UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY & BIOINFORMATICS

ANDREA KAROLINA DUARTE DUQUE
BSc in Biotechnology

Computational tools to assess adaptation and resilience to thermal stress in intertidal fish: a functional genomics approach

Sep, 2025

ITQB NOVA
NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL
UNIVERSIDADE NOVA DE LISBOA

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

COMPUTATIONAL TOOLS TO ASSESS ADAPTATION AND RESILIENCE THERMAL STRESS IN INTERTIDAL FISH A FUNCTIONAL GENOMICS APPROACH

ANDREA KAROLINA DUARTE DUQUE

BSc in Biotechnology

Adviser: Sara Carolina Gusmão Coito Madeira
Assistant Professor, NOVA University Lisbon

Co-advisers: Pedro Manuel Brôa Costa
Assistant Professor, *NOVA* University Lisbon

Examination Committee:

Chair: Paula Maria Theriaga Mendes Bernardo Gonçalves,
Associate Professor, NOVA University Lisbon

Rapporteurs: Francisco Rente de Pina Martins,
Assistant Professor, Escola Superior de Tecnologia do Barreiro

Adviser: Sara Carolina Gusmão Coito Madeira,
Assistant Professor, NOVA University Lisbon

Members: Pedro Manuel Brôa Costa,
Assistant Professor, NOVA University Lisbon

MASTER IN MSC COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2025

Computational tools to assess adaptation and resilience to thermal stress In Intertidal fish: a functional genomics approach

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisers, Carolina Madeira and Pedro Costa, for giving me the opportunity to carry out this research. Their guidance, patience, and ability to bring clarity whenever I tend to overcomplicate things have been truly invaluable. I am also grateful to all the members of the SeaTox team for their continuous support, insightful suggestions, and encouragement throughout this journey. My heartfelt thanks go to my family for always believing in me and standing by my side through every challenge. To my friends, thank you for reminding me of the importance of balance, joy, and focusing on the positives along the way. Finally, a very special thank you to Mariana Gomes, Carolina Costa, Filipe Morgado, and my little sister for always supporting me, and helping me make the right decisions when I needed it most. I acknowledge Fundação para a Ciência e Tecnologia (FCT, I.P.) for funding project ExtremeOceans (PTDC/BIA-BMA/1494/2020, doi: 10.54499/PTDC/BIA-BMA/1494/2020) through national funds, which supported the research work of this thesis. I also thank the financial support provided via strategic projects of UCIBIO (UID/04378) and i4HB (LA/P/0140/2020). Lastly, I thank Câmara Municipal de Oliveira do Bairro for awarding me an individual scholarship to complete my MSc degree at NOVA School of Science and Technology.

"The creatures which can stand the 'storm and stress' of […] changes which occur in the environment, by undergoing modification of […] the structures which they get congenitally - these creatures will live; while those which cannot, will not." Baldwin (1896).

ABSTRACT

Thermal variability associated with climate change poses significant challenges to intertidal species, demanding both phenotypic plasticity and local adaptation. This study investigates the molecular bases of thermal resilience in *Pomatoschistus microps*, an intertidal fish distributed along the Portuguese coast. We hypothesized that populations from different latitudes exhibit divergent coping mechanisms to seasonal warming, reflecting alternative strategies to cope with environmental variability and thermal stress. Using RNA-seq, we compared the transcriptomic profiles of three populations (Ria de Aveiro, Troia, and Ria Formosa) across spring and summer following a de novo transcriptome assembly and R-based bioinformatics analyses. Differential gene expression and pathway enrichment analyses were conducted to identify molecular mechanisms underlying thermal adaptation and acclimation. Troia populations exhibit a highly dynamic, stress-preparatory strategy, characterized by frontloading of stress-related genes, while Ria Formosa populations demonstrate basal stabilization via enhanced ribosomal and mitochondrial pathways. Ria de Aveiro seasonal variation reveals metabolic priming and summer profiles dominated by reproductive activation. These results demonstrate that *P. microps* populations employ distinct molecular strategies to achieve resilience, integrating local adaptation and phenotypic plasticity across spatial and temporal scales. By linking transcriptomic responses to environmental context, this study advances understanding of mechanisms driving organismal survival under variable thermal regimes. The findings have implications for biodiversity conservation, fisheries management, and aquaculture, offering molecular insights to support strategies that enhance thermal tolerance and population persistence under climate change.

Keywords: Seasonal variation, Latitudinal Gradient, Transcriptomics, Intertidal, *P. microps*, Molecular mechanisms, Thermal acclimatization & adaptation

RESUMO

A variabilidade térmica associada às alterações climáticas representa desafios significativos para as espécies intertidais, exigindo simultaneamente plasticidade fenotípica e adaptação local. Este estudo investiga as bases moleculares da resiliência térmica em *Pomatoschistus microps*, um peixe intertidal distribuído ao longo da costa portuguesa. Hipotetizamos que populações de diferentes latitudes apresentam mecanismos distintos de resposta ao aquecimento sazonal, refletindo estratégias alternativas para lidar com a variabilidade ambiental e o stress térmico. Recorrendo a RNA-seq, comparamos os perfis transcritos de três populações (Ria de Aveiro, Troia e Ria Formosa) na primavera e no verão, após a montagem *de novo* do transcriptôma e análises bioinformáticas em R. Foram conduzidas análises de expressão gênica diferencial e de enriquecimento de vias metabólicas para identificar os mecanismos moleculares subjacentes à adaptação e aclimatização térmica. As populações de Troia evidenciaram uma estratégia altamente dinâmica e de preparação para o stress, caracterizada por *frontloading* de genes relacionados com respostas ao stress, enquanto as populações da Ria Formosa demonstraram estabilização basal através do reforço de vias ribossômicas e mitocondriais. A variação sazonal na Ria de Aveiro revelou uma preparação metabólica, com perfis de verão dominados pela ativação reprodutiva. Estes resultados demonstram que as populações de *P. microps* empregam estratégias moleculares distintas para alcançar resiliência, integrando adaptação local e plasticidade fenotípica em diferentes escalas espaciais e temporais. Ao relacionar respostas transcriptômicas com o contexto ambiental, este estudo contribui para a compreensão dos mecanismos que sustentam a sobrevivência dos organismos sob regimes térmicos variáveis. Os resultados têm implicações para a conservação da biodiversidade, a gestão das pescas e a aquacultura, oferecendo uma base molecular que poderá apoiar estratégias destinadas a aumentar a tolerância térmica e a persistência populacional perante as alterações climáticas.

Palavras chave: Variação sazonal, Gradiente latitudinal, Transcriptômica, Intertidal, *P. microps*, Mecanismos moleculares, Aclimatização & adaptação térmica

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

%ID	percentage identity.
BLAST	Basic Local Alignment Search Tool.
bp	Base pair.
cDNA	Complementary DNA.
Csa	Mediterranean climate with dry hot summers.
Csb	Mediterranean climate with dry mild summers.
CTmax	Critical Thermal maximum.
CTmin	Critical Thermal minimum.
DEGs	Differentially expressed genes.
DNA	Deoxyribonucleic acid.
***e*-value**	Expectation value or Expect value.
FDR	False Discovery Rate.
GLMs	Generalize linear models.
GSEA	Gene Set Enrichment Analysis.
HSPs	Heat Shock Proteins.
KEGG	Kyoto Encyclopaedia of Genes and Genomes.
log_2FC	log_2 fold change.
M	Million.
MCL	Markov Cluster Algorithm.
NES	Normalized Enrichment Score.
ORFs	Open Reading Frames.

***p*-value**	Probability Value.
PCA	Principal Component Analysis.
PPI	Protein-Protein interactions.
PSU	Practical Salinity Unit.
R	R Programming Language.
RA	Ria de Aveiro.
RF	Ria Formosa.
RIN	RNA Integrity Number.
RNA	Ribonucleic Acid.
RNA-seq	RNA sequencing.
Sp	Spring.
Su	Summer.
Swiss-prot	"UniProtKB/Swiss-Prot" (reviewed, manually annotated).
TMM	Trimmed mean of M-values.
TR	Troia.

1

INTRODUCTION

1.1 Understanding Stress in Marine Organisms

The survival of marine organisms in the face of changes in oceanic conditions is intrinsically dependent on their ability to adequately respond to stress. The concept of *stress* has been widely debated over time and is defined as any threat to homeostasis, that is, the ability of an organism to maintain a stable internal balance in the face of environmental variations (Molina & Blasina, 2025). These threats, known as *stressors*, include biotic factors such as food availability, presence of predators, pathogen infections, and intraspecific interactions, as well as abiotic factors such as temperature, salinity fluctuations, hydrodynamic conditions (e.g., river flow, tidal emersion and immersion cycles) and exposure to toxins (Schulte, 2014). Organisms detect stressful environmental shifts through their sensory systems, hence triggering a stress response (Molina & Blasina, 2025). This response can be understood as a set of physiological modifications occurring at different levels of biological organization (cells, tissue, organ, whole organism) (Schulte, 2014). Primary responses involve the activation of physiological and biochemical systems, for instance through changes in hormone levels (e.g., cortisol and catecholamines), whereas secondary responses involve the redistribution of energy toward cellular defense and repair mechanisms, and immune activation (Islam et al., 2022). Lastly, tertiary responses involve the adoption of behavioral changes (Islam et al., 2022; Molina & Blasina, 2025).

1.2 Environmental Challenges Faced by Intertidal Species

Intertidal zones represent some of the most physiologically demanding habitats on Earth (Han et al., 2019). Positioned at the interface between marine and terrestrial ecosystems, these zones are defined as the area between the lowest low tide and the highest high tide (Leeuwis &

Gamperl, 2022). Organisms inhabiting this environment are subject to extreme and recurrent fluctuations in temperature, pH, dissolved oxygen, salinity, and nutrient availability across daily and seasonal timescales (Place et al., 2012). These environmental variations are primarily driven by tidal cycles, freshwater inputs, wind stress, and thermal equilibrium, and their intensity and frequency are further modulated by geomorphological differences between intertidal and adjacent coastal environments (e.g., rocky shore, sandy/muddy shore and estuary) (Leeuwis & Gamperl, 2022) Figure 1.1.

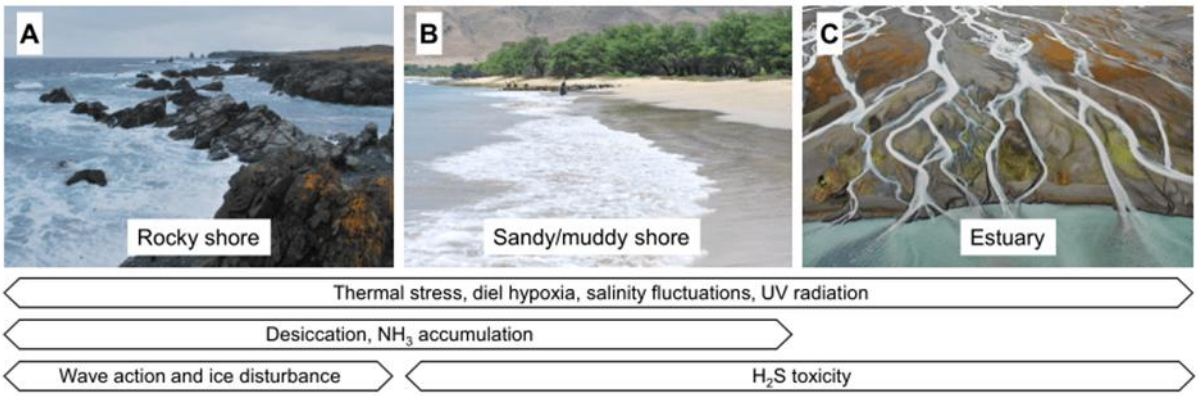


Figure 1.1 - The abiotic stressors encountered by animals depends on the type of high intertidal habitat. Source: Leeuwis & Gamperl 2022.

Faced with such natural challenges, many intertidal species have evolved coping mechanisms that sustain life under fluctuating and (sometimes) extreme conditions (Leeuwis & Gamperl, 2022). For example, the Japanese mudskipper (air-breathing, amphibious gobies of the subfamily Oxudercinae) protects their eggs from extremely hypoxic water conditions, by storing fresh air in an egg chamber. The burrow-guarding males transport mouthfuls of air during each low tide, creating an oxygen-rich environment essential for embryonic development (Ishimatsu & Graham, 2011). These kinds of evolutionary adaptations confer great ecological importance on intertidal zones, establishing them as natural laboratories for several interdisciplinary research (Nascimento et al., 2021; Rigal et al., 2008). Studying these organisms provides critical insights into the mechanisms of stress tolerance and resilience. Such knowledge is increasingly important for predicting and mitigating the effects of global climate change on marine biodiversity and coastal ecosystems (Han et al., 2019; J. Wang et al., 2023).

1.3 Climate Change (Global stressor): A New Thermal Reality

The intensification of climate change is now amplifying natural environmental stressors, often pushing species beyond their adaptive limits (Seebacher et al., 2014). Over the past century, the global mean surface temperature has risen by approximately 1.1 °C, with projections suggesting increases of 1.5 °C to 4.5 °C by the end of this century depending on emission scenarios (Calvin et al., 2023). Moreover, warming profoundly alters both the chemistry and the physical structure of the world's oceans. It drives more frequent, intense, and prolonged marine heatwaves (MHW), accelerates acidification and promotes widespread deoxygenation (Stillman & Tagmount, 2009). These challenges, either alone or in combination, pose great threats to marine organisms and coastal ecosystems.

Populations living at the edge of their geographical distribution are particularly vulnerable, as they are traditionally assumed to exist close to their upper thermal tolerance limits, with narrower thermal-safety margins than central populations. However, growing evidence suggests that local adaptation and acclimatization can elevate tolerance thresholds in rear-edge (i.e., low latitude and warm) populations, consequently, both rear-edge and central populations may have similar thermal-safety margins and sensitivity to climate impacts. If populations do not have locally adapted thermal limits, responses to high temperatures among populations are explained by absolute maximum temperature. By contrast, if populations have locally adapted thermal limits, thermal-safety margins would remain similar throughout a species range and we would expect responses to be better explained by stress anomalies, rather than the absolute temperature (Bennett et al., 2015). Likewise, organisms inhibiting habitats characterized by high natural variability, such as intertidal zones, face heightened vulnerability as climate change amplifies environmental stressors. In these areas, rapid temperature fluctuations have a considerable impact on animal's stress susceptibility (Han et al., 2019; Sasaki & Dam, 2019) particularly in ectothermic organisms whose metabolic rates and body temperatures are dictated by ambient conditions, namely seawater-air temperature during immersion and emersion periods, respectively (Elmer et al., 2025).

Evidence of climate-driven ecological change in the ocean is already widespread. Shifts in phenology, alterations in metabolic function, oxidative stress responses, and changes in behavioural patterns have been documented across marine taxa (Gerlich et al., 2025; Logan & Cox, 2020; Von Schmalensee et al., 2023). However, ocean warming is strongly heterogeneous and exhibits a pronounced latitudinal gradient (Baptista et al., 2018), with polar regions

experiencing temperature increases four times greater than the global mean, while tropical and temperate zones (harbouring most of the Earth's biodiversity) undergo warming at a comparatively slower pace (Serreze & Barry, 2011; Vinagre et al., 2011). This uneven distribution of heat in the ocean implies that ecological responses to climate change will also be spatially heterogeneous (Han et al., 2019; Sasaki & Dam, 2019). Species or populations adapted to heat and native to low-latitude regions are often predicted to be more vulnerable to climate change, as they already experience temperatures near their upper thermal limits (Han et al., 2019; Sasaki & Dam, 2019). In contrast, species from higher latitudes, particularly those inhabiting areas with high seasonal variability such as temperate ones, may exhibit greater phenotypic plasticity and thus possess a higher potential to cope with rising temperatures (Sasaki & Dam, 2019; Schaum et al., 2022). Recognizing these contrasts is essential for forecasting differential vulnerabilities and for devising targeted conservation strategies that reflect each region's unique stress profile (Sasaki & Dam, 2019).

1.4 Stress, Phenotypic Plasticity and Evolution

Rapid and often unpredictable environmental changes challenge organismal survival (Moritz & Agudo, 2013; Stollewerk et al., 2025). Predicting the coping mechanisms of marine species under these conditions requires critical understanding of how they cope, through both phenotypic plasticity and genetic adaptation, which altogether shape their resilience (i.e., the capacity of a system to persist and recover following perturbations) to climate stressors (Sydeman et al., 2015).

Phenotypic plasticity is the capacity of an organism to alter its phenotype in response to environmental input without changes in its genotype (Mallard et al., 2020; Stollewerk et al., 2025). Such plastic changes can alter morphology (e.g., body size and shape), physiology (e.g., stoichiometric homeostasis, metabolic and excretion rates, aerobic scope), and behaviour (e.g., habitat, diet selection, orientation) (Molina & Blasina, 2025; Stollewerk et al., 2025). This short-term flexibility often acts as the first line of defense against environmental stress and is one of the most important mechanisms by which organisms can respond to and survive environmental changes (i.e., acclimatization) (Clark et al., 2018; Sydeman et al., 2015). However, plasticity is neither unlimited nor uniform, it varies across individuals, populations, and species, it can itself evolve under selection, and its expression is constrained by underlying genetic architecture, developmental modules, physiological limits, and ecological context (Stollewerk et al., 2025). Such variation in plasticity arises from differences in genetic norms of reaction,

epigenetic regulation that can be transmitted across generations, ontogenesis and population specific evolutionary histories (Adrian-Kalchhauser et al., 2020; Chevin et al., 2013; Stollewerk et al., 2025).

Phenotypic plasticity can manifest in different ways, each with ecological and evolutionary implications. It may be anticipatory, when organisms respond to an environmental cue that predicts a future condition (e.g., subtle alterations in temperature and photoperiod precede seasonal change). It can also be responsive, as a direct reaction to an environmental stimulus, which may be active (via activation of specific molecular pathways) or passive (arising through environmental effects on e.g., the speed of biochemical and metabolic processes). Plasticity can further be reversible, when adjustments are labile and can be undone within the organism's lifetime, or irreversible, when changes remain fixed after a certain ontogenetic stage, Figure 1.2 (Stollewerk et al., 2025).

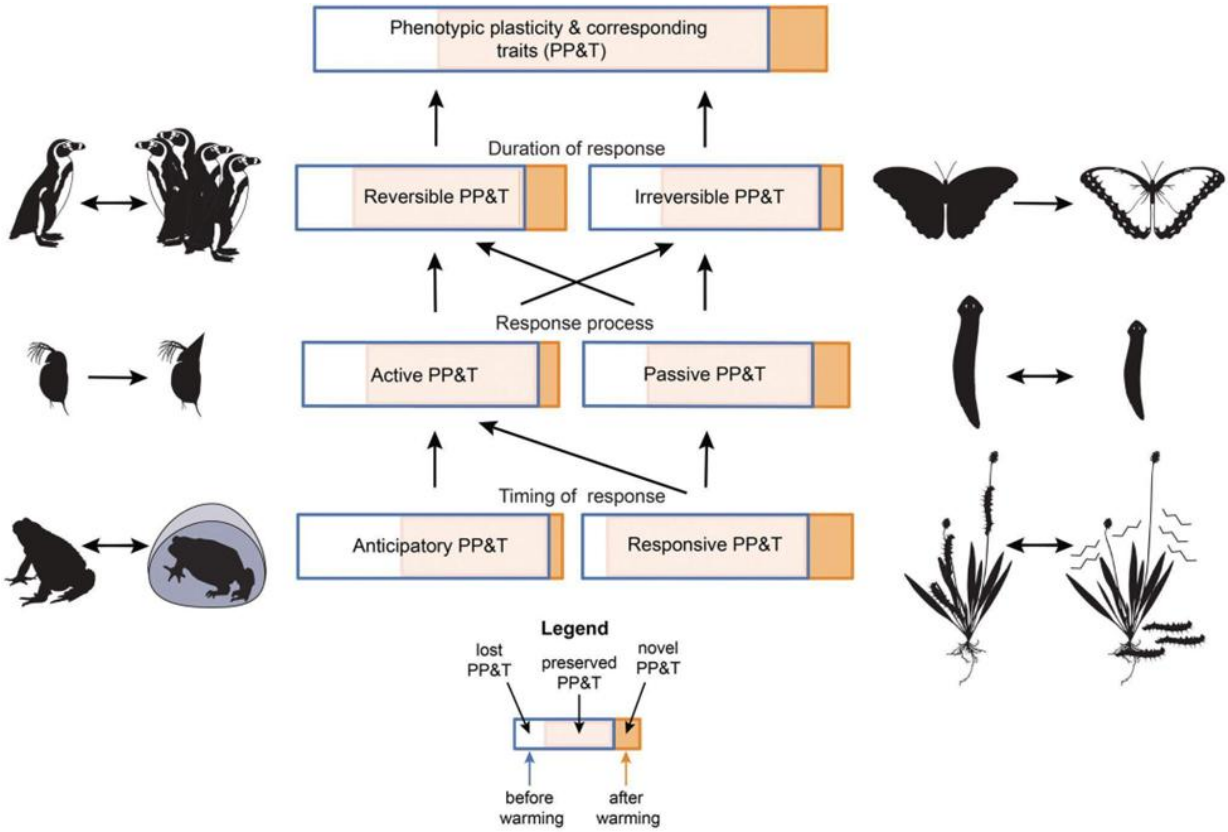


Figure 1.2 - Conceptual representation of the relationships among different forms of phenotypic plasticity and their associated traits (PP&T), as well as potential shifts induced by climate warming. Each rectangle depicts the diversity and quantitative expression of plastic traits prior to (blue, unfilled areas) and following (orange, filled areas) exposure to warming. The total area of each rectangle reflects the cumulative magnitude and number of expressed plastic traits. Arrows connecting the rectangles indicate possible transitions between distinct types of plasticity. In this hypothetical scenario, climate warming results in a net reduction in phenotypic plasticity, as a greater number of traits become non-plastic or exhibit diminished expression (lost PP&T), relative to the emergence of novel plastic traits or increased plasticity (novel PP&T). The proportion of traits that remain unaffected (preserved PP&T) is

not evenly distributed across plasticity types. For instance, anticipatory and irreversible plasticity are more likely to be lost than gained, whereas responsive plasticity may show the opposite trend. Silhouettes illustrate representative examples of plastic traits potentially influenced by climate warming: Left (top to bottom): reversible plasticity (huddling behaviour in penguins), active plasticity (defensive morphology in *Daphnia*), and anticipatory plasticity (hibernation in toads). Right (top to bottom): irreversible plasticity (seasonal pigmentation in butterflies), passive plasticity (degrowth in planarians under resource limitation), and responsive plasticity (induced toxin release in plants following herbivory). Source: Stollewerk et al., 2025.

Evolutionary adaptation is a longer-term process in which the genetic makeup of a population changes over one or more generations as a result of differential reproductive success or survival among genotypes (DeBiasse & Kelly, 2016; Sydeman et al., 2015). This mechanism becomes particularly critical under conditions of environmental stress, such as rising temperatures, and is more likely to occur in species characterized by short generation times and high fecundity, which facilitate the rapid spread of advantageous genetic variants (Moritz & Agudo, 2013; Sydeman et al., 2015). A specific manifestation of this process is local adaptation, wherein genetically differentiated populations exhibit higher average fitness within their native environments compared to foreign ones (Porcelli et al., 2015). This form of adaptive response is particularly prevalent in species with broad geographic distributions. In populations characterized by sufficiently low gene flow, divergent selection pressures across distinct environments can drive local adaptation (Porcelli et al., 2015).

The interaction between stress, plasticity, and evolution is a central theme in contemporary ecological and evolutionary biology. Stress can act as both a constraint and a driver of adaptation (Moritz & Agudo, 2013). On one hand, persistent stress may reduce fitness, divert energy from growth and reproduction, and increase mortality (Jesus et al., 2016). On the other hand, stressful conditions can act as evolutionary forces, leading to adaptative changes in populations (Madeira et al., 2014). The intensity and duration of stress are critical determinants of these outcomes: mild or short-term stress can induce beneficial plastic responses that enhance tolerance, whereas chronic stress or recurrent stress, as well as multi-stressors often overwhelm protective mechanisms, leading to harmful effects such as impaired physiological function, reduced reproductive success, and long-term population decline (Glibert et al., 2022; Marigómez et al., 2017; Sommer et al., 2024).

1.5 Advancing Our Understanding of the Stress Response: Transcriptomics

Traditional investigations of adaptation and phenotypic plasticity in marine organisms have predominantly employed physiological assays, such as measurements of metabolic rates and enzyme activities (e.g., the glycolytic enzyme lactate dehydrogenase and the mitochondrial enzymes citrate synthase and cytochrome c oxidase) (Seebacher et al., 2014), as well as assessments of thermal tolerance limits including critical thermal maximum (CTmax) and minimum (CTmin), and thermal performance curves (Elmer et al., 2025; Missionário et al., 2024; Sasaki & Dam, 2019). Biochemical analyses, including the quantification of heat shock proteins (HSPs), oxidative damage markers, and antioxidant enzyme activity, have also been extensively utilized (Chen et al., 2022; Han et al., 2019; Stillman & Tagmount, 2009). In addition, behavioural traits such as locomotor performance, thermoregulatory behaviours, and reproductive cycles are frequently analysed to provide complementary perspectives on organismal responses to environmental stress (Angiulli et al., 2020; Bertolo et al., 2011). While powerful, these approaches cannot fully capture the molecular basis of plasticity and adaptation. Omics technologies – encompassing genomics, transcriptomics, proteomics, and metabolomics — have revolutionized this field by enabling a better understanding on how the molecular mechanisms that underpin production traits such as growth rate, resilience to environmental stressors or how diseases are developed (Rise et al., 2019). Among these approaches, transcriptomics can reveal details of an organism's physiological regulation and function by revealing gene expression changes and gene networks that cannot be discerned through more targeted approaches or traditional phenotypic response measurements, providing insight into the regulation of key cellular processes, molecular pathways, and functional mechanisms through specific gene expression (Page & Lawley, 2022).

1.6 Tools to Address the Challenges of Reduced Genomic Resources for Non-model Organisms

Despite the remarkable advancements, omics technologies still face several limitations. A significant technical limitation is the scarcity of genomic information for aquatic species (Zhang et al., 2025). Non-model organisms – organisms of substantial ecological or evolutionary importance but with limited prior characterization – often lack high-quality reference genomes (Haas et al., 2013). In such cases, *de novo* assembly strategies enable the reconstruction of

transcriptomes directly from RNA-seq reads without the need for a reference genome (Grabherr et al., 2011; Haas et al., 2013). Once the sequences are obtained, they must be annotated to uncover their potential biological functions (Raghavan et al., 2022). The most effective approach typically involves a sequence homology search to compare the new sequences against those from phylogenetically related species with a well-characterized genome. However, for many newly studied transcriptomes, such well-annotated reference species are unavailable, posing a significant challenge to accurate functional prediction. In such cases, using the latest nonredundant protein database is an appropriate alternative (Haas et al., 2013).

Bioinformatics tools help overcome these obstacles. For example, in mixed-species samples, tag- or hybridization-based methods are used to simultaneously detect the response of both the host and the pathogen (Westermann et al., 2012). Functional insights can be derived using protein interaction networks (e.g., STRING), which integrate diverse data sources (Szklarczyk et al., 2023).

1.7 Field Case Study: The Portuguese Coast

As mentioned earlier, although climate change is a global phenomenon, its effects are unevenly distributed, with some regions being particularly vulnerable due to their geographic and climatic characteristics (Baptista et al., 2018). The Iberian Peninsula (a landmass in southwestern Europe primarily occupied by Spain and Portugal) has been identified as one of Europe's most climate-sensitive regions (Böhnisch et al., 2021). Situated at the interface of Atlantic and Mediterranean climate systems, the region is subject to strong seasonal variability, recurrent droughts, and frequent heatwaves, all of which are expected to intensify under future scenarios (Pereira et al., 2021). Portugal, in particular, experiences different climatic regimes: dry, hot summers under Mediterranean conditions (Csa) along the southwestern and southern coast and cooler, more humid Atlantic-influenced summers along the north- and central-western coasts (Csb) (*Atlas Climático Ibérico*, n.d.) Figure 1.3.

This fine-scale climatic mosaic compresses environmental gradients over relatively short geographic distances, exposing local populations to strikingly different thermal and hydrological regimes (Cardoso et al., 2019). Thus, this creates a rather suitable north-south gradient with ideal environmental characteristics for latitudinal field studies. Additionally, the particularly high vulnerability of Iberian marine systems to climate change and extreme weather events makes them an ideal natural laboratory for testing hypotheses related to adaptation, plasticity, and evolutionary potential.

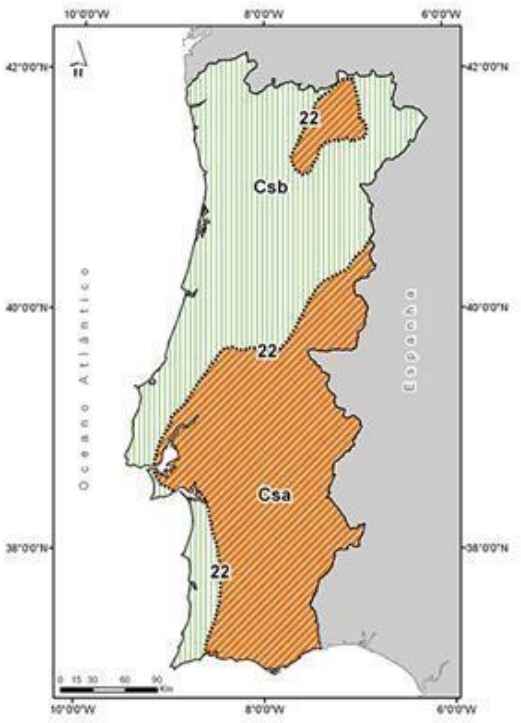


Figure 1.3 - Clima de Portugal Continental, segundo a classificação de Koppen. Source: ipma.pt

1.8 *Pomatoschistus microps* as a Model Species

The common goby, *Pomatoschistus microps* (Krøyer, 1838) Figure 1.4, is a gobiid fish abundant in estuaries, lagoons, and coastal waters and is distributed from the Baltic Sea to the western Mediterranean (Missionário et al., 2024; Salgado et al., 2004), displaying remarkable tolerance to a wide range of temperatures and salinities (i.e., eurythermal and euryhaline, respectively) (Dolbeth et al., 2007; Missionário et al., 2024). Its ecological relevance as an intermediate predator — consuming plankton and small invertebrates while serving as prey for larger fish and birds — combined with key biological traits (e.g., a 1–2 year lifespan, small body size, and ease of collection), establishes *P. microps* as an exceptional model organism for environmental research (Dolbeth et al., 2007; Missionário et al., 2024). Additionally, its abundance in estuaries and shallow coastal waters, which experience high temperature fluctuations, renders it a relevant species for climate change research, particularly the investigation of acclimation and adaptation processes to abiotic stressors. While some studies suggest that *P. microps* is well adapted to temperature in its natural environment, exhibiting a large thermal safety margin (defined by the difference between the organism's thermal tolerance and the maximum habitat temperature, see Missionário et al., 2024; and Sunday et al., 2014) and demonstrating some adaptive mechanisms (e.g., variations in reproductive effort and egg size,

see Dolbeth et al., 2007); it has also been reported to exhibit very low acclimation capacity in heat tolerance and cellular stress responses during environmental changes (Missionário et al., 2024). This potential vulnerability reinforces the necessity of understanding the underlying molecular adaptation mechanisms towards a changing ecosystem to develop predictive models of natural population trajectories, both in terms of distribution range expansions or contractions, biomass and local extinction risk (Dolbeth et al., 2007).

Despite this, *P. microps* is still an unconventional model for molecular studies. Its genome has been sequenced after recent joint efforts from multiple consortia (e.g., Earth BioGenome Project, Vertebrate Genomes Project) to sequence a vast diversity of organism genomes, including marine ones. However, *P. microps* genome deposited in GenomeArk database (available from https://www.genomeark.org/genomeark-all/Pomatoschistus_microps.html) does not yet provide the genome assembly or annotation for the species. With the current limitations in the omics resources available, this species poses specific challenges in terms of data analysis, requiring adapted computational pipelines, and caution in biological interpretation of the results.

Figure 1.4 - *Pomatoschistus microps*. Credits: Julien Renoult, some rights reserved (CC BY). Source: iNaturalist (https://www.inaturalist.org/photos/8568065).

2

OBJECTIVES AND HYPOTHESES

This thesis investigates the molecular mechanisms that shape how *Pomatoschistus microps* populations respond to local environmental regimes and seasonal thermal variation. Using a functional genomics approach, it aims to identify genes and genetic pathways that underly phenotypic plasticity, adaptive capacity, and ultimately resilience in this intertidal fish. To address this aim, the work pursues two complementary objectives. First, it compares transcriptomic profiles across three fish populations distributed along a latitudinal thermal gradient (Ria de Aveiro, Troia, and Ria Formosa) and across two seasonal regimes (spring and summer). This comparison will reveal whether populations and seasons are associated with distinct patterns of differential gene expression that may reflect local adaptation or short-term acclimation. Second, it identifies the canonical pathways and biological processes that are activated or suppressed under seasonal warming. By examining enrichment patterns across the transcriptome, the study will determine whether variation in gene expression translates into functional differences in physiological pathways linked to thermal tolerance and resilience.

The central hypothesis is that *P. microps* populations from different latitudes exhibit divergent transcriptomic responses to seasonal warming, reflecting alternative physiological strategies for coping with environmental variability and thermal stress. From this hypothesis arise two guiding questions:

1. How do seasonal thermal regimes influence transcriptomic stress responses in *P. microps*?
2. To what extent do populations along a latitudinal gradient differ in their transcriptional responses to local environmental conditions, particularly thermal regimes?

By linking genomic responses to ecological context, this thesis aims to advance understanding of how intertidal species tolerate and adapt to both long term local conditions as well as rapid environmental changes. Beyond its contribution to fundamental macrophysiological, ecological and evolutionary knowledge, the findings hold applied relevance for biodiversity conservation, with potential extrapolations to aquaculture, and fisheries management, where the identification of molecular variants that enhance tolerance or allow the persistence of populations may inform conservation strategies or selective breeding programs for aquaculture.

This work is supported by ExtremeOceans - Unravelling evolutionary physiology landscapes of coastal marine fauna under extreme temperatures using a multi-layer Systems Biology approach (PTDC/BIA-BMA/1494/2020), a multidisciplinary collaborative project that employs a multi-layer Systems Biology approach to move the discovery of climate change physiology forward, beyond effects-driven research and into the molecular and cellular mechanisms controlling physiological systems of marine organisms under temperature stress.

3

METHODOLOGY

The methodologies described represent a collaborative effort between NOVA-FCT, University of Aveiro and CCMAR within the ExtremeOceans project (FCT Ref. PTDC/BIA-BMA/1494/2020). Specifically, animal collection and sample harvesting was performed by all three institutions. The comprehensive transcriptomics pipeline – which includes RNA extraction, pre-processing, RNA-sequencing, transcriptomics assembly and functional annotation (detailed in Methods Sections 3.1 to 3.4) – was carried out by project members from the SeaTox research group at NOVA-FCT.

3.1 Animal Collection and Sample Harvesting

Specimens of the common goby *Pomatoschistus microps* (n_{total}=60 adult fish) were collected from three distinct coastal systems in Portugal: Ria de Aveiro (40°37′44″ N, 8°44′42″ W; northwestern coast), Troia Peninsula (38°29′07″ N, 8°53′15″ W; central western coast), and Ria Formosa (37°00′ N, 7°59′ W; southern coast). Sampling was performed during two seasonal periods in 2021: spring (May) and summer (July), Figure 3.1a-b.

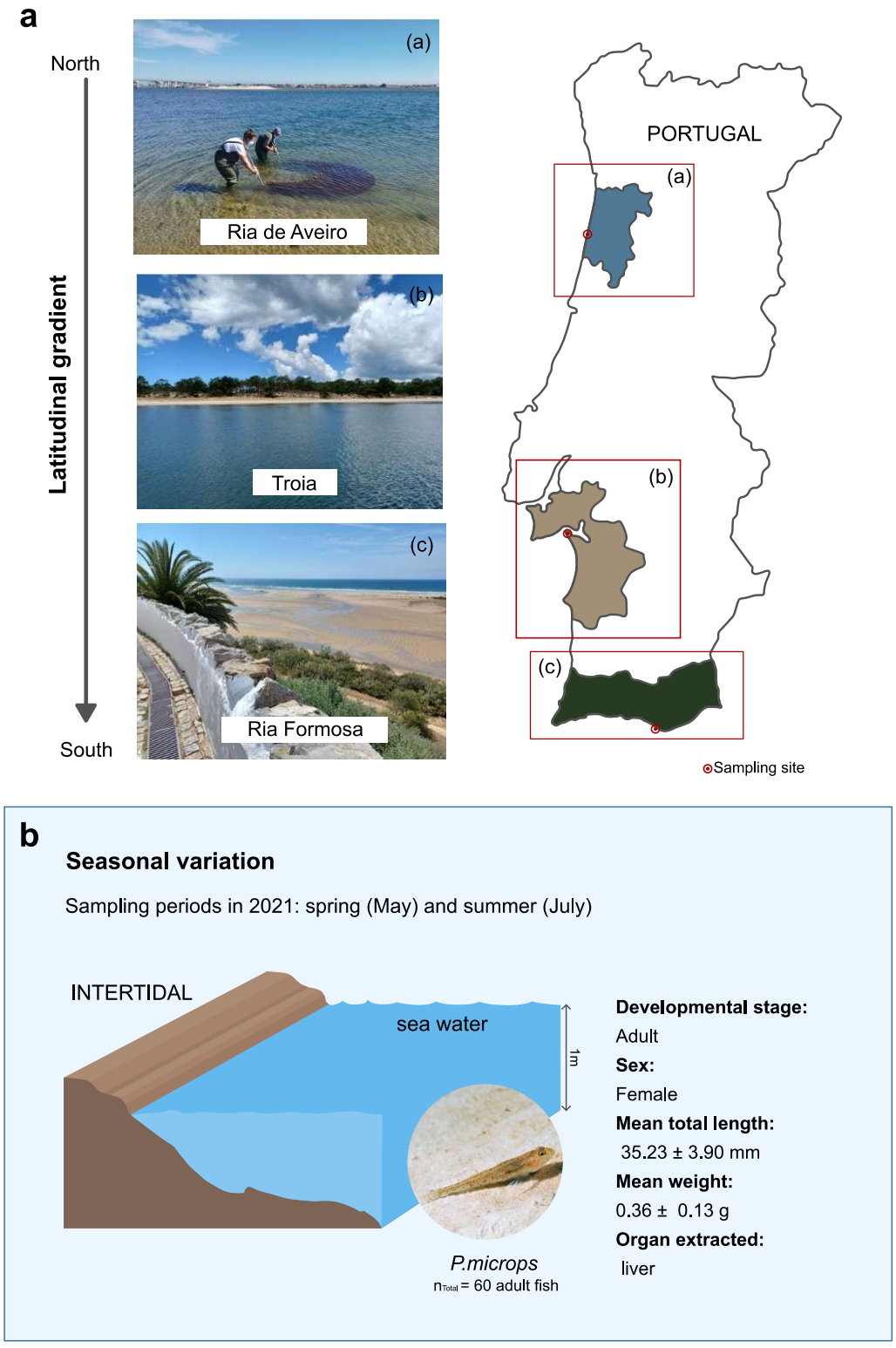


Figure 3.1 - Experimental design. Panel (a) illustrates the latitudinal gradient of sampling sites along the Portuguese coast, including: (a) Ria de Aveiro (north), (b) Troia Peninsula (central), and (c) Ria Formosa (south). Panel (b) provides an overview of the intertidal sampling environment and biological specifications of the target species (*Pomatoschistus microps)*. Details include developmental stage, sex, mean total length (cm), mean body weight (g), the specific organ used for RNA extraction, and the corresponding sampling periods. Figure made using Inkscape.

At each site and season, in situ environmental parameters were measured in shallow waters (approximately 1 m depth). These included water temperature (°C), salinity, dissolved oxygen (mg L^{-1}), and pH, as well as air temperature (°C). Full records of local environmental conditions are provided in Appendix A (Table A.2).

A total of 10 individuals per location per season were collected using hand-operated trawl nets deployed in shallow coastal habitats at approximately 1 m depth. Captured individuals were immediately transferred to aerated seawater tanks for transport to the laboratory. Subsets of specimens were allocated to different purposes, including histopathological and molecular analyses, within the context of the ExtremeOceans project.

For transcriptomic profiling, 18 adult females were selected (3 biological replicates from each of the three locations and both seasons). Only fish of uniform size (mean total length 35.23 ± 3.90 mm; mean weight 0.36 ± 0.13 g) were considered. Females were selected because they could be sampled consistently across sites and seasons.

3.2 RNA Extraction and Pre-processing

Total RNA was extracted from 5-15mg of liver tissue using the AllPrep DNA-RNA Mini kit (Qiagen, Hilden, Germany) following manufacturer's guidelines. Residual DNA was eliminated on-column using the RNase-Free DNase set (Qiagen, Hilden, Germany).

RNA quantity and purity were initially assessed by spectrophotometry (Nanodrop ND 1000, Thermo Fisher Scientific, Waltham, MA, USA). Integrity was subsequently verified on an Agilent TapeStation 4200, where integrity numbers (RIN) were recorded. Only samples with RIN > 5 were retained for library preparation.

3.3 RNA-seq and Transcriptome assembly

cDNA libraries were sequenced in the Illumina Novaseq platform (Illumina, San Diego, CA, USA) with 150 bp paired-end reads (short-reads) and produce 40 million (M) reads. Integrity verification, library preparation and sequencing were performed by StabVida (Caparica, Portugal).

Raw sequencing reads were initially quality checked with FastQC (v0.11.9). For all libraries, adapter sequences and low-quality bases were removed using TrimGalore (v.0.6.0) with default parameters, discarding reads shorter than 20 bp after trimming.

Because no *P. microps* transcriptomes were available in public databases for read mapping, we performed *de novo* transcriptome assembly using Trinity v2.14.0. All samples (40 M

reads) were used to construct the reference transcriptome. Assembly quality was evaluated with multiple metrics, including contig N50, Ex90N50, and read content. Assemblies are generally considered high-quality when the Contig N50 and Ex90N50 exceed 1000 bp, and the read content is approximately 70–80% (Seyednasrollah et al., 2013). Read mapping statistics were generated with Bowtie2 v2.4.2 and Samtools v1.8.

To assess the individual expression levels, reads from all samples were pseudo-aligned to the respective assembled transcriptome and transcript abundance quantified using Kallisto v0.48.0 and imported into R v4.5.0 via the tximport package 1.37.0.

3.4 Functional Annotation

Protein-coding regions were predicted from the assembled transcriptome using Transdecoder v5.5.0, with default parameters. Predicted ORFs were then functionally annotated by homology-matching against three protein databases: (i) The complete Swiss-Prot database (release 2023-04), providing a well-curated, broad range of high-confidence annotations, (ii) A zebrafish-specific subset of Swiss-Prot (*Danio rerio*) enabling functional inference based on the best-annotated teleost genome available (Tan et al., 2016). Although *P .microps* and *D. rerio* are from phylogenetically distant lineages within the teleost's, zebrafish remains the closest taxon with a thoroughly curated genomic resource, and is therefore widely adopted as a reference for less characterized fish species (McCluskey & Postlethwait, 2014) and (iii) A *platyhelminthes*-specific subset of Swiss-Prot, to identify potential parasite associated transcripts (parasite presence in these populations confirmed histologically; Guerreiro *et al.* (unpublished)).

Homology-matching between sequences were performed with BLASTp from NCBI BLAST+ v2.12.0 using an *e*-value cut-off of 1E-5 and default parameters. The use of multiple reference databases aimed to maximise annotation accuracy.

Annotation results from the three databases were integrated separately with Kallisto output, resulting in three independent annotated datasets corresponding to each database (henceforth also referred to as the Swiss-Prot dataset, the Zebrafish dataset, and the Plat dataset). Following the merge between the annotation results and the Kallisto output, redundant entries were removed, retaining only the highest-confidence hits for each transcript (smallest *e*-value and bigger %ID).

These datasets were retained for downstream analysis, with the Swiss-Prot dataset primarily used in differential expression and assessment of stress-related candidate genes (detailed in Methods Sections 3.5 and 3.6), the Zebrafish dataset primarily used in gene set

enrichment analysis, protein–protein interaction networks (detailed in Methods Sections 5, 6, 7, 8 and 9), and the Plat dataset used to identify potential parasite associated transcripts (detailed in Methods Sections 3.5).

3.5 Exploratory Data Analysis

Annotation quality was then assessed using a structured, multi-step analytical workflow. Alignment quality metrics – percentage identity (%ID) and expectation values (*e*-values) – were analysed for all the data cleaning steps for each dataset. Higher %ID values indicate greater sequence similarity, while lower *e*-values reflect more statistically significant alignments. Distribution of these metrics were examined to evaluate overall annotation confidence.

Taxonomic composition was characterized by quantifying the proportion of matches assigned to fish species and Platyhelminthes, based on the BLAST hit taxonomy. Putative parasite transcripts were identified using a comparative annotation approach: transcripts were classified as parasite-associated if they had high sequence similarity (%ID > 70) and strong alignment significance (*e*-value < 1E-70) in both the Swiss-Prot and *Platyhelminthes*-specific datasets, and if the match was taxonomically assigned to a *platyhelminthes*- species. Nevertheless, we acknowledge that many genes are highly conserved across metazoans, limiting the certainty of definitive taxonomic origin (Borner & Burmester, 2017). Consequently, transcripts identified here as "parasite-associated" should be interpreted as putative parasite sequences, with classification strengthened by consistent top hits to *platyhelminthes*- taxa across datasets.

Principal component Analysis (PCA) was conducted with the PCAtools package v2.21.0 in R to investigate potential population variance. The input matrix comprised $\log_2$ transformed transcripts abundance.

For targeted analyses, abundance matrices were subset according to biological categories of interest, such as (i) putative parasite associated transcripts and (ii) taxonomically assigned sequence matches to fish species. For each subset, a covariance matrix was computed, and principal components were extracted via eigenvalue decomposition. The first two principal components (capturing the largest variance proportions) were visualized to assess clustering patterns. A global PCA with all the abundance data was also done.

To facilitate visual interpretation of groupings in the PCA plots, ellipses were drawn around clusters of samples using the *ggforce::geom_mark_ellipse()* function. These ellipses were used solely as graphical guides to represent the spatial extent of groups and were not intended to imply statistical significance.

3.6 Differential Gene Expression Analysis

Following verification of functional annotation quality, differential expression analysis was performed using edgeR v4.7.2. (Y. Chen et al., 2024). This package was selected over alternatives (such as DESeq2, limma-voom, NOISeq, and Cuffdiff) due to its computational speed, higher power of detection, and its statistically rigorous negative binomial modelling framework, which provides reliable performance even with small numbers of replicates (Seyednasrollah et al., 2013).

Only transcripts annotated against Swiss-Prot and Zebrafish reference databases were included in the analysis. Transcripts considered parasite-associated (criteria previously described in Section 3.5 – Exploratory data analysis) were excluded to ensure a focus on host-derived expression, which was the model organism of interest in the present study.

For each annotation set a raw count matrix, with genes as rows and samples as columns, was generated. Genes with low expression across samples were filtered using the *edgeR::filterByExpr()* function, which improves statistical power by removing features unlikely to be informative (Chen et al., 2024).

Library sizes were normalized using the trimmed mean of M-values (TMM) method. Dispersion parameters were estimated within edgeR's empirical Bayes framework (Chen et al., 2024). Differential expression testing was performed using generalized linear models (GLMs) fitted to the experimental design, which include nine contrasts (Table 3.**1**). These contrasts were designed to assess (i) seasonal differences within each site and (ii) site-specific differences within each season.

Table 3.1 - Multiple comparisons used to assess differences in gene expression. Group comparisons were defined as nine different contrasts that encompass seasonal differences at each location and population differences in each season.

Contrast	Season	Site	Comparison
C1	Spring vs. Summer	Ria de Aveiro	Seasonal difference
C2	Spring vs. Summer	Troia Peninsula	Seasonal difference
C3	Spring vs. Summer	Ria Formosa	Seasonal difference
C4	Spring	Ria de Aveiro vs. Troia	Site differences
C5	Spring	Ria de Aveiro vs. Ria Formosa	Site differences
C6	Spring	Troia vs. Ria Formosa	Site differences
C7	Summer	Ria de Aveiro vs. Troia	Site differences
C8	Summer	Ria de Aveiro vs. Ria Formosa	Site differences
C9	Summer	Troia vs. Ria Formosa	Site differences

Statistical significance was assessed using the quasi-likelihood F-test, providing a more conservative and rigorous type I error rate control (Chen et al., 2025). *P*-values were adjusted for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) procedure. Genes with FDR < 0.05 were considered significantly differentially expressed.

Quality assessment of the input count data and model outputs included principal component analysis (settings previously described in Section 3.5 – Exploratory data analysis). These diagnostics were used to detect outliers and assess the presence of batch effects.

For visualization, volcano plots were generated to display $\log_2$ fold changes ($\log_2$FC) versus -$\log_{10}$(*p*-values), highlighting significant genes (FDR < 0.05). Heatmaps were produced with pheatmap v1.0.13, using normalized expression values, with hierarchical clustering applied to both genes and samples. The heatmap colour gradient represents relative expression levels.

Differentially expressed genes (DEGs) obtained from each contrast were further examined for associations with a predefined list of biologically relevant keywords: *heat, stress, thermal, oxygen, salinity, HSP, hypoxia, adaptation, oxidative stress, and acid-base balance.*

Functional annotations for all DEGs were retrieved from Uniprot using *UniprotR::GetProteinFunction()* function. Keyword mapping was conducted by scanning the UniProt-derived functional descriptions for occurrences of these terms, using case-insensitive pattern matching and permitting partial matches to capture term variants (e.g., "thermally"). In cases where multiple keywords matched a single gene, only the first match in the predefined keyword list was retained.

For each keyword, the number of associated over- and under-expressed DEGs was quantified, and their corresponding $\log_2$FC values were extracted. This data was visualized by (i) a

matrix of pie charts, where each chart represents the proportion of DEGs over- and under-expressed assigned to the specific keyword; and (ii) bar plots displaying the log_2FC distribution for all the DEGs associated with a specific keyword.

3.7 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was performed using multiGSEA R package v.1.19.0. Zebrafish differential expression results were used as input, including log_2 fold changes and associated *p*-values for each gene.

Genes were ranked according to the signal-to-noise metric, combining the direction (sign) of the log_2 fold change and the statistical significance (*p*-value) of differential expression, as implemented by fgsea (Canzler & Hackermüller, 2020). Gene sets were obtained from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database, restricted to canonical pathways, and mapped to *Danio rerio* Entrez Gene identifiers with org.Dr.eg.db R package v3.21.0.

Enrichment scores for each pathway were calculated by the multiGSEA algorithm, and statistical significance was assessed using permutation testing (n = 1,000 permutations) with Benjamini-Hochberg false discovery rate (FDR) correction. Pathways with an adjusted FDR < 0.05 were considered significantly enriched. For visualization, a bubble plot was generated to display enrichment scores for each pathway in the two most relevant contrasts. Bubble colour represents the normalized enrichment score, with blue indicating negative enrichment and red indicating positive enrichment. Bubble size corresponds to statistical significance, with larger dots indicating lower FDR-adjusted *p*-values (i.e., more significant results).

3.8 Protein-Protein Interactions Networks

To investigate the potential mechanistic relationships among the Zebrafish dataset DEGs, direction specific PPI networks were constructed separately for over- and under-expressed gene sets identified in each contrast. Network analysis was conducted using the STRING web-based database (version 12.0; https://string-db.org/), with *Danio rerio* as the selected model organism.

STRING integrates evidence from multiple data sources, including text mining, experimental data, curated databases, gene co-expression, genomic neighbourhood, gene fusion events, and co-occurrence analyses. Only interactions with a minimum required confidence score of 0.700 (high confidence) were retained.

Network clustering was performed using the Markov Cluster Algorithm (MCL) with an inflation parameter of three to optimize cluster granularity. Clusters containing fewer than three proteins were excluded from downstream analysis.

3.9 Data Visualization and Code Availability

Figures were produced with ggplot2 v3.5.2 and complementary packages. Custom layouts and composite figures were arranged using Inkscape v1.4. Figure 3.2 presents a schematic overview of the transcriptomics pipeline used in this study, summarizing the workflow from specimen collection to functional and network analyses.

All data processing and visualization scripts used in this study are available in the GitHub repository https://github.com/ak-duque. A comprehensive summary table listing all software, R packages, and their versions – organized by analysis stage – is provided in Appendix A (Table A.1).

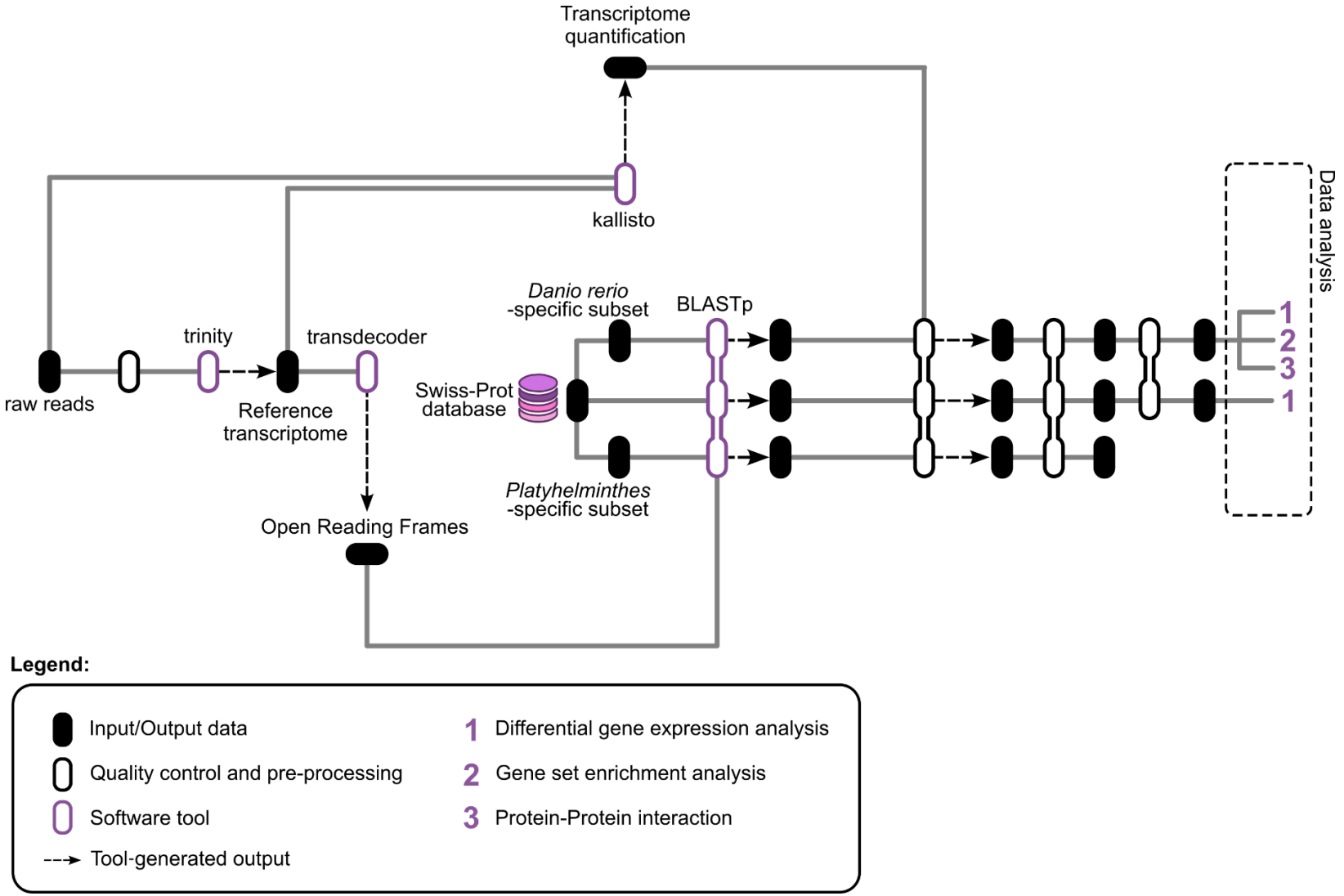


Figure 3.2 - Schematic overview of the transcriptomics pipeline applied in this study. Overview of the methodological workflow used in this study, from *P. microps* specimen collection across three Portuguese coastal habitats (spring and summer 2021) to RNA-seq, transcriptomics assembly, functional annotation, Differential gene expression analysis, Gene set enrichment analysis and Protein-Protein interactions networks.

4

RESULTS

4.1 Transcriptome Profiling of Liver Tissue in *P. microps*

The assembled transcriptome from the liver of *Pomatoschistus microps* resulted in a total of 487,125 transcripts. Among these, 212,794 transcripts were identified as potential open-reading frames (ORFs). Through homology-matching against the complete Swiss-Prot database 161,919 ORFs were annotated, representing (76.1%) of the total transcripts.
Further ORF annotation against the two curated Swiss-prot subsets showed that: (i) within the zebrafish-specific dataset 36.8% of ORFs (78,410 sequences) were annotated, and (ii) within the *platyhelminth*-specific dataset, 1.8% of ORFs (3,743 sequences) were annotated.

After merging the annotation results with the Kallisto output and removing redundant entries (retaining only the highest-confidence hits, i.e., lowest *e*-value and highest %ID), the final datasets comprised 15,937 unique Swiss-Prot annotations, 2,974 zebrafish-specific annotations, and 145 *platyhelminth*-specific annotations.

Nine transcripts displayed concordant high-confidence hits to *platyhelminth* species across both the Swiss-Prot and *platyhelminth* datasets (*e*-value < 1E-70; %ID > 70). These included conserved genes such as ACT2, TPM2, and SM20 (Table 4.1) (Denz et al., 2004; Moschella et al., 1999; Schwob & Martin, 1992). As these hits likely represent non-host origin, all *platyhelminth*-associated transcripts were excluded from downstream host expression analyses.

Table 4.1 - High-confidence transcripts with consistent annotations to *platyhelminth* species in both Swiss-Prot and *platyhelminth*-specific datasets.

	Swiss-prot		*Platyhelminthes*		
Gene Name	**%ID**	***e*-value**	**%ID**	***e*-value**	**Organism**
ACT2	98	0	98	0	*Schistosoma mansoni*
AT1A	79	0	79	0	*Taenia solium*
ENO	81	0	81	0	*Fasciola hepatica*
MYPH	75	1,21E-105	75	4,5E-109	*Echinococcus granulosus*
MYSP	83	0	83	0	*Schistosoma japonicum*
PFKA	74	0	74	0	*Schistosoma mansoni*
PGK	78	0	78	0	*Schistosoma mansoni*
SM20	94	7,68E-98	94	2,86E-101	*Schistosoma mansoni*
TPM2	90	3,2E-173	90	1,19E-176	*Schistosoma mansoni*

To investigate potential population structure and overall variance, PCA was performed on the three annotated dataset (complete Swiss-Prot, zebrafish-specific, and *platyhelminth*-specific).

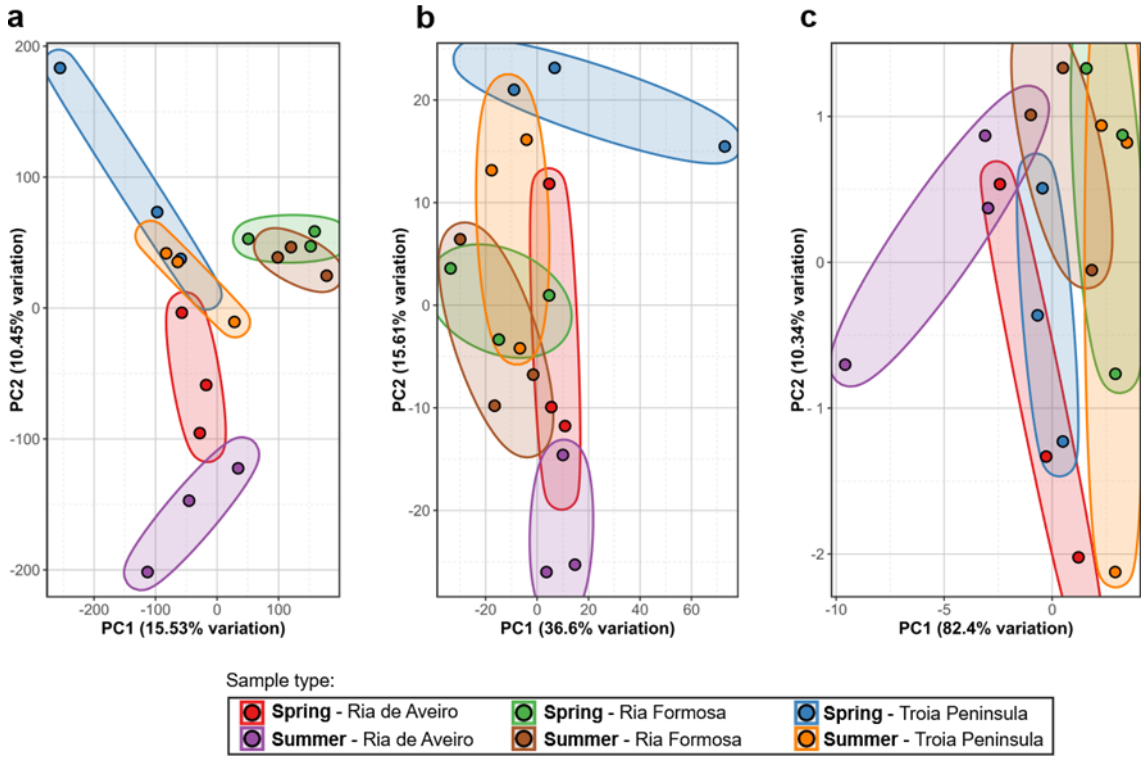


Figure 4.1 - Principal component analysis (PCA) of *P. microps* females' liver transcriptomic profiles. Based on (a) Swiss-Prot annotated transcripts, (b) taxonomically assigned sequence matches to fish species, and (c) platyhelminth-associated transcripts. Axes indicate the proportion of variance explained by the first two principal components (PC1 and PC2), with ellipses used solely as graphical guides to represent the spatial extent of groups and do not imply statistical significance.

PCA analysis performed on the complete Swiss-Prot dataset (Figure 4.1a) revealed clear clustering of samples by both site and season. The PC1, which explained 15.53% of the total variance, effectively separated spring and summer samples from the Troia site from all other locations. In contrast, the PC2, accounting for 10.45% of the variance, primarily captured the separation between spring and summer samples within the Ria de Aveiro site.

When analysis was restricted to fish matches, population clusters became less pronounced, with most populations partially overlapping along PC1 (36.6% variation). Spring Troia and Summer Ria de Aveiro samples were more separated from the other sites, along PC2 (15.61% variation) (Figure 4.1b).

In the case of *Platyhelminth*-associated transcripts (Figure 4.1c), PCA did not reveal clear clustering patterns despite PC1 and PC2 together explaining 93.7% of the variance, suggesting substantial inter-individual variability within this dataset.

4.2 Population level differences in Gene Expression

Differential gene expression analysis was conducted using edgeR with a false discovery rate threshold of less than 0.05. Following the removal of lowly expressed genes, the analysis retained 9,655 genes annotated in the Swiss-Prot dataset and 2,473 genes from the zebrafish-specific dataset to be tested.

Differentially expressed gene counts varied across nine predefined contrasts, encompassing both seasonal and site comparisons (see Methods Table 3.1), and while the magnitude of DEGs differed between annotation sets, the overall patterns of differential expression remained consistent (Table 4.2).

Table 4.2 - DEG counts identified per contrast. Contrasts C4-C6 refer to population comparisons in Spring, whereas contrasts C7-C9 refer to population comparisons in summer, namely, Ria de Aveiro vs. Troia, Ria de Aveiro vs. Ria Formosa and Troia vs. Ria Formosa and Troia vs. Ria Formosa, respectively, for both cases.

	Spring			**Summer**		
	C4	**C5**	**C6**	**C7**	**C8**	**C9**
Complete Swiss-Prot	36	40	1012	308	494	192
Zebrafish-specific	13	24	279	12	116	78

Principal component analyses restricted to DEGs further validated the transcriptomic profile PCA patterns (Figure 4.2). For both annotation sets, PC1 and PC2 together explained approximately 58% of the total variance, an improvement compared to the previous analysis.

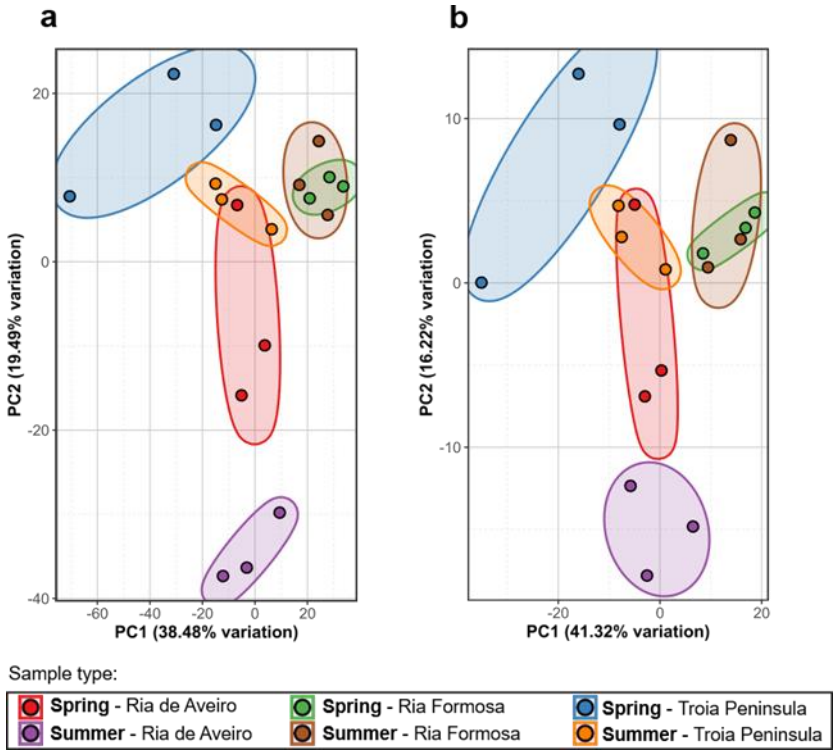


Figure 4.2 - Principal component analysis of differentially expressed genes. PCA plots Illustrate sample clustering based on transcriptomics profiles derived from differentially expressed genes. Each plot corresponds to a specific curated database: (a) swiss-prot (b) zebrafish-specific. Axes indicate the proportion of variance explained by the first two principal components.

Importantly, the separation between spring and summer groups at Ria de Aveiro, previously detectable along PC2 in the complete Swiss-Prot dataset (10.45% variance explained), became more distinct when using only DEGs, with PC2 accounting for 19,49% of the variance. Although the proportion of explained variance remained low, this seasonal divergence was supported by hierarchical clustering on heatmap analyses on the DEGs obtained when comparing spring and summer in Ria de Aveiro, which confirmed that summer and spring samples from Ria de Aveiro occupy opposite clusters (Figure 4.3).

Since almost all DEGs identified in the zebrafish-specific dataset were also captured in the more comprehensive Swiss-Prot dataset and given that Swiss-Prot represents a well-curated and reliable annotation, downstream interpretations focused primarily on the DEGs from the Swiss-Prot dataset. Nonetheless, PCA results from the zebrafish-specific DEGs are also presented, as these data from the basis for subsequent functional analyses.

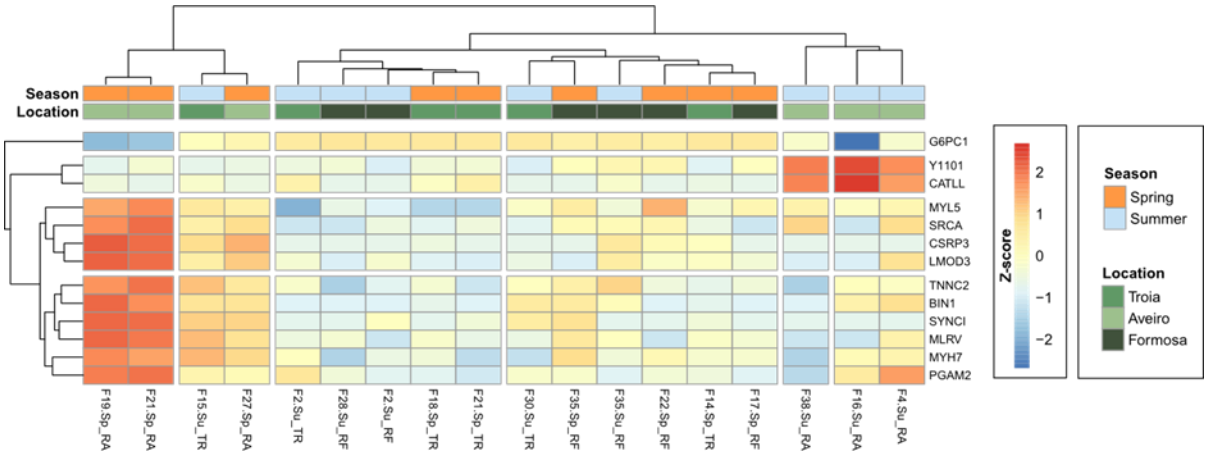


Figure 4.3 - Heatmap of relative gene expression for 13 differentially expressed genes (shortlisted from seasonal differences in Ria de Aveiro). Relative gene expression changes are shown by different colours: blue for lower relative gene expression and red for higher relative gene expression. The horizontal dendrogram illustrates hierarchical clustering of the eighteen samples, whereas the vertical dendrogram shows the association between genes.

When site-specific differences were examined by season, the strongest spatial divergence was observed in spring between Troia and Ria Formosa, yielding 1,012 DEGs, representing approximately 10.5% of the tested genes. Of these, 299 were under-expressed in Troia (and thus overexpressed in Ria Formosa), while 713 were overexpressed in Troia (under-expressed in Ria Formosa). These results match our previous findings along PC1 in the Swiss-Prot dataset, evidencing separate clusters for these populations and explaining 38.48% of the variance (Figure 4.2a). By contrast, the spring comparisons of Ria de Aveiro versus Troia and Ria de Aveiro versus Ria Formosa produced far fewer DEGs, with only 36 and 45 significantly different expressed genes, respectively. During summer, although none of the site contrasts reached the magnitude of DEGs identified in the Troia versus Ria Formosa Spring comparison, the number of DEGs was more evenly distributed across sites, with all contrasts yielding more than 100 and less than 600 significant genes (Figure 4.4a).

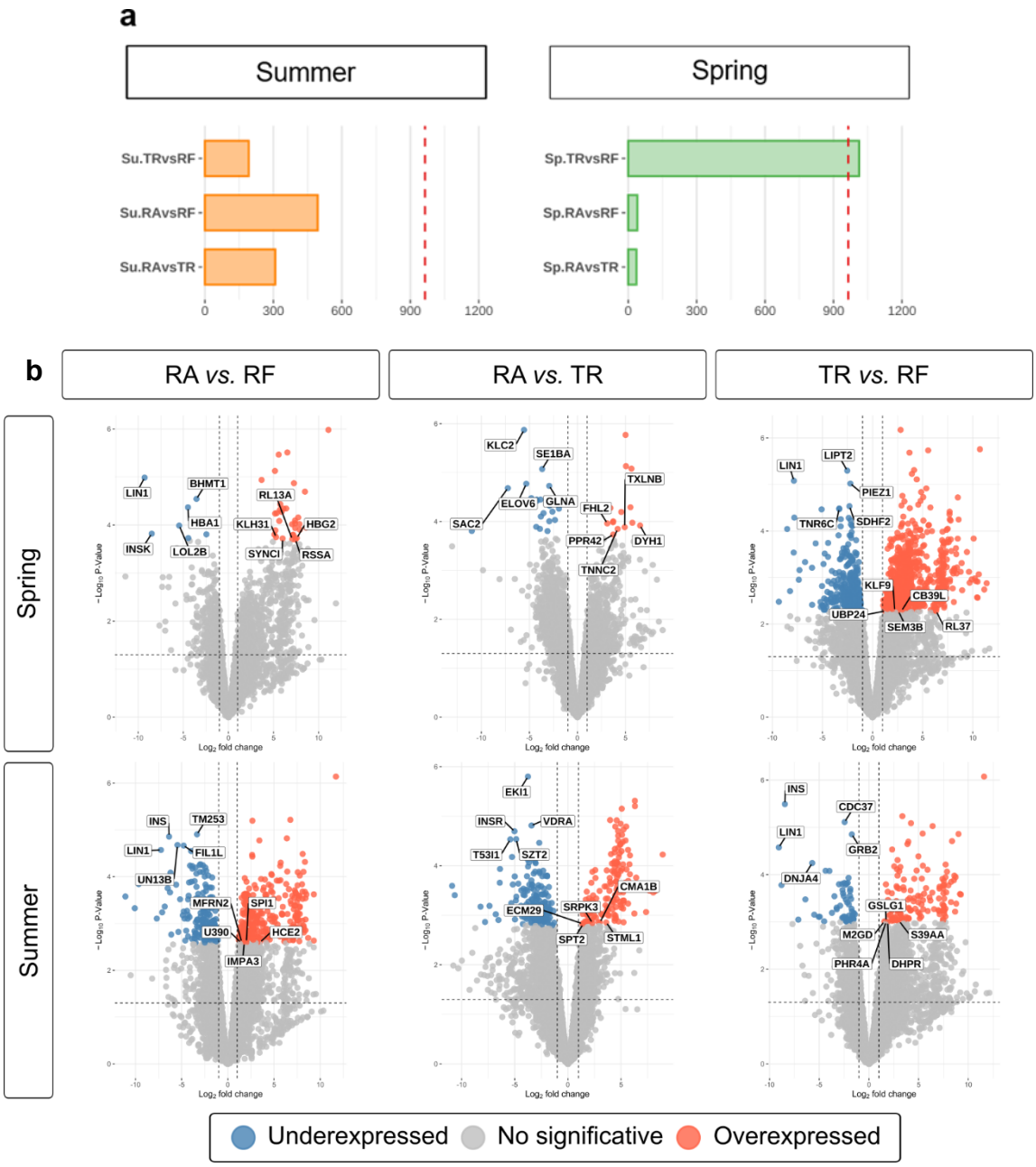


Figure 4.4 - Site-specific differences in differentially expressed genes across seasons. (a) Bar plot showing the number of DEGs detected in each contrast, with the dashed line indicating 10% of the total genes tested (944.5). (b) Volcano plots displaying the distribution of transcripts based on –log10(*p*-value) versus $\log_2$(fold change). Grey points represent non-significant transcripts, while coloured points denote significantly differentially expressed genes (FDR-adjusted $p < 0.05$). Labels highlight the five most significant transcripts (lowest adjusted FDR values) within each contrast. Abbreviations: RA - Ria de Aveiro, TR - Troia, RF - Ria Formosa.

Volcano plots highlighted the distribution of significant genes and top 10 most significant genes comprising both overexpressed and under expressed genes on each contrast (Figure 4.4b).

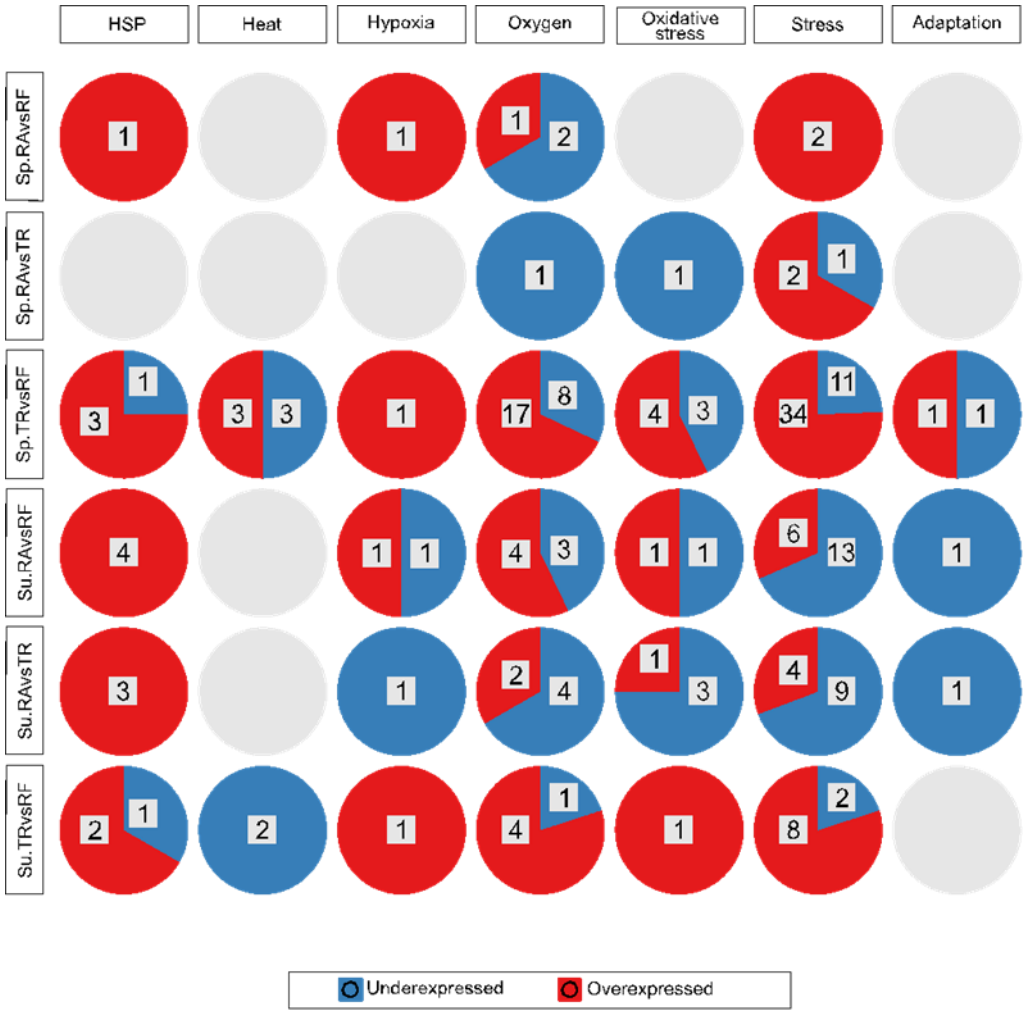


Figure 4.5 - Distribution of differentially expressed genes across curated keywords. Pie charts represent the number of differentially expressed genes associated with each of seven defined biological keywords: HSP, Heat, Hypoxia, Oxygen, Oxidative Stress and Adaptation. Functional annotations for each gene were retrieved using the uniprotR package, and keyword mapping was performed by screening UniProt-derived gene function descriptions. Genes were categorized based on the presence of selected terms within these descriptions. Red and blue slices indicate overexpressed (logFC > 0) and under expressed (logFC < 0) genes, respectively. Abbreviations: RA - Ria de Aveiro, TR - Troia, RF - Ria Formosa, Sp - Spring, Su - Summer.

Following our keyword mapping strategy using relevant biological terms, results show that several highly significant genes were detected across multiple contrasts. Notably, heat shock proteins (e.g., HS90B, HPBP1) and ubiquitin's (e.g., UBQL1) showed consistent patterns of differential expression.

Among all comparisons, the greatest divergence was observed between Troia and Ria Formosa during the spring season. For deeper insights we further examined genes associated with the keyword "stress". This category encompassed the largest number of DEGs (34; See Figure 4.5), possibly reflecting not only the biological relevance of stress responses but also the overall high DEG count in this contrast compared to others.

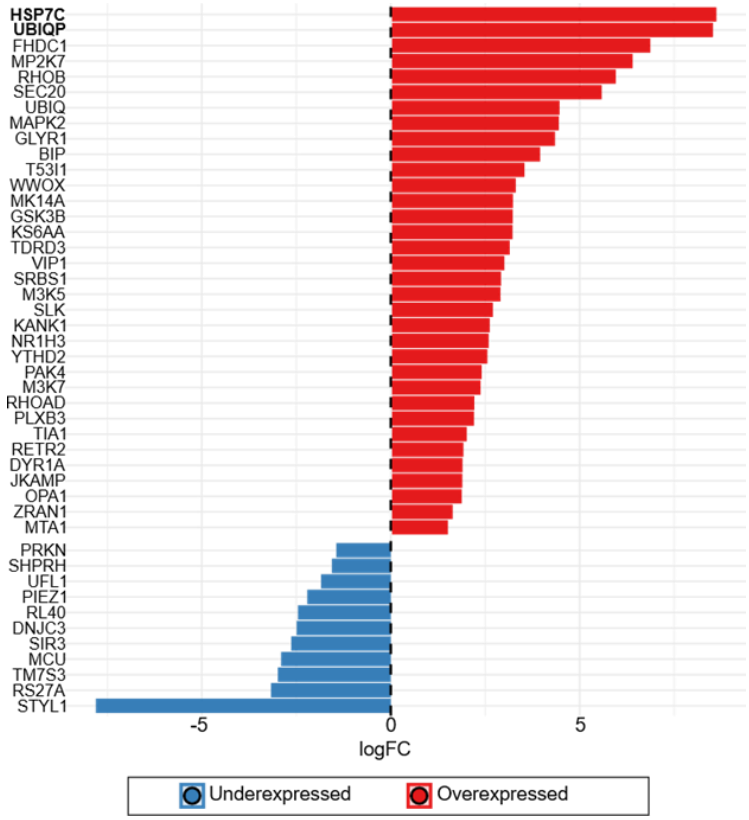


Figure 4.6 - Stress-related differentially expressed genes in Troia vs. Ria Formosa during spring. Bar plot represents $\log_2$ (fold change) values of genes associated with the curated keyword stress. Red and blue bars indicate overexpressed (logFC > 0) and under expressed genes, respectively.

The majority of stress-matched genes were overexpressed in Troia relative to Ria Formosa. Classic stress effectors, such as HSP70 and UBIQP, displayed the highest positive fold changes, whereas STYL1, a gene involved in cytoskeletal organization, emerged as the most strongly under-expressed in Ria Formosa (Figure 4.6).

4.3 Seasonal Gene Expression Dynamics

To complement this analysis and inform the distinction between phenotypic plasticity and local adaptation, seasonal gene expression dynamics were examined within sites. Notably, significant seasonal differential expression was detected exclusively in Ria de Aveiro, where 13 genes showed differential expression between spring and summer. Of these, 11 genes were overexpressed and two were under-expressed in spring relative to summer. No significant seasonal DEGs were detected in Ria Formosa or Troia. These findings correspond with PCA and hierarchical clustering patterns, which demonstrated clear seasonal divergence only for Ria de Aveiro (Figures 4.2 and 4.3).

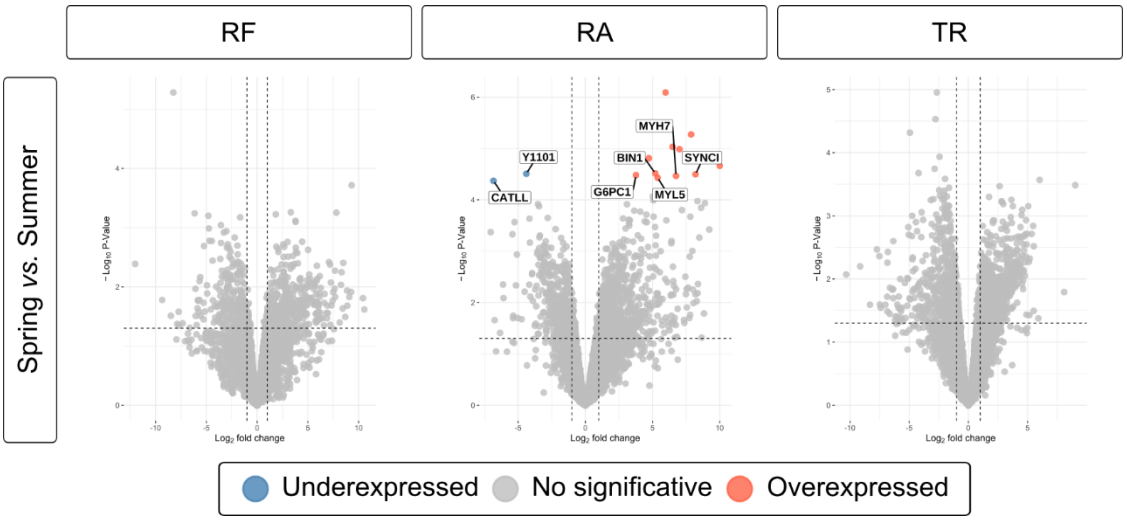


Figure 4.7 - Season differences in differentially expressed genes across sites. Volcano plots displaying the distribution of transcripts based on $-\log_{10}(p\text{-value})$ versus $\log_2(\text{fold change})$. Grey points represent non-significant transcripts, while coloured points denote significantly differentially expressed genes (FDR-adjusted $p < 0.05$). Labels highlight the five most significant transcripts (lowest adjusted FDR values) within each contrast. Abbreviations: RA - Ria de Aveiro, TR - Troia, RF - Ria Formosa.

Volcano plots highlighted the distribution of significant genes and top 10 most significant genes comprising both overexpressed and under-expressed genes on each contrast (Figure 4.7).

Keyword mapping analysis of seasonal DEGs in Ria de Aveiro revealed only two stress-matched genes, CSRP3 and SYNCI, both overexpressed in spring (Figure 4.8).

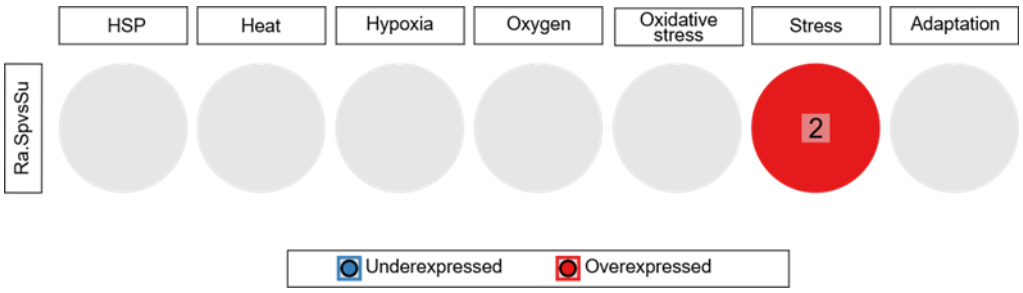


Figure 4.8 - Distribution of differentially expressed genes across curated keywords. Pie charts represent the number of differentially expressed genes associated with each of seven defined biological keywords: HSP, Heat, Hypoxia, Oxygen, Oxidative Stress, and Adaptation. Functional annotations for each gene were retrieved using the uniprotR package, and keyword mapping was performed by screening UniProt-derived gene function descriptions. Genes were categorized based on the presence of selected terms within these descriptions. Red and blue slices indicate overexpressed (logFC > 0) and under expressed (logFC < 0) genes, respectively. Abbreviations: RA - Ria de Aveiro, Sp - Spring, Su - Summer.

4.4 Enrichment and Functional Analysis of Selected Contrasts

Gene set enrichment analysis identified one significantly enriched pathway (FDR < 0.05): lipoic acid metabolism in the comparison between Troia and Ria Formosa in spring (FDR = 0.008). No additional pathways reached the significance threshold. Nevertheless, to explore broader enrichment patterns, the top 10 ranked pathways based on FDR-adjusted *p*-values were examined for two key contrasts: Troia versus Ria Formosa in spring, and Ria de Aveiro spring versus summer (Figure 4.9).

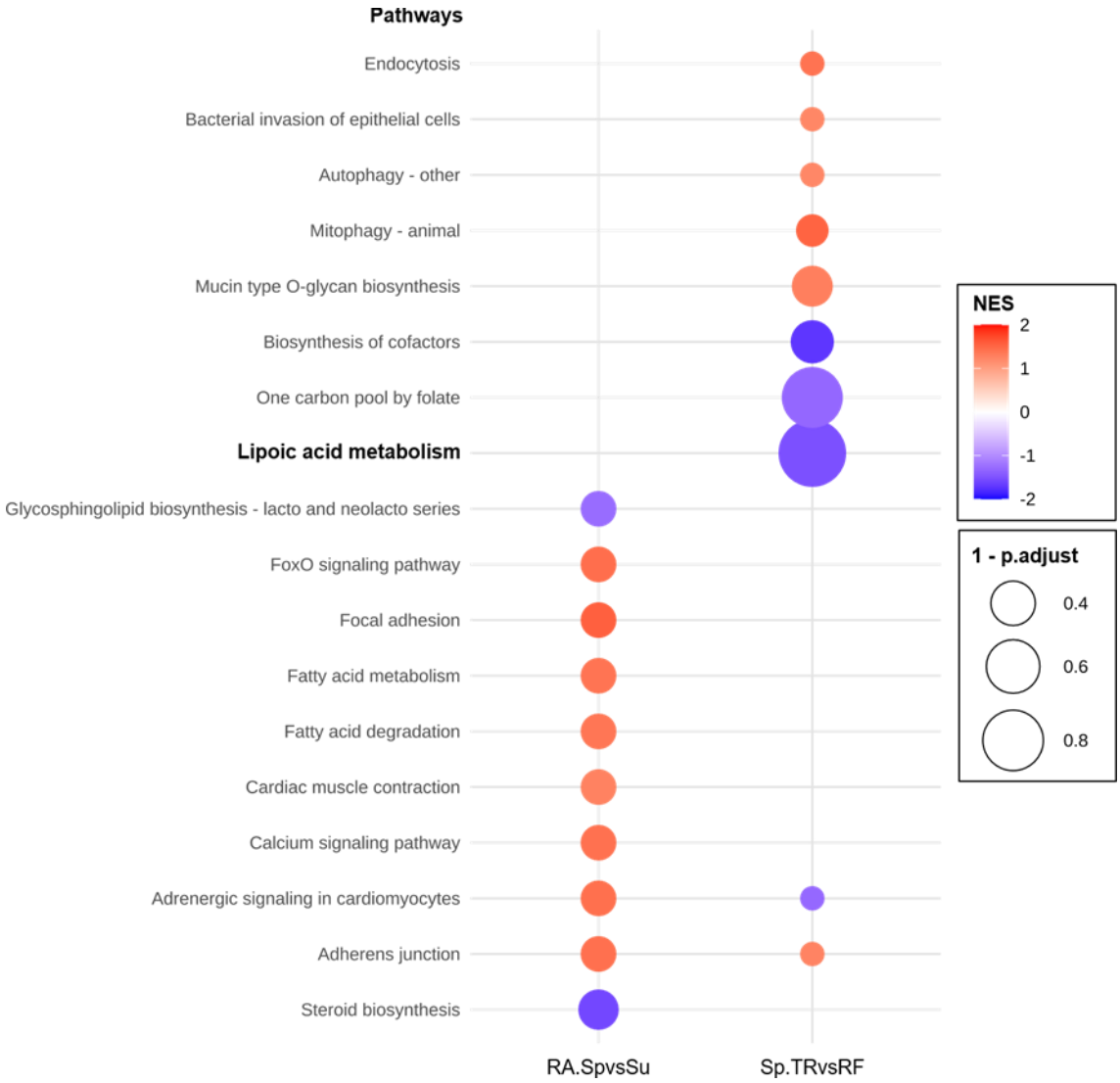


Figure 4.9 - GSEA highlighting strong latitudinal (Spring Troia vs. Ria Formosa) and seasonal (Ria de Aveiro Spring vs. Summer) transcriptional shifts. The y-axis shows the top 10 gene sets ranked by smallest FDR-adjusted *p*-values, and the x-axis represents the two key comparisons: *Spring Troia vs. Ria Formosa* and *Ria de Aveiro Spring vs. Summer*. Bubble colour indicates the normalized enrichment score (NES), with blue representing negative enrichment and red representing positive enrichment. Bubble size reflects statistical significance, with larger bubbles corresponding to lower FDR-adjusted *p*-values. Abbreviations: RA – Ria de Aveiro, TR – Troia, RF – Ria Formosa, Sp – Spring, Su – Summer.

Among the top 10 pathways, adrenergic signalling in cardiomyocytes and adherens junction were shared between the two contrasts, with the former showing opposite enrichment directions and the latter showing the same enrichment direction. In Troia versus Ria Formosa (spring), lipoic acid metabolism, one carbon pool by folate, biosynthesis of cofactors, and adrenergic signalling in cardiomyocytes were negatively enriched, indicating under-representation or down-regulation in Troia. Conversely, in Ria de Aveiro (spring versus summer), negative enrichment was observed for steroid biosynthesis and glycosphingolipid biosynthesis, while the remaining gene sets demonstrated positive enrichment trends.

To further explore potential mechanistic relationships, protein–protein interaction networks were constructed separately for over- and under-expressed gene sets identified in each contrast. After excluding clusters containing fewer than three nodes, a total of five networks were obtained: (i) Ria de Aveiro vs. Ria Formosa in spring (over-expressed genes), (ii) Ria de Aveiro vs. Ria Formosa in summer (over-expressed genes), (iii) Troia vs. Ria Formosa in spring (over-expressed genes), (iv) Troia vs. Ria Formosa in spring (under-expressed genes), and (v) Troia vs. Ria Formosa in summer (over-expressed genes). Several clusters within these networks were associated with more than one functional pathway.

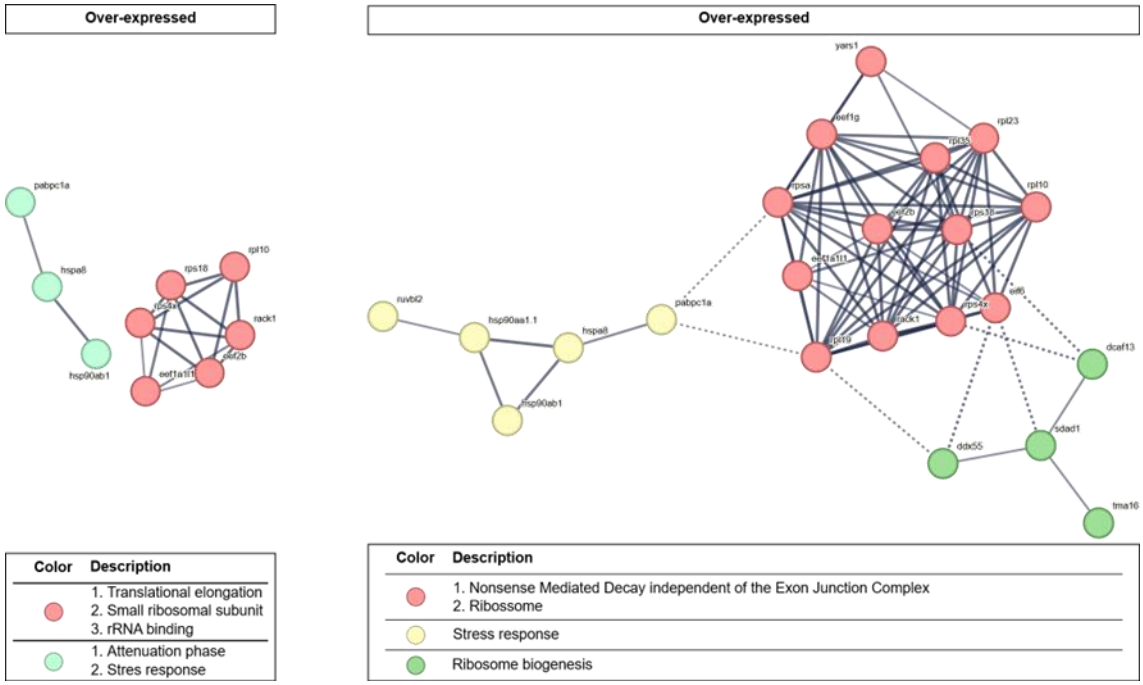


Figure 4.10 - Protein-protein interaction (PPI) network constructed from differentially expressed genes between Ria de Aveiro vs. Ria Formosa. (a) genes overexpressed during spring and (b) genes overexpressed during summer. Nodes represent proteins, and edges indicate interactions supported by multiple sources with a confidence score ≥ 0.700. Edge thickness reflects the strength of evidence

underlying each interaction. Network clustering was performed using the Markov Cluster Algorithm (MCL), with nodes sharing the same colour belonging to the same cluster.

In the Ria de Aveiro versus Ria Formosa comparison during spring, overexpressed genes formed two distinct clusters associated with the attenuation phase and translational elongation (Figure 4.10a). In contrast, overexpressed genes in summer are organized into three functional networks related to stress response, nonsense-mediated decay independent of the exon junction complex (EJC), and ribosome biogenesis (Figure 4.10b).

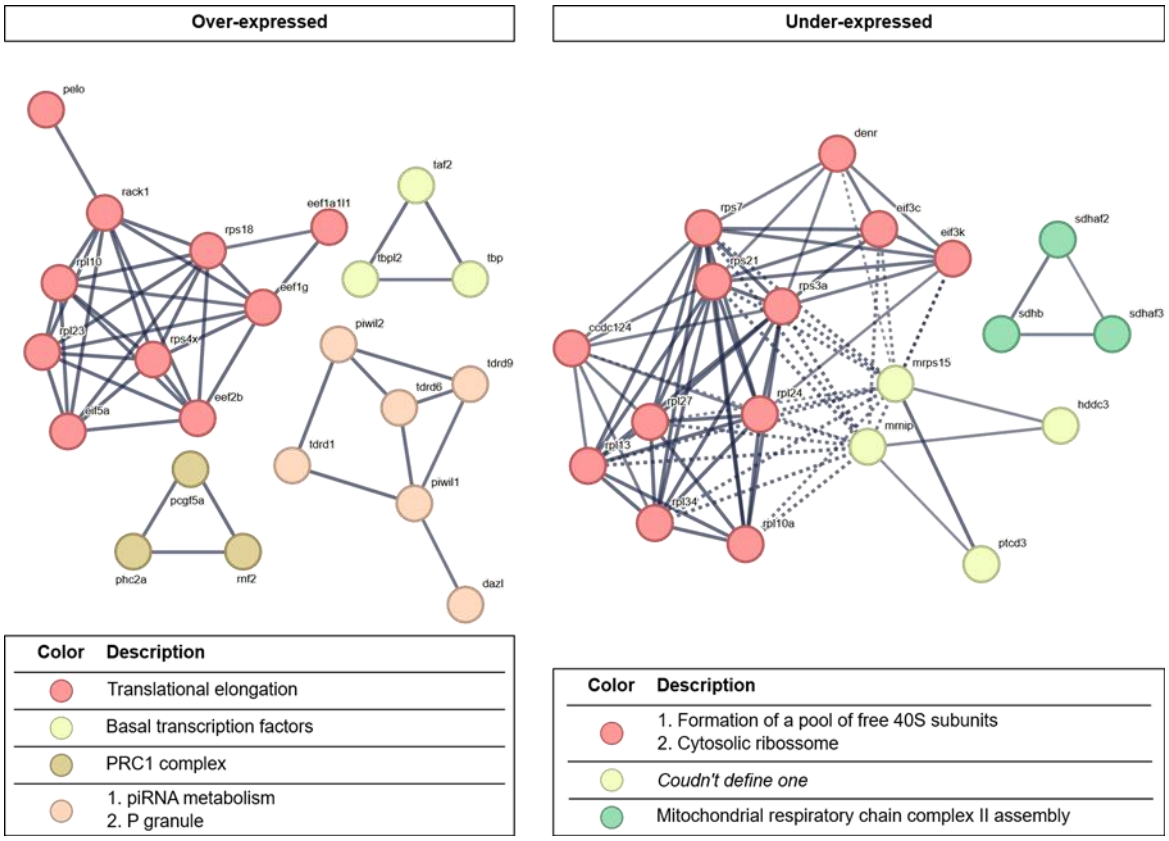


Figure 4.11 - Protein-protein interaction (PPI) network constructed from differentially expressed genes between Troia vs. Ria Formosa during spring: (a) overexpressed and (b) under expressed genes. Nodes represent proteins, and edges indicate interactions supported by multiple sources with a confidence score ≥ 0.700. Edge thickness reflects the strength of evidence underlying each interaction. Network clustering was performed using the Markov Cluster Algorithm (MCL), with nodes sharing the same colour belonging to the same cluster.

In the Troia versus Ria Formosa comparison during spring, overexpressed genes grouped into four clusters associated with translational elongation, basal transcription factors, the PRC1 complex, and piRNA metabolism (Figure 4.11a). By contrast, under-expressed genes formed three clusters linked to the formation of the free 40S ribosomal subunit pool and the assembly of mitochondrial respiratory chain complex II (Figure 4.11b). One additional cluster (the yellow cluster) could not be confidently assigned to a specific pathway; however, it

included the proteins *mrps15* and *mrnip*, which displayed extensive interactions with proteins in another defined cluster, suggesting a potential functional relationship despite the absence of a direct pathway annotation. Functionally, *mrps15* is a mitochondrial ribosomal protein involved in protein synthesis within the mitochondria, while *mrnip* is a regulator of DNA repair and cell cycle checkpoints.

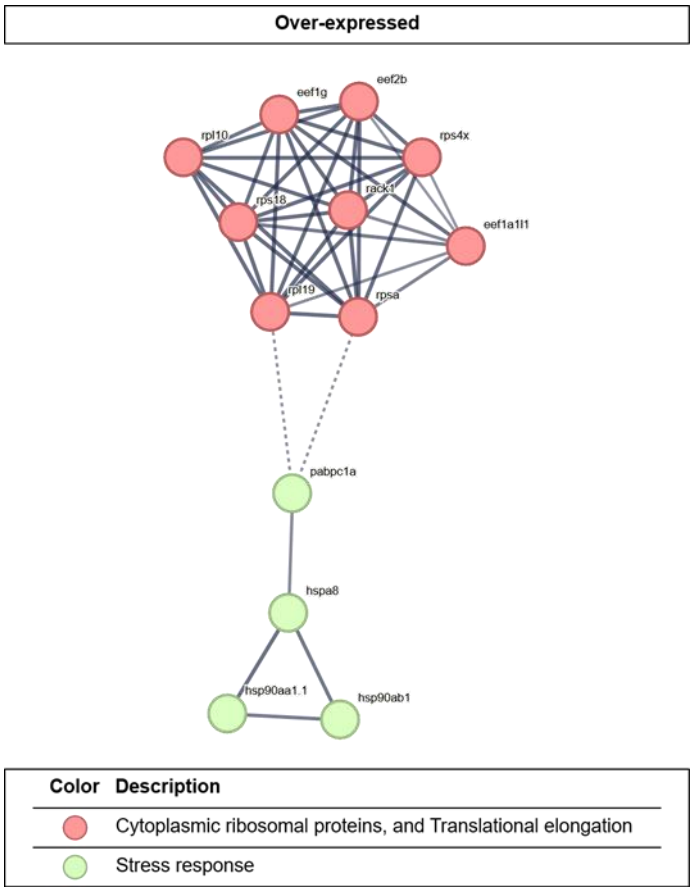


Figure 4.12 - Protein-protein interaction (PPI) network constructed from differentially expressed genes overexpressed in Troia vs. Ria Formosa during summer. Nodes represent proteins, and edges indicate interactions supported by multiple sources with a confidence score ≥ 0.700. Edge thickness reflects the strength of evidence underlying each interaction. Network clustering was performed using the Markov Cluster Algorithm (MCL), with nodes sharing the same colour belonging to the same cluster.

Finally, the Troia versus Ria Formosa comparison during summer (Figure 4.12) revealed a network of over-expressed genes that formed two main clusters, primarily associated with cytoplasmic ribosomal proteins, translational elongation, and stress response.

Notably, across contrasts, the STRING analysis consistently highlighted the recurrent involvement of pathways related to stress response and translational elongation.

5

DISCUSSION

We investigated molecular responses in populations distributed along a latitudinal thermal gradient on the Portuguese coast (northern vs. central regions vs. southern) and across seasonal thermal regimes (spring vs. summer). Our results demonstrate some degree of population-level divergence in gene expression, in response to the specific local environmental regimes, whereas only one population (Ria de Aveiro) showed differences between seasons. This highlights the complexity of molecular responses to external drivers. Understanding such mechanisms is critical for predicting how species persist under variable and changing conditions.

Overall, our principal component analysis revealed two main axes of variation that appear to capture distinct aspects of population variance. The first principal component likely reflects underlying genetic differentiation. In contrast, the second principal component is most plausibly associated with environmental differences, particularly those linked to local thermal regimes.

Among the sites examined, Ria de Aveiro emerged as the most distinct along PC2, particularly in summer. While shallow water temperatures in the southern sites - Troia Peninsula and Ria Formosa - were consistently warm (25.9 ± 3.5 °C and 25.9 ± 0.5 °C, respectively), summer seawater in Ria de Aveiro remained comparatively cool (18.6 ± 1.6 °C). This is below the reported optimal temperature for *P. microps*, which is around 20 °C (Cereja et al., 2018). This nearly 7 °C difference highlights the influence of the Iberian coastal upwelling system, driven by persistent northerly winds that bring cold subsurface waters to the western coast (Alvarez et al., 2013). Other physicochemical variables varied less across sites. In our measurements, salinity values were generally within the typical oceanic range (34–38 PSU; (Rigal et al., 2008). Except in Ria de Aveiro during spring, where salinity dropped to ~28 PSU, which can be explained by the freshwater inflow reported for this system (Lopes et al., 2019; Vargas et al.,

2017). However, this fluctuation likely exerts reduced physiological stress on *P. microps*, since it is a species known for having a wide tolerance for salinity variation, withstanding salinities ranging from 0 to 51 PSU (Souza et al., 2018; Tougard et al., 2014), and maintaining stable blood osmolality across 10–40 PSU (Rigal et al., 2008). Similarly, in our measurements, pH and dissolved oxygen remained within ranges typical of well-oxygenated coastal systems, consistent with values previously reported for such environments (George et al., 2024; Gobler & Baumann, 2016), suggesting that hypoxia or acidification were not key structuring drivers during our study. Taken together, temperature emerges as the dominant environmental driver underlying the divergence captured by PC2, separating cooler sites (Ria de Aveiro) from warmer sites (Troia and Ria Formosa).

In contrast, PC1 differentiated Ria Formosa from the other sites. This separation may indicate local adaptation, possibly associated with the region's persistently high temperatures. Studies in related taxa have shown that populations inhabiting chronically warm environments can evolve enhanced thermal plasticity, which in turn facilitates adaptive evolution (Chen et al., 2022; Dilworth et al., 2024). For example, *Daphnia magna* populations from warmer climates exhibit stronger upregulation of hemoglobin in response to acclimation at elevated temperatures, indicating local adaptation for increased plasticity in hemoglobin expression (Yampolsky et al., 2014). Interestingly, the largest site-specific divergence in our dataset occurred in spring between Troia and Ria Formosa, where transcriptional profiles diverged more strongly than between any other population pair. This separation was consistent across both PCA clustering and differential gene expression analyses, with more than 1,000 DEGs (~10.5% of the tested genes) distinguishing the two populations. Although their average shallow seawater temperature summer conditions appear broadly similar, with overlapping mean temperatures (~25.9°C), the seasonal dynamics of shallow seawater warming diverge sharply between these two sites. Specifically, the magnitude of seasonal increase in water temperature is more pronounced in Troia (ΔT spring–summer ≈ 3.9 °C) than in Ria Formosa or Ria de Aveiro ($\Delta T \approx 2.5$ °C in both). This suggests that divergence may be driven not solely by mean summer temperature conditions, but by the trajectory of seasonal thermal regimes, as supported by Logan and Cox (2020) and Schmalensee et al. (2022).

One plausible explanation for the higher number of differentially expressed genes observed in spring is anticipatory plasticity. Anticipatory plasticity is a well-documented strategy in ectotherms, enabling adjustment of cellular processes in advance of more extreme conditions (Stollewerk et al., 2025). For instance, mosquito populations exhibit greater geographic variation in thermal reaction norms under warm versus cool seasons, with timing of diapause

strongly modulating these differences (Ragland & Kingsolver, 2007). In live-bearing lizards, climatic effects on life history and between-year fluctuations in temperature have driven divergence in sex-determining mechanisms, with lowland populations relying on temperature-dependent systems while highland populations maintain genetic sex determination (Pen et al., 2010).

Troia has also experienced a disproportionately high incidence of marine heatwaves in recent years, with six events documented between 2010–2022 (each spanning 6 to 16 days) (see *IPMA*, 2025). Marine heatwaves are increasingly recognized as acute selective pressures shaping molecular and phenotypic resilience in coastal organisms (Oliver et al., 2018; Smale et al., 2019). It is plausible that the transcriptional distinctiveness observed in Troia compared to the other two populations represents not only anticipatory plasticity but is also the result of selection due to cumulative exposure to recurrent marine heatwaves. However, this links to an important evolutionary consideration: phenotypic plasticity is most advantageous in predictable (even if fluctuating) environments (Ghalambor et al., 2007; Stollewerk et al., 2025). Increasingly stochastic extremes, such as irregular heatwaves, may erode the reliability of certain plastic responses in guaranteeing survival if their adaptive value is lost or they become maladaptive (Stollewerk et al., 2025).

At the molecular level, our analyses consistently revealed the presence of differential gene expression of highly conserved stress-response genes between populations, most notably heat shock proteins and ubiquitin-related genes. Heat shock proteins represent a nearly universal response to thermal stress across taxa (Collins, Peck, et al., 2021). Acting as molecular chaperones, they stabilize unfolded proteins and promote correct refolding under conditions that threaten protein integrity, such as acute heat exposure (Payton et al., 2016). The ubiquitin–proteasome system provides a complementary layer of protection, ensuring the removal of irreparably damaged proteins that escape refolding. In this process, ubiquitin ligases recognize misfolded proteins, label them with ubiquitin chains, and target them for degradation by the 26S proteasome, recycling their amino acids for new protein synthesis (Fan & Jespersen, 2025).

Within our dataset, distinct population patterns provide further insight into how *Pomatoschistus microps* mobilizes its molecular defence mechanisms. In spring, individuals from Troia displayed elevated expression of HSP7C (i.e., the constitutively expressed form of HSP70, synonym to heat shock cognate 70) and ubiquitin-related transcripts, consistent with a strong reliance on ubiquitous mechanisms. This is complemented by suppression of lipoic acid metabolism, a pathway that albeit providing antioxidant protection also carries high

metabolic and energetic costs (Kütter et al., 2014). Its downregulation therefore likely reflects a strategic energy distribution in which resources are conserved and redirected toward more immediate needs, such as stress defence, protein transport or cellular repair (Fan & Jespersen, 2025; Ros et al., 2021). This pattern is consistent with a constitutive frontloading strategy, in which organisms maintain higher baseline levels of protective proteins as a preparatory defence against sudden stress (Collins et al., 2023; Dilworth et al., 2024; Payton et al., 2016). Comparable mechanisms have been observed in corals, such as *Acropora* spp., which exhibit frontloading of thermal stress response genes and, adaptively, suppress expression of metabolic and ribosomal processing genes in populations exposed to more thermally stressful environments (Dilworth et al., 2024). Additional pathways overexpressed in Troia compared to Ria Formosa, including translational elongation, basal transcription factors, the PRC1 complex, and piRNA metabolism, point to a cellular program adapted toward rapid stress responsiveness. Upregulation of translational elongation and basal transcription factors suggest that cells are maintaining a high capacity for rapid protein synthesis, a feature consistent with organisms that must repeatedly synthesize stress-response proteins during episodes of acute environmental change (Alagar Boopathy et al., 2022). The enrichment of PRC1 complex components indicates modulation of chromatin architecture and transcriptional plasticity, enabling the fine-tuned regulation of stress-inducible genes (Bordet et al., 2022). In parallel, increased activity of piRNA metabolism pathways highlights investment in genome integrity maintenance, as the piRNA pathway suppresses transposable element activation under stress and contributes to germline stability, that is the ability of genetic or epigenetic changes to be preserved and passed on from parents to their offspring via the egg and sperm cells. (Czech et al., 2018; Kelly, 2014; Weick & Miska, 2014). Taken together, these processes indicate that Troia populations adopt a highly dynamic strategy. Importantly, this is not a static adaptation but a flexible, stress-preparatory regulation (Collins, Clark, et al., 2021). Indeed, in subsequent seasons we observed continued upregulation of cytoplasmic ribosomal proteins and translational elongation factors, reinforcing the notion that Troia fish sustain an enhanced capacity for protein turnover and rapid stress-response deployment throughout the year or at least during the warmer seasons (spring and summer) (Seluzicki et al., 2025). Such patterns align with evidence from other ectotherms in variable habitats, where ribosomal and translational plasticity have been linked to resilience under thermal fluctuation (Chen et al., 2023; Collins, Clark, et al., 2021).

Conversely, Ria Formosa exhibited higher expression of STYL1, a gene implicated in cytoskeletal organization, which may contribute to cellular structural stability under

environmental stress. Actually, a study on an intertidal amphipod found that frontloaded genes also included those related to immune and cytoskeletal functions (Collins et al., 2023; Somero, 2020). Additionally, a notable gene expressed in Ria Formosa fish provides valuable insights into possible biological mechanisms, DNJC3 a co-chaperone of HSPA87/HSC70, is a heat shock cognate that enhances protein-folding capacity (Stricher et al., 2013) and could also represent a form of frontloading by maintaining elevated levels of molecular chaperones. Supporting this interpretation, in Ria Formosa we observed overexpression of pathways related to the formation of the free 40S ribosomal subunit pool and the assembly of mitochondrial respiratory chain complex II. These enrichments suggest a fundamentally different adaptive strategy, characterized by stable investment in translational efficiency through increased ribosomal capacity, and energy production through enhanced mitochondrial capacity (Gerber et al., 2020; Muir et al., 2022). The 40S ribosomal subunit pool provides the foundation for efficient initiation of protein synthesis, ensuring consistent capacity to sustain cellular growth and maintenance under warm but predictable conditions (Liu & Qian, 2014). A parallel example is found in the Atlantic salmon (*Salmo salar*), where juveniles exposed to elevated temperatures (+4°C) during development display increased abundance of ribosomal proteins in cardiac tissue. This enhancement of protein synthesis capacity supports a hypertrophic growth model of the compact myocardium, representing a compensatory mechanism to maintain cardiac function under thermal stress. Moreover, these salmon also show increased mitochondrial respiratory chain proteins and higher mitochondrial density, leading to improved oxidative phosphorylation and ATP production efficiency (Muir et al., 2022). Meanwhile, the upregulation of mitochondrial respiratory chain complex II assembly, also known as succinate dehydrogenase (SDH), signifies a more efficient oxidative phosphorylation system capable of sustaining ATP production even under the stress of chronically elevated temperatures. When temperatures remain high over prolonged periods, cells adapt by enhancing the machinery responsible for oxidative phosphorylation, including Complex II (Bezawork-Geleta et al., 2017; Goetzman et al., 2023). Together, these pathways indicate that Ria Formosa populations adopt a strategy of basal stabilization, where cellular homeostasis is maintained through robust 40S ribosomal subunit and efficient mitochondrial function rather than acute stress preparedness.

Latitudinal variation thus emerges as a factor driving molecular divergence among *P. microps* (Troia and Ria Formosa) populations. In contrast, seasonal differences were only detected in Ria de Aveiro, where individuals exhibited transcriptional shifts between spring and summer. Although the number of differentially expressed genes was small (13 DEGs, of which

11 were overexpressed) and no pathway reached statistical significance, these results are not devoid of biological meaning, especially when we interpret alongside literature.

Ria de Aveiro experiences a relatively low seasonal thermal amplitude. In Ria de Aveiro, the measured abiotic parameters throughout seasons remained within ranges considered optimal for *P. microps* performance, as mentioned before. The key environmental distinction between seasonal differences across populations relies in the fact that Ria de Aveiro maintains relatively more stable conditions that are nonetheless cooler than those recorded in Troia and Ria Formosa – Troia with a strong seasonal thermal amplitude and records of marine heat waves and Ria Formosa which maintains relatively stable, but overall, much warmer conditions than Ria de Aveiro.

The complete absence of seasonal DEGs in Troia and Ria Formosa could reflect the influence of environmental (un)predictability and overall warmer thermal regimes in shaping molecular responses. Previous field studies on fish, such as Feidantsis et al., (2013) reveal that in the liver of *Sparus aurata* inhabiting shallow waters, the most pronounced changes in the concentration of proteins associated with cellular stress responses, specifically nitrogen-activated protein kinases (MAPKs) and extracellular signal regulated kinases (ERKs), occurred in April and May. Following these peaks, MAPK levels remained stable throughout the summer, while ERK levels declined. This may help explain the absence of transcriptional shifts in Troia between spring and summer, suggesting that they may actually occur earlier in the year (as the season progresses from winter to spring) as a 'preparative defence strategy' for the warmer seasons, which include spring and summer (see Dong et al., 2008).

Experimental evidence further substantiates the limited seasonal transcriptional variation observed between spring and summer. Laboratory simulations mimicking both current and projected summer conditions revealed only minor shifts in critical thermal maximum (CTmax), with warm-acclimatized individuals exhibiting slightly higher tolerance than cold-acclimatized ones (Missionário et al., 2024). These findings could indicate that *P. microps* is well adapted to the substantial temperature variability typical of shallow coastal habitats, exhibiting relative insensitivity to temperature fluctuations within the species' operating thermal niche, with transcriptional and physiological adjustments likely triggered only once critical thresholds are exceeded (Missionário et al., 2024). This adaptive strategy allows these organisms to be readily prepared to respond to frequent environmental fluctuations, reducing the lag time between the detection of external cues and the induction of an appropriate response, allowing for their physiology to quickly match their environment.

At the molecular level, seasonal transcriptomic differences in the Ria de Aveiro population reveal distinct physiological priorities between summer and spring. During summer, steroid biosynthesis emerged as a central pathway. In teleost's, steroidogenesis is a highly conserved process that regulates reproduction, stress adaptation, osmoregulation, and immune function (Kocerha & Denslow, 2025). This enrichment is consistent with the reproductive phenology of the local population, as spawning activity in Ria de Aveiro typically peaks in summer (Arruda et al., 1993). The upregulation of steroidogenic pathways therefore likely reflects the endocrine activation required to support gonadal maturation and spawning. In parallel, the upregulation of cathepsin L-like (CATLL) indicates intensified proteolytic activity. Cathepsin L has been widely implicated in vitellogenin processing and yolk protein mobilization in maturing females, as well as in tissue remodelling during reproductive cycles (Jiang et al., 2022). Its elevated expression in summer therefore provides molecular evidence of active vitellogenesis and reproductive readiness within the population. In contrast, spring samples showed enrichment of fatty acid metabolism and degradation. Lipids, a major energy reserve in fish, support growth, locomotion, reproduction, and migration (Olivares-Rubio & Vega-López, 2016). Photoperiod has been proposed as a cue modulating lipid metabolism (Wang et al., 2023) partly through effects on liver oxidative stress. Aquatic organisms rely on circadian clocks to synchronize physiology with rhythmic light cycles. Light modulates feeding, energy use, and growth; for instance, long-day photoperiods stimulate feeding, enhancing hepatic lipid metabolism and growth in juvenile gibel carp. Seasonal variation in lipid metabolism enzymes has also been documented in yellowtail (*Seriola quinqueradiata*) (Wang et al., 2023). Spring upregulation of FoxO signalling further suggests responses to environmental stress. FoxO transcription factors, regulated by insulin and growth factors, control cell proliferation, apoptosis, metabolism, and stress resistance. In aquaculture, there is a growing body of research on the response of FoxO to environmental factor stresses. For example, FoxO activity changes under hypoxia, heat stress (e.g., *Brachymystax lenok tsinlingensis*), salinity (e.g., *Luciobarbus capito*), some studies have also demonstrated that it can upregulate the expression of autophagy-related genes (Wang et al., 2023). Seasonal variation thus indicates energy accumulation and stress defence, while summer profiles reflect endocrine and hepatic adjustments supporting reproduction. This seasonal alteration between metabolic preparation and reproductive investment seems to align with the species' phenology. However, as no pathways reached statistical significance, these patterns should be interpreted cautiously and cannot be taken as definitive evidence of such processes.

The current study is, nonetheless, baffled by challenges and pitfalls. Firstly, *P. microps* is a non-model organism, which presents inherent challenges for transcriptome assembly and annotation (Jackson et al., 2024). Despite leveraging the Swiss-Prot database and a zebrafish-specific database to maximize annotation accuracy, the absence of species-specific genomic resources may contribute to potential annotation errors (Jackson et al., 2024; Shomron, 2021). Nevertheless, we focused on the liver due to its central role in homeostasis and metabolic regulation of whole-body physiology, such as covering bile production, amino acid regulation, glucose metabolism, vitamin storage, detoxification, and immune responses (Daniel Schlenk & William H. Benson, 2017; Heymann & Tacke, 2016). Liver gene expression changes thus reflect both the impacts of environmental stress and the organism's adaptive responses (Chien & Hwang, 2001; Tao et al., 2018). Additionally, the liver often shows stronger gene expression responses to temperature extremes than other tissues, making it suitable for studying thermal stress in challenging habitats like intertidal zones (Das et al., 2005; Song et al., 2015). Secondly, our reliance on wild-caught individuals introduces inherent environmental confounding factors that can challenge the interpretation of transcriptomic data (Keagy et al., 2023). Despite efforts to characterize physicochemical conditions, confounding variables such as capture stress or prior stress history, can significantly influence gene expression profiles (Keagy et al., 2023). Nevertheless, conducting transcriptomic analyses on wild specimens holds the advantage of capturing a realistic snapshot of molecular networks as well as natural inter-individual variability of biologically and ecologically relevant responses to complex, natural environments that laboratories cannot fully replicate. Specifically, this approach allows detection of genes expressed exclusively under natural conditions and links molecular responses across cellular, population, community, and ecosystem levels (Keagy et al., 2023). Moreover, landscape transcriptomics serves as a powerful bioindicator, revealing acute, chronic, and historical exposure to multiple and interacting stressors simultaneously, thus elucidating the specific landscape factors shaping organismal responses (Keagy et al., 2023). Another important consideration is our use of principal component analysis and complementary visualization methods, such as heatmaps and hierarchical clustering, for exploratory data analysis. PCA is a standard initial step for high-dimensional RNA-seq data, summarizing major patterns of variation (Koch et al., 2018). However, it has well-documented limitations: it can be distorted by outliers, batch effects, and technical variation unrelated to biology (Chen et al., 2020; Yu et al., 2024). More importantly, there is a risk of apophenia (i.e., the tendency to perceive meaningful patterns in noisy data) leading to overinterpretation of apparent clusters or trends without robust supporting evidence (Hanson et al., 2021). Therefore, we treat PCA results as

exploratory and interpret them cautiously, corroborating patterns with additional analyses before drawing conclusions.

Notwithstanding the above mentioned limitations, our study provides valuable insights. From a climate change perspective, these divergent strategies have contrasting implications. As marine heatwaves become more frequent and unpredictable, populations such as Troia, which rely on anticipatory plasticity, may be at risk if environmental cues no longer reliably predict future conditions. Conversely, populations adapted to stable environments like Ria Formosa may be highly vulnerable to sudden thermal extremes. In Ria de Aveiro, no clear stress-adaptive strategy emerged, but its northern location and the mitigating influence of upwelling may delay the onset of climate-driven impacts relative to southern populations. By identifying distinct molecular strategies, our findings provide a basis for predicting population resilience under future scenarios.

6

CONCLUSION

This thesis investigated the molecular mechanisms that shape how *Pomatoschistus microps* populations respond to local environmental regimes and seasonal thermal variation, using a functional genomics approach. By comparing transcriptomic profiles across three populations along a latitudinal gradient and across two seasonal regimes three key conclusions emerge:

1. **Divergent responses are shaped by thermal regimes.** We demonstrated that gene expression divergence among populations is primarily associated with local temperature regimes rather than salinity, oxygen, or pH, which remained within tolerable ranges. Ria de Aveiro emerged as distinct due to the cooling effect of coastal upwelling, while Ria Formosa populations exhibited signatures of persistent local adaptation to chronically warm environments.
2. **Alternative physiological strategies.** Our transcriptomic analyses revealed that Troia fish adopt a highly dynamic stress-preparatory strategy, characterized by frontloading of stress-related genes, consistent with recurrent exposure to marine heatwaves. In contrast, Ria Formosa populations displayed a basal stabilization strategy, with sustained investment in ribosomal and mitochondrial efficiency, suggesting adaptation to persistently warm but predictable thermal regimes. This contrast illustrates how populations facing different thermal environments may rely on alternative molecular responses to achieve resilience.
3. **Seasonal plasticity is restricted to Ria de Aveiro.** Unlike southern populations, only Ria de Aveiro fish displayed seasonal transcriptional adjustments, with spring characterized by metabolic priming (FoxO signalling, lipid catabolism) and summer dominated by reproductive activation (steroidogenesis, vitellogenesis). This seasonal alternation between metabolic preparation and reproductive investment seems to align with the species phenology.

Future directions should integrate long-term monitoring with controlled experiments to disentangle plasticity from genetic adaptation, incorporate other tissues beyond the liver, and assess carry-over effects between life stages. Such efforts are essential to predict the resilience of coastal fish under intensifying climate change and increasingly stochastic marine heat-waves.

REFERENCES

Adrian-Kalchhauser, I., Blomberg, A., Larsson, T., Musilova, Z., Peart, C. R., Pippel, M., Solbakken, M. H., Suurväli, J., Walser, J.-C., Wilson, J. Y., Alm Rosenblad, M., Burguera, D., Gutnik, S., Michiels, N., Töpel, M., Pankov, K., Schloissnig, S., & Winkler, S. (2020). The round goby genome provides insights into mechanisms that may facilitate biological invasions. *BMC Biology*, *18*(1), 11. https://doi.org/10.1186/s12915-019-0731-8

Alagar Boopathy, L. R., Jacob-Tomas, S., Alecki, C., & Vera, M. (2022). Mechanisms tailoring the expression of heat shock proteins to proteostasis challenges. *Journal of Biological Chemistry*, *298*(5), 101796. https://doi.org/10.1016/j.jbc.2022.101796

Alvarez, I., Dias, J. M., deCastro, M., Vaz, N., Sousa, M. C., & Gómez-Gesteira, M. (2013). Influence of upwelling events on the estuaries of the north-western coast of the Iberian Peninsula. *Marine and Freshwater Research*, *64*(12), 1123–1134. https://doi.org/10.1071/MF12298

Angiulli, E., Pagliara, V., Cioni, C., Frabetti, F., Pizzetti, F., Alleva, E., & Toni, M. (2020). Increase in environmental temperature affects exploratory behaviour, anxiety and social preference in Danio rerio. *Scientific Reports*, *10*(1), 5385. https://doi.org/10.1038/s41598-020-62331-1

Arruda, L. M., Azevedo, J. N., & Neto, A. I. (1993). Abundance, Age-structure and Growth, and Reproduction of Gobies (Pisces; Gobiidae) in the Ria de Aveiro Lagoon (Portugal). *Estuarine, Coastal and Shelf Science*, *37*, 509–523.

Atlas Climático Ibérico. (sem data). Agencia Estatal de Meteorología Ministerio de Medio Ambiente y Medio Rural y Marino.

Baptista, V., Silva, P. L., Relvas, P., Teodósio, M. A., & Leitão, F. (2018). Sea surface temperature variability along the Portuguese coast since 1950. *International Journal of Climatology*, *38*(3), 1145–1160. https://doi.org/10.1002/joc.5231

Bennett, S., Wernberg, T., Arackal Joy, B., De Bettignies, T., & Campbell, A. H. (2015). Central and rear-edge populations can be equally vulnerable to warming. *Nature Communications*, *6*(1), 10280. https://doi.org/10.1038/ncomms10280

Bertolo, A., Pépino, M., Adams, J., & Magnan, P. (2011). Behavioural Thermoregulatory Tactics in Lacustrine Brook Charr, Salvelinus fontinalis. *PLOS ONE*, *6*(4), e18603. https://doi.org/10.1371/journal.pone.0018603

Bezawork-Geleta, A., Rohlena, J., Dong, L., Pacak, K., & Neuzil, J. (2017). Mitochondrial Complex II: At the Crossroads. *Trends in Biochemical Sciences*, *42*(4), 312–325. https://doi.org/10.1016/j.tibs.2017.01.003

Böhnisch, A., Mittermeier, M., Leduc, M., & Ludwig, R. (2021). Hot Spots and Climate Trends of Meteorological Droughts in Europe–Assessing the Percent of Normal Index in a Single-Model Initial-Condition Large Ensemble. *Frontiers in Water*, *3*. https://doi.org/10.3389/frwa.2021.716621

Bordet, G., Couillault, C., Soulavie, F., Filippopoulou, K., & Bertrand, V. (2022). PRC1 chromatin factors strengthen the consistency of neuronal cell fate specification and maintenance in C. elegans. *PLOS Genetics*, *18*(5), e1010209. https://doi.org/10.1371/journal.pgen.1010209

Borner, J., & Burmester, T. (2017). Parasite infection of public databases: A data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. *BMC Genomics*, *18*(1), 100. https://doi.org/10.1186/s12864-017-3504-1

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., Cheung, W. W. L., Connors, S., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., … Péan, C. (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (First). Intergovernmental Panel on Climate Change (IPCC). https://doi.org/10.59327/IPCC/AR6-9789291691647

Canzler, S., & Hackermüller, J. (2020). multiGSEA: A GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics*, *21*(1), 561. https://doi.org/10.1186/s12859-020-03910-x

Cardoso, R. M., Soares, P. M. M., Lima, D. C. A., & Miranda, P. M. A. (2019). Mean and extreme temperatures in a warming climate: EURO CORDEX and WRF regional climate high-resolution projections for Portugal. *Climate Dynamics*, *52*(1), 129–157. https://doi.org/10.1007/s00382-018-4124-4

Cereja, R., Mendonça, V., Dias, M., Vinagre, C., Gil, F., & Diniz, M. (2018). Physiological effects of cymothoid parasitization in the fish host Pomatoschistus microps (Krøyer, 1838) under increasing ocean temperatures. *Ecological Indicators*, *95*, 176–182. https://doi.org/10.1016/j.ecolind.2018.07.045

Chen, X., Ji, Y., Cheng, Y., Hao, Y., Lei, X., Song, G., Qu, Y., & Lei, F. (2022). Comparison between short-term stress and long-term adaptive responses reveal common paths to molecular adaptation. *iScience*, *25*(3), 103899. https://doi.org/10.1016/j.isci.2022.103899

Chen, X., Zhang, B., Wang, T., Bonni, A., & Zhao, G. (2020). Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinformatics*, *21*(1), 269. https://doi.org/10.1186/s12859-020-03608-0

Chen, Y., McCarthy, D., Baldoni, P., Robinson, M., & Smyth, G. (2024). *edgeR: differential analysis of sequence read count data User's Guide*.

Chen, Y., Wu, X., Lai, J., Liu, Y., Song, M., Li, F., & Gong, Q. (2023). Integrated biochemical, transcriptomic and metabolomic analyses provide insight into heat stress response in Yangtze sturgeon (Acipenser dabryanus). *Ecotoxicology and Environmental Safety*, *249*, 114366. https://doi.org/10.1016/j.ecoenv.2022.114366

Chevin, L.-M., Collins, S., & Lefèvre, F. (2013). Phenotypic plasticity and evolutionary demographic responses to climate change: Taking theory out to the field. *Functional Ecology*, *27*(4), 967–979. https://doi.org/10.1111/j.1365-2435.2012.02043.x

Chien, L. T., & Hwang, D. F. (2001). Effects of thermal stress and vitamin C on lipid peroxidation and fatty acid composition in the liver of thornfish Terapon jarbua. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, *128*(1), 91–97. https://doi.org/10.1016/s1096-4959(00)00299-2

Clark, M. S., Thorne, M. A. S., King, M., Hipperson, H., Hoffman, J. I., & Peck, L. S. (2018). Life in the intertidal: Cellular responses, methylation and epigenetics. *Functional Ecology*, *32*(8), 1982–1994. https://doi.org/10.1111/1365-2435.13077

Collins, M., Clark, M. S., Spicer, J. I., & Truebano, M. (2021). Transcriptional frontloading contributes to cross-tolerance between stressors. *Evolutionary Applications*, *14*(2), 577–587. https://doi.org/10.1111/eva.13142

Collins, M., Clark, M. S., & Truebano, M. (2023). The environmental cellular stress response: The intertidal as a multistressor model. *Cell Stress and Chaperones*, *28*(5), 467–475. https://doi.org/10.1007/s12192-023-01348-7

Collins, M., Peck, L. S., & Clark, M. S. (2021). Large within, and between, species differences in marine cellular responses: Unpredictability in a changing environment. *Science of The Total Environment*, *794*, 148594. https://doi.org/10.1016/j.scitotenv.2021.148594

Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E. L., Fabry, M. H., Kneuss, E., & Hannon, G. J. (2018). piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annual Review of Genetics*, *52*(1), 131–157. https://doi.org/10.1146/annurev-genet-120417-031441

Daniel Schlenk & William H. Benson. (2017). *Target Organ Toxicity in Marine and Freshwater Teleosts* (1.ª ed.). CRC Press.

Das, P., Gupta, A., & Manna, S. K. (2005). Heat shock protein 70 expression in different tissues of Cirrhinus mrigala (Ham.) following heat stress. *Aquaculture Research*, *36*(6), 525–529. https://doi.org/10.1111/j.1365-2109.2005.01214.x

DeBiasse, M. B., & Kelly, M. W. (2016). Plastic and Evolved Responses to Global Change: What Can We Learn from Comparative Transcriptomics?: Table 1. *Journal of Heredity*, *107*(1), 71–81. https://doi.org/10.1093/jhered/esv073

Denz, C. R., Narshi, A., Zajdel, R. W., & Dube, D. K. (2004). Expression of a novel cardiac-specific tropomyosin isoform in humans. *Biochemical and Biophysical Research Communications*, *320*(4), 1291–1297. https://doi.org/10.1016/j.bbrc.2004.06.084

Dilworth, J., Million, W. C., Ruggeri, M., Hall, E. R., Dungan, A. M., Muller, E. M., & Kenkel, C. D. (2024). Synergistic response to climate stressors in coral is associated with genotypic variation in baseline expression. *Proceedings of the Royal Society B: Biological Sciences*, *291*(2019), 20232447. https://doi.org/10.1098/rspb.2023.2447

Dolbeth, M., Martinho, F., Leitão, R., Cabral, H., & Pardal, M. A. (2007). Strategies of Pomatoschistus minutus and Pomatoschistus microps to cope with environmental instability. *Estuarine, Coastal and Shelf Science*, *74*(1–2), 263–273. https://doi.org/10.1016/j.ecss.2007.04.016

Donelson, J. M., Sunday, J. M., Figueira, W. F., Gaitán-Espitia, J. D., Hobday, A. J., Johnson, C. R., Leis, J. M., Ling, S. D., Marshall, D., Pandolfi, J. M., Pecl, G., Rodgers, G. G., Booth, D. J., & Munday, P. L. (2019). Understanding interactions between plasticity, adaptation and range shifts in response to marine environmental change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1768), 20180186. https://doi.org/10.1098/rstb.2018.0186

Dong, Y., Miller, L. P., Sanders, J. G., & Somero, G. N. (2008). Heat-shock protein 70 (Hsp70) expression in four limpets of the genus Lottia: Interspecific variation in constitutive and inducible synthesis correlates with in situ exposure to heat stress. *The Biological Bulletin*, *215*(2), 173–181. https://doi.org/10.2307/25470698

Elmer, M. C., Monro, K., Thompson, H., Stuckey, A., & Kellermann, V. (2025). Phenotypic plasticity underlies seasonal and latitudinal variation in thermal tolerance in a native bee. *Ecology*, *106*(9), e70183. https://doi.org/10.1002/ecy.70183

Seyednasrollah, Laiho, A., & Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, *16*(1), 59–70. https://doi.org/10.1093/bib/bbt086

Fan, Q., & Jespersen, D. (2025). Proteases and the ubiquitin-proteasome system: Understanding protein degradation under heat stress in plants. *Environmental and Experimental Botany*, *237*, 106174. https://doi.org/10.1016/j.envexpbot.2025.106174

Feidantsis, K., Antonopoulou, E., Lazou, A., Pörtner, H. O., & Michaelidis, B. (2013). Seasonal variations of cellular stress response of the gilthead sea bream (Sparus aurata). *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology*, *183*(5), 625–639. https://doi.org/10.1007/s00360-012-0735-y

George, R., Job, S., Semba, M., Monga, E., Lugendo, B., Tuda, A. O., & Kimirei, I. (2024). High-frequency dynamics of pH, dissolved oxygen, and temperature in the coastal ecosystems of the Tanga-Pemba Seascape: Implications for upwelling-enhanced ocean acidification and deoxygenation. *Frontiers in Marine Science*, *10*, 1286870. https://doi.org/10.3389/fmars.2023.1286870

Gerber, L., Clow, K. A., Mark, F. C., & Gamperl, A. K. (2020). Improved mitochondrial function in salmon (Salmo salar) following high temperature acclimation suggests that there are cracks in the proverbial 'ceiling'. *Scientific Reports*, *10*(1), 21636. https://doi.org/10.1038/s41598-020-78519-4

Gerlich, H. S., Holmstrup, M., Schmidt, N. M., & Høye, T. T. (2025). Phenological responses to climate change across taxa and local habitats in a high-Arctic arthropod community. *Ecological Monographs*, *95*(1), e1643. https://doi.org/10.1002/ecm.1643

Ghalambor, C. K., McKAY, J. K., Carroll, S. P., & Reznick, D. N. (2007). Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology*, *21*(3), 394–407. https://doi.org/10.1111/j.1365-2435.2007.01283.x

Glibert, P. M., Cai, W.-J., Hall, E. R., Li, M., Main, K. L., Rose, K. A., Testa, J. M., & Vidyarathna, N. K. (2022). Stressing over the Complexities of Multiple Stressors in Marine and Estuarine Systems. *Ocean-Land-Atmosphere Research*, *2022*, 2022/9787258. https://doi.org/10.34133/2022/9787258

Gobler, C. J., & Baumann, H. (2016). Hypoxia and acidification in ocean ecosystems: Coupled dynamics and effects on marine life. *Biology Letters*, *12*(5), 20150976. https://doi.org/10.1098/rsbl.2015.0976

Goetzman, E., Gong, Z., Zhang, B., & Muzumdar, R. (2023). Complex II Biology in Aging, Health, and Disease. *Antioxidants*, *12*(7), 1477. https://doi.org/10.3390/antiox12071477

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Han, G., Cartwright, S. R., Ganmanee, M., Chan, B. K. K., Adzis, K. A. A., Hutchinson, N., Wang, J., Hui, T. Y., Williams, G. A., & Dong, Y. (2019). High thermal stress responses of Echinolittorina snails at their range edge predict population vulnerability to future warming. *Science of The Total Environment*, *647*, 763–771. https://doi.org/10.1016/j.scitotenv.2018.08.005

Hanson, N. A., Lavallee, M. B., & Thiele, R. H. (2021). Apophenia and anesthesia: How we sometimes change our practice prematurely. *Canadian Journal of Anesthesia/Journal Canadien d'anesthésie*, *68*(8), 1185–1196. https://doi.org/10.1007/s12630-021-02005-2

Heymann, F., & Tacke, F. (2016). Immunology in the liver—From homeostasis to disease. *Nature Reviews Gastroenterology & Hepatology*, *13*(2), 88–110. https://doi.org/10.1038/nrgastro.2015.200

Hoffmann, A. A., & Bridle, J. (2022). The dangers of irreversibility in an age of increased uncertainty: Revisiting plasticity in invertebrates. *Oikos*, *2022*(4), e08715. https://doi.org/10.1111/oik.08715

Instituto Português do Mar e da Atmosfera. (2025). https://www.ipma.pt/pt/oclima/ondascalor/

Ishimatsu, A., & Graham, J. B. (2011). Roles of Environmental Cues for Embryonic Incubation and Hatching in Mudskippers. *Integrative and Comparative Biology*, *51*(1), 38–48. https://doi.org/10.1093/icb/icr018

Islam, M. J., Kunzmann, A., & Slater, M. J. (2022). Responses of aquaculture fish to climate change-induced extreme temperatures: A review. *Journal of the World Aquaculture Society*, *53*(2), 314–366. https://doi.org/10.1111/jwas.12853

Jackson, D. J., Cerveau, N., & Posnien, N. (2024). De novo assembly of transcriptomes and differential gene expression analysis using short-read data from emerging model organisms – a brief guide. *Frontiers in Zoology*, *21*(1), 17. https://doi.org/10.1186/s12983-024-00538-y

Jesus, T. F., Grosso, A. R., Almeida-Val, V. M. F., & Coelho, M. M. (2016). Transcriptome profiling of two Iberian freshwater fish exposed to thermal stress. *Journal of Thermal Biology*, *55*, 54–61. https://doi.org/10.1016/j.jtherbio.2015.11.009

Jiang, S., Xiong, Y., Zhang, W., Zhu, J., Cheng, D., Gong, Y., Wu, Y., Qiao, H., & Fu, H. (2022). Molecular Characterization of a Novel Cathepsin L in Macrobrachium nipponense and Its Function in Ovary Maturation. *Frontiers in Endocrinology*, *12*, 816813. https://doi.org/10.3389/fendo.2021.816813

Keagy, J., Drummond, C. P., Gilbert, K. J., Grozinger, C. M., Hamilton, J., Hines, H. M., Lasky, J., Logan, C. A., Sawers, R., & Wagner, T. (2023). Landscape transcriptomics as a tool for addressing global change effects across diverse species. *Molecular Ecology Resources*, *25*(5), e13796. https://doi.org/10.1111/1755-0998.13796

Kelly, W. G. (2014). Transgenerational epigenetics in the germline cycle of Caenorhabditis elegans. *Epigenetics & Chromatin*, *7*(1), 6. https://doi.org/10.1186/1756-8935-7-6

Kocerha, J., & Denslow, N. D. (2025). Steroidogenic acute regulatory protein in fish. *Journal of Endocrinology*, *265*(1), e240232. https://doi.org/10.1530/JOE-24-0232

Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A Beginner's Guide to Analysis of RNA Sequencing Data. *American Journal of Respiratory Cell and Molecular Biology*, *59*(2), 145–157. https://doi.org/10.1165/rcmb.2017-0430TR

Kütter, M. T., Romano, L. A., Ventura-Lima, J., Tesser, M. B., & Monserrat, J. M. (2014). Antioxidant and toxicological effects elicited by alpha-lipoic acid in aquatic organisms. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, *162*, 70–76. https://doi.org/10.1016/j.cbpc.2014.03.008

Leeuwis, R. H. J., & Gamperl, A. K. (2022). ADAPTATIONS AND PLASTIC PHENOTYPIC RESPONSES OF MARINE ANIMALS TO THE ENVIRONMENTAL CHALLENGES OF THE HIGH INTERTIDAL ZONE. *Oceanography and Marine Biology*, *60*.

Liu, B., & Qian, S. (2014). Translational reprogramming in cellular stress response. *WIREs RNA*, *5*(3), 301–305. https://doi.org/10.1002/wrna.1212

Logan, M. L., & Cox, C. L. (2020). Genetic Constraints, Transcriptome Plasticity, and the Evolutionary Response to Climate Change. *Frontiers in Genetics*, *11*, 538226. https://doi.org/10.3389/fgene.2020.538226

Lopes, J. F., Lopes, C. L., & Dias, J. M. (2019). Climate Change Impact in the Ria de Aveiro Lagoon Ecosystem: A Case Study. *Journal of Marine Science and Engineering*, *7*(10), 352. https://doi.org/10.3390/jmse7100352

Madeira, D., Vinagre, C., Costa, P., & Diniz, M. (2014). Histopathological alterations, physiological limits, and molecular changes of juvenile Sparus aurata in response to thermal stress. *Marine Ecology Progress Series*, *505*, 253–266. https://doi.org/10.3354/meps10794

Mallard, F., Nolte, V., & Schlötterer, C. (2020). The Evolution of Phenotypic Plasticity in Response to Temperature Stress. *Genome Biology and Evolution*, *12*(12), 2429–2440. https://doi.org/10.1093/gbe/evaa206

Marigómez, I., Múgica, M., Izagirre, U., & Sokolova, I. M. (2017). Chronic environmental stress enhances tolerance to seasonal gradual warming in marine mussels. *PLOS ONE*, *12*(3), e0174359. https://doi.org/10.1371/journal.pone.0174359

McCluskey, B. M., & Postlethwait, J. H. (2014). Phylogeny of Zebrafish, a "Model Species," within Danio, a "Model Genus". *Molecular Biology and Evolution*, *32*(3), 635–652. https://doi.org/10.1093/molbev/msu325

Missionário, M., Almeida, C., Fernandes, J. F., Vinagre, C., Madeira, C., & Madeira, D. (2024). Shallow water fish display low phenotypic plasticity to ocean warming and extreme weather events regardless of previous thermal history. *Estuarine, Coastal and Shelf Science*, *305*, 108849. https://doi.org/10.1016/j.ecss.2024.108849

Molina, J. M., & E.Blasina, G. (2025). *Volume 2: Marine Ecology: An Ecosystemic View of Anthropogenic Impacts*. CRC Press. https://www.routledge.com/Volume-2-Marine-Ecology-An-Ecosystemic-View-of-Anthropogenic-Impacts/Molina-Blasina/p/book/9780367456603

Moritz, C., & Agudo, R. (2013). The Future of Species Under Climate Change: Resilience or Decline? *Science*, *341*(6145), 504–508. https://doi.org/10.1126/science.1237190

Moschella, M. C., Menzies, K., Tsao, L., Lieb, M. A., Kohtz, J. D., Kohtz, D. S., & Taubman, M. B. (1999). SM-20 is a novel growth factor-responsive gene regulated during skeletal muscle development and differentiation. *Gene Expression*, *8*(1), 59–66.

Muir, C. A., Bork, B. S., Neff, B. D., & Damjanovski, S. (2022). Proteomic analysis of temperature-dependent developmental plasticity within the ventricle of juvenile Atlantic salmon (Salmo salar). *Current Research in Physiology*, *5*, 344–354. https://doi.org/10.1016/j.crphys.2022.07.005

Nascimento, Â., Biguino, B., Borges, C., Cereja, R., Cruz, J. P. C., Sousa, F., Dias, J., Brotas, V., Palma, C., & Brito, A. C. (2021). Tidal variability of water quality parameters in a mesotidal estuary (Sado Estuary, Portugal). *Scientific Reports*, *11*(1), 23112. https://doi.org/10.1038/s41598-021-02603-6

Olivares-Rubio, H. F., & Vega-López, A. (2016). Fatty acid metabolism in fish species as a biomarker for environmental monitoring. *Environmental Pollution*, *218*, 297–312. https://doi.org/10.1016/j.envpol.2016.07.005

Oliver, E. C. J., Donat, M. G., Burrows, M. T., Moore, P. J., Smale, D. A., Alexander, L. V., Benthuysen, J. A., Feng, M., Sen Gupta, A., Hobday, A. J., Holbrook, N. J., Perkins-Kirkpatrick, S. E., Scannell, H. A., Straub, S. C., & Wernberg, T. (2018). Longer and more frequent marine heatwaves over the past century. *Nature Communications*, *9*(1), 1324. https://doi.org/10.1038/s41467-018-03732-9

Page, T. M., & Lawley, J. W. (2022). The Next Generation Is Here: A Review of Transcriptomic Approaches in Marine Ecology. *Frontiers in Marine Science*, *9*, 757921. https://doi.org/10.3389/fmars.2022.757921

Payton, S. L., Johnson, P. D., & Jenny, M. J. (2016). Comparative physiological, biochemical, and molecular thermal stress response profiles for two Unionid freshwater mussel species. *Journal of Experimental Biology*, jeb.140129. https://doi.org/10.1242/jeb.140129

Pen, I., Uller, T., Feldmeyer, B., Harts, A., While, G. M., & Wapstra, E. (2010). Climate-driven population divergence in sex-determining systems. *Nature*, *468*(7322), 436–438. https://doi.org/10.1038/nature09512

Pereira, R. J., Lima, T. G., Pierce-Ward, N. T., Chao, L., & Burton, R. S. (2021). Recovery from hybrid breakdown reveals a complex genetic architecture of mitonuclear incompatibilities. *Molecular Ecology*, *30*(23), 6403–6416. https://doi.org/10.1111/mec.15985

Place, S. P., Menge, B. A., & Hofmann, G. E. (2012). Transcriptome profiles link environmental variation and physiological response of *Mytilus californianus* between Pacific tides. *Functional Ecology*, *26*(1), 144–155. https://doi.org/10.1111/j.1365-2435.2011.01924.x

Porcelli, D., Butlin, R. K., Gaston, K. J., Joly, D., & Snook, R. R. (2015). The environmental genomics of metazoan thermal adaptation. *Heredity*, *114*(5), 502–514. https://doi.org/10.1038/hdy.2014.119

Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to *de novo* transcriptome assembly and annotation. *Briefings in Bioinformatics*, *23*(2), bbab563. https://doi.org/10.1093/bib/bbab563

Ragland, G. J., & Kingsolver, J. G. (2007). Influence of seasonal timing on thermal ecology and thermal reaction norm evolution in Wyeomyia smithii. *Journal of Evolutionary Biology*, *20*(6), 2144–2153. https://doi.org/10.1111/j.1420-9101.2007.01428.x

Rigal, F., Chevalier, T., Lorin-Nebel, C., Charmantier, G., Tomasini, J.-A., Aujoulat, F., & Berrebi, P. (2008). Osmoregulation as a potential factor for the differential distribution of two cryptic gobiid species, Pomatoschistus microps and P. marmoratus in French Mediterranean lagoons. *Scientia Marina*, *72*(3), 469–476. https://doi.org/10.3989/scimar.2008.72n3469

Rise, M. L., Martyniuk, C. J., & Chen, M. (2019). Comparative physiology and aquaculture: Toward Omics-enabled improvement of aquatic animal health and sustainable production. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *31*, 100603. https://doi.org/10.1016/j.cbd.2019.100603

Ros, M., Guerra-García, J. M., Lignot, J.-H., & Rivera-Ingraham, G. A. (2021). Environmental stress responses in sympatric congeneric crustaceans: Explaining and predicting the context-dependencies of invader impacts. *Marine Pollution Bulletin*, *170*, 112621. https://doi.org/10.1016/j.marpolbul.2021.112621

Sabino-Pinto, J., Goedbloed, D. J., Sanchez, E., Czypionka, T., Nolte, A. W., & Steinfartz, S. (2019). The Role of Plasticity and Adaptation in the Incipient Speciation of a Fire Salamander Population. *Genes*, *10*(11), 875. https://doi.org/10.3390/genes10110875

Salgado, J. P., Cabral, H. N., & Costa, M. J. (2004). *Feeding ecology of the gobies Pomatoschistus minutus (Pallas, 1770) and Pomatoschistus microps (Krøyer, 1838) in the upper Tagus estuary, Portugal*.

Sasaki, M. C., & Dam, H. G. (2019). Integrating patterns of thermal tolerance and phenotypic plasticity with population genetics to improve understanding of vulnerability to warming in a widespread copepod. *Global Change Biology*, *25*(12), 4147–4164. https://doi.org/10.1111/gcb.14811

Schaum, C.-E., Buckling, A., Smirnoff, N., & Yvon-Durocher, G. (2022). Evolution of thermal tolerance and phenotypic plasticity under rapid and slow temperature fluctuations. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1980), 20220834. https://doi.org/10.1098/rspb.2022.0834

Schulte, P. M. (2014). What is environmental stress? Insights from fish living in a variable environment. *The Journal of Experimental Biology*, *217*(Pt 1), 23–34. https://doi.org/10.1242/jeb.089722

Schwob, E., & Martin, R. P. (1992). New yeast actin-like gene required late in the cell cycle. *Nature*, *355*, 179–180.

Seebacher, F., White, C. R., & Franklin, C. E. (2014). Physiological plasticity increases resilience of ectothermic animals to climate change. *Nature Climate Change*, *5*(1), 61–66. https://doi.org/10.1038/nclimate2457

Seluzicki, C. M., Razavi-Mohseni, M., Türker, F., Patel, P., Hua, B., Beer, M. A., Goff, L., & Margolis, S. S. (2025). Regulation of translation elongation and integrated stress response in heat-shocked neurons. *Cell Reports*, *44*(5), 115639. https://doi.org/10.1016/j.celrep.2025.115639

Serreze, M. C., & Barry, R. G. (2011). Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change*, *77*(1), 85–96. https://doi.org/10.1016/j.gloplacha.2011.03.004

Shomron, N. (Ed.). (2021). *Deep Sequencing Data Analysis* (Vol. 2243). Springer US. https://doi.org/10.1007/978-1-0716-1103-6

Smale, D. A., Wernberg, T., Oliver, E. C. J., Thomsen, M., Harvey, B. P., Straub, S. C., Burrows, M. T., Alexander, L. V., Benthuysen, J. A., Donat, M. G., Feng, M., Hobday, A. J., Holbrook, N. J.,

Perkins-Kirkpatrick, S. E., Scannell, H. A., Sen Gupta, A., Payne, B. L., & Moore, P. J. (2019). Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nature Climate Change*, *9*(4), 306–312. https://doi.org/10.1038/s41558-019-0412-1

Somero, G. N. (2020). The cellular stress response and temperature: Function, regulation, and evolution. *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, *333*(6), 379–397. https://doi.org/10.1002/jez.2344

Sommer, B., Hodge, J. M., Lachs, L., Cant, J., Pandolfi, J. M., & Beger, M. (2024). Decadal demographic shifts and size-dependent disturbance responses of corals in a subtropical warming hotspot. *Scientific Reports*, *14*(1), 6327. https://doi.org/10.1038/s41598-024-56890-w

Song, Y.-F., Luo, Z., Huang, C., Chen, Q.-L., Pan, Y., & Xu, Y.-H. (2015). Endoplasmic Reticulum Stress–Related Genes in Yellow Catfish Pelteobagrus fulvidraco: Molecular Characterization, Tissue Expression, and Expression Responses to Dietary Copper Deficiency and Excess. *G3 Genes|Genomes|Genetics*, *5*(10), 2091–2104. https://doi.org/10.1534/g3.115.019950

Souza, A. T., Ilarri, M. I., Timóteo, S., Marques, J. C., & Martins, I. (2018). Assessing the effects of temperature and salinity oscillations on a key mesopredator fish from European coastal systems. *Science of The Total Environment*, *640–641*, 1332–1345. https://doi.org/10.1016/j.scitotenv.2018.05.348

Stillman, J. H., & Tagmount, A. (2009). Seasonal and latitudinal acclimatization of cardiac transcriptome responses to thermal stress in porcelain crabs, *Petrolisthes cinctipes*. *Molecular Ecology*, *18*(20), 4206–4226. https://doi.org/10.1111/j.1365-294X.2009.04354.x

Stollewerk, A., Kratina, P., Sentis, A., Chaparro-Pedraza, C., Decaestecker, E., De Meester, L., Eyice, O., Govaert, L., Jones, J. I., Laforsch, C., Madeira, C., Narwani, A., Oostra, V., Raeymaekers, J. A. M., Rossberg, A. G., Schott, M., Stoks, R., Van Velzen, E., & Boukal, D. (2025). Plasticity in climate change responses. *Biological Reviews*, brv.70056. https://doi.org/10.1111/brv.70056

Stricher, F., Macri, C., Ruff, M., & Muller, S. (2013). HSPA8/HSC70 chaperone protein: Structure, function, and chemical targeting. *Autophagy*, *9*(12), 1937–1954. https://doi.org/10.4161/auto.26448

Sunday, J. M., Bates, A. E., Kearney, M. R., Colwell, R. K., Dulvy, N. K., Longino, J. T., & Huey, R. B. (2014). Thermal-safety margins and the necessity of thermoregulatory behavior across latitude and elevation. *Proceedings of the National Academy of Sciences*, *111*(15), 5610–5615. https://doi.org/10.1073/pnas.1316145111

Sydeman, W. J., Poloczanska, E., Reed, T. E., & Thompson, S. A. (2015). Climate change and marine vertebrates. *Science*, *350*(6262), 772–777. https://doi.org/10.1126/science.aac9874

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

Tan, H., Onichtchouk, D., & Winata, C. (2016). DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish. *Zebrafish*, *13*(1), 54–60. https://doi.org/10.1089/zeb.2015.1179

Tao, Y.-F., Qiang, J., Bao, J.-W., Chen, D.-J., Yin, G.-J., Xu, P., & Zhu, H.-J. (2018). Changes in Physiological Parameters, Lipid Metabolism, and Expression of MicroRNAs in Genetically Improved Farmed Tilapia (Oreochromis niloticus) With Fatty Liver Induced by a High-Fat Diet. *Frontiers in Physiology*, *9*. https://doi.org/10.3389/fphys.2018.01521

Tougard, C., Folly, J., & Berrebi, P. (2014). New Light on the Evolutionary History of the Common Goby (Pomatoschistus microps) with an Emphasis on Colonization Processes in the Mediterranean Sea. *PLoS ONE*, *9*(3), e91576. https://doi.org/10.1371/journal.pone.0091576

Vargas, C. I. C., Vaz, N., & Dias, J. M. (2017). An evaluation of climate change effects in estuarine salinity patterns: Application to Ria de Aveiro shallow water system. *Estuarine, Coastal and Shelf Science*, *189*, 33–45. https://doi.org/10.1016/j.ecss.2017.03.001

Vinagre, C., Santos, F. D., Cabral, H., & Costa, M. J. (2011). Impact of climate warming upon the fish assemblages of the Portuguese coast under different scenarios. *Regional Environmental Change*, *11*(4), 779–789. https://doi.org/10.1007/s10113-011-0215-z

Von Schmalensee, L., Caillault, P., Gunnarsdóttir, K. H., Gotthard, K., & Lehmann, P. (2023). Seasonal specialization drives divergent population dynamics in two closely related butterflies. *Nature Communications*, *14*(1), 3663. https://doi.org/10.1038/s41467-023-39359-8

Wang, J., Ma, L.-X., & Dong, Y.-W. (2023). Coping with harsh heat environments: Molecular adaptation of metabolic depression in the intertidal snail Echinolittorina radiata. *Cell Stress and Chaperones*, *28*(5), 477–491. https://doi.org/10.1007/s12192-022-01295-9

Wang, W., Su, S., Dong, P., Feng, W., Li, J., Zhang, C., & Tang, Y. (2023). Effects of Seasonal Photoperiod on Growth, Lipid Metabolism, and Antioxidant Response in the Huanghe Carp (Cyprinus carpio haematopterus). *Fishes*, *8*(12), 595. https://doi.org/10.3390/fishes8120595

Weick, E.-M., & Miska, E. A. (2014). piRNAs: From biogenesis to function. *Development*, *141*(18), 3458–3471. https://doi.org/10.1242/dev.094037

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*(9), 618–630. https://doi.org/10.1038/nrmicro2852

Yampolsky, L. Y., Schaer, T. M. M., & Ebert, D. (2014). Adaptive phenotypic plasticity and local adaptation for temperature tolerance in freshwater zooplankton. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776), 20132744. https://doi.org/10.1098/rspb.2013.2744

Yu, Y., Mai, Y., Zheng, Y., & Shi, L. (2024). Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology*, *25*(1), 254. https://doi.org/10.1186/s13059-024-03401-9

Zhang, J., Li, M., Meng, D., Xu, S., Teame, T., Yao, Y., Yang, Y., Zhang, Z., Ran, C., Jijakli, M. H., Ding, Q., & Zhou, Z. (2025). Review on Omics Approaches in Aquatic Animal Nutrition: Current Status, Limitations, and Perspectives. *The Journal of Nutrition*, S0022316625004973. https://doi.org/10.1016/j.tjnut.2025.08.019

A

AN APPENDIX

A.1 Tables

Table A.1 - Methods steps and softwares

Step	Software and Algorithms	Source/URL
Quality control	FastQC v0.11.9	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Pre-processing	TrimGalore v0.6.0	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Transcript assembly	Trinity v2.14.0	https://github.com/trinityrnaseq/trinityrnaseq
Read alignment	Bowtie2 v2.4.2, Samtools v1.8	https://github.com/BenLangmead/bowtie2
ORF prediction	TransDecoder v5.5.0	https://github.com/TransDecoder/TransDecoder
Functional annotation	BLASTp v2.12.0, Swiss-Prot	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Expression quantification	Kallisto v0.48.0, tximport	https://pachterlab.github.io/kallisto
Differential expression analysis	edgeR v4.7.2, limma v3.65.0,	https://bioconductor.org/packages/edgeR
Gene set enrichment analysis	multiGSEA v1.19.0, org.Dr.eg.db	https://bioconductor.org/packages/multiGSEA
Protein-Protein interaction analysis	STRING v12.0	https://string-db.org/
Data visualization	ggplot2 v3.5.2, Inkscape v1.4	https://ggplot2.tidyverse.org/, https://inkscape.org/

Table A.2. Environmental data collection. Source: Madeira et al., 2025 (under revision).

Climate type	**Csb**		**Csb/Csa**		**Csa**	
Coastal region	**Northwestern**		**Southwestern**		**Southern**	
Location	**Ria de Aveiro**		**Troia Peninsula**		**Ria Formosa**	
Season	**Spring (A-J)**	**Summer (J-S)**	**Spring (A-J)**	**Summer (J-S)**	**Spring (A-J)**	**Summer (J-S)**
Avg. Shallow seawater temp. (ºC, <1m depth, collection days)	21.1±0.8	18.6±1.6	22.0±0.9	25.9±3.5	23.4±0.9	25.9±0.5
Air temp. (ºC, 1,5 m from ground level, collection days)	18.5	22.6	20.8	28.2	28.0	26.0
Avg. max. shallow seawater temp. (ºC, collection sites)	22.8±1.6	22.8±0.7	28.6±1.4	30.0±1.1	–	32.9±1.3
Abs. max. Shallow seawater temp. (ºC, collection sites)	23.9	23.6	31.6	34.1	–	35.1
Avg. seasonal seawater temp. (ºC, collection sites)	17.5±0.7	18.8±1.0	17.8±1.9	20.2±0.2	–	25.0±0.8
Avg. monthly seawater temp. amplitude (ºC, max-min, collection sites)	7.9±1.6	8.1±0.6	17.5±2.2	14.4±1.4	–	15.1±1.8
Air temp. range (avg., min. To max., ºC, CSN 1981-2010)	12.6-20.4	15.8-24.1	11.6-23.7	15.8-29.2	14.6-18.8	19.1-28,2
Highest air temperature extreme, ºC and (year) during 1981-2010		39,3 (2010)		43,5 (1995)		44,3 (2004)
No. of heatwave events and (no. days) since 2010-2022		1 (6d)		6 (6-16d)		0 (na)
Salinity	27.6±1.03	34.0±0.4	38.0±1.6	36±1.8	35.2±0.4	35.0±0.0
[O] ppm2	NA	7.6±0.3	7.2±0.9	8.3±1.1	8.0±0.7	7.5±1.1
pH	8.2±0.006	8.1±0.04	7.9±0.08	8.1±0.16	7.62±0.4	8.1±0.07

A.2 Scripts

A.2.1 Main Script

```
================================================================================
====
        # TASK 1
        #
================================================================================
====

        # 1. Load Blastp file
        swissprot <- load_blastp_data("01-Input/Blastp/ExtremeOceans_Blastp_SwissProt.csv")
        dim(swissprot) # 161.919 15
        zebrafish <- load_blastp_data("01-Input/Blastp/ExtremeOceans_Blastp_Zebrafish.csv")
        dim(zebrafish) # 78.410 15
        plat <- load_blastp_data("01-Input/Blastp/ExtremeOceans_Blastp_Plat.csv")
        dim(plat) # 13.905 15

        # Check for missing values
        sum(is.na(swissprot))          # Total count of missing values
        colSums(is.na(swissprot))      # Count per column
        anyNA(swissprot)               # TRUE if any missing values exist



        # 2. Load Kallisto files

        ## File paths to access each sample's quantification data
        abundance_files <- file.path("./01-Input/Kallisto", sample_name, "abundance.tsv")
        ### output e.g.,
        ###
        ###    [1] "./01-Input/Kallisto/F14.Sp_TR/abundance.tsv"
        ###    [2] "./01-Input/Kallisto/F15.Su_TR/abundance.tsv"
        ###    [3] "./01-Input/Kallisto/F16.Su_RA/abundance.tsv"

        names(abundance_files) <- sample_name

        ## Import kallisto files (1st time)

        abundance_tsv <- tximport(
          abundance_files,
          type = "kallisto",
          txOut = TRUE,
          countsFromAbundance = "lengthScaledTPM"
          )

        tx2gene_df <-unlist(sapply(rownames(abundance_tsv$counts),
                           function(x) unlist(strsplit(x, "i", fixed = TRUE))[1]))
        ### logic e.g.,
        ###
        ### TranscriptID - TRINITY_DN168445_c0_g1_i1 (i.e., [c]luster| [g]ene| [i]soform])
```

```
###
### sapply(rownames(abundance_tsv$counts) - For every transcript ID, the code:
### function(x) unlist(strsplit(x, "i", fixed = TRUE))[1])) - splits the word at "i"
###                                          [TRINITY_DN168445_c0_g1_] + [i1]
###                                          and select the first "[1]" part
### GenesID - TRINITY_DN168445_c0_g1

### data frame => | TranscriptID | GeneID |
tx2gene_df <-data.frame(rownames(abundance_tsv$counts), tx2gene_df)
colnames(tx2gene_df)<-c("TranscriptID", "GeneID")
print(head(tx2gene_df))

## Import Kallisto files (2nd time) but know with tx2gene @param
abundance_tsv <- tximport(
  abundance_files,
  type = "kallisto",
  tx2gene = tx2gene_df,
  countsFromAbundance = "lengthScaledTPM"
  )

## Explore the object created with tximport
names(abundance_tsv)
View(head(abundance_tsv$abundance)) # matrix containing TPM's
View(head(abundance_tsv$counts)) # matrix containing read counts
View(head(abundance_tsv$length))

save(abundance_tsv, file = "03-Output/01-DEG-Analysis/analysis-ready-data/abundance_tsv")

# 3. Merge Blastp(1.) + Kallisto(2.) counts

## Preparing data to merge | Purpose .: Be easy to merge
### Kallisto data - Create a new data frame that contains a "GeneID" column
counts <- abundance_tsv$counts
counts <- cbind(as.data.frame(row.names(counts)), counts)
colnames(counts)[1]<-"GeneID"

### Blastp data - Add "GeneID" column + clean Blastp data

swissprot <- select_best_hits(swissprot)
dim(swissprot) # 37.824 16

zebrafish <- select_best_hits(zebrafish)
dim(zebrafish) # 17.932 16

plat <- select_best_hits(plat)
dim(plat) # 3.743 16

## Merge (Blastp + kallisto counts) | [Bp]Blastp + [k]kallisto - BpK
BpK_swissprot <- merge(swissprot, counts, by = "GeneID")
dim(BpK_swissprot) # 37.824 34

BpK_zebrafish <- merge(zebrafish, counts, by = "GeneID")
dim(BpK_zebrafish) # 17.932 34

BpK_plat <- merge(plat, counts, by = "GeneID")
dim(BpK_plat) # 3.743 34
```

```
## Save files
save_to_excel(BpK_swissprot, "03-Output/01-DEG-Analysis/preprocessed-data/BpK-Swissprot.xlsx")
save_to_excel(BpK_zebrafish, "03-Output/01-DEG-Analysis/preprocessed-data/BpK-Zebrafish.xlsx")
save_to_excel(BpK_plat, "03-Output/01-DEG-Analysis/preprocessed-data/BpK-Plat.xlsx")

#
# TASK 2
#

# Data needed:
BpK_swissprot <- read.xlsx("03-Output/01-DEG-Analysis/preprocessed-data/BpK-Swissprot.xlsx")
BpK_zebrafish <- read.xlsx("03-Output/01-DEG-Analysis/preprocessed-data/BpK-Zebrafish.xlsx")
BpK_plat <- read.xlsx("03-Output/01-DEG-Analysis/preprocessed-data/BpK-Plat.xlsx")

# Process data:
Data_SP <- process_BpK_data(BpK_swissprot) #15947
dim(Data_SP)
#save_to_excel(Data_SP, "./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-SP.xlsx")
Data_Z <- process_BpK_data(BpK_zebrafish) #2975
dim(Data_Z)
#save_to_excel(Data_Z, "./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-Z.xlsx")
Data_P <- process_BpK_data(BpK_plat) #145
dim(Data_P)
#save_to_excel(Data_P, "./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-P.xlsx")


#
# TASK 2.1 - Global annotation performance
#

# 2.1.1 e-value and pident behavior (before and after processing)

## pident raw data :
#barData_swissprot <- process_barplot_data(swissprot, "swissprot", metric = "pident")
#barData_zebrafish <- process_barplot_data(zebrafish, "zebrafish", metric = "pident")
#barData_plat <- process_barplot_data(plat, "platyhelminthes", metric = "pident")

#data_list <- list("Swiss-prot" = barData_swissprot,
```

```
#               "Zebrafish" = barData_zebrafish,
#               "Platyhelminth" = barData_plat)

#barData <- prepare_summary(data_list)

## Bar plot
#plot_barplot(barData, save_plot = TRUE,
#             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-pident-add-geneID.pdf" )

#plot_barplot(barData, save_plot = TRUE,
#             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-pident-raw-data.pdf" )

## Pident before processing :

# Prepare each data set
barData_BpKsp <- process_barplot_data(BpK_swissprot, "swissprot", metric = "pident")
barData_BpKz <- process_barplot_data(BpK_zebrafish, "zebrafish", metric = "pident")
barData_BpKp <- process_barplot_data(BpK_plat, "platyhelminthes", metric = "pident")

data_list <- list("Swiss-prot" = barData_BpKsp,
                  "Zebrafish" = barData_BpKz,
                  "Platyhelminth" = barData_BpKp)

barData <- prepare_summary(data_list)

## Bar plot
plot_barplot(barData, save_plot = TRUE,
             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-pident-before-processing.pdf" )

## Pident after processing :

# Prepare each data set
barData_sp <- process_barplot_data(Data_SP, "swissprot", metric = "pident")
barData_z <- process_barplot_data(Data_Z, "zebrafish", metric = "pident")
barData_p <- process_barplot_data(Data_P, "platyhelminthes", metric = "pident")

data_list <- list("Swiss-prot" = barData_sp,
                  "Zebrafish" = barData_z,
                  "Platyhelminth" = barData_p)

barData <- prepare_summary(data_list)

## Bar plot
plot_barplot(barData, save_plot = TRUE,
             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-pident-after-processing.pdf" )

## E-value raw data :

#barData_swissprot <- process_barplot_data(swissprot, "swissprot", metric = "evalue")
#barData_zebrafish <- process_barplot_data(zebrafish, "zebrafish", metric = "evalue")
#barData_plat <- process_barplot_data(plat, "platyhelminthes", metric = "evalue")
```

```
	#data_list <- list("Swiss-prot" = barData_swissprot,
	#		"Zebrafish" = barData_zebrafish,
	#		"Platyhelminth" = barData_plat)

	#barData <- prepare_summary(data_list)

	## Bar plot

	#plot_barplot(barData, save_plot = TRUE,
	#		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-evalue-add-geneID.pdf" )

	#plot_barplot(barData, save_plot = TRUE,
	#		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-evalue-raw-data.pdf" )

	## E-value before processing :

	barData_BpKsp <- process_barplot_data(BpK_swissprot, "swissprot", metric = "evalue")
	barData_BpKz <- process_barplot_data(BpK_zebrafish, "zebrafish", metric = "evalue")
	barData_BpKp <- process_barplot_data(BpK_plat, "platyhelminthes", metric = "evalue")

	data_list <- list("Swiss-prot" = barData_BpKsp,
			"Zebrafish" = barData_BpKz,
			"Platyhelminth" = barData_BpKp)

	barData <- prepare_summary(data_list)

	## Bar plot
	plot_barplot(barData, save_plot = FALSE)

	plot_barplot(barData, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-evalue-before-processing.pdf" )

	## E-value after processing :

	barData_sp <- process_barplot_data(Data_SP, "swissprot", metric = "evalue")
	barData_z <- process_barplot_data(Data_Z, "zebrafish", metric = "evalue")
	barData_p <- process_barplot_data(Data_P, "platyhelminthes", metric = "evalue")

	data_list <- list("Swiss-prot" = barData_sp,
			"Zebrafish" = barData_z,
			"Platyhelminth" = barData_p)

	barData <- prepare_summary(data_list)

	## Bar plot
	plot_barplot(barData, save_plot = FALSE)

	plot_barplot(barData, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-evalue-after-processing.pdf" )

	# 2.1.2 Common GeneNameID's on the different BLAST's outputs
```

```
	# Before processing
	Overlap_BpKSp_and_BpKZ <- analyse_overlap(BpK_swissprot, BpK_zebrafish,"swissprot",
"zebrafish")
	Overlap_BpKSp_and_BpKP <- analyse_overlap(BpK_swissprot, BpK_plat,"swissprot", "plat")
	Overlap_BpKZ_and_BpKP <- analyse_overlap(BpK_zebrafish, BpK_plat,"zebrafish", "plat")

	# After processing
	Overlap_Sp_and_Z <- analyse_overlap(Data_SP, Data_Z,"swissprot", "zebrafish")
	Overlap_Sp_and_P <- analyse_overlap(Data_SP, Data_P,"swissprot", "plat")
	Overlap_Z_and_P <- analyse_overlap(Data_Z, Data_P,"zebrafish", "plat")


	GeneID_Overlap_Sp_and_P <- analyse_overlap(Data_SP, Data_P,"swissprot", "plat",
                                column_name = "GeneID")
	GeneNameID_Overlap_Sp_and_P <- analyse_overlap(Data_SP, Data_P,"swissprot", "plat",
                                  column_name = "GeneNameID")

	## Venn Diagram
	vennData <- list("Swiss-prot" = BpK_swissprot$GeneID,
	            "Zebra-fish" = BpK_zebrafish$GeneID,
	            "Platyhelminth" = BpK_plat$GeneID)

	plot_vennDiagram(vennData, save_plot=TRUE,
	            output_file="./03-Output/01-DEG-Analysis/annotation-results/figures/vennDia-
gram-BpKcommonGeneID-before-processing.pdf")

	#
==========================================================================
====
	# TASK 2.2 - Taxonomic representation \ Putative parasitic genes
	#
==========================================================================
====

	# 2.2.1 Taxonomic representation

	taxData <- list(SP_Organisms = Data_SP, P_Organisms = Data_P, Z_Organisms = Data_Z)

	taxResults <- analyse_taxonomic_representation(taxData)
	#save_to_excel(taxResults,
	#             output_file = "./03-Output/01-DEG-Analysis/annotation-results/tables/taxonomic-
representation.xlsx")

	## Pie chart

	# 2.2.2 Fish organisms

	# fish species: ----
	fish <- c('Acipenser baerii',
	      'Acipenser transmontanus',
	      'Anguilla anguilla',
	      'Anguilla japonica',
	      'Anoplopoma fimbria',
	      'Argyrosomus regius',
	      'Boreogadus saida',
	      'Brachyopsis segaliensis',
```

'Carassius auratus',
'Catostomus commersonii',
'Chaenocephalus aceratus',
'Chauliodus sloani',
'Chelon auratus',
'Chelon ramada',
'Chionodraco hamatus',
'Ctenopharyngodon idella',
'Cynoscion nebulosus',
'Cyprinus carpio',
'Danio rerio',
'Devario aequipinnatus',
'Dicentrarchus labrax',
'Diplobatis ommata',
'Dissostichus eleginoides',
'Electrophorus electricus',
'Epinephelus akaara',
'Epinephelus coioides',
'Esox lucius',
'Formosania lacustris',
'Fundulus heteroclitus',
'Gadus morhua',
'Galeus melastomus',
'Gasterosteus aculeatus',
'Gillichthys mirabilis',
'Gillichthys seta',
'Haplochromis burtoni',
'Haplochromis nubilus',
'Haplochromis xenognathus',
'Harpagifer antarcticus',
'Hemitripterus americanus',
'Heterodontus francisci',
'Heteropneustes fossilis',
'Hippocampus comes',
'Hippoglossus hippoglossus',
'Ichthyomyzon unicuspis',
'Ictalurus punctatus',
'Katsuwonus pelamis',
'Labeo rohita',
'Lepomis macrochirus',
'Leucoraja ocellata',
'Lophius americanus',
'Makaira nigricans',
'Megalobrama amblycephala',
'Meiacanthus atrodorsalis',
'Micropogonias undulatus',
'Micropterus salmoides',
'Misgurnus fossilis',
'Monopterus albus',
'Mullus surmuletus',
'Myoxocephalus octodecemspinosus',
'Notemigonus crysoleucas',
'Oncorhynchus keta',
'Oncorhynchus kisutch',
'Oncorhynchus masou',
'Oncorhynchus mykiss',

```
        'Oncorhynchus nerka',
        'Oncorhynchus tshawytscha',
        'Oplegnathus fasciatus',
        'Opsanus tau',
        'Oreochromis mossambicus',
        'Oreochromis niloticus',
        'Oryzias javanicus',
        'Oryzias latipes',
        'Oryzias luzonensis',
        'Osmerus mordax',
        'Pagrus major',
        'Paralichthys olivaceus',
        'Paramisgurnus dabryanus',
        'Petromyzon marinus',
        'Pleuronectes platessa',
        'Poecilia reticulata',
        'Poeciliopsis lucida',
        'Pomatoschistus minutus',
        'Prionace glauca',
        'Protopterus aethiopicus',
        'Psalidodon fasciatus',
        'Pseudopleuronectes americanus',
        'Salmo salar',
        'Salmo trutta',
        'Salvelinus fontinalis',
        'Scomber japonicus',
        'Scomber scombrus',
        'Scomberomorus niphonius',
        'Scorpaena plumieri',
        'Scyliorhinus stellaris',
        'Seriola quinqueradiata',
        'Siganus canaliculatus',
        'Siniperca chuatsi',
        'Solea senegalensis',
        'Sparus aurata',
        'Squalus acanthias',
        'Synanceia horrida',
        'Synanceia verrucosa',
        'Takifugu pardalis',
        'Takifugu rubripes',
        'Tetraodon fluviatilis',
        'Tetraodon miurus',
        'Tetraodon nigroviridis',
        'Tetronarce californica',
        'Thalassophryne nattereri',
        'Thunnus obesus',
        'Torpedo marmorata',
        'Trachurus japonicus',
        'Trematomus bernacchii',
        'Trichopodus trichopterus',
        'Xiphophorus hellerii',
        'Xiphophorus maculatus')

# -------------------------

fish_SP <- Data_SP %>% filter(OS %in% fish)
```

```
fish_occurrance <- fish_SP %>%
  count(OS, name = "Count", sort = TRUE) %>%
  as.data.frame()

View(fish_occurrance)

dim(fish_SP) # 2420
View(fish_SP)

fishGeneNameID <- fish_SP$GeneNameID
fishGeneID <- fish_SP$GeneID
length(fishGeneID)

common_GeneNameID_fishSP_P <- get_common_GeneNameID(fish_SP, Data_P)
nrow(common_GeneNameID_fishSP_P) # 9
View(common_GeneNameID_fishSP_P)

common_GeneID_fishSP_P <- get_unique_GeneID(common_GeneNameID_fishSP_P)
length(common_GeneID_fishSP_P) # 17

fish_SP <- fish_SP[!fish_SP$GeneID %in% common_GeneID_fishSP_P, ]
dim(fish_SP) # 2410




# 2.2.3 Putative parasitic genes

platGenes <- identify_putative_plat_genes(Data_SP, Data_P)
View(platGenes)

platGeneNameID <- platGenes$GeneNameID

platGeneID <- unique(c(platGenes$GeneID.Ref, platGenes$GeneID.Plat))

save_to_excel(platGeneNameID,      "./03-Output/01-DEG-Analysis/analysis-ready-data/puta-
tive-plat-genes.xlsx")

#
==========================================================================
===
# TASK 2.3 - Principal Component Analysis (PCA)
#
==========================================================================
====
# PCA to be done:
# - with all abundance/counts (2)
# - abundance/counts subsets:
#      raw ----------------------
#      - platGeneID (2)
#      - swiss-prot fishGeneID (2)
#      - all swiss-prot (2)
#
#      deg (under- over-) --------
```

```
        #       with filterByExpr ---------
        #       - all zebrafish (2)
        #       - all swiss-prot (2)
        #
        #       without filterByExpr ------
        #       - all zebrafish (2)
        #       - all swiss-prot (2)
        #
        #       dt (under- over- noSig)----
        #       with filterByExpr ---------
        #       - all zebrafish (2)
        #       - all swiss-prot (2)
        #
        #       without filterByExpr ------
        #       - all zebrafish (2)
        #       - all swiss-prot (2)
        # ------------------------------------------------------------------------

        # Data needed:
        load("03-Output/01-DEG-Analysis/analysis-ready-data/abundance_tsv")
        View(abundance_tsv)

        # PCA data (subsets of counts \or abundance)
        # tsv_datatype
        # counts :
        counts <- abundance_tsv$counts
        # abundance :
        abundance <- abundance_tsv$abundance

        # 2.3.1 All count
        PCAdata <- process_PCAdata(tsv_data = counts)
        dim(PCAdata) # 266105
        pca_result <- runPCA(PCAdata)
        plot_pca(pca_result, save_plot = TRUE,
                output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-all-
count.pdf")

        #pca_res <- process_PCAloadings(pca_result, Data_SP)
        #plot_pcaloadings(pca_res, save_plot = FALSE)


        # 2.3.2 All abundance
        PCAdata <- process_PCAdata(tsv_data = abundance)
        dim(PCAdata) # 266105
        pca_result <- runPCA(PCAdata)
        plot_pca(pca_result, save_plot = TRUE,
                output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-all-
abundance.pdf")

        # 2.3.3 Putative plat genes | counts
        PCAdata <- process_PCAdata(GeneID_vector = platGeneID, tsv_data = counts)
        dim(PCAdata) # 10
        pca_result <- runPCA(PCAdata)
        plot_pca(pca_result, save_plot = TRUE,
                output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-puta-
tive-plat-genes-count.pdf")
```

```
# 2.3.4 Putative plat genes | abundance
PCAdata <- process_PCAdata(GeneID_vector = platGeneID, tsv_data = abundance)
dim(PCAdata) # 10
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-putative-plat-genes-abundance.pdf")

# 2.3.5 Fish genes | counts
PCAdata <- process_PCAdata(df1 = fish_SP, df2 = Data_Z, tsv_data = counts)
dim(PCAdata) # 2254
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-fish-genes-count.pdf")

# 2.3.6 Fish genes | abundance
PCAdata <- process_PCAdata(df1 = fish_SP, df2 = Data_Z, tsv_data = abundance)
dim(PCAdata) # 2254
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-fish-genes-abundance.pdf")

#
# TASK 2.4 - Remove putative plat genes
#

dim(Data_SP) # 15947
Data_SP <- Data_SP[!Data_SP$GeneID %in% platGeneID, ]
dim(Data_SP) # 15937
save_to_excel(Data_SP, "./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-SP[v2].xlsx")

dim(Data_Z) # 2975
Data_Z <- Data_Z[!Data_Z$GeneID %in% platGeneID, ]
dim(Data_Z) # 2974
save_to_excel(Data_Z, "./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-Z[v2].xlsx")

#
# TASK 3 - DEG analysis
#
# Perform DEG analysis for swiss-prot and zebrafish datasets (with) and (without)
# filterByExpr filter

# Data needed:
```

```
Data_SP <- read.xlsx("./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-SP[v2].xlsx")
dim(Data_SP) # 15 937

Data_Z <- read.xlsx("./03-Output/01-DEG-Analysis/analysis-ready-data/BpK-data-processed-Z[v2].xlsx")
dim(Data_Z) # 2 974

load("03-Output/01-DEG-Analysis/analysis-ready-data/abundance_tsv")
# counts :
counts <- abundance_tsv$counts
# abundance :
abundance <- abundance_tsv$abundance

#
================================================================================
====
# TASK 3.1 - swiss-prot (with/without filterByExpr)
#
================================================================================
====

# 3.1.1 with filterByExpr ----

deg_SP_with_filter <-run_and_save_DEG(Data_SP,
                                      "SP",
                                      model_type = "QLF" ,
                                      use_filter = TRUE,
                                      use_lfc = FALSE)

# Available results and contrasts
names(deg_SP_with_filter)
names(deg_SP_with_filter$DEG_result)
names(deg_SP_with_filter$decideTest_result)

# Example: see specific contrast and columns
deg_SP <- deg_SP_with_filter$DEG_result
Sp.RAvsTR <- deg_SP[["Sp.RAvsTR"]]
View(Sp.RAvsTR)
colnames(Sp.RAvsTR)
View(Sp.RAvsTR[, c("logFC", "FDR", "evalue", "pident")])
Sp.RAvsTR <- Sp.RAvsTR %>%
  # from lowest FDR to highest
  arrange(FDR)
View(Sp.RAvsTR)

# Get top 10
top10_SP_with_filter <- get_top10_degs(deg_SP, use_filter = TRUE, "SP")

# Get UniProt information
TaxaInfo <- get_uniprot_taxainfo(deg_SP)
functionInfo <- get_uniprot_proteinFunction(deg_SP)
Goterms<- get_uniprot_GOinfo(deg_SP)
```

```
save_to_excel(TaxaInfo, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/TaxaInfo-SP.xlsx")
save_to_excel(functionInfo , "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/functionInfo-SP.xlsx")
save_to_excel(Goterms, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/Goterms-SP.xlsx")

functionInfo <- read_excel("./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/functionInfo-SP.xlsx")

keyword_filtered <- datasubset_keyword_filtered(
  deg_list = functionInfo,
  keyword_pattern = "heat|stress|thermal|oxygen|salinity|HSP|hypoxia|adaptation|oxidative stress|acid base balance",
  column_name = "Function..CC."
)

View(keyword_filtered)

save_to_excel(keyword_filtered, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/keyword-filtered-SP.xlsx")

# Read and analyse keyword filtered
keyword_filtered <- read_excel("./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/keyword-filtered-SP.xlsx")

# analysis example:

# spring
table(keyword_filtered$Sp.TRvsRF$keyword)

hsp_filtered <- keyword_filtered$Sp.TRvsRF %>%
  filter(keyword == "HSP") %>%
  View()

heat_filtered <- keyword_filtered$Sp.TRvsRF %>%
  filter(keyword == "heat") %>%
  View()

oxidativeStress_filtered <- keyword_filtered$Sp.TRvsRF %>%
  filter(keyword == "oxidative stress") %>%
  View()
Stress_filtered <- keyword_filtered$Su.RAvsRF %>%
  filter(keyword == "stress") %>%
  View()
Stress_filtered <- keyword_filtered$Su.TRvsRF %>%
  filter(keyword == "stress") %>%
  View()
table(keyword_filtered$Sp.RAvsRF$keyword)
table(keyword_filtered$Sp.RAvsTR$keyword)

# summer
table(keyword_filtered$Su.TRvsRF$keyword)
table(keyword_filtered$Su.RAvsRF$keyword)
table(keyword_filtered$Su.RAvsTR$keyword)
```

```
# Ria de Aveiro
table(keyword_filtered$RA.SpvsSu$keyword)

hsp_filtered <- keyword_filtered$Su.RAvsTR%>%
  filter(keyword == "HSP")
View(hsp_filtered)
hsp_filtered$GeneNameID

# common degs between seasons

# spring vs summer
dim(deg_SP[["Sp.TRvsRF"]]) # 1012

dim(deg_SP[["Su.TRvsRF"]]) # 192

common <- merge(deg_SP[["Sp.TRvsRF"]], deg_SP[["Su.TRvsRF"]], by = "Accession")
dim(common) # 110
overSp <- common %>% filter(logFC.x > 0)
dim(overSp)
overSu <- common %>% filter(logFC.y > 0)
dim(overSu)

#-------------------------------------------------------------------------------

# 3.1.1 without filterByExpr ----

deg_SP_without_filter <-run_and_save_DEG(Data_SP,
                                "SP",
                                model_type = "QLF",
                                use_filter = FALSE,
                                use_lfc = FALSE)

# Get top 10 + summary
top10_SP_without_filter <- get_top10_degs(deg_SP_without_filter$DEG_result,
                                 use_filter = FALSE, "SP")




# 3.1.2 continuation .... Principal Component Analysis

# 3.1.2.1 degs | with filterByExpr | counts
deg_SP <- deg_SP_with_filter$DEG_result

PCAdata <- process_PCAdata(data = deg_SP, tsv_data = counts)
dim(PCAdata) # 1625
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
      output_file   =   "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-
with-filterByExpr-count-SP.pdf")

# 3.1.2.2 degs | with filterByExpr | abundance
PCAdata <- process_PCAdata(data = deg_SP, tsv_data = abundance)
dim(PCAdata) # 1625
```

```
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-with-filterByExpr-abundance-SP.pdf")

# 3.1.2.3 degs | without filterByExpr | counts
deg_SP <- deg_SP_without_filter$DEG_result

PCAdata <- process_PCAdata(data = deg_SP, tsv_data = counts)
dim(PCAdata) # 1298
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-without-filterByExpr-count-SP.pdf")

# 3.1.2.4 degs | without filterByExpr | abundance
PCAdata <- process_PCAdata(data = deg_SP, tsv_data = abundance)
dim(PCAdata) # 1298
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-without-filterByExpr-abundance-SP.pdf")

# 3.1.2.5 dt | with filterByExpr | counts
dt_SP <- deg_SP_with_filter$decideTest_result

PCAdata <- process_PCAdata(data = dt_SP, tsv_data = counts)
dim(PCAdata) # 9645
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-filterByExpr-count-SP.pdf")

# 3.1.2.6 dt | with filterByExpr | abundance
PCAdata <- process_PCAdata(data = dt_SP, tsv_data = abundance)
dim(PCAdata) # 9645
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-filterByExpr-abundance-SP.pdf")

# 3.1.2.7 dt | without filterByExpr | counts
dt_SP <- deg_SP_without_filter$decideTest_result

PCAdata <- process_PCAdata(data = dt_SP, tsv_data = counts)
dim(PCAdata) # 15937
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-without-filterByExpr-count-SP.pdf")


# 3.1.2.8 dt | without filterByExpr | abundance
PCAdata <- process_PCAdata(data = dt_SP, tsv_data = abundance)
dim(PCAdata) # 15937
pca_result <- runPCA(PCAdata)
```

```
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-
out-filterByExpr-abundance-SP.pdf")

# 3.1.3.1 with filterByExpr

dt_SP <- deg_SP_with_filter$decideTest_result

View(deg_SP_with_filter$DEG_result)
names(deg_SP_with_filter$decideTest_result)

res <- process_volcano_data(dt_SP[["Sp.RAvsTR"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Spring-RAvsTR-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["Sp.RAvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Spring-RAvsRF-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["Sp.TRvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Spring-TRvsRF-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["Su.RAvsTR"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Summer-RAvsTR-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["Su.RAvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Summer-RAvsRF-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["Su.TRvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Summer-TRvsRF-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["RA.SpvsSu"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-RiaAveiro-SpvsSu-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["TR.SpvsSu"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-Troia-SpvsSu-with-filterByExpr-SP.pdf")

res <- process_volcano_data(dt_SP[["RF.SpvsSu"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
        output_file = "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/fig-
ures/volcanoPlot-RiaFormosa-SpvsSu-with-filterByExpr-SP.pdf")
```

```
        # 3.1.3.2 without filterByExpr

        View(deg_SP_without_filter$DEG_result)
        names(deg_SP_without_filter$decideTest_result)

        dt_SP <- deg_SP_without_filter$decideTest_result

        res <- process_volcano_data(dt_SP[["Sp.RAvsTR"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Spring-RAvsTR-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["Sp.RAvsRF"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Spring-RAvsRF-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["Sp.TRvsRF"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Spring-TRvsRF-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["Su.RAvsTR"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Summer-RAvsTR-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["Su.RAvsRF"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Summer-RAvsRF-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["Su.TRvsRF"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Summer-TRvsRF-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["RA.SpvsSu"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-RiaAveiro-SpvsSu-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["TR.SpvsSu"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-Troia-SpvsSu-without-filterByExpr-SP.pdf")

        res <- process_volcano_data(dt_SP[["RF.SpvsSu"]])
        plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                         output_file = "./03-Output/01-DEG-Analysis/DEG-results/without-filterByExpr/figures/volcanoPlot-RiaFormosa-SpvsSu-without-filterByExpr-SP.pdf")

        #
================================================================================
====
        # TASK 3.2 - Zebrafish (with/without filterByExpr)
```

```
#
================================================================================
====

# 3.2.1 with filterByExpr ----------------------------------------------------

deg_Z_with_filter <-run_and_save_DEG(Data_Z,
                                     "Z",
                                     model_type = "QLF",
                                     use_filter = TRUE,
                                     use_lfc = FALSE)

deg_Z <- deg_Z_with_filter$DEG_result

# Get top 10 + summary
top10_Z_with_filter <- get_top10_degs(deg_Z_with_filter$DEG_result,
                                      use_filter = TRUE, "Z")




# 3.2.2 without filterByExpr -------------------------------------------------

deg_Z_without_filter<-run_and_save_DEG(Data_Z,
                                       "Z",
                                       model_type = "QLF",
                                       use_filter = FALSE,
                                       use_lfc = FALSE)

# Get top 10 + summary
top10_Z_without_filter <- get_top10_degs(deg_Z_without_filter$DEG_result,
                                         use_filter = FALSE, "Z")

# 3.2.2 continuation .... Principal Component Analysis

# 3.2.2.1 degs | with filterByExpr | counts
deg_Z <- deg_Z_with_filter$DEG_result

PCAdata <- process_PCAdata(data = deg_Z, tsv_data = counts)
dim(PCAdata) # 407
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
      output_file   =   "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-
with-filterByExpr-count-Z.pdf")

# 3.2.2.2 degs | with filterByExpr | abundance
PCAdata <- process_PCAdata(data = deg_Z, tsv_data = abundance)
dim(PCAdata) # 407
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
      output_file   =   "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-
with-filterByExpr-abundance-Z.pdf")

# 3.2.2.3 degs | without filterByExpr | counts
deg_Z <- deg_Z_without_filter$DEG_result
```

```
PCAdata <- process_PCAdata(data = deg_Z, tsv_data = counts)
dim(PCAdata) # 404
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-without-filterByExpr-count-Z.pdf")

# 3.2.2.4 degs | without filterByExpr | abundance
PCAdata <- process_PCAdata(data = deg_Z, tsv_data = abundance)
dim(PCAdata) # 404
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-degs-without-filterByExpr-abundance-Z.pdf")

# 3.2.2.5 dt | with filterByExpr | counts
dt_Z <- deg_Z_with_filter$decideTest_result

PCAdata <- process_PCAdata(data = dt_Z, tsv_data = counts)
dim(PCAdata) # 2473
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-filterByExpr-count-Z.pdf")

# 3.2.2.6 dt | with filterByExpr | abundance
PCAdata <- process_PCAdata(data = dt_Z, tsv_data = abundance)
dim(PCAdata) # 2473
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-filterByExpr-abundace-Z.pdf")

# 3.2.2.7 dt | without filterByExpr | counts
dt_Z <- deg_Z_without_filter$decideTest_result

PCAdata <- process_PCAdata(data = dt_Z, tsv_data = counts)
dim(PCAdata) # 2974
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-out-filterByExpr-count-Z.pdf")

# 3.2.2.8 dt | without filterByExpr | abundance
PCAdata <- process_PCAdata(data = dt_Z, tsv_data = abundance)
dim(PCAdata) # 2974
pca_result <- runPCA(PCAdata)
plot_pca(pca_result, save_plot = TRUE,
    output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/pcaPlot-dt-with-out-filterByExpr-abundance-Z.pdf")

# 3.1.3 volcano plots

# 3.1.3.1 with filterByExpr
```

```
View(deg_Z_with_filter$DEG_result)
names(deg_Z_with_filter$decideTest_result)

dt_Z <- deg_Z_with_filter$decideTest_result

res <- process_volcano_data(dt_Z[["Sp.RAvsTR"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-RAvsTR-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["Sp.RAvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-RAvsRF-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["Sp.TRvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-TRvsRF-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["Su.RAvsTR"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-RAvsTR-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["Su.RAvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-RAvsRF-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["Su.TRvsRF"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-TRvsRF-with-filterByExpr-Z.pdf")

res <- process_volcano_data(dt_Z[["RA.SpvsSu"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-RiaAveiro-SpvsSu-with-filterByExpr-Z.pdf")

#res <- process_volcano_data(dt_Z[["TR.SpvsSu"]])
#plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE)

res <- process_volcano_data(dt_Z[["RF.SpvsSu"]])
plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
                 output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-RiaFormosa-SpvsSu-with-filterByExpr-Z.pdf")

# 3.1.3.2 without filterByExpr

View(deg_Z_without_filter$DEG_result)
names(deg_Z_without_filter$decideTest_result)

dt_Z <- deg_Z_without_filter$decideTest_result

res <- process_volcano_data(dt_Z[["Sp.RAvsTR"]])
```

```
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-RAvsTR-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["Sp.RAvsRF"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-RAvsRF-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["Sp.TRvsRF"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Spring-TRvsRF-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["Su.RAvsTR"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-RAvsTR-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["Su.RAvsRF"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-RAvsRF-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["Su.TRvsRF"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-Summer-TRvsRF-without-filterByExpr-Z.pdf")

	res <- process_volcano_data(dt_Z[["RA.SpvsSu"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-RiaAveiro-SpvsSu-without-filterByExpr-Z.pdf")

	#res <- process_volcano_data(dt_Z[["TR.SpvsSu"]])
	#plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE)

	res <- process_volcano_data(dt_Z[["RF.SpvsSu"]])
	plot_volcanoPlot(res$volcano_data, res$highlighted, save_plot = TRUE,
		output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/volcano-Plot-RiaFormosa-SpvsSu-without-filterByExpr-Z.pdf")

	#
================================================================================
====
	# TASK  - Swiss-prot vs. Zebrafish analysis
	#
================================================================================
====

	SPvsZ <- common_deg_SP_Z(deg_SP, deg_Z)

	save_to_excel(SPvsZ$common, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/degs-common[v2].xlsx")
	save_to_excel(SPvsZ$unique_SP, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/degs-unique-SP[v2].xlsx")
```

```
save_to_excel(SPvsZ$unique_Z, "./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/degs-unique-Z[v2].xlsx")

SPvsZ$common$Sp.RAvsTR
SPvsZ$common$Sp.RAvsRF
SPvsZ$common$Sp.TRvsRF
SPvsZ$common$Su.RAvsTR
SPvsZ$common$Su.RAvsRF
SPvsZ$common$Su.TRvsRF
SPvsZ$common$RF.SpvsSu
SPvsZ$common$RA.SpvsSu

SPvsZ$unique_SP$Sp.RAvsTR
SPvsZ$unique_SP$Sp.RAvsRF
SPvsZ$unique_SP$Sp.TRvsRF
SPvsZ$unique_SP$Su.RAvsTR
SPvsZ$unique_SP$Su.RAvsRF
SPvsZ$unique_SP$Su.TRvsRF
SPvsZ$unique_SP$RF.SpvsSu
SPvsZ$unique_SP$RA.SpvsSu

SPvsZ$unique_Z$Sp.RAvsTR
SPvsZ$unique_Z$Sp.RAvsRF
SPvsZ$unique_Z$Sp.TRvsRF
SPvsZ$unique_Z$Su.RAvsTR
SPvsZ$unique_Z$Su.RAvsRF
SPvsZ$unique_Z$Su.TRvsRF
SPvsZ$unique_Z$RF.SpvsSu
SPvsZ$unique_Z$RA.SpvsSu

# bar plot's

#barData_common <- process_barplot_data(SPvsZ$common$Sp.TRvsRF, "common", metric = "pident") # It doesn't work cus pident == pident.SP and pident.Z
barData_unique_SP <- process_barplot_data(SPvsZ$unique_SP$Sp.TRvsRF, "swissprot", metric = "pident")
barData_unique_Z <- process_barplot_data(SPvsZ$unique_Z$Sp.TRvsRF, "zebrafish", metric = "pident")

data_list <- list("unique_SP" = barData_unique_SP,
                  "unique_Z" = barData_unique_Z)

barData <- prepare_summary(data_list)

## Bar plot
plot_barplot(barData, save_plot = FALSE)

plot_barplot(barData, save_plot = TRUE,
             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-Sp.TRvsRF-pident-uniqueSP-uniqueZ.pdf" )

#barData_common <- process_barplot_data(SPvsZ$common$Sp.TRvsRF, "common", metric = "evalue")
barData_unique_SP <- process_barplot_data(SPvsZ$unique_SP$Sp.TRvsRF, "swissprot", metric = "evalue")
```

```
barData_unique_Z <- process_barplot_data(SPvsZ$unique_Z$Sp.TRvsRF, "zebrafish", metric =
"evalue")

data_list <- list("unique_SP" = barData_unique_SP,
                  "unique_Z" = barData_unique_Z)

barData <- prepare_summary(data_list)

## Bar plot
plot_barplot(barData, save_plot = FALSE)

plot_barplot(barData, save_plot = TRUE,
             output_file = "./03-Output/01-DEG-Analysis/annotation-results/figures/barPlot-
Sp.TRvsRF-evalue-uniqueSP-uniqueZ.pdf" )

#
==========================================================================
====
# TASK  - Gene set enrichment analysis (GSEA)
#
==========================================================================
====

## Data (with filterByExpr):
gseaobj <- read_excel("./03-Output/01-DEG-Analysis/DEG-results/with-filterByExpr/decide-
Test-Z.xlsx")
View(gseaobj)

# ----- to use multiGSEA package, we need a df with "Accession", "logFC", "PValue" columns ----
-
GSEA_data <- process_GSEAdata(gseaobj)
save_to_excel(GSEA_data, "./03-Output/02-Pathway-Enrichment/analysis-ready-data/with-
filterByExpr/GSEA-data.xlsx")
pathways <- perform_GSEA(GSEA_data)

View(pathways$ES_processed_result$Sp.TRvsRF)

save_to_excel(pathways$ES_processed_result, "./03-Output/02-Pathway-Enrichment/GSEA-
results/with-filterByExpr/pathways.xlsx")
#save_to_excel(pathways$ES_processed_result, "./03-Output/02-Pathway-Enrichment/GSEA-
results/without-filterByExpr/pathways.xlsx")

# Plot top10
RA_SpvsSu <- process_GSEA_plotdata(pathways$top_10_result$RA.SpvsSu, "RA.SpvsSu")
Sp_TRvsRF <- process_GSEA_plotdata(pathways$top_10_result$Sp.TRvsRF, "Sp.TRvsRF")

data_to_plot <- list(c(RA_SpvsSu,Sp_TRvsRF))
data_to_plot <- bind_rows(RA_SpvsSu, Sp_TRvsRF)
View(data_to_plot)

plot_dotPlot(data_to_plot,save_plot = TRUE, output_file = "./03-Output/02-Pathway-Enrich-
ment/GSEA-results/with-filterByExpr/dotplot-Ra-SpvsSu-Sp-TRvsRF.pdf" )

#
==========================================================================
====
```

```
        # TASK  - STRING analysis
        #
================================================================================
====

        # String (version 12.0):
        # software link:
        # https://string-db.org/cgi/input?sessionId=bAHDVX8em5Is&input_page_show_search=on

        # steps:
        # search -> Multiple proteins -> List Of Name -> Organisms -> Advance Settings
        #                                                |-> Required score
        #                                                   |-> high confidence (0.700)

        # tips:
        # List Of Names -> one-per-line, use Gene.Names..primary from UniProt, if you try to use Ac-
cessions, it doesn't work!
        #            -> do not use all the Gene.Names..primary! It only work up to 2.000 entry
        #            -> try for instance:
        #                 - only the degs (per-contrast)
        #                 - only the under-expressed (per-contrast)
        #                 - only the over-expressed (per-contrast)
        #                 - only possible specific genes of interests selected (per-contrast)
        #             -> "per contrast" means that you have to search individually for each contrast, not all
together
        #            -> you cannot use it if the organisms are all different, as in BLAST against Swiss-Prot!
it fails
        # Organisms -> use auto-detect, or the actual organism, in my case "Danio rerio"

        #
================================================================================
====

        # Data needed:

        # Running UniProt may take approximately 5min (degs) to 3hours (dt).
        # Alternatively, you can skip this step and read the saved file instead.
        # ------------------------------------------------------------------------
        # DataZ decide test without filterByExpr
        #decideTest_Z <- deg_Z_without_filter$decideTest_result
        #SpvsSu <- decideTest_Z[c("RA.SpvsSu", "TR.SpvsSu", "RF.SpvsSu" )]
        #Data_SpvsSu <- get_uniprot_taxainfo(SpvsSu)
        #Su <- decideTest_Z[c("Su.RAvsTR", "Su.RAvsRF", "Su.TRvsRF" )]
        #Data_Su <- get_uniprot_taxainfo(Su)
        #Sp <- decideTest_Z[c("Sp.RAvsTR", "Sp.RAvsRF", "Sp.TRvsRF" )]
        #Data_Sp <- get_uniprot_taxainfo(Sp)
        #uniprot_taxaobj <- c(Data_SpvsSu,Data_Su,Data_Sp)
        #View(uniprot_taxaobj)
        #save_to_excel(uniprot_taxaobj,              "./03-Output/02-Pathway-Enrichment/preprocessed-
data/without-filterByExpr/uniprot-taxaobj-dtZ.xlsx")

        # DataZ decide test with filterByExpr
        #decideTest_Z <- deg_Z_with_filter$decideTest_result
        #SpvsSu <- decideTest_Z[c("RA.SpvsSu", "TR.SpvsSu", "RF.SpvsSu" )]
        #Data_SpvsSu <- get_uniprot_taxainfo(SpvsSu)
        #Su <- decideTest_Z[c("Su.RAvsTR", "Su.RAvsRF", "Su.TRvsRF" )]
```

```
#Data_Su <- get_uniprot_taxainfo(Su)
#Sp <- decideTest_Z[c("Sp.RAvsTR", "Sp.RAvsRF", "Sp.TRvsRF" )]
#Data_Sp <- get_uniprot_taxainfo(Sp)
#uniprot_taxaobj <- c(Data_SpvsSu,Data_Su,Data_Sp)
#View(uniprot_taxaobj)
#save_to_excel(uniprot_taxaobj, "./03-Output/02-Pathway-Enrichment/preprocessed-data/with-filterByExpr/uniprot-taxaobj-dtZ.xlsx")

# DataZ degs without filterByExpr
#deg_Z <- deg_Z_without_filter$DEG_result
#SpvsSu <- deg_Z[c("RA.SpvsSu", "TR.SpvsSu", "RF.SpvsSu" )]
#Data_SpvsSu <- get_uniprot_taxainfo(SpvsSu)
#Su <- deg_Z[c("Su.RAvsTR", "Su.RAvsRF", "Su.TRvsRF" )]
#Data_Su <- get_uniprot_taxainfo(Su)
#Sp <- deg_Z[c("Sp.RAvsTR", "Sp.RAvsRF", "Sp.TRvsRF" )]
#Data_Sp <- get_uniprot_taxainfo(Sp)
#uniprot_taxaobj <- c(Data_SpvsSu,Data_Su,Data_Sp)
#View(uniprot_taxaobj)
#save_to_excel(uniprot_taxaobj, "./03-Output/02-Pathway-Enrichment/preprocessed-data/without-filterByExpr/uniprot-taxaobj-degZ.xlsx")

# DataZ degs with filterByExpr
deg_Z <- deg_Z_with_filter$DEG_result
SpvsSu <- deg_Z[c("RA.SpvsSu", "TR.SpvsSu", "RF.SpvsSu" )]
Data_SpvsSu <- get_uniprot_taxainfo(SpvsSu)
Su <- deg_Z[c("Su.RAvsTR", "Su.RAvsRF", "Su.TRvsRF" )]
Data_Su <- get_uniprot_taxainfo(Su)
Sp <- deg_Z[c("Sp.RAvsTR", "Sp.RAvsRF", "Sp.TRvsRF" )]
Data_Sp <- get_uniprot_taxainfo(Sp)
uniprot_taxaobj <- c(Data_SpvsSu,Data_Su,Data_Sp)
uniprot_taxaobj <- lapply(uniprot_taxaobj, function(df) {
  df[order(df$logFC, decreasing = TRUE), ]
})
View(uniprot_taxaobj)
save_to_excel(uniprot_taxaobj, "./03-Output/02-Pathway-Enrichment/preprocessed-data/with-filterByExpr/uniprot-taxaobj-degZ.xlsx")
# ------------------------------------------------------------------------------

# prepare data to use in string software:

## Data (with filterByExpr) decide test:
#uniprot_taxaobj_with <- read_excel("./03-Output/02-Pathway-Enrichment/preprocessed-data/with-filterByExpr/uniprot-taxaobj-dtZ.xlsx")

# ----- to use STRING, we need the column "Gene.Names..primary." -----
#STRING_data <- process_STRINGdata(uniprot_taxaobj_with)
#save_to_excel(STRING_data, "./03-Output/02-Pathway-Enrichment/analysis-ready-data/with-filterByExpr/STRING-data.xlsx")

## Data (with filterByExpr) degs:
uniprot_taxaobj_with <- read_excel("./03-Output/02-Pathway-Enrichment/preprocessed-data/with-filterByExpr/uniprot-taxaobj-degZ.xlsx")

# ----- to use STRING, we need the column "Gene.Names..primary." -----
STRING_data <- process_STRINGdata(uniprot_taxaobj)
```

save_to_excel(STRING_data, "./03-Output/02-Pathway-Enrichment/analysis-ready-data/with-filterByExpr/STRING-data.xlsx")

A.2.2 Functions Script

```
# package dependencies
## CRAN
library(showtext)
library(openxlsx)
library(stringr)
## BiocManager
library(tximport)
library(ggrepel)
library(PCAtools) # require package: ggrepel
library(limma)
library(edgeR) # require package: limma
library(UniprotR)
library(multiGSEA)
library(org.Dr.eg.db)
## tidyverse
library(tidyr)
library(dplyr)
## Plot
library(ggplot2)
library(ggVennDiagram)
library(pheatmap)
library(ggthemes)

# Session info
sessionInfo()
# Constants and Global settings
sample_name <- c( "F14.Sp_TR","F15.Su_TR","F16.Su_RA",
                  "F17.Sp_RF","F18.Sp_TR","F19.Sp_RA",
                  "F2.Su_RF" ,"F2.Su_TR" ,"F21.Sp_RA",
                  "F21.Sp_TR","F22.Sp_RF","F27.Sp_RA",
                  "F28.Su_RF","F30.Su_TR","F35.Sp_RF",
                  "F35.Su_RF","F38.Su_RA","F4.Su_RA" )

#
================================================================================
====
# TASK 1
#
================================================================================
====

load_blastp_data <- function(filename){
  # Read data (table format)
  data <- read.delim(filename, sep = ";", header = FALSE)
  # Assign column names
  colnames(data) <- c(
    "qseqid", "sseqid","pident","length","mismatch",
    "gapopen", "qstart", "qend", "sstart", "send",
    "evalue", "bitscore","sacc","qcovs","stitle")
  return(data)}
```

```
Add_GeneID_column <- function(blast_data) {
  # Debugging:
  # Example of correct qseqid: TRINITY_DN0_c10_g1_i1
  not_trinity <- blast_data$qseqid[!grepl("TRINITY_D", blast_data$qseqid)]
  if (length(not_trinity) > 0) {
    #cat("qseqid sem 'TRINITY_D':\n")
    #print(not_trinity)
  } else {
    #cat("Todos os qseqid têm 'TRINITY_D'.\n")
  }
  blast_data <- blast_data[grepl("TRINITY_D", blast_data$qseqid), ]
  blast_data$GeneID <- sapply(
    blast_data$qseqid,
    function(x) strsplit(x, "i", fixed = TRUE)[[1]][1]
  )
  return(blast_data)}


select_best_hits <- function(blast_data){
  # version 1
  # Add GeneID column
  blast_data <- Add_GeneID_column(blast_data)
  # Debugging step
  # Example: TRINITY_DN0_c10_g1_i1
  #example <- blast_data %>%
  #  filter(GeneID == "TRINITY_DN100212_c0_g1_") %>%
  #  select(c("qseqid","GeneID","evalue","pident"))
  #View(example)
  # Order: lowest evalue and highest pident
  blast_data <- blast_data[order(blast_data$evalue, -blast_data$pident),]
  # Best hits (first row)
  best_hits <- blast_data[match(unique(blast_data$GeneID), blast_data$GeneID), ]
  # Debugging step
  # Example: TRINITY_DN0_c10_g1_i1
  #example_best_hit <- best_hits %>%
  #  filter(GeneID == "TRINITY_DN100212_c0_g1_") %>%
  #  select(c("qseqid","GeneID","evalue","pident"))
  #View(example_best_hit)
  return(best_hits)
}

save_to_excel <- function(data, output_file) {

  work_book <- createWorkbook()
  if (is.data.frame(data)) {
    sheet_name <- deparse(substitute(data))  # Get the variable name
    addWorksheet(work_book, sheet_name)
    writeData(work_book, sheet_name, data)
  } else if (is.list(data) && all(sapply(data, is.data.frame))) {
    for (name in names(data)) {
      df <- data[[name]]
      if (nrow(df) > 0) {
        addWorksheet(work_book, name)
        writeData(work_book, name, df)
      }
```

```
  }
 }
 saveWorkbook(work_book, file = output_file, overwrite = TRUE)
 cat("Excel created successfully:", output_file, "\n")}

read_excel <- function(path){
 # getting data from sheets
 sheets <- openxlsx::getSheetNames(path)
 dataframe <- lapply(sheets, openxlsx::read.xlsx, xlsxFile=path)
 # assigning names to dataframe
 names(dataframe) <- sheets
 return(dataframe)}

process_BpK_data <- function(BpK_data){
 # Test
 #BpK_data <- BpK_swissprot
 #dim(BpK_data)
 BpK_data <- BpK_data %>%
  mutate(GeneNameID = str_extract(sseqid, "(?<=\\|)[^|]+(?=_)"),
      # sp|P79762|ZP3_CHICK -> ZP3
      Accession = str_extract(sacc, "(?<=\\|)[^\\|]+(?=\\|)"),
      # sp|P79762|ZP3_CHICK -> P79762
      OS = str_extract(stitle, "(?<=OS=)[^ ]+ [^ ]+")
      # sp|Q90508|VIT1_FUNHE Vitellogenin-1 OS=Fundulus heteroclitus OX=8078 GN=vtg1
PE=1 SV=2 -> Fundulus heteroclitus
      )
 #table(BpK_data$GeneNameID)
 # Debugging step
 # Example: ABCA1 (SwissProt)
 #example <- BpK_data %>%
 #  filter(GeneNameID== "ABCA1") %>%
 #  select(c("GeneNameID","evalue","pident"))
 #View(example)
 # Order: lowest evalue and highest pident
 BpK_data <- BpK_data[order(BpK_data$evalue, -BpK_data$pident), ]
 # Select the best hit (1st row)
 BpK_data <- BpK_data[match(unique(BpK_data$GeneNameID), BpK_data$GeneNameID), ]
 #dim(BpK_data)

 # Debugging step
 # Example: ABCA1  (SwissProt)
 #example1 <- BpK_data %>%
 #  filter(GeneNameID== "ABCA1") %>%
 #  select(c("GeneNameID","evalue","pident"))
 #View(example1)
 #sum(is.na(BpK_data)) # 3 NA
 #colSums(is.na(BpK_data)) # Column "OS"
 #View(BpK_data[is.na(BpK_data$OS), ])
 # the extraction fail because stitle was incomplete (e.g., sp|Q3SYR3|ABEC2_BOVIN Probable
C- )
 # how to fix? (solution specific to my data results/problems)
 # BpK_swissprot -> missing: HUMAN (Homo sapiens) | BOVIN (Bos taurus)  | MOUSE (Mus
musculus)
 # BpK_zebrafish -> missing: DANRE (Danio rerio)
 # BpK_plat -> all ok!
 if (anyNA(BpK_data)) {
```

```
  BpK_data <- BpK_data %>%
    # ifelse(test, yes, no)
    # test
    mutate(OS = ifelse(is.na(OS),
                       case_when(
                         grepl("HUMAN", sseqid) ~ "Homo sapiens",
                         grepl("BOVIN", sseqid) ~ "Bos taurus",
                         grepl("MOUSE", sseqid) ~ "Mus musculus",
                         grepl("DANRE", sseqid) ~ "Danio rerio"),
  }
  #sum(is.na(BpK_data))
  #head(BpK_data$OS, 20)
  # confirm if it is well done:
  example3 <- BpK_data %>%
    filter(Accession %in% c("Q99J72","P14060", "Q3SYR3")) %>%
    select(OS, stitle)
  #View(example3)
  return(BpK_data)
}


evaluate_sequence_quality <- function(value, metric) {
  # By annotation
  if (metric == "pident") {
    # pident -> percentage of identical positions
    case_when(
      value < 30 ~ "Bad",
      value >= 30 & value <= 70 ~ "Good",
      value > 70 ~ "Excellent"
    )
  } else if (metric == "evalue") {
    # evalue -> expect value
    # The lower the E value, the more significant the score and the alignment.
    case_when(
      value > 1e-10 ~ "Bad",
      value <= 1e-70 ~ "Excellent",
      TRUE ~ "Good"
    )
  # By expression
  } else if (metric == "logFC") {
    case_when(
      value < -1 ~ "Underexpressed",
      value > 1 ~ "Overexpressed",
      TRUE ~ "No Significant"
    )
  } else if (metric == "FDR") {
    case_when(
      value < 0.01 ~ "Excellent",
      value <= 0.05 ~ "Good",
      TRUE ~ "Bad"
    )
  }
}
format_table <- function(table) {
  paste(names(table), table, sep=": ", collapse=" | ")
}
```

```
identify_putative_plat_genes <- function(reference_data, plat_data){
 # Test
 #reference_data <- Data_SP
 #plat_data <- Data_P
 matched_GeneNameID <- merge(reference_data, plat_data, by = "GeneNameID",
                      suffixes = c(".Ref", ".Plat"))
                      # R -> reference
                      # P -> plat
 #View(matched_GeneNameID)
 if(nrow(matched_GeneNameID)>0){
  matched_GeneNameID_processed <- matched_GeneNameID %>%
   mutate(
    pident_category.Ref = evaluate_sequence_quality(pident.Ref, "pident"),
    evalue_category.Ref = evaluate_sequence_quality(evalue.Ref, "evalue"),
    pident_category.Plat = evaluate_sequence_quality(pident.Plat, "pident"),
    evalue_category.Plat = evaluate_sequence_quality(evalue.Plat, "evalue")
   ) %>%
   select(c(GeneNameID,
         pident.Ref,
         pident_category.Ref,
         evalue.Ref,
         evalue_category.Ref,
         pident.Plat,
         pident_category.Plat,
         evalue.Plat,
         evalue_category.Plat,
         OS.Ref,OS.Plat,
         GeneID.Ref, GeneID.Plat))


  #View(matched_GeneNameID_processed)
  putative_plat_data <- matched_GeneNameID_processed %>%
   #filter(
   #  evalue.Plat < evalue.Ref, #só com e-value -> 36
   #  OS.Ref == OS.Plat #evalue e mesmo OS -> 32
   #)
   # New strategy (Sat May  3 18:51:55 2025)
   filter(
    pident_category.Ref == "Excellent",
    evalue_category.Ref == "Excellent",
    pident_category.Plat == "Excellent",
    evalue_category.Plat == "Excellent",

    OS.Ref == OS.Plat
   )
  #View(putative_plat_data[,  c("evalue.Plat",  "evalue.Ref","OS.Ref","OS.Plat")])  #  ->  32
GeneNameID
  #putative_plat_data%>% count(OS.Ref)
  # ------------------------------------------------------------------
  # Print summary ----------------------------------------------------
  summary <- paste(
   "--- SUMMARY ---",
   paste("Total GeneNameID in reference data:", nrow(reference_data)),
   paste("Total GeneNameID in plat data:", nrow(plat_data)),
   paste("GeneNameID in common:", nrow(matched_GeneNameID)),
```

```
            "",
            "--- Categories ---",
            paste("pident (Ref):", format_table(table(matched_GeneNameID_processed$pident_cate-
gory.Ref))),
            paste("evalue (Ref):", format_table(table(matched_GeneNameID_processed$evalue_cate-
gory.Ref))),
            paste("pident (Plat):", format_table(table(matched_GeneNameID_processed$pident_cate-
gory.Plat))),
            paste("evalue (Plat):", format_table(table(matched_GeneNameID_processed$evalue_cate-
gory.Plat))),
            "",
            "--- OS in common (examples) ---",
            paste("Common GeneNameID with same OS:",
                sum(matched_GeneNameID$OS.Ref == matched_GeneNameID$OS.Plat)),
            paste("Most common organism in OS.Ref:",
                names(which.max(table(matched_GeneNameID$OS.Ref)))),
            paste("Most common organism in OS.Plat:",
                names(which.max(table(matched_GeneNameID$OS.Plat)))),
            "",
            "--- Putative parasite genes ---",
            "",
            "criteria:",
            "e-value == Excellent (evalue <= 1e-70)",
            "pident == Excellent (pident >= 70)",
            "Equal organism",
            "",
            paste("We identify a total of", nrow(putative_plat_data), "putative parasite genes"),
            paste(putative_plat_data$GeneNameID, collapse = ", "),
            "",
            paste("pident (Ref): Min:", min(putative_plat_data$pident.Ref), " | Max:", max(puta-
tive_plat_data$pident.Ref)),
            paste("evalue (Ref): Min:", min(putative_plat_data$evalue.Ref), " | Max:", max(puta-
tive_plat_data$evalue.Ref)),
            paste("pident (Plat): Min:", min(putative_plat_data$pident.Plat), " | Max:", max(puta-
tive_plat_data$pident.Plat)),
            paste("evalue (Plat): Min:", min(putative_plat_data$evalue.Plat), " | Max:", max(puta-
tive_plat_data$evalue.Plat)),
            "---------------",
            "",
            sep = "\n"
          )
          # Print summary
          cat(summary, sep = "\n")
          # ------------------------------------------------------------------------
          # ------------------------------------------------------------------------
        } else {
          cat("Coun't identify putative platyhelminthes genes")
        }
        return(putative_plat_data)}

      analyse_overlap <- function(data1, data2, name1, name2, column_name = "GeneID"){
        #test
        #data1 <- BpK_swissprot
        #data2 <- BpK_zebrafish
        #name1 <- "swissprot"
        #name2 <- "zebrafish"
```

```
  #column_name <- "GeneID"
  dim(data1) # 37824
  value1 <- unique(data1[[column_name]])
  num1 <- length(value1) # 37824
  num1
  head(value1)
  dim(data2) # 17932
  value2 <- unique(data2[[column_name]])
  num2 <- length(value2) # 17932
  num2
  # find common values (e.g., GenesID | GeneNameID)
  common <- intersect(value1, value2)
  num_common <- length(common)
  num_common
  percentage_in_1 <- 100*(num_common/num1)
  summary <- paste(
    "--- SUMMARY ---",
    paste("Overlap analysis:", name1, "vs", name2),
    paste(name1, ":", num1),
    paste(name2, ":", num2),
    paste("Common:", num_common),
    paste("% of", name1, "in", name2, ":", round(percentage_in_1,2)),
    paste("% of", name2, "in", name1, ":", round(percentage_in_2,2)),
    sep = "\n"
  )
  # Print summary
  cat(summary)
  # Data frames for common and unique values
  # common
  common_df <- merge(data1, data2, by = column_name, suffixes = c(".1", ".2"))
  # unique to data1
  unique_to_1 <- anti_join(data1, data2, by = column_name)
  #  unique to data2
  unique_to_2 <- anti_join(data2, data1, by = column_name)
  #all rows from data2 whose value in column_name does not appear in data1
  return(list(common_values = common, common_df = common_df,
              unique_to_1= unique_to_1,
              unique_to_2 = unique_to_2))}

analyse_taxonomic_representation <- function(data){
  taxa_occurrance <- list()
  if(is.list(data)){
    for (df_name in names(data)) {
      df <- data[[df_name]]
      occurrance_df <- df %>%
        count(OS, name = "Count", sort = TRUE) %>%
        as.data.frame()
      num_os <- nrow(occurrance_df)
      top_taxa <- head(occurrance_df, 5)
      total_entries <- nrow(df)
      top_taxon <- top_taxa$OS[1]
      top_taxon_count <- top_taxa$Count[1]
      top_taxon_percentage <- round(100 * top_taxon_count / total_entries, 2)
      summary <- paste(
        "--- SUMMARY ---",
        paste0("Total entries: ", total_entries),
```

```
            paste0("Number of unique taxa (OS): ", num_os),
            paste0("Most represented taxon: ", top_taxon, " (", top_taxon_count, " entries, ",
top_taxon_percentage, "%)"),
            "Top 5 taxa by representation:",
            paste0(
              apply(top_taxa, 1, function(row) {
                sprintf("  - %s: %s entries (%.2f%%)",
                        row["OS"],
                        row["Count"],
                        100 * as.numeric(row["Count"]) / total_entries)
              }),
              collapse = "\n"
            ),
            sep = "\n"
          )
          cat(summary, "\n\n")
          taxa_occurrance[[df_name]] <- occurrance_df
        }
      return(taxa_occurrance)}

    perform_DEG <-function(data, model_type, use_filter = FALSE, use_lfc =FALSE){
      data <- data
      # constants
      p_value <- 0.05
      log_fold_change <- 1.5
      adjust_method <- "fdr"
      experimental_groups <- c(
        "Sp_TR","Su_TR","Su_RA",
        "Sp_RF","Sp_TR","Sp_RA",
        "Su_RF" ,"Su_TR" ,"Sp_RA",
        "Sp_TR","Sp_RF","Sp_RA",
        "Su_RF","Su_TR","Sp_RF",
        "Su_RF","Su_RA","Su_RA")
      # data preparation --------------------------------------------------------
      dge <- DGEList(counts = data[ ,17:34],
                     group = experimental_groups,
                     genes = data[ ,1])
      dge$samples
      # pre-processing ----------------------------------------------------------
      ##  Filtering to remove low counts
      if(use_filter){
        keep <- filterByExpr(dge, group = experimental_groups)
        #table(keep)
        dge <- dge[keep, ,keep.lib.sizes = FALSE]
      }
      ##  Normalization for composition bias
         #  Normalization by trimmed mean of M values (TMM)
         #  eliminate composition biases between libraries
      dge <- calcNormFactors(dge)
      dge$samples
      # Model setup -------------------------------------------------------------
        design <- model.matrix(~0+experimental_groups, data = dge$samples)
      colnames(design) <- levels(dge$samples$group)
      design
      # Dispersion estimation ---------------------------------------------------
        dge <- estimateDisp(dge, design)
```

```
#plotBCV(dge)
# Model fitting ------------------------------------------------------------
# Model: Quasi-likelihood negative binomial generalized log-linear model
# Fit  : quasi-likelihood F-test
if (model_type == "QLF"){
  fit <- glmQLFit(dge, design, robust = TRUE)
  head(fit$coefficients)
  #plotQLDisp(fit)
# Model: Negative binomial generalized log-linear model
# Fit  : likelihood ratio test
} else {
  fit <- glmFit(dge, design)
  head(fit$coefficients)
}
# Testing for differential expression ----------------------------------------
result_list <-list()
decideTest_list <- list()
## (2 seasons + 3 sites)
## seasons: spring | summer
## sites: Ria de Aveiro | Troia | Ria Formosa
## single contrasts:
## season vs season in site
## site vs site in season
# Define contrast pairs-------
my_contrasts <- makeContrasts(
  Sp.RAvsTR = Sp_RA - Sp_TR,
  Sp.RAvsRF = Sp_RA - Sp_RF,
  Sp.TRvsRF = Sp_TR - Sp_RF,
  Su.RAvsTR = Su_RA - Su_TR,
  Su.RAvsRF = Su_RA - Su_RF,
  Su.TRvsRF = Su_TR - Su_RF,
  RA.SpvsSu = Sp_RA - Su_RA,
  TR.SpvsSu = Sp_TR - Su_TR,
  RF.SpvsSu = Sp_RF - Su_RF,
  levels = design
)
# ----------------------------
contrast_name <- colnames(my_contrasts)
for(name in contrast_name){
  cat("Performing contrast: ", name, "\n")
  if(model_type == "QLF"){
    result <- glmQLFTest(fit, contrast = my_contrasts[,name])
  } else {
    result <- glmRT(fit, contrast = my_contrasts[,name])
  }
  # Assessing Results  -------------------------------------------------------
  if(use_lfc){
    # Identify which genes are significantly differentially expressed for each
    # contrast from a fit object containing p-values and test statistics.
    summary_result <- summary(decideTests(result,
                                          p.value = p_value,
                                          lfc = log_fold_change,
                                          adjust.method = adjust_method))
    decideTest_result <- as.data.frame(decideTests(result,
                                                   p.value = p_value,
                                                   lfc = log_fold_change,
```

```
                                    adjust.method = adjust_method))
  }else{
    summary_result <- summary(decideTests(result,
                                p.value = p_value,
                                adjust.method = adjust_method))
    decideTest_result <- as.data.frame(decideTests(result,
                                    p.value = p_value,
                                    adjust.method = adjust_method))
  }
  # Extracts the most differentially expressed genes
  DEGs <- topTags(result,
                  n = nrow(result$table),
                  p.value = p_value,
                  adjust.method = adjust_method)$table
  # Processing Results  ----------------------------------------------------
  if(is.null(DEGs)){
    cat("No significant DEG's found", "\n")
    result_list[[name]] <- data.frame()
    decideTest_list[[name]] <- data.frame()
    decideTest <- cbind(result, decideTest_result)
    decideTest_with_data <- merge(decideTest, data, by.x = "genes", by.y="GeneID")
    decideTest_list[[name]] <- decideTest_with_data
    next
  }
  merge_degs_with_data <- merge(DEGs, data, by.x = "genes", by.y="GeneID")
  merge_degs_with_data <- merge_degs_with_data %>% arrange(desc(logFC))
  result_list[[name]] <- merge_degs_with_data
  decideTest <- cbind(result, decideTest_result)
  decideTest_with_data <- merge(decideTest, data, by.x = "genes", by.y="GeneID")
  decideTest_list[[name]] <- decideTest_with_data
  print(summary_result)
 }
 return(list(DEG_result = result_list, decideTest_result = decideTest_list))}

run_and_save_DEG <- function(data, blast_siglas, model_type,
                     use_filter=TRUE,use_lfc = FALSE){
 # --- run deg ---
 results <- perform_DEG(data,
                    model_type,
                    use_filter = use_filter,
                    use_lfc = use_lfc)

 deg <- results[["DEG_result"]]
 dt <- results[["decideTest_result"]]
 # --- save deg ---
 out_dir <- "./03-Output/01-DEG-Analysis/DEG-results"
 filter_label <- ifelse(use_filter, # test (use_filter = TRUE or FALSE)
                    "with-filterByExpr", # yes (use_filter = TRUE)
                    "without-filterByExpr") # no (use_filter = FALSE)
 save_to_excel(deg, file.path(out_dir, filter_label,
                       paste0("deg-", blast_siglas, ".xlsx")))
 save_to_excel(dt, file.path(out_dir, filter_label,
                       paste0("decideTest-", blast_siglas, ".xlsx")))
 return(list(DEG_result=deg, decideTest_result= dt))
}
```

```
get_top10_degs <- function(result_list, use_filter = TRUE, blast_siglas){
  top10 <- list()
  for (contraste_name in names(result_list)) {
    contrast <- result_list[[contraste_name]]
    if (nrow(contrast) > 0) {
      cat("\nPerforming contrast: ", contraste_name, "\n")
      top10_over <- contrast %>%
        filter(logFC > 0) %>%
        arrange(desc(PValue)) %>%
        head(10)
      # Top 10 under-expressed|downregulated (logFC < 0)
      top10_under <- contrast %>%
        filter(logFC < 0) %>%
        arrange(PValue) %>%
        head(10)
      # Top 10 global
      top10_global <- contrast %>%
        arrange(PValue) %>%
        head(10)
      min_fdr_gene <- top10_global %>%
        filter(FDR == min(FDR)) %>%
        pull(GeneNameID)
      max_fdr_gene <- top10_global %>%
        filter(FDR == max(FDR)) %>%
        pull(GeneNameID)
      # Dataframe with the top 10 under- plus the top 10 overexpressed genes
      top10_over_top10_under <- rbind(top10_over, top10_under)
      top10[[contraste_name]] <- top10_over_top10_under
      # Print parameter summary ----------------------------------------------------
      summary_top10 <- paste(
        "--- SUMMARY ---",
        "",
        "--- total counts ---",
        paste("Total DEGs:", nrow(contrast)),
        paste("Total over:", nrow(contrast%>%filter(logFC>0))),
        paste("Total under:", nrow(contrast%>%filter(logFC<0))),
        "",
        "--- over-expressed (logFC > 0) ---",
        paste(top10_over$GeneNameID, collapse = " |"),
        "",
        "--- under-expressed (logFC < 0) ---",
        paste(top10_under$GeneNameID, collapse = " |"),
        "",
        "--- Global Top 10 (Independent of logFC) ---",
        paste(top10_global$GeneNameID, collapse = " |"),
        "",
        "--- FDR Summary ---",
        paste("FDR (Over): Min:", min(top10_over$FDR), " | Max:", max(top10_over$FDR)),
        paste("FDR (Under): Min:", min(top10_under$FDR), " | Max:", max(top10_under$FDR)),
        paste("FDR (Global): Min:", min(top10_global$FDR), " | Max:", max(top10_global$FDR)),
        # Nota.: The min /or max GeneNameID, coud be several, there is different
        #        geneNameID with the same FDR
        "-----------------------------------------------------------------------",
        "",
        sep = "\n"
      )
```

```
    # Print summary
    cat(summary_top10, sep ="\n\n")
    # ------------------------------------------------------------------------
   }
  }
  # --- save deg ---
  out_dir <- "./03-Output/01-DEG-Analysis/DEG-results"
  filter_label <- ifelse(use_filter,"with-filterByExpr","without-filterByExpr")
  save_to_excel(top10, file.path(out_dir, filter_label, paste0("top10-", blast_siglas, ".xlsx")))
  return(top10)}

get_GeneNameID_contrasts <- function(deg_list) {
  # Remove empty data frames from the list
  deg_list <- deg_list[sapply(deg_list, function(df) nrow(df) > 0)]
  # For each contrast, extract GeneNameID and rename the column to the contrast name
  gene_dfs <- lapply(names(deg_list), function(contrast) {
    deg_list[[contrast]] %>%
      select(GeneNameID) %>%
      rename(!!contrast := GeneNameID)
  })
  # Combine all the data frames into one
  contrasts_GeneNameID <- bind_rows(gene_dfs)
  # Convert to long format and summarize which contrasts each GeneNameID appears in
  result <- contrasts_GeneNameID %>%
    pivot_longer(
      cols = everything(),
      names_to = "contrast",
      values_to = "GeneNameID",
      values_drop_na = TRUE
    ) %>%
    group_by(GeneNameID) %>%
    summarize(
      contrast = toString(unique(contrast)),
      .groups = "drop"
    )
  b <- a %>% count(contrast) %>% arrange(desc(n))
  return(result)}

common_deg_SP_Z <- function(deg_SP, deg_Z) {
  merged_list <- list()
  not_in_common_list_SP <- list()
  not_in_common_list_Z <- list()
  for (contrast_name_Z in names(deg_Z)) {
    for (contrast_name_SP in names(deg_SP)) {
      if (contrast_name_SP == contrast_name_Z) {
        contrast_z  <- deg_Z[[contrast_name_Z]]
        contrast_sp <- deg_SP[[contrast_name_SP]]
        # Check if either data frame is empty
        if (nrow(contrast_z) == 0 || nrow(contrast_sp) == 0) {
          next
        }
        common_degs <- merge(contrast_sp, contrast_z, by = "GeneNameID",
                             suffixes = c(".SP", ".Z"))
        common_degs  <- common_degs  %>%
          select(GeneNameID,
                 logFC.SP,logFC.Z,
```

```
            PValue.SP, PValue.Z,
            FDR.SP, FDR.Z,
            pident.SP, pident.Z,
            evalue.SP, evalue.Z,
            OS.SP,OS.Z) %>%
          arrange(logFC.SP) %>% #(menor -> maior)
          mutate(
            pident_category.SP = evaluate_sequence_quality(pident.SP, "pident"),
            pident_category.Z = evaluate_sequence_quality(pident.Z, "pident"),
            evalue_category.SP = evaluate_sequence_quality(evalue.SP, "evalue"),
            evalue_category.Z = evaluate_sequence_quality(evalue.Z, "evalue"),
            FDR_category.SP = evaluate_sequence_quality(FDR.SP, "FDR"),
            FDR_category.Z = evaluate_sequence_quality(FDR.Z, "FDR"))
        # Genes NOT in common ----------------------------------------------------
        not_in_common_SP <- anti_join(contrast_sp, contrast_z, by = "GeneNameID")
        not_in_common_Z  <- anti_join(contrast_z, contrast_sp, by = "GeneNameID")
        not_in_common_SP <- not_in_common_SP %>%
          select(GeneNameID,logFC, PValue, FDR, pident, evalue, OS) %>%
          arrange(logFC) %>%
          mutate(
            pident_category = evaluate_sequence_quality(pident, "pident"),
            evalue_category = evaluate_sequence_quality(evalue, "evalue"),
            FDR_category = evaluate_sequence_quality(FDR, "FDR"))
        not_in_common_Z <- not_in_common_Z %>%
          select(GeneNameID,logFC, PValue, FDR, pident, evalue, OS) %>%
          arrange(logFC) %>%
          mutate(
            pident_category = evaluate_sequence_quality(pident, "pident"),
            evalue_category = evaluate_sequence_quality(evalue, "evalue"),
            FDR_category = evaluate_sequence_quality(FDR, "FDR"))
        # ------------------------------------------------------------------------
        # Save the data frame in the list
        merged_list[[contrast_name_SP]] <- common_degs
        not_in_common_list_SP[[contrast_name_SP]] <- not_in_common_SP
        not_in_common_list_Z[[contrast_name_Z]]   <- not_in_common_Z
        # Mini summary
        cat("\nContrast:", contrast_name_SP)
        cat("\ndeg swiss-prot:", nrow(contrast_sp))
        cat("\ndeg zebrafish:", nrow(contrast_z))
        cat("\nIn common: ", nrow(common_degs))
        cat("\nunder: ",nrow(common_degs %>% filter(logFC.SP < 0)))
        cat("\nover: ", nrow(common_degs %>% filter(logFC.SP > 0)),"\n")
      }
    }
  }

  return(list(
    common = merged_list,
    unique_SP = not_in_common_list_SP,
    unique_Z = not_in_common_list_Z
  ))}

get_uniprot_taxainfo <- function(deg_list, specific_accession = NULL) {
  uniprot_info <- list()
  for (contrast_name in names(deg_list)) {
    contrast <- deg_list[[contrast_name]]
```

```
    if (nrow(contrast) > 0) {
      # Decide which accession(s) to use
      Accession <- if (!is.null(specific_accession)) specific_accession else contrast$Accession
      ## 1. Get Taxonomy Information
       TaxaObj <- GetNamesTaxa(Accession)
       merge_data <- merge(contrast, TaxaObj, by.x="Accession", by.y="Entry", all.x = TRUE)
      uniprot_info[[contrast_name]] <- merge_data
    }
   }
   return(uniprot_info)}

 get_uniprot_proteinFunction <- function(deg_list, specific_accession = NULL) {
   uniprot_info <- list()
   for (contrast_name in names(deg_list)) {
    contrast <- deg_list[[contrast_name]]
    if (nrow(contrast) > 0) {
      # test
      #Accession <- deg_SP[["Sp.RAvsTR"]]$Accession
      # Decide which accession(s) to use
      Accession <- if (!is.null(specific_accession)) specific_accession else contrast$Accession
      ## 1. Get protein function Information
      Obj <- GetProteinFunction(Accession)
      merge_data <- merge(contrast, Obj, by.x="Accession", by.y=0, all.x = TRUE)
      uniprot_info[[contrast_name]] <- merge_data
    }
   }
   return(uniprot_info)}

 get_uniprot_GOinfo <- function(deg_list, specific_accession = NULL) {
   uniprot_info <- list()
    for (contrast_name in names(deg_list)) {
    contrast <- deg_list[[contrast_name]]
    if (nrow(contrast) > 0) {
      # test
      #Accession <- deg_SP[["Sp.RAvsTR"]]$Accession
      # Decide which accession(s) to use
      Accession <- if (!is.null(specific_accession)) specific_accession else contrast$Accession
      ## 1. Get GO terms
      Obj <- GetProteinGOInfo(Accession)
      #merge_data <- merge(contrast, Obj, by.x="Accession", by.y=0, all.x = TRUE)
      uniprot_info[[contrast_name]] <- Obj
      #uniprot_info[[contrast_name]] <- merge_data
    }
   }
   return(uniprot_info)
 }


 datasubset_keyword_filtered <- function(deg_list, keyword_pattern = NULL, column_name =
NULL) {
   result <- list()
   for (contrast_name in names(deg_list)) {
    contrast <- deg_list[[contrast_name]]
    if (nrow(contrast) > 0) {
      data_filtered <- contrast %>%
       filter(grepl(keyword_pattern, .data[[column_name]], ignore.case = TRUE)) %>%
```

```
            mutate(keyword = str_extract(.data[[column_name]], regex(keyword_pattern, ignore_case
= TRUE))) %>%
            select(Accession, GeneNameID, logFC, FDR, pident, evalue, Function..CC., keyword )
           result[[contrast_name]] <- data_filtered
          }
         }
         return(result)}

        process_STRINGdata <- function(uniprot_taxaobj){
         preprocessing <- function(df){
          df %>%
           select("Gene.Names..primary.", "logFC")%>%
           arrange(logFC)}
         STRING_data <- lapply(uniprot_taxaobj, preprocessing)
         return(STRING_data)}

        process_GSEAdata <- function(uniprot_taxaobj){
         preprocessing <- function(df){
          df %>%
           select("Accession", "logFC", "PValue")%>%
           arrange(PValue)}
         GSEA_data <- lapply(uniprot_taxaobj, preprocessing)
         return(GSEA_data)}

        process_EnrichmentScores <- function(enrichment_scores){
         # convert to a data frame
         ES <- as.data.frame(enrichment_scores)
         # leadingEdge -> [1] "A8E7C5" "Q6PC64" -> list
         # collapse leadingEdge list to string (to facilitate data manipulation)
         ES$leadingEdge <- sapply(ES$leadingEdge,
                                  function(x) paste(x, collapse = ";"))
         # out: leadingEdge -> "A8E7C5;Q6PC64" -> string
         ES <- ES %>%
          arrange(padj)
         # remove pathway prefix if present
         pathway_prefix <- "^\\(KEGG\\)"
         if ("pathway" %in% colnames(ES)){
          ES$pathway <- sub(pathway_prefix, "", ES$pathway)
         }
         return(ES)}

        perform_GSEA <- function(GSEA_data){
         databases <- c("kegg") # options: kegg, reactome, GO
         pathways <- getMultiOmicsFeatures(
          dbs = databases,
          layer = "transcriptome",
          returnTranscriptome = "UNIPROT", # options: SYMBOL, ENTREZID, UNIPROT, ENSEMBL,
REFSEQ
          useLocal = FALSE,
          organism = "drerio"
         )
         pathways_short <- lapply(names(pathways), function(name){
          head(pathways[[name]], 2)
         })
         names(pathways_short) <- names(pathways)
         #pathways_short
```

```
        # ES - [E]nrichemnt [S]core
        raw_ES <- list()
        processed_ES <- list()
        top_10 <- list()
        for (contraste_name in names(GSEA_data)) {
          # test
          contrast <- GSEA_data[["Sp.RAvsTR"]]
          contrast <- GSEA_data[[contraste_name]]
          if (nrow(contrast) > 0) {
            omics_data <- initOmicsDataStructure(layer = c("transcriptome"))
            layers <- names(omics_data)
            print(layers)
            # ranks
            omics_data$transcriptome <- rankFeatures(
              contrast$logFC,
              contrast$PValue
            )
            #print(head(omics_data$transcriptome))
            names(omics_data$transcriptome) <- contrast$Accession
            omics_data$transcriptome <- sort(omics_data$transcriptome)
            #print(head(omics_data$transcriptome))
            set.seed(42)
            enrichment_scores <- multiGSEA(pathways,
                                           omics_data) # ranks
            #print(head(enrichment_scores))
            #View(enrichment_scores[["transcriptome"]])
            # Save the results
            if(is.null(enrichment_scores)){
              raw_ES[[contraste_name]] <- data.frame()
              processed_ES[[contraste_name]] <- data.frame()
              top_10[[contraste_name]]<-data.frame()
              next
            }
            raw_ES[[contraste_name]] <- enrichment_scores
            processed_enrichment_scores   <-   process_EnrichmentScores(enrichment_scores$transcrip-
tome)
            processed_ES[[contraste_name]] <- processed_enrichment_scores %>% filter(padj < 0.05)
            top_10[[contraste_name]]<- processed_enrichment_scores %>% head(10)
          }
        }
        return(list(ES_raw_result = raw_ES,
                    ES_processed_result = processed_ES,
                    top_10_result = top_10))
      }
```

Nota.: No presente trabalho foram utilizadas ferramentas de inteligência generativa, nomeadamente https://www.perplexity.ai/ , para revisão gramatical, reformulação de texto. Estas ferramentas foram empregues sob a supervisão do autor e todos os conteúdos gerados foram verificados quanto à sua precisão. A autoria e a validação dos conteúdos permanecem inteiramente sob a responsabilidade do autor, estando a utilização da inteligência artificial em conformidade com as normas institucionais de integridade académica

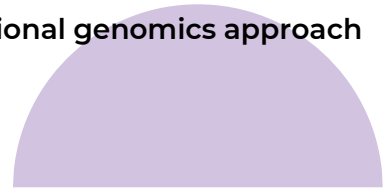
UNIVERSIDADE NOVA
DE LISBOA

2025

Computational tools to assess adaptation and resilience to thermal stress in intertidal fish: a functional genomics approach

ANDREA KAROLINA DUARTE DUQUE

MASTER IN **COMPUTATIONAL BIOLOGY & BIOINFORMATICS**



UNIVERSIDADE NOVA DE LISBOA