Martins, que sempre me apoiaram e me deram uma motivação extra. A todos eles, o meu enorme obrigado!

O trabalho desenvolvido nesta dissertação está incorporado no projeto: EdgeOmics - Freshwater Bivalves at the edge: Adaptation genomics under climate-change scenarios (http://doi.org/10.54499/PTDC/CTA-AMB/3065/2020).

## DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

## Abstract

Due to global climate change, the temperatures of streams and rivers are increasing, negatively affecting aquatic life, including bivalve species. Freshwater mussels are vital components of rivers, streams, and lake ecosystems, participating in essential ecological roles such as nutrient cycling, and increasing water quality. Furthermore, they serve as essential ecosystem engineers, providing habitat to other organisms and supporting intricate food webs. Besides their biological importance, freshwater mussels are poorly studied in terms of genomics. In the present work, the Iberian dolphin freshwater mussel *Unio delphinus* Spengler, 1793 (Bivalvia: Unionoida) was used as a model species to investigate the effects of climate change in freshwater mussels. The primary objective of this thesis was to determine the gene expression patterns in a model species of freshwater mussels under the effects of thermal stress exacerbated by climate change, with an overall goal of understanding the potential consequences for freshwater mussel populations. Two different ecological experiments were performed: chronic and acute. The chronic experiments where temperatures were gradually increased to simulate a scenario of progressive increasing temperatures. The acute experiments where temperatures were rapidly increased to replicate the effects of a briefer extreme climatic event. To achieve this main goal, a comprehensive bioinformatic pipeline focused on transcriptomics analysis was developed using the R Bioconductor package to generate the differential gene expression profiles of these individuals under thermal stress. The bioinformatic methodology of this work differs from the past studies, by developing an R code compilation of three methods, EdgeR, limma, and DESeq2 for differential gene expression analysis in these organisms. The output of the present work provides a comprehensive overview of gene expression profile responses of *U. delphinus* under climate change scenarios. Additionally, the results revealed a wide range of pathways and the corresponding genes that are impacted by thermal stress, with a particular emphasis on the up-regulation of the genes ATP6V1A, ATP6V0A1, ATP6V0A, and ATP6V1. In the chronic experiments, and high temperatures, mussels expressed these genes and, interestingly, all the pathways that these genes included appeared up-regulated. The discovered genes and pathways provide vital insights into these organisms' adaptation tactics and identify prospective targets for monitoring and conservation efforts.



**Keywords**: Freshwater mussels; transcriptomics; bioinformatics; gene expression; climate change.

## Resumo

Devido às alterações climáticas globais, a temperatura dos rios e ribeiros está a aumentar, afectando a vida aquática, incluindo as espécies de bivalves. Os mexilhões de água doce são componentes dos rios, ribeiros e lagos, participando em funções vitais dos ecossistemas, como a reciclagem de nutrientes e melhorando a qualidade da água. Além disso, atuam como engenheiros de ecossistemas, providenciando habitat a outros organismos e fazendo parte de teias alimentares complexas. Apesar da sua importância biológica, os mexilhões de água doce são pouco estudados em termos genómicos. No presente trabalho, o mexilhão de água doce *Unio delphinus* Spengler, 1793 (Bivalvia: Unionoida) é uti-lizado como espécie modelo para investigar os efeitos das mudanças climáticas em mexilhões de água doce. O objetivo principal desta dissertação foi determinar os padrões de expressão genética numa es-pécie modelo de mexilhões de água doce sob os efeitos de stress térmico exacerbado pelas mudanças climáticas, com o objetivo geral de comprender as consequências para populações de mexilhões de água doce. Duas experiências foram realizadas: crónica e aguda. As experiências crónicas onde a temperatura foi gradualmente aumentada para simular um cenário de aumento progres-sivo da temperatura. As experiências agudas onde a temperatura foi aumentada rapidamente de modo a reproduzir os efeitos de um extremo climático breve. De modo a atingir esse objetivo, uma pipeline de bioinformática focada na análise transcriptómica foi desenvolvido usando o pacote R Bioconductor para gerar os perfis diferenciais de expressão genética desses indivíduos sob stress térmico. A metodologia bioinformática deste trabalho difere dos estudos anteriores, por desenvolver uma compilação de código em R de três métodos, EdgeR, limma e DESeq2 para análise diferencial de expressão genética nesses organismos. O resultado do presente trabalho fornece uma visão abrangente das respostas do perfil de expressão genética de *U. delphinus* num cenário de mudanças climáticas. Os resultados revelaram uma ampla gama de vias e os genes correspondentes relacionados com o stress térmico, com ênfase particular na regulação positiva dos genes ATP6V1A, ATP6V0A1, ATP6V0A e ATP6V1. Nas experiências crónicas, a uma temperatura elevada, os mexilhões expressaram estes genes e, curiosamente, todas as vias que estes genes incluíam pareciam reguladas positivamente. Os genes e vias descobertos fornecem informações vitais sobre adaptação destes organismos e identificam alvos potenciais para esforços de conservação.



**Palavras-chave**: Mexilhões de água doce; transcriptómica; bioinformática; expressão genética; alterações climática.

# Contents


**1 Introduction** **1**

1.1 Context/motivation . . . 1

1.2 Objectives . . . 2

1.3 Thesis Structure . . . 2

**2 State of the Art** **4**

2.1 Transcriptomics . . . 4

2.1.1 A brief history . . . 4

2.1.2 Next Generation Sequencing . . . 6

2.2 RNA Sequencing . . . 11

2.2.1 RNA-Seq workflow . . . 11

2.2.2 Applications of RNA-Seq data . . . 15

2.2.3 Differential gene expression analysis . . . 17

2.3 Freshwater mussels as a model organism . . . 20

2.4 The effects of climate change on freshwater ecosystems . . . 22

2.4.1 Physiological processes . . . 23

2.4.2 Phenological processes . . . 24

2.4.3 Genetic variation . . . 25

2.5 Bioinformatics Integration . . . 26

**3 Methodologies** **28**

3.1 Brief overview . . . 28

3.1.1 Sampling and laboratory acclimatization . . . 28

3.1.2 Ecological Experiment . . . 28

3.1.3 RNA extraction and NGS sequencing . . . 30

3.2 Bioinformatics . . . 30

3.2.1 Processing raw data . . . 31
3.3 Mapping Reads . . . 31
3.3.1 Expression quantification . . . 32
3.4 Differential expression . . . 32
3.4.1 DEApp . . . 32
3.4.2 R/Bioconductor package . . . 33
3.4.3 Comparative Analysis . . . 35
3.5 Enrichment Analysis . . . 37

**4 Results and discussion** **38**
4.1 RNA extraction and NGS sequencing . . . 38
4.2 Processing raw data – FastQC and Trimmomatic . . . 38
4.3 Mapping reads – HISAT2 . . . 43
4.4 Differential Expression - DEApp and R code . . . 45
4.4.1 Data pre-processing . . . 46
4.4.2 Comparative analysis . . . 50
4.4.3 Enrichment Analysis - g: Profiler . . . 58

**5 Conclusions and Future Work** **73**

**Bibliography** **75**

**A Supplementary material for chapter 4** **105**
A.0.1 RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of the two experiments. . . . 105
A.0.2 FASTQC and Trimmomatic . . . 106
A.0.3 AGAT and HISAT2 . . . 107
A.0.4 HISAT2 . . . 108
A.0.5 Stringtie . . . 109

**B Details of results** **110**
B.0.1 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 20 °C of the chronic experiments, using g: Profiler - Table 12. . . . 110

B.0.2 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 25 °C of the chronic experiments, using g: Profiler - Table 13. . . . . 110
B.0.3 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 30 °C of the chronic experiments, using g: Profiler - Table 14. . . . . 110
B.0.4 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of North vs South of the chronic experiments, using g: Profiler - Table 15. . . . . 110
B.0.5 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of North vs South of the acute experiments, using g: Profiler - Table 16. . . . . 110

# Acronyms

| | |
|---|---|
| **ABI** | AB SOLiD |
| **ATP** | Adenosine triphosphate |
| **BAM** | Binary Alignment/Map |
| **BH** | Benjamini-Hochberg |
| **BLASTP** | Protein BLAST |
| **BWA** | Burrows-Wheeler Aligners |
| **CCAT** | CCAT-seq |
| **ChIP-Seq** | Chromatin Immunoprecipitation followed by sequencing |
| **CIIMAR** | Interdisciplinary Centre of Marine and Environmental Research |
| **CPU** | Central Processing Unit |
| **CPM** | Count Per Million |
| **CTH** | Cystathionine gamma-lyase |
| **CTM** | Critical thermal maxima methodology |
| **DGE** | Differential Gene Expression |
| **DNA** | Deoxyribonucleic acid |
| **DNA-Seq** | DNA Sequencing |
| **FDR** | False Discovery Rates |
| **FPKM** | Transcript per million mapped |
| **GC-content** | Guanine-cytosine content |
| **GCNT3** | Glucosaminyl (N-acetyl) transferase 3 |
| **GO** | Gene Ontology |
| **GTF** | General Feature Format |
| **H2Se** | Hydrogen selenide |
| **HGP** | Human Genome Project |

| | |
|---|---|
| **HT-NGS** | High-throughput Next Generation Sequencing |
| **HSP** | Heat shock proteins |
| **IUCN** | International Union for Conservation of Nature |
| **KEGG** | Kyoto Encyclopaedia of Genes and Genomes |
| **lncRNA** | Long non-coding RNAs |
| **logFC** | Log Fold Change |
| **MPS** | Massively Parallel Sequencing |
| **mRNA** | Messenger RNA |
| **ncRNA** | Non-coding RNA |
| **NGS** | Next Generation Sequencing |
| **PacBio** | Pacific Biosciences |
| **PA** | Poly-A-tailed mRNA selection |
| **PCR** | Polymerase chain reaction |
| **PE** | Paired-end |
| **PES1** | Pescadillo ribosomal biogenesis factor 1 |
| **qPCR** | Quantitative PCR |
| **RD** | rRNA depletion |
| **RNA** | Ribonucleic acid |
| **RNA-Seq** | RNA Sequencing |
| **RPKM** | Transcript per million mapped |
| **RT-PCR** | Reverse Transcription PCR |
| **SAM** | Sequence Alignment/Map |
| **SE** | Single-end |
| **SBS** | Sequencing By Synthesis |
| **SNP** | Single Nucleotide Polymorphisms |
| **SNPs** | Single-nucleotide polymorphisms |
| **ssDNA** | Single-stranded DNA |
| **TMM** | Trimmed Mean of M value |
| **tRNA** | Transfer RNA |
| **WGS** | Whole Genome Sequencing |

# List of Figures


1 Timeline with a brief history of DNA/RNA sequencing with an emphasis on transcriptomics. 6

2 Framework about NGS Library Construction. Adapted from Head et al. (2014). . . . . 9

3 RNA-Seq sample preparation workflow. Adapted from Pease et al. (2010). . . . . 12

4 Mapping of a paired-end read with different reference files: genome and transcriptome. On the left, it represents an alignment of a paired-end read to the genome. The read needs to be mapped across introns (spliced alignment). On the right, an alignment of a paired-end read to the transcriptome. Adapted from Deshpande et al. (2023). . . . . 13

5 RNA-Seq workflow. Adapted from Han et al.(2015), EMBL et al.(2022). . . . . 15

6 *Unio delphinus* in the Sabor River (Portugal). Photo credit: Ronaldo Sousa. . . . . 20

7 Scheme of freshwater mussel's life cycle. *Adapted from Modesto et al. (2018).* . . . . . 21

8 Effects of climate change in freshwater ecosystems. Adapted from Bellard et al. (2012). 23

9 *Unio delphinus* used in the acute experiments. . . . . 29

10 *Unio delphinus* used in chronic experiments. . . . . 29

11 Bioinformatic methodological overview: light blue boxes represent the input/output files, dark blue boxes represent tasks and tools to perform the analysis and green boxes represent the sources files. . . . . 30

12 FastQC quality scores across all bases reports for the sample of North population under a temperature condition of 20°C before pre-processing. Figure 12 A and B represent the forward and the reverse fragments, respectively. . . . . 39

13 FastQC quality scores across all bases reports for the sample of North population under a temperature condition of 20°C after pre-processing. Figure 13 A and B represent the forward and the reverse fragments, respectively. The absence of whiskers in the processed data's graph indicates that the quality ratings of the processed data are greater, as shown in the graphs. . . . . 40

14 FastQC adapter content reports for the sample of the North population under a temperature condition of 20°C before pre-processing. Figure 14 A and B represent the forward and the reverse fragments, respectively. . . . . 41

15 FastQC adapter content reports for the sample of the North population under a temperature condition of 20°C after pre-processing. Figure 15 A and B represent the forward and the reverse fragments, respectively. The adapters graph of the processed data no longer shows the red line, indicating that the adapters seen in the raw data have likewise been eliminated. . . . . 42

16 Boxplot (Figure 16. A) and MDS plot (Figure 16. B) of normalized samples distribution of chronic data, with N20G4 sample. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery. Both representations reveal the presence of the 'N20G4' sample as an outlier within the North population at a temperature of 20 °C. Outliers can distort statistical measurements and may provide incorrect results. It was deemed necessary to exclude the N20G4 sample from the subsequent analysis. . . . . 47

17 Boxplot (Figure 17. A) and MDS plot (Figure 17. B) of normalized samples distribution of chronic data, without N20G4 sample. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery. . . . . 48

18 Boxplot (Figure 18. A) and MDS plot (Figure 18. B) of normalized samples distribution of acute data. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery. . . . . 49

19 Estimated dispersion plots for chronic experiments, using three distinct methods: EdgeR, limma, DESeq2. . . . . 51

20 Veen diagram of the comparisons North 20 °C ∩ 25 °C ∩ 30 °C and South 20 °C ∩ 25 °C ∩ 30 °C. . . . . 53
21 Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 20 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes. . . . . 54
22 Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 25 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes. . . . . 54
23 Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 30 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes. . . . . 55
24 Volcano plot and heatmap for the comparative analysis of North vs South chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes. . . . . 55
25 Estimated dispersion plots for acute experiments, using three distinct methods: EdgeR, limma, DESeq2. . . . . 56
26 Volcano plot and heatmap for acute experiments for the comparative analysis: North vs South. Along with the volcano plot is represented the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. Each data point on the plot corresponds to a unique gene. The blue points represent the down-regulated genes and the red ones the up-regulated genes. . . . . 57
27 Overview of g: Profiler results in chronic experiments at three temperatures (20 °C, 25 °C, and 30 °C), highlighting statistically significant enrichments, and temperature-specific patterns. . . . . 58
28 Overview of g: Profiler results in chronic experiments in the comparative analysis of North vs South, highlighting statistically significant enrichments, and temperature-specific patterns. . . . . 59

29 Overview of g: Profiler results in acute experiments, highlighting statistically significant enrichments, temperature-specific patterns. . . . . 69

30 RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of acute experiments. The blue table represents the northern population and the orange table the southern population. . . . . 105

31 RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of chronic experiments. The blue table represents the northern population and the orange table the southern population. . . . . 106

# List of Tables


1 Software packages mostly used to identify gene expression. . . . . 19

2 Comparative analysis conducted in chronic and acute experiments using both DEApp and R/Bioconductor package. The comparisons highlighted in blue were performed exclusively using the R/Bioconductor packages. . . . . 36

3 Alignment rate for each read sample of chronic experiments against the *Unio delphinus* reference genome. Samples were named according to two variables: population and temperature. North population samples were accessed with "N" and south populations with "S", and next to this first letter with the temperature. In the case of chronic experiments, it can be 20, 25, or 30. Following the temperature are the replicates, which are five per condition, and are represented by "G1", "G2", "G3", "G4" and "G5". . . . . 44

4 Alignment rates for each read sample of acute experiment against the *Unio delphinus* reference genome. Samples were named according to two variables: population and temperature. North population samples were accessed with "N" and south populations with "S", and next to this first letter with the temperature. In the case of chronic experiments, can be 20, 25, or 30. Following the temperature are the replicates, which are five per condition, and are represented by "G1", "G2", "G3", "G4" and "G5". . . . . 45

5 Number of DGEs using the software DEApp and R/Bioconductor code in all the comparative analyses developed using the three methods, EdgeR, Limma, and DESeq2, and number of DGEs resulting from the overlap of the three methods in the analysis that showed consistency in the results. . . . . 52

6 Number of DGEs using the software DEApp and R/Bioconductor code in the comparative analysis developed using the three methods, EdgeR, Limma, and DESeq2, and number of DGEs resulting from the overlap of the three methods in the analysis that showed consistency in the results. . . . . 57

7 Gene Ontology (GO) Molecular Function (GO: MF), the database of manually annotated protein complexes (CORUM), Cellular Component (GO: CC), Human Phenotype Ontology (HP) and Reactome pathways (REAC) enrichment analysis of the comparative output analysis of North vs South at the temperature of 20 °C in chronic experiments, using g: Profiler. . . . . 60
8 Gene Ontology (GO) Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), and the database of manually annotated protein complexes (CORUM) enrichment analysis of the comparative output analysis of North vs South at the temperature of 25 °C in chronic experiments, using g: Profiler. . . . . 62
9 Gene Ontology (GO) Molecular Function (MF), Biological Process (BP), Cellular Components (CC), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REAC) enrichment analysis of the comparative output analysis of North vs South at the temperature of 30 °C in chronic experiments, using g: Profiler. . . . . 66
10 Gene Ontology (GO) Molecular Function (MF), Biological Process (BP), Cellular Components (CC), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REAC) enrichment analysis of the comparative output analysis of North vs South in chronic experiments, using g: Profiler. . . . . 71
11 Gene Ontology (GO) Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), Reactome (REAC), and Protein databases (CORUM) enrichment analysis of the comparative output analysis of North vs South in acute experiments, using g: Profiler. . . . . 72

12 List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 20 °C. . . . . 110
13 List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 25 °C. . . . . 111
14 List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 30 °C. . . . . 112
15 List of go terms and their associate gene names and descriptions of the comparative analysis of North vs South of chronic experiments. . . . . 115
16 List of go terms and their associate gene names and descriptions of the comparative analysis of acute experiments. . . . . 117

# Chapter 1

# Introduction

## 1.1 Context/motivation

Climate change is an escalating global problem, affecting all living organisms across ecosystems (1). Freshwater mussels are no exception and are increasingly vulnerable to environmental stressors, including climate change. These organisms, as vital components of rivers, streams, and lakes ecosystems, play essential ecological roles such as nutrient cycling and increasing water quality (2; 3). Moreover, they serve as essential ecosystem engineers, providing habitat to other organisms and supporting intricate food webs (4). Freshwater mussels also serve as critical environmental change indicators, offering insights into historical and current conditions by, for example, analyzing their shells, and forecasting impending water quality shifts (5).

Bivalves, and in particular freshwater mussels, are still poorly studied in terms of genomics, with most studies still focusing on species with considerable commercial value (6). This lack of information could be worrisome given the ecological roles of bivalves and also because predictions are crucial for alerting scientists and researchers about potential hazards in the future, providing a way to support the link between biological changes at different levels (from individuals to ecosystems) and climate change (7). Bioinformatics tools and other molecular approaches are widely used to study the genetic variation among organisms. Given the crucial ecological importance and poor conservation status of many freshwater mussel species, this taxonomic group cannot be disregarded considering this research topic. In this context, RNA-Sequencing (RNA-Seq) technology emerges as an invaluable tool for investigating the intricate molecular responses of freshwater mussels to thermal stress, shedding light on their adaptive mechanisms under climate change scenarios. This involves identifying which genes are up-regulated or down-regulated in response to fluctuating temperatures. Ultimately, the insights gained from this research will enhance the understanding of how climate change impacts freshwater mussels and will provide valuable knowledge for implementing effective strategies to conserve and manage these organisms in a rapidly changing

environment.

## 1.2 Objectives

The main goal of this master's thesis was to determine the gene expression patterns on freshwater mussels, using the species of *U. delphinus* Spengler, 1793 as a model organism, under the effects of climate change. A comprehensive bioinformatic pipeline focused on transcriptomics analysis was developed to generate the differential gene expression profiles of these individuals under thermal stress. The study focuses on the application of diverse bioinformatics tools and methodologies.

To achieve this goal the following specific objectives were set up:

- Conduct a review of the current state of knowledge and relevant concepts, laying the foundation for subsequent stages of the study.
- Perform data pre-processing generated by sequencing, which involves quality control and filtering the raw sequencing data.
- Execute read mapping to the reference genome, enabling the accurate alignment of sequencing reads.
- Investigate differential gene expression, aiming to identify genes that exhibit significant changes in expression under thermal stress conditions.
- Undertake a comparative analysis of gene expression profiles, comparing different experimental conditions to gain insights into how *U. delphinus* responds to climate-induced thermal stress.
- Conduct an enrichment analysis to elucidate the biological processes and pathways that are significantly impacted by temperature-induced stress in freshwater mussels.

## 1.3 Thesis Structure

**Chapter 2. State of the Art** This chapter revised the background information and literature.

**Chapter 3. Methodologies** This chapter described the data processing methods used in each task of the study.

**Chapter 4. Results and Discussion** This chapter described key findings obtained through the CIIMAR cluster, and R/Bioconductor package, discussing their main implications.

**Chapter 5. Conclusion** The final chapter summarizes the main findings and concludes the research.

# Chapter 2

# State of the Art

## 2.1 Transcriptomics

Genomic approaches have been recognized as fundamental tools to study biodiversity (8; 9). Genomics tools, such as, transcriptomics are key resources to understand individuals' evolutionary and adaptive characteristics (10). Transcriptomics represents one of the most developed areas in the post-genomic age (11) and refers to the study of the whole transcriptome (12). The transcriptome, which encompasses protein-coding messenger RNA (mRNA) and non-coding RNA [ncRNA: ribosomal RNA (rRNA), transfer RNA (tRNA), and other ncRNAs], represents the entire set of ribonucleic acid (RNA) transcripts in a particular cell type or tissue at a certain developmental stage and/or under a specific physiological environment (13; 14). Consequently, transcriptomic analyses are an effective method to link the relationship between genotype and phenotype (15). The application of transcriptomic approaches allows not only to quantify changes in expression level for each gene, but also to map, annotate, and determine the functional structure of each gene in the genome (16; 17). However, the complexity of studying the transcriptome has led to the need for the development and advancement of new and more effective high-throughput methods. Over the last decades, several technologies have been developed to study transcriptomes.

### 2.1.1 A brief history

Over the last 60 years, sequencing technologies have advanced significantly. It began with the advent of first-generation sequencing, then led to next-generation sequencing (NGS), and advanced further into the domain of third-generation sequencing technologies (Figure 1) (18). The discovery of deoxyribonucleic acid's (DNA) double helix structure, in 1953 by Watson and Crick (19), served as the foundation for a new field of study that focuses on the molecular biology of the cell (20; 21). Building upon this foundational knowledge, in 1975, Sanger sequencing, a low-throughput method, was released (22). The first-generation

sequencing technology, also referred to as Sanger sequencing or dideoxy chain-termination sequencing, has been used for single-stranded DNA sequencing (DNA-Seq) (23) and it has been employed in diverse fields with an accuracy of around 99.99% (24). Sanger sequencing remains useful, due to its practical applications and has become a 'gold standard' method (25; 26), encouraging scientists to study genomic code in living organisms (27). Although other sequencing methods emerged around the same time (e.g. Maxam and Gilbert in 1977 (28)), none were able to exceed Sanger sequencing. In the same year, Sanger was capable of sequencing the first genome, bacteriophage $\phi$X 174, which is 5375 bases in length (29). In subsequent years, Frederick Sanger's research group conducted the completion of various genome sequences (30). The introduction of the complete Human Genome Project (HGP) was a big step in the field of genomics (31). It was reported in 1990, and for over a decade, it was developed at great cost and with intense human effort (29). The goal of HGP was to map and sequence the entire human genome (32). Sanger sequencing played a significant role in this project, leading to the identification of thousands of genes (33). As the HGP proceeded, it became clear that to improve genetic analysis and have quick access to tools to evaluate gene expression data (34), there was a need for faster, greater throughput and cheaper technologies (35). This led to the development of DNA microarrays in the late 1990s (36). One of the major applications of DNA microarrays is to measure gene expression levels (35). DNA microarrays use thousands of nucleic acid sequences attached to a surface to assess the relative concentration of nucleic acid sequences in a mixture by hybridization and subsequent detection of the hybridization events (37). This technology allowed researchers to simultaneously analyze the expression of thousands of genes in a single experiment (38). The completion of the first human genome was only the beginning of the modern DNA sequencing era, which resulted in further invention and improved development towards new advanced high-throughput DNA sequencing strategies known as "high-throughput next-generation sequencing" (HT-NGS or simply NGS) (39).

In the 2000s, NGS technologies marked a revolutionary advancement in genomics and transcriptomics (40). NGS provided a high-throughput, cost-effective way to sequence entire genomes and transcriptomes (41). RNA Sequencing was one of the most significantly advanced approaches led by NGS. Therefore, it opened up new approaches for studying gene expression, alternative splicing, and non-coding RNAs with unprecedented detail. Nowadays, this technology is widely used as researchers can comprehensively analyze the transcriptome. While both RNA-Seq and microarrays have their merits and may be chosen based on specific research goals, the flexibility, sensitivity, and ability to uncover previously unknown aspects of the transcriptome have positioned RNA-Seq as the preferred choice for many transcriptomics studies in the 21st century (16; 42). This technology has played an important role in advancing the

understanding of gene expression, regulation, and functional genomics (43).

The first- and second-generation sequencing technologies had a significant contribution to scientific advances, including a better understanding of whole genome sequences, a more complete characterization of the transcriptome, and a better understanding of interactions between proteins and DNA (44). The third-generation sequencing emerged in the early to mid-2010s and represents a newer generation of sequencing methods that began to gain importance (45). These technologies introduced the concept of single-molecule sequencing, which allowed individual DNA or RNA molecules to be sequenced directly, resulting in extremely large read lengths ranging from hundreds to tens of thousands of bases (46).

Transcriptomics has an extensive history, built together by innovative scientific advancements. Each significant achievement marks a crucial turning point in the comprehension of the transcriptome, from the discovery of the double helix structure of DNA to the enormous accomplishment of sequencing the whole human genome. The development of RNA-Seq technology revolutionized the capacity to explore RNA molecules with unprecedented depth and accuracy. Transcriptomics was further pushed to its limits with the advent of third-generation sequencing technology, which allowed for real-time, single-molecule analysis.

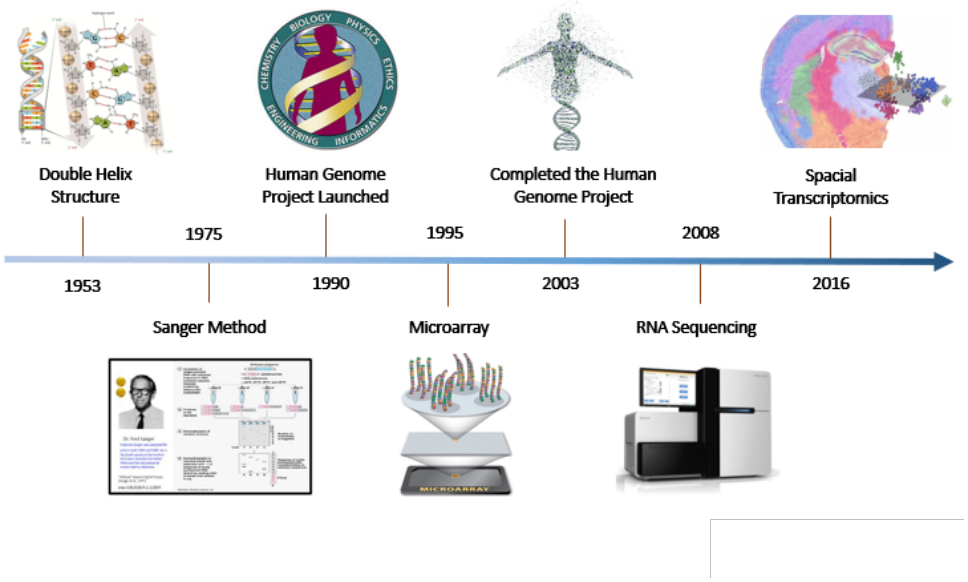



Figure 1: Timeline with a brief history of DNA/RNA sequencing with an emphasis on transcriptomics.

### 2.1.2 Next Generation Sequencing

Next-generation sequencing, an innovative DNA/RNA sequencing approach, has transformed molecular biology and genomics (47). Remarkably, it is now possible to sequence an entire human genome in less than a day using NGS, an achievement that took nearly a decade with the prior Sanger sequencing tech-

nology (48). This dramatic reduction in sequencing time has been driven by the increasing demand for cost-effective and rapid sequencing techniques since the completion of the first human genome sequence (49). The true revolution in genomic research was initiated in 2004 with the advent of massively parallel sequencing platforms (MPS), the precursor to NGS (16). The designation of NGS is used to distinguish this innovative approach from its predecessors in DNA sequencing technologies (50). Next-generation sequencing technologies have enabled complete genome sequencing and facilitated advancements in transcriptomics, epigenomics, metagenomics, and other omics investigations (51). Moreover, in a single NGS sequencing run, millions of small DNA fragments are generated (52). These fragments vary in read lengths, typically ranging from 25 to 450 base pairs depending on the specific NGS technology (53). This high-throughput capability has transformed the landscape of genomic research by providing researchers with unprecedented data in a single sequencing experiment (54). Companies like Roche (454 GS FLX), Illumina (Genome Analyzer II), and ABI (AB SOLiD) introduced MPS technologies, revolutionizing gene expression quantification and paving the way for diverse experimental protocols (55; 56; 57). Over the past two decades, Illumina has emerged as the most widely used and successful NGS platform (58). The most commonly used technologies for NGS are Illumina and Ion Torrent when referring to short-read sequencing procedures, whereas Pacific Biosciences (PacBio) and Oxford Nanopore offer long-read sequencing approaches (59). Today, there is a plethora of options for high-throughput functional transcriptomics (59). Researchers can choose a sequencing platform with distinct technologies to match their specific needs, resulting in varying sequencing capabilities and read lengths. Regardless of the chosen sequencing technology, the NGS workflow remains defined by three fundamental steps: sample preparation, library preparation, and sequencing/analysis (60).

Before performing the NGS workflow, some parameters are important to consider (61):

1. Single-end (SE) and paired-end (PE) sequencing: SE sequencing produces reads from only one end of complementary DNA (cDNA) fragments and is often faster and less expensive, while PE sequencing produces reads from both ends of cDNA fragments and gives greater depth and quality of sequencing (62). Increased coverage gives more information on splicing events, single-nucleotide polymorphisms (SNPs), and gene fusions, as well as enhances *de novo* assembly and accurate read mapping to the reference assembly (63).

2. Sequencing read length: The read length is the number of cycles (base pairs) conducted in the sequencing device at one time. Longer readings are often more accurate and better, especially from PE runs (64). While longer reads are more expensive than shorter reads, they are frequently favoured for better genome or transcriptome assemblies (65).

3. Depth or coverage: The Lander/Waterman equation may be used to calculate the coverage of an assembly; coverage = number of readings read length/assembly length (66). Because various transcripts are expressed at varying amounts, determining the coverage for RNA sequencing is difficult. The depth is usually determined by the experimental design, gene expression level, and size of the reference genome/transcriptome (67).

4. Sequence Variance: Variability can occur at the technological and biological levels (68). As a result, tests must be designed to control this variance. To minimize biological variation, samples must be used for biological replication at the same developmental stage and cultivated under similar settings. Integrating replicated samples from the same species can effectively reduce the occurrence of erroneous predictions resulting from sequencing errors (69). So, it is recommended at least two biological duplicates. An independent approach, such as quantitative polymerase chain reaction (qPCR), can be used to confirm significant differences (70).

**NGS Workflow**

A pre-requisite for all NGS technologies is a good-quality library, so library preparation (represented in Figure 2) is a crucial first step in any NGS workflow (71). Next Generation Sequencing can be classified as DNA-Seq and RNA-Seq (59). Two fundamental processes are involved in creating a sequencing library from RNA or DNA samples: target sequences will be produced by amplification and attachment of sequencing adapters that will subsequently communicate with the NGS platform (72). If RNA serves as the initial template, a further step is required in which the RNA is first reverse transcripted to create cDNA (RT-PCR) (73). Polymerase chain reaction (PCR) amplification produces a library of specially sized DNA fragments that are compatible with the sequencing system to be used (74). Primers for library creation are created depending on the sequences of interest, which might range from the entire genome to specific RNA transcripts (75). The adaptor ligation stage effectively pairs amplified DNA or cDNA fragments, known as amplicons, with particular oligonucleotide sequences that will interface with the surface of a sequencing flow cell (72). If many samples are to be sequenced in a single run, a unique identification, or barcode, is ligated to the amplicon as well (76).

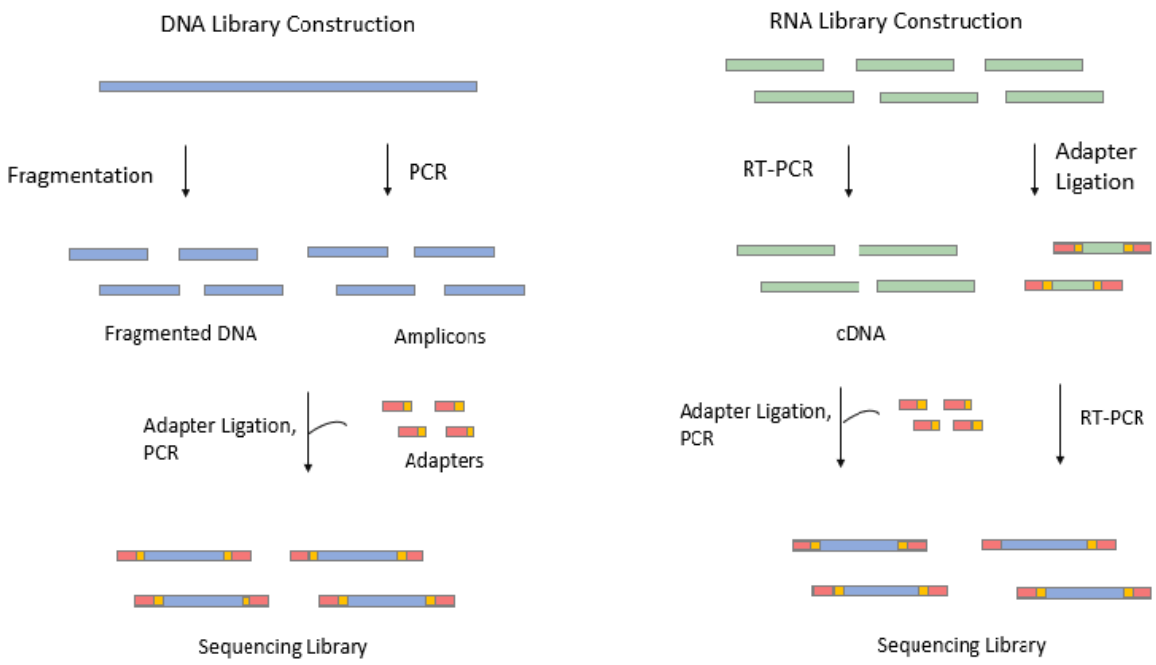



Figure 2: Framework about NGS Library Construction. Adapted from Head et al. (2014).

The next step is sequencing, also referred to as NGS sequencing. As described above, several platforms exist to perform this step. It is possible to divide the technologies in terms of read length, normally associated with short-read sequencing of the platforms: Illumina and Ion Torrent, and long-read sequencing of the platforms: PacBio and Oxford Nanopore (59). Here, is described the workflow as the most commonly used technology, i.e. Illumina. In the NGS workflow's sequencing stage, libraries from multiple samples may be loaded and sequenced simultaneously (77). Once the library molecules holding the desired genetic information are immobilized by surface-tethered probes that are specially tailored to match their adaptor sequences, a strong amplification process known as bridge PCR occurs. This process turns a single DNA molecule into millions of identical copies of single-stranded DNA (ssDNA), improving the quantity of genetic material accessible for subsequent study tenfold (78). Illumina's NGS technology is established on sequencing by synthesis (SBS) with a fluorescent-labelled reversible terminator technology (47). Chemically altered nucleotides connect to the DNA template strand by natural complementarity in the SBS procedure (79). Each nucleotide has a fluorescent tag and a terminator that may be reversed to prevent the inclusion of the next base (79). The fluorescent signal identifies the nucleotide that has been introduced, and the terminator is split open to allow the binding of the following base (80). The reads are then removed after reading the forward DNA strand, and the procedure is then repeated for the reverse

strand. PE sequencing is the name of this technique (47).

The last step of NGS's workflow is data analysis. Given the vast amounts of generated data, simplified bioinformatics data analysis is critical to apply these technologies (81). The data analysis pipeline is often separated into three stages. Following sequencing, the software performs primary analysis, which includes base-calling for each clonal amplified DNA fragment (82). This step also includes quality control techniques such as read filtering. The sequence information is saved in FASTQ format together with the quality ratings. PE sequencing provides two FASTQ files linked by sequence IDs for the Illumina platform (83). On the Illumina technology, sequencing run quality is measured using three indicators: cluster density, percentage of clusters clearing filters, and percentage of base call with a quality score of at least Q30 (1 in 1000 chance of an inaccurate base call) (84). The second part of the analysis includes read alignment. Short readings are stored in FASTQ files, either SE or PE and are first aligned to a reference genome depending on the availability. Binary alignment/map (BAM) format is used to store the alignment results (85). Coverage is another quality metric to consider at this stage, which comprises a depth of coverage (the number of times a base is sequenced) and breadth of coverage (the percentage of a reference genome covered) (86). Coverage guarantees that sensitivity and specificity are maintained (87). Finally, NGS analysis involves variant annotation and biological insights.

**Applications of NGS**

The adaptability of NGS platforms' has expanded the scope of genomics and trascriptomics research, allowing for investigations on uncommon genetic illnesses, cancer genomics, microbiome analyses, infectious diseases, and population genetics. Furthermore, NGS has enabled the development of tailored medicines, precision medicine techniques, and enhanced diagnostic procedures (40). The introduction of NGS technology into biological research has turned biological science into a big data science (88) in which computations complement short reads and high error rates of experimental research with high-depth coverage sequencing data. Many biologists and computer scientists are emphasizing the critical need for computing power, storage, and various bioinformatics software for analyzing large amounts of sequence data that this technology produces (89). The fast development of NGS technologies has provided benefits in assessing the functional complexity of the entire transcriptome (90). As NGS technology becomes more affordable, RNA-Seq is being employed more often. With RNA-Seq, it is possible to determine a species' transcription profile and analyze changes that occur in response to environmental factors (91). RNA-Sequencing is an emerging NGS platform-based technique for identifying and quantifying transcripts that is independent of any identified sequence characteristic (92). Moreover, transcriptomics allows the

assessment of gene expression and signalling pathways in specific tissues or cells. Gene expression profiles correspond to the measurement of the activity of thousands of genes at once and thus represent an emerging approach to understanding the responses of individuals exposed to thermal stress (93).

## 2.2 RNA Sequencing

RNA Sequencing, also named whole-transcriptome shotgun sequencing, refers to the use of high-throughput sequencing technologies for characterizing the RNA content (94). In molecular biology, RNA-Seq represents an effective method for examining and quantifying gene expression at the transcriptome level (95). This high-throughput technology has transformed the knowledge of gene regulation, the identification of new transcripts, and the finding of biomarkers by offering a comprehensive perspective of gene expression (96). It is a vital tool for expanding our understanding of gene expression in many biological systems because of its adaptability and capacity to identify both known and new transcripts (97). RNA-Seq can be used to analyze genomes from species that lack entire genomic sequence information, but subsequent data processing in these circumstances is sometimes impeded by a lack of gene annotation or a sequenced genome (98). In contrast, the microarray technique requires prior knowledge of the sequences to be discovered (99). Finally, RNA-Seq provides a huge dynamic range of quantification levels, allowing for the measurement of extremely low and high amounts of gene expression, which is essential for capturing the full spectrum of transcriptional activity within a biological sample.(100). Recent developments in RNA-Seq have made it possible to quickly characterize and quantify transcriptomes (101). RNA-Seq allows the identification of novel transcripts, small RNAs, SNPs, alternate splicing products, sense and antisense transcripts, and fusion transcripts, (102). RNA Sequencing can be used for a wide range of applications and it is highly adjustable through the use of different pipelines. Depending on the design experiments scientists can adapt different strategies to the organism under study (93).

### 2.2.1 RNA-Seq workflow

Han et al., (2015) described a typical RNA-Seq workflow and divided it into three main parts: Experimental Biology, Computational Biology, and Systems Biology (103).

1. Experimental Biology

   After the RNA extraction, the following step is the library construction and sequencing (Figure 3). The number of sequenced reads for a sample is referred to as sequencing depth or library size

(100). The generation of a sequencing library for RNA-Seq analysis is a complicated and multi-step procedure that can result in substantial variance (104). Ribosomal RNA is very abundant and must be removed from total RNA before sequencing to enable effective transcript/gene identification. Standard methods including poly-A-tailed mRNA selection (PA) or rRNA depletion (RD) are commonly used to solve this issue (Figure 3) (105). Mitochondrial RNA selection is useful for analyzing gene expression at the mRNA level because it selects polyadenylated mRNA molecules (106). Ribosomal RNA depletion eliminates rRNA molecules and is used to enrich a wider spectrum of RNA species other than mRNA and reduce sequencing costs (107). Next-generation sequencing of cDNA enables qualitative and quantitative transcriptome studies. It may also be used to identify new genes more precisely, characterize alternative splicing or fusion transcripts, and detect RNA edits and mutations (13).

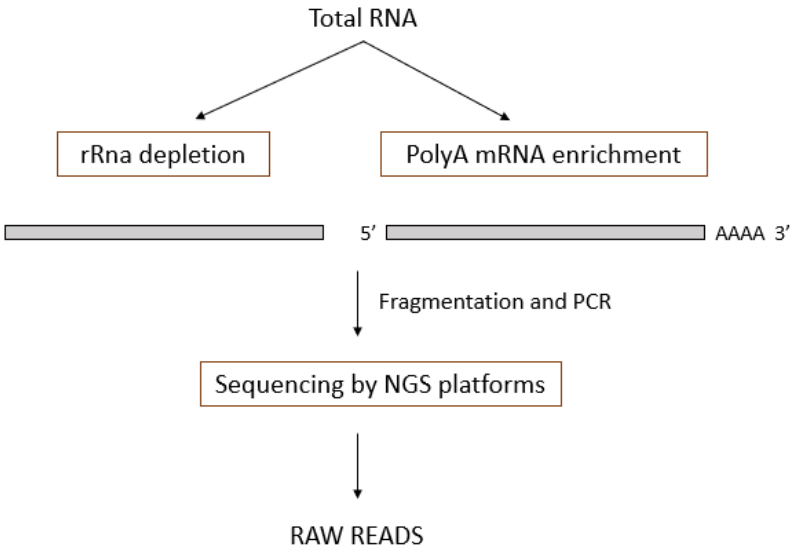



Figure 3: RNA-Seq sample preparation workflow. Adapted from Pease et al. (2010).

2. Computational Biology

- Quality control of sequence reads

As previously stated, high-throughput sequencers may create massive volumes of data in a single run. Before studying the collected sequences, it is necessary to assess library quality and sequencing performance. As a result, the low-quality bases must be eliminated prior to any alignment or assembly of the reads. Base calling on any of the regularly used NGS platforms might be

hampered by low-quality starting material, chemical mistakes, or instrument flaws. Quality Control (QC) typically considers duplication rate, as most sequences are expected to occur only once when the universe is larger than the sample size (108); rRNA abundance, which should be low; strand specificity, defined as the number of reads mapping to known transcribed regions at the expected strand; coverage continuity at annotated transcripts, and performance at 5' and 3' ends, defined as agreement with known end annotation (109).

- Mapping reads

After filtering the reads from the raw cDNA sequence reads, the short-sequenced reads must be mapped to a reference genome (genome mapping) or transcriptome (transcriptome mapping), depending on the availability (110). As a result, the primary purpose of this step is to find the locations where each read occurs on a specific reference (111). Mapping the reads to the reference data can be difficult since millions of small reads must be mapped to genomes that are often quite big. This means that mapping methods must be extremely efficient in terms of Central Processing Unit (CPU) and memory use (112). Furthermore, because repeated sequences account for more than 50% of the genome in complex species (113), mapping techniques must be capable of handling numerous mapping sites. Read alignment against the genome is slower than against the transcriptome because it must account for all non-coding locations (Figure 4) (114). Various alignment programs that can do spliced alignments, including TopHat2 (115), SOAPSplice (116), and Blat (117). Aligners such as Bowtie 2 (118), BWA (119), and others, on the other hand, specialize in the alignment of short reads to a reference.

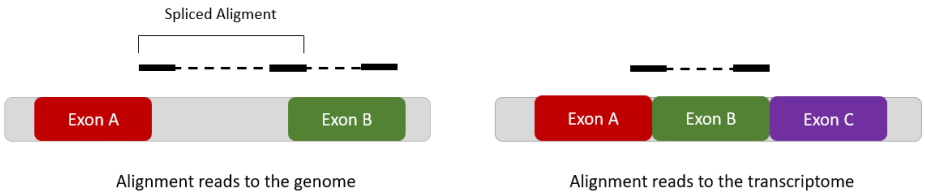



Figure 4: Mapping of a paired-end read with different reference files: genome and transcriptome. On the left, it represents an alignment of a paired-end read to the genome. The read needs to be mapped across introns (spliced alignment). On the right, an alignment of a paired-end read to the transcriptome. Adapted from Deshpande et al. (2023).

Researchers often use a mapping technique to analyze transcriptome data for species with known reference genomes. However, this technique is ineffective when the reference sequence is unavailable or incomplete (120). Thus, a *de novo* assembly is required to give a feasible approach for transcriptome analysis in species with unsequenced genomes (121). It is required when a high-quality reference genome is unavailable, such as for many non-model species, when analyzing complex microbial communities, in metatranscriptome investigations, or when examining unculturable microorganisms (122). The use of *de novo* assembly of short sequence reads into transcripts, allows researchers to rebuild the sequences of the whole transcriptome, identify and catalog all expressed genes, differentiate isoforms, and capture transcript expression levels (123).

The output file of the referenced mappers is in Sequence Alignment/Map (SAM) format and provides information about both aligned and non-aligned reads (124). This file provides information on the genomic portion where the read was mapped and its respective score for aligned reads (125).

- Expression quantification and normalization

Following the mapping of RNA-Seq reads, the data must be transformed into a quantitative measure of gene expression (126). For reads that were mapped to the transcriptome, the easiest solution is to add the number of readings that fall within the coordinates of each element (127). On the other hand, for reads that were aligned to the genome, the measurements of gene expression must be performed by HTSeq package (128). It also requires normalization for each sequencing run to eliminate the potential genes that seem to be differentially expressed. The reads per kilobase method is one method for normalizing RNA-Seq data (129). The transcript per million mapped (RPKM) measure normalizes a transcript's read by both its length and its size as well as the overall amount of mapped reads in the sample (100). Likewise, the fragments per kilobase of the transcripts per million mapped (FPKM) statistic is used to normalize paired-end data and it obtains estimates of meaningful expressions (130).

3. Systems Biology

Once the reads have been aligned or assembled, researchers face a vast amount of data generated through high-throughput techniques. This is where differential gene expression (DGE) analysis can be applied. Differential gene expression analysis is a computational strategy that aims to measure and discover genes whose expression levels change significantly between two or more experimental conditions (131). This study assists researchers in identifying genes that are upregulated or down-

regulated under certain conditions, offering vital insights into biological processes, disease causes, or therapeutic responses (132).

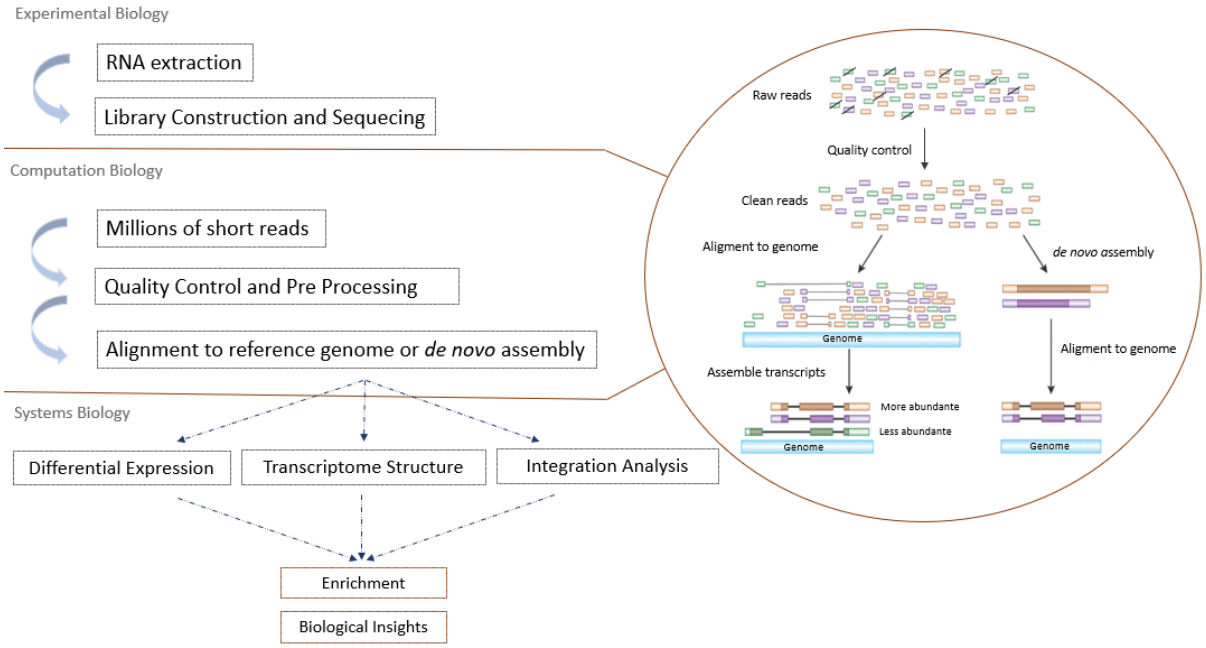



Figure 5: RNA-Seq workflow. Adapted from Han et al.(2015), EMBL et al.(2022).

### 2.2.2 Applications of RNA-Seq data

Transcriptomics has applications in a wide range of scientific fields, from fundamental research to medicine and biotechnology (133). As the knowledge of transcriptomics grows, priceless new information on the complex molecular processes emerges.

- Diagnostics and disease profiling

  In the past years, transcriptomics has gradually emerged as a useful tool for rare disease diagnosis and discovery (134). Transcriptional start sites have been extensively found using RNA-Seq methods (135). Gene fusions, allele-specific expression, and disease-related Single Nucleotide Polymorphisms (SNP) can be detected using RNA-Seq (136). Since these regulatory components have a key role in human illness, the identification of these polymorphisms is crucial for the interpretation of disease association studies. In the context of cancer, it permits the detection and categorization of distinct cancer types based on gene expression patterns (137; 138). In addition to cancer, transcriptomics is useful in disease subtyping. Diabetes and Alzheimer's are frequently heterogeneous diseases with several molecular subtypes (139). Transcriptome analysis has been widely

employed to analyze tumour heterogeneity, categorizing tumours into molecular subgroups, and developing signatures that predict therapeutic response and patient outcomes (140). Furthermore, transcriptomics is at the forefront of biomarker discovery, constantly revealing new markers that might increase diagnostic accuracy and prognosis prediction (141).

- Responses to environment

  Transcriptomics represents an effective method for studying how organisms respond to environmental changes and stresses (142). High-throughput technologies allow monitoring reactions in minute detail, exposing the genes and pathways that allow adaptability and survival (143). Studying the entire mRNA reflects actively expressed genes in a cell, and consequently provides answers on how to respond to changes in the external environment (144). Significant effort has been directed towards characterizing shifts in mRNA abundance triggered by changes in key environmental variables such as temperature (145), salinity (146), oxygen (147), and pH (148). Transcriptomics has been utilized effectively in the field of environmental physiology to address a wide variety of concerns involving how or if organisms can acclimatize or adapt to the abiotic circumstances associated with living in distinct environments (149). For example, transcriptomics helps climate change studies by exposing how changing gene expression impacts ecosystems and species. This knowledge is critical for conservation efforts and techniques that aid species in adapting to changing environmental conditions (150).

- Human and pathogen transcriptomes

  Transcriptomics is an effective technique for studying host and pathogen mechanisms in infectious illnesses (151). RNA-Seq of human pathogens is a proven method for analyzing gene expression variations, finding novel virulence factors, predicting drug resistance, and exposing host-pathogen immunological interactions (152; 153; 154). Dual RNA-Seq has recently been utilized to characterize the RNA expression in the pathogen and host throughout the infection process (155). Researchers may use these methods to understand how viruses attack host cellular machinery, exposing the strategies pathogens use to escape the immune system and establish infection. This understanding is critical for creating medicines that disrupt these efforts.

- Gene discovery and annotation

  The creation of large-scale transcriptome data, such as RNA-Seq, is the first step towards gene discovery (156). Transcriptomics offers the potential to discover new genes, including previously

undiscovered or poorly characterized genes (157). Transcriptomics is crucial in the process of gene annotation because it provides data on gene expression patterns and isoforms, which can be utilized to infer their functions and regulatory roles (97). Comparative analysis facilitates functional categorization of genes, allowing genes to be classified based on their roles in biological processes and activities (158). Transcriptomic data may also be used to map genes onto established biological pathways, offering information on their participation in certain cellular processes and signalling cascades. Furthermore, gene annotation contributes to a better knowledge of gene regulation (159).

- Non-coding RNA

  The importance of ncRNAs in gene regulation and their functions in numerous biological processes and disorders has been highlighted through transcriptomics (160). MicroRNAs are short ncRNAs that play an important role in post-transcriptional regulation (161). Transcriptomic investigations aid in the identification of particular microRNA target genes, revealing information on their roles in processes like development, immunology, and cancer (162). Long non-coding RNAs (lncRNAs) are another type of non-coding RNA that has been widely explored using transcriptomics (163). They have a role in gene regulation and have been linked to a variety of illnesses (164).

### 2.2.3 Differential gene expression analysis

One of the most common uses of RNA-Seq data is DGE analysis. This technique is used in various RNA-Seq data processing applications, enabling the identification of genes that are differentially expressed between two or more conditions (165). The obtained data from RNA-Seq technology is submitted to statistical and bioinformatics methods for rigorously interpreting the vast amount of data generated through high-throughput techniques to identify differentially expressed genes (166). Different software for DGE analysis apply distinct approaches for normalizing data, and their performance varies depending on the type of experimental data and the number of sample replicates (167). Specific tools have been developed to determine which genes are differentially expressed. Table 1 shows several software for DGE analysis, with the different normalizations used, read count distribution, and differential expression levels.

The most common and efficient used statistical packages are EdgeR, DESeq2, and Limma (168).

**edgeR** (169) uses empirical Bayes estimation and precise testing based on a negative binomial model to determine differential expression. The software was created to analyze trials with a minimal number of duplicates. An empirical Bayes approach, in particular, is utilized to reduce the degree of overdispersion among genes by borrowing information among genes. To measure differential expression for each gene,

an exact test similar to Fisher's exact test but tailored for overdispersed data is utilized. The Trimmed Mean of M-values (TMM) normalization approach is employed by default to account for differences in sequencing depths between samples, whilst the Benjamini-Hochberg procedure is utilized to regulate the false discovery rate (FDR) (170).

**DESeq2** (171) employs a Wald test for significance testing: the shrunken estimate of Log fold change (logFC) is divided by its standard error, yielding a z-statistic that is compared to a conventional normal distribution. The Wald test allows the assessment of individual coefficients or contrasts of coefficients without having to build a simplified model like the likelihood ratio test, but the likelihood ratio test is still an option in DESeq2. The Wald test P values from the subset of genes that pass an independent filtering step are corrected for multiple testing using the Benjamini-Hochberg (BH) approach. DESeq and edgeR perform similarly in differentially expressed, however, DESeq has been proven to have more conservative FDRs, compared to EdgeR (131).

**Limma** (172) is predicated on linear modelling. It was created to analyze microarray data, but it has also recently been expanded to RNA-Seq data. According to the Limma user guide, the current recommendation is to use the edgeR package's TMM normalization and the so-called 'voom'-conversion, which essentially transforms the normalized counts to logarithmic (base 2) scale and estimates their mean-variance relationship to assign a weight to each observation before linear modelling (173). The Benjamini-Hochberg technique is employed by default to estimate the FDR (170).

As a final step, the differentially expressed genes resulting from the statistical testing may be classified into common pathways. This approach finds differentially active pathways and, eventually, the possibility of linking molecular information (transcriptome) with an organism's phenotype (active pathways) (179). The approaches used to analyze the behavior of differentially expressed genes and concatenate them into cellular pathways make use of pathway information accessible in public sources such as the Gene Ontology (GO) (180) or the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (181). This is accomplished by associating pathway information in these databases with gene expression patterns, resulting in the transformation of a list of individual genes into a collection of pathways.

Table 1: Software packages mostly used to identify gene expression.

| **Method** | **Normalization** | **Read count distribution** | **Differential expression test** | **Reference** |
|---|---|---|---|---|
| edgeR | TMM/Upper quartile/ RLE (DESeq-like)/None | Negative binomial distribution | Exact test | (169) |
| DESeq | DESeq sizeFactors | Negative binomial distribution | Exact test | (174) |
| baySeq | Scaling factors (quantile/TMM/total) | Negative binomial distribution | Assesses the posterior probabilities via empirical Bayesian | (175) |
| SAMseq (samr) | SAMseq | Nonparametric method | Wilcoxon rank statistic and a resampling strategy | (176) |
| Limma | TMM | voom transformation of counts | Empirical Bayes method | (172) |
| Cuffdiff 2 (Cufflinks) | Geometric (DESeq-like/quartile /classic-fpkm) | Beta negative binomial distribution | *t-test* | (177) |
| EBSeq | DESeq median normalization | Negative binomial distribution | Evaluates the posterior probability via empirical Bayesian methods | (178) |

## 2.3 Freshwater mussels as a model organism

Molluscs constitute one of the ancient and most diverse groups of invertebrates, with over 90,000 described species. It is estimated that 50.000 to 55.000 are marine, 25.000 to 30.000 terrestrial, and 6.000 to 7.000 freshwater species (182).

Bivalvia is the second most diverse class of Mollusca after Gastropoda, with around 20,000 species (183). A great part of them - clams, oysters, scallops, and mussels - are a vital source of food and raw materials (pearls and nacre, for example) for humans, and have a significant economic value (184). Of a total of 1209 known species of freshwater bivalves, 1178 integrate a single order (i.e., Unionida) divided into eight families: *Unionidae*, *Margaritiferidae*, *Hyriidae*, *Mycetopodidae*, *Iridinidae*, *Etheriidae* (all Unionida), *Sphaeriidae* and *Cyrenidae* (both Venerida) (185). Freshwater mussels, from the order Unionida, are distributed all over the globe (except Antarctica), inhabiting rivers, lakes, and even artificial structures (e.g., channels) (186).



Figure 6: *Unio delphinus* in the Sabor River (Portugal). Photo credit: Ronaldo Sousa.

Besides the economic value and services provided to humans, these animals are responsible for important ecosystem functions (187). Healthy mussel beds are essential in freshwater ecosystems, as they are important filter feeders, capable of removing particles from the water column to the benthos, being in this way responsible for water purification (188; 187). Freshwater mussels actively participate in the process of nutrient cycling, increasing water quality and potentially participating in the nutrient reduction process (2; 3). Furthermore, they are an important group of ecosystem engineers, providing habitat to other organisms and support to food webs (4). Freshwater mussels can act as crucial biomonitors of environmental change, providing information about historical (using shells, for example) and current conditions and keeping track of impending changes in water quality (189; 190).

Additionally, freshwater mussels are fascinating because of their unique life cycle. Freshwater mussel's life cycle incorporates parental care and larval parasitism (Figure 7). The parasitic stage is a unique characteristic of freshwater mussels that distinguishes them from all other groups of bivalves (191). Their life cycle includes a free-living adult and a short-term larval stage called glochidium (192). Glochidia are obligate parasites present on the gills or fins of fish (and rarely other vertebrates)(193). The interaction between the freshwater mussel and their fish hosts their important for the mussel population, mostly because of glochidia's reduced lifespan and its limited dispersal capacity. However, these interactions can also contribute to the freshwater mussels' poor conservation status, as their complicated life cycle makes them particularly vulnerable to changing environmental conditions (191).

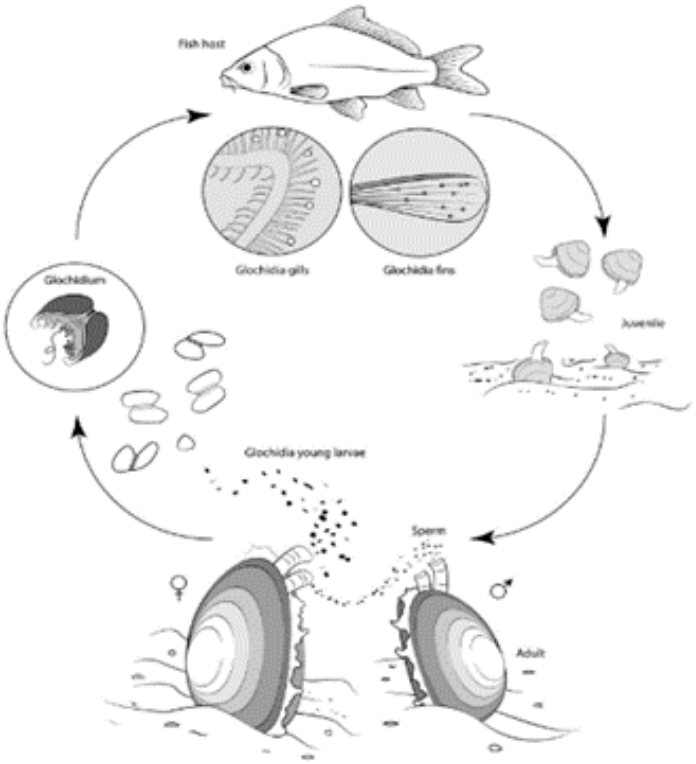



Figure 7: Scheme of freshwater mussel's life cycle. *Adapted from Modesto et al. (2018)*.

*Unio delphinus* is endemic to the Iberian Peninsula and is present in a wide diversity of habitats, from small to large rivers, subjected to a great variety of ecological conditions, ranging from oligotrophic waters to semiarid and more eutrophic streams (194). The genus *Unio* is part of the biggest freshwater mussel family (i.e., Unionidae), which comprises 674 out of 840 species (195). *Unio delphinus* is one of the species inside this family and its distribution comprises all major Atlantic Iberian River basins.

Furthermore, most freshwater mussel species are decreasing at a catastrophic rate worldwide. Over the last 30 years, freshwater biodiversity has decreased faster than in terrestrial or marine ecosystems

(196). The International Union for Conservation of Nature (IUCN, 2017) classified 217 of 535 freshwater mussel species as 'Near Threatened', 'Vulnerable', 'Endangered', or 'Critically Endangered', and 89 species as 'Data Deficient' (197). Globally, known major threats to freshwater biodiversity include loss, fragmentation, and degradation of habitat, overexploitation, pollution, the introduction of non-native invasive species, and climate change. European freshwater mussels are vulnerable to all these threats (198; 191; 199).

## 2.4 The effects of climate change on freshwater ecosystems

Climate change is a complex combination of stressors, including temperature and precipitation changes, rising atmospheric $CO_2$, rising frequency and intensity of droughts, floods, heatwaves, and cold spells (200). Climate change poses a serious threat to life on Earth, affecting aquatic and terrestrial ecosystems. The effects of climate change on biodiversity are expected to be unprecedented. It is estimated that 15-37% of terrestrial species will potentially become extinct due to climate change in the next 50 years and a similarly bleak future is suggested for freshwater species in the next decades (1). Although climatic variations occur naturally on seasonal to millennial time scales, the rates of warming observed in recent decades threaten the normal function of ecosystems (200). The rapid progress of climate change (201) has particularly affected aquatic ecosystems, further accentuated by anthropogenic stresses such as pollution, the introduction of invasive species, and habitat loss and fragmentation. When species are subjected to local climate changes, they frequently adapt by altering, for instance, their dispersal phenology or their physiology to reflect new climatic patterns (202). One of the most crucial questions in the debate on the ecological effects of climate change is whether the species will be able (or not) to adapt quickly enough to keep up with the rapid speed of changing temperatures (203).

According to Bellard et al. (2012), the various components of climate change are expected to have an impact on all levels of biodiversity (204). Organisms subjected to the effects of climate change will be affected in their genetics, physiology, and phenology (Figure 8).

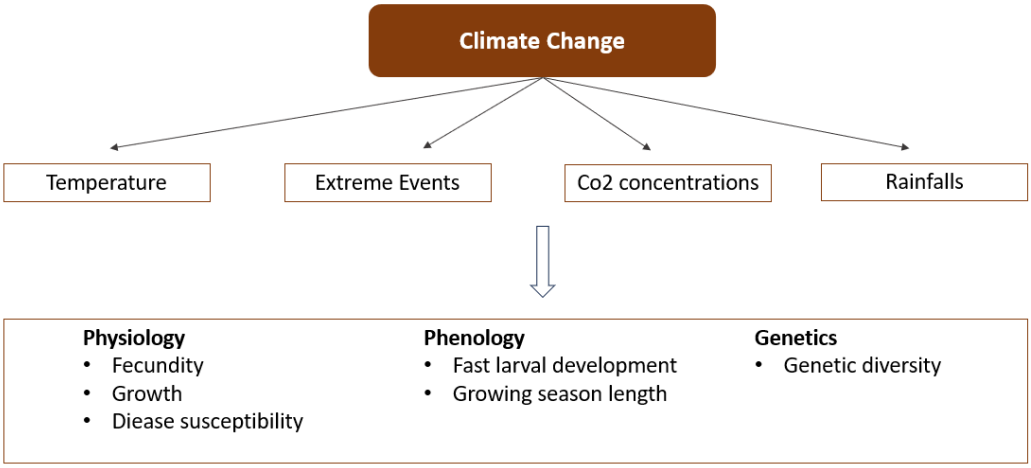



Figure 8: Effects of climate change in freshwater ecosystems. Adapted from Bellard et al. (2012).

### 2.4.1 Physiological processes

Thermal stress has critical physiological implications for the well-being of an animal (205). Freshwater mussels are thermo-conformers, meaning that their physiological processes are governed by the temperature of the surrounding water. In some freshwater ecosystems, summer water temperatures can reach 35 °C and exceed 40 °C, which is at or above the thermal limit of many mussel species (206). Due to their sedentary lifestyle, inability to control body temperature, and contaminants bioaccumulation, freshwater mussels are sensitive to temperature stress and water quality (207). Recent studies indicate that many species of freshwater mussels may already be living close to their upper thermal limits (208).

The primary factor constraining mussel feeding, respiration, and growth is temperature (209). Elevated temperatures are capable of altering metabolic rates in native mussels and decreasing the amount of energy that is available for growth (210). During unstressed conditions, the supply of Adenosine triphosphate (ATP) is adequate to meet the animal's baseline maintenance expenses as well as growth and reproduction. More of the limiting ATP supply is diverted to basal maintenance under moderately stressful conditions, which may result in less energy for growth and reproduction. During times of high stress, all ATP is directed into basic metabolism, leaving little or no energy for growth and reproduction (208).

Disease susceptibility is another concern relative to climate change and these organisms (211). Environmental conditions and poor water quality can reduce bivalves' heat tolerance, resulting in the sickness or mortality of the most sensitive species. Bivalve populations can suffer from significant mortality due to several illnesses induced by the presence of a wide range of pathogens. Furthermore, in the context of climate change, outbreaks of infectious diseases are projected to become increasingly common. According

to recent research, water pathogens increased by global warming will impair bivalves' immune systems and reduce their susceptibility to severe temperatures. As a result of physiological dysfunction, huge mortalities and significant losses in hatcheries, as well as natural ecosystems, are predicted (212). Although much less studied than marine species, it is expected that freshwater species will be more susceptible to parasites and diseases in the context of climate change.

### 2.4.2 Phenological processes

Changes in phenology are a major response of temperate aquatic organisms to climate change (213). Phenotypic plasticity is an individual's ability to modify their phenotype in response to environmental changes. Phenotypic plasticity allows an organism to adapt to changing environmental conditions and is widely regarded as a crucial method for coping with changing environmental conditions in nature (214).

The reproductive timing of mussels is directly affected and regulated by water temperature. The complex life cycle of freshwater mussels makes them more susceptible to disturbances caused by environmental stressors. The necessity of an obligatory larval parasitic stage on fish hosts makes unionid mussels more sensitive to the consequences of climate change (215). In many species, glochidia are released from the female in spring and early summer, from April to July, showing that they are sexually mature during this period (216). A study developed by Galbraith and Vaughn (2009) showed a peak in female egg size slightly earlier than a peak in sperm concentrations. Since females have to be ready for males to release their sperm, these observations suggest that female mussels reproductively mature earlier than males. Furthermore, this study also indicated that females present a larger concentration of developed eggs once in cool than in warm temperatures (217). It is expected that increasing temperatures will anticipate the reproduction of freshwater mussels (218). Thus, an extreme change in the reproductive timing of these populations can have consequences such as difficulties in gonad maturation, and unstable spawning season, that ultimately will affect recruitment and cause a decrease in the abundance of this species (219).

The relationship between freshwater mussels and their host fish is directly affected by warm temperatures. During the larval stage, many freshwater mussels can use a variety of different fish species to complete metamorphosis. However, some mussel species require specific fish species (191). Climate change can reduce the abundance of fish hosts, shift their distribution, and/or change the time of the year when they are present. Consequently, mussels that specialize in those fish may be negatively affected in future generations, especially mussels that are already threatened (220).

### 2.4.3 Genetic variation

Many studies have shown links between intraspecific genetic variation and stress tolerance (150; 221; 222; 223). Recently, studies have moved to link genetic diversity with the response in gene expression, under environmental stress. Thermal tolerance is frequently reflected in the differential expression of specific genes under thermal exposure. The variation among populations is frequently indicative of local adaptation. Studying populations along a thermal gradient may be particularly instructive because the differences in expression under heat stress, reflect local adaptation to the environment (224).

Intraspecific genetic variation plays an important role in understanding evolutionary changes. Climate change affects intraspecific diversity and can influence evolutionary adaptation to changing environmental conditions. In many cases, these changes will limit genetic variation in populations and species, to the point where low genetic diversity will lead to reduced population viability and extinction (150). According to simulations, climate change will significantly limit genetic diversity and connectedness between populations. The findings of a study using the freshwater mussel *Cumberlandia monodonta* (Say, 1829), indicated that the species will be further divided into smaller populations, each of which will be isolated and begin to differentiate genetically (202).

Species responding to climate change can adapt by shifting in abundance and distribution, by going extinct, or by evolving. Climate change is capable of decreasing the genetic diversity of populations. Species could genetically adapt to climate change through mutations or even the selection of existing genotypes (204). It is expected that warm temperatures will increase metabolism in most species, and result in an increase in mutation rates (150). Many species are threatened by climate change because they may not be able to adjust quickly enough to changing biotic and abiotic factors. If populations survive, they can handle climatic changes through phenotypic plasticity, or they can genetically adjust to changing environments based on genetic variety allowing for *de novo* mutations. The ability of a species to adapt to changing environmental conditions is determined by the degree of genetic diversity of a population in a given habitat, the extent of migration and gene flow between populations, and the pace of genetic innovation (e.g., by *de novo* mutations) (225). If incoming gene flow and dispersion are limited, abrupt environmental changes can cause severe demographic shifts (particularly in small populations), and only genetic adaptation can restore positive population growth and allow for the so-called evolutionary rescue of the population (226; 227).

Adaptation from standing genetic variation is essential under fast environmental change (as caused by GCC) since evolution can operate on alleles that already segregate at greater frequencies in the population (the pace of selection depends on effective population size) (228). Genomic data can give information on

genetic variation within species, on species' demography, and on effective population size. These population genomic characteristics may be computed with high confidence and serve as proxies or estimators of a population or species' sensitivity to changing climate, which is an important step in predicting evolutionary responses (226; 229).

Research on the prediction of how biodiversity would respond to climate change has become very active. Predictions are crucial for alerting scientists about potential hazards in the future, providing a way to support the link between biological changes and climate change (7).

## 2.5 Bioinformatics Integration

Bioinformatic tools are widely used to evaluate the genetic diversity among organisms, and transcriptomics has evolved into an effective technique for determining the molecular basis of biological processes. Transcription research necessitates a thoughtful selection of a model organism that aligns with the research goals. In this regard, freshwater mussels emerge as a valuable yet relatively understudied model in genomic research, especially in comparison to species of considerable commercial value that have received more attention (230). Anyway, freshwater mussel transcriptomes harbour a wealth of information, potentially encompassing key genes and pathways associated with critical stressors, such as temperature fluctuations (10). The choice of freshwater mussels as a model organism in a biological context highlights the untapped potential of these organisms in contributing to the understanding of evolutionary and adaptive characteristics. Nonetheless, their full genomic and transcriptomic potential remains largely unexplored due to a scarcity of available data. In the realm of computational biology, this scarcity of genomics resources presents a substantial challenge. While well-established model organisms benefit from comprehensive genomic datasets, the majority of non-model organisms, including freshwater mussels, still lack whole genome or transcriptome assemblies. The dearth of genomics resources in this context is striking. For nearly 1000 known species of freshwater mussels, only seven genomes (231; 232; 230; 233; 234; 235; 236) and less than 20 whole transcriptomes have been sequenced (91; 237; 238; 239; 240; 241; 242; 243; 244; 245; 246; 247; 248). This data scarcity underscores the pressing need for innovative computational approaches to unlock the hidden genomics potential within these organisms.

To date, not many studies using transcriptomic methodologies have been applied to freshwater mussels to assess the effects of climate change. However, some efforts have been made to understand what kinds of genes are expressed under thermal stress. Mussels increase the expression of transcripts cod-

ing for proteins when exposed to stress. These include many molecular chaperones such as heat shock proteins (249). The heat shock proteins - HSPs - are molecular chaperones, which assist proteins in appropriately folding during translation and facilitate movement across membranes (250). Heat shock proteins play a crucial role in environmental stress adaptation through protein folding, membrane translocation, misfolded protein degradation, and other regulatory activities (251). These proteins are also known as stress proteins and are present in all organisms (252). Heat shock proteins genes are expressed in response to stress, and their products are required for cell viability (253). The HSP70 family is the most researched group of HSPs in aquatic organisms, particularly in the study of thermotolerance (254). Because it is induced in aquatic species upon exposure to temperature change, salinity variation, handling, toxins, and other stresses, HSP70 is a potential bioindicator of environmental disturbance. Characterizations and functions of HSPs have become increasingly important in the adaptation process. Currently, different sequences of HSPs have been isolated from bivalves (255).

The study of transcriptomics, in particular by advanced sequencing techniques such as NGS, has revolutionized the understanding of biological systems. RNA Sequencing workflows enable a comprehensive examination of gene expression patterns, facilitating advanced studies like differential gene expression analysis. RNA Sequencing techniques applied in the study of DGEs provide powerful tools to discern key molecular pathways involved in response to stress. This powerful methodology may be applied in diverse fields but the lack of information on the genetic responses in freshwater mussels through climate change, emphasizes how important it is the need to advance this particular topic of study. The present work will help to identify which genes are differentially expressed in response to climate change. Ultimately, the insights gained from this research will enhance the understanding of how climate change impacts freshwater mussels and will provide valuable knowledge for implementing effective strategies to conserve and manage these organisms in a rapidly changing environment.

# Chapter 3

# Methodologies

## 3.1 Brief overview

The work developed in this master's thesis is part of the project: EdgeOmics - Freshwater Bivalves at the edge: Adaptation genomics under climate-change scenarios (http://doi.org/10.54499/PTDC/CTA-AMB/3065/2020). Hence, previous work was done prior to the beginning of the thesis, which includes animal sampling, ecological experiments, RNA extraction, and sequencing.

### 3.1.1 Sampling and laboratory acclimatization

A total of forty individuals of *U. delphinus* were collected in June 2022. Twenty individuals from the Rabaçal River, North of Portugal, and twenty individuals from the same species from the Guadiana River, South of Portugal. Individuals were transported alive to the University of Minho (Ecology Laboratory in IB-S - Institute of Science and Innovation to Bio-Sustainability of the University of Minho) where the ecological experiments were conducted. At the lab, mussels were placed in tanks (20cm/30cm) for acclimatization to a target temperature of 15°C. For that, the initial temperature of each tank was increased or decreased by about $\sim 2.5\,^{\circ}\mathrm{C\,d^{-1}}$ until the target temperature was reached. After this, individuals were maintained at the acclimatized temperature for two weeks, being fed daily with a solution of *Chlorella* (phytoplankton species).

### 3.1.2 Ecological Experiment

Two different types of temperature exposure experiments were performed: acute and chronic (Figure 9 and 10, respectively).

For the acute experiment, a total of ten individuals, five from the Northern population and five from the Southern population were used. The acute experiment consisted of a Critical thermal maxima methodology

(CTM) experiment, where 35°C was established as the maximum temperature. Briefly, starting at the acclimated temperature, individuals were subjected to a rapid progressive temperature increment (0,3 °C/min) until either the target temperature was reached or individuals exhibited extreme gaping behavior (i.e., adductor muscles relaxed, mantle edges and syphons split, and the foot becomes partially or fully extended (256)). When either of these conditions was reached, the experiment was finished, and gill tissue samples were collected and stored in RNAlater Stabilization Solution (ThermoFisher) at 4°C for 24h and then placed at -80°C.

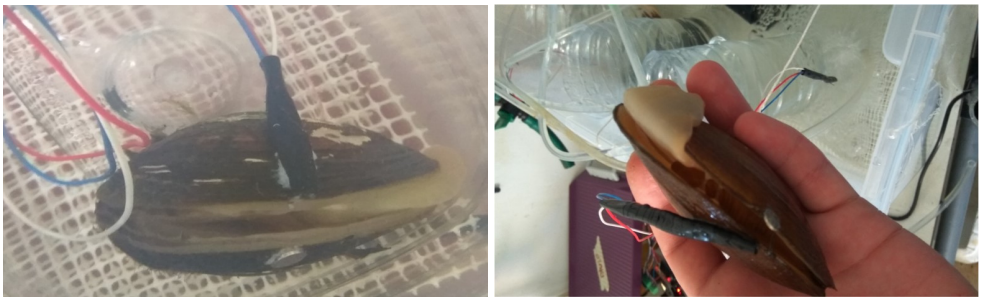


Figure 9: *Unio delphinus* used in the acute experiments.

The chronic experiments lasted for two weeks and involved a total of thirty individuals, fifteen for each population. As mentioned, the individuals started at the acclimatization temperature and subsequently, the temperature was gradually increased $\sim 1\ ^{\circ}\mathrm{C\ d^{-1}}$ until it reached 30°C. During this period of gradual increase in temperature, three distinct sampling stages were defined, i.e. 20°C, 25°C, and 30°C. At each temperature treatment, five individuals from each population were collected, gill tissues sampled, and placed in RNAlater as described above. All the tissues were stored at 80°C, at CIIMAR (Interdisciplinary Centre of Marine and Environmental Research).



Figure 10: *Unio delphinus* used in chronic experiments.

### 3.1.3 RNA extraction and NGS sequencing

Total RNA was extracted from the gill of all the sampled individuals, from each experiment, from the two populations (North and South) using the NZYTech total RNA isolation kit. RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) were obtained using a DeNovix DS-11 Series Spectrophotometer/Fluorometer. Due to the paired-end reads used in the present work, the extracted total RNA from the fourteen samples was sent to Macrogen, Inc. to build strand-specific libraries. The insert size was 250–300bp, and sequenced using 150bp paired-end reads on the Illumina HiSeq 4000 platform.

## 3.2 Bioinformatics

A diagram illustrating the pipeline used in this work is represented in Figure 11. The methods of quality control, alignment, and measurement of the level of gene expression were carried out using a high-performance cluster at CIIMAR. The code developed in these tasks corresponds to a series of commands executed within a Singularity container environment. Singularity is a containerization tool that allows the package and run applications and their dependencies in isolated environments.

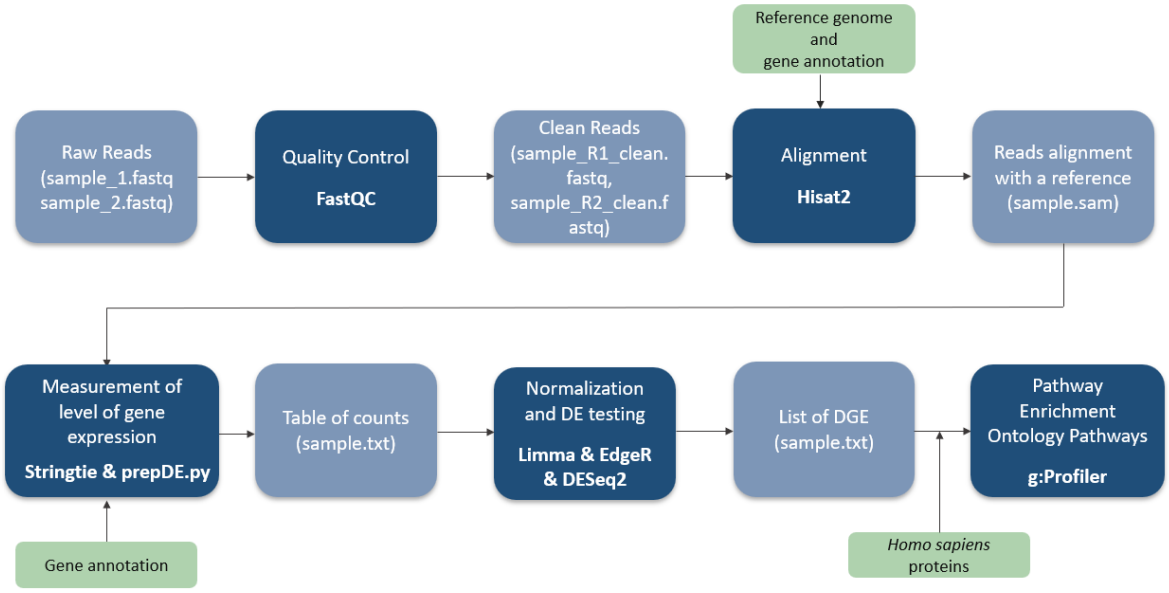



Figure 11: Bioinformatic methodological overview: light blue boxes represent the input/output files, dark blue boxes represent tasks and tools to perform the analysis and green boxes represent the sources files.

### 3.2.1 Processing raw data

Raw sequencing reads underwent a thorough comprehensive quality assessment using FASTQC software with the default parameters (version 0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This software systematically evaluates various aspects of sequencing data, including base quality scores, sequence length distribution, guanine-cytosine content (GC-content), and the presence of potential sequencing artefacts. Following the initial quality assessment, an additional preprocessing step was applied to improve the overall quality of the reads. Poor-quality bases, related to sequencing errors or technical issues, were removed from the raw reads using Trimmomatic, with the PE -threads 28 -phred33 parameters (version 0.38). Trimmomatic employs a range of quality filters and trimming parameters to enhance the accuracy and reliability of the sequence data by eliminating low-quality regions (257). After trimming the reads, FASTQC was applied again. The output of each FASTQC run is a set of HTML format reports for each sample.

## 3.3 Mapping Reads

The filtered *U. delphinus* reads were mapped to the *U. delphinus* reference genome assembly (available at NCBI under accession number GCA_029339505.1) (232) using the software HISAT2 with the parameters –dta -q -p 30 -x (version 2.2.1). HISAT2 represents a fast and sensitive alignment tool for mapping NGS reads to a reference genome. Previously, before initiating the mapping process, the reference annotation file was converted into a General Feature Format (GTF). This conversion was carried out using the software AGAT, employing the script agat_convert_sp_gxf2gxf.pl for seamless file format transition (version 1.0.0) and indexed with the software HISAT2 with the default parameters (version 2.2.1). Converting the reference genome file format to GTF has several advantages, including a more compact representation and a clearer illustration of the hierarchical link between genes, transcripts, and exons. This format is more suitable for bioinformatics analysis and tools.

After the file conversion, four commands were performed, two for generating the extracted splice sites file, another one for the exons file, and a last command to perform HISAT2-build. For the first two commands, the gene annotation file of *U. delphinus* was used. To perform the indexing, HISAT2-build was used with the parameters -p 30 –ss –exon (version 2.2.1). HISAT2-build constructs an HISAT2 index using the extracted splice sites and the exons files from the set of DNA sequences in the reference file. The output of the HISAT2-build is a summary file (summary.txt) that contains information about the genome index. This file provides statistics and details about the index, such as the total number of sequences and

total sequence length. These files are essential for the HISAT2 alignment process, and they collectively provide a comprehensive index of the reference genome or transcriptome. Once the index is built, the original sequence FASTA file is no longer used by Hisat2. The output of HISAT2 is one bam file for each set of paired-end read files.

### 3.3.1 Expression quantification

Following RNA-Seq read alignment, the data was quantified using the software Stringtie with the parameters -e -B -p 30 -G (version 2.1.2) and prepDE.py to get the read counts per gene. StringTie is a software tool for assembling transcripts and quantifying gene and isoform expression from RNA-Seq data. It assembles aligned reads (often in BAM format) into transcripts or gene models. It measures the levels of expression of various transcripts, revealing how much each transcript is expressed in different samples or situations. prepDE.py is a Python script that is frequently used in conjunction with other tools to prepare data for downstream DGE analysis. It aids in the organization of output from tools such as StringTie into a format appropriate for DGE. With the gene annotation file of *U. delphinus* it was possible to perform this task.

## 3.4 Differential expression

Gene counts produced by prepDE.py were employed in the differential expression analysis, which serves the purpose of identifying genes that exhibit differential expression across the study conditions. To carry out this analysis, it was used two distinct methodologies: one involved the use of an online tool, DEApp, while the other entailed the development of code in the R/Bioconductor package.

### 3.4.1 DEApp

Three methods available in DEApp were used: edgeR, limma-voom, and DESeq2 (https://github.com/yan-cri/DEApp). In the 'DE Analysis Filtering Criteria' parameters of the software, it was set FDR adjusted p-value, with the values 0.05 and 1.5 for p-value or FDR adjusted p-value and Fold Change, respectively. This software proves its utility by enabling the comparison of two conditions using a specific method or the simultaneous comparison of two conditions using two or three methods. This software enables the visualization of the output through graphic representations, such as an MDS plot, a volcano plot, a veen diagram, and a list of genes that are differentially expressed, i.e., up and down-regulated. Since combining the three methods has been shown to produce the most compelling results (258; 259), for this work, the list of genes resulting from the overlap of the three methods was used. The use of these genes

can offer several benefits such as robustness and reliability. Each method is based on its statistical model and assumptions and it is possible to cross-validate the results and lessen the risk of making conclusions based on assumptions that may not hold for the individual dataset by employing various approaches. The output findings will be more reliable if all three approaches agree on a group of differentially expressed genes. Also, the use of the three methods may reduce false positives and false negatives. By combining results from multiple methods, it is possible to reduce the risk of false positives and false negatives. Genes that are consistently identified as differentially expressed by multiple methods are less likely to be false discoveries.

### 3.4.2 R/Bioconductor package

A code in a R/Bioconductor package was developed for each method (edgeR (version 3.40.2), limma-voom (version 3.54.2), and DESeq2 (1.38.2)). Developing code offers the capability to compare two or more conditions at the same time, a feature not possible through DEApp. This feature provides significant advantages, especially in terms of flexibility and manipulating multiple conditions at the same time, enabling a more comprehensive exploration of the data.

The script's development progressed through the subsequent steps:

- **Read data**

  Two files were uploaded for the analysis: countData and colData. The countData file consists of a matrix containing gene expression values for all genes and samples. On the other hand, colData is a data frame that provides comprehensive information about the individual samples. This information includes sample name or ID and condition or treatment. A DGE list object was created with the two files and calculated the Count Per Million (CPM) values.

- **Data pre-processing**

  Data pre-processing was performed, including filtering out low-count genes and normalizing the data. Two steps for filtering data were implemented. First, a minimum CPM threshold was set to one. Genes with CPM values below this threshold were considered as lowly expressed and excluded from further analysis. The second step involved a requirement that genes must meet this CPM threshold in a minimum number of samples to be retained. This minimum sample threshold, gThreshold, was set to five because each condition has five samples. Genes that did not meet the CPM threshold in at least five samples were excluded. Following the filtering step, the data was normalized to account for variations in library size and sequencing depth across samples. The TMM

normalization method was employed to calculate normalization factors in the three methods. This method is essential for making the data directly comparable and for enabling accurate differential expression analysis.

- **Statistical Analysis**

  This step involved the identification of differentially expressed genes across experimental conditions or treatments. Differential gene expression analysis was performed using the R/Bioconductor packages EdgeR, Limma-Voom, and DESeq2. These packages offer robust statistical methods. For the three methods. The common step of each method was to create a design matrix. Next, for EdgeR a generalized linear model (GLM) approach was applied to evaluate differences in gene expressions between different conditions. The glmFit function was used for fitting the GLM to the data. In limma, the voom function was applied to the data. This transformation is an integral part of the Limma methodology, as it converts count data into a format suitable for linear modeling. The next step involves fitting linear models to the data using the lmFit function. These linear models take into account the design matrix, providing a framework to understand the relationships between the samples and the defined groups. The contrasts.fit function was applied to compute the desired contrasts, followed by empirical Bayes moderation using the eBayes function. For DESeq2 the foundation of the analysis begins with the creation of a DESeqDataSet matrix. DESeq2 employs a negative binomial distribution model to account for the inherent count nature of the data. The DESeq function was applied to the DESeqDataSet, using the Wald test as the chosen method for hypothesis testing.

- **Multiple Testing Correction**

  Multiple testing correction is a crucial consideration in any analysis involving a high number of genes, helping to manage the reduction of FDR. To control the FDR, both EdgeR and Limma employed the BH correction method. In EdgeR, the topTags function was utilized with specific parameter settings, including n = Inf, adjust. method = ”BH”, and sort.by = ”none”. Additionally, to further refine the results, EdgeR applied thresholds for retaining genes with FDR < 0.05 and $|\log2FoldChange| \geq 1.00$. For Limma, the topTable function was employed with the following parameter configurations: number=Inf, adjust=”BH”, and ‘sort.by=”none”. These settings ensured that the FDR was effectively controlled, and the most significant results were reported for further analysis. In Limma, a similar filtering approach was implemented as EdgeR with the same parameters. In DESeq2, the p-values obtained by the Wald test are corrected for multiple testing using the BH method by default.

- **Overlap between methods**

  After getting the list of genes that were differentially expressed for each method, similar to the DEApp analysis, an overlap of the three lists was performed to have a final list with the genes that were common to the three methods. For the overlap, it was set the values of logFC and padj of the method limma.

### 3.4.3 Comparative Analysis

Table 2 represents the comparative analysis conducted in chronic and acute experiments using both DEApp and R code. All the comparisons in the table were reproduced using the code in R, but only five: North 20 °C vs South 20 °C, North 25 °C vs South 25 °C, North 30 °C vs South 30 °C, North vs South (chronic and acute) were reproduced with DEApp. The two comparative analyses represented in the table in blue were developed exclusively with R code because they use more than two conditions at the same time and, thus are not possible through DEApp. These comparative analyses, North 20 °C ∩ 25°C ∩ 30°C and South 20 °C ∩ 25°C ∩ 30°C, correspond to the intersection of the two individual comparisons of each population (20 vs 25 and 20 vs 30) with the control of the temperature of 20 °C. This analysis aims to understand how genes increase or decrease in expression within the population.

The output of both, the software DEApp and R/Bioconductor package, is a table, which represents each gene that is differentially expressed for the condition that is being analyzed. For each gene, the output table contains information about the mean expression level as a joint estimate from both conditions and estimated separately for each condition, the fold change from the first to the second condition, the logarithm (to basis 2) of the FC, the p-value for statistical significance of this change and the p-value adjusted with BH procedure that controls the percentage of FDR.

Table 2: Comparative analysis conducted in chronic and acute experiments using both DEApp and R/Bioconductor package. The comparisons highlighted in blue were performed exclusively using the R/Bioconductor packages.

| Comparison | Description |
|---|---|
| **Chronic experiments** | |
| North 20 °C vs South 20°C | Comparison of gene expression between North and South population at 20°C. |
| North 25 °C vs South 25 °C | Comparison of gene expression between North and South population at 25 °C. |
| North 30 °C vs South 30°C | Comparison of gene expression between North and South population at 30 °C. |
| North vs South | Global comparison of gene expression between North and South population. |
| North 20 °C ∩ 25°C ∩ 30°C | Comparison of gene expression at different temperatures within the North population. |
| South 20°C ∩ 25°C ∩ 30°C | Comparison of gene expression at different temperatures within the South population. |
| **Acute experiments** | |
| North vs South | Comparison of gene expression between North and South population. |

## 3.5 Enrichment Analysis

The functional enrichment analysis was performed using g:Profiler (version e110_eg57_p18_4b54a898) with g:GOSt multiple testing correction method applying a significance threshold of 0.05 (260).

Before introducing the list of genes in g:Profiler we performed a protein BLAST (BLASTP), with the parameters -max_target_seqs 1 -max_hsps 1 -evalue 1e-5, against the reference proteins from the genome of *Homo sapiens* since the genome of *U. delphinus* was not available. A FASTA file containing the proteins of *Homo sapiens* was downloaded from Ensembl. After the BLAST, two lists were used: up and down-regulated gene ids of *Homo sapiens* that resulted from the BLAST.

g:Profiler is a powerful bioinformatic tool and web server used for gene and protein set enrichment analysis (260). It presents different categories depending on the aim, such as g:GOSt, g:Convert, g:ORrth and g:SNPense. For this work, g:GOSt was used, which performs functional profiling, specifying *Homo sapiens* (Human) as the model organism. The human genome is one of the most extensively studied and well-known genomes in genomics. In consequence, the human genome is well-annotated, meaning that the locations and functions of many human genes have been identified and characterized. The category used in this work maps genes to known functional information sources and detects statistically significantly enriched terms. g:Profiler supports various annotation databases, including GO, KEGG pathways, Reactome, and many more. These databases help in associating the input gene or protein list with known biological functions and pathways. g:Profiler employs statistical methods to assess the enrichment of input genes or proteins in various biological categories. It calculates p-values and enrichment scores to determine the significance of the associations. g:Profiler generates reports that list the enriched terms, associated statistics, and visualizations.

# Chapter 4

# Results and discussion

## 4.1 RNA extraction and NGS sequencing

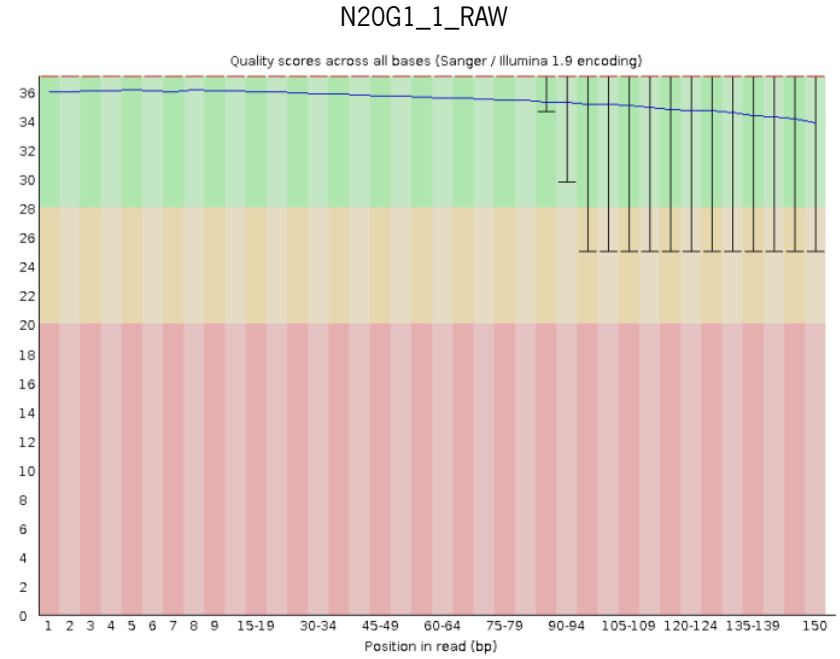
The concentrations and extraction quality measurements are in Appendix A, Section A.0.1. All values are in accordance with the requirements for NGS RNA-seq.

## 4.2 Processing raw data – FastQC and Trimmomatic

Reads are available upon request to the author of the thesis.

The reports generated by the FASTQC software, encompassing both the raw reads and the outputs from the Trimmomatic software, are accessible in https://figshare.com/projects/Identify_gene_expression_profiles_in_frehwater_mussels_under_thermal_stress/191853. Figure 12 to 15 illustrates the modules 'Per Base Sequence Quality' (Figure 12 and 13) and 'Adapter Contents' (Figure 14 and 15) of the FASTQC software output, showcasing data from the North population under a temperature condition of 20°C. The module 'Per Base Sequence Quality' provides a comprehensive overview of the quality of each base position in the sequence reads. Figure 12 presents some base quality variations across sequences, especially at the ends. After trimming the data with Trimmomatic software, the base quality at the ends of the sequences appears more uniform as regions of low quality have been trimmed (Figure 13). Figure 14 shows a red line that represents the level of adapter content in the sequencing data. This indicates that a portion of the sequences contains adapter sequences or other contaminants and eliminating adapters and low-quality regions can result in a higher average base quality across the entire sequence (Figure 15). Overall, the results of the FASTQ reports consistently demonstrated a high quality of the sequenced samples. Additionally, the application of pre-processing procedures has shown to be crucial in improving the data's quality.

**(A)**




**(B)**

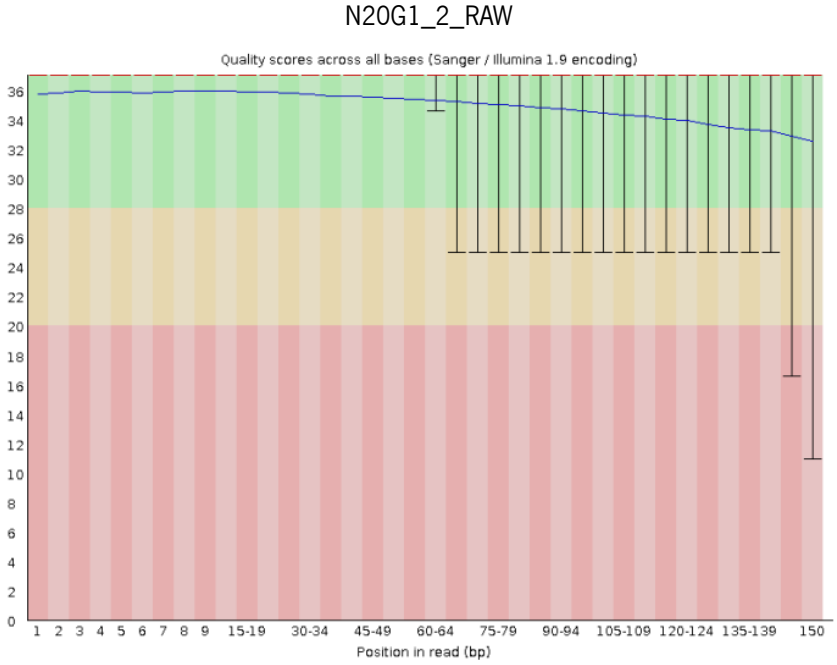
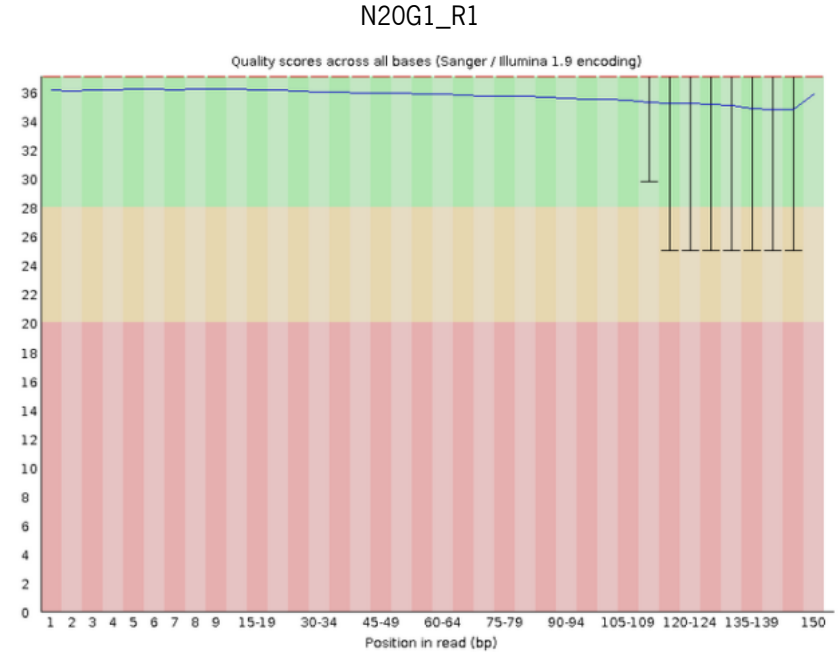



Figure 12: FastQC quality scores across all bases reports for the sample of North population under a temperature condition of 20°C before pre-processing. Figure 12 A and B represent the forward and the reverse fragments, respectively.

**(A)**




**(B)**

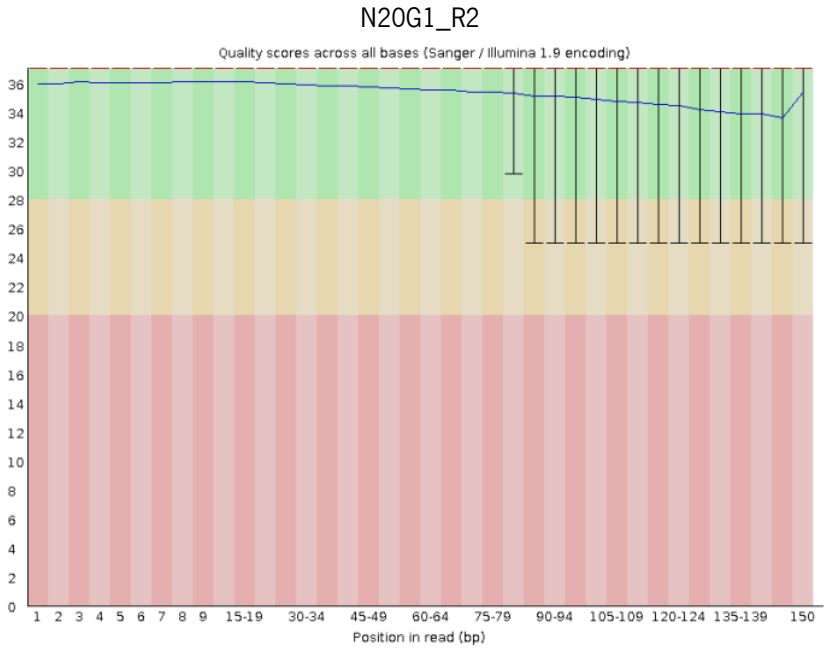
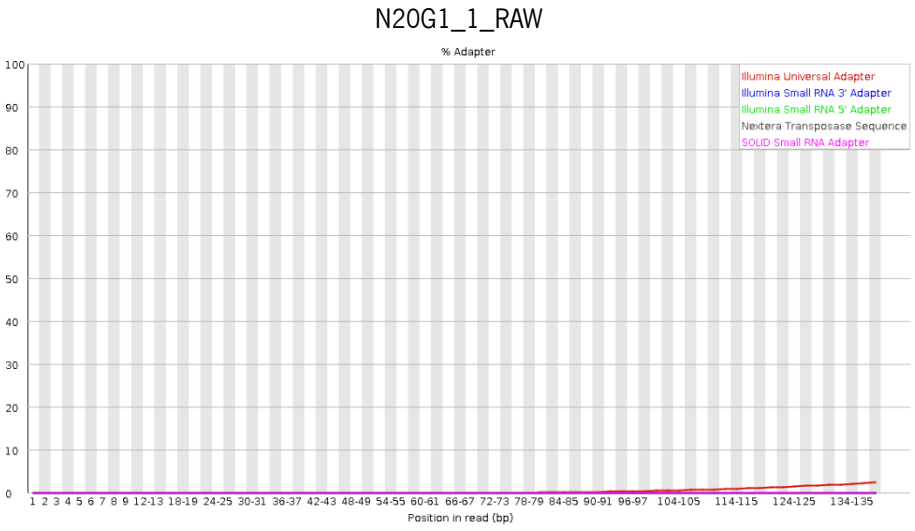



Figure 13: FastQC quality scores across all bases reports for the sample of North population under a temperature condition of 20°C after pre-processing. Figure 13 A and B represent the forward and the reverse fragments, respectively. The absence of whiskers in the processed data's graph indicates that the quality ratings of the processed data are greater, as shown in the graphs.

**(A)**




**(B)**

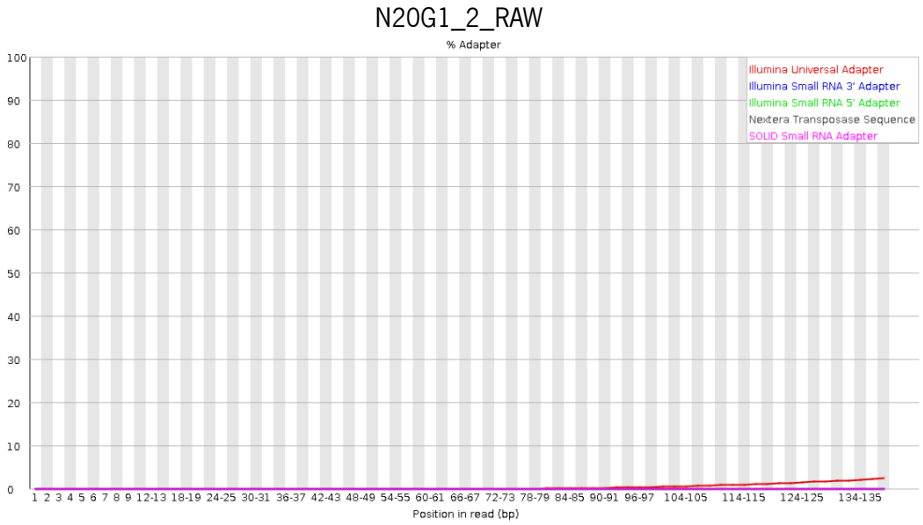
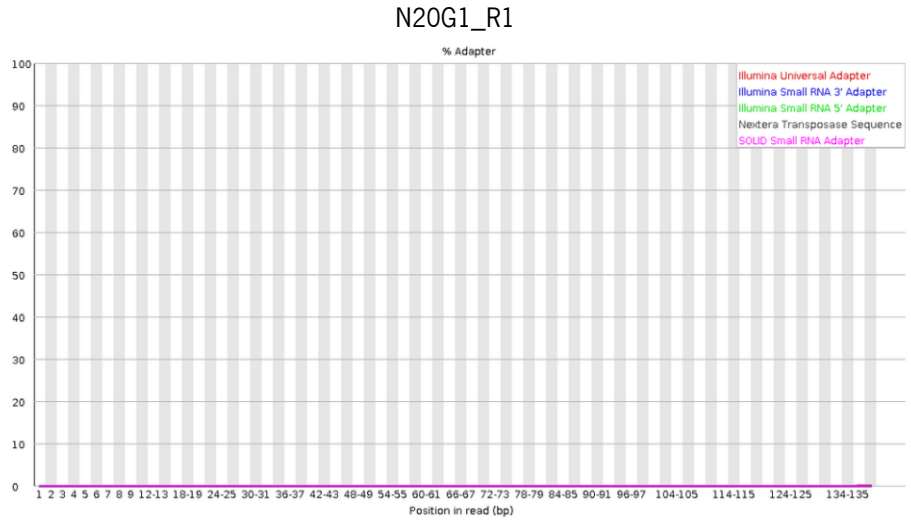



Figure 14: FastQC adapter content reports for the sample of the North population under a temperature condition of 20°C before pre-processing. Figure 14 A and B represent the forward and the reverse fragments, respectively.

**(A)**




**(B)**

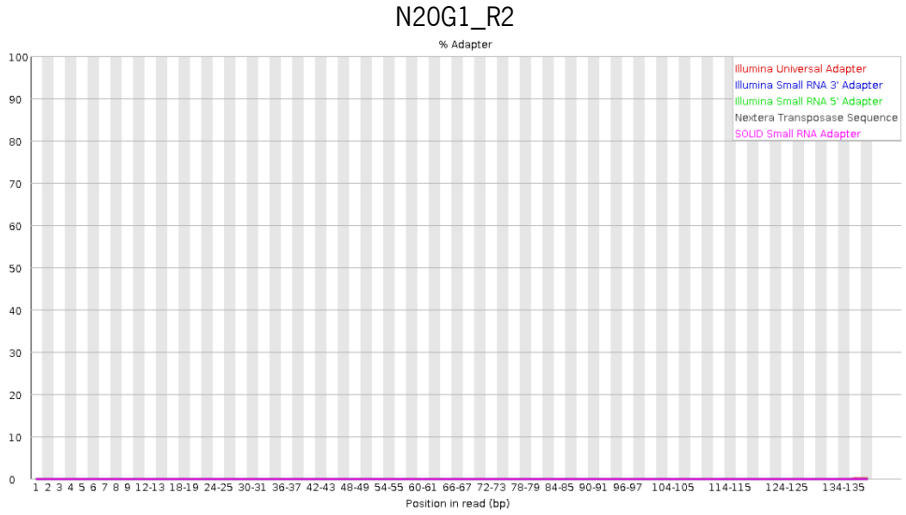



Figure 15: FastQC adapter content reports for the sample of the North population under a temperature condition of 20°C after pre-processing. Figure 15 A and B represent the forward and the reverse fragments, respectively. The adapters graph of the processed data no longer shows the red line, indicating that the adapters seen in the raw data have likewise been eliminated.

## 4.3 Mapping reads – HISAT2

The alignment rates, as represented in Tables 3 and 4, play a crucial role in assessing the quality and reliability of the results obtained in the RNA sequencing mapping against the *U. delphinus* genome. These rates, produced by the Hisat2 software, reached an impressive mark of 90%. The robustness of the alignment results obtained in this study is reinforced when contextualized with findings from other studies employing the same tool and methodology. Notably, studies conducted by Gomes-dos-Santos et al. (2023) and Smith et al. (2021) performed RNA sequencing analyses of freshwater mussel genomes using the Hisat2 software (231; 230). The consistency in the use of Hisat2 across these studies underscores the reliability of the alignment process. The high level of concordance suggests minimal errors or biases in the sequencing process. Furthermore, reliable alignments are foundational for subsequent steps in data analysis. The elevated alignment rates contribute to the overall trustworthiness of downstream analyses and interpretations. The code for the development of this task is available in Appendix A, Section A.0.3 and A.0.4.

Table 3: Alignment rate for each read sample of chronic experiments against the *Unio delphinus* reference genome. Samples were named according to two variables: population and temperature. North population samples were accessed with "N" and south populations with "S", and next to this first letter with the temperature. In the case of chronic experiments, it can be 20, 25, or 30. Following the temperature are the replicates, which are five per condition, and are represented by "G1", "G2", "G3", "G4" and "G5".

| | **North Samples** | **Alignment Rate (%)** | **South Samples** | **Alignment Rate (%)** |
|---|---|---|---|---|
| **Control** | N20G1 | 91.36 | S20G1 | 94.14 |
| | N20G2 | 93.07 | S20G2 | 91.42 |
| | N20G3 | 92.26 | S20G3 | 91.91 |
| | N20G4 | 91.67 | S20G4 | 92.86 |
| | N20G5 | 94.60 | S20G5 | 92.14 |
| **Treatment** | N25G1 | 94.02 | S25G1 | 93.96 |
| | N25G2 | 94.29 | S25G2 | 91.50 |
| | N25G3 | 92.50 | S25G3 | 91.36 |
| | N25G4 | 93.88 | S25G4 | 92.04 |
| | N25G5 | 95.32 | S25G5 | 92.57 |
| | N30G1 | 93.43 | S30G1 | 92.71 |
| | N30G2 | 94.43 | S30G2 | 95.06 |
| | N30G3 | 93.50 | S30G3 | 94.02 |
| | N30G4 | 94.19 | S30G4 | 92.86 |
| | N30G5 | 94.13 | S30G5 | 92.82 |

Table 4: Alignment rates for each read sample of acute experiment against the *Unio delphinus* reference genome. Samples were named according to two variables: population and temperature. North population samples were accessed with "N" and south populations with "S", and next to this first letter with the temperature. In the case of chronic experiments, can be 20, 25, or 30. Following the temperature are the replicates, which are five per condition, and are represented by "G1", "G2", "G3", "G4" and "G5".

| **North Samples** | **Alignment Rates (%)** | **South Samples** | **Alignment Rates (%)** |
|---|---|---|---|
| NG1 | 94.79 | SG1 | 93.45 |
| NG2 | 94.80 | SG2 | 93.50 |
| NG3 | 94.77 | SG3 | 93.18 |
| NG4 | 94.51 | SG4 | 93.07 |
| NG5 | 94.93 | SG5 | 90.48 |

## 4.4 Differential Expression - DEApp and R code

Ganser et al. (2015) investigated the effects of increasing water temperatures on the physiological processes of four species of freshwater mussels in North America (208). This study is the only reference that is currently available in the literature on the current subject, and as such, it provides a standard comparison to the present work. In their research, mussels underwent exposure to four distinct temperature treatments—20 °C, 25 °C, 30 °C, and 35 °C—each replicated five times in a randomized approach, with the 20 °C treatment serving as the baseline control. The present work's laboratory experiment design uses the same temperature conditions as Ganser et al. (2015), providing a reliable and comparable methodology. The two investigations employ the same temperatures, but the present work diverges by categorizing them into two separate experiments: chronic and acute. As described, the chronic experiments lasted for two weeks and employed gradually increasing temperatures,i.e., 20 °C, 25 °C, and 30 °C, while the acute experiments specifically used a temperature of 35 °C. This change allows a comprehensive exploration of the consequences of short-term versus prolonged thermal stress on freshwater mussels, expanding the understanding of their genetic responses to increasing temperatures caused by climate change. Both experiments were designed to recreate the events of heat waves but with different durations. The chronic experiments were conducted gradually to simulate the progressively increasing temperatures, which means a prolongated extreme event, while the acute experiments were designed to replicate the effects of a briefer extreme event. The following analyses explored in this section were produced separately for each experiment.

The source code developed in the R/Bioconductor package present in this chapter is available in the corresponding GitHub repository: https://github.com/beatrizsilva16/DGE_Analysis. The repository contains all the graphical representations and the code associated.

### 4.4.1 Data pre-processing

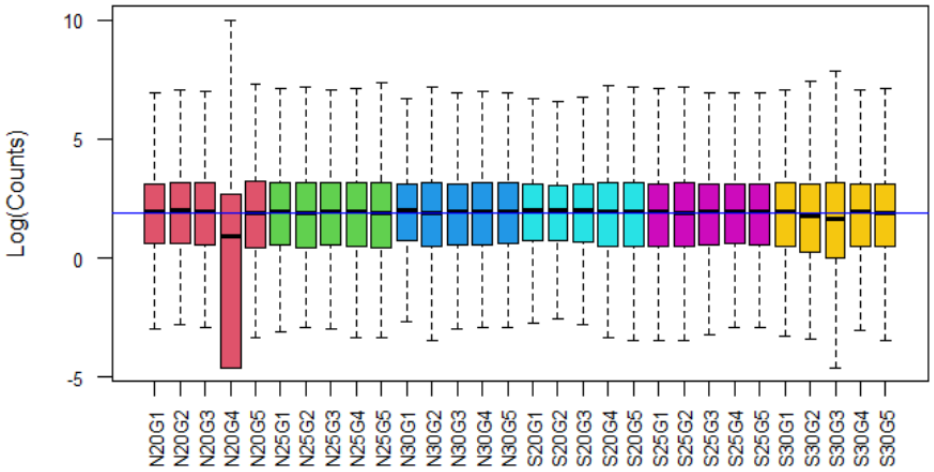
- Chronic experiments

The initial count of reads was significantly reduced from 39,060 to 22,696. This reduction in read counts reflects the removal of low-quality or non-relevant data, ensuring the retention of high-quality, informative reads for the subsequent analyses. The boxplot (Figure 16. A) and MDS plot (Figure 16. B) shown in Figure 16 reveal the presence of the 'N20G4' sample as an outlier within the North population at a temperature of 20 °C. Outliers can distort statistical measurements and may provide incorrect results. It was deemed necessary to exclude the N20G4 sample from the subsequent analysis. Removing this outlier produced a series of repercussions. The normalization thresholds were adjusted to the modified dataset, with a value of 4 to account for the remaining samples in the impacted condition. The visual representations after exclusion are shown in Figure 17. The credibility of the study findings is strengthened by the careful treatment and identification of outliers, which guarantees that the analyses that follow are based on a more accurate and representative dataset (261).

- Acute Experiments

For the acute experiments, the same procedure was applied, and the initial count reads were reduced from 39,060 to 18,509. A thorough study of the normalized data is presented in Figure 18, which also includes a boxplot (Figure 18. A) and an MDS plot (Figure 18. B). These graphical depictions act as visual confirmations, provide strong proof that the data is reliable, and accurately capture the subtleties of the experimental setup. The reliability and consistency of the findings in the acute experiment are supported by the decreased count readings and the well-organized patterns shown in the boxplot and MDS plot, contributing to the overall robustness of the analytical approach.

**(A)**




**(B)**

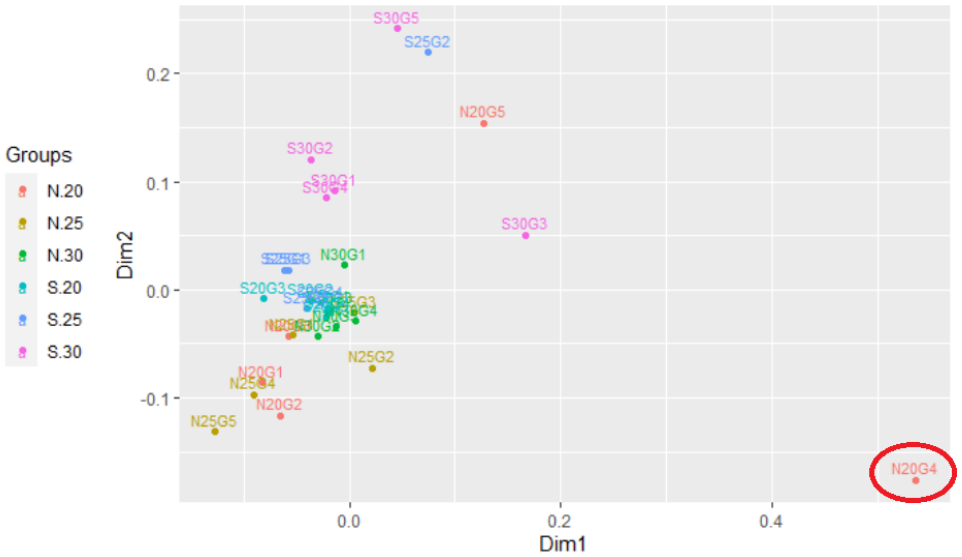
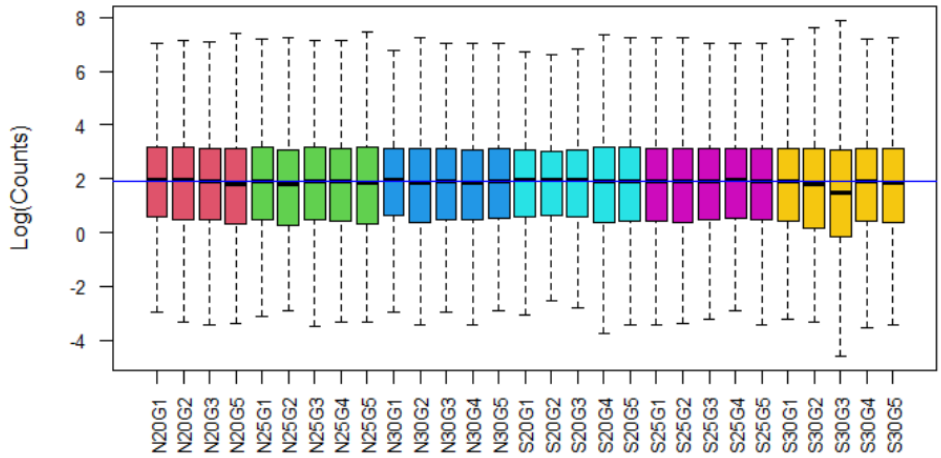



Figure 16: Boxplot (Figure 16. A) and MDS plot (Figure 16. B) of normalized samples distribution of chronic data, with N20G4 sample. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery. Both representations reveal the presence of the 'N20G4' sample as an outlier within the North population at a temperature of 20 °C. Outliers can distort statistical measurements and may provide incorrect results. It was deemed necessary to exclude the N20G4 sample from the subsequent analysis.

**(A)**




**(B)**

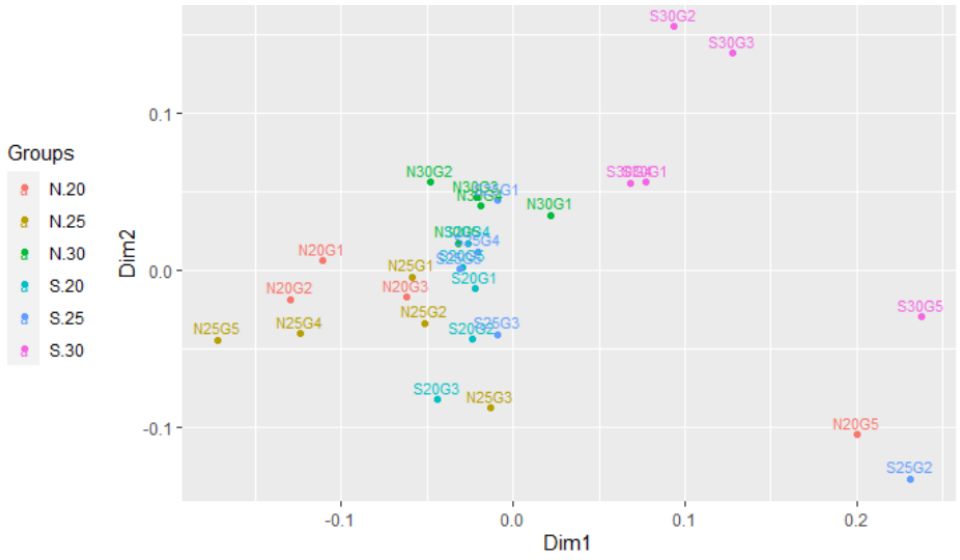
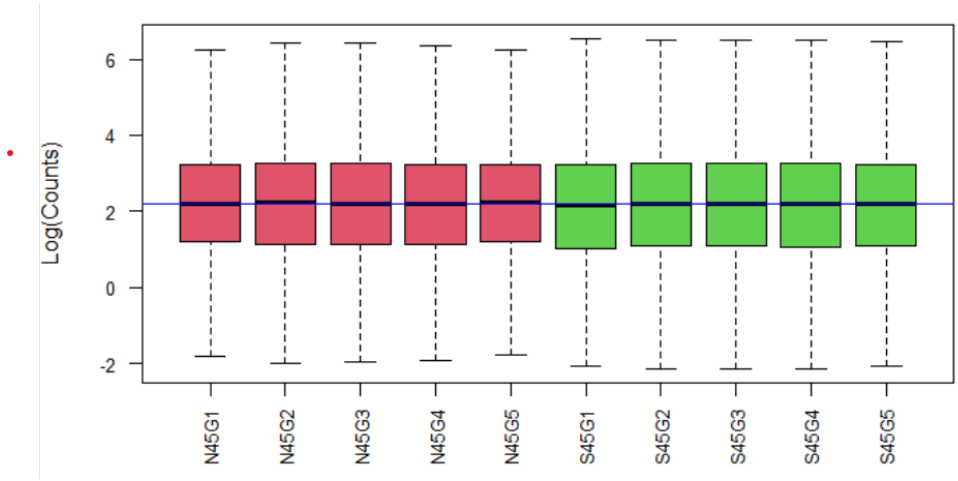



Figure 17: Boxplot (Figure 17. A) and MDS plot (Figure 17. B) of normalized samples distribution of chronic data, without N20G4 sample. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery.

**(A)**




**(B)**

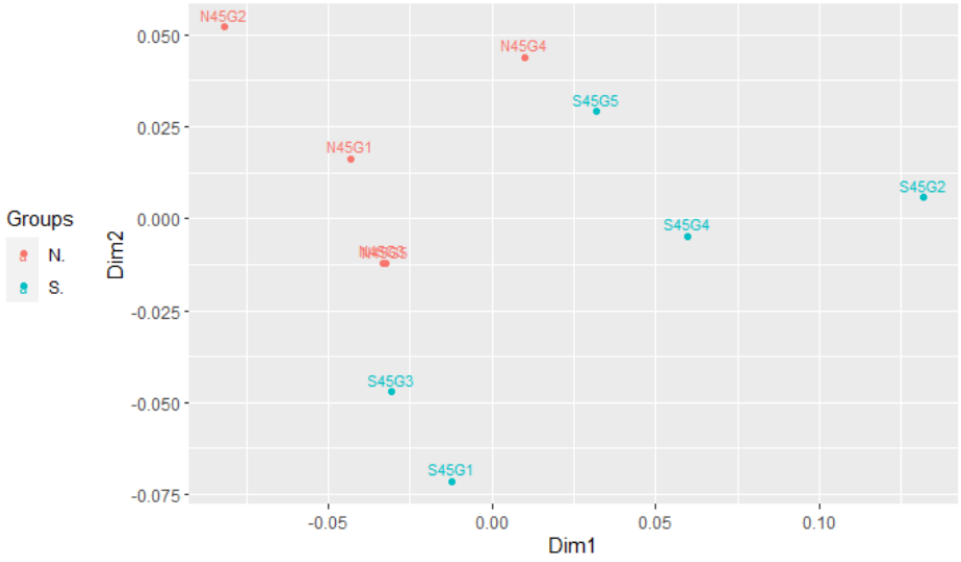



Figure 18: Boxplot (Figure 18. A) and MDS plot (Figure 18. B) of normalized samples distribution of acute data. A boxplot is a representation in which it is possible to visualize dataset distribution, highlighting the interquartile range (box), median (line), and data range (whiskers). An MDS plot projects multidimensional data into a simplified space, reflecting similarities and dissimilarities between points. Proximity indicates similarity, facilitating pattern discovery.

### 4.4.2 Comparative analysis

In recent years, an increasing amount of research has adopted a methodology to examine DGE, utilizing the most commonly used techniques such as edgeR, limma, and DESeq2 (259; 262). Combining several analytical techniques is a powerful strategy that facilitates a deeper robust investigation of gene expression patterns and a more sophisticated comprehension of their biological meaning (263). Since each method has its advantages and limitations, the use of edgeR, limma, and DESeq2 simultaneously improves the overall analytical depth of the study while making the analysis of gene expression more robust and effective. Until today, the simultaneous use of these three approaches has never been applied in a study using freshwater mussels. The present work fills this gap by investigating freshwater mussel gene expression patterns using the three methods at the same time. The combination of edgeR, limma, and DESeq2 is a new way to study differential gene expression in freshwater mussels, diverging from the traditional methodologies that have been employed in past studies (249; 243).

The values of FDR and p.adj in the output of the overlap were obtained using the limma method. Several studies, using RNA-seq data, have consistently reported that limma when compared to other methods, produces superior results and is often considered the most balanced method (264; 265). Limma uses a solid statistical approach by employing an empirical Bayes moderation technique. Furthermore, limma is flexible and can handle different experimental designs, even those with multiple factors (266). In addiction to its flexibility, limma integrates built-in methods to correct errors that may arise when simultaneously testing numerous genes (267). Its robust performance and balanced approach make it a widely favored choice for gene expression analysis, contributing to the reliability and accuracy of the obtained results.

Similar to the section above, the following analysis was divided into two different experiments, chronic and acute.

## Chronic Experiments

Figure 19 represents the estimated dispersion plots, for chronic experiments, resulting from applying three distinct methods—edgeR, limma, and DESeq2. Dispersion plots are essential for comprehending how gene expression varies between samples and are critical to the precision of differential expression analysis. The dispersion plots provide information about the degree of variability within the dataset by showing the spread of expression levels for each gene. Examining these graphs helps determine if the selected statistical techniques adequately capture the intrinsic biological variability and technological noise in the experiment data, which is the case of the results here presented, i.e. the statistics fit the data.

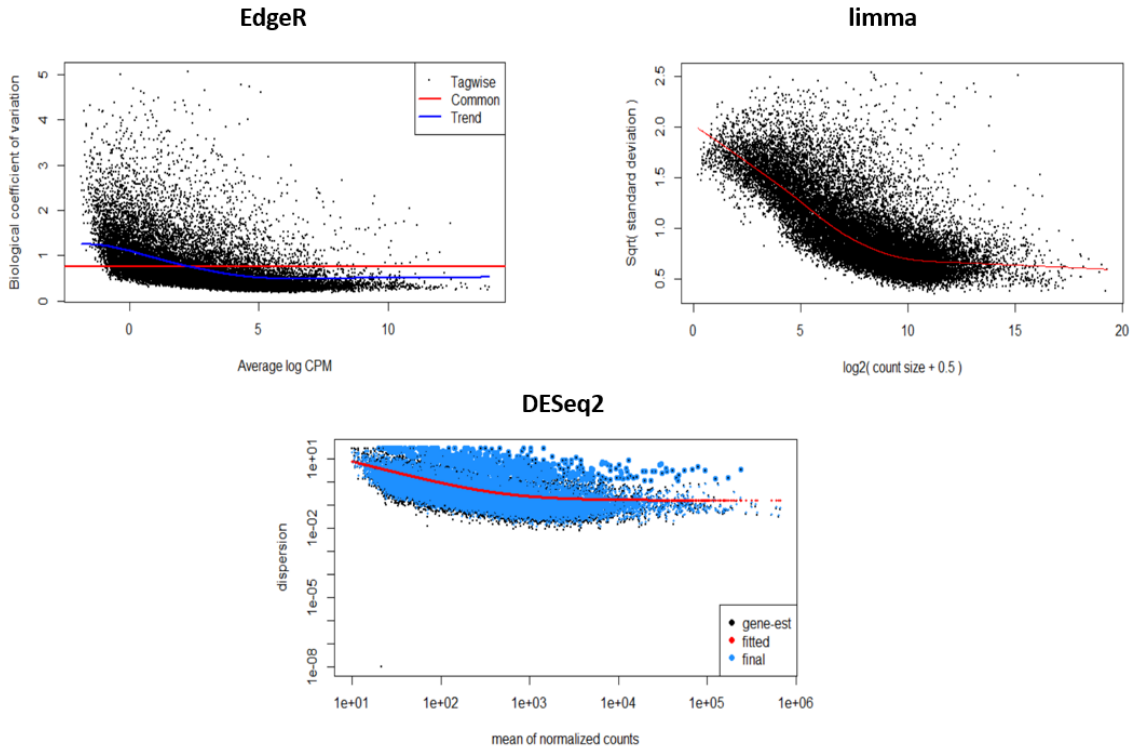



Figure 19: Estimated dispersion plots for chronic experiments, using three distinct methods: EdgeR, limma, DESeq2.

Table 5 provides an overview of the number of DGEs obtained in two different stages of the chronic experiments analysis. The first part illustrates all the comparative analyses with both DEApp and R/Bioconductor packages using the three methods separately. Additionally, the second part shows the number of DGEs obtained through the overlap between the three methods using exclusively the R/Bioconductor package. In the table, the number of DGEs in the two analyses (North 20 °C ∩ 25 °C ∩ 30 °C and South 20 °C ∩ 25 °C ∩ 30 °C) in DEApp is not represented, since this software does not allow the use of three conditions at the same time.

Table 5: Number of DGEs using the software DEApp and R/Bioconductor code in all the comparative analyses developed using the three methods, EdgeR, Limma, and DESeq2, and number of DGEs resulting from the overlap of the three methods in the analysis that showed consistency in the results.

| Comparisons | Number of DGEs | | | | | |
|---|---|---|---|---|---|---|
| | EdgeR | | Limma | | DESeq2 | |
| | DEApp | R code | DEApp | R code | DEApp | R code |
| North 20 °C vs South 20 °C | 641 | 538 | 166 | 169 | 850 | 1115 |
| North 25 °C vs South 25 °C | 1523 | 1527 | 746 | 718 | 1605 | 1912 |
| North 30 °C vs South 30 °C | 2924 | 2853 | 2093 | 2090 | 2758 | 3259 |
| North vs South | 2891 | 2891 | 1936 | 3030 | 1815 | 3095 |
| North 20 °C ∩ 25 °C ∩ 30 °C | - | 75 | - | 38 | - | 37 |
| South 20 °C ∩ 25 °C ∩ 30 °C | - | 534 | - | 266 | - | 77 |
| | Overlap between the three methods using R code | | | | | |
| North 20 °C vs South 20 °C | 130 | | | | | |
| North 25 °C vs South 25 °C | 594 | | | | | |
| North 30 °C vs South 30 °C | 1609 | | | | | |
| North vs South | 2150 | | | | | |

The obtained results align with expectations. As shown in Table 5, the number of DGEs in the comparisons is similar to the number produced by the R code. The similarity between the observed values highlights the robustness and consistency of the analytical techniques used in this work and the validation of the approach using Rcode. The similarity is a consequence of the use of the same methods, parameters, and thresholds in the two approaches, which lead to consistent findings in the analysis of DGE for this study. However, in the analysis of South 20 °C ∩ 25 °C ∩ 30 °C the number of DGE for the methods EdgeR, limma, and DESeq2 was 534, 266, and 77, respectively. When performing the overlap between the three methods, as shown in Figure 20, the number of the DGE is too low. In the Northern population, only 2 DGEs are resulting from the overlap, while in the Southern population, the number is 32.

Furthermore, when examining the Venn diagrams, a lack of DGE in the overlap between DESeq2 and Limma was observed. This observation aligns with the well-documented conservative behaviour of the DESeq2 method, which tends to be cautious in declaring genes as differentially expressed (268). These results are consistent with some studies that confirm that DESeq2 is extremely conservative and consequently loses sensitivity compared to the other methods (269; 270). Indeed, this feature is not universally applied to every analysis. In fact, in the rest of the comparative analyses represented, DESeq2 demonstrated a notable advantage, successfully identifying a greater number of differentially expressed

genes, as illustrated in Table 5. This attribute of DESeq2 comproves its sensitivity, which is different in the specific context. Depending on the experiment's circumstances and parameter settings, DESeq2 may exhibit higher sensitivity, resulting in the detection of either an increased or decreased number of genes as differentially expressed. The results highlight the complex behavior of DESeq2, demonstrating its extreme sensitivity in certain situations and its conservative character in others. These findings emphasize the significance of employing three methods, as each has been shown to exhibit certain limitations (271). Therefore the comparisons South 20 °C ∩ 25 °C ∩ 30 °C and North 20 °C ∩ 25 °C ∩ 30 °C were excluded from the further analysis.

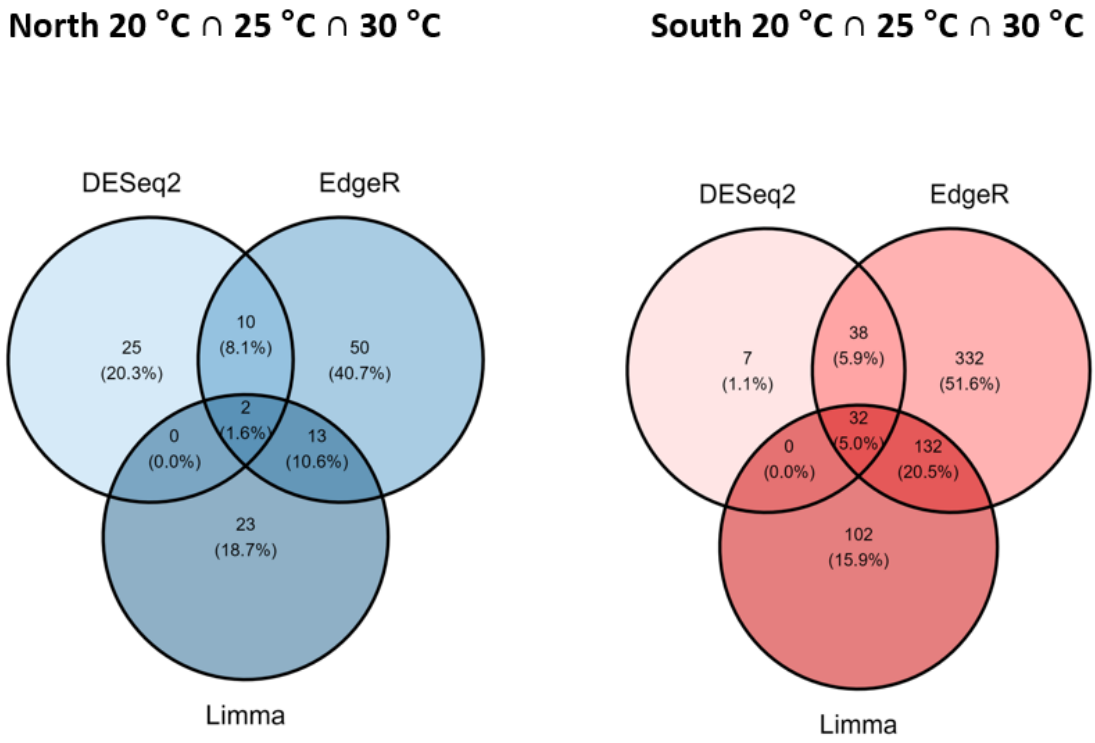



Figure 20: Veen diagram of the comparisons North 20 °C ∩ 25 °C ∩ 30 °C and South 20 °C ∩ 25 °C ∩ 30 °C.

For each comparative analysis a volcano plot and a heatmap (Figure 21 to 24) were employed, to visualize the results. Volcano plots were used to show importance with statistical significance and fold change criteria specified. Complementing the volcano plots, heatmaps were performed to provide a comprehensive overview of the differential expression of each comparative analysis.

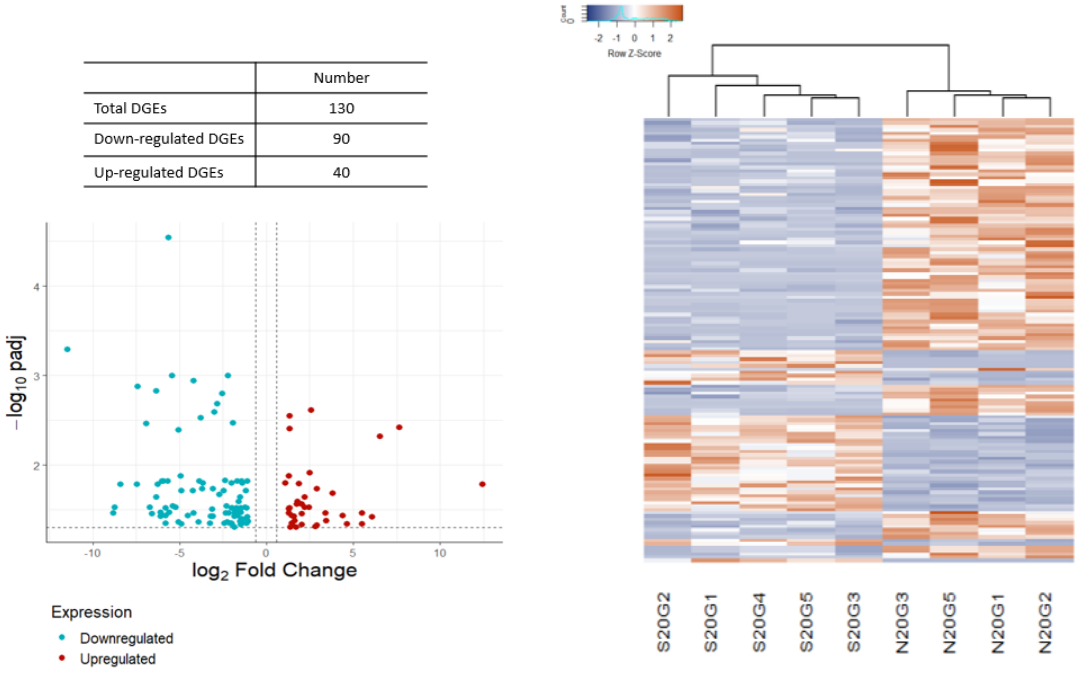



Figure 21: Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 20 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes.

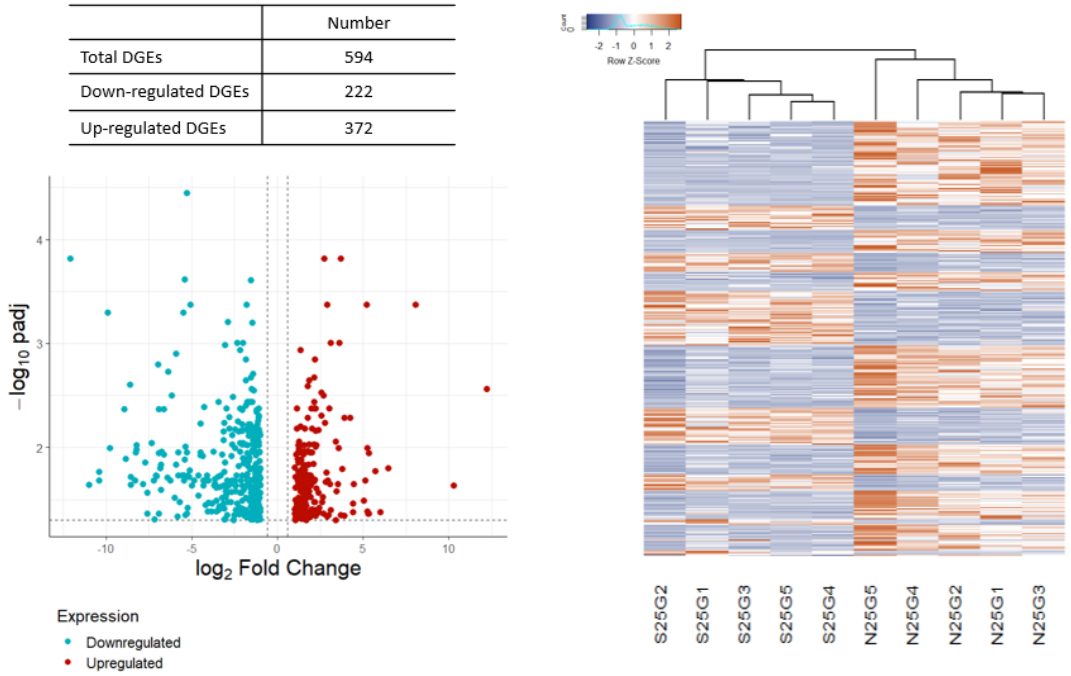



Figure 22: Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 25 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes.

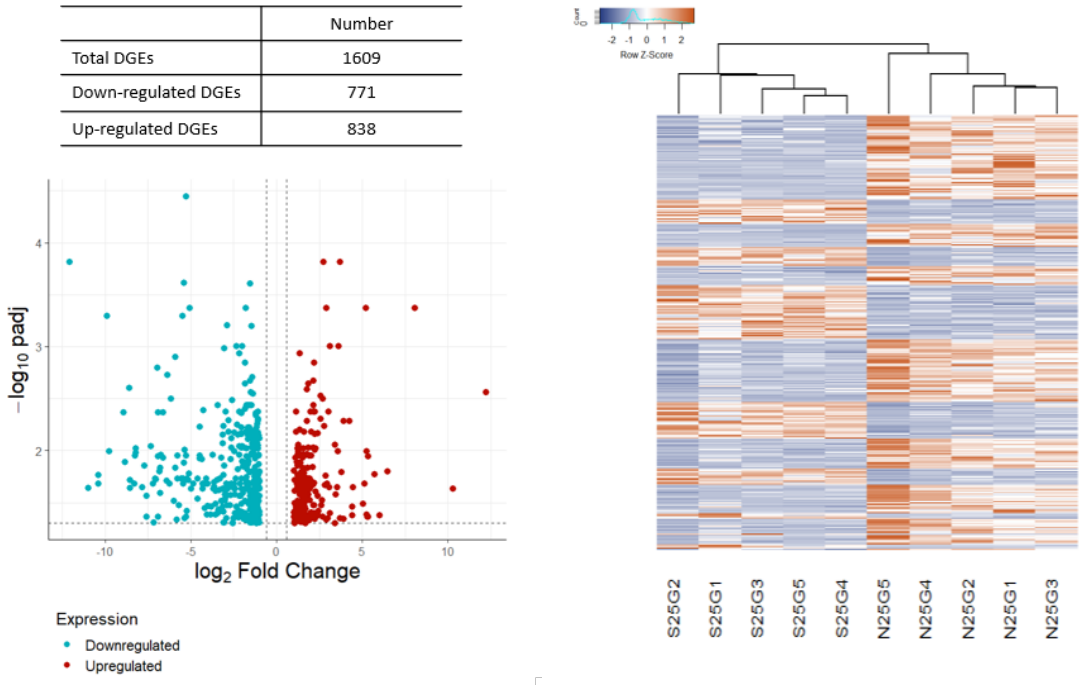



Figure 23: Volcano plot and heatmap for the comparative analysis of North vs South at the temperature of 30 °C of chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes.

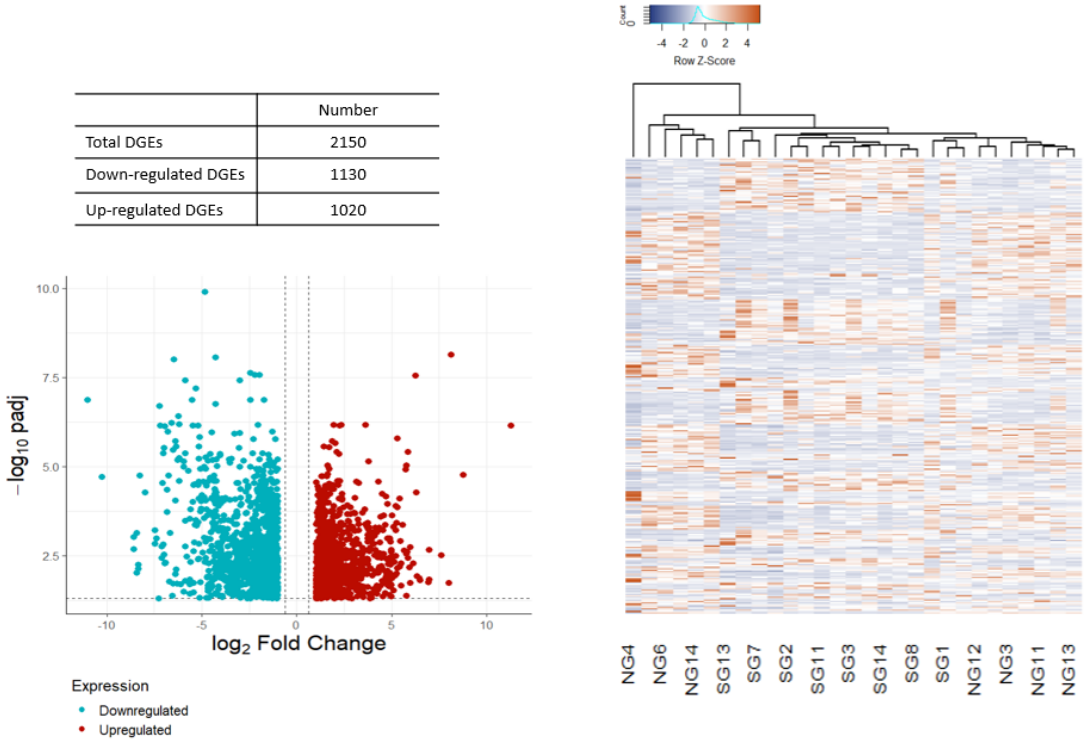



Figure 24: Volcano plot and heatmap for the comparative analysis of North vs South chronic experiments. The volcano plot includes the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. The blue points represent the down-regulated genes and the red ones the up-regulated genes.

## Acute Experiments

Figure 25 represents the estimated dispersion plots resulting from three distinct methods—edgeR, limma, and DESeq2- in the acute experiments. Once again the dispersion plot shows that statistically fit the data.

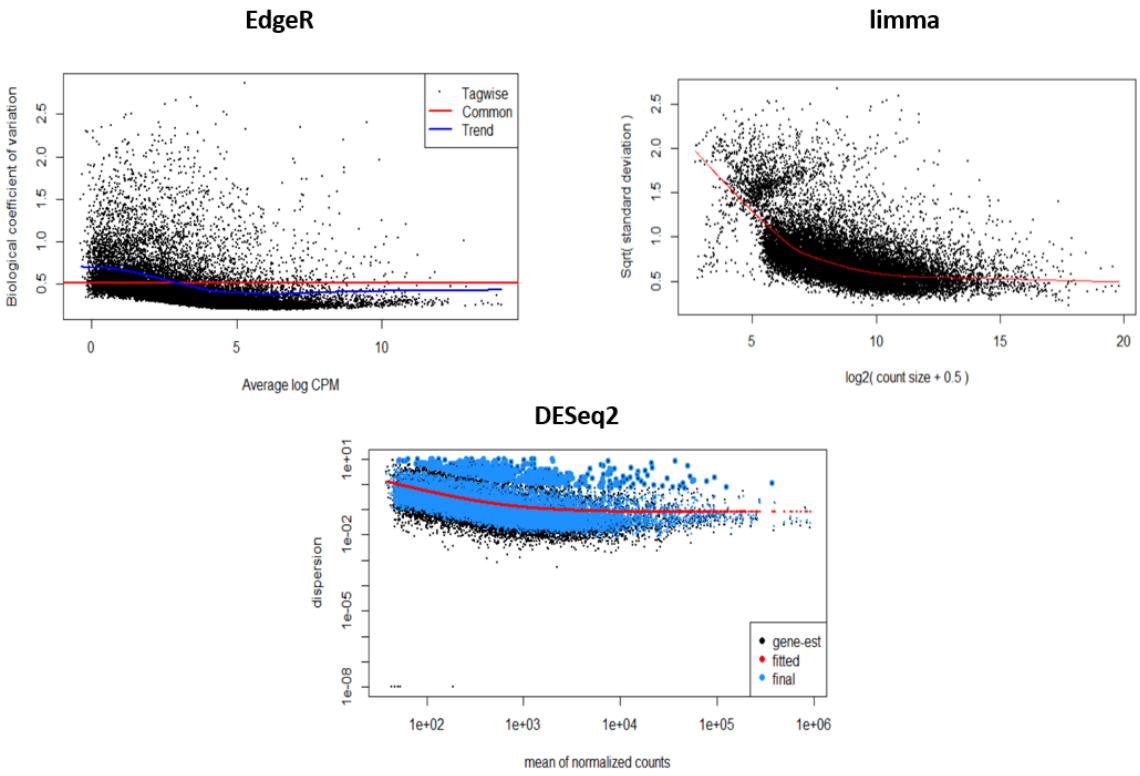



Figure 25: Estimated dispersion plots for acute experiments, using three distinct methods: EdgeR, limma, DESeq2.

Table 6 provides an overview of the number of DGEs obtained in the acute experiments. The first part illustrates the comparative analysis with both DEApp and R/Bioconductor packages using the three methods separately. Additionally, the second part shows the number of DGEs obtained through the overlap between the three methods using the R/Bioconductor package. The results generated are similar between the two approaches, which probes the reliability of the findings since it was applied the same methods and thresholds in both.

Table 6: Number of DGEs using the software DEApp and R/Bioconductor code in the comparative analysis developed using the three methods, EdgeR, Limma, and DESeq2, and number of DGEs resulting from the overlap of the three methods in the analysis that showed consistency in the results.

| Comparisons | Number of DGEs | | | | | |
|---|---|---|---|---|---|---|
| | EdgeR | | Limma | | DESeq2 | |
| | DEApp | R code | DEApp | R code | DEApp | R code |
| North vs South | 1433 | 1427 | 798 | 798 | 1467 | 1583 |
| | Overlap between the three methods using R code | | | | | |
| North vs South | 735 | | | | | |

For the comparative analysis a volcano plot and a heatmap (Figure 26) were employed, to visualize the results.

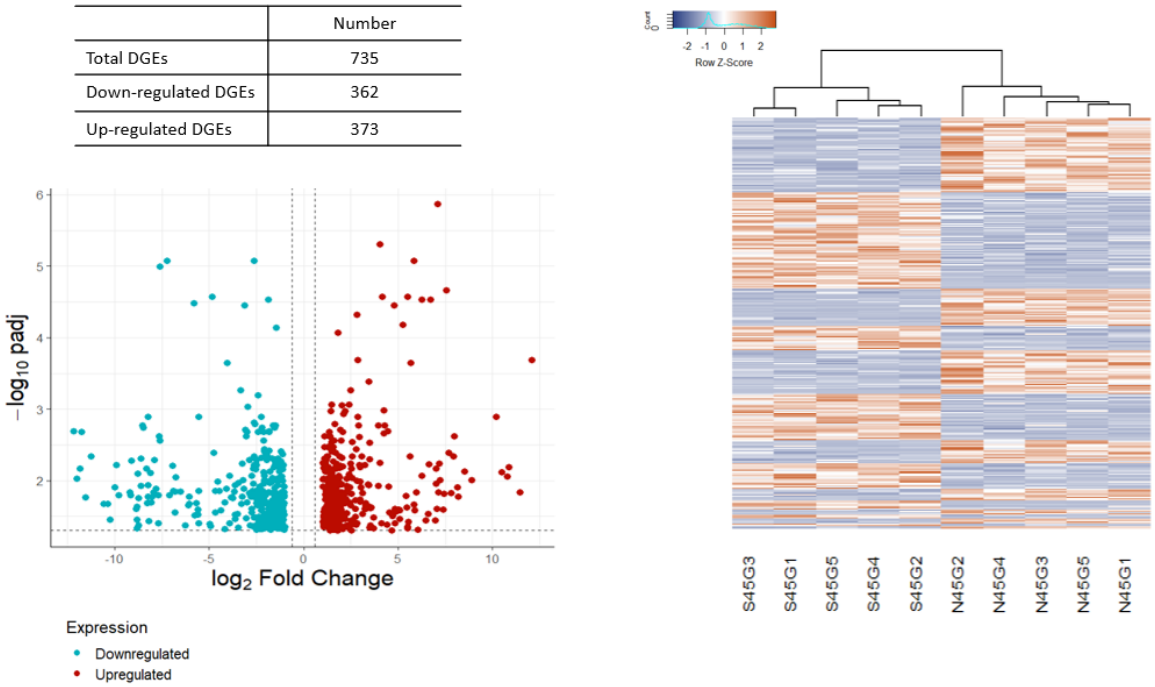



Figure 26: Volcano plot and heatmap for acute experiments for the comparative analysis: North vs South. Along with the volcano plot is represented the total number of DGEs, along with the down-regulated and up-regulated number of DGEs. Each data point on the plot corresponds to a unique gene. The blue points represent the down-regulated genes and the red ones the up-regulated genes.

### 4.4.3 Enrichment Analysis - g: Profiler

Using the pre-set options defined in the developed pipeline, g: Profiler identified the statistically significant ontology terms that differ between up and down-regulated lists of differentially expressed genes in the chronic and acute experiments. g: Profiler provides a strong method for functional enrichment analysis by methodically comparing the input differentially expressed gene sets against a wide range of functionally annotated databases. This allows the identification of biological terms, pathways, and molecular activities that are overrepresented and linked to the target genes. In the following subsections, a summary of the obtained results is described and compared of chronic and acute experiments.

**Chronic experiments**

Figures 27 and 28, provide an overview of g: Profiler results from chronic experiments conducted at three different temperatures, 20 °C, 25 °C, and 30 °C, and the comparative analysis of North vs South. The images in Figure 27 provide a clear visualization of molecular dynamics in response to different temperatures, revealing that as the temperature increases, the representations in each overview also intensify. This visual trend suggests a correlation between thermal magnitude and the complexity of molecular responses, underscoring the sensitivity of organisms to thermal stress.

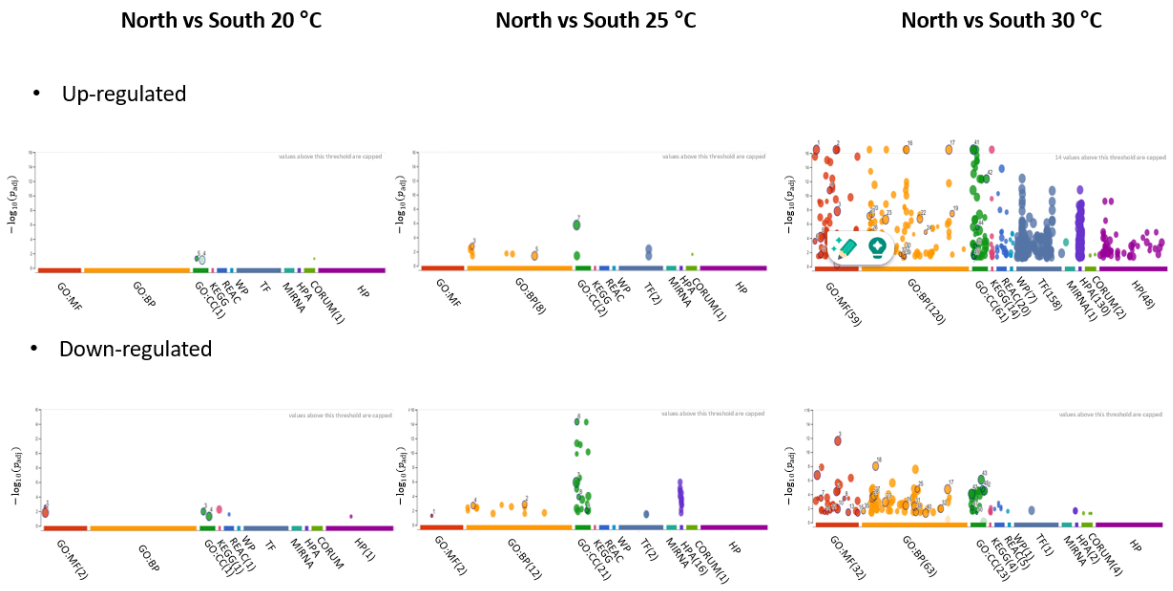
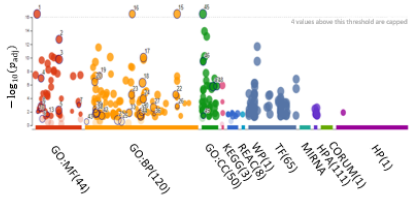



Figure 27: Overview of g: Profiler results in chronic experiments at three temperatures (20 °C, 25 °C, and 30 °C), highlighting statistically significant enrichments, and temperature-specific patterns.

- Up-regulated




- Down-regulated

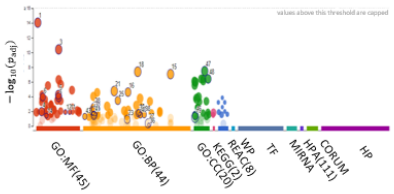



Figure 28: Overview of g: Profiler results in chronic experiments in the comparative analysis of North vs South, highlighting statistically significant enrichments, and temperature-specific patterns.

- North vs South 20 °C

  Table 7 lists the most significant ontology terms with their corresponding padj values, using g: Profiler. The list of source go: terms encompasses three distinct categories of ontology terms: Molecular Function (GO: MF), the database of manually annotated protein complexes (CORUM), Cellular Component (GO: CC), Human Phenotype Ontology (HP), and Reactome pathways (REAC). Table 7 includes some of the genes involved in each pathway. Section B.01 shows a detailed list of the go terms and the respective Ensembl ID, gene name, and description of the comparative analysis at 20 °C.

  **Molecular Function**

  The term GO:0003829 ($\beta$-1,3-galactosyl-O-glycosyl-glycoprotein $\beta$ -1,6-N-acetyl glucosaminyltransferase activity) emerges with a low padj of $2.291 \times 10^{-3}$. This ontology term refers to a particular enzymatic activity used in the processing or modification of molecules that include $\beta$-1,3-galactose. This molecular function is linked with the transferase activity involved in glycosylation (272). The down-regulation suggests a diminished ability to alter glycoproteins. As a result, some cellular functions like adhesion, and cell signaling may be compromised (273).

  **Cellular Component**

  The 90S pre-ribosome pathway (padj: $4.807 \times 10^{-2}$) is generated during co-transcriptional ribosome production, an early biogenesis intermediate (274). The up-regulation of this pathway may suggest a higher requirement of ribosome synthesis, which may be connected to the necessity for a cell to produce more proteins.

### Reactome

The Metabolism of ingested SeMet, Sec, MeSec into H2Se pathway (padj: $2.549 \times 10^{-2}$) is an integral component of Selenium Metabolism and Selenoprotein. Selenium corresponds to a vital trace mineral for human health, including antioxidative and anti-inflammatory properties (275). The down-regulation of this pathway may point to a decline in the metabolism of amino acids that include selenium (SeMet, Sec, and MeSec) into hydrogen selenide (H2Se).

In the present work, it was found that the genes glucosaminyl (N-acetyl) transferase 3 (GCNT3), pescadillo ribosomal biogenesis factor 1 (PES1), and cystathionine gamma-lyase (CTH) are involved in the pathways described above. Currently, and as far as we know, no published studies specifically investigates the correlation between these genes resulting from this pathway, and the impacts of rising temperatures on freshwater mussels. These findings highlight the importance of this work in elucidating previously undiscovered genes in this field of study.

Table 7: Gene Ontology (GO) Molecular Function (GO: MF), the database of manually annotated protein complexes (CORUM), Cellular Component (GO: CC), Human Phenotype Ontology (HP) and Reactome pathways (REAC) enrichment analysis of the comparative output analysis of North vs South at the temperature of 20 °C in chronic experiments, using g: Profiler.

| **Source** | **Term ID** | **Term Name** | **Gene Name** | **padj (up-regulated)** | **padj (down-regulated)** |
|---|---|---|---|---|---|
| GO:MF | GO:0003829 | $\beta$-1,3-galactosyl-O-glycosyl-glycoprotein $\beta$ -1,6-N-acetyl glucosaminyltransferase activity | GCNT3 GCNT1 | | $2.291 \times 10^{-3}$ |
| CORUM | CORUM:7197 | Methionine adenosyl-transferase $\alpha 1$ $\beta$-v1 | MAT1A | $4.982 \times 10^{-2}$ | |
| GO:CC | GO:0030686 | 90S preribosome | PES1 NOL6 | $4.807 \times 10^{-2}$ | |
| HP | HP:0010068 | Broad first metatarsal | NEK1 FGFR2 | | $4.967 \times 10^{-2}$ |
| REAC | R-HSA-2408508 | Metabolism of ingested SeMet,Sec, MeSec into H2Se | CTH NNMT | | $2.549 \times 10^{-2}$ |

- North vs South 25 °C

Table 8 lists the most significant ontology terms with their corresponding padj values, using g: Profiler. The list of source go: terms encompasses four distinct categories of ontology terms: Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), and the database of manually annotated protein complexes (CORUM). Table 8 includes some of the genes involved in each pathway. Section B.02 shows a detailed list of go terms and the respective Ensembl ID, gene name, and description of the comparative analysis at 25 °C.

**Molecular Function**

The term GO:0016823 (hydrolase activity acting on acid carbon-carbon bonds in ketonic substances) was identified with the padj value of $4.999 \times 10^{-2}$. The reduction in the expression of genes linked to hydrolase activity implies a possible contribution to the specific degradation of certain substances (276). The targeted molecules could function as energy stores or take part in metabolic processes that are essential to freshwater mussels' survival. The down-regulation suggests a tactical reaction to lessen the effects of environmental stress. This pathway may help mussels produce signal molecules for dealing with stress. In the present work, it was found that the genes fumarylacetoacetate hydrolase domain containing 1 (FAHD1) and fumarylacetoacetate hydrolase (FAH) participate in this pathway. Currently, and as far as we know, no published studies specifically investigates the correlation between these genes resulting from this pathway, and the impacts of rising temperatures on freshwater mussels. These findings highlight the importance of this work in elucidating previously undiscovered genes in this field of study.

**Biological Processes**

The regulation of viral-induced cytoplasmic pattern recognition pathway (padj: $1.794 \times 10^{-2}$) refers to the biological mechanisms by which cells detect and react to viral infections. The up-regulation of this pathway might indicate an improved immune response to any viral infections brought on by stress (277). In the present work, it was found that the gene toll-like receptor 6 (TLR6) is involved in this pathway. Wu et al., 2023 described this gene in a comparative transcriptomic analysis of a marine mussel during high-temperatures (278).

The pathway demonstrating a significant padj value includes cilium movement involved in cell motility (padj: $2.566 \times 10^{-2}$). Cilia or cilium are hair-like projections that emerge from the surface of certain living cells and are involved in various cellular functions such as movement and sensory processes (279). The down-regulation of this pathway suggests that the mussels are deliberately

changing how they move and perceive the environment, appearing to be less sensitive to temperature changes as a result of their decreased expression.

Moreover, increased sensitivity to environmental stimuli is shown by the up-regulation of genes involved in the cytosolic pattern recognition receptor signaling pathway (padj: $1.720 \times 10^{-3}$). The up-regulation of this pathway suggests that the organism is becoming more responsive to potential threats (280). This heightened sensitivity may be a protective measure that might be necessary to deal with challenges arising from changing environmental conditions, such as increasing temperatures, where a stronger immune response is required. Benítez et al., 2023 and Ertl et al., 2019 reported the genes interferon induced with helicase C domain 1 (IFIH1) and solute carrier family 15 member 4 (SLC15A4) as differentially regulated genes in the study of increasing temperatures on bivalves, respectively (281; 282). The present work, also reports these genes relative to this pathway.

### Cellular Components

The pathway dyneim complex (padj: $1.227 \times 10^{-4}$) is responsible for force generation and motility towards the minus ends of microtubules (MTs) (283). Motor proteins are essential for the operation of structures like cilia and flagella as well as for intracellular transport and organelle placement. Dynein is involved in the transport of vesicles and organelles toward the microtubule-organizing center or the cell center (284). The down-regulation of this pathway may suggest that intracellular transport and organelle placement may be affected.

Table 8: Gene Ontology (GO) Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), and the database of manually annotated protein complexes (CORUM) enrichment analysis of the comparative output analysis of North vs South at the temperature of 25 °C in chronic experiments, using g: Profiler.

| Source | Term ID | Term Name | Gene Name | padj (up-regulated) | padj (down-regulated) |
|---|---|---|---|---|---|
| GO:MF | GO:0016823 | hydrolase activity, acting on acid carbon-carbon bonds, in ketonic substances | FAHD1 FAH | | $4.999 \times 10^{-2}$ |
| GO:BP | GO:0039531 | regulation of viral-induced cytoplasmic pattern recognition | BIRC2 TLR6 | $1.794 \times 10^{-2}$ | |

Continued in the next page.

| Source | Term ID | Term Name | Gene Name | padj (up-regulated) | padj (down-regulated) |
|---|---|---|---|---|---|
| GO:BP | GO:0035082 | axoneme assembly | CFAP206<br>DRC1 | | $1.579 \times 10^{-3}$ |
| GO:BP | GO:0001578 | microtubule bundle formation | TEKT2<br>DNAJB13 | | $8.994 \times 10^{-3}$ |
| GO:BP | GO:0016823 | cytosolic pattern recognition receptor signaling pathway | IFIH1<br>SLC15A4 | $1.720 \times 10^{-3}$ | |
| GO:BP | GO:0060294 | cilium movement involved in cell motility | TEKT4<br>CFAP206 | | $2.566 \times 10^{-2}$ |
| GO:CC | GO:0097228 | sperm principal piece | CFAP206<br>CFAP20 | | $9.234 \times 10^{-3}$ |
| GO:CC | GO:0005879 | axonemal microtubule | ENKUR<br>RIBC2 | | $1.340 \times 10^{-10}$ |
| GO:CC | GO:0030286 | dynein complex | DRC1<br>DYNLT2B | | $1.227 \times 10^{-4}$ |
| GO:CC | GO:0005881 | cytoplasmic microtubule | TEKT4<br>CFAP206 | | $1.421 \times 10^{-6}$ |
| CORUM | CORUM:5529 | TRAF2-cIAP1/BIRC2 complex | BIRC2<br>TRAF2 | $2.394 \times 10^{-2}$ | |

- North vs South 30 °C

  Table 9 lists the most significant ontology terms with their corresponding padj values, using g: Profiler. The list of source go: terms encompasses five distinct categories of ontology terms: Molecular Function (MF), Biological Process (BP), Cellular Components (CC), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REAC). Table 9 includes some of the genes involved in each pathway. Section B.03 shows a detailed list of go terms and the respective Ensembl ID, gene name, and description of the comparative analysis at 30 °C.

  **Molecular Function**

  Elevated temperatures can disrupt the process of protein folding, leading to the misfolding and aggregation of proteins. The upregulation of the protein folding chaperone pathway (padj: $9.084 \times 10^{-3}$) suggests an adaptive strategy to preserve proper protein conformation, prevent aggregation, and ensure the functional integrity of essential cellular proteins under temperature-induced stress. Hu et al. (2022) and Tan et al. (2023) identified numerous genes including, heat shock protein family A (Hsp70) member 4 (HSPA4) and heat shock protein family B (small) member 6 (HSPB6)

respectively, as differentially expressed in the study of bivalves exposed to thermal stress (285; 286). The two mentioned genes were detected in the present study, and are linked to the protein folding chaperone pathway.

### Biological Processes

The Golgi lumen acidification pathway (padj: $1.418 \times 10^{-5}$) is important for enzyme activation and protein transport to occur (287). Cells may up-regulate this mechanism when temperatures rise to maintain optimal Golgi function and the up-regulation observed may influence protein processing and secretion. In the present work, it was found that the genes ATPase H+ transporting V0 subunit a1 (ATP6V1A) and ATPase H+ transporting accessory protein 1 (ATP6AP1) are involved in this pathway. Barret et al., 2022 described these genes relative to molecular responses in the mussel *Mytilus edulis* in response to thermal stress (288).

The vacuolar acidification pathway emerges with a low padj value of $2.281 \times 10^{-3}$. Vacuolar acidification is a homeostatic process that regulates intracellular pH, ion, chemical balance, and protein recycling, all of which are essential for optimal cellular function (289). When exposed to higher temperatures, organisms increase vacuolar acidity as part of their response mechanism (290). The up-regulation of this pathway may suggest efficiency in the process of elimination of metabolic exposure to higher temperatures.

In addition, the amino acid activation pathway emerges with a very low padj value of $5.068 \times 10^{-9}$. Moreover, amino acids are the basic building elements of proteins (291). The demand for protein synthesis frequently rises when an organism is exposed to higher temperatures (292). The up-regulation of this pathway may suggest an effective amino acid utilization, guaranteeing an adequate supply for protein synthesis. Furthermore, the cellular response to leucine, an important amino acid, emerges with the padj value of $4.359 \times 10^{-2}$. Leucine is essential for protein synthesis and cellular metabolism (293). Cells may up-regulate their sensitivity to leucine as temperatures rise to support increased metabolic activity (294). This pathway may be up-regulated as a cellular mechanism to address the increased demand for energy and biosynthetic activities associated with higher temperatures.

### KEGG

The up-regulation of the collecting duct acid secretion pathway (padj: $1.813 \times 10^{-3}$) suggests an adaptive response to rising temperatures. Acid secretion in the collecting ducts is critical for maintaining acid-base homeostasis, which is required for cellular activity (295). The activation of

this pathway may indicate an increased capacity to eliminate excess acids.

Warm temperatures frequently cause metabolic changes, and activation of the fatty acid degradation pathway (padj: $7.081 \times 10^{-4}$) indicates the organism's reliance on alternate energy sources. The fatty acid degradation is a catabolic process that uses fatty acids to create energy (296). This adaptation enables the organism to optimize its energy metabolism in response to increased temperature stress, guaranteeing a steady supply of energy for critical cellular functions. Alonso et al., 2020 described the gene cytochrome P450 family 2 subfamily U member 1 (CYP2U1) involved in this pathway in the species of *Physella acuta* (297). In the present work, this gene was also found.

Protein synthesis, a crucial biological function, depends on the aminoacyl-tRNA biosynthesis pathway (padj: $2.688 \times 10^{-10}$) (298). The upregulation of this process in the setting of high temperatures shows that there is a greater requirement for protein synthesis to support diverse cellular activities, such as stress response and repair mechanisms. Increased aminoacyl-tRNA biosynthesis may help the organism create proteins required for heat adaptation and resistance (299).

The comparative analysis of North vs South at the temperature of 30 °C demonstrated a pattern of genes related to the vacuolar ATPase (V-ATPase) complex, including ATP6V1A, ATP6V0A1, ATP6V0A, and ATP6V1F. These genes are present in several pathways such as ATPase activity, coupled to transmembrane movement of ions, rotational mechanism, endosomal lumen acidification, Golgi lumen acidification, synaptic vesicle lumen acidification, protein import into mitochondrial matrix, lysosomal lumen acidification, vacuolar acidification, proton-transporting V-type ATPase, V1 domaincomplex and Collecting duct acid secretion. Interestingly, all these pathways are upregulated. The vacuolar ATPases (V-ATPases) are a sort of proton pump that links ATP hydrolysis to proton transport within cells and across the plasma membrane. They participate in a wide range of cellular functions, including membrane traffic, protein processing and degradation, and small molecule linked transport, as well as physiological processes including urine acidity and bone resorption (300).

Table 9: Gene Ontology (GO) Molecular Function (MF), Biological Process (BP), Cellular Components (CC), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REAC) enrichment analysis of the comparative output analysis of North vs South at the temperature of 30 °C in chronic experiments, using g: Profiler.

| **Source** | **Term ID** | **Term Name** | **Gene Name** | **padj (up-regulated)** | **padj (down-regulated)** |
|---|---|---|---|---|---|
| GO:MF | GO:0010854 | adenylate cyclase regulator activity | CALM3<br>GRM7 | | $3.627 \times 10^{-2}$ |
| GO:MF | GO:0050135 | NAD(P)+ nucleosidase activity | TLR2<br>TLR4 | | $3.694 \times 10^{-4}$ |
| GO:MF | GO:0022848 | acetylcholine-gated monoatomic cation-selective channel | CHRNA3<br>CHRNA2 | | $2.146 \times 10^{-2}$ |
| GO:MF | GO:0044769 | ATPase activity, coupled to transmembrane movement of ions, rotational mechanism | ATP6V0A1<br>ATP6V1A | $5.787 \times 10^{-5}$ | |
| GO:MF | GO:0016875 | ligase activity, forming carbon-oxygen bonds | RARS1<br>WARS1 | $6.280 \times 10^{-10}$ | |
| GO:MF | GO:0044183 | protein folding chaperone | HSPA4<br>HSPB6 | $9.084 \times 10^{-3}$ | |
| GO:BP | GO:0061795 | Golgi lumen acidification | ATP6V1A<br>ATP6AP1 | $1.418 \times 10^{-5}$ | |
| GO:BP | GO:0036295 | cellular response to increased oxygen levels | CAV1 | | $1.942 \times 10^{-2}$ |
| GO:BP | GO:0048388 | endosomal lumen acidification | ATP6V0A1<br>ATP6V1F | $2.780 \times 10^{-5}$ | |
| GO:BP | GO:0043201 | response to leucine | LARS1<br>RRAGD | $4.359 \times 10^{-2}$ | |
| GO:BP | GO:0097401 | synaptic vesicle lumen acidification | ATP6V1A<br>ATP6AP1 | $2.339 \times 10^{-6}$ | |

Continued in the next page.

| Source | Term ID | Term Name | Gene Name | padj (up-regulated) | padj (down-regulated) |
|---|---|---|---|---|---|
| GO:BP | GO:0030150 | protein import into mitochondrial matrix | ATP6V1A ATP6AP1 | $1.191 \times 10^{-2}$ | |
| GO:BP | GO:0007042 | lysosomal lumen acidification | ATP6V1A ATP6AP1 | $1.175 \times 10^{-3}$ | |
| GO:BP | GO:0007035 | vacuolar acidification | ATP6V1A ATP6AP1 | $2.281 \times 10^{-3}$ | |
| GO:BP | GO:0043038 | amino acid activation | RARS1 WARS1 | $5.068 \times 10^{-9}$ | |
| GO:CC | GO:0035354 | Toll-like receptor 1-Toll-like receptor 2 protein complex | TLR2 TLR1 | | $4.439 \times 10^{-2}$ |
| GO:CC | GO:0033180 | proton- -transporting V-type ATPase, V1 domaincomplex | ATP6V1A ATP6V1D | $7.555 \times 10^{-4}$ | |
| GO:CC | GO:0017101 | aminoacyl-tRNA synthetase multienzyme complex | RARS1 FARSB | $1.271 \times 10^{-3}$ | |
| KEGG | KEGG:04966 | Collecting duct acid secretion | ATP6V0A ATP6V1A | $1.813 \times 10^{-3}$ | |
| KEGG | KEGG:00071 | Fatty acid degradation | CYP2U1 ECI1 | $7.081 \times 10^{-4}$ | |
| KEGG | KEGG:00970 | Aminoacyl-tRNA biosynthesis | RARS1 MARS1 | $2.688 \times 10^{-10}$ | |
| REAC | R-HSA-629597 | Highly calcium permeable ACh receptors | CHRNA3 CHRNA2 | | $1.584 \times 10^{-3}$ |

- North vs South

  Table 10 lists the most significant ontology terms with their corresponding padj values, using g: Profiler. The list of source go: terms encompasses five distinct categories of ontology terms: Molecular Function (MF), Biological Process (BP), Cellular Components (CC), and Kyoto Encyclopedia of Genes and Genomes (KEGG). Table 10 includes some of the genes involved in each pathway. Section B.04 shows a detailed list of go terms and the respective Ensembl ID, gene name, and description of the comparative analysis of North vs South.

### Molecular Function

The aryl sulfotransferase activity pathway (padj: $9.486 \times 10^{-3}$) refers to the enzymatic activity of aryl sulfotransferases. These enzymes transfer a sulfate group to aryl compounds, playing a role in the metabolism and detoxification of various substances (301). The up-regulation of this pathway may suggest a dynamic cellular response aimed at reducing the impacts of stress-induced alterations.

### Biological Processes

Macroautophagy is a cellular process that degrades and recycles cellular components (302). Negative regulation of macroautophagy pathway (padj: $1.051 \times 10^{-2}$) implies that this process is being suppressed. Thermal stress-induced up-regulation may reflect a cellular mechanism to prioritize other tasks over self-degradation during stressful situations.

The alcohol catabolic process pathway (padj: $2.339 \times 10^{-6}$) refers to the degradation of alcohol molecules. The up-regulation of this pathway during thermal stress might suggest an increased metabolic demand or altered energy requirements associated with the stress response.

The membrane depolarization pathway (padj: $1.392 \times 10^{-3}$) is the process by which the electrical potential across a cell membrane changes. Thermal stress-induced up-regulation may imply a cellular response targeted at changing membrane characteristics, potentially to adapt to the stressor.

### KEGG

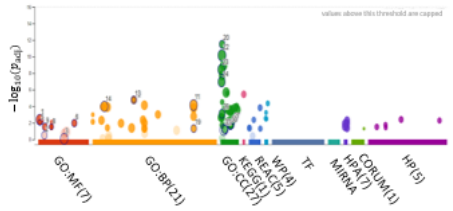
The ATP-binding cassette (ABC) transporters pathway (padj: $3.295 \times 10^{-4}$) is a widespread superfamily of integral membrane proteins that are responsible for the ATP-powered translocation of several substrates across membranes (303). The up-regulation of this pathway may indicate an increased cellular requirement for active transport of molecules across cell membranes.

In the present work, it was found that the genes sulfotransferase family 1A member 2 (SULT1A2), ring finger protein 41 (RNF41), glycerol kinase 5 (GK5) and purinergic receptor P2X 4 (P2RX4) are involved in the pathways described above. Currently, and as far as we know, no published studies specifically investigates the correlation between these genes resulting from this pathway, and the impacts of rising temperatures on freshwater mussels. These findings highlight the importance of this work in elucidating previously undiscovered genes in this field of study.

**Acute experiments**

Figure 27, provides an overview of g: Profiler results from acute experiments. The images provide a clear visualization of molecular dynamics in the present analysis.

- Up-regulated




- Down-regulated

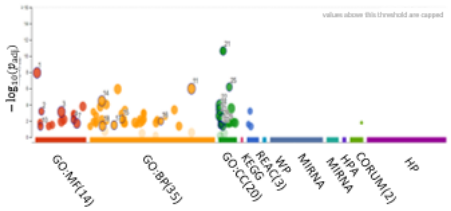



Figure 29: Overview of g: Profiler results in acute experiments, highlighting statistically significant enrichments, temperature-specific patterns.

Table 11 lists the most significant ontology terms with their corresponding p-values, using g: Profiler in the acute experiments. The list of source ontology terms encompasses four distinct categories: Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), Reactome (REAC), and Protein databases (CORUM). Table 11 includes some of the genes involved in each pathway. Section B.05 shows a detailed list of the go terms and the respective Ensembl ID, gene name, and description of the comparative analysis of acute experiments.

**Molecular Function**

The protein disulfide isomerase activity pathway (padj: $8.697 \times 10^{-3}$) refers to a protein disulfide isomerase's molecular function. Protein disulfide isomerases (PDIs) promote correct protein folding by catalyzing the rearrangement of disulfide links within proteins (304). The up-regulation of this pathway may indicate a greater necessity for appropriate protein folding and disulfide bond rearrangement. In the present work, it was found that the genes protein disulfide isomerase family A member 6 (PDIA6) and protein disulfide isomerase family A member 4 (PDIA4) are involved in this pathway. Kong et al., 2022 described these genes relative to molecular responses in the bivalve *Sinonovacula constricta* under thermal stress (305).

The oxidoreductase activity, acting on single donors by incorporating molecular oxygen pathway (padj: $3.441 \times 10^{-2}$) operates on single donors to incorporate molecular oxygen into the reaction. This enzyme participates in oxidation-reduction processes in which oxygen is a reactant. The up-regulation of this pathway may suggest an increase in the demand for molecular oxygen processes. This might be related to activities like cellular respiration, hypoxia response, or other oxygen-dependent metabolic pathways.

**Biological Processes**

Proteins require folding into an exact three-dimensional shape to function properly. The endoplasmic reticulum is a cellular compartment that is involved in protein production and folding (306). The up-regulation of protein folding in endoplasmic reticulum pathway (padj: $1.637 \times 10^{-5}$) may indicate a cellular adaptation to maintain proper protein conformation and mitigate potential misfolding stress under thermal challenges. In the present work, it was found that the gene heat shock protein 70 protein 5 (HSPA5) is involved in this pathway. Bultelle et al., 2021 described these genes relative to molecular responses in the bivalve *Mytilus edulis platensis* under thermal stress (307).

The regulation of viral-induced cytoplasmic pattern recognition signaling pathway (padj: $2.889 \times 10^{-2}$) refers to a complex cellular mechanism that affects the detection and subsequent signaling responses that are triggered by the presence of viral material in a cell's cytoplasm (308). This pathway is an important part of the innate immune system, which is the body's first line of defense against viral infections. The down-regulation of this pathway suggests a potential compromise in the cell's ability to recognize and respond to viral threats.

**CORUM**

The TLR1-TLR2 complex pathway (padj: $1.664 \times 10^{-2}$) is related to the generation and control of a particular molecular complex including Toll-like Receptor 1 (TLR1) and Toll-like Receptor 2 (TLR2). Toll-like receptors are a type of pattern recognition receptor that detects diverse microbial components and initiates immunological responses in the innate immune system (309). The down-regulation of this pathway in response to thermal stress may suggest a modulation of the immune response, potentially reflecting an adaptation to altered environmental conditions. In the present work, it was found that the genes TLR1 and TLR2 are involved in this pathway. Xu et al., 2018 described these genes relative to molecular responses in the bivalve *Mytilus edulis coruscus* under thermal stress (310).

Table 10: Gene Ontology (GO) Molecular Function (MF), Biological Process (BP), Cellular Components (CC), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome (REAC) enrichment analysis of the comparative output analysis of North vs South in chronic experiments, using g: Profiler.

| **Source** | **Term ID** | **Term Name** | **Gene Name** | **padj (up-regulated)** | **padj (down-regulated)** |
|---|---|---|---|---|---|
| GO:MF | GO:0004062 | aryl sulfotransferase activity | SULT1A2 SULT1B1 | $9.486 \times 10^{-3}$ | |
| GO:MF | GO:0061809 | NAD+ nucleotidase, cyclic ADP-ribose generating | TLR1 TLR4 | | $1.219 \times 10^{-2}$ |
| GO:MF | GO:0022848 | acetylcholine-gated monoatomic cation- - selective activity | CHRNA9 CHRNA4 | | $6.576 \times 10^{-4}$ |
| GO:MF | GO:0005231 | excitatory extracellular ligand-gated monoatomic ion channel activity | CHRNA9 CHRNA4 | | $1.350 \times 10^{-2}$ |
| GO:MF | GO:0072349 | modified amino acid transmembrane transporter activity | SLC16A9 SLC25A13 | $8.170 \times 10^{-3}$ | |
| GO:MF | GO:0140359 | ABC-type transporter activity | ABCA5 ABCG8 | $9.794 \times 10^{-4}$ | |
| GO:BP | GO:0016242 | negative regulation of macroautophagy | RNF41 RUBCN | $1.051 \times 10^{-2}$ | |
| GO:BP | GO:0046164 | alcohol catabolic process | GK5 SNX17 | $2.339 \times 10^{-6}$ | |
| GO:BP | GO:0072337 | modified amino acid transport | SLC16A9 LRP2 | $2.954 \times 10^{-3}$ | |
| GO:BP | GO:1901616 | organic hydroxy compound catabolic process | GK5 CYP3A4 | $2.339 \times 10^{-6}$ | |
| GO:BP | GO:0051899 | membrane depolarization | P2RX4 SLMAP | $1.392 \times 10^{-3}$ | |
| GO:BP | GO:0010507 | negative regulation of autophagy | LRP2 CYP3A4 | $6.581 \times 10^{-3}$ | |
| GO:CC | GO:0005892 | acetylcholine-gated channel complex | CHRNA9 CHRNA4 | | $1.937 \times 10^{-4}$ |
| KEGG | KEGG:02010 | ABC transporters | ABCA5 ABCG8 | $3.295 \times 10^{-4}$ | |
| KEGG | KEGG:00590 | Arachidonic acid metabolism | CYP2U1 CYP4F2 | | $3.046 \times 10^{-2}$ |

Table 11: Gene Ontology (GO) Molecular Function (GO: MF), Biological Processes (GO: BP), Cellular Component (GO: CC), Reactome (REAC), and Protein databases (CORUM) enrichment analysis of the comparative output analysis of North vs South in acute experiments, using g: Profiler.

| **Source** | **Term ID** | **Term Name** | **Gene Name** | **padj (up-regulated)** | **padj (down-regulated)** |
|---|---|---|---|---|---|
| GO:MF | GO:0003756 | protein disulfide isomerase activity | PDIA6 PDIA4 | $8.697 \times 10^{-3}$ | |
| GO:MF | GO:0016864 | intramolecular oxidoreductase activity, transposing S-S bonds | PDIA6 PDIA4 | $8.697 \times 10^{-3}$ | |
| GO:MF | GO:0016701 | oxidoreductase activity, acting on single donors with incorporation of molecular oxygen | CDO1 ALOX12B | $3.441 \times 10^{-2}$ | |
| GO:BP | GO:0000098 | sulfur amino acid acting on single catabolic process | CDO1 CSAD | $1.043 \times 10^{-3}$ | |
| GO:BP | GO:0034975 | protein folding acting on single in endoplasmic reticulum | HSP90B1 HSPA5 | $1.637 \times 10^{-5}$ | |
| GO:BP | GO:0000096 | sulfur amino acid metabolic process | CDO1 CSAD | $7.382 \times 10^{-3}$ | |
| GO:BP | GO:0039531 | regulation of viral-induced cytoplasmic pattern recognition signaling pathway | BIRC2 TRIM3 | | $2.889 \times 10^{-2}$ |
| GO:CC | GO:0035354 | Toll-like receptor 1-Toll-like receptor 2 protein complex | TLR1 TLR2 | | $9.438 \times 10^{-3}$ |
| GO:CC | GO:0033018 | sarcoplasmic reticulum lumen | HSP90B1 MANF | $1.366 \times 10^{-2}$ | |
| GO:CC | GO:0034663 | endoplasmic reticulum chaperone complex | HSP90B1 HYOU1 | $2.579 \times 10^{-6}$ | |
| GO:CC | GO:0005614 | interstitial matrix | ADAMTSL4 VIT | | $1.123 \times 10^{-2}$ |
| GO:CC | GO:0071682 | endocytic vesicle lumen | HSP90B1 HYOU1 | $6.601 \times 10^{-3}$ | |
| GO:CC | GO:0005790 | smooth endoplasmic reticulum | ERP29 HSP90B1 | $3.111 \times 10^{-5}$ | |
| REAC | R-HSA-381183 | ATF6 (ATF6-alpha) activates chaperone gene | HSP90B1 HSPA5 | $3.787 \times 10^{-2}$ | |
| REAC | R-HSA-381038 | XBP1(S) activates chaperone genes | HYOU1 PDIA6 | $4.978 \times 10^{-2}$ | |
| REAC | R-HSA-381070 | IRE1alpha activates chaperones | HYOU1 PDIA6 | $4.334 \times 10^{-3}$ | |
| CORUM | CORUM:7048 | TLR1-TLR2 complex | TLR1 TLR2 | | $1.664 \times 10^{-2}$ |

# Chapter 5

# Conclusions and Future Work

The present work aimed to determine and analyse the gene expression patterns in *U. delphinus* Spengler, 1793 under the effects of climate change. To accomplish this objective, a comprehensive bioinformatic pipeline focused on transcriptomics analysis was developed to generate the differential gene expression profiles of these individuals under thermal stress.

First, a complete pipeline to analyze RNA-Seq data was implemented. The pipeline initiates with a thorough data quality assessment, followed by the alignment of cleaned reads to a reference genome (*U. delphinus* genome). Subsequently, the gene expression levels were quantified and tested for differential gene expression. Ultimately, the results were consolidated into GO terms. The final output of this pipeline is a table for each comparative analysis, highlighting differentially active pathways among the processed RNA-Seq samples. This work differs from the past studies, by developing an R code compilation of three methods, EdgeR, Limma, and DESeq2 for gene expression analysis. The combination of the three methods is a new, and in our opinion creative, way to study differential gene expression in freshwater mussels, diverging from the traditional methodologies that have been employed in past studies. The present work fills this gap by investigating freshwater mussel gene expression patterns using the three methods at the same time. One of the advantages of this approach was demonstrated in the results of the comparative analysis of North 20 °C ∩ 25°C ∩ 30°C and South 20 °C ∩ 25°C ∩ 30°C. As stated, a limitation in the use of DESeq2 was found during these analyses, contrasting with the performance of the other two methods. The detection of this limitation was possible by employing all three methods since the comparative assessment of DGEs revealed a significantly lower count associated with the DESeq2 method. In opposition limma and EdgeR generate a higher number of DGEs in this specific analysis. Apart from this observation, the rest of the analysis demonstrated to produce good and reliable results from this approach.

Due to the lack of information about the effects of the development of climate change on freshwater mussels, the present work demonstrated to be a great advance in this field. The importance of the findings of this work is further highlighted by the absence of the gene profiles found in the literature relative to

freshwater mussels exposed to thermal stress. Just a few genes were found in the literature. Nevertheless, it was possible to extract meaningful biological considerations from the present work. Namely, the results demonstrate a correlation between water temperature and gene expression divergence, with a notable increase in the number of differentially expressed genes as temperatures rise. This trend, observed in chronic experiments comparing North vs South populations at three different temperatures, suggests a heightened genetic response to environmental stimuli in warmer conditions. In acute experiments, the southern population emerges as more resilient to abrupt increases in water temperature. The findings indicate that the southern population exhibits enhanced capabilities in protein production and response to oxidative stress. Moreover, regardless of water temperature, differences in the transport of proteins and inflammatory responses persist between the northern and southern populations. The adaptive strategies of southern populations, such as increased protein production, become more pronounced with rising temperatures.

Despite the results obtained, some future work from the outputs of this work can be done, including:

- Integrate proteomics or metabolomics data to provide a more comprehensive understanding of molecular mechanisms.

- Create ecological models using gene expression data to predict the distribution of mussels in the future under different scenarios of climate change.

- Develop and evaluate conservation plans, like as aided migration and habitat restoration, based on gene expression data.

Overall, these findings here reported shed light on the intricate biological processes underlying the adaptation of freshwater mussels to environmental disturbances. The present work reveals details about part of the evolutionary history of *U. delphinus* populations and holds implications for future biodiversity conservation efforts. The knowledge found not only makes it possible to make accurate decisions but also emphasizes how important it is to understand how genetic responses face changing environmental conditions, with bioinformatics being a vital tool for deciphering these complex biological reactions.

# Bibliography


[1] J. Heino, R. Virkkala, and H. Toivonen, "Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions," *Biological Reviews*, vol. 84, no. 1, pp. 39–54, 2009.

[2] D. L. Strayer, "Understanding how nutrient cycles and freshwater mussels (unionoida) affect one another," *Hydrobiologia*, vol. 735, pp. 277–292, 2014.

[3] C. L. Atkinson, H. M. Halvorson, K. A. Kuehn, M. Winebarger, A. Hamid, and M. N. Waters, "Filter-feeders have differential bottom-up impacts on green and brown food webs," *Oecologia*, vol. 195, pp. 187–198, 2021.

[4] M. Ilarri, A. Souza, L. Amorim, and R. Sousa, "Decay and persistence of empty bivalve shells in a temperate riverine system," *Science of the Total Environment*, vol. 683, pp. 185–192, 2019.

[5] J. Pander and J. Geist, "Ecological indicators for stream restoration success," *Ecological indicators*, vol. 30, pp. 106–118, 2013.

[6] F. Nicolini, F. Ghiselli, A. Luchetti, and L. Milani, "Bivalves as emerging model systems to study the mechanisms and evolution of sex determination: a genomic point of view," *Genome Biology and Evolution*, vol. 15, no. 10, p. evad181, 2023.

[7] H. M. Pereira, P. W. Leadley, V. Proença, R. Alkemade, J. P. Scharlemann, J. F. Fernandez-Manjarrés, M. B. Araújo, P. Balvanera, R. Biggs, W. W. Cheung *et al.*, "Scenarios for global biodiversity in the 21st century," *Science*, vol. 330, no. 6010, pp. 1496–1501, 2010.

[8] F. W. Allendorf, P. A. Hohenlohe, and G. Luikart, "Genomics and the future of conservation genetics," *Nature reviews genetics*, vol. 11, no. 10, pp. 697–709, 2010.

[9] G. Formenti, K. Theissinger, C. Fernandes, I. Bista, A. Bombarely, C. Bleidorn, C. Ciofi, A. Crottini, J. A. Godoy, J. Höglund *et al.*, "The era of reference genomes in conservation genomics," *Trends in Ecology & Evolution*, vol. 37, no. 3, pp. 197–202, 2022.

[10] A. Gomes-dos Santos, A. M. Machado, L. F. C. Castro, V. Prié, A. Teixeira, M. Lopes-Lima, and E. Froufe, "The gill transcriptome of threatened european freshwater mussels," *Scientific Data*, vol. 9, no. 1, p. 494, 2022.

[11] K. Sowndarya, "Biotechnology: Transcriptomics."

[12] D. Skerrett-Byrne Anthony, C. Jiang Chen, B. Nixon, and H. Hondermarck, "Transcriptomics," in *Encyclopedia of Cell Biology (Second Edition)*, second edition ed., R. A. Bradshaw, G. W. Hart, and P. D. Stahl, Eds. Oxford: Academic Press, 2023, pp. 363–371. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128216187001577

[13] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[14] J. Lindberg and J. Lundeberg, "The plasticity of the mammalian transcriptome," *Genomics*, vol. 95, no. 1, pp. 1–6, 2010.

[15] X. Qian, Y. Ba, Q. Zhuang, and G. Zhong, "Rna-seq technology and its application in fish transcriptomics," *Omics: a journal of integrative biology*, vol. 18, no. 2, pp. 98–110, 2014.

[16] V. Costa, C. Angelini, I. De Feis, A. Ciccodicola *et al.*, "Uncovering the complexity of transcriptomes with rna-seq," *BioMed Research International*, vol. 2010, 2010.

[17] Y. Ruan, P. Le Ber, H. H. Ng, and E. T. Liu, "Interrogating the transcriptome," *Trends in biotechnology*, vol. 22, no. 1, pp. 23–30, 2004.

[18] E. E. Hilt and P. Ferrieri, "Next generation and other sequencing technologies in diagnostic microbiology and infectious diseases," *Genes*, vol. 13, no. 9, p. 1566, 2022.

[19] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[20] A. Gondane and H. M. Itkonen, "Revealing the history and mystery of rna-seq," *Current Issues in Molecular Biology*, vol. 45, no. 3, pp. 1860–1874, 2023.

[21] R. Olby, "Quiet debut for the double helix," *Nature*, vol. 421, no. 6921, pp. 402–405, 2003.

[22] E. Y. Chan, "Advances in sequencing technology," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 573, no. 1-2, pp. 13–40, 2005.

[23] K. Light, J. Brooks, and Y.-S. Choi, "Manipulative-based activity using pop beads for demonstration of sanger sequencing," *Journal of chemical education*, vol. 100, no. 8, pp. 3138–3143, 2023.

[24] S. Kalayinia, H. Goodarzynejad, M. Maleki, and N. Mahdieh, "Next generation sequencing applications for cardiovascular disease," *Annals of Medicine*, vol. 50, no. 2, pp. 91–109, 2018.

[25] A. Totomoch-Serra, M. F. Marquez, and D. E. Cervantes-Barragán, "Sanger sequencing as a first-line approach for molecular diagnosis of andersen-tawil syndrome," *F1000Research*, vol. 6, 2017.

[26] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy *et al.*, "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome biology*, vol. 10, no. 3, pp. 1–13, 2009.

[27] M. Jamal, S. Raza, S. M. A. U. S. Bukhari, S. Andleeb, M. A. Nawaz, S. Pervez, L. Shah, and Redaina, "First generation–the sanger shotgun approach," in *Sequencing Technologies in Microbial Food Safety and Quality*. CRC Press, 2021, pp. 287–296.

[28] A. M. Maxam and W. Gilbert, "A new method for sequencing dna." *Proceedings of the National Academy of Sciences*, vol. 74, no. 2, pp. 560–564, 1977.

[29] T. P. Niedringhaus, D. Milanova, M. B. Kerby, M. P. Snyder, and A. E. Barron, "Landscape of next-generation sequencing technologies," *Analytical chemistry*, vol. 83, no. 12, pp. 4327–4341, 2011.

[30] M. Verma, S. Kulshrestha, and A. Puri, "Genome sequencing," *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, pp. 3–33, 2017.

[31] Z. Dong and Y. Chen, "Transcriptomics: advances and approaches," *Science China Life Sciences*, vol. 56, pp. 960–967, 2013.

[32] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro Jr, "Human genome project," *The American journal of surgery*, vol. 165, no. 2, pp. 258–264, 1993.

[33] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, members of the DOE, and N. planning groups, "New goals for the us human genome project: 1998-2003," *science*, vol. 282, no. 5389, pp. 682–689, 1998.

[34] T. Lenoir and E. Giannella, "The emergence and diffusion of dna microarray technology," *Journal of biomedical discovery and collaboration*, vol. 1, no. 1, pp. 1–39, 2006.

[35] N. L. Van Hal, O. Vorst, A. M. Van Houwelingen, E. J. Kok, A. Peijnenburg, A. Aharoni, A. J. Van Tunen, and J. Keijer, "The application of dna microarrays in gene expression analysis," *Journal of Biotechnology*, vol. 78, no. 3, pp. 271–280, 2000.

[36] P. O. Brown and D. Botstein, "Exploring the new world of the genome with dna microarrays," *Nature genetics*, vol. 21, no. 1, pp. 33–37, 1999.

[37] R. Bumgarner, "Overview of dna microarrays: types, applications, and their future," *Current protocols in molecular biology*, vol. 101, no. 1, pp. 22–1, 2013.

[38] M. Kurella, L.-L. Hsiao, T. Yoshida, J. D. Randall, G. Chow, S. S. Sarang, R. V. Jensen, and S. R. Gullans, "Dna microarray analysis of complex biologic processes," *Journal of the American Society of Nephrology*, vol. 12, no. 5, pp. 1072–1078, 2001.

[39] C. S. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," *Journal of applied genetics*, vol. 52, pp. 413–435, 2011.

[40] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das *et al.*, "Next-generation sequencing technology: Current trends and advancements," *Biology*, vol. 12, no. 7, p. 997, 2023.

[41] E. Dias-Neto, D. N. Nunes, R. J. Giordano, J. Sun, G. H. Botz, K. Yang, J. C. Setubal, R. Pasqualini, and W. Arap, "Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis," *PloS one*, vol. 4, no. 12, p. e8338, 2009.

[42] S. Marguerat and J. Bähler, "Rna-seq: from technology to biology," *Cellular and molecular life sciences*, vol. 67, pp. 569–579, 2010.

[43] M. Jain, "Next-generation sequencing technologies for gene expression profiling in plants," *Briefings in functional genomics*, vol. 11, no. 1, pp. 63–70, 2012.

[44] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing," *Human molecular genetics*, vol. 19, no. R2, pp. R227–R240, 2010.

[45] G. M. V. De Siqueira, F. M. Pereira-dos Santos, R. Silva-Rocha, and M.-E. Guazzaroni, "Nanopore sequencing provides rapid and reliable insight into microbial profiles of intensive care units," *Frontiers in Public Health*, vol. 9, p. 710985, 2021.

[46] H. Lv, F.-Y. Dao, D. Zhang, H. Yang, and H. Lin, "Advances in mapping the epigenetic modifications of 5-methylcytosine (5mc), n6-methyladenine (6ma), and n4-methylcytosine (4mc)," *Biotechnology and Bioengineering*, vol. 118, no. 11, pp. 4204–4216, 2021.

[47] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[48] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[49] A. Grada and K. Weinbrecht, "Next-generation sequencing: methodology and application," *Journal of Investigative Dermatology*, vol. 133, no. 8, pp. 1–4, 2013.

[50] B. Bruijns, R. Tiggelaar, and H. Gardeniers, "Massively parallel sequencing techniques for forensics: A review," *Electrophoresis*, vol. 39, no. 21, pp. 2642–2654, 2018.

[51] S. E. Levy and R. M. Myers, "Advancements in next-generation sequencing," *Annual review of genomics and human genetics*, vol. 17, pp. 95–115, 2016.

[52] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, "Overview of next-generation sequencing technologies," *Current protocols in molecular biology*, vol. 122, no. 1, p. e59, 2018.

[53] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current opinion in biotechnology*, vol. 24, no. 1, pp. 22–30, 2013.

[54] M. E. McCue and A. M. McCoy, "The scope of big data in one medicine: unprecedented opportunities and challenges," *Frontiers in veterinary science*, vol. 4, p. 194, 2017.

[55] S. J. Emrich, W. B. Barbazuk, L. Li, and P. S. Schnable, "Gene discovery and annotation using lcm-454 transcriptome sequencing," *Genome research*, vol. 17, no. 1, pp. 69–73, 2007.

[56] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra, "Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing," *Biotechniques*, vol. 45, no. 1, pp. 81–94, 2008.

[57] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel *et al.*, "Stem cell transcriptome profiling via massive-scale mrna sequencing," *Nature methods*, vol. 5, no. 7, pp. 613–619, 2008.

[58] B. T. Wilhelm and J.-R. Landry, "Rna-seq—quantitative measurement of expression through massively parallel rna-sequencing," *Methods*, vol. 48, no. 3, pp. 249–257, 2009.

[59] T. Hu, N. Chitnis, D. Monos, and A. Dinh, "Next-generation sequencing technologies: An overview," *Human Immunology*, vol. 82, no. 11, pp. 801–811, 2021.

[60] J. F. Hess, T. Kohl, M. Kotrová, K. Rönsch, T. Paprotka, V. Mohr, T. Hutzenlaub, M. Brüggemann, R. Zengerle, S. Niemann *et al.*, "Library preparation for next generation sequencing: A review of automation strategies," *Biotechnology advances*, vol. 41, p. 107537, 2020.

[61] J. Dalal, E. Land, N. Vasani, L. He, C. Smith, M. Rodriguez-Welsh, I. Y. Perera, and H. Sederoff, "Methods for rna profiling of gravi-responding plant tissues," *Plant gravitropism: methods and protocols*, pp. 91–117, 2015.

[62] R. J. Mackenzie, "Rna-seq: Basics, applications and protocol," *Retrieved from Technology Networks: https://www. technologynetworks. com/genomics/articles/rna-seqbasics-applications-and-protocol-299461*, 2018.

[63] H. Jiang, R. Lei, S.-W. Ding, and S. Zhu, "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads," *BMC bioinformatics*, vol. 15, pp. 1–12, 2014.

[64] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 671–682, 2011.

[65] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Unicycler: resolving bacterial genome assemblies from short and long sequencing reads," *PLoS computational biology*, vol. 13, no. 6, p. e1005595, 2017.

[66] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu *et al.*, "Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph," *Briefings in functional genomics*, vol. 11, no. 1, pp. 25–37, 2012.

[67] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[68] C. A. Fassbinder-Orth, "Methods for quantifying gene expression in ecoimmunology: from qpcr to rna-seq," *Integrative and Comparative Biology*, vol. 54, no. 3, pp. 396–406, 2014.

[69] P. Guan and W.-K. Sung, "Structural variation detection using next-generation sequencing data: a comparative technical review," *Methods*, vol. 102, pp. 36–49, 2016.

[70] S. Taylor, M. Wakem, G. Dijkman, M. Alsarraj, and M. Nguyen, "A practical approach to rt-qpcr—publishing data that conform to the miqe guidelines," *Methods*, vol. 50, no. 4, pp. S1–S5, 2010.

[71] S. Bruinsma, J. Burgess, D. Schlingman, A. Czyz, N. Morrell, C. Ballenger, H. Meinholz, L. Brady, A. Khanna, L. Freeberg *et al.*, "Bead-linked transposomes enable a normalization-free workflow for ngs library preparation," *BMC genomics*, vol. 19, no. 1, pp. 1–16, 2018.

[72] S. R. Head, H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, and P. Ordoukhanian, "Library construction for next-generation sequencing: overviews and challenges," *Biotechniques*, vol. 56, no. 2, pp. 61–77, 2014.

[73] D. D. Luan and T. H. Eickbush, "Rna template requirements for target dna-primed reverse transcription by the r2 retrotransposable element," *Molecular and cellular biology*, vol. 15, no. 7, pp. 3882–3891, 1995.

[74] E. L. Van Dijk, Y. Jaszczyszyn, and C. Thermes, "Library preparation methods for next-generation sequencing: tone down the bias," *Experimental cell research*, vol. 322, no. 1, pp. 12–20, 2014.

[75] R. Hrdlickova, M. Toloue, and B. Tian, "Rna-seq methods for transcriptome analysis," *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 1, p. e1364, 2017.

[76] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, and Y. Wang, "Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes," *BMC genomics*, vol. 16, no. 1, pp. 1–12, 2015.

[77] D. Duncan and N. Patel, "Next-generation sequencing in the clinical laboratory," in *Diagnostic Molecular Pathology*. Elsevier, 2017, pp. 25–33.

[78] R. A. White, P. C. Blainey, H. C. Fan, and S. R. Quake, "Digital pcr provides sensitive and absolute calibration for high throughput sequencing," *BMC genomics*, vol. 10, no. 1, pp. 1–12, 2009.

[79] J. Guo, L. Yu, N. J. Turro, and J. Ju, "An integrated system for dna sequencing by synthesis using novel nucleotide analogues," *Accounts of chemical research*, vol. 43, no. 4, pp. 551–563, 2010.

[80] G. Turcatti, A. Romieu, M. Fedurco, and A.-P. Tairi, "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for dna sequencing by synthesis," *Nucleic acids research*, vol. 36, no. 4, pp. e25–e25, 2008.

[81] R. K. Patel and M. Jain, "Ngs qc toolkit: a toolkit for quality control of next generation sequencing data," *PloS one*, vol. 7, no. 2, p. e30619, 2012.

[82] C. Ledergerber and C. Dessimoz, "Base-calling for next-generation sequencing platforms," *Briefings in bioinformatics*, vol. 12, no. 5, pp. 489–497, 2011.

[83] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen, "Performance comparison of benchtop high-throughput sequencing platforms," *Nature biotechnology*, vol. 30, no. 5, pp. 434–439, 2012.

[84] V. Bhola, A. S. Bopardikar, R. Narayanan, K. Lee, and T. Ahn, "No-reference compression of genomic data stored in fastq format," in *2011 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2011, pp. 147–150.

[85] T. Carver, U. Böhme, T. D. Otto, J. Parkhill, and M. Berriman, "Bamview: viewing mapped read alignment data in the context of the reference sequence," *Bioinformatics*, vol. 26, no. 5, pp. 676–677, 2010.

[86] S. S. Ajay, S. C. Parker, H. O. Abaan, K. V. F. Fajardo, and E. H. Margulies, "Accurate and comprehensive sequencing of personal genomes," *Genome research*, vol. 21, no. 9, pp. 1498–1505, 2011.

[87] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, and E. Lyon, "Acmg clinical laboratory standards for next-generation sequencing," *Genetics in medicine*, vol. 15, no. 9, pp. 733–747, 2013.

[88] B. Schmidt and A. Hildebrandt, "Next-generation sequencing: big data meets high performance computing," *Drug discovery today*, vol. 22, no. 4, pp. 712–717, 2017.

[89] T. Kwon, W. G. Yoo, W.-J. Lee, W. Kim, and D.-W. Kim, "Next-generation sequencing data analysis on cloud computing," *Genes & Genomics*, vol. 37, pp. 489–501, 2015.

[90] N. T. Tran, Z.-X. Gao, H.-H. Zhao, S.-K. Yi, B.-X. Chen, Y.-H. Zhao, L. Lin, X.-Q. Liu, and W.-M. Wang, "Transcriptome analysis and microsatellite discovery in the blunt snout bream (megalobrama

amblycephala) after challenge with aeromonas hydrophila," *Fish & shellfish immunology*, vol. 45, no. 1, pp. 72–82, 2015.

[91] I. Roznere, B. T. Sinn, and G. T. Watters, "The amblema plicata transcriptome as a resource to assess environmental impacts on freshwater mussels," *Freshwater Mollusk Biology and Conservation*, vol. 21, no. 2, pp. 57–64, 2018.

[92] L. Solieri, T. C. Dakal, and P. Giudici, "Next-generation sequencing and its potential impact on food microbial genomics," *Annals of Microbiology*, vol. 63, no. 1, pp. 21–37, 2013.

[93] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang *et al.*, "A survey of best practices for rna-seq data analysis," *Genome biology*, vol. 17, no. 1, pp. 1–19, 2016.

[94] Z. Jiang, X. Zhou, R. Li, J. J. Michal, S. Zhang, M. V. Dodson, Z. Zhang, and R. M. Harland, "Whole transcriptome analysis with sequencing: methods, challenges and potential solutions," *Cellular and molecular life sciences*, vol. 72, pp. 3425–3439, 2015.

[95] U. Nagalakshmi, K. Waern, and M. Snyder, "Rna-seq: a method for comprehensive transcriptome analysis," *Current protocols in molecular biology*, vol. 89, no. 1, pp. 4–11, 2010.

[96] D. Kumar, G. Bansal, A. Narang, T. Basak, T. Abbas, and D. Dash, "Integrating transcriptome and proteome profiling: Strategies and applications," *Proteomics*, vol. 16, no. 19, pp. 2533–2544, 2016.

[97] J. B. Wolf, "Principles of transcriptome analysis and gene expression quantification: an rna-seq tutorial," *Molecular ecology resources*, vol. 13, no. 4, pp. 559–572, 2013.

[98] J. C. Vera, C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden, "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing," *Molecular ecology*, vol. 17, no. 7, pp. 1636–1647, 2008.

[99] J. Shendure, "The beginning of the end for microarrays?" *Nature methods*, vol. 5, no. 7, pp. 585–587, 2008.

[100] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, pp. 621–628, 2008.

[101] P. A. McGettigan, "Transcriptomics in the rna-seq era," *Current opinion in chemical biology*, vol. 17, no. 1, pp. 4–11, 2013.

[102] F. Ozsolak and P. M. Milos, "Rna sequencing: advances, challenges and opportunities," *Nature reviews genetics*, vol. 12, no. 2, pp. 87–98, 2011.

[103] Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou, "Advanced applications of rna sequencing and challenges," *Bioinformatics and biology insights*, vol. 9, pp. BBI–S28 991, 2015.

[104] S. Li, S. W. Tighe, C. M. Nicolet, D. Grove, S. Levy, W. Farmerie, A. Viale, C. Wright, P. A. Schweitzer, Y. Gao *et al.*, "Multi-platform assessment of transcriptome profiling using rna-seq in the abrf next-generation sequencing study," *Nature biotechnology*, vol. 32, no. 9, pp. 915–925, 2014.

[105] A. Kumar, M. Kankainen, A. Parsons, O. Kallioniemi, P. Mattila, and C. A. Heckman, "The impact of rna sequence library construction protocols on transcriptomic profiling of leukemia," *BMC genomics*, vol. 18, pp. 1–13, 2017.

[106] C. P. Wigington, K. R. Williams, M. P. Meers, G. J. Bassell, and A. H. Corbett, "Poly (a) rna-binding proteins and polyadenosine rna: new members and novel functions," *Wiley Interdisciplinary Reviews: RNA*, vol. 5, no. 5, pp. 601–622, 2014.

[107] T. Raz, P. Kapranov, D. Lipson, S. Letovsky, P. M. Milos, and J. F. Thompson, "Protocol dependence of sequencing-based gene expression measurements," *PloS one*, vol. 6, no. 5, p. e19287, 2011.

[108] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, vol. 27, no. 6, pp. 863–864, 2011.

[109] J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev, "Comprehensive comparative analysis of strand-specific rna sequencing methods," *Nature methods*, vol. 7, no. 9, pp. 709–715, 2010.

[110] R. Tripathi, P. Sharma, P. Chakraborty, and P. K. Varadwaj, "Next-generation sequencing revolution through big data analytics," *Frontiers in life science*, vol. 9, no. 2, pp. 119–149, 2016.

[111] M. C. Schatz, "Cloudburst: highly sensitive read mapping with mapreduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[112] S. Pawar, A. Stanam, and Y. Zhu, "Evaluating the computing efficiencies (specificity and sensitivity) of graphics processing unit (gpu)-accelerated dna sequence alignment tools against central processing unit (cpu) alignment tool," *Journal of Bioinformatics and Sequence Analysis*, vol. 9, no. 2, pp. 10–14, 2018.

[113] U. D. J. G. I. H. T. . B. E. . P. P. . R. P. . W. S. . S. T. . D. N. . C. J.-F. . O. A. . L. S. . E. C. . U. E. . F. M. 4, R. G. S. C. S. Y. . F. A. . H. M. . Y. T. . T. A. . I. T. . K. C. . W. H. . T. Y. . T. T. 9, Genoscope, C. U.-. W. J. . H. R. . S. W. . A. F. . B. P. . B. T. . P. E. . R. C. . W. P. 10, I. o. M. B. R. A. . P. M. . N. G. . T. S. . R. A. . Department of Genome Analysis, G. S. C. S. D. R. . D.-S. L. . R. M. . W. K. . L. H. M. . D. J. 11, B. G. I. G. C. Y. H. . Y. J. . W. J. . H. G. . G. J. 15 *et al.*, "Initial sequencing and analysis of the human genome," *nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[114] I. Ulitsky, "Evolution to the rescue: using comparative genomics to understand long non-coding rnas," *Nature Reviews Genetics*, vol. 17, no. 10, pp. 601–614, 2016.

[115] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome biology*, vol. 14, no. 4, pp. 1–13, 2013.

[116] S. Huang, J. Zhang, R. Li, W. Zhang, Z. He, T.-W. Lam, Z. Peng, and S.-M. Yiu, "Soapsplice: genome-wide ab initio detection of splice junctions from rna-seq data," *Frontiers in genetics*, vol. 2, p. 46, 2011.

[117] W. J. Kent, "Blat—the blast-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.

[118] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[119] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[120] H. Lopez-Maestre, L. Brinza, C. Marchet, J. Kielbassa, S. Bastien, M. Boutigny, D. Monnin, A. E. Filali, C. M. Carareto, C. Vieira *et al.*, "Snp calling from rna-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence," *Nucleic Acids Research*, vol. 44, no. 19, pp. e148–e148, 2016.

[121] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, "Comparison of rna-seq by poly (a) capture, ribosomal rna depletion, and dna microarray for expression profiling," *BMC genomics*, vol. 15, no. 1, pp. 1–11, 2014.

[122] B. Tjaden, "De novo assembly of bacterial transcriptomes from rna-seq data," *Genome biology*, vol. 16, pp. 1–10, 2015.

[123] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber *et al.*, "De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis," *Nature protocols*, vol. 8, no. 8, pp. 1494–1512, 2013.

[124] I. Ochoa, M. Hernaez, and T. Weissman, "Aligned genomic data compression via improved modeling," *Journal of bioinformatics and computational biology*, vol. 12, no. 06, p. 1442002, 2014.

[125] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup, "The sequence alignment/map format and samtools," *bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[126] F. Finotello and B. Di Camillo, "Measuring differential gene expression with rna-seq: challenges and strategies for data analysis," *Briefings in functional genomics*, vol. 14, no. 2, pp. 130–142, 2015.

[127] B. T. Wilhelm, S. Marguerat, I. Goodhead, and J. Bähler, "Defining transcribed regions using rna-seq," *Nature protocols*, vol. 5, no. 2, pp. 255–266, 2010.

[128] S. Anders, P. T. Pyl, and W. Huber, "Htseq—a python framework to work with high-throughput sequencing data," *bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[129] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome biology*, vol. 11, no. 3, pp. 1–9, 2010.

[130] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[131] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for rna-seq data," *Genome biology*, vol. 14, no. 9, pp. 1–13, 2013.

[132] D. Seo, G. S. Ginsburg, and P. J. Goldschmidt-Clermont, "Gene expression analysis of cardiovascular diseases: novel insights into biology and clinical applications," *Journal of the American College of Cardiology*, vol. 48, no. 2, pp. 227–235, 2006.

[133] L. Ballarin, B. Rinkevich, K. Bartscherer, A. Burzynski, S. Cambier, M. Cammarata, I. Domart-Coulon, D. Drobne, J. Encinas, U. Frank *et al.*, "Maristem—stem cells of marine/aquatic invertebrates: From basic research to innovative applications," *Sustainability*, vol. 10, no. 2, p. 526, 2018.

[134] A. Fernandez-Marmiesse, S. Gouveia, and M. L. Couce, "Ngs technologies as a turning point in rare disease research, diagnosis and treatment," *Current medicinal chemistry*, vol. 25, no. 3, pp. 404–432, 2018.

[135] C. M. Sharma and J. Vogel, "Differential rna-seq: the approach behind and the biological insight gained," *Current opinion in microbiology*, vol. 19, pp. 97–105, 2014.

[136] V. Costa, M. Aprile, R. Esposito, and A. Ciccodicola, "Rna-seq and human complex diseases: recent accomplishments and future perspectives," *European Journal of Human Genetics*, vol. 21, no. 2, pp. 134–142, 2013.

[137] B. Weigelt, H. Horlings, B. Kreike, M. Hayes, M. Hauptmann, L. Wessels, D. De Jong, M. Van de Vijver, L. V. Veer, and J. Peterse, "Refinement of breast cancer classification by molecular characterization of histological special types," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 216, no. 2, pp. 141–150, 2008.

[138] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors *et al.*, "A pathology atlas of the human cancer transcriptome," *Science*, vol. 357, no. 6352, p. eaan2507, 2017.

[139] R. A. Neff, M. Wang, S. Vatansever, L. Guo, C. Ming, Q. Wang, E. Wang, E. Horgusluoglu-Moloch, W.-m. Song, A. Li *et al.*, "Molecular subtyping of alzheimer's disease using rna sequencing data reveals novel mechanisms and targets," *Science advances*, vol. 7, no. 2, p. eabb5398, 2021.

[140] F. Valdes-Mora, K. Handler, A. M. Law, R. Salomon, S. R. Oakes, C. J. Ormandy, and D. Gallego-Ortega, "Single-cell transcriptomics in cancer immunobiology: the future of precision oncology," *Frontiers in immunology*, vol. 9, p. 2582, 2018.

[141] X. Yang, L. Kui, M. Tang, D. Li, K. Wei, W. Chen, J. Miao, and Y. Dong, "High-throughput transcriptome profiling in drug and biomarker discovery," *Frontiers in genetics*, vol. 11, p. 19, 2020.

[142] M. Feder and J.-C. Walser, "The biological limitations of transcriptomics in elucidating stress and stress responses," *Journal of evolutionary biology*, vol. 18, no. 4, pp. 901–910, 2005.

[143] Y. H. Lee, M.-S. Kim, M. Wang, R. K. Bhandari, H. G. Park, R. S.-S. Wu, and J.-S. Lee, "Epigenetic plasticity enables copepods to cope with ocean acidification," *Nature Climate Change*, vol. 12, no. 10, pp. 918–927, 2022.

[144] K. Schirmer, B. B. Fischer, D. J. Madureira, and S. Pillai, "Transcriptomics in ecotoxicology," *Analytical and bioanalytical chemistry*, vol. 397, pp. 917–923, 2010.

[145] C. A. Logan and B. A. Buckley, "Transcriptomic responses to environmental temperature in eurythermal and stenothermal fishes," *The Journal of experimental biology*, vol. 218, no. 12, pp. 1915–1924, 2015.

[146] T. G. Evans and G. N. Somero, "A microarray-based transcriptomic time-course of hyper-and hypoosmotic stress signaling events in the euryhaline fish gillichthys mirabilis: osmosensors to effectors," *Journal of Experimental Biology*, vol. 211, no. 22, pp. 3636–3649, 2008.

[147] A. Y. Gracey, T.-H. Lee, R. M. Higashi, and T. Fan, "Hypoxia-induced mobilization of stored triglycerides in the euryoxic goby gillichthys mirabilis," *Journal of Experimental Biology*, vol. 214, no. 18, pp. 3005–3012, 2011.

[148] I. Benner, R. E. Diner, S. C. Lefebvre, D. Li, T. Komada, E. J. Carpenter, and J. H. Stillman, "Emiliania huxleyi increases calcification but not expression of calcification-related genes in long-term exposure to elevated temperature and p co2," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1627, p. 20130049, 2013.

[149] T. G. Evans and G. E. Hofmann, "Defining the limits of physiological plasticity: how gene expression can assess and predict the consequences of ocean change," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1596, pp. 1733–1745, 2012.

[150] S. U. Pauls, C. Nowak, M. Bálint, and M. Pfenninger, "The impact of global climate change on genetic diversity within populations and species," *Molecular ecology*, vol. 22, no. 4, pp. 925–946, 2013.

[151] H. J. Lee, A. Georgiadou, T. D. Otto, M. Levin, L. J. Coin, D. J. Conway, and A. J. Cunnington, "Transcriptomic studies of malaria: a paradigm for investigation of systemic host-pathogen interactions," *Microbiology and Molecular Biology Reviews*, vol. 82, no. 2, pp. 10–1128, 2018.

[152] Y. Tang, G. Xin, L.-M. Zhao, L.-X. Huang, Y.-X. Qin, Y.-Q. Su, W.-Q. Zheng, B. Wu, N. Lin, and Q.-P. Yan, "Novel insights into host-pathogen interactions of large yellow croakers (larimichthys crocea) and pathogenic bacterium pseudomonas plecoglossicida using time-resolved dual rna-seq of infected spleens," *Zoological Research*, vol. 41, no. 3, p. 314, 2020.

[153] J. Jeukens, L. Freschi, I. Kukavica-Ibrulj, J.-G. Emond-Rheault, N. P. Tucker, and R. C. Levesque, "Genomics of antibiotic-resistance prediction in pseudomonas aeruginosa," *Annals of the New York Academy of Sciences*, vol. 1435, no. 1, pp. 5–17, 2019.

[154] A. J. Westermann, S. A. Gorski, and J. Vogel, "Dual rna-seq of pathogen and host," *Nature Reviews Microbiology*, vol. 10, no. 9, pp. 618–630, 2012.

[155] A. J. Westermann, L. Barquist, and J. Vogel, "Resolving host–pathogen interactions by dual rna-seq," *PLoS pathogens*, vol. 13, no. 2, p. e1006033, 2017.

[156] G. Haseneyer, T. Schmutzer, M. Seidel, R. Zhou, M. Mascher, C.-C. Schön, S. Taudien, U. Scholz, N. Stein, K. F. Mayer *et al.*, "From rna-seq to large-scale genotyping-genomics resources for rye (secale cereale l.)," *BMC Plant Biology*, vol. 11, pp. 1–13, 2011.

[157] S. Saha, A. B. Sparks, C. Rago, V. Akmaev, C. J. Wang, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu, "Using the transcriptome to annotate the genome," *Nature biotechnology*, vol. 20, no. 5, pp. 508–512, 2002.

[158] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.

[159] A. E. Garcia-Vedrenne, M. Groner, A. Page-Karjian, G.-F. Siegmund, S. Singhal, J. Sziklay, and S. Roberts, "Development of genomic resources for a thraustochytrid pathogen and investigation of temperature influences on gene expression," *PLoS One*, vol. 8, no. 9, p. e74196, 2013.

[160] X. Shi, M. Sun, H. Liu, Y. Yao, and Y. Song, "Long non-coding rnas: a new frontier in the study of human diseases," *Cancer letters*, vol. 339, no. 2, pp. 159–166, 2013.

[161] B. Kamalidehghan, M. Habibi, S. S. Afjeh, M. Shoai, S. Alidoost, R. Almasi Ghale, N. Eshghifar, and F. Pouresmaeili, "The importance of small non-coding rnas in human reproduction: a review article," *The application of clinical genetics*, pp. 1–11, 2020.

[162] K. Grillone, C. Riillo, F. Scionti, R. Rocca, G. Tradigo, P. H. Guzzi, S. Alcaro, M. T. Di Martino, P. Tagliaferri, and P. Tassone, "Non-coding rnas in cancer: Platforms and strategies for investigating the genomic "dark matter"," *Journal of Experimental & Clinical Cancer Research*, vol. 39, no. 1, pp. 1–19, 2020.

[163] S. Jalali, G. G. Jayaraj, and V. Scaria, "Integrative transcriptome analysis suggest processing of a subset of long non-coding rnas to small rnas," *Biology direct*, vol. 7, pp. 1–13, 2012.

[164] M.-T. Melissari and P. Grote, "Roles for long non-coding rnas in physiology and disease," *Pflügers Archiv-European Journal of Physiology*, vol. 468, pp. 945–958, 2016.

[165] A. McDermaid, B. Monier, J. Zhao, B. Liu, and Q. Ma, "Interpretation of differential gene expression results of rna-seq data: review and integration," *Briefings in bioinformatics*, vol. 20, no. 6, pp. 2044–2054, 2019.

[166] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "Degseq: an r package for identifying differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–138, 2010.

[167] J. Westwood, "Using transcriptomics to study behavior," in *Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research*. Elsevier, 2018, pp. 267–288.

[168] L. Shiyi, W. Zitao, Z. Ronghui, W. Feiyan, C. Yanxiang, and L. Yeqiang, "Three differential expression analysis methods for rna sequencing: limma, edger, deseq2," 2021.

[169] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[170] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[171] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[172] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, 2004.

[173] F. Seyednasrollah, A. Laiho, and L. L. Elo, "Comparison of software packages for detecting differential expression in rna-seq studies," *Briefings in bioinformatics*, vol. 16, no. 1, pp. 59–70, 2015.

[174] S. Anders, "Huber w ((2010)) differential expression analysis for sequence count data," *Genome Biol*, vol. 11, no. 10, p. R106.

[175] T. Hardcastle, "und kelly, ka (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data," *BMC bioinformatics*, p. 422.

[176] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data," *Statistical methods in medical research*, vol. 22, no. 5, pp. 519–536, 2013.

[177] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with rna-seq," *Nature biotechnology*, vol. 31, no. 1, pp. 46–53, 2013.

[178] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski, "Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments," *Bioinformatics*, vol. 29, no. 8, pp. 1035–1043, 2013.

[179] F. Emmert-Streib and G. V. Glazko, "Pathway analysis of expression data: deciphering functional building blocks of complex diseases," *PLoS computational biology*, vol. 7, no. 5, p. e1002053, 2011.

[180] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[181] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "Kegg for integration and interpretation of large-scale molecular data sets," *Nucleic acids research*, vol. 40, no. D1, pp. D109–D114, 2012.

[182] "Molluscabase eds. (2023)," 2023, accessed on January 2nd, 2023. [Online]. Available: https://www.molluscabase.org/

[183] T. Regan, L. Stevens, C. Penaloza, R. D. Houston, D. Robledo, and T. P. Bean, "Ancestral physical stress and later immune gene family expansions shaped bivalve mollusc evolution," *Genome biology and evolution*, vol. 13, no. 8, p. evab177, 2021.

[184] L. Song, L. Wang, L. Qiu, and H. Zhang, "Bivalve immunity," *Invertebrate immunity*, pp. 44–65, 2010.

[185] D. L. Graf, "Patterns of freshwater bivalve global diversity and the state of phylogenetic studies on the unionoida, sphaeriidae, and cyrenidae," *American Malacological Bulletin*, vol. 31, no. 1, pp. 135–153, 2013.

[186] X.-C. Huang, J.-H. Su, J.-X. Ouyang, S. Ouyang, C.-H. Zhou, and X.-P. Wu, "Towards a global phylogeny of freshwater mussels (bivalvia: Unionida): Species delimitation of chinese taxa, mitochondrial phylogenomics, and diversification patterns," *Molecular Phylogenetics and Evolution*, vol. 130, pp. 45–59, 2019.

[187] C. C. Vaughn and C. C. Hakenkamp, "The functional role of burrowing bivalves in freshwater ecosystems," *Freshwater Biology*, vol. 46, no. 11, pp. 1431–1446, 2001.

[188] C. C. Vaughn and T. J. Hoellein, "Bivalve impacts in freshwater and marine ecosystems," *Annual Review of Ecology, Evolution, and Systematics*, vol. 49, pp. 183–208, 2018.

[189] T. L. Rocha, T. Gomes, V. S. Sousa, N. C. Mestre, and M. J. Bebianno, "Ecotoxicological impact of engineered nanomaterials in bivalve molluscs: an overview," *Marine environmental research*, vol. 111, pp. 74–88, 2015.

[190] C. L. Atkinson and C. C. Vaughn, "Biogeochemical hotspots: temporal and spatial scaling of the impact of freshwater mussels on ecosystem function," *Freshwater Biology*, vol. 60, no. 3, pp. 563–574, 2015.

[191] V. Modesto, M. Ilarri, A. T. Souza, M. Lopes-Lima, K. Douda, M. Clavero, and R. Sousa, "Fish and mussels: importance of fish for freshwater mussel conservation," *Fish and Fisheries*, vol. 19, no. 2, pp. 244–259, 2018.

[192] K. Douda, M. Lopes-Lima, M. Hinzmann, J. Machado, S. Varandas, A. Teixeira, and R. Sousa, "Biotic homogenization as a threat to native affiliate species: fish introductions dilute freshwater mussel's host resources," *Diversity and Distributions*, vol. 19, no. 8, pp. 933–942, 2013.

[193] A. Terui, Y. Miyazaki, A. Yoshioka, T. Kadoya, F. Jopp, and I. Washitani, "Dispersal of larvae of margaritifera laevis by its host fish," *Freshwater Science*, vol. 33, no. 1, pp. 112–123, 2014.

[194] P. Morais, M. M. Rufino, J. Reis, E. Dias, and R. Sousa, "Assessing the morphological variability of unio delphinus spengler, 1783 (bivalvia: Unionidae) using geometric morphometry," *Journal of Molluscan Studies*, vol. 80, no. 1, pp. 17–23, 2014.

[195] E. Froufe, D. V. Gonçalves, A. Teixeira, R. Sousa, S. Varandas, M. Ghamizi, A. Zieritz, and M. Lopes-Lima, "Who lives where? molecular and morphometric analyses clarify which unio species (unionida, mollusca) inhabit the southwestern palearctic," *Organisms Diversity & Evolution*, vol. 16, pp. 597–611, 2016.

[196] M. Jenkins, "Prospects for biodiversity," *Science*, vol. 302, no. 5648, pp. 1175–1177, 2003.

[197] N. Ferreira-Rodríguez, Y. B. Akiyama, O. V. Aksenova, R. Araujo, M. C. Barnhart, Y. V. Bespalaya, A. E. Bogan, I. N. Bolotov, P. B. Budha, C. Clavijo *et al.*, "Research priorities for freshwater mussel conservation assessment," *Biological Conservation*, vol. 231, pp. 77–87, 2019.

[198] R. Cazzolla Gatti, "Freshwater biodiversity: A review of local and global threats," *International Journal of Environmental Studies*, vol. 73, no. 6, pp. 887–904, 2016.

[199] J. Geist, "Integrative freshwater ecology and biodiversity conservation," *Ecological Indicators*, vol. 11, no. 6, pp. 1507–1516, 2011.

[200] G. Woodward, D. M. Perkins, and L. E. Brown, "Climate change and freshwater ecosystems: impacts across multiple levels of organization," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1549, pp. 2093–2106, 2010.

[201] N. Poff, M. M. Brinson, and J. Day, "Aquatic ecosystems and global climate change," *Pew Center on Global Climate Change, Arlington, VA*, vol. 44, pp. 1–36, 2002.

[202] K. Inoue and D. J. Berg, "Predicting the effects of climate change on population connectivity and genetic diversity of an imperiled freshwater mussel, cumberlandia monodonta (bivalvia: Margaritiferidae), in riverine systems," *Global change biology*, vol. 23, no. 1, pp. 94–107, 2017.

[203] M. E. Visser, "Keeping up with a warming world; assessing the rate of adaptation to climate change," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1635, pp. 649–659, 2008.

[204] C. Bellard, C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp, "Impacts of climate change on the future of biodiversity," *Ecology letters*, vol. 15, no. 4, pp. 365–377, 2012.

[205] J. S. Johnson, "Heat stress: impact on livestock well-being and productivity and mitigation strategies to alleviate the negative effects," *Animal Production Science*, vol. 58, no. 8, pp. 1404–1413, 2018.

[206] D. E. Spooner and C. C. Vaughn, "A trait-based approach to species' roles in stream ecosystems: climate change, community structure, and material cycling," *Oecologia*, vol. 158, pp. 307–317, 2008.

[207] H. Falfushynska, L. Gnatyshyna, I. Yurchak, A. Ivanina, O. Stoliar, and I. Sokolova, "Habitat pollution and thermal regime modify molecular stress responses to elevated temperature in freshwater mussels (anodonta anatina: Unionidae)," *Science of the Total Environment*, vol. 500, pp. 339–350, 2014.

[208] A. M. Ganser, T. J. Newton, and R. J. Haro, "Effects of elevated water temperature on physiological responses in adult freshwater mussels," *Freshwater Biology*, vol. 60, no. 8, pp. 1705–1716, 2015.

[209] K. Luck and J. D. Ackerman, "Threats to freshwater mussels: The interactions of water temperature, velocity and total suspended solids on ecophysiology and growth," *Science of The Total Environment*, vol. 821, p. 153101, 2022.

[210] A. M. Stevens and C. J. Gobler, "Interactive effects of acidification, hypoxia, and thermal stress on growth, respiration, and survival of four north atlantic bivalves," *Marine Ecology Progress Series*, vol. 604, pp. 143–161, 2018.

[211] M. P. Simmonds and W. J. Eliott, "Climate change and cetaceans: concerns and recent developments," *Journal of the Marine biological Association of the United Kingdom*, vol. 89, no. 1, pp. 203–210, 2009.

[212] A. Zgouridou, E. Tripidaki, I. A. Giantsis, J. A. Theodorou, M. Kalaitzidou, D. E. Raitsos, A. Lattos, A.-M. Mavropoulou, S. Sofianos, D. Karagiannis *et al.*, "The current situation and potential effects of climate change on the microbial load of marine bivalves of the greek coastlines: An integrative review," *Environmental Microbiology*, vol. 24, no. 3, pp. 1012–1034, 2022.

[213] R. Adrian, S. Wilhelm, and D. Gerten, "Life-history traits of lake plankton species may govern their phenological response to climate warming," *Global change biology*, vol. 12, no. 4, pp. 652–661, 2006.

[214] T. A. Blewett, S. A. Binning, A. M. Weinrauch, C. M. Ivy, G. S. Rossi, B. G. Borowiec, G. Y. Lau, S. L. Overduin, I. Aragao, and T. Norin, "Physiological and behavioural strategies of aquatic animals living in fluctuating environments," *Journal of Experimental Biology*, vol. 225, no. 9, p. jeb242503, 2022.

[215] B. Gallardo, A. E. Bogan, S. Harun, L. Jainih, M. Lopes-Lima, M. Pizarro, K. A. Rahim, R. Sousa, S. G. Virdis, and A. Zieritz, "Current and future effects of global change on a hotspot's freshwater diversity," *Science of the Total Environment*, vol. 635, pp. 750–760, 2018.

[216] L. A. Helfrich, R. J. Neves, and H. Chapman, "Sustaining america's aquatic biodiversity: freshwater mussel biodiversity and conservation," 2019.

[217] H. S. Galbraith and C. C. Vaughn, "Temperature and food interact to influence gamete development in freshwater mussels," *Hydrobiologia*, vol. 636, pp. 35–47, 2009.

[218] I. Ibáñez, J. S. Clark, M. C. Dietze, K. Feeley, M. Hersh, S. LaDeau, A. McBride, N. E. Welch, and M. S. Wolosin, "Predicting biodiversity change: outside the climate envelope, beyond the species–area curve," *Ecology*, vol. 87, no. 8, pp. 1896–1906, 2006.

[219] K. Hiscock, A. Southward, I. Tittley, and S. Hawkins, "Effects of changing temperature on benthic marine life in britain and ireland," *Aquatic Conservation: marine and freshwater Ecosystems*, vol. 14, no. 4, pp. 333–362, 2004.

[220] C. D. Harley, A. Randall Hughes, K. M. Hultgren, B. G. Miner, C. J. Sorte, C. S. Thornber, L. F. Rodriguez, L. Tomanek, and S. L. Williams, "The impacts of climate change in coastal marine systems," *Ecology letters*, vol. 9, no. 2, pp. 228–241, 2006.

[221] H. Tsujimoto, Q. Sohail, and Y. Matsuoka, "Broadening the genetic diversity of common and durum wheat for abiotic stress tolerance breeding," in *Advances in Wheat Genetics: From Genome to Field: Proceedings of the 12th International Wheat Genetics Symposium*. Springer Japan, 2015, pp. 233–238.

[222] A. S. Huffmyer, N. K. Bean, E. Majerová, C. I. Harris, and C. Drury, "Variable intraspecific genetic diversity effects impact thermal tolerance in a reef-building coral," *Coral Reefs*, pp. 1–11, 2022.

[223] P. Gillette, "Intraspecific genetic variability in temperature tolerance in the coral pocillopora damicornis: Effects on growth, photosynthesis and survival," Ph.D. dissertation, University of Miami, 2012.

[224] S. U. Franssen, J. Gu, N. Bergmann, G. Winters, U. C. Klostermeier, P. Rosenstiel, E. Bornberg-Bauer, and T. B. Reusch, "Transcriptomic resilience to global warming in the seagrass zostera marina, a marine foundation species," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19 276–19 281, 2011.

[225] M. Wang, *Novel Strategy for Disease Control of Rice and Examination of the Associated Microbiome*. North Carolina State University, 2019.

[226] A.-M. Waldvogel, B. Feldmeyer, G. Rolshausen, M. Exposito-Alonso, C. Rellstab, R. Kofler, T. Mock, K. Schmid, I. Schmitt, T. Bataillon *et al.*, "Evolutionary genomics can improve prediction of species' responses to climate change," *Evolution Letters*, vol. 4, no. 1, pp. 4–18, 2020.

[227] C. Schonewald-Cox, "Genetics, minimum population size, and the island preserve," *Hawaii's terrestrial ecosystems: preservation and managment. University of Hawaii, Honolulu*, pp. 432–458, 1985.

[228] R. D. Barrett and D. Schluter, "Adaptation from standing genetic variation," *Trends in ecology & evolution*, vol. 23, no. 1, pp. 38–44, 2008.

[229] M. C. Fitzpatrick and S. R. Keller, "Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation," *Ecology letters*, vol. 18, no. 1, pp. 1–16, 2015.

[230] C. H. Smith, "A high-quality reference genome for a parasitic bivalve with doubly uniparental inheritance (bivalvia: Unionida)," *Genome Biology and Evolution*, vol. 13, no. 3, p. evab029, 2021.

[231] A. Gomes-dos Santos, M. Lopes-Lima, A. M. Machado, A. Marcos Ramos, A. Usié, I. N. Bolotov, I. V. Vikhrev, S. Breton, L. F. C. Castro, R. R. da Fonseca *et al.*, "The crown pearl: a draft genome assembly of the european freshwater pearl mussel margaritifera margaritifera (linnaeus, 1758)," *DNA research*, vol. 28, no. 2, p. dsab002, 2021.

[232] A. Gomes-dos Santos, M. Lopes-Lima, A. M. Machado, A. Teixeira, L. F. C. Castro, and E. Froufe, "Pacbio hi-fi genome assembly of the iberian dolphin freshwater mussel unio delphinus spengler, 1793," *bioRxiv*, pp. 2023–01, 2023.

[233] R. L. Rogers, S. L. Grizzard, J. E. Titus-McQuillan, K. Bockrath, S. Patel, J. P. Wares, J. T. Garner, and C. C. Moore, "Gene family amplification facilitates adaptation in freshwater unionid bivalve megalonaias nervosa," *Molecular Ecology*, vol. 30, no. 5, pp. 1155–1173, 2021.

[234] S. Renaut, D. Guerra, W. R. Hoeh, D. T. Stewart, A. E. Bogan, F. Ghiselli, L. Milani, M. Passamonti, and S. Breton, "Genome survey of the freshwater mussel venustaconcha ellipsiformis (bivalvia: Unionida) using a hybrid de novo assembly approach," *Genome biology and evolution*, vol. 10, no. 7, pp. 1637–1646, 2018.

[235] Z. Bai, Y. Lu, H. Hu, Y. Yuan, X. Liu, G. Wang, D. Huang, Z. Wang, Y. Mao, H. Wang *et al.*, "Chromosome-level genome assembly of freshwater pearl mussel, hyriopsis cumingii, provides insights into outstanding biomineralization ability," *Authorea Preprints*, 2022.

[236] A. Gomes-dos Santos, M. Lopes-Lima, A. M. Machado, T. Forest, G. Achaz, A. Teixeira, V. Prié, L. F. C. Castro, and E. Froufe, "The crown pearl v2: an improved genome assembly of the european freshwater pearl mussel margaritifera margaritifera (linnaeus, 1758)," *GigaByte*, vol. 2023, 2023.

[237] R. Wang, C. Li, J. Stoeckel, G. Moyer, Z. Liu, and E. Peatman, "Rapid development of molecular resources for a freshwater mussel, villosa lienosa (bivalvia: Unionidae), using an rna-seq-based approach," *Freshwater Science*, vol. 31, no. 3, pp. 695–708, 2012.

[238] Y. Luo, C. Li, A. G. Landis, G. Wang, J. Stoeckel, and E. Peatman, "Transcriptomic profiling of differential responses to drought in two freshwater mussel species, the giant floater pyganodon grandis and the pondhorn uniomerus tetralasmus," *PLoS One*, vol. 9, no. 2, p. e89481, 2014.

[239] B. B. Patnaik, T. H. Wang, S. W. Kang, H.-J. Hwang, S. Y. Park, E. B. Park, J. M. Chung, D. K. Song, C. Kim, S. Kim *et al.*, "Sequencing, de novo assembly, and annotation of the transcriptome of the endangered freshwater pearl bivalve, cristaria plicata, provides novel insights into functional genes and marker discovery," *PLoS One*, vol. 11, no. 2, p. e0148622, 2016.

[240] X. Wang, Z. Liu, and W. Wu, "Transcriptome analysis of the freshwater pearl mussel (cristaria plicata) mantle unravels genes involved in the formation of shell and pearl," *Molecular genetics and genomics*, vol. 292, pp. 343–352, 2017.

[241] Q. Yang, K. Guo, X. Zhou, X. Tang, X. Yu, W. Yao, and Z. Wu, "Histopathology, antioxidant responses, transcriptome and gene expression analysis in triangle sail mussel hyriopsis cumingii after bacterial infection," *Developmental & Comparative Immunology*, vol. 124, p. 104175, 2021.

[242] A. Bertucci, F. Pierron, J. Thébault, C. Klopp, J. Bellec, P. Gonzalez, and M. Baudrimont, "Transcriptomic responses of the endangered freshwater mussel margaritifera margaritifera to trace metal

contamination in the dronne river, france," *Environmental Science and Pollution Research*, vol. 24, pp. 27 145–27 159, 2017.

[243] L. S. Robertson, H. S. Galbraith, D. Iwanowicz, C. J. Blakeslee, and R. S. Cornman, "Rna sequencing analysis of transcriptional change in the freshwater mussel elliptio complanata after environmentally relevant sodium chloride exposure," *Environmental toxicology and chemistry*, vol. 36, no. 9, pp. 2352–2366, 2017.

[244] C. Capt, S. Renaut, F. Ghiselli, L. Milani, N. A. Johnson, B. E. Sietman, D. T. Stewart, and S. Breton, "Deciphering the link between doubly uniparental inheritance of mtdna and sex determination in bivalves: clues from comparative transcriptomics," *Genome biology and evolution*, vol. 10, no. 2, pp. 577–590, 2018.

[245] D. Huang, J. Shen, J. Li, and Z. Bai, "Integrated transcriptome analysis of immunological responses in the pearl sac of the triangle sail mussel (hyriopsis cumingii) after mantle implantation," *Fish & Shellfish Immunology*, vol. 90, pp. 385–394, 2019.

[246] C. Capt, S. Renaut, D. T. Stewart, N. A. Johnson, and S. Breton, "Putative mitochondrial sex determination in the bivalvia: insights from a hybrid transcriptome assembly in freshwater mussels," *Frontiers in genetics*, vol. 10, p. 840, 2019.

[247] X. Chen, Z. Bai, and J. Li, "The mantle exosome and micrornas of hyriopsis cumingii involved in nacre color formation," *Marine Biotechnology*, vol. 21, pp. 634–642, 2019.

[248] R. S. Cornman, L. S. Robertson, H. Galbraith, and C. Blakeslee, "Transcriptomic analysis of the mussel elliptio complanata identifies candidate stress-response genes and an abundance of novel or noncoding transcripts," *PLoS One*, vol. 9, no. 11, p. e112420, 2014.

[249] I. Roznere, B. T. Sinn, M. Daly, and G. T. Watters, "Freshwater mussels (unionidae) brought into captivity exhibit up-regulation of genes involved in stress and energy metabolism," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[250] C. Georgopoulos and W. Welch, "Role of the major heat shock proteins as molecular chaperones," *Annual review of cell biology*, vol. 9, no. 1, pp. 601–634, 1993.

[251] W. Wang, B. Vinocur, O. Shoseyov, and A. Altman, "Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response," *Trends in plant science*, vol. 9, no. 5, pp. 244–252, 2004.

[252] B. M. Sanders, “Stress proteins in aquatic organisms: an environmental perspective,” *Critical reviews in Toxicology*, vol. 23, no. 1, pp. 49–75, 1993.

[253] E. De Nadal, G. Ammerer, and F. Posas, “Controlling gene expression in response to stress,” *Nature Reviews Genetics*, vol. 12, no. 12, pp. 833–845, 2011.

[254] H. Nakamoto and T. Hiyama, “Heat-shock proteins and temperature stress,” *Handbook of plant and crop stress. Marcel Dekker, New York*, pp. 399–416, 1999.

[255] X. Xia, S. Xue, X. Wang, Q. Zhang, C. Huang, L. Guo, and L. Yao, “Response a chronic effects of pbde-47: Up-regulations of hsp60 and hsp70 expression in freshwater bivalve anodonta woodiana,” *Fish & Shellfish Immunology*, vol. 65, pp. 213–225, 2017.

[256] H. S. Galbraith, C. J. Blakeslee, and W. A. Lellis, “Recent thermal history influences thermal tolerance in freshwater mussel species (bivalvia: Unionoida),” *Freshwater Science*, vol. 31, no. 1, pp. 83–92, 2012.

[257] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[258] C. W. Law, M. Alhamdoosh, S. Su, X. Dong, L. Tian, G. K. Smyth, and M. E. Ritchie, “Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger,” *F1000Research*, vol. 5, 2016.

[259] S. Liu, Z. Wang, R. Zhu, F. Wang, Y. Cheng, and Y. Liu, “Three differential expression analysis methods for rna sequencing: limma, edger, deseq2,” *JoVE (Journal of Visualized Experiments)*, no. 175, p. e62528, 2021.

[260] L. Kolberg, U. Raudvere, I. Kuzmin, P. Adler, J. Vilo, and H. Peterson, “g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update),” *Nucleic Acids Research*, p. gkad347, 2023.

[261] D. Barbará, C. Domeniconi, and J. P. Rogers, “Detecting outliers using transduction and statistical testing,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 55–64.

[262] A. Stupnikov, C. McInerney, K. Savage, S. McIntosh, F. Emmert-Streib, R. Kennedy, M. Salto-Tellez, K. Prise, and D. McArt, “Robustness of differential gene expression analysis of rna-seq,” *Computational and structural biotechnology journal*, vol. 19, pp. 3470–3481, 2021.

[263] L. Väremo, J. Nielsen, and I. Nookaew, "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods," *Nucleic acids research*, vol. 41, no. 8, pp. 4378–4391, 2013.

[264] L. A. Corchete, E. A. Rojas, D. Alonso-López, J. De Las Rivas, N. C. Gutiérrez, and F. J. Burguillo, "Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis," *Scientific reports*, vol. 10, no. 1, p. 19737, 2020.

[265] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[266] G. K. Smyth, N. Thorne, and J. Wettenhall, "Limma: linear models for microarray data user's guide," *Software manual available from http://www. bioconductor. org*, 2003.

[267] S. Li, P. P. Łabaj, P. Zumbo, P. Sykacek, W. Shi, L. Shi, J. Phan, P.-Y. Wu, M. Wang, C. Wang *et al.*, "Detecting and correcting systematic variation in large-scale rna sequencing data," *Nature biotechnology*, vol. 32, no. 9, pp. 888–895, 2014.

[268] D. J. Lorenz, R. S. Gill, R. Mitra, and S. Datta, "Using rna-seq data to detect differentially expressed genes," *Statistical analysis of next generation sequencing data*, pp. 25–49, 2014.

[269] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham *et al.*, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, pp. 1–18, 2017.

[270] T. Mou, W. Deng, F. Gu, Y. Pawitan, and T. N. Vu, "Reproducibility of methods to detect differentially expressed genes from single-cell rna sequencing," *Frontiers in genetics*, vol. 10, p. 1331, 2020.

[271] C. R. Yendrek, E. A. Ainsworth, and J. Thimmapuram, "The bench scientist's guide to statistical analysis of rna-seq data," *BMC research notes*, vol. 5, no. 1, pp. 1–10, 2012.

[272] C. Boscher, J. W. Dennis, and I. R. Nabi, "Glycosylation, galectins and cellular signaling," *Current opinion in cell biology*, vol. 23, no. 4, pp. 383–392, 2011.

[273] J. Gu, T. Isaji, Q. Xu, Y. Kariya, W. Gu, T. Fukuda, and Y. Du, "Potential roles of n-glycosylation in cell adhesion," *Glycoconjugate journal*, vol. 29, pp. 599–607, 2012.

[274] M. Kornprobst, M. Turk, N. Kellner, J. Cheng, D. Flemming, I. Koš-Braun, M. Koš, M. Thoms, O. Berninghausen, R. Beckmann *et al.*, "Architecture of the 90s pre-ribosome: a structural view on the birth of the eukaryotic ribosome," *Cell*, vol. 166, no. 2, pp. 380–393, 2016.

[275] R. MP, "Selenium and human health," *Lancet*, vol. 379, pp. 1256–1268, 2012.

[276] A. E. Stütz and T. M. Wrodnigg, "Imino sugars and glycosyl hydrolases: historical context, current aspects, emerging trends," *Advances in carbohydrate chemistry and biochemistry*, vol. 66, pp. 187–298, 2011.

[277] S. Muralidharan and P. Mandrekar, "Cellular stress response and innate immune signaling: integrating pathways in host defense and inflammation," *Journal of leukocyte biology*, vol. 94, no. 6, pp. 1167–1184, 2013.

[278] Z. Wu, Y. Yang, L. Zhou, C. Chi, X. Sun, B. Wu, Z. Liu, Y. Wang *et al.*, "Transcriptome sequencing reveals the response to acute thermal stress in the pacific abalone, haliotis discus hannai," *Aquaculture Research*, vol. 2023, 2023.

[279] C. H. Frie, "The structure and activity of the cilia in the gills of some fresh-water mussels," Ph.D. dissertation, Bowling Green State University, 1948.

[280] C. D. Wilson, G. Arnott, and R. W. Elwood, "Freshwater pearl mussels show plasticity of responses to different predation risks but also show consistent individual differences in responsiveness," *Behavioural processes*, vol. 89, no. 3, pp. 299–303, 2012.

[281] S. Benítez, Á. Figueroa, N. A. Lagos, A. X. Silva, C. Duarte, C. A. Vargas, M. A. Lardies, and L. Cárdenas, "Differential gene expression analysis in the scallop argopecten purpuratus exposed to altered ph and temperature conditions in an upwelling-influenced farming area," *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, vol. 45, p. 101046, 2023.

[282] N. G. Ertl, W. A. O'Connor, and A. Elizur, "Molecular effects of a variable environment on sydney rock oysters, saccostrea glomerata: Thermal and low salinity stress, and their synergistic effect," *Marine genomics*, vol. 43, pp. 19–32, 2019.

[283] J. T. Canty, R. Tan, E. Kusakci, J. Fernandes, and A. Yildiz, "Structure and mechanics of dynein motors," *Annual review of biophysics*, vol. 50, pp. 549–574, 2021.

[284] N. Pavin and I. M. Tolić-Nørrelykke, "Dynein, microtubule and cargo: a menage a trois," *Biochemical Society Transactions*, vol. 41, no. 6, pp. 1731–1735, 2013.

[285] Z. Hu, H. Song, J. Feng, C. Zhou, M.-J. Yang, P. Shi, Z.-L. Yu, Y.-R. Li, Y.-J. Guo, H.-Z. Li *et al.*, "Massive heat shock protein 70 genes expansion and transcriptional signatures uncover hard clam adaptations to heat and hypoxia," *Frontiers in Marine Science*, vol. 9, p. 898669, 2022.

[286] Y. Tan, Y.-X. Sun, Y.-J. Zhu, M.-L. Liao, and Y.-W. Dong, "The impacts of thermal heterogeneity across microhabitats on post-settlement selection of intertidal mussels," *Iscience*, 2023.

[287] Y. Maeda and T. Kinoshita, "The acidic environment of the golgi is critical for glycosylation and transport," *Methods in enzymology*, vol. 480, pp. 495–510, 2010.

[288] N. J. Barrett, J. Thyrring, E. M. Harper, M. K. Sejr, J. G. Sørensen, L. S. Peck, and M. S. Clark, "Molecular responses to thermal and osmotic stress in arctic intertidal mussels (mytilus edulis): the limits of resilience," *Genes*, vol. 13, no. 1, p. 155, 2022.

[289] T. D. Nguyen, M. E. Walker, J. M. Gardner, and V. Jiranek, "Appropriate vacuolar acidification in saccharomyces cerevisiae is associated with efficient high sugar fermentation," *Food microbiology*, vol. 70, pp. 262–268, 2018.

[290] K. S. Keuenhof, L. Larsson Berglund, S. Malmgren Hill, K. L. Schneider, P. O. Widlund, T. Nyström, and J. L. Höög, "Large organellar changes occur during mild heat shock in yeast," *Journal of cell science*, vol. 135, no. 5, p. jcs258325, 2022.

[291] D. J. Dietzen, "Amino acids, peptides, and proteins," in *Principles and Applications of Molecular Diagnostics*. Elsevier, 2018, pp. 345–380.

[292] G. N. Somero, "Proteins and temperature," *Annual review of physiology*, vol. 57, no. 1, pp. 43–68, 1995.

[293] Y. Duan, F. Li, Y. Li, Y. Tang, X. Kong, Z. Feng, T. G. Anthony, M. Watford, Y. Hou, G. Wu *et al.*, "The role of leucine and its metabolites in protein and energy metabolism," *Amino acids*, vol. 48, pp. 41–51, 2016.

[294] R. R. Mayer, J. H. Cherry, and D. Rhodes, "Effects of heat shock on amino acid metabolism of cowpea cells," *Plant physiology*, vol. 94, no. 2, pp. 796–810, 1990.

[295] C. A. Wagner, O. Devuyst, S. Bourgeois, and N. Mohebbi, "Regulated acid–base transport in the collecting duct," *Pflügers Archiv-European Journal of Physiology*, vol. 458, pp. 137–156, 2009.

[296] W. D. Nunn, "A molecular view of fatty acid catabolism in escherichia coli," *Microbiological reviews*, vol. 50, no. 2, pp. 179–192, 1986.

[297] M. Alonso-Trujillo, A.-B. Muñiz-González, and J.-L. Martínez-Guitarte, "Endosulfan exposure alters transcription of genes involved in the detoxification and stress responses in physella acuta," *Scientific Reports*, vol. 10, no. 1, p. 7847, 2020.

[298] M. Ibba and D. Söll, "Aminoacyl-trna synthesis," *Annual review of biochemistry*, vol. 69, no. 1, pp. 617–650, 2000.

[299] C. Lorenz, C. E. Lünse, and M. Mörl, "trna modifications: impact on structure and thermal adaptation," *Biomolecules*, vol. 7, no. 2, p. 35, 2017.

[300] L. Stransky, K. Cotter, and M. Forgac, "The function of v-atpases in cancer," *Physiological reviews*, vol. 96, no. 3, pp. 1071–1091, 2016.

[301] E. Chapman, M. D. Best, S. R. Hanson, and C.-H. Wong, "Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility," *Angewandte Chemie International Edition*, vol. 43, no. 27, pp. 3526–3548, 2004.

[302] Y. Feng, D. He, Z. Yao, and D. J. Klionsky, "The machinery of macroautophagy," *Cell research*, vol. 24, no. 1, pp. 24–41, 2014.

[303] D. C. Rees, E. Johnson, and O. Lewinson, "Abc transporters: the power to change," *Nature reviews Molecular cell biology*, vol. 10, no. 3, pp. 218–227, 2009.

[304] H. F. Gilbert, "[3] protein disulfide isomerase," in *Methods in enzymology*. Elsevier, 1998, vol. 290, pp. 26–50.

[305] X. Kong, H. Yao, J. Ren, W. Dai, Z. Lin, C. Li, and Y. Dong, "Pdia6 involves the thermal stress response of razor clam, sinonovacula constricta," *Fish & Shellfish Immunology*, vol. 131, pp. 766–774, 2022.

[306] I. Braakman and N. J. Bulleid, "Protein folding and modification in the mammalian endoplasmic reticulum," *Annual review of biochemistry*, vol. 80, pp. 71–99, 2011.

[307] F. Bultelle, I. Boutet, S. Devin, F. Caza, Y. St-Pierre, R. Péden, P. Brousseau, P. Chan, D. Vaudry, F. Le Foll *et al.*, "Molecular response of a sub-antarctic population of the blue mussel (mytilus edulis platensis) to a moderate thermal stress," *Marine Environmental Research*, vol. 169, p. 105393, 2021.

[308] X. Cao, "Self-regulation and cross-regulation of pattern-recognition receptor signalling in health and disease," *Nature Reviews Immunology*, vol. 16, no. 1, pp. 35–50, 2016.

[309] K. Takeda and S. Akira, "Toll-like receptors in innate immunity," *International immunology*, vol. 17, no. 1, pp. 1–14, 2005.

[310] M. Xu, J. Wu, D. Ge, C. Wu, C. Chi, Z. Lv, Z. Liao, and H. Liu, "A novel toll-like receptor from mytilus coruscus is induced in response to stress," *Fish & Shellfish Immunology*, vol. 78, pp. 331–337, 2018.

# Appendix A
# Supplementary material for chapter 4

## A.0.1 RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of the two experiments.

Figures 30 and 31 represent the RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of the acute and chronic experiments, respectively.

| Sample - North | 260 Raw | 280 Raw | 320 Raw | 260 | 280 | 260/280 | ng/µL |
|---|---|---|---|---|---|---|---|
| Gill 1 | 1.781 | 0.881 | 0.092 | 1.685 | 0.786 | 2.142 | 1347.733 |
| Gill 2 | 2.375 | 1.194 | 0.092 | 2.279 | 1.099 | 2.073 | 1822.933 |
| Gill 3 | 1.842 | 0.897 | 0.069 | 1.769 | 0.825 | 2.143 | 1414.933 |
| Gill 4 | 2.021 | 1.001 | 0.078 | 1.939 | 0.92 | 2.106 | 1550.933 |
| Gill 5 | 1.922 | 0.938 | 0.066 | 1.852 | 0.869 | 2.13 | 1481.333 |

| Sample - South | 260 Raw | 280 Raw | 320 Raw | 260 | 280 | 260/280 | ng/µL |
|---|---|---|---|---|---|---|---|
| Gill 1 | 1.717 | 0.868 | 0.113 | 1.6 | 0.753 | 2.126 | 1280.267 |
| Gill 2 | 2.703 | 1.486 | 0.081 | 2.618 | 1.402 | 1.867 | 2094.6 |
| Gill 3 | 1.127 | 0.566 | 0.076 | 1.047 | 0.488 | 2.148 | 837.867 |
| Gill 4 | 1.954 | 0.97 | 0.095 | 1.855 | 0.873 | 2.126 | 1484.267 |
| Gill 5 | 1.373 | 0.693 | 0.089 | 1.28 | 0.602 | 2.128 | 1024.267 |

Figure 30: RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of acute experiments. The blue table represents the northern population and the orange table the southern population.

| Sample - North | 260 Raw | 280 Raw | 320 Raw | 260 | 280 | 260/280 | ng/µL |
|---|---|---|---|---|---|---|---|
| N20G1 | 1.548 | 0.764 | 0.091 | 1.449 | 0.67 | 2.163 | 1159 |
| N20G2 | 1.57 | 0.807 | 0.116 | 1.446 | 0.688 | 2.102 | 1156.6 |
| N20G3 | 2.128 | 1.074 | 0.132 | 1.988 | 0.939 | 2.117 | 1590.2 |
| N20G4 | 0.497 | 0.295 | 0.111 | 0.378 | 0.181 | 2.09 | 302.2 |
| N20G5 | 1.394 | 0.713 | 0.084 | 1.302 | 0.626 | 2.08 | 1041.4 |
| N25G1 | 2.044 | 1.092 | 0.131 | 1.905 | 0.958 | 1.989 | 1523.8 |
| N25G2 | 1.284 | 0.633 | 0.055 | 1.221 | 0.575 | 2.124 | 976.6 |
| N25G3 | 1.439 | 0.725 | 0.084 | 1.347 | 0.638 | 2.112 | 1077.4 |
| N25G4 | 1.847 | 0.925 | 0.112 | 1.727 | 0.81 | 2.132 | 1381.4 |
| N25G5 | 1.483 | 0.736 | 0.064 | 1.411 | 0.669 | 2.11 | 1128.6 |
| N30G1 | 2.18 | 1.148 | 0.127 | 2.045 | 1.018 | 2.009 | 1635.8 |
| N30G2 | 0.868 | 0.447 | 0.075 | 0.785 | 0.369 | 2.128 | 627.8 |
| N30G3 | 1.181 | 0.589 | 0.07 | 1.103 | 0.516 | 2.138 | 882.2 |
| N30G4 | 1.277 | 0.651 | 0.094 | 1.175 | 0.554 | 2.121 | 939.8 |
| N30G5 | 0.8 | 0.435 | 0.093 | 0.699 | 0.339 | 2.063 | 559 |

| Sample - Acute | 260 Raw | 280 Raw | 320 Raw | 260 | 280 | 260/280 | ng/µL |
|---|---|---|---|---|---|---|---|
| S20G1 | 1.627 | 0.823 | 0.102 | 1.517 | 0.718 | 2.113 | 1213.4 |
| S20G2 | 1.634 | 0.859 | 0.099 | 1.527 | 0.757 | 2.018 | 1221.4 |
| S20G3 | 1.742 | 0.887 | 0.071 | 1.663 | 0.813 | 2.046 | 1330.2 |
| S20G4 | 1.712 | 0.834 | 0.053 | 1.651 | 0.778 | 2.122 | 1320.6 |
| S20G5 | 2.473 | 1.293 | 0.087 | 2.378 | 1.203 | 1.977 | 1902.2 |
| *S25G1* | *2.165* | *1.198* | *0.157* | *2.004* | *1.038* | *1.93* | *1603.4* |
| S25G2 | 2.526 | 1.533 | 0.151 | 2.367 | 1.379 | 1.717 | 1893.4 |
| S25G3 | 1.441 | 0.741 | 0.09 | 1.343 | 0.648 | 2.073 | 1074.2 |
| S25G4 | 2.709 | 1.564 | 0.072 | 2.629 | 1.489 | 1.766 | 2103 |
| S25G5 | 2.14 | 1.131 | 0.134 | 1.998 | 0.994 | 2.01 | 1598.2 |
| S30G1 | 2.444 | 1.235 | 0.061 | 2.375 | 1.171 | 2.028 | 1899.8 |
| S30G2 | 1.7 | 0.844 | 0.074 | 1.618 | 0.767 | 2.11 | 1294.2 |
| S30G3 | 1.257 | 0.667 | 0.097 | 1.152 | 0.567 | 2.032 | 921.4 |
| S30G4 | 2.716 | 1.421 | 0.105 | 2.603 | 1.313 | 1.983 | 2082.2 |
| S30G5 | 2.629 | 1.307 | 0.074 | 2.547 | 1.23 | 2.071 | 2037.4 |

Figure 31: RNA concentration (ng/$\mu$l) and quality measurements (OD260/280 ratio values) of chronic experiments. The blue table represents the northern population and the orange table the southern population.

### A.0.2 FASTQC and Trimmomatic

```
#!/bin/bash
#$ -pe smp 32
#$ -N RNAShi
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -l hostname=!(compute-0-15)
#$ -o EdgeRNA.output
#$ -e EdgeRNA.output
#$ -t 1-2

FILE1=$(head -n $SGE_TASK_ID rawdata/rawlist | tail -n1|cut -f1 )
FILE2=$(head -n $SGE_TASK_ID rawdata/rawlist | tail -n1|cut -f2 )
NAME=$(head -n $SGE_TASK_ID rawdata/rawlist | tail -n1|cut -f3 )

```

```
echo "running fastqc"

singularity exec /share/apps/singularity/orca-4.simg fastqc
rawdata/$FILE1 rawdata/$FILE2 -o fastqc_raw -t 30

echo "running trimmomatic"
singularity exec /share/apps/singularity/orca-6.simg trimmomatic
PE-threads 28 -phred33 \
rawdata/$FILE1 \
rawdata/$FILE2 \
"$NAME"_R1_P.fq \
"$NAME"_R1_U.fq \
"$NAME"_R2_P.fq \
"$NAME"_R2_U.fq \
ILLUMINACLIP:/home/andresantos/illumina_adaptares.fa:2:30:10
LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36

echo "running fastqc"
singularity exec /share/apps/singularity/orca-4.simg fastqc
"$NAME"_R1_P.fq "$NAME"_R2_P.fq -o fastqc_corrected -t 30

echo "zipping outputs"
singularity exec /share/apps/singularity/pigz_v2.4.simg pigz
-p 30 "$NAME"_R1_P.fq
singularity exec /share/apps/singularity/pigz_v2.4.simg pigz
-p 30 "$NAME"_R1_U.fq
singularity exec /share/apps/singularity/pigz_v2.4.simg pigz
-p 30 "$NAME"_R2_P.fq
singularity exec /share/apps/singularity/pigz_v2.4.simg pigz
-p 30 "$NAME"_R2_U.fq
mv *fq.gz Trimmomatic_out
exit
```

### A.0.3 AGAT and HISAT2

```
#!/bin/bash
#$ -pe smp 32
#$ -N H2Index
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -l h=(compute-0-1)
#$ -o H2Index-stdo.output
#$ -e H2Index-stderr.output

singularity exec /share/apps/singularity/agat_v1.0.0.simg
agat_convert_sp_gff2gtf.pl --gff Ude_annotation_sorted.gff3
-o Ude_annotation_sorted.gtf 1>Ude_annotation_sorted.log
2>Ude_annotation_sorted.err

```

```
singularity exec /share/apps/singularity/hisat_v2.2.1.simg
extract_splice_sites.py
Ude_annotation_sorted.gtf > Ude_gff.ss

singularity exec /share/apps/singularity/hisat_v2.2.1.simg
extract_exons.py Ude_annotation_sorted.gtf > Ude_gff.exons

singularity exec /share/apps/singularity/hisat_v2.2.1.simg
hisat2-build -p 30 --ss Ude_gff.ss --exon Ude_gff.exons
Ude_SM.fa index_Ude
exit
```

### A.0.4 HISAT2

```
#!/bin/bash
#$ -pe smp 32
#$ -N H2Mapp
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -l h=!(compute-0-15)
#$ -o H2Mapp-stdo.output1
#$ -e H2Mapp-stderr.output
#$ -t 1-2

#Hisat2
INFILE1=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f1)
INFILE2=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f2)
INFILE3=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f3)
INFILE4=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f4)
INFILE5=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f5)
INFILE6=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f6)
INFILE7=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f7)
INFILE8=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f8)
INFILE9=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f9)
INFILE10=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f10)
INFILE11=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f11)


#singularity exec /share/apps/singularity/pigz_v2.4.simgpigz
-p 30 -d $INFILE10
#singularity exec /share/apps/singularity/pigz_v2.4.simg pigz
-p 30 -d $INFILE11

singularity exec /share/apps/singularity/hisat_v2.2.1.simg hisat2
--dta -q -p 30 -x index_Ude -1 $INFILE1 -2 $INFILE2 -S $INFILE3

singularity exec /share/apps/singularity/orca-6.simg samtools
view -@ 30 -Sb $INFILE3 > $INFILE4

```

```
singularity exec /share/apps/singularity/orca-6.simg samtools
sort -@ 30 $INFILE4 -o $INFILE5

singularity exec /share/apps/singularity/orca-6.simg samtools
index $INFILE5

singularity exec /share/apps/singularity/orca-6.simg samtools
flagstat $INFILE5 > $INFILE6

rm $INFILE3
rm $INFILE4
```

### A.0.5 Stringtie

```
#!/bin/bash
#$ -pe smp 32
#$ -N stringtie_count
#$ -cwd
#$ -j y
#$ -S /bin/bash
#$ -l h=!(compute-0-0|compute-0-1|compute-0-2|compute-0-3
# |compute-0-4|compute-0-5|compute-0-6)
#$ -e StingtieC-stderr.output
#$ -t 1-40


INFILE1=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f1)
INFILE2=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f2)
INFILE3=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f3)
INFILE4=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f4)
INFILE5=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f5)
INFILE6=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f6)
INFILE7=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f7)
INFILE8=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f8)
INFILE9=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f9)
INFILE10=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f10)
INFILE11=$(head -n $SGE_TASK_ID ConfigFile1.txt | tail -n1|cut -f11)


singularity exec /share/apps/singularity/stringtie_v2.1.2.sif
stringtie $INFILE5 -e -B -p 30 -G Ude_annotation_sorted.gtf
-o $INFILE8 -A $INFILE9
```

# Appendix B
# Details of results

B.0.1 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 20 °C of the chronic experiments, using g: Profiler - Table 12.

B.0.2 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 25 °C of the chronic experiments, using g: Profiler - Table 13.

B.0.3 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of 30 °C of the chronic experiments, using g: Profiler - Table 14.

B.0.4 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of North vs South of the chronic experiments, using g: Profiler - Table 15.

B.0.5 Detailed list of go terms and their associated Ensembl ID, gene name, and description of the comparative analysis of North vs South of the acute experiments, using g: Profiler - Table 16.

Table 12: List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 20 °C.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0003829 | ENSG00000140297 | GCNT3 | glucosaminyl (N-acetyl) transferase 3 |
| | ENSG00000187210 | GCNT1 | glucosaminyl (N-acetyl) transferase 1 |
| CORUM:7197 | ENSG00000151224 | MAT1A | methionine adenosyltransferase 1A |
| GO:0030686 | ENSG00000100029 | PES1 | pescadillo ribosomal biogenesis factor 1 |
| | ENSG00000165271 | NOL6 | nucleolar protein 6 |
| HP:0010068 | ENSG00000137601 | NEK1 | NIMA related kinase 1 |
| | ENSG00000066468 | FGFR2 | fibroblast growth factor receptor 2 |
| R-HSA-2408508 | ENSG00000116761 | CTH | cystathionine gamma-lyase |
| | ENSG00000166741 | NNMT | nicotinamide N-methyltransferase |

Table 13: List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 25 °C.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0016823 | ENSG00000180185 | FAHD1 | fumarylacetoacetate hydrolase domain containing 1 |
| | ENSG00000103876 | FAH | fumarylacetoacetate hydrolase |
| GO:0039531 | ENSG00000110330 | BIRC2 | baculoviral IAP repeat containing 2 |
| | ENSG00000023445 | BIRC3 | baculoviral IAP repeat containing 3 |
| | ENSG00000149476 | TKFC | triokinase and FMN cyclase |
| | ENSG00000139370 | SLC15A4 | solute carrier family 15 member 4 |
| | ENSG00000110171 | TRIM3 | tripartite motif containing 3 |
| | ENSG00000174130 | TLR6 | toll like receptor 6 |
| GO:0035082 | ENSG00000272514 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000157856 | DRC1 | dynein regulatory complex subunit 1 |
| | ENSG00000119661 | DNAL1 | dynein axonemal light chain 1 |
| | ENSG00000172426 | RSPH9 | radial spoke head component 9 |
| | ENSG00000197889 | MEIG1 | meiosis/spermiogenesis associated 1 |
| | ENSG00000092850 | TEKT2 | tektin 2 |
| | ENSG00000187726 | DNAJB13 | DnaJ heat shock protein family (Hsp40) member B13 |
| GO:0001578 | ENSG00000272514 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000157856 | DRC1 | dynein regulatory complex subunit 1 |
| | ENSG00000119661 | DNAL1 | dynein axonemal light chain 1 |
| | ENSG00000172426 | RSPH9 | radial spoke head component 9 |
| | ENSG00000197889 | MEIG1 | meiosis/spermiogenesis associated 1 |
| | ENSG00000092850 | TEKT2 | tektin 2 |
| | ENSG00000187726 | DNAJB13 | DnaJ heat shock protein family (Hsp40) member B13 |
| GO:0002753 | ENSG00000110330 | BIRC2 | baculoviral IAP repeat containing 2 |
| | ENSG00000023445 | BIRC3 | baculoviral IAP repeat containing 3 |
| | ENSG00000152223 | EPG5 | ectopic P-granules 5 autophagy tethering factor |
| | ENSG00000149476 | TKFC | triokinase and FMN cyclase |
| | ENSG00000139370 | SLC15A4 | solute carrier family 15 member 4 |
| | ENSG00000115267 | IFIH1 | interferon induced with helicase C domain 1 |
| | ENSG00000110171 | TRIM3 | tripartite motif containing 3 |
| | ENSG00000174130 | TLR6 | toll like receptor 6 |
| GO:0060294 | ENSG00000163060 | TEKT4 | tektin 4 |
| | ENSG00000151023 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000196118 | CFAP20 | cilia and flagella associated protein 20 |
| | ENSG00000203666 | RSPH9 | radial spoke head component 9 |
| GO:0097228 | ENSG00000163060 | TEKT4 | tektin 4 |
| | ENSG00000272514 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000070761 | CFAP20 | cilia and flagella associated protein 20 |
| | ENSG00000172426 | RSPH9 | radial spoke head component 9 |
| | ENSG00000197889 | MEIG1 | meiosis/spermiogenesis associated 1 |
| | ENSG00000151023 | ENKUR | enkurin, TRPC channel interacting protein |
| | ENSG00000092850 | TEKT2 | tektin 2 |
| GO:0005879 | ENSG00000163060 | TEKT4 | tektin 4 |
| | ENSG00000272514 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000186973 | CFAP144 | cilia and flagella associated protein 144 |
| | ENSG00000070761 | CFAP20 | cilia and flagella associated protein 20 |
| | ENSG00000151023 | ENKUR | enkurin, TRPC channel interacting protein |
| | ENSG00000183690 | EFHC2 | EF-hand domain containing 2 |
| | ENSG00000179902 | ENKUR | enkurin, TRPC channel interacting protein |
| | ENSG00000092850 | CFAP276 | cilia and flagella associated protein 276 |
| | ENSG00000128408 | RIBC2 | RIB43A domain with coiled-coils 2 |

Continued in the next page.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0030286 | ENSG00000157856 | DRC1 | dynein regulatory complex subunit 1 |
| | ENSG00000213123 | DYNLT2B | dynein light chain Tctex-type 2B |
| | ENSG00000100246 | DNAL4 | dynein axonemal light chain 4 |
| | ENSG00000119661 | DNAL1 | dynein axonemal light chain 1 |
| | ENSG00000152760 | DYNLT5 | dynein light chain Tctex-type family member 5 |
| | ENSG00000264364 | DYNLL2 | dynein light chain LC8-type 2 |
| GO:0005881 | ENSG00000163060 | TEKT4 | tektin 4 |
| | ENSG00000272514 | CFAP206 | cilia and flagella associated protein 206 |
| | ENSG00000186973 | CFAP144 | cilia and flagella associated protein 144 |
| | ENSG00000070761 | CFAP20 | cilia and flagella associated protein 20 |
| | ENSG00000151023 | ENKUR | enkurin, TRPC channel interacting protein |
| | ENSG00000183690 | EFHC2 | EF-hand domain containing 2 |
| | ENSG00000179902 | CFAP276 | cilia and flagella associated protein 276 |
| | ENSG00000092850 | TEKT2 | tektin 2 |
| | ENSG00000128408 | RIBC2 | RIB43A domain with coiled-coils 2 |
| CORUM:5529 | ENSG00000110330 | BIRC2 | baculoviral IAP repeat containing 2 |
| | ENSG00000127191 | TRAF2 | TNF receptor associated factor 2 |

Table 14: List of go terms and their associate gene names and descriptions of the comparative analysis of chronic experiments in the temperature of 30 °C.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0010854 | ENSG00000160014 | CALM3 | calmodulin 3 |
| | ENSG00000196277 | GRM7 | glutamate metabotropic receptor 7 |
| | ENSG00000143933 | CALM2 | calmodulin 2 |
| GO:0050135 | ENSG00000196083 | IL1RAP | interleukin 1 receptor accessory protein |
| | ENSG00000137462 | TLR2 | toll like receptor 2 |
| | ENSG00000136869 | TLR4 | toll like receptor 4 |
| | ENSG00000174125 | TLR1 | toll like receptor 1 |
| | ENSG00000174123 | TLR10 | toll like receptor 10 |
| GO:0022848 | ENSG00000080644 | CHRNA3 | cholinergic receptor nicotinic alpha 3 subunit |
| | ENSG00000120903 | CHRNA2 | cholinergic receptor nicotinic alpha 2 subunit |
| | ENSG00000147432 | CHRNB3 | cholinergic receptor nicotinic beta 3 subunit |
| | ENSG00000147434 | CHRNA6 | cholinergic receptor nicotinic alpha 6 subunit |
| GO:0044769 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000159720 | ATP6V0D1 | ATPase H+ transporting V0 subunit d1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| | ENSG00000136888 | ATP6V1G1 | ATPase H+ transporting V1 subunit G1 |
| GO:0016875 | ENSG00000113643 | RARS1 | arginyl-tRNA synthetase 1 |
| | ENSG00000140105 | WARS1 | tryptophanyl-tRNA synthetase 1 |
| | ENSG00000116120 | FARSB | phenylalanyl-tRNA synthetase subunit beta |
| | ENSG00000166986 | MARS1 | methionyl-tRNA synthetase 1 |
| | ENSG00000139131 | YARS2 | tyrosyl-tRNA synthetase 2 |
| | ENSG00000133706 | LARS1 | leucyl-tRNA synthetase 1 |
| | ENSG00000031698 | SARS1 | seryl-tRNA synthetase 1 |
| | ENSG00000106105 | GARS1 | glycyl-tRNA synthetase 1 |
| | ENSG00000196305 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000179115 | FARSA | phenylalanyl-tRNA synthetase subunit alpha |
| | ENSG00000113407 | TARS1 | threonyl-tRNA synthetase 1 |
| | ENSG00000134440 | NARS1 | asparaginyl-tRNA synthetase 1 |
| | ENSG00000110619 | CARS1 | cysteinyl-tRNA synthetase 1 |
| | ENSG00000115866 | DARS1 | aspartyl-tRNA synthetase 1 |

Continued in the next page.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0044183 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000170606 | HSPA4 | heat shock protein family A (Hsp70) member 4 |
| | ENSG00000004776 | HSPB6 | heat shock protein family B (small) member 6 |
| | ENSG00000187522 | HSPA14 | heat shock protein family A (Hsp70) member 14 |
| | ENSG00000115541 | HSPE1 | heat shock protein family E (Hsp10) member 1 |
| | ENSG00000144381 | HSPD1 | heat shock protein family D (Hsp60) member 1 |
| | ENSG00000143256 | PFDN2 | prefoldin subunit 2 |
| | ENSG00000179218 | CALR | calreticulin |
| | ENSG00000113013 | HSPA9 | heat shock protein family A (Hsp70) member 9 |
| GO:0048388 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000071553 | ATP6AP1 | ATPase H+ transporting accessory protein 1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| GO:0061795 | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000065923 | SLC9A7 | solute carrier family 9 member A7 |
| | ENSG00000071553 | ATP6AP1 | ATPase H+ transporting accessory protein 1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| GO:0036295 | ENSG00000105974 | CAV1 | caveolin 1 |
| GO:0048388 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000071553 | ATP6AP1 | ATPase H+ transporting accessory protein 1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| GO:0043201 | ENSG00000133706 | LARS1 | leucyl-tRNA synthetase 1 |
| | ENSG00000025039 | RRAGD | Ras related GTP binding D |
| | ENSG00000121879 | PIK3CA | phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha |
| | ENSG00000141564 | RPTOR | regulatory associated protein of MTOR complex 1 |
| | ENSG00000108443 | RPS6KB1 | ribosomal protein S6 kinase B1 |
| GO:0097401 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000159720 | ATP6V0D1 | ATPase H+ transporting V0 subunit d1 |
| | ENSG00000071553 | ATP6AP1 | ATPase H+ transporting accessory protein 1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| | ENSG00000136888 | ATP6V1G1 | ATPase H+ transporting V1 subunit G1 |
| GO:0030150 | ENSG00000154174 | TOMM70 | translocase of outer mitochondrial membrane 70 |
| | ENSG00000120675 | DNAJC15 | DnaJ heat shock protein family (Hsp40) member C15 |
| | ENSG00000126768 | TIMM17B | translocase of inner mitochondrial membrane 17B |
| | ENSG00000104980 | TIMM44 | translocase of inner mitochondrial membrane 44 |
| | ENSG00000109519 | GRPEL1 | GrpE like 1, mitochondrial |
| | ENSG00000130204 | TOMM40 | translocase of outer mitochondrial membrane 40 |
| GO:0030150 | ENSG00000154174 | TOMM70 | translocase of outer mitochondrial membrane 70 |
| | ENSG00000120675 | DNAJC15 | DnaJ heat shock protein family (Hsp40) member C15 |
| | ENSG00000126768 | TIMM17B | translocase of inner mitochondrial membrane 17B |
| | ENSG00000104980 | TIMM44 | translocase of inner mitochondrial membrane 44 |
| | ENSG00000109519 | GRPEL1 | GrpE like 1, mitochondrial |
| | ENSG00000130204 | TOMM40 | translocase of outer mitochondrial membrane 40 |

Continued in the next page.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0007042 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D of outer mitochondrial membrane 70 |
| | ENSG00000071553 | ATP6AP1 | ATPase H+ transporting accessory protein 1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| GO:0007035 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D of outer mitochondrial membrane 70 |
| | ENSG00000159720 | ATP6V0D1 | ATPase H+ transporting V0 subunit d1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000117410 | ATP6V0B | ATPase H+ transporting V0 subunit b |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| GO:0043038 | ENSG00000113643 | RARS1 | arginyl-tRNA synthetase 1 |
| | ENSG00000140105 | WARS1 | tryptophanyl-tRNA synthetase 1 |
| | ENSG00000116120 | FARSB | phenylalanyl-tRNA synthetase subunit beta |
| | ENSG00000166986 | MARS1 | methionyl-tRNA synthetase 1 |
| | ENSG00000139131 | YARS2 | tyrosyl-tRNA synthetase 2 |
| | ENSG00000133706 | LARS1 | leucyl-tRNA synthetase 1 |
| | ENSG00000031698 | SARS1 | seryl-tRNA synthetase 1 |
| | ENSG00000106105 | GARS1 | glycyl-tRNA synthetase 1 |
| | ENSG00000196305 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000179115 | FARSA | phenylalanyl-tRNA synthetase subunit alpha |
| | ENSG00000113407 | TARS1 | threonyl-tRNA synthetase 1 |
| | ENSG00000134440 | NARS1 | asparaginyl-tRNA synthetase 1 |
| | ENSG00000110619 | CARS1 | cysteinyl-tRNA synthetase 1 |
| | ENSG00000115866 | DARS1 | aspartyl-tRNA synthetase 1 |
| GO:0035354 | ENSG00000137462 | TLR2 | toll like receptor 2 |
| | ENSG00000174125 | TLR1 | toll like receptor 1 |
| GO:0033180 | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000047249 | ATP6V1H | ATPase H+ transporting V1 subunit H |
| | ENSG00000136888 | ATP6V1G1 | ATPase H+ transporting V1 subunit G1 |
| GO:0017101 | ENSG00000113643 | RARS1 | arginyl-tRNA synthetase 1 |
| | ENSG00000166986 | MARS1 | methionyl-tRNA synthetase 1 |
| | ENSG00000133706 | LARS1 | leucyl-tRNA synthetase 1 |
| | ENSG00000196305 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000115866 | DARS1 | aspartyl-tRNA synthetase 1 |
| KEGG:00970 | ENSG00000113643 | RARS1 | arginyl-tRNA synthetase 1 |
| | ENSG00000140105 | MARS1 | methionyl-tRNA synthetase 1 |
| | ENSG00000116120 | LARS1 | leucyl-tRNA synthetase 1 |
| | ENSG00000166986 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000139131 | DARS1 | aspartyl-tRNA synthetase 1 |
| | ENSG00000133706 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000031698 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000106105 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000196305 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000179115 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000113407 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000134440 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000110619 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000115866 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000166986 | IARS1 | isoleucyl-tRNA synthetase 1 |
| | ENSG00000103707 | IARS1 | isoleucyl-tRNA synthetase 1 |

Continued in the next page.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| KEGG:04966 | ENSG00000033627 | ATP6V0A1 | ATPase H+ transporting V0 subunit a1 |
| | ENSG00000114573 | ATP6V1A | ATPase H+ transporting V1 subunit A |
| | ENSG00000100554 | ATP6V1D | ATPase H+ transporting V1 subunit D |
| | ENSG00000159720 | ATP6V0D1 | ATPase H+ transporting V0 subunit d1 |
| | ENSG00000128524 | ATP6V1F | ATPase H+ transporting V1 subunit F |
| | ENSG00000136888 | ATP6V1G1 | ATPase H+ transporting V1 subunit G1 |
| KEGG:00071 | ENSG00000155016 | CYP2U1 | cytochrome P450 family 2 subfamily U member 1 |
| | ENSG00000167969 | ECI1 | enoyl-CoA delta isomerase 1 |
| | ENSG00000122971 | ACADS | acyl-CoA dehydrogenase short chain |
| | ENSG00000115361 | ACADL | acyl-CoA dehydrogenase long chain |
| | ENSG00000072778 | ACADVL | acyl-CoA dehydrogenase very long chain |
| | ENSG00000138029 | HADHB | hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit beta |
| | ENSG00000157184 | CPT2 | carnitine palmitoyltransferase 2 |
| | ENSG00000167315 | ACAA2 | acetyl-CoA acyltransferase 2 |
| | ENSG00000105607 | GCDH | glutaryl-CoA dehydrogenase |
| REAC:R-HSA-629597 | ENSG00000080644 | CHRNA3 | cholinergic receptor nicotinic alpha 3 subunit |
| | ENSG00000120903 | CHRNA2 | cholinergic receptor nicotinic alpha 2 subunit |
| | ENSG00000147432 | CHRNB3 | cholinergic receptor nicotinic beta 3 subunit |
| | ENSG00000147434 | CHRNA6 | cholinergic receptor nicotinic alpha 6 subunit |

Table 15: List of go terms and their associate gene names and descriptions of the comparative analysis of North vs South of chronic experiments.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0004062 | ENSG00000197165 | SULT1A2 | sulfotransferase family 1A member 2 |
| | ENSG00000173597 | SULT1B1 | sulfotransferase family 1B member 1 |
| | ENSG00000196502 | SULT1A1 | sulfotransferase family 1A member 1 |
| | ENSG00000198203 | SULT1C2 | sulfotransferase family 1C member 2 |
| GO:0061809 | ENSG00000196083 | IL1RAP | interleukin 1 receptor accessory protein |
| | ENSG00000174125 | TLR1 | toll like receptor 1 |
| | ENSG00000174130 | TLR6 | toll like receptor |
| | ENSG00000136869 | TLR4 | toll like receptor 4 |
| | ENSG00000174125 | TLR1 | toll like receptor 1 |
| GO:0022848 | ENSG00000174343 | CHRNA9 | cholinergic receptor nicotinic alpha 9 subunit |
| | ENSG00000101204 | CHRNA4 | cholinergic receptor nicotinic alpha 4 subunit |
| | ENSG00000120903 | CHRNA2 | cholinergic receptor nicotinic alpha 2 subunit |
| | ENSG00000080644 | CHRNA3 | cholinergic receptor nicotinic alpha 3 subunit |
| | ENSG00000147434 | CHRNA6 | cholinergic receptor nicotinic alpha 6 subunit |
| GO:0005231 | ENSG00000174343 | CHRNA9 | cholinergic receptor nicotinic alpha 9 subunit |
| | ENSG00000101204 | CHRNA4 | cholinergic receptor nicotinic alpha 4 subunit |
| | ENSG00000120903 | CHRNA2 | cholinergic receptor nicotinic alpha 2 subunit |
| | ENSG00000080644 | CHRNA3 | cholinergic receptor nicotinic alpha 3 subunit |
| | ENSG00000147434 | CHRNA6 | cholinergic receptor nicotinic alpha 6 subunit |
| GO:0072349 | ENSG00000165449 | SLC16A9 | solute carrier family 16 member 9 |
| | ENSG00000004864 | SLC25A13 | solute carrier family 25 member 13 |
| | ENSG00000152779 | SLC16A12 | SLC16A12 |
| | ENSG00000142494 | SLC47A1 | solute carrier family 47 member 1 |
| | ENSG00000268104 | SLC6A14 | solute carrier family 6 member 14 |
| | ENSG00000138074 | SLC5A6 | solute carrier family 5 member 6 |
| | ENSG00000103064 | SLC7A6 | solute carrier family 7 member 6 |
| GO:0140359 | ENSG00000154265 | ABCA5 | ATP binding cassette subfamily A member 5 |
| | ENSG00000143921 | ABCG8 | ATP binding cassette subfamily G member 8 |
| | ENSG00000119688 | ABCD4 | ATP binding cassette subfamily D member 4 |
| | ENSG00000165029 | ABCA1 | ATP binding cassette subfamily A member 1 |
| | ENSG00000103222 | ABCC1 | ATP binding cassette subfamily C member 1 |
| | ENSG00000135776 | ABCB10 | ATP binding cassette subfamily B member 10 |
| | ENSG00000160179 | ABCG1 | ATP binding cassette subfamily G member 1 |
| | ENSG00000167972 | ABCA3 | ATP binding cassette subfamily A member 3 |

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0016242 | ENSG00000181852 | RNF41 | ring finger protein 41 |
| | ENSG00000145016 | RUBCN | rubicon autophagy regulator |
| | ENSG00000121879 | PIK3CA | phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha |
| | ENSG00000188906 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000171055 | FEZ2 | fasciculation and elongation protein zeta 2 |
| | ENSG00000198793 | MTOR | mechanistic target of rapamycin kinase |
| | ENSG00000141458 | NPC1 | NPC intracellular cholesterol transporter 1 |
| GO:0046164 | ENSG00000175066 | GK5 | glycerol kinase 5 |
| | ENSG00000115234 | SNX17 | sorting nexin 17 |
| | ENSG00000197165 | SULT1A2 | sulfotransferase family 1A member 2 |
| | ENSG00000173597 | SULT1B1 | sulfotransferase family 1B member 1 |
| | ENSG00000196502 | SULT1A1 | sulfotransferase family 1A member 1 |
| | ENSG00000149476 | TKFC | triokinase and FMN cyclase |
| | ENSG00000140263 | SORD | orbitol dehydrogenase |
| | ENSG00000198074 | AKR1B10 | aldo-keto reductase family 1 member B10 |
| GO:0072337 | ENSG00000165449 | SLC16A9 | solute carrier family 16 member 9 |
| | ENSG00000081479 | LRP2 | LDL receptor related protein 2 |
| | ENSG00000004864 | SLC25A13 | solute carrier family 25 member 13 |
| | ENSG00000152779 | SLC16A12 | solute carrier family 16 member 12 |
| | ENSG00000142494 | SLC47A1 | solute carrier family 47 member 1 |
| | ENSG00000268104 | SLC6A14 | solute carrier family 6 member 14 |
| | ENSG00000138074 | SLC5A6 | solute carrier family 5 member 6 |
| | ENSG00000103064 | SLC7A6 | solute carrier family 7 member 6 |
| GO:1901616 | ENSG00000175066 | GK5 | glycerol kinase 5 |
| | ENSG00000160868 | CYP3A4 | cytochrome P450 family 3 subfamily A member 4 |
| | ENSG00000166816 | LDHD | lactate dehydrogenase D |
| | ENSG00000115234 | SNX17 | sorting nexin 17 |
| | ENSG00000160868 | CYP3A4 | cytochrome P450 family 3 subfamily A member 4 |
| | ENSG00000197165 | SULT1A2 | sulfotransferase family 1A member 2 |
| | ENSG00000173597 | SULT1B1 | sulfotransferase family 1B member 1 |
| | ENSG00000196502 | SULT1A1 | sulfotransferase family 1A member 1 |
| | ENSG00000149476 | TKFC | triokinase and FMN cyclase |
| | ENSG00000140263 | SORD | sorbitol dehydrogenase |
| | ENSG00000198074 | AKR1B10 | aldo-keto reductase family 1 member B10 |
| GO:0051899 | ENSG00000135124 | P2RX4 | purinergic receptor P2X 4 |
| | ENSG00000163681 | SLMAP | sarcolemma associated protein |
| | ENSG00000049759 | NEDD4L | NEDD4 like E3 ubiquitin protein ligase |
| | ENSG00000188906 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000001084 | GCLC | glutamate-cysteine ligase catalytic subuni |
| GO:0010507 | ENSG00000181852 | P2RX4 | purinergic receptor P2X 4 |
| | ENSG00000145016 | SLMAP | sarcolemma associated protein |
| | ENSG00000132405 | NEDD4L | NEDD4 like E3 ubiquitin protein ligase |
| | ENSG00000025039 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000121879 | GCLC | glutamate-cysteine ligase catalytic subuni |
| | ENSG00000188906 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000146872 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000171055 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000198793 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000141564 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000188906 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000141458 | LRRK2 | leucine rich repeat kinase 2 |

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0005892 | ENSG00000174343 | CHRNA9 | cholinergic receptor nicotinic alpha 9 subunit |
| | ENSG00000101204 | CHRNA4 | cholinergic receptor nicotinic alpha 4 subunit |
| | ENSG00000120903 | CHRNA2 | cholinergic receptor nicotinic alpha 2 subunit |
| | ENSG00000080644 | CHRNA3 | cholinergic receptor nicotinic alpha 3 subunit |
| | ENSG00000147434 | CHRNA6 | cholinergic receptor nicotinic alpha 6 subunit |
| KEGG:02010 | ENSG00000154265 | ABCA5 | ATP binding cassette subfamily A member 5 |
| | ENSG00000143921 | ABCG8 | ATP binding cassette subfamily G member 8 |
| | ENSG00000119688 | ABCD4 | ATP binding cassette subfamily D member 4 |
| | ENSG00000165029 | ABCA1 | ATP binding cassette subfamily A member 1 |
| | ENSG00000107331 | ABCA2 | ATP binding cassette subfamily A member 2 |
| | ENSG00000103222 | ABCC1 | ATP binding cassette subfamily C member 1 |
| | ENSG00000135776 | ABCB10 | ATP binding cassette subfamily B member 10 |
| | ENSG00000160179 | ABCG1 | ATP binding cassette subfamily G member 1 |
| | ENSG00000167972 | ABCA3 | ATP binding cassette subfamily A member 3 |
| KEGG:00590 | ENSG00000155016 | CYP2U1 | cytochrome P450 family 2 subfamily U member 1 |
| | ENSG00000186115 | CYP4F2 | cytochrome P450 family 4 subfamily F member 2 |
| | ENSG00000179593 | ALOX15B | arachidonate 15-lipoxygenase type B |
| | ENSG00000179148 | ALOXE3 | arachidonate lipoxygenase 3 |
| | ENSG00000138109 | CYP2C9 | cytochrome P450 family 2 subfamily C member 9 |
| | ENSG00000159228 | CBR1 | carbonyl reductase 1 |

Table 16: List of go terms and their associate gene names and descriptions of the comparative analysis of acute experiments.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0003756 | ENSG00000143870 | PDIA6 | protein disulfide isomerase family A member 6 |
| | ENSG00000167004 | PDIA3 | protein disulfide isomerase family A member 3 |
| | ENSG00000155660 | PDIA4 | protein disulfide isomerase family A member 4 |
| | ENSG00000184164 | CRELD2 | cysteine rich with EGF like domains 2 |
| GO:0016864 | ENSG00000143870 | PDIA6 | protein disulfide isomerase family A member 6 |
| | ENSG00000167004 | PDIA3 | protein disulfide isomerase family A member 3 |
| | ENSG00000155660 | PDIA4 | protein disulfide isomerase family A member 4 |
| | ENSG00000184164 | CRELD2 | cysteine rich with EGF like domains 2 |
| GO:0016701 | ENSG00000129596 | CDO1 | cysteine dioxygenase type 1 |
| | ENSG00000179477 | ALOX12B | arachidonate 12-lipoxygenase, 12R type |
| | ENSG00000012779 | ALOX5 | arachidonate 5-lipoxygenase |
| | ENSG00000182551 | ADI1 | acireductone dioxygenase 1 |
| GO:0000098 | ENSG00000129596 | CDO1 | cysteine dioxygenase type 1 |
| | ENSG00000139631 | CSAD | cysteine sulfinic acid decarboxylase |
| | ENSG00000151224 | MAT1A | methionine adenosyltransferase 1A |
| | ENSG00000108578 | BLMH | bleomycin hydrolase |
| GO:0034975 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000167004 | PDIA3 | protein disulfide isomerase family A member 3 |
| | ENSG00000044574 | HSPA5 | heat shock protein family A (Hsp70) member 5 |
| | ENSG00000179218 | CALR | calreticulin |
| | ENSG00000102580 | DNAJC3 | DnaJ heat shock protein family (Hsp40) member C3 |
| GO:0000096 | ENSG00000129596 | CDO1 | cysteine dioxygenase type 1 |
| | ENSG00000139631 | CSAD | cysteine sulfinic acid decarboxylase |
| | ENSG00000151224 | MAT1A | methionine adenosyltransferase 1A |
| | ENSG00000108578 | BLMH | bleomycin hydrolase |
| | ENSG00000182551 | ADI1 | acireductone dioxygenase 1 |

Continued in the next page.

| Term ID | Ensembl ID | Gene Name | Gene Description |
|---|---|---|---|
| GO:0039531 | ENSG00000110330 | BIRC2 | baculoviral IAP repeat containing 2 |
| | ENSG00000110171 | TRIM3 | tripartite motif containing 3 |
| | ENSG00000105974 | CAV1 | caveolin 1 |
| | ENSG00000134644 | PUM1 | pumilio RNA binding family member 1 |
| | ENSG00000101966 | XIAP | X-linked inhibitor of apoptosis |
| | ENSG00000075240 | GRAMD4 | GRAM domain containing |
| GO:0035354 | ENSG00000174125 | TLR1 | toll like receptor 1 |
| | ENSG00000137462 | TLR2 | toll like receptor 2 |
| GO:0033018 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000145050 | MANF | mesencephalic astrocyte derived neurotrophic factor |
| GO:0034663 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000149428 | HYOU1 | hypoxia up-regulated 1 |
| | ENSG00000166794 | PPIB | peptidylprolyl isomerase B |
| | ENSG00000143870 | PDIA6 | protein disulfide isomerase family A member 6 |
| | ENSG00000044574 | HSPA5 | heat shock protein family A (Hsp70) member 5 |
| GO:0005614 | ENSG00000143382 | ADAMTSL4 | ADAMTS like 4 |
| | ENSG00000205221 | VIT | vitrin |
| | ENSG00000188730 | VWC2 | von Willebrand factor C domain containing 2 |
| GO:0071682 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member |
| | ENSG00000149428 | HYOU1 | hypoxia up-regulated 1 |
| | ENSG00000008438 | PGLYRP1 | peptidoglycan recognition protein 1 |
| | ENSG00000179218 | CALR | calreticuli |
| GO:0005790 | ENSG00000089248 | ERP29 | endoplasmic reticulum protein 29 |
| | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000149428 | HYOU1 | hypoxia up-regulated 1 |
| | ENSG00000166794 | PPIB | peptidylprolyl isomerase B |
| | ENSG00000179218 | CALR | calreticulin |
| | ENSG00000102580 | DNAJC3 | DnaJ heat shock protein family (Hsp40) member C3 |
| GO:0005901 | ENSG00000094631 | HDAC6 | histone deacetylase |
| | ENSG00000188906 | LRRK2 | leucine rich repeat kinase 2 |
| | ENSG00000105974 | CAV1 | caveolin 1 |
| | ENSG00000003436 | TFPI | tissue factor pathway inhibitor |
| | ENSG00000128602 | SMO | smoothened, frizzled class receptor |
| R-HSA-381183 | ENSG00000166598 | HSP90B1 | heat shock protein 90 beta family member 1 |
| | ENSG00000044574 | HSPA5 | heat shock protein family A (Hsp70) member 5 |
| | ENSG00000179218 | CALR | calreticulin |
| R-HSA-381038 | ENSG00000149428 | HYOU1 | hypoxia up-regulated 1 |
| | ENSG00000143870 | PDIA6 | protein disulfide isomerase family A member 6 |
| | ENSG00000074842 | MYDGF | myeloid derived growth factor |
| | ENSG00000128590 | DNAJB9 | DnaJ heat shock protein family (Hsp40) member B9 |
| | ENSG00000102580 | DNAJC3 | DnaJ heat shock protein family (Hsp40) member C |
| R-HSA-381070 | ENSG00000149428 | HYOU1 | hypoxia up-regulated 1 |
| | ENSG00000143870 | PDIA6 | protein disulfide isomerase family A member 6 |
| | ENSG00000074842 | MYDGF | myeloid derived growth factor |
| | ENSG00000044574 | HSPA5 | heat shock protein family A (Hsp70) member 5 |
| | ENSG00000128590 | DNAJB9 | DnaJ heat shock protein family (Hsp40) member B9 |
| | ENSG00000102580 | DNAJC3 | DnaJ heat shock protein family (Hsp40) member C3 |
| CORUM:7048 | ENSG00000174125 | TLR1 | toll like receptor 1 |
| | ENSG00000137462 | TLR2 | toll like receptor 2 1 |