Development of a computational toolbox for data integration in environmental pathways analysis

Diana Catarina Seco Simões
Master in Bioinformatics and Computational Biology
Department of Computer Science
Faculty of Sciences of the University of Porto
2025

Development of a computational toolbox for data integration in environmental pathways analysis

Diana Catarina Seco Simões
Dissertation carried out as part of the Master in Bioinformatics and Computational Biology
Department of Computer Science
2025

Supervisor
Teresa Neuparth, Researcher, CIIMAR

Co-supervisor
João Carneiro, Researcher, CBMA/IB-S

Co-supervisor
Mónica Lopes-Marques, Researcher, CIIMAR

Dedicated to Avó Paivita and my parents, for their unconditional support. I love you <3

Author Statement

The work developed in this dissertation was entirely carried out by the candidate, closely monitored and reviewed by the supervisor team. The work developed was communicated in several scientific meetings during its development period in which I was the main presenting author.

- Poster Communication: “Transgenerational Effects of Environmental Contaminants in *Gammarus locusta*: Insights from Multi-Omic Analyses” Simões, D., Machado, A., Lopes-Marques, M., Carneiro, J., Neuparth., T. XX ENBE Annual Meeting of the Portuguese Association for Evolutionary Biology. December 2024
- Poster Communication: “Transcriptomic Insights into Transgenerational Effects of Environmental Contaminants in *Gammarus locusta*” Simões, D., Machado, A., Lopes-Marques, M., Carneiro, J., Neuparth., T. DCC Master’s Day. April 2025
- Poster Communication: “Development of a computational toolbox for data integration in environmental pathway analysis” Simões, D., Machado, A., Lopes-Marques, M., Carneiro, J., Neuparth., T. IJUP Encontro de Investigação Jovem da U.Porto. May 2025

Acknowledgements

Em primeiro lugar, quero expressar a minha profunda gratidão à minha orientadora, **Teresa Neuparth**, que desde o início me acolheu e me fez sentir que podia sempre contar com o seu apoio. Agradeço por todas as vezes em que me explicou com paciência os mecanismos dos efeitos transgeracionais, pelo interesse constante no progresso do meu trabalho e pela forma como me guiou ao longo de todo o percurso. Obrigada por estar sempre disponível, mesmo fora de horas, para me ajudar, especialmente nesta fase final, que foi tão exigente. Agradeço ainda a motivação para procurar ajuda sempre que me sentia bloqueada numa análise. Um simples "obrigada" não é suficiente para retribuir toda a dedicação, compreensão e paciência que mostrou.

Em segundo lugar, quero agradecer ao **André Machado**, que foi um verdadeiro mentor nesta tese de mestrado. Mesmo no meio da escrita da sua própria tese de doutoramento, dedicou parte do seu tempo para me acompanhar desde o início deste trabalho. Foi ele quem me guiou nos primeiros passos em bioinformática, explicou conceitos fundamentais, mostrou como funcionava o cluster e partilhou referências e artigos essenciais. Ajudou-me a compreender programas, o seu funcionamento e as suas aplicações, incentivando-me constantemente a desenvolver pensamento crítico sobre os resultados e ferramentas que utilizava. Sou-lhe profundamente grata pela paciência, pela forma acessível e generosa como dedicou horas a explicar-me tudo até eu compreender verdadeiramente, mesmo tendo tantas outras responsabilidades e pessoas a recorrer à sua ajuda.

Em terceiro lugar, quero agradecer à **Mónica Marques**, a minha segunda co-orientadora, por me ajudar a compreender melhor alguns conceitos bioinformáticos em transcriptómica que inicialmente eram pouco claros para mim. Agradeço por ter estado sempre disponível, por oferecer insights valiosos no desenvolvimento da minha aplicação e por demonstrar interesse em ajudar quando eu enfrentava dificuldades.

Em quarto lugar, agradeço ao **João Carneiro**, o meu primeiro co-orientador, por, mesmo à distância, incentivar a participação da equipa de orientação em projetos e atividades de team-building, promovendo a interação entre todos. Sou grata também pelo seu apoio nas dúvidas sobre parâmetros específicos a usar e pelos insights valiosos que a sua experiência em bioinformática trouxe para aprimorar a minha aplicação.

Quero agradecer às pessoas do CIIMAR que me apoiaram, em especial ao **Diogo**, que dedicou horas a instalar programas no meu computador, a esclarecer dúvidas e a fazer-me companhia, e ao **André**, que também esteve presente quando precisei e que junto do Diogo tornou as tardes no CIIMAR mais agradáveis.

Quero também agradecer às pessoas que estiveram mais próximas de mim ao longo do mestrado: **Ana Luísa**, **Tomás** e **Cecília**, que me acompanharam praticamente desde o primeiro dia nesta faculdade nova para mim. Guardo com muito carinho todas as nossas memórias; o nosso grupo é muito especial para mim e vocês marcaram de forma única estes dois anos de mestrado. Pelas idas a karaokes, pelas francesinhas e por simplesmente estarmos juntos nas aulas ou qualquer lado, cada momento foi incrível e tenho um enorme carinho por todos vocês.

A outros colegas do mestrado, como **Pedro**, **David** e os **Diogos**, agradeço pelas idas a hackathons, que me permitiram crescer profissionalmente no desenvolvimento de aplicações e no uso de novas linguagens de programação, e pela oportunidade de aprender com pessoas tão fascinantes como vocês. Um obrigado também às pessoas que conheci nesses eventos, em especial ao **Daniel Cardoso**, que me deixaram uma marca positiva.

À **Marta**, ao **Pedro** e à **Eneia**, por terem marcado também o meu percurso na FCUP logo desde o início e que me proporcionaram momentos muito felizes.

A todos os **colegas de eventos de bioinformática**, que me deram insights valiosos e com quem pude discutir métodos e análises do meu trabalho.

Agradeço às **minhas amigas e amigos de Coimbra**, que, apesar de não terem acompanhado tão de perto estes últimos dois anos de mestrado, fizeram parte de momentos inesquecíveis durante os três anos da minha licenciatura, que foram, sem dúvida, alguns dos mais bonitos da minha vida.

Agradeço também à **Filipa**, por me fazer companhia neste último ano de mestrado, pelas idas ao ginásio juntas, pelos jantares na tua casa, onde me ajudaste a resolver problemas da tese e por até me quereres dar comida nesta fase final. E à **Alexandra**, minha amiga mais antiga, pelas longas caminhadas, conversas sobre a vida e por me ajudar a distrair nos momentos mais difíceis.

Agradeço ao **Tiago**, o meu querido namorado, que esteve comigo de perto ao longo de todo este percurso no Porto e na FCUP, praticamente desde que me mudei para cá. Obrigada por te preocupares sempre comigo, por perceberes os momentos difíceis desta fase final da tese e por me motivares a dar o meu melhor. Obrigada por cuidares de mim em todos os aspetos possíveis, por seres a minha família aqui no Porto e por todo o carinho e acolhimento que tu e a tua família sempre me deram. És uma pessoa muito especial na minha vida e sou muito grata por te ter ao meu lado.

Por último, mas não menos importante, quero agradecer aos **meus pais** por me proporcionarem a oportunidade de realizar este mestrado no Porto, e à **minha mãe**, em particular, por se preocupar tanto comigo, sempre me incentivar a dar o meu melhor e a querer o melhor para mim, oferecendo-me continuamente o seu apoio. Agradeço também à **minha irmã**, por ser quem é e por trazer alegria à minha vida. Aos **meus familiares**, em especial ao meu **avô Simões**, que desde a minha infância sempre disse acreditar que eu teria um futuro brilhante, e a todos os outros familiares que, de forma direta ou indireta, impactam a minha vida, contribuindo positivamente para que eu possa dar sempre o meu melhor.

Resumo

O consumo global de contaminantes tem aumentado significativamente, incluindo fármacos como a Sinvastatina (SIM), um medicamento amplamente prescrito para redução do colesterol em humanos. Devido ao seu uso extensivo, a SIM é frequentemente libertada para o ambiente através de efluentes farmacêuticos e águas residuais urbanas. As estações de tratamento de águas residuais convencionais geralmente não são projetadas para remover completamente esses contaminantes, resultando na sua deteção generalizada em sistemas aquáticos, incluindo efluentes, estuários e ambientes marinhos. Organismos não-alvo são, consequentemente, expostos a concentrações subletais, podendo sofrer efeitos moleculares, fisiológicos e reprodutivos.

Gammarus locusta, um anfípode marinho, constitui um organismo modelo ecologicamente relevante para estudar os impactos de contaminantes, devido ao seu papel-chave nas cadeias alimentares costeiras e à sua sensibilidade a stressores químicos. Este estudo avaliou os efeitos diretos e transgeracionais da exposição à SIM em machos de *G. locusta* ao nível molecular e correlacionou-os com fenótipos observados em estudos anteriores. Foi gerado *de novo* um transcriptoma de referência para machos (V1), devidamente anotado e agrupado em clusters, permitindo análises detalhadas ao nível de genes. A expressão génica diferencial foi avaliada com o **TransDegAnalyser,** uma aplicação desenvolvida para este estudo baseada em Shiny para análise e exploração interativa de dados de RNA-Seq.

A exposição à SIM induziu alterações transcriptómicas persistentes, afetando a formação do exoesqueleto, o metabolismo e as vias de resposta ao stress, consistentes com a redução do tamanho corporal observada nos machos F0 e F3. Estes resultados fornecem uma compreensão mecanística dos efeitos transgeracionais de contaminantes ambientais e destacam o valor de ferramentas transcriptómicas integrativas para a investigação ecotoxicológica.

Palavras-chave: Sinvastatina, fármacos, contaminantes ambientais, *Gammarus locusta*, efeitos transgeracionais, transcriptómica, RNA-Seq, ecotoxicologia, efeitos moleculares

Abstract

The global consumption of contaminants has increased substantially, including pharmaceuticals such as simvastatin (SIM), a widely prescribed lipid-lowering drug in humans. Due to its extensive use, SIM is frequently released into the environment through pharmaceutical effluents and urban wastewater. Conventional wastewater treatment plants are generally not designed to fully remove such contaminants, resulting in their widespread detection in aquatic systems, including effluents, estuaries and marine environments. Non-target organisms are consequently exposed to sub-lethal concentrations, potentially causing molecular, physiological and reproductive effects.

Gammarus locusta, a marine amphipod, represents an ecologically relevant model organism for studying contaminant impacts due to its key role in coastal food webs and its sensitivity to chemical stressors. This study evaluated direct and transgenerational effects of SIM exposure in male *G. locusta* at the molecular level and related them to observed phenotypes in previous studies. A high-quality male reference transcriptome (V1) was generated *de novo*, annotated and clustered to enable detailed gene-level analyses.

Differential gene expression was assessed using **TransDegAnalyser**, a Shiny-based application developed in this study for analysis and interactive exploration of RNA-Seq data. SIM exposure induced persistent transcriptomic changes affecting cuticle formation, metabolism and stress-response pathways, consistent with the reduced body size observed in F0 and F3 males. These findings provide mechanistic insight into transgenerational effects of environmental contaminants and highlight the value of integrative transcriptomic tools for ecotoxicological research.

Keywords: Simvastatin, pharmaceuticals, environmental contaminants, *Gammarus locusta*, transgenerational effects, transcriptomics, RNA-Seq, ecotoxicology, molecular effects

Table of Contents

List of Tables

List of Figures

List of Abbreviations

FCUP	FACULTY OF SCIENCES OF THE UNIVERSITY OF PORTO
UP	UNIVERSITY OF PORTO
WWTP	WASTEWATER TREATMENT PLANT
CEC	CONTAMINANT OF EMERGING CONCERN
LDL	LOW-DENSITY LIPOPROTEIN
SIM	SIMVASTATIN
G. LOCUSTA	*GAMMARUS LOCUSTA*
MF	METHYL FARNESOATE
OECD	ORGANIZATION FOR ECONOMIC COOPERATION DEVELOPMENT
OORF	OECD OMICS REPORTING FRAMEWORK
AOP	ADVERSE OUTCOME PATHWAY
MIES	MOLECULAR INITIATING EVENTS
KES	KEY EVENTS
AO	ADVERSE OUTCOMES
TPOD	TRANSCRIPTOMIC POINTS OF DEPARTURE
TK-TD	TOXICOKINETIC – TOXICODYNAMIC
PE	PAIRED-END READS
RNA-SEQ	RNA SEQUENCING
DEGS	DIFFERENTIALLY EXPRESSED GENES
GO	GENE ONTOLOGY
KEGG	KYOTO ENCYCLOPEDIA OF GENES AND GENOMES
ORFS	OPEN READING FRAMES
HPC	HIGH PERFORMANCE COMPUTING
BUSCO	BENCHMARKING UNIVERSAL SINGLE-COPY ORTHOLOGS
SAM	SEQUENCE ALIGNMENT MAP
BAM	BINARY ALIGNMENT MAP
GTF	GENE TRANSFER FORMAT
BP	BIOLOGICAL PROCESS
MF	MOLECULAR FUNCTION
CC	CELLULAR COMPONENT
PCA	PRINCIPAL COMPONENT ANALYSIS
COG	CLUSTER OF ORTHOLOGOUS GENES
EC	ENZYME COMMISSION NUMBERS

V0	VERSION 0 OF TRANSCRIPTOME
V1	VERSION 1 OF TRANSCRIPTOME
GC	GUANINE–CYTOSINE CONTENT
EC	ENZYME COMISSION

1. Introduction

1.1. Ecotoxicological Context and Biological Background

1.1.1. Pharmaceuticals as Emerging Contaminants

Pharmaceutical compounds have become essential components of modern healthcare systems worldwide, playing a pivotal role in the treatment and prevention of a wide range of human and veterinary diseases (Aus der Beek et al., 2016). Over the past few decades, their consumption has significantly increased, driven by factors such as population growth, aging demographics, expanded access to healthcare and advances in pharmaceutical research and global distribution (Van Boeckel et al., 2014). It is currently estimated that over 3,000 active pharmaceuticals are in use worldwide, with total production volumes reaching several hundred tons annually (Carvalho & Santos, 2016; Grenni et al., 2018).

This exponential growth in pharmaceutical use has led to the widespread occurrence of parental pharmaceuticals and their transformation products in the environment, raising emerging environmental concerns regarding their ecological impact (Ortúzar et al., 2022).

Following their administration, pharmaceuticals are often excreted in urine and feces as active substances or metabolites (Sui et al., 2015; Aus der Beek et al., 2016). As a result, these substances enter wastewater systems via domestic sewage, hospital discharges and effluents from pharmaceutical manufacturing, ultimately reaching surface waters (Michael et al., 2013; Verlicchi et al., 2012). This comes from the fact that conventional wastewater treatment plants (WWTPs) were not specifically designed to remove pharmaceutical compounds (Patel et al., 2019). WWTPs primarily target organic matter that is easily or moderately biodegradable, such as sugars, proteins and simple fats, which can be readily broken down by microorganisms and are typically present at mg/L concentrations (Stackelberg et al., 2004). In contrast, pharmaceuticals often resist microbial degradation and are biologically active even at much lower concentrations (ng/L–µg/L), making them difficult to remove in conventional treatment systems. (Aus der Beek et al., 2016; Patel et al., 2019; Verlicchi et al., 2012; Michael et al., 2013; Le-Minh et al., 2010; Ziylan & Ince, 2011). Consequently, many bioactive substances, particularly antibiotics, analgesics and lipid regulators are only partially degraded during wastewater treatment processes and enter in surface waters. A comprehensive review by Aus der

Beek et al. (2016) identified the presence of 631 pharmaceutical compounds in at least 71 countries. These substances have been detected in various water bodies, including rivers, lakes, estuaries and coastal marine waters, highlighting their persistent and diffuse presence in aquatic environments (Patel et al., 2019; Ortúzar et al., 2022).

Although pharmaceuticals are designed to target specific biological pathways in humans or domestic animals, many of these compounds can also affect non-target aquatic organisms (Neuparth et al., 2014; Aus der Beek et al., 2016). Growing evidence indicates that even at environmentally relevant concentrations, often in the nanogram to low microgram per liter range (ng/L-µg/L), pharmaceutical compounds can trigger a range of sublethal effects in non-target organisms (Neuparth et al., 2014; Aus der Beek et al., 2016). This sensitivity is particularly concerning given the chronic and often lifelong exposure experienced by aquatic species, especially in ecosystems continuously receiving wastewater effluents near to urban areas (Paut Kusturica et al., 2022). Moreover, even brief or single exposures to certain pharmaceuticals may trigger biological changes with transgenerational consequences, affecting unexposed offspring across successive generations (Neuparth et al., 2020).

Given their persistence, biological activity at low concentrations and widespread detection in aquatic environments, pharmaceuticals are now highly recognized as a major group of Contaminants of Emerging Concern (CECs) (Aus der Beek et al., 2016; Richardson & Ternes, 2021). This classification reflects their potential to harm ecosystems and human health, as well as the current lack of comprehensive monitoring and regulatory control (Patel et al., 2019; Fakhri et al., 2024). In fact, current regulatory frameworks remain insufficient and many compounds were approved for medical use long before their environmental fate and ecological effects were properly evaluated (Ortúzar et al., 2022).

1.1.2. Simvastatin as an Emerging Contaminant, Occurrence in Aquatic Environments and Potential for Bioaccumulation

Among the CECs detected in aquatic environments, statins as lipid regulator pharmaceuticals, have emerged as compounds of significant ecotoxicological concern due to their widespread use, environmental persistence and potent biological activity at low concentrations (Tete et al., 2020).

Statins are a class of hypolipidemic agents extensively prescribed to humans for the prevention and treatment of hypercholesterolemia and cardiovascular diseases. This

condition is associated with elevated levels of circulating low-density lipoprotein (LDL) cholesterol in the bloodstream (Stossel, 2008; Endo, 2010; Nunes, 2017; Sirtori, 2014).

Among all the statins, simvastatin (SIM) (Figure 1) is one of the most widely prescribed pharmaceutical in Europe and the USA and consequently the most **frequently detected in aquatic environments**, particularly near urban effluents and surface waters, often at concentrations ranging from ng/L to µg/L (Miao and Metcalfe, 2003; Walley et al., 2005; Ellesat et al., 2010; Santos et al., 2016; Tete et al., 2020).

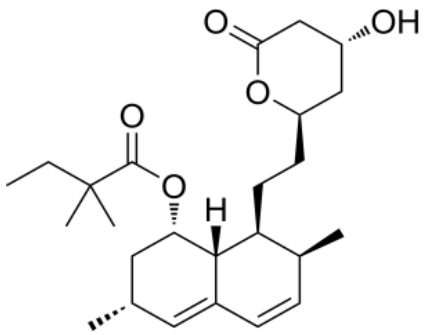


Figure 1 – Simvastatin.

SIM enters wastewater systems in substantial amounts and has been frequently detected in effluents and receiving waters across different countries. In Portugal, influents to WWTPs have been reported to contain SIM at median concentrations of 2652.1 ng/L (Pereira et al., 2015), while in the southeastern United States, sewage influents from a medium-sized WWTP contained 1230 ng/L (Ottmar, Colosi and Smith, 2012). Concentrations in WWTP effluents are generally lower than in influents but remain detectable, ranging from 5 ng/L in the UK (Kasprzyk-Hordern et al., 2009; Boleda et al., 2011) to 2.65 ± 0.8 µg/L in South Africa (Tete et al., 2020) and up to 369.8 ng/L in Portuguese effluents (Salgado et al., 2010), confirming that conventional treatment does not completely remove SIM from wastewater and illustrating its pervasive occurrence globally (Rebelo et al., 2021).

SIM has also been detected in surface waters at lower concentrations, ranging from 2.3 to 16.7 ng/L in the Jiulong River Basin and estuary in Southeast China (Lv et al., 2014) and in Norwegian freshwater ecosystems, a predicted environmental concentration of 630 ng/L have been estimated (Santos et al., 2016) highlighting its presence even at ng/L.

The octanol-water partition coefficient (log Kow) is a widely used indicator of a chemical's hydrophobicity and its tendency to accumulate in lipid-rich tissues. Compounds with

higher log Kow values (≥ 3) are more lipophilic and therefore have a greater potential for bioaccumulation and sorption to organic matter and sediments (Mitchell et al., 2013). With a log Kow of 4.68, SIM shows high potential for bioaccumulation. Its moderate hydrophobicity and structural stability promote strong partitioning to particulate matter and sediments, where it can persist for prolonged periods and accumulate in benthic zones (Santos et al., 2016). Such sorption enhances bioavailability to sediment-dwelling organisms, including marine crustaceans like amphipods, either through direct contact or ingestion of contaminated particles (Ritter & Bourne, 2024).

1.1.2.1. Simvastatin's Mode of Action in Vertebrates and Invertebrates

SIM, like other statins, inhibits Hydroxymetylglutaryl-CoA reductase (HMGR),, a key enzyme of the mevalonate pathway (Goldstein & Brown, 1990; Endo, 2010; Zhou & Liao, 2009; Attardo et al., 2022; Costa et al., 2025; Neuparth et al. 2020). While this pathway is conserved across metazoans, its downstream products and physiological roles differ between vertebrates and invertebrates (Figure 2) (Bellés et al., 2005; Lombard & Moreira, 2010; Neuparth et al., 2020). Functional studies have shown that HMGR is competitively inhibited by SIM in both vertebrates and arthropods (Zapata et al., 2002a,b; Li et al.2003; Santos et al., 2016; Costa et al., 2025).

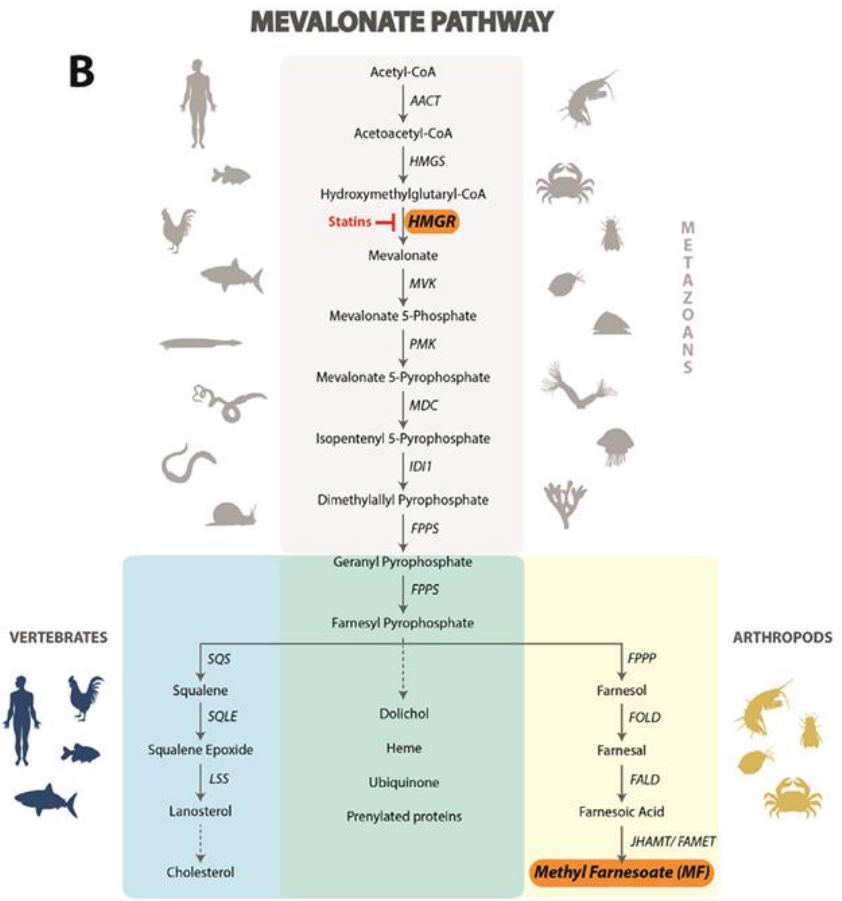


Figure 2 - Schematic representation of the mevalonate pathway across metazoans, highlighting in orange the point of statin inhibition and showing the metabolic divergence between vertebrates and arthropods. *Figure from T. Neuparth, et al., 2020.*

In vertebrates, the mevalonate pathway primarily leads to cholesterol synthesis, an essential component of cell membranes and a precursor for other important isoprenoids (Bloch, 1992; Alberts et al., 1980; Miziorko, 2011). By inhibiting HMGR, mevalonate production is reduced, leading not only to lower cholesterol levels (Figure 2 - blue), but also to alterations in key intermediates, such as ubiquinone (coenzyme Q10) and dolichols. (Littarru & Langsjoen, 2007; Vaklavas et al., 2009; Santos et al., 2016; Chen et al., 2024). These intermediates are critical for mitochondrial electron transport, protein prenylation and glycoprotein synthesis, meaning that SIM exposure can influence cellular energy metabolism, protein function and overall cellular homeostasis (Ericsson & Gustav Dallner, 1993; Rauthan & Pilon, 2011; Guerra et al., 2021; Pisanti et al., 2022).

In contrast to vertebrates, arthropods do not synthesize cholesterol *de novo* and instead obtain it largely from their diet (Li et al., 2003). In crustaceans, the mevalonate pathway, which shares most of the enzymatic steps with vertebrates (Figure 2 - grey), produces the isoprenoid methyl farnesoate (MF) (Figure 2 - yellow) (Bellés et al., 2005; Neuparth

et al., 2014). By inhibiting HMGR, mevalonate production is reduced, leading to lower MF levels. MF is a juvenile hormone analogue crucial for regulating reproduction, development and molting in crustaceans (Bellés et al., 2005; Goodman & Cusson, 1976; Neuparth et al., 2014).

Although SIM inhibits HMGR in both vertebrates and arthropods, its downstream effects differ: in arthropods, inhibition reduces farnesyl pyrophosphate production (the MF precursor), disrupting endocrine signaling and impacting growth, molting and reproduction (Neuparth et al., 2020; Santos et al., 2016).

Although SIM was originally developed to modulate lipid metabolism in humans, the structural and functional conservation of the mevalonate pathway across metazoans means that aquatic arthropods exposed to environmentally relevant concentrations of SIM may experience sublethal and reproductive effects, despite lacking a cholesterol metabolism equivalent to vertebrates. This susceptibility has been demonstrated in several ecotoxicological studies, including transgenerational effects in marine sediment-dwelling crustaceans such as *G. locusta* (Neuparth et al., 2014; Neuparth et al., 2020; Neuparth et al., 2022).

The detection of SIM in coastal waters and sediments worldwide (Pereira et al., 2015; Ritter & Bourne, 2024), together with the dependence of these taxa on endocrine signals derived from the mevalonate pathway for reproduction and development, underscores a tangible mechanistic risk for non-target crustacean species. By interfering with both metabolic processes and endocrine regulation, SIM can trigger effects that cascade across life stages and persist through generations (Neuparth et al., 2020; Neuparth et al., 2022).

Given these widespread environmental exposures and the critical role of the mevalonate pathway in crustacean reproduction, *G. locusta* has emerged as a relevant model species to investigate the sublethal, reproductive and transgenerational effects of SIM, as demonstrated in studies by Neuparth et al. and collaborators (2014; 2020). In this context, **SIM was selected in this study as a model compound** to assess its effects on the amphipod *Gammarus locusta*.

1.1.3. *Gammarus locusta* as a Model Organism

Gammarids are small, laterally compressed, shrimp-like crustacean amphipods characterized by the absence of a carapace and a segmented body covered by a flexible coxal plates, divided into three main regions: head, thorax (in red in Figure 3) and

abdomen (in blue and green in Figure 3) (Gerhardt et al., 2011). They are widely distributed in aquatic environments, including marine, freshwater and estuarine habitats (Whiteley et al., 2011).

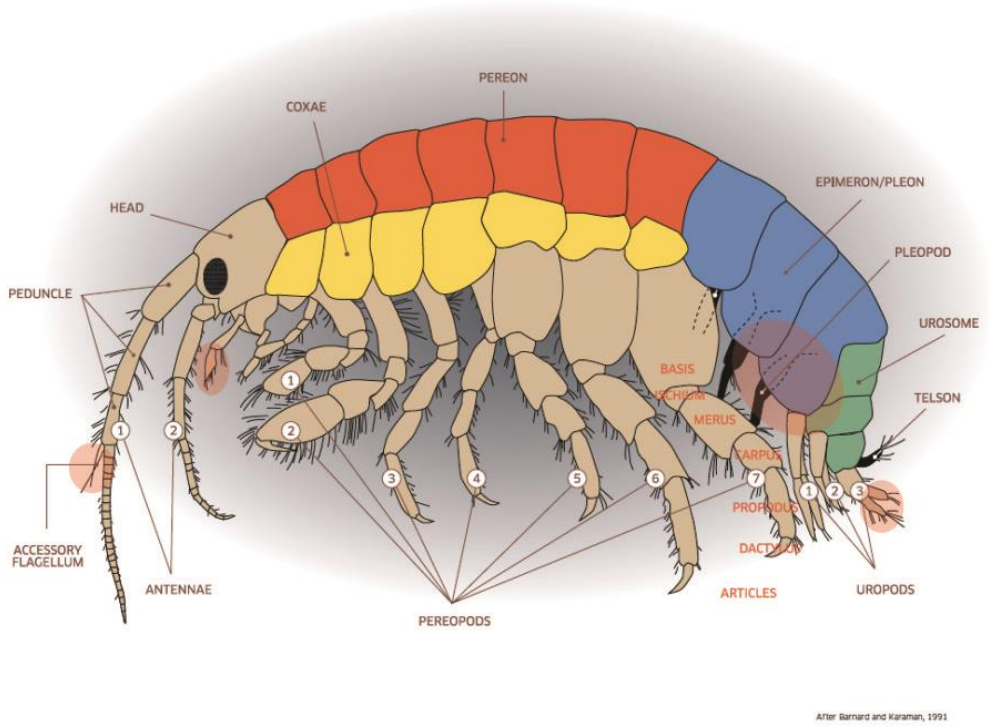


Figure 3 - Model of a general amphipod. Characters used in identification of *Gammarus* highlighted. *Adapted from Ola Reibo based on Barnard & Karaman, 1991.*

G. locusta (Figure 4) (Linnaeus, 1758) belongs to the phylum Arthropoda, class Malacostraca, order Amphipoda and family Gammaridae. The genus *Gammarus* comprises over 100 described species worldwide, with *G. locusta* being one of the most ecologically significant and experimentally valuable (Gerhardt et al., 2011). This species has gained prominence as a model organism in marine ecotoxicology due to its sensitivity to a broad range of environmental contaminants including SIM, its ecological

Figure 4 – *Gammarus locusta.*

relevance and ease of laboratory maintenance (Neuparth et al., 2002; Neuparth et al., 2014; Santos et al., 2016; Ritter & Bourne, 2024).

G. locusta is naturally distributed along the northeastern Atlantic coasts of Europe, from northern Norway to southern Portugal and is particularly abundant in the Sado estuary (Costa & Costa, 1999). It inhabits intertidal and shallow subtidal zones, typically associated with macroalgal beds, seagrass meadows and rocky substrates (Costa et al., 1998; Costa & Costa, 1999; Costa et al., 2004; Whiteley et al., 2011; Neuparth et al., 2014). These structurally complex habitats are ecologically productive yet vulnerable to anthropogenic pressures, including pharmaceutical residues from wastewater effluents and urban runoff (Neuparth et al., 2014; Santos et al., 2016; Neuparth et al., 2020).

As a benthic detritivore and omnivore, *G. locusta* contributes significantly to nutrient cycling, organic matter decomposition and algal biomass regulation (Peschke et al., 2014; De Lange et al., 2006; Bossus et al., 2014). Additionally, it serves as an essential prey species for various fish, birds, seals and invertebrate species, which greatly contributes to its high ecological value in coastal and transitional ecosystems such as estuaries (Costa and Costa, 1999; Whiteley et al., 2011; Vellinger et al., 2012; Peschke et al., 2014).

G. locusta is easy to work with in the laboratory due to its manageable size (≈10–15 mm), robustness and tolerance to varying salinities. It reaches sexual maturity in about 35 - 45 days and females produce multiple broods of 40 - 80 embryos over their lifetime (Neuparth et al., 2022). Its short generation time and high reproductive rate make it ideal for multigenerational and transgenerational ecotoxicological experiments (Costa & Costa, 1999; Neuparth et al., 2002).

Compared to fish and mollusks, *G. locusta* offers several advantages for ecotoxicological studies, including logistical simplicity for breeding and testing, short life cycles, high reproductive output and reliable responses to contaminants, while avoiding many of the ethical concerns associated with vertebrate testing (Ré et al., 2009; Geffard et al., 2010; Whiteley et al., 2011; Peschke et al., 2014; Neuparth et al., 2014).

G. locusta has been extensively employed in ecotoxicological research due to its ecological relevance, sensitivity to contaminants, broad geographical distribution and ease of use in laboratory. These characteristics makes *G. locusta* an ideal organism for the study presented in this thesis.

Understanding the nature and implications of contaminant effects across multiple generations is essential for ecotoxicological assessments. The following section focuses

on the transgenerational effects of SIM in *G. locusta*, highlighting the complexity of inheritance and exposure dynamics.

1.1.4. Transgenerational Effects of SIM in *G. locusta*

In ecotoxicology, it is important to distinguish between multigenerational, intergenerational and transgenerational effects.

Multigenerational studies involve the continuous exposure of several consecutive generations (F0, F1, F2, etc.) to a stressor, such as SIM.

In contrast, **intergenerational effects** are observed in generations that were not directly exposed but still exhibit alterations due to parental exposure to the stressor. For example, when a **pregnant female (F0)** is exposed, F1 embryos are directly exposed during embryonic development within the contaminated F0 female. F2 individuals are also indirectly exposed because their germ cells were formed in the F1 (Figure 5b). In **male-mediated inheritance**, exposure of the F0 male affects the germ cells that generate F1 offspring. As a result, F1 phenotypes reflect indirectly exposure (Figure 5a) (Hime et al., 2021).

True transgenerational effects are defined as phenotypes that occur in generations never directly or indirectly exposed (typically F3 and beyond in females, and F2 onward in males), reflecting heritable changes passed through the germline (Skinner, 2008; Skinner, 2014; Hime et al., 2021).

To better illustrate the conceptual framework of inter- and transgenerational inheritance, an illustrative schematic showing male and female germline inheritance in a general amphipod model is shown in Figure 5 to summarize generational definitions.

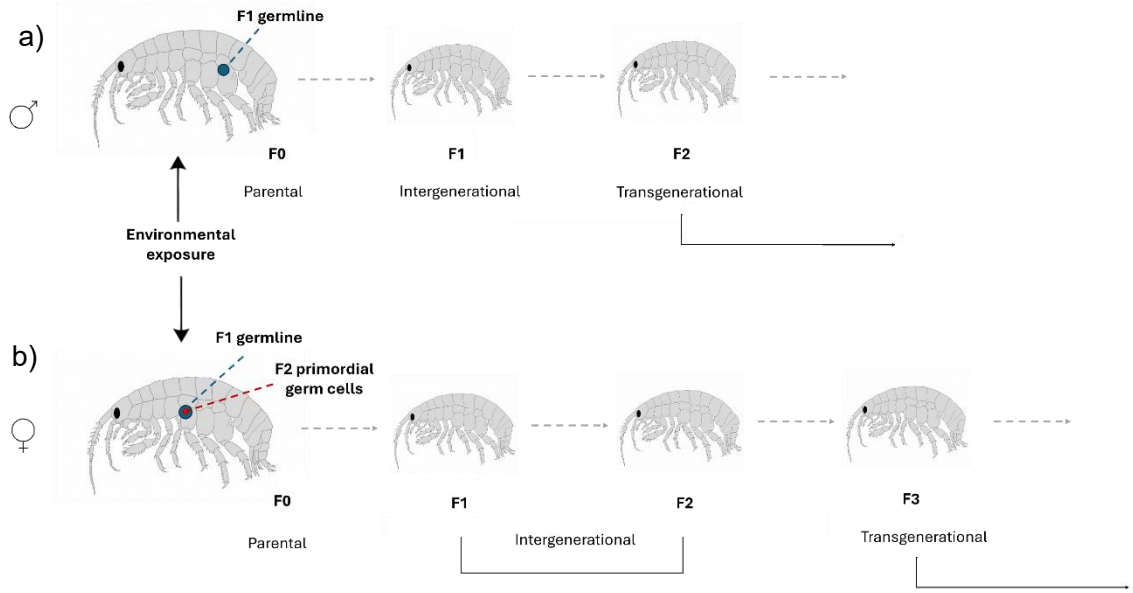


Figure 5 - Illustrative schematic of intergenerational and transgenerational inheritance through the male and female germlines in a general amphipod model.

Transgenerational studies are critical in ecotoxicology because they allow researchers to assess not only the direct effects of contaminants, such as SIM on the exposed parental generation (F0), but also the persistence of these effects in subsequent generations (F1–F3), even in the absence of the original stressor. Such approaches provide insight into heritable physiological, reproductive and molecular disruptions that can have long-term population-level consequences (Bhandari et al., 2015).

In studies by Neuparth et al. with *G. locusta*, SIM was applied to the parental generation (F0) at environmentally relevant concentrations and the F1 through F3 generations were subsequently monitored in SIM-free water to detect any lasting inter- or transgenerational effects. These studies evaluated reproductive output, growth, metabolic pathways, gene expression changes and epigenetic modifications to assess indirect impacts of parental SIM exposure across generations (Neuparth et al., 2014, 2020, 2022; Santos et al., 2016; Alves et al., 2021; Rodrigues et al., 2024). Research has shown that exposure to SIM induces significant physiological, reproductive and molecular disturbances in *G. locusta*, with clear differences between male and female responses (Neuparth et al., 2014, 2020, 2022; Santos et al., 2016; Alves et al., 2021; Rodrigues et al., 2024).

1.2. Omics Approaches in Ecotoxicology and Environmental Risk Assessment

Over the last two decades, omics sciences have emerged as transformative tools in ecotoxicology, providing mechanistic insights into the effects of environmental contaminants that go beyond classical apical endpoints such as mortality, growth and reproductive output (Snape et al., 2004; Misra et al., 2018). The traditional approach of ecotoxicology used only a few biochemical and physiological biomarkers (e.g., oxidative stress enzymes, acetylcholinesterase activity vitellogenin induction) to assess chemical impacts (van der Oost et al., 2003; Newman & Newman, 2014). This regulatory endpoints continue to be important but they do not identify minor initial effects or explain the biological processes that occur at the molecular level (Li et al., 2025).

Omics technologies, by contrast, offer high-throughput, hypothesis-generating approaches capable of simultaneously profiling thousands of molecular features, allowing the discovery of novel biomarkers and previously unrecognized toxicity pathways (Ankley et al., 2006; Garcia-Reyero & Perkins, 2010; Mirbahai & Chipman, 2014). Having high-throughput technologies of varying kinds opened opportunities for these branches of science to produce large-scale data for every biological system (Snape et al., 2004; Li et al., 2025). Omics technologies developed for different systems include **genomics**, which include the study of the architecture, function and evolution of the genome, and thus a blueprint of the genetic makeup and possibilities of the organism. **Transcriptomics** involves the study of all the transcripts in the cell and regulation mechanisms, reflecting and capturing the sign of the molecular responses as they unfold to different stimuli. Focusing on proteins, their quantity, modifications and interactions, and thus their functional dynamics, **proteomics** becomes a vital functional layer. In elucidating small metabolites and their transitions that drive a physiological state, cellular processes and their dynamics, **metabolomics** provides a deeper understanding of the cell. Other branches as **epigenomics** study the reversible, heritable and functional changes that genes display without a sequence alteration to their nuclear DNA (e.g., changes by DNA methylation and histone modifications) (Sauer et al., 2017; Lind & Spagopoulou, 2018). This way, omics technologies have rapidly advanced from exploratory tools into valuable instruments for chemical hazard characterization and ecological risk assessment (OECD, 2023; Villeneuve & Garcia-Reyero, 2010). Increasingly, regulatory bodies recognize that omics data can support hazard identification, elucidation of modes of action and translation of molecular changes into

ecologically relevant endpoints, provided that data are collected and interpreted under standardized frameworks such as the Organisation for Economic Co-operation and Development (OECD) Omics Reporting Framework (OORF) (OECD, 2025; Buesen et al., 2017).

One of the primary uses of omics is biomarker discovery. Biomarkers of exposure indicate that an organism has come into contact with a chemical, whereas biomarkers of effect reflect molecular alterations mechanistically linked to adverse outcomes (Graham Van Aggelen et al., 2010; Gosh & Corneliu Duca, 2021; Shields, 2023). Untargeted omics approaches, especially transcriptomics and metabolomics, are powerful tools for generating new hypothesis. They allow us to spot unexpected patterns and coordinated molecular responses that simpler, single-endpoint tests could easily overlook. These discoveries can then guide the development of more focused assays, useful for biomonitoring and regulatory applications (Song et al., 2023).

A key difficulty in risk assessment is linking early molecular changes to an organism's or population's growth, reproduction, and survival. The Adverse Outcome Pathway (AOP) framework was developed to address this issue, providing a structured way to link molecular initiating events (MIEs) to a series of key events (KEs) and ultimately to adverse outcomes (AOs) that are relevant for risk management (Ankley et al., 2010; Tollefsen et al., 2014). Omics data are particularly well suited to detect MIEs and KEs, offering mechanistic evidence that strengthens causal inference. For example, transcriptomic changes indicating disrupted steroidogenesis or altered energy metabolism can help explain observed reproductive impairments at the organismal level (Neuparth et al., 2022).

Omics methods can help with quantitative risk assessment in addition to qualitative insights. The derivation of transcriptomic points of departure (tPODs), which serve as conservative thresholds for chemical safety assessment, is made possible by transcriptomic concentration–response modeling. Several studies have demonstrated that tPODs often align with traditional apical toxicity thresholds, supporting their use in chemical prioritization, grouping and read-across strategies (Thomas et al., 2013; Costa et al., 2021).

Omics datasets are vast and highly complex, as their high dimensionality makes them prone to false discoveries. On top of that, natural biological differences, such as sex, life stage, or environmental history, add further layers of variability. For non-model species, the lack of complete genomic resources makes it harder to assign functions to genes or map biological pathways (Ekblom & Wolf, 2014; Conesa et al., 2016). Moreover,

integration across different omics layers and with classical endpoints requires advanced bioinformatic and systems biology approaches, which are still under development (Smith et al., 2025).

Even so, real-world applications are already showing how valuable omics can be for a better ecology understanding. Studies on non-model organisms such as earthworms, zebrafish and amphipods have consistently revealed conserved responses to contaminants, including disruptions in oxidative stress regulation, lipid metabolism and endocrine signaling, which correlate with sublethal phenotypes and reproductive impairments (Blalock et al., 2018; Camilo Escobar-Sierra et al., 2024; Russo et al., 2020; López-Landavery et al., 2023, Neuparth et al., 2020). A particularly relevant example is the case of SIM exposure in *G. locusta* females, where transcriptomic analyses revealed transgenerational effects on reproduction, demonstrating how molecular-level data can extend hazard characterization beyond single-generation endpoints (Neuparth et al., 2020b; Neuparth et al., 2022).

In the future, standardizing experimental design, bioinformatic workflows and mechanistic interpretation will be necessary for the regular incorporation of omics into regulatory risk assessment. Linking omics signatures to AOPs, deriving tPODs and embedding molecular key events into population-level or toxicokinetic-toxicodynamic (TK-TD) models represent the next frontier in bridging molecular data to ecological protection goals (Villeneuve & Garcia-Reyero, 2010; Villeneuve et al., 2024).

Among the different omics approaches, transcriptomics has proven especially practical and insightful for ecotoxicology, offering a powerful way to study molecular responses in non-model aquatic organisms.

1.2.1. Transcriptomics

Transcriptomics is an analysis of the total complement of RNA molecules, both messenger RNAs (mRNAs) and non-coding RNAs, within a cell, tissue, or organism under specified conditions (Conesa et al., 2016). By providing a snapshot of cellular activity, transcriptomics enables researchers to investigate how biological systems respond to environmental stressors, including chemical contaminants, at a genome-wide scale (Conesa et al., 2016; Ozsolak & Milos, 2010; Wang et al., 2009). In ecotoxicology, it is valued as a sensitive early-warning tool because molecular alterations often precede observable phenotypic or population-level effects (Villeneuve & Garcia-Reyero, 2010; Head et al., 2024).

According to Lockhart and Winzeler (2000) and Nookaew et al. (2012), transcriptomic profiling was previously dependent on **microarrays**, which enabled the simultaneous quantification of thousands of predefined sets of transcripts. However, these systems had limitations, including a limited dynamic range, low sensitivity for rare transcripts, and a reliance on probe design. The field underwent a revolution with the introduction of **RNA sequencing (RNA-seq),** which made it possible to detect all transcripts, including alternative splice variants and novel genes, objectively and in large quantities. For RNA-Seq analysis, **short-read sequencing** (e.g., Illumina platforms) remains the most widely used approach due to its high accuracy, sensitivity and relatively low cost, allowing robust quantification of transcript abundance and *de novo* transcriptome assembly in species without reference genomes (Grabherr et al., 2011; Martin & Wang, 2011; Neuparth et al., 2020a; Russo et al., 2020; Wang et al., 2023; Tan et al., 2023).

Emerging transcriptomic technologies offer additional resolution. **Long-read RNA-seq** (PacBio, Oxford Nanopore) generates full-length transcript sequences, resolving complex isoforms and splice variants but generally has higher error rates, lower throughput and increased costs (Byrne et al., 2017; Wang et al., 2021). **Single-cell RNA-seq** enables transcriptomic profiling at the level of individual cells, revealing cellular heterogeneity and rare cell populations, although it remains technically challenging and less commonly applied in ecotoxicology because of its high cost (Tang et al., 2009; Svensson et al., 2018; Jovic et al., 2022). In the context of aquatic invertebrates such as *G. locusta*, the absence of reference genomes and standardized dissociation protocols further constrains the feasibility of applying these emerging approaches. Consequently, **short-read RNA-seq remains the most robust and cost-effective choice** for ecotoxicogenomic studies in such species (Orsini et al., 2016; Péden et al., 2019; Neuparth et al., 2020a,b).

Even in the absence of reference genomes, transcriptomic profiling in non-model organisms is possible through *de novo* assembly tools, such as Trinity. These tools allow the reconstruction of transcriptomes from RNA data, enabling high-quality gene expression analysis in species like *G. locusta* without the need for a reference genome (Grabherr et al., 2011; Russo et al., 2020). As a result, transcriptomics has become an increasingly valuable approach in ecotoxicogenomics, especially for uncovering molecular responses to environmental stressors in organisms with limited genomic resources.

In this context, **RNA-seq** data are commonly used to identify differentially expressed genes (DEGs), construct co-expression networks, and perform functional enrichment

and pathway analyses (e.g., Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG)), thus linking transcriptomic expression changes to affected biological processes (Conesa et al., 2016; Langfelder & Horvath, 2008). In RNA-Seq, best practices include careful experimental design, rigorous quality control and accurate annotation strategies, which increase the reliability and biological interpretability of the data (Neuparth et al., 2020a; Burns et al., 2024).

In *G. locusta*, transcriptomic approaches have provided valuable insights into the molecular mechanisms underlying responses to environmental contaminants. For example, research on *G. locusta* females has demonstrated that exposure to contaminants, such as SIM, disrupts neuroendocrine signaling and induces transcriptional alterations consistent with observed effects on reproduction and growth, highlighting the value of transcriptomics for mechanistic interpretation of apical endpoints (Rodrigues et al., 2024; Neuparth et al., 2020b).

Overall, transcriptomics, particularly **short-read RNA-seq**, provides a **practical, sensitive and mechanistically informative tool** that will be used in the current study with *G. locusta* males to assess the transgenerational effects of SIM and for linking molecular perturbations to organismal and ecological outcomes, observed in the Neuparth et al. 2014, 2020 studies. When combined with robust experimental design, appropriate bioinformatic pipelines and mechanistic interpretation, transcriptomics strengthens inference in ecological risk assessment and enhances the potential integration of molecular evidence into environmental decision-making.

1.2.2. Bioinformatics Approaches for Transcriptomic Analysis in the Present Study

In this study, transcriptomics analyses were carried out using bioinformatics tools, which provided the computational capacity to millions of RNA-seq reads into meaningful biological insights. In *G. locusta*, where no reference genome is available, bioinformatics enabled the ***de novo* reconstruction of the transcriptome**, thereby generating a reference framework that made it possible to capture the complete set of expressed genes. This step was essential to establish the molecular basis for subsequent analyses.

Once assembled, the transcriptome was subjected to **functional annotation**, which linked the recovered sequences to putative biological functions and enrolled pathways through comparison with curated protein databases, GO terms and KEGG pathways. Although **annotation in non-model organisms is often constrained by the limited**

representation of invertebrates in these databases, it nonetheless provides crucial insight into gene functions and their involvement in metabolic and signalling pathways.

On this basis, assembled transcriptomes could be further explored to identify **DEGs**, revealing which genes are significantly up- or downregulated in response to environmental stressors. DEG analysis provides a direct connection between molecular perturbations and organismal responses to environmental stressors, particularly when followed by **functional enrichment analyses**. Enrichment approaches highlight biological processes and pathways that are disproportionately affected, such as energy metabolism, reproduction or stress responses, thereby supporting mechanistic interpretations of toxicant effects in an ecotoxicological context.

To streamline these analyses in the present study, a dedicated bioinformatics platform, **TransDegAnalyser**, was developed. This interactive application integrates transcript counts, experimental metadata and functional annotations into a coherent workflow that performs DEG analysis, functional enrichment and data visualization within a single environment. By combining robust statistical analysis with intuitive, user-friendly visualizations, this tool made it easier to explore transcriptomic responses across treatments and generations in *G. locusta* males, giving biological meaning to the annotated DEGs. Beyond enhancing reproducibility and interpretability, **TransDegAnalyser** represents a flexible resource that can be adapted to other ecotoxicogenomic transgenerational studies and species, bridging the gap between raw sequencing data and ecological risk assessment.

2. Objectives

2.1. General Objective

To evaluate the direct and transgenerational effects of SIM exposure in *G. locusta* males, combining transcriptomic profiling, functional interpretation and ecological endpoints and to develop a user-friendly bioinformatics tool for facilitating data visualization, analysis and interpretation linking molecular responses to observed phenotypic outcomes observed in previous studies.

2.2. Specific Objectives

O1 - Construct and validate a male reference transcriptome
Assemble and evaluate the quality and completeness of the *G. locusta* male transcriptome to provide a robust reference for annotation and differential expression analysis.

O2 – Predict protein-coding regions from the reference transcriptome and conduct functional annotation
Identify coding sequences and assign putative biological functions through comparison with curated databases. This annotation not only identifies the coding regions but also provides the foundation for interpreting transcriptomic responses, including GO and KEGG enrichment analyses to highlight biological processes and pathways potentially affected by SIM exposure.

O3 - Develop an interactive tool (TransDegAnalyser) for differential expression analysis and functional enrichment visualization
Implement a Shiny-based application to perform differential gene expression analysis between F0 (parental) and F3 (transgenerational) generations, comparing SIM-exposed and control groups, while integrating functional enrichment results in accessible and reproducible visualizations.

O4 - Integrate molecular and apical endpoints
Link transcriptomic data with observed phenotypic effects observed in previous studies (e.g.: growth and reproduction) to propose mechanistic explanations for the transgenerational impacts of SIM and to identify potential molecular markers of effect.

3. Materials and Methods

This chapter describes the experimental design and the bioinformatic workflow used for RNA-Seq analysis in *G. locusta* males. It includes the generation of RNA-Seq data, transcriptome processing and the development of an interactive tool to perform differential gene expression analysis and integrate functional annotation.

3.1. Experimental Design

The methodology for RNA-seq data generation, including exposure conditions and sample collection, followed the protocol described by Neuparth et al. (2020). Male individuals from the parental generation (F0) were divided into a control group (unexposed) and an exposed group (SIM 320 ng/L). Offspring from the F0 exposed group were reared in SIM-free water for three successive generations (F1–F3) to assess transgenerational effects in F3 generation, while offspring from the F0 control group were maintained under identical SIM-free conditions to serve as controls for the F3 generation.

For RNA-seq analyses, samples were collected specifically from F0 and F3 males. This included four biological replicates for the F0 control (n = 4), three for the F0 exposed group (n = 3) and three replicates each from the F3 control and F3 transgenerational groups (n = 3 per group), totaling 13 samples for sequencing.

All experimental procedures were conducted under controlled laboratory conditions, including a semi-static aquarium system with filtered seawater and sediment, temperature maintained at 20°C, a 16:8 h light:dark photoperiod and regular monitoring of water quality parameters, as detailed in Neuparth et al. (2020). RNA samples were processed for library preparation using Illumina TruSeq kits, following the manufacturer's protocols. Following library preparation, RNA-Seq was performed commercially at Novogene (Hong Kong) on an Illumina HiSeq-2500 platform with paired-end 150 bp (PE150) reads, targeting over 30 million reads per sample. This workflow generated high-quality raw RNA-seq data in FASTA format, with reads separated into forward and reverse pairs (paired-end), which were subsequently used for downstream gene expression analyses. Forward read files end with *1 (e.*g., S320_1G_1_1.fq) and reverse read files with *2* (e.g., S320_1G_1_2.fq). The experimental design is shown in Figure 6.

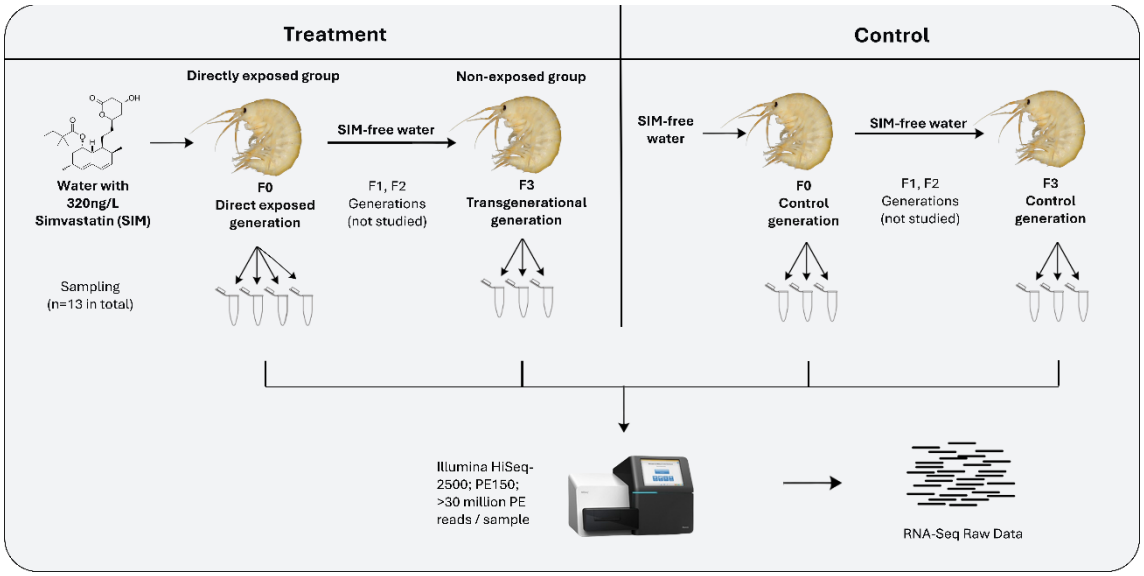


Figure 6 - Experimental setup for assessing direct and transgenerational effects of SIM in *G. locusta* males via RNA-Seq. In total, 13 samples from the F0 and F3 generations of both groups were collected and processed for RNA-Seq analysis.

3.2. Bioinformatic Approach for RNA-seq Data Analysis

RNA-seq data from *G. locusta* samples were analyzed using a comprehensive bioinformatic pipeline designed to transform raw sequencing reads into biologically interpretable results (Conesa et al., 2016; Pola-Sánchez et al., 2024). The workflow encompasses multiple stages, including pre-processing of raw RNA-Seq reads, transcriptome assembly, Open Reading Frame (ORF) prediction, functional annotation of the longest ORFs and the development of an interactive Shiny application (TransDegAnalyser) for the identification and exploration of differentially expressed genes and their associated biological pathways.

Computationally demanding steps, including transcriptome assembly, backmapping of reads to transcriptome and ORF prediction were executed on a high-performance computing (HPC) environment with up to 30 CPU threads and 500 GB of RAM to ensure efficiency and reproducibility. The Galaxy Europe platform (usegalaxy.eu), which offers a cloud-based environment with standardized tools and computational resources for repeatable analyses, was used to perform functional annotation. The development and execution of the Shiny application for differential expression analysis were carried out locally on a personal workstation (HP laptop) running Windows 11, equipped with an Intel® Core™ i7 processor, 16 GB of RAM, and 475 GB SSD storage. The overall

workflow is summarized in **Figure 7**. All the scripts used to perform this bioinformatic approach are available in Github and can be accessed through this link: https://github.com/abiliaframboesa/Thesis_RNASeq.

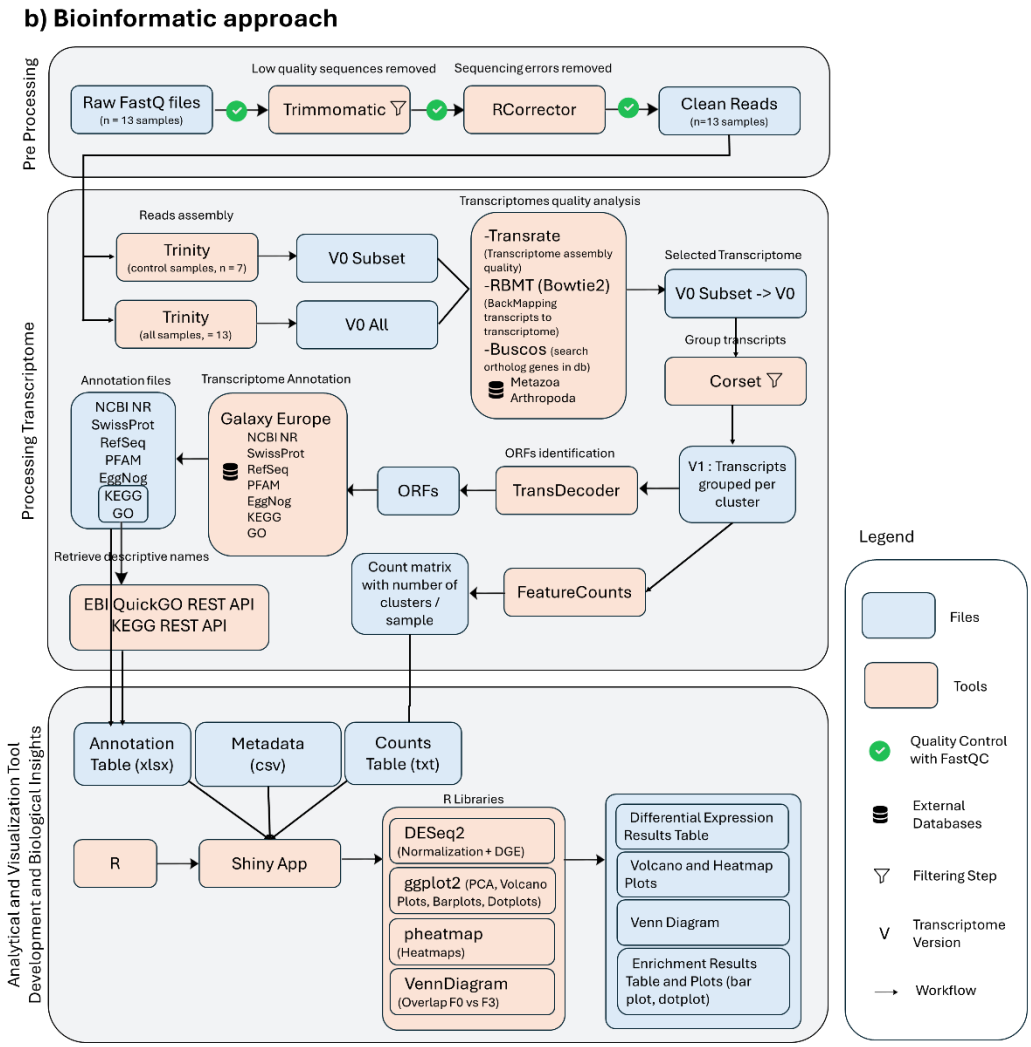


Figure 7 - Bioinformatic workflow for RNA-seq data analysis of *G. locusta* samples. Pre-processing, *de novo* transcriptome assembly, functional annotation and the creation of an interactive tool for identifying and visualizing genes with differential expression and the pathways that are linked to them are all included in the workflow. Each step's software is displayed in orange boxes, input/output files are displayed in blue, and green checkmarks indicate quality control procedures.

3.2.1. Pre-processing and Quality Control

The first step in the pipeline focused on ensuring high-quality input data for downstream analyses. Raw RNA-seq FASTA files (eg: S320_1G_1_1.fq) were subjected to a rigorous pre-processing and quality control workflow, including initial quality assessment, adapter and low-quality base removal and sequencing error correction. These steps are essential

to maximize read quality, reduce technical artefacts and improve the accuracy of *de novo* transcriptome assembly, particularly in non-model organisms.

3.2.1.1. Initial Quality Assessment

Raw reads were first evaluated using **FastQC (v.0.11.8)** (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), a standard tool for high-throughput sequencing quality control. Key metrics assessed included per-base quality scores, GC content, sequence length distribution, duplication levels and the presence of adapter contamination (Andrews, 2010). Early quality assessment is critical to detect common Illumina sequencing artefacts, such as 3′ quality decay and residual adapter sequences, which can lead to incorrect base calls, misaligned reads and artificial contigs, ultimately affecting the accuracy of transcriptome assembly and downstream gene expression analyses (Minoche et al., 2011).

3.2.1.2. Quality Trimming and Filtering

Adapter removal and trimming of low-quality bases were performed using **Trimmomatic (v.0.38)** (Bolger et al., 2014), chosen for its robust sliding-window algorithm and widespread adoption in RNA-seq workflows. The following parameters were applied:

- **LEADING:5** and **TRAILING:5** – remove bases with Phred quality <5 from read ends.
- **SLIDINGWINDOW:4:5** – dynamically trim reads when the average quality within a 4-bp window falls below Phred 5.
- **MINLEN:36** – discard reads shorter than 36 bp after trimming.
- **ILLUMINACLIP** – remove adapter sequences using the built-in TruSeq3 database.

Post-trimming read quality was reassessed using **FastQC**. These parameters were consistent with previously published pipelines for *G. locusta* females (Neuparth et al., 2020), ensuring comparability.

3.2.1.3. Error Correction

Sequencing errors were corrected using **Rcorrector (v.1.0.3)** (Song & Florea, 2015) with default parameters. Rcorrector employs **k-mer frequency** analysis via a De Bruijn graph approach to distinguish between **trusted** and **erroneous** k-mers: those that appear frequently in the dataset are considered reliable, while rare k-mers are flagged as potential sequencing errors and corrected by aligning them to the most probable high-frequency counterparts (Song & Florea, 2015; NIH HPC, 2015). Because sequencing errors can result in transcripts that are fragmented or misassembled, this step is especially crucial. Error correction has been shown to significantly improve transcriptome contiguity and reduce assembly artefacts, especially in *de novo* assemblies from non-model organisms (MacManes & Eisen, 2013; MacManes, 2014; Song & Florea, 2015).

The excellent quality of the processed reads was validated by a final **FastQC** run. The high-quality FASTA files produced by these quality control procedures (shown by the green checkmarks in Figure 7) were subsequently utilized for transcriptome assembly. These programs were run with scripts present in https://github.com/abiliaframboesa/Thesis_RNASeq.

3.2.2. Transcriptome Assembly - V0

Due to the absence of a reference genome for *G. locusta*, a *de novo* transcriptome assembly was performed with the pre-processed FASTA files using **Trinity v2.15.1** (following the methodology described by Haas et al., 2013).

This initial assembly (Version -V0) provided a provisional reference framework to which RNA-seq reads from individual samples could be mapped, enabling preliminary assessment of transcript representation, read distribution and overall assembly quality. Due to possible high inter-individual variability across samples, including heterozygosity and condition-specific expression profiles, a single assembly may not fully capture the transcriptomic diversity or could introduce artefacts such as chimeric contigs or inflated redundancy (Ojeda et al., 2019; Morteza Sheikh-Assadi et al., 2022).

To address these challenges, two V0 assemblies were generated to explore different strategies for representing the transcriptome, allowing informed selection of the most reliable reference for downstream analyses:

- **V0 All:** including all 13 samples, to capture the full transcriptomic diversity across conditions.
- **V0 Subset:** using only control-condition samples (n = 7), representing a simpler and cleaner, baseline transcriptome.

For each assembly, reads from the selected samples were first combined into two paired files, corresponding to forward (ending in 1.cor.fq.gz) and reverse reads (ending in 2.cor.fq.gz). For the V0 All assembly, reads from all 13 samples were concatenated, while for the V0 Subset assembly only the seven control-condition samples were combined.

For example, for the full assembly (V0 All):

cat All_*_1.cor.fq.gz > All_paired_1.cor.fq.gz

cat All_*_2.cor.fq.gz > All_paired_2.cor.fq.gz

These paired files were then used as input for Trinity, which was executed with strand-specific parameters (**SS_lib_type RF**), a minimum **contig length of 300 bp**, and ***in silico* read normalization** to optimize both assembly quality and computational efficiency. The assembly was performed on a HPC cluster using **30 CPU cores of Intel® Xeon® E5-2650 v2 processors (2 sockets, 8 cores per socket, 2 threads per core, 2.60 GHz)** and **500 GB of RAM memory**, following standard Trinity best practices (Haas et al., 2013). These parameters were consistent with previously published pipelines for *G. locusta* females (Neuparth et al., 2020), ensuring comparability and reproducibility.

These assemblies correspond to **Version 0 (V0) of the transcriptome**. Both assemblies were subsequently evaluated for quality using TransRate (Smith-Unna et al., 2016), BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015; Manni et al., 2021) and Bowtie2 (Langmead & Salzberg, 2012). The results of these assessments were used to guide the selection of the V0 assembly that would serve as the reference for generating the refined Version 1 (V1) transcriptome for downstream differential gene expression and functional analyses. The script used for transcriptome assembly, along with all other steps described in this methodology, is available on GitHub, with the link provided in https://github.com/abiliaframboesa/Thesis_RNASeq.

3.2.2.1. Assembly Quality Assessment

To evaluate the structural quality, completeness and representativity of the *de novo* transcriptome assemblies (V0 All vs. V0 Subset), a multi-step quality assessment was conducted. This included structural scoring with TransRate (Smith-Unna et al., 2016), completeness evaluation using BUSCO (Simão et al., 2015; Manni et al., 2021) and representativity assessment through read backmapping with Bowtie2 (Langmead & Salzberg, 2012). The results of these analyses were used to determine which V0 assembly would serve as the reference for generating the refined, final Version 1 (V1) transcriptome for downstream differential gene expression and functional analyses (Haas et al., 2013; Smith-Unna et al., 2016).

3.2.2.1.1. TransRate Metrics

Assembly structural quality was evaluated using **TransRate v1.0.3** (Smith-Unna et al., 2016), a widely used reference-free tool for *de novo* transcriptome assessment. In this analysis, TransRate was run using **default parameters without input reads**, focusing exclusively on contig-level structural metrics rather than read evidence to provide an unbiased, reference-free evaluation of assembly architecture. The same parameters were applied in the previously published pipeline for *G. locusta* females (Neuparth et al., 2020), ensuring comparability.

Metrics computed included total number of transcripts, N50, N90, largest transcript length, number of transcripts over 1 kb, total assembled bases and GC content. These parameters provide information on contig length distribution, assembly completeness, redundancy and base composition. For example, N50 and N90 indicate contig continuity, while total transcript counts reflect potential redundancy or fragmentation. GC content can reveal sequence composition biases or potential contamination (Li & Dewey, 2011; Vijay et al., 2013).

3.2.2.1.2. BUSCO Completeness Analysis

To evaluate the completeness of the assembled transcriptomes with respect to expected gene content, **BUSCO v5.5.0** (Simão et al., 2015; Manni et al., 2021) was employed in **transcriptome mode (-m tran)**, which is optimized for assemblies derived from RNA-Seq data. BUSCO assesses completeness of the transcriptome based on the detection of highly conserved single-copy orthologs, providing a biologically meaningful estimate

of transcriptome coverage along with the genes present in different databases selected for the analysis.

Two lineage datasets of increasing taxonomic breadth were selected:

- **arthropoda_odb10** – specific to arthropods, allowing detection of conserved genes and potential lineage-specific losses;
- **metazoa_odb10** – a broader dataset providing a general baseline for transcriptome completeness across metazoans (Ungaro et al., 2017; Waterhouse et al., 2017)**.**

BUSCO was executed in **offline mode** to ensure reproducibility and to avoid potential discrepancies due to database updates. Thirty CPU threads (-c 30) were allocated to accelerate computation.

The percentage of complete (single/duplicated), fragmented and missing BUSCOs was reported and used in conjunction with structural metrics to assess assembly completeness.

3.2.2.1.3. Read BackMapping (Bowtie2 + Samtools)

To evaluate the representativity and structural usability of the transcriptome assemblies, **read backmapping** was performed using **Bowtie2 v2.3.5** (Langmead & Salzberg, 2012). This approach estimates the proportion of RNA-seq reads aligning to the assembled transcripts, serving as a proxy for transcriptome completeness, redundancy and the ability of the assembly to represent transcripts across all experimental conditions (Haas et al., 2013; MacManes, 2014).

Bowtie2 indices were first generated for each assembly (V0 Subset and V0 All) using the **bowtie2-build** function. These indices create an index of the transcriptome that allows Bowtie2 to efficiently and accurately align RNA-seq reads during the backmapping step.

Subsequently, individual samples were mapped to the corresponding transcriptome assembly using the parameters **paired-end mode**, --no mixed --no-discordant --end-to-end --all --score-min L,-0.1,-0.1, which allow reporting of multiple valid alignments and are recommended for transcript clustering (Davidson & Oshlack, 2014).Also **30 CPU threads** allocated per run to ensure efficient computation.

Three mapping strategies were performed:

1. All concatenated cleaned reads from all experimental conditions (n = 13) mapped to the global assembly (V0 All).
2. Only concatenated control-condition reads (n = 7) mapped to the control-derived assembly (V0 Subset).
3. Individual mapping of each sample (n = 13) to the control-derived assembly (V0 Subset) to assess whether the assembly represents transcriptomic diversity across conditions.

Alignments were generated in **Sequence Aligment Map (SAM) format**, then converted to **Binary Alignment Map (BAM)**, sorted and indexed using **Samtools v1.13** (Li et al., 2009; Danecek et al., 2021). Mapping statistics were generated using the **flagstat** command of Samtools, reporting the proportion of reads successfully aligned to each transcriptome, providing a quantitative measure of assembly representativity.

The results from this backmapping step, combined with structural metrics from TransRate and completeness metrics from BUSCO, informed the selection of the final **Version 0 (V0) transcriptome** for downstream functional annotation and differential expression analyses.

3.2.3. Final Reference Transcriptome (V1)

After selecting the V0 reference transcriptome assembly, **Corset v1.09** (Davidson & Oshlack, 2014) was used to cluster transcripts into putative gene groups, thereby reducing transcript redundancy and enabling gene-level analyses.

Corset was then executed with **biological replicate grouping** enabled (-g parameter) to group biological replicates (1,1,1,1,2,2,2,…), which clusters transcripts based on shared read support across replicate groups rather than individual samples, reducing transcript redundancy and possible technical noise enabling gene-level analyses.Next, the **fetchClusterSeqs.py** script, included in the Trinity package, was used to extract representative sequences for each cluster from the selected transcriptome (V0 Subset), generating the reduced reference transcriptome **(V1 Transcriptome)**.

This step produced the V1 transcriptome, which served as the reference transcriptome for downstream analyses. Specifically, it was used to generate the count matrix of transcript clusters per sample for differential expression analysis, as well as for ORF prediction and subsequent functional annotation. The script for this analysis is present in https://github.com/abiliaframboesa/Thesis_RNASeq.

3.2.4. Generation of Count Matrix from V1 Transcriptome for Differential Expression Analysis

With the final V1 reference transcriptome established, the next step was to quantify transcript cluster expression levels across all samples. For this purpose, raw read counts per transcript cluster were generated to produce a count matrix, which serves as the primary input for differential expression analysis.

To generate this matrix, two key files derived from the Trinity assembly were required:

1. **Trinity gene-transcript map** – This file, produced by Trinity during assembly, lists which transcripts (isoforms) belong to each gene cluster, allowing grouping of isoforms into single biological units.
2. **GTF (Gene Transfer Format) annotation file** – Generated from the gene-transcript map, the GTF file provides the coordinates and structure of each transcript in a format compatible with counting tools.

These files were used in **featureCounts v2.0.2** (Liao et al., 2013) to quantify how many RNA-seq reads map to each transcript cluster present in V1 Transcriptome. This method correctly assigns reads that align to multiple isoforms within the same cluster and generates a count matrix of each cluster for all experimental samples.

The resulting matrix was then used as an input of the app for differential expression analyses. The script used to run this process is presented in https://github.com/abiliaframboesa/Thesis_RNASeq.

3.2.5. ORF Prediction with TransDecoder

To enable functional annotation at the protein level, ORFs were predicted from the refined V1 transcriptome using **TransDecoder**, following the author guidelines (v5.5.0; Haas et al., 2013; https://transdecoder.github.io/). In *de novo* transcriptome analysis, ORF prediction is an essential step because it identifies putative protein-coding regions, which facilitates function annotation, pathway mapping, and subsequent homology searches.

TransDecoder was executed in **LongOrfs** mode to extract candidate coding regions, retaining the longest peptide per transcript as the most likely protein product. This was the last step ran in HPC computing environment. No additional pre-annotation filtering

was applied at this stage to maximize sensitivity. The resulting predicted peptide sequences were counted and then splited into smaller chunks using Python scripts (see attachment) to enable efficient parallelized processing in subsequent BLAST similarity searches and functional annotation with EggNOG-mapper.

3.2.6. Functional Annotation of Predicted ORFs

To functionally annotate the predicted protein-coding sequences from the V1 transcriptome, a multi-step homology approach and orthology-based assignment was applied using **Galaxy Europe (Afgan et al., 2018)**. The analysis leveraged chunked ORF sequences to enable parallelized processing and reduce computational load.

Step 1: Sequence Similarity Searches

Protein sequences, split into chunks for parallelized processing (see https://github.com/abiliaframboesa/Thesis_RNASeq), were aligned against major reference databases - NCBI NR (2023-07-28) and UniProtKB/SwissProt (2023-03), using DIAMOND **v2.1.13+galaxy0** BLASTP (Buchfink et al., 2014; Buchfink et al., 2021).

To maximize the reliability of homology detection, searches were restricted to crustaceans **(--taxonlist 6656)** and weak or partial matches were filtered out by applying stringent thresholds on **E-value (≤0.001)**, **sequence identity (≥30%)** and **minimum bidirectional coverage (≥50%**) (Altschul et al., 1990). **Sensitive** alignment mode was chosen to capture distant homologs, while still maintaining computational efficiency (Buchfink et al., 2014). To increase the likelihood of finding several biologically relevant homologs, up to **25 hits** were initially kept for each query; however, only the best hit was later used for downstream functional annotation (see https://github.com/abiliaframboesa/Thesis_RNASeq). Low-complexity regions were **masked** to prevent spurious alignments and improve the reliability of homology-based functional annotation (Li & Godzik, 2006) and the **BLOSUM62 scoring matrix** was applied as a robust standard for detecting homologous proteins at moderate evolutionary distances (Henikoff & Henikoff, 1992; Buchfink et al., 2014).

Step 2: Orthology-based Functional Annotation

Orthology-based annotations were performed on the predicted ORFs using **eggNOG-mapper** (Galaxy v2.1.8; Huerta-Cepas et al., 2017) to assign orthologous groups, functional categories, GO terms and KEGG pathways. The analysis was restricted to the **Arthropoda lineage (taxid 6656)** to reduce spurious assignments from distant taxa, while **sensitive alignment mode** ensured detection of more divergent homologs. Only hits with ≥50% **query coverage**, ≥30% **sequence identity** and an **E-value** ≤0.001 were retained to ensure high-confidence functional assignments.

Since GO terms and KEGG pathways were initially provided as identifiers (e.g., GO:0008150 for GO terms and ko04080 for KEGG pathways), a custom Python script was implemented to retrieve descriptive names via the **EBI QuickGO REST API** (The Gene Ontology Consortium, 2018; EMBL-EBI, 2025) and the **KEGG REST API** (Kanehisa et al., 2020; *KEGG API*, 2024) (see https://github.com/abiliaframboesa/Thesis_RNASeqt).

This step converts identifier-based annotations into interpretable functional information for each transcript cluster, including:

- **GO annotation** – associated **biological processes (BP), molecular functions (MF) and cellular components (CC)**.
- **KEGG pathway annotation** – indicates the **biological pathways** (metabolic, signaling, etc.) in which the gene/protein participates.

These functional annotations enhance biological interpretability and provide the basis for downstream enrichment analyses.

The outputs from BlastP and eggNOG-mapper, including GO terms and KEGG pathways, were then integrated to produce a **comprehensive functional annotation** Excel file for the V1 transcriptome through Python scripts (see attachment). These functional annotations were subsequently used as input for the user-friendly toolbox developed in this study, supporting a biological interpretation for the differentially expressed genes.

3.2.7. Development of an Analytical and Visualization App (TransDegAnalyser) to Perform Differential Gene Expression Analysis and Functional Interpretation Visualization

To facilitate the exploration and interpretation of the transcriptomic dataset, a dedicated analytical and visualization application, **TransDegAnalyser**, was developed. This platform integrates transcript counts, experimental metadata and functional annotations

to provide a complete workflow for differential gene expression analysis and functional interpretation visualization. Through an interactive and user-friendly interface, the application guides the user from raw input files to statistical analyses and biological insights, combining reproducibility with accessibility. The following subsections describe the technical implementation of the application, the required input data and the underlying computational framework.

3.2.7.1. R and Shiny Framework

The platform **TransDegAnalyser** was specifically designed to analyse differential gene expression across generations and treatments in the transcriptome of male *G. locusta* and to interpret the biological functions of significant genes.

The implementation of **TransDegAnalyser** was developed in R programming language carried out in **RStudio (v.4.4.0)** (R Core Team, 2025) using the **Shiny framework (v.1.11.1)** (*Web Application Framework for R*, 2021), which enables the development of interactive applications. Shiny allows users to explore and visualize results without requiring direct execution of R code, **integrating a robust backend for statistical analysis** with a **modern, user-friendly frontend for data visualization and interpretation of the results.**

The backend statistical analyses, including normalization, data manipulation and differential gene expression testing, were performed with **DESeq2 package (v.1.46.0)** (Love et al., 2014), while visualization of the results and user interactivity were supported by a combination of specialized R packages.

To facilitate a dynamic exploration and an interpretation of transcriptomic data, the application was designed with an **interactive interface** that allows users to:

- Explore differential expression results dynamically.
- Filter by generation, treatment group or significance thresholds.
- Visualize top enriched GO terms and KEGG pathways.

The choice of **R and Shiny** was motivated by their **widespread adoption in bioinformatics**, their compatibility with **high-dimensional transcriptomic datasets** and the availability of **extensive packages** for data manipulation, statistical analysis and visualization of the results (Kasprzak et al., 2020).

The application integrates **three input files** - counts, metadata and functional annotation - which serve as the basis for performing normalization, dimensionality reduction (PCA), differential expression analysis, transgenerational gene comparisons and functional enrichment analyses of GO terms and KEGG pathways. Results are shown as **plots and tables**, enabling dynamic and user-friendly exploration of expression patterns and their biological significance.

Specifically, the packages used were package **ggplot2 (v.3.5.2)** (Wickham, 2016) that was used for static visualizations such as PCA, volcano plots and barplots, while **ggrepel (v.0.9.6)** was employed to improve labeling of points in these plots. **Pheatmap (v.1.0.13)** supported clustered heatmaps of differentially expressed genes, while **DT (v.0.33)** provided interactive tables of the results from the specific analysis. Transgenerational comparisons were implemented with **VennDiagram (v.1.7.3)** and **grid (attached to R, v.4.4.0)** and functional annotation was handled via **readxl (v.1.4.5)**. Interface customization and usability improvements were accomplished with **shinythemes (v.1.2.0)** and **shinycssloaders (v.1.1.0)**. The code used to create the app is available in the Github and can be accessed with this link: https://github.com/abiliaframboesa/Thesis_RNASeq.

3.2.7.2. Input Data

The application integrates **three types of input files**, each serving a distinct role in the analysis workflow:

1. Counts table

- TXT file containing raw read counts per transcript cluster per sample (counts matrix), derived from the V1 reference transcriptome following the clustering (Corset) and quantification (featureCounts) pipeline described in Section 3.2.4. An additional column “Protein” was added using a custom script, providing the corresponding protein identifier from NCBI for each cluster (see attachment).

- Columns include:
    - **Cluster**: transcript cluster identifier (e.g., Cluster-60951.1)
    - **Protein**: corresponding protein identifier from NCBI NR (e.g., XP_045121053.1)

- **Sample columns**: one column per sample (e.g., S320_1G_1_sorted), containing raw read counts per cluster

- The counts table was used directly to construct a DESeq2 dataset via DESeqDataSetFromMatrix(), which was subsequently employed to perform differential expression analyses as it allows direct mapping of clusters to annotated proteins.

2. Metadata table

- A CSV file providing detailed experimental information for each sample, allowing the mapping of count data to biological and experimental factors.
- Columns include:
  - **sample_id**: matches exactly the column names in the counts table (e.g., S320_1G_1_sorted)
  - **group**: experimental condition (treatment: S320 or control: SC)
  - **generation**: experimental generation (parental: F0 or transgenerational: F3)
  - **combined_group**: combined factor of generation and treatment (e.g., F0_S320), useful for combined comparisons
- The metadata table was used to construct the design formula in DESeq2 (e.g., ~ group or ~ generation + group) when creating the DESeqDataSetFromMatrix() object. It was also used to subset samples for specific comparisons (e.g., F0_S320 vs F0_SC) and generate group-level visualizations such as PCA plots and heatmaps, linking expression patterns to experimental conditions.

3. Annotation table

- A XLSX file providing functional annotation data from EggNOG-mapper (Huerta-Cepas et al., 2017), including GO terms and KEGG pathways. This file was created through python scripts which are present in the attachment.
- Columns include:
  - **Gene**: cluster identifier (e.g., Cluster-60951.1)

- **EggNOG**: ortholog description
- **Category**: COG (Cluster of Orthologous Genes) functional category
- **GO**: Gene Ontology terms (merged BP, MF and CC)
- **KEGG_pathway**: KEGG pathway identifiers (KO codes)
- **EC_number**: Enzyme Commission (EC) numbers

- The annotation table provides reference for functional interpretation and enrichment analyses of DEGs.

3.2.7.3. Data Preprocessing and Normalization

Upon input on the app, counts and metadata files were validated and filtered to remove incomplete or inconsistent entries. Genes with low total counts across all samples (<10 reads) were excluded to improve DEG detection sensitivity and reduce noise (Sha et al., 2015).

Normalization was performed using **DESeq2's variance-stabilizing transformation (vst())** (Love et al., 2014), which accounts for differences in total read counts and stabilizes variance across the range of counts, reducing the impact of highly expressed genes. The resulting normalized data were used for downstream analyses, including PCA (Principal Component Analysis), heatmaps and functional enrichment.

3.2.7.4. Principal Component Analysis

To summarize the high-dimensional gene expression data and explore overall patterns of variation across samples, **PCA** was performed on variance-stabilized counts generated by DESeq2. By converting correlated gene expression variables into a smaller collection of uncorrelated principal components, each of which accounts for a portion of the samples' overall variance, PCA lowers the dimensionality of the dataset. This approach enables the detection of clustering patterns related to treatment and generation. PCA coordinates were obtained using the **plotPCA()** function, while visualizations were created with **ggplot2**.

3.2.7.5. Differential Expression Analysis

Differential expression analyses were conducted using DESeq2 for within-generation comparisons. Genes showing statistically significant differences in expression between treatment and control groups within each generation (F0 and F3) were identified in these analyses. These results were interpreted using the following key metrics:

- **Log2 fold change (log2FC):** shows how much and in which direction gene expression varies between conditions. Genes with an absolute log2FC above a predefined threshold are considered to show biologically meaningful changes. (Positive values = upregulation; negative values = downregulation).
- **Adjusted p-value (padj):** calculated using the Benjamini-Hochberg method to control for multiple testing. A padj below the chosen cutoff indicates that the observed difference is statistically significant.

Users can select generations and treatments through interactive UI elements (**uiOutput()** function) and significance thresholds (log2FC and padj) are adjustable. Clusters meeting the thresholds of log2FC ≥ 1 and padj ≤ 0.05 were considered significantly differentially expressed. Results were visualized as **volcano plots** (ggplot2) and presented in **interactive tables (renderDT())**, allowing simultaneous assessment of both the magnitude and statistical significance of gene expression changes, facilitating the identification of the most biologically relevant DEGsand providing insight into the molecular responses associated with the experimental treatment (Love et al., 2014; McDermaid et al., 2018). Volcano plots and interactive filters enable identification of the most biologically relevant genes per comparison (McDermaid et al., 2018).

All DESeq2 steps were fully reproducible in the app using:

- **DESeqDataSetFromMatrix()** to construct the dataset
- **DESeq()** for statistical modeling
- **results()** to extract differential expression metrics
- Dynamic filtering based on user-defined thresholds

3.2.7.6. Heatmap Visualization

Significant genes were visualized using heatmaps generated with **pheatmap package** via **pheatmap()**. Counts were row-normalized (scale = "row") to emphasize relative expression differences across samples, ensuring that variation patterns could be

compared independently of absolute expression levels (Bioinformatics at the National Cancer Institute, 2025). Sample annotations indicating treatment group and generation were added to facilitate interpretation of the heatmap. Users could dynamically select generations, treatments and significance thresholds. This approach allows the identification of co-regulated gene clusters, revealing groups of genes that show similar expression patterns across conditions and highlights treatment- or generation-specific expression profiles, thereby providing insights into coordinated molecular responses and potential functional relationships among genes (Pola-Sánchez et al., 2024).

3.2.7.7. Transgenerational Comparison

To investigate **potential direct (F0)** and **transgenerational (F3) effects** of SIM exposure, genes identified as significantly differentially expressed in the F0 and F3 generations when compared to control generations were identified and compared. This comparative analysis aimed at determine which transcriptional changes were specific to a single generation and which were consistently regulated across generations, providing insights into potential heritable or epigenetic responses. Users could select thresholds and choose which genes to display (F0, F3 or intersection).

To illustrate the overlap and uniqueness of DEGs across generations, Venn diagrams were created using the **VennDiagram package** and the **VennDiagram::draw.pairwise.venn()** function. Both generation-specific and gene sets that might be regulated across generations can be quickly identified using this diagram.

In addition, **summary tables (renderDT())** cataloging the unique and shared genes were created. These tables serve as a structured resource for downstream analyses, including functional enrichment of GO terms and KEGG pathways and facilitate the identification of candidate genes potentially involved in direct (F0) and transgenerational (F3) responses. By integrating visual and tabular representations, this approach provides an overview of how gene expression patterns may be transmitted over generations due to SIM exposure in F0 male individuals. These outputs were fully reproducible and could be downloaded for downstream analyses.

3.2.7.8. Functional Enrichment Analysis

To gain biological insights from the DEGs, functional enrichment analysis was performed by mapping significant genes to a curated annotation table described in Methods section

3.2.6. The annotation table file contained EggNOG functions, COG categories, GO terms, KEGG pathways and enzyme codes. Since not all genes were annotated, because *G. locusta* lacks a reference genome and the presence of invertebrates in databases is small, only DEGs with available functional annotations were considered for enrichment analysis. This approach allowed the identification of GO terms associated with biological processes, molecular functions cellular components and KEGG pathways associated with metabolic pathways that are overrepresented in the DEG set relative to the full transcriptome.

To perform enrichment analysis, associated functional terms for each DEG were extracted andterm frequencies calculated. Results of this analysis were visualized using barplots generated with **ggplot2**, highlighting the most enriched functional categories and enabling rapid identification of key biological mechanisms. Comprehensive tables listing DEGs and their functional annotations facilitated detailed exploration and interactive filtering by generation, significance threshold or functional category in the TransDegAnalyser application.

In summary, this functional enrichment framework links differential expression patterns to biological processes and pathways, supporting mechanistic interpretation and identification of candidate pathways potentially involved in physiological and transgenerational responses to SIM exposure.

4. Results and Discussion

This chapter presents the results obtained from RNA-seq analysis of *G. locusta* males and discusses the DEGs observed in both F0 (exposed) and F3 (transgenerational) generations compared to their respective controls, along with their potential biological implications. High-quality RNA-seq data were successfully generated for all samples and processed through the bioinformatic workflow developed in the current study described in Section 3, including pre-processing, quality control, *de novo* transcriptome assemblies, transcript cluster counts matrix, ORF prediction and functional annotation. DEGs analyses were performed using the R/Shiny application also developed for this study, which used DESeq2 to identify significantly altered transcripts and provided an interactive interface to explore gene expression patterns across generations and treatments. Functional enrichment analyses of GO terms and KEGG pathways were then performed using DEGs identified by the application, with functional annotations provided via a pre-prepared input file, enabling mechanistic insights into the biological processes and pathways affected by SIM and its potential transgenerational impacts.

4.1. Rationale for Focusing on the SIM impacts on Male *G. locusta* and perform a Transcriptome Assembly

In the previous multigenerational and transgenerational study by Neuparth et al. (2020), SIM at 320 ng/L caused the most pronounced adverse effects on ecologically relevant endpoints, including growth and reproduction in *G. locusta* (males and females). That study also reported significant transcriptomic alterations in females exposed to SIM. Building on those findings, the present work applied RNA-seq analyses to males from the F0 generation and transgenerational lineage (F3) from 320 ng/L SIM treatment, as this concentration produced the clearest and most consistent effects in key apical endpoints in the Neuparth et al. (2020) study. The approach presented in this study allowed for a more targeted and mechanistic investigation of the molecular basis of SIM-induced effects in *G. locusta* males across generations, an aspect that had not been previously studied.

The exclusive focus on males addressed an existing knowledge gap of previous SIM studies in *G. locusta* that largely focused on female-specific endpoints (Neuparth et al.,

2020). Moreover, male-specific molecular responses in direct exposed (F0) generation to contaminants and in transgenerational lineages (F2, F3, and subsequent no exposed generations) remain poorly characterized.

A *de novo* transcriptome assembly was generated for this study because no annotated reference genome is currently available for *G. locusta* (Neuparth et al., 2020). Although a female transcriptome has been published (Neuparth et al., 2020), its use was deemed insufficient. Transcriptomic profiles can differ markedly between sexes, with many genes exhibiting sex-biased expression or producing sex-specific isoforms (Ellegren & Parsch, 2007; Zhang et al., 2022). Using a female-only transcriptomic reference could omit transcripts essential for understanding male-specific molecular responses to SIM.

4.2. RNA-Seq Data Quality and Transcriptome Assembly

High-throughput sequencing of male *G. locusta* samples generated a total of **445 million paired-end reads**, corresponding to an average of ~33 million reads per sample. Table 1 summarizes the number of raw reads, reads removed during trimming and the final number of high-quality reads retained for downstream analyses across all samples.

But before trimming, initial quality control of the raw reads was performed using FastQC (Andrews, 2010) showing generally good per-base quality scores, with only a slight decrease towards the 3' ends in all samples, a common feature of Illumina data (Andrews, 2010; Conesa et al., 2016).

Table 1 - Summary of sequencing output and pre-processing results for male *G. locusta* RNA-seq samples.

Samples	Raw sequencing reads	Final number of reads
S320_1G_1	30,860,545	30,321,544 (98,25%)
S320_1G_3	36,866,621	36,152,331 (98,06%)
S320_1G_4	33,412,246	33,006,389 (98,79%)
S320_1G_5	30,905,928	30,531,022 (98,79%)
S320T_4G_2	40,370,541	39,716,419 (98,38%)
S320T_4G_3	41,918,743	41,125,744 (98,11%)
S320T_4G_5	31,966,257	31,592,258 (98,83%)
SC_1G_2	31,295,071	30,685,351 (98,05%)
SC_1G_3	39,663,314	38,889,620 (98,05%)
SC_1G_5	31,849,255	31,469,249 (98,81%)
SC_4G_1	30,526,392	30,022,895 (98,35%)
SC_4G_2	30,628,501	30,083,235 (98,22%)
SC_4G_3	35,246,752	34,521,736 (97,94%)
Totals	445,510,166	438,117,793 (98,34%)

Figure 8a illustrates one sample as an example for these results. Adapter contamination was negligible in all samples, GC content was consistent with an average of ~42% per sample and sequence length distributions were uniform (150 bp) in all samples, indicating good sequencing performance. To correct the small fraction of bases with quality scores below 30 (illustrated in yellow in Figure 8a), Trimmomatic (Bolger et al., 2014) was applied to trim low-quality regions and remove any residual adapters, ensuring high-quality reads for downstream analyses.

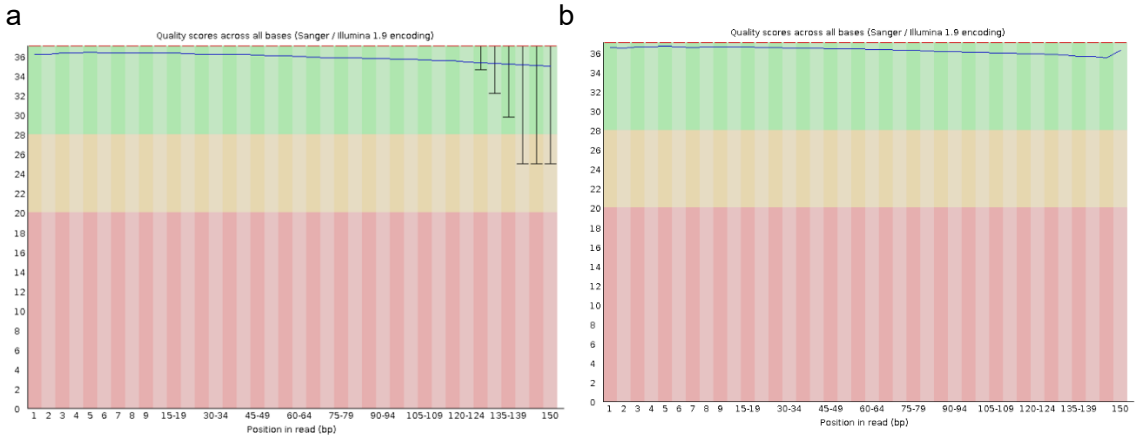


Figure 8a,b - Per-base sequence quality scores for sample S320_F0_4 before (a) and after (b) trimming and filtering with Trimmomatic. The raw data (a) show a slight decrease in Phred quality scores (<30) towards the 3' end of the reads, a common Illumina artefact, whereas the processed data (b) show uniformly high Phred scores (>30) across the entire read length, indicating successful removal of low-quality bases and improved overall read quality.

Trimmomatic was then performed resulting in a substantial improvement in overall read quality, as confirmed by post-processing FastQC reports showing uniformly high Phred scores (>30) across read lengths (Figure 8b).

On average, **98.34 % of reads were retained** after trimming (range: 97.9–98.8 % between samples), with only ~1.66 % of reads removed across the dataset (Table 1). This retention rate is considered excellent and consistent with other well-optimized RNA-seq studies in non-model crustaceans and other invertebrates (Meyer et al., 2011; Machado et al., 2022).

Subsequent error correction with Rcorrector (Song & Florea, 2015) further enhanced the dataset by identifying and correcting low-frequency k-mers likely representing sequencing errors, while preserving true biological variation. Given the complexity of the *G. locusta* transcriptome and the lack of a reference genome, this step was especially crucial (Neuparth et al., 2020). These adjustments helped to produce more accurate and

coherent *de novo* assembly by lowering possible artifacts like incorrectly named bases. Importantly, this process did not remove any reads, but corrected likely erroneous bases that can arise during Illumina sequencing.

Overall, the pre-processing and quality control workflow yielded **438 million high-quality paired reads** across all samples, providing a robust foundation for *de novo* transcriptome assembly and subsequent differential expression analysis (Table 1). These curated reads were then used to generate two transcriptome assemblies (V0 All and V0 Subset), whose metrics are reported and compared in the next Section.

4.3. Transcriptome Assembly and Quality Assessment

De novo transcriptome assemblies were performed using Trinity with normalized reads from the pre-processed samples to produce two reference transcriptomes and select the best one. As explained in the Methods section 3.2.2, two transcriptome assemblies were generated: **V0 All (assembly with all samples)** and **V0 Subset (assembly with control samples only)**. This dual approach allowed us to assess the impact of including treatment-specific samples on transcriptome complexity and to select the most representative reference for downstream analyses. Key assembly metrics are summarized in Table 2.

Table 2 – Summary of transcriptome assembly metrics for the V0 All (global) and V0 Subset (control-only) assemblies of male *G. locusta*.

Input Data	**Dataset (all samples)**	**Subset (control samples)**
Number of raw sequencing reads (PE)	438,117,793	199,209,285
Selected reads during normalization stage (Trinity)	408,111,404	195,672,086
Transcriptome Assembly Versions	**Global Assembly**	**Control Assembly**
Transrate Transcriptome Quality		
Total number of Transcripts	575,712	393,397
n50 transcript length (bp)	1,365	1,490
n90 transcript length (bp)	429	450
Mean sequence length (bp)	982.14	1,046.72
Largest transcript (bp)	21,631	23,735
Number of transcripts over 1k nn	162,451	123,350
GC %	41	41
BUSCO (Met/Art)		
Complete (%)	91.4/88.6	90.3/85.9
Single (%)	4.0/5.1	6.5/6.7
Multi (%)	87.4/83.5	83.8/79.2
Fragment (%)	3.7/5.0	3.8/5.6

Missing (%)	4.9/6.4	5.9/8.5
Total Busco Groups (%)	95.1/93.6	94.1/91.5
Mapping Statistics % (Bowtie2)		
All_samples	97.13	-
All_Control_samples	-	97.13
S320_1G_1	-	95.93
S320_1G_3	-	95.88
S320_1G_4	-	95.62
S320_1G_5	-	96.01
S320_4G_2	-	96.02
S320_4G_3	-	95.73
S320_4G_5	-	96.53
SC_1G_2	-	96.05
SC_1G_3	-	96.19
SC_1G_5	-	97.06
SC_4G_1	-	96.47
SC_4G_2	-	96.28
SC_4G_3	-	96.45

During *in silico* normalization, Trinity retained 408,111,404 reads (≈92%) for the global assembly and 195,672,086 reads (≈98%) for the control-only assembly, effectively reducing redundancy from highly expressed transcripts while preserving transcript diversity (Table 2 – Input Data). These high retention rates indicate excellent dataset quality and complexity, consistent with other non-model crustacean transcriptomes (Escobar-Sierra et al., 2025). For context, in the female *G. locusta* transcriptome spanning four generations (F0–F3) (Neuparth et al., 2020b), 559 million reads were used for assembly out of 563 million sequenced, corresponding to a retention rate >99%. These observations suggest that our read retention (~92–98%) is within the expected range for high-quality, complex *de novo* assemblies in *G. locusta*, supporting the reliability of downstream transcript reconstruction.

Our resulting assemblies exhibited notable quality differences with Transrate (Table 2 – Transrate Transcriptome Quality). The global assembly comprised 575,712 transcripts (N50 = 1,365 bp; N90 = 429 bp; mean = 982 bp), whereas the control-only assembly contained 393,397 (N50 = 1,490 bp; N90 = 450 bp; mean = 1,047 bp). The largest transcripts had 21,631 bp and 23,735 bp, respectively, with consistent GC content (41%) and high numbers of transcripts exceeding 1 kb (162,451 vs. 123,350). The larger transcript count in the global assembly likely reflects increased complexity by including rare, treatment-specific isoforms and allelic variants, whereas the control-only assembly yielded longer, more contiguous sequences, indicative of a “cleaner” baseline transcriptome.

For context, the female *G. locusta* raw transcriptome (F0, F1, F2 and F3 generations) (Neuparth et al., 2020b) contained 1,607,378 sequences (N50 = 1,062 bp; mean length = 690 bp; largest transcript = 43,269 bp). The markedly higher transcript diversity in

females likely reflects the inclusion of multiple generations, which broadens isoform representation, as well as the expression of reproduction-related and regulatory genes (e.g., vitellogenesis, oogenesis and endocrine control pathways) that might be less active in males. Thus, the lower number of assembled transcripts in males (393k–575k) may reflect a sex-specific expression profile focused on somatic and spermatogenic processes rather than reduced assembly quality.

Overall, our assembly statistics of the male transcriptomes (N50 = 1,365–1,490 bp; mean length = 982–1,047 bp) are consistent with high-quality *de novo* assemblies for females *G. locusta* (Neuparth et al,. 2020b) and other non-model crustaceans, including the gonad transcriptome of *Paracentrotus lividus* (N50 ≈ 1,842 bp; Machado et al., 2022) and *Gammarus pulex* (N50 > 1,500 bp; Escobar-Sierra et al., 2025).

BUSCO analyses against the Metazoa and Arthropoda lineages performed in the current study, indicated high assembly completeness for both global and control only transcriptomes (Table 2 - BUSCO). Considering the sum of Complete and Fragmented BUSCOs, the global assembly recovered 95.1% of Metazoa and 93.6% of Arthropoda orthologs, while the control-only assembly showed slightly lower values (94.1% and 91.5%, respectively). These findings support the reliability of both assemblies by showing that most conserved single-copy orthologs are represented (>90%) (Keiichi Kakui et al., 2021).

As anticipated, completeness analyses of the transcriptomes using the Arthropoda database showed lower completeness scores than with the Metazoa database. This is commonly observed because the Arthropoda lineage contains a higher number of clade-specific orthologs, some of which may not be expressed under the experimental conditions, resulting in apparent absence and consequently a lower percentage of orthologs identified with BUSCO in transcriptome assemblies (Simão et al., 2015; Manni et al., 2021). Additionally, the dataset may lack closely related species to *Gammarus*, which can further reduce the detection of expected orthologs. Therefore, the lower Arthropoda scores do not indicate poor assembly quality but rather reflect both the biological specificity of the dataset and the limited representation of closely related reference species.

Comparison with the previously published female *G. locusta* transcriptome (Neuparth et al., 2020) indicates that our male assemblies present slightly lower BUSCO completeness percentages for both Metazoa and Arthropoda datasets. The female transcriptome reported total BUSCO found values of 99.7% (Metazoa) and 97.5%

(Arthropoda) for the raw assembly, whereas our male global assembly recovered 95.1% and 93.6%, respectively (Neuparth et at., 2020 Annex 2).

This difference is likely due to methodological variations: the female transcriptome pipeline included an additional taxonomic filtering step using Centrifuge, which removed reads not classified as Arthropoda, thereby reducing background contamination and non-target sequences. Our male transcriptome, on the other hand, used reads from the entire organism without this filtering, which could have included transcripts from symbionts, intestinal bacteria, or other environmental sources (Zhou et al., 2024). Consequently, a larger pool of non-target genes could lower the apparent percentage of conserved single-copy orthologs in Metazoa and Arthropoda datasets (Simão et al., 2015; Manni et al., 2021). Additionally, some of the difference may reflect biological variation, since females might express additional reproduction-related and regulatory genes (e.g., vitellogenesis, oogenesis, endocrine control) that might be absent or downregulated in males, contributing to a broader representation of orthologs.

Despite this, the observed BUSCO scores for both male assemblies remain high, supporting their reliability as references for functional annotation and differential expression analyses, as scores above 90% are generally considered indicative of good transcriptome coverage and assembly quality (Keiichi Kakui et al., 2021).

Backmapping of RNA-seq reads to the assembled transcriptomes confirmed that both assemblies captured transcriptome diversity with high efficiency (>95% mapping; Table 2 - Mapping Statistics Bowtie2). The **global assembly (V0 All)** showed a mapping rate of **97.13% across all concatenated samples** and the **control-only assembly (V0 Subset)** reached the same mapping rate of **97.13% for concatenated control samples.** When mapping all individual samples (control, exposed and transgenerational) to the control-only assembly, alignment rates remained high (95.62–97.06%), demonstrating that the subset assembly adequately represents transcriptomic content across treatments. These results are consistent with other research on crustaceans, which found that mapping rates of 88% to 93% were a sign of high-quality assemblies (Shyamal et al., 2018).

Despite similar representativity, the assemblies differed in redundancy and structural quality. The global assembly contained 575,712 transcripts and a higher proportion of duplicated BUSCOs (87.4% Metazoa / 83.5% Arthropoda), reflecting redundancy likely driven by heterozygosity and treatment-specific isoforms (Manni et al., 2021). In contrast, the control-only assembly reduced redundancy to 393,397 transcripts with fewer duplicated orthologs (83.8% / 79.2%), while retaining >94% total conserved genes. It

also exhibited slightly higher contiguity (N50 = 1,490 bp; N90 = 450 bp) and longer maximum transcripts, indicating greater assembly coherence.

Assemblies incorporating multiple individuals or treatments can artificially inflate transcript and isoform numbers by assembling allelic variants as separate contigs (Grabherr et al., 2011). Restricting the assembly to controls mitigates this redundancy while preserving transcriptome completeness, yielding a cleaner baseline transcriptome that improves the accuracy of functional annotation and expression analyses (Grabherr et al., 2011; Conesa et al., 2016). Based on these results, the **V0 control-only assembly was selected as the V0 assembly reference**.

4.4. Transcriptome Reduction and Final Reference (V1)

Taken together, these results supported the selection of the V0 control-only assembly as the most suitable baseline for downstream analyses, as it combines high completeness with lower redundancy and improved contiguity compared to the global assembly. Nevertheless, with nearly 400,000 transcripts, the raw V0 Subset still contained a considerable degree of redundancy, reflecting alternative isoforms, allelic variants and assembly artefacts inherent to *de novo* transcriptome reconstruction. To address this, we implemented a transcriptome reduction pipeline aimed at collapsing redundant contigs into clusters while retaining the most biologically representative isoforms. The final reference transcriptome (V1) was produced as a result of this step and was used as the foundation for further analyses.

The refinement of the control-only V0 transcriptome resulted in a final, high-confidence reference transcriptome (V1) for *G. locusta* males. **Transcripts were clustered based on shared read support across biological replicates** using Corset v1.09 (Davidson & Oshlack, 2014), resulting in a **V1 transcriptome exhibiting substantial reduction of redundancy** while retaining the majority of biologically relevant isoforms.

After filtering, the total number of transcripts decreased from 393,397 in the V0 control assembly to 136,740 in V1 (−65%) (Table 3). This reduction markedly improved interpretability and assembly coherence by consolidating redundant isoforms and assembly artefacts into coherent transcript clusters, a crucial step for non-model organisms with high heterozygosity such as *G. locusta*. This was evident in the quality assessment of the assembled V0 transcriptome of *G. locusta* females, where BUSCO

analysis also reported a high proportion of duplicated orthologs (>90%) (Neuparth et al., 2020b Annex 2) and post-assembly redundancy reduction has likewise been recommended in non-model RNA-seq studies (Yang & Smith, 2013). Indeed, the combined effects of heterozygosity, isoform complexity and assembly artefacts are known to inflate transcript counts, fragment contigs and introduce artificial duplications in marine invertebrate transcriptomes (Lima et al., 2017).

Table 3. Assembly statistics and BUSCO completeness of the control-derived transcriptome before (V0) and after (V1) redundancy reduction and filtering.

Transcriptome Assembly Versions	**V0 (control samples)**	**V1**
Total number of Transcripts	393,397	136,740
n50 transcript length (bp)	1,490	1,405
Mean sequence length (bp)	1,046.72	1,087
Largest transcript (bp)	23,735	23,735
Number of transcripts over 1k nn	123,350	44,705
GC %	41	41
BUSCO (Met/Art)		
Complete (%)	90.3/85.9	89.2/85.2
Single (%)	6.5/6.7	62.7/60.9
Multi (%)	83.8/79.2	26.5/24.3
Fragment (%)	3.8/5.6	3.7/4.7
Missing (%)	5.9/8.5	7.1/10.1
Total Busco Groups (%)	94.1/91.5	92.9/89.9

Importantly, this reduction preserved most of the biologically relevant content. BUSCO analysis confirmed that V1 maintained high completeness (92.9% Metazoa; 89.9% Arthropoda), with only a slight decrease relative to V0 (94.1% and 91.5%, respectively). The marked drop in duplicated BUSCOs (Metazoa: 83.8% → 26.5%; Arthropoda: 79.2% → 24.3%) indicates successful removal of redundancy likely caused by allelic variants or mis-assembled paralogs, resulting in a transcriptome that better represents the gene space.

N50 length and mean transcript length remained stable (N50 = 1,405 bp; mean = 1,087 bp), indicating that transcript contiguity was not compromised. Although the number of transcripts >1 kb decreased (123,350 → 44,705), this likely reflects the elimination of low-confidence isoforms rather than loss of core genes, consistent with the maintained BUSCO completeness.

Following transcriptome refinement with Corset v1.09 (Davidson & Oshlack, 2014), read counts for each transcript cluster across all samples were quantified using FeatureCounts v.2.0.2 (Liao et al., 2014) and a count matrix with clusters as rows and samples as columns was generated (Table 4). Each row represents a coherent transcript cluster appropriate for further gene expression analysis thanks to the method of

clustering transcripts by sample similarity using Corset, which also minimizes technical noise from potential alternative isoforms that could appear and make the analysis redundant or potential assembly artifacts.

Table 4 - Excerpt from count matrix (FeatureCounts output, post-Corset clustering).

Cluster	S320_1G_1	S320_1G_3	S320_1G_4	S320_1G_5	SC_1G_2	SC_1G_3
Cluster-0.0	5	2	3	10	50	40
Cluster-1.0	3	5	7	4	10	15
Cluster-27.1	2	14	0	2	10	46
Cluster-69.0	14	79	96	21	18	390
Cluster-120.13	24	197	62	97	51	94

Note: Only a subset of clusters and samples are shown for illustration; full count matrix includes all clusters and biological replicates.

The V1 transcriptome makes it easier to quantify gene expression by grouping redundant isoforms into clusters. In non-model organisms with high heterozygosity, individual isoforms often reflect alternative alleles or assembly artefacts rather than biologically distinct transcripts. In order to improve interpretability and statistical power in differential expression analyses, clustering makes sure that expression is measured per gene or biologically relevant unit rather than per potentially redundant isoform. This approach aligns with best practices for *de novo* transcriptome refinement in non-model species (Haas et al., 2013; MacManes, 2018), offering a robust foundation for investigating SIM-induced transcriptional changes across generations.

4.5. ORF Prediction

ORFs were predicted from the refined V1 male transcriptome using TransDecoder (v5.5.0; Haas et al., 2013), a standard tool for *de novo* transcriptome analyses that identifies putative protein-coding sequences. This step enables downstream functional annotation, including homology searches, domain identification and pathway mapping (Conesa et al., 2016; Grabherr et al., 2011).

From the **136,740 clusters** present in the V1 transcriptome, a total of **75,042 ORFs** were identified, with the following distribution by type: **complete ORFs: 37,324 (49.7%)**, **internal ORFs: 10,388 (13.8%)**, **5'-partial ORFs: 20,967 (27.9%)** and **3'-partial ORFs: 6,363 (8.5%)** (Table 5). While partial ORFs may indicate truncated transcripts as a result of incomplete assembly or biological isoforms, complete ORFs correspond to sequences that contain both start and stop codons (Moreno-Santillán et al., 2019; Gilchrist et al.,

2015). The observed proportion of complete ORFs (~50%) is consistent with other *de novo* transcriptome studies in non-model invertebrates, where assembly fragmentation and variable transcript expression often lead to partial ORFs (Grabherr et al., 2011; Smith-Unna et al., 2016).

Table 5 – Summary of predicted ORFs in the V1 transcriptome of *G. locusta* males.

ORF Type	Number of ORFs	Percentage (%)
Complete	37,324	49.74
Internal	10,388	13.84
5'-partial	20,967	27.94
3'-partial	6,363	8.48
Total	75,042	100

All ORFs, both complete and partial, were retained for functional annotation. By maximizing sensitivity, this method guarantees that biologically significant sequences, including truncated transcripts and isoforms with low expression, are included in subsequent analyses (Moreno-Santillán et al., 2019; Gilchrist et al., 2015). Retaining partial ORFs is a valid practice in transcriptomes of crustaceans, particularly in non-model species, as these sequences can contain important functional information, such as conserved domains, homology to known proteins and functional annotations including GO terms and KEGG pathways (Huerlimann et al., 2018; Vysakh et al., 2025). Equally, Neuparth et al. (2020) retained all predicted ORFs to functionally annotate the female *G. locusta* transcriptome.

The ORF prediction results support the V1 transcriptome's suitability as a reference for RNA-seq analyses and differential expression studies by showing a high-quality transcriptome assembly with a significant proportion of transcripts encoding putative proteins. The distribution of complete and partial ORFs (49.7% complete, 13.8% internal, 27.9% 5'-partial, 8.5% 3'-partial) reflects both technical factors, such as assembly fragmentation and read coverage and biological complexity, highlighting the capacity of the male V1 transcriptome to capture coding potential across different tissue types and developmental processes. (Conesa et al., 2016; Huerlimann et al., 2018).

4.6. Functional Annotation of the V1 Transcriptome

Functional annotation of the 75,042 predicted ORFs resulted in a large proportion of sequences with significant matches to crustacean proteins (taxid: 6656), yielding a comprehensive dataset of putative protein-coding genes with homology, domain and pathway information (Table 6).

Table 6 – Summary of functional annotation of predicted ORFs in the V1 males *G. locusta* transcriptome, before and after redundancy removal.

Annotation Database / Feature	Pre-curation (nr clusters)	% of total ORFs	Post-curation (removal of duplicated clusters)	% of curated clusters
NCBI NR hits	17,766	23.7%	–	–
SwissProt hits	4,458	5.9%	–	–
eggNOG hits	6,208	8.3%	5,950	95.8%
COG category	6,051	8.3%	5,801	97.5%
eggNOG description	6,051	97.5% of 6,208	5,801	97.5%
GO terms	4,132	66.6% of 6,208	3,950	66.4%
KEGG pathways	2,916	47.0% of 6,208	2,791	46.9%
EC numbers	1,934	32.50%	1,755	29.50%

In total, of the 75,042 predicted ORFs, **17,766** had significant BLASTP hits against NCBI NR and **4,458 ORFs** matched entries in SwissProt. Among the EggNOG-annotated ORFs, **6,051 sequences (97.5%)** had descriptive annotations and respective COG categories, **4,132 (≈66.6%)** were associated with GO terms, **2,916 (≈47.0%)** were linked to KEGG pathways and 1,934 (32.5%) had enzyme commission numbers.

To avoid redundancy, a curated dataset of 5,950 unique gene clusters was generated by removing 258 duplicated clusters with identical annotations in EggNOG, GO, KEGG

and EC annotations. After curation, 5,801 clusters (97.5%) retained EggNOG annotation and COG categories, 3,950 (≈66.4%) had at least one GO term, 2,791 (≈46.9%) were mapped to KEGG pathways and 1,755 (29.5%) had EC number. After duplicates were eliminated, the percentage of annotated clusters stayed constant, indicating that functional information that was biologically significant was retained.

The NCBI nr annotations were used to create a new column in the counts table file to assign to each cluster a protein identifier. SwissProt annotations were not used, as the NCBI nr database was selected instead, following the approach of Neuparth et al. (2020).

The information of the Cluster combined with EggNog Description, GO terms, COG categories, KEGG pathways and EC number was retrieved to a single XLSX file to serve as input for the app developed in this study. The annotation file is as follows in Table 7.

Table 7 - Summary of selected clusters from the annotation file used in this study. Only a subset of clusters is shown for illustration purposes.

Gene	EggNOG	Categoria	GO terms	KEGG pathway	EC number
Cluster-28429.1	ligand-gated ion channel (TC 1.A.9) family	T	plasma membrane protein complex; regulation of cellular process; transmitter-gated ion channel activity	Neuroactive ligand-receptor interaction (ko04080)	-
Cluster-9904.1	Transmembrane transporter activity	P	intracellular chemical homeostasis; regulation of pH; sulfate transmembrane transport	-	-
Cluster-36097.8	actin binding	Z	-	Lysosome (ko04142); Endocytosis (ko04144)	-

This curated, non-redundant dataset served as a robust reference for downstream analyses, including functional enrichment and pathway reconstruction. Despite COG categories and EC numbers are annotated, we didn't use them for the enrichment analyses. To enable interactive exploration of these curated data and facilitate differential expression and functional enrichment analyses, we subsequently developed a custom Shiny application.

4.7. Development of the TransDegAnalyser Tool for Differential Expression Analysis and Functional Enrichment Visualization

As part of this work, a custom web-based Shiny application, **TransDegAnalyser**, was developed in R (v.4.4.0; R Core Team, 2025) using the Shiny framework (v.1.11.1; *Web Application Framework for R*, 2021) to enable **interactive exploration of *G. locusta* male RNA-seq data** (Figure 9). The platform integrates three primary data files – counts matrix, experimental metadata and functional annotations (EggNOG, GO and KEGG) - into a single environment, facilitating dynamic and reproducible transcriptomic analyses across generations. The app also includes downloadable example files which can serve as input to perform the analyses and each plot/table have an interactive description summarizing parameters used for the analysis and the outputs. The application is available in GitHub and can be accessed through this link: https://github.com/abiliaframboesa/Thesis_RNASeq.

This computational toolbox provides a **user-friendly interface**, making it accessible to users without bioinformatic experience, for performing multilevel transcriptomic analyses between generations.

The interface of **TransDegAnalyser** is organized into six main tabs, the first corresponding to a welcome page and the subsequent each corresponding to a stage of transcriptomic analysis (Figure 9):

- **Welcome** – introduction to the platform and the biological context of transgenerational effects, overview of the available functionalities and a quick-start guide with instructions on uploading input files and navigating the results.
- **PCA** – visualization of global gene expression patterns that enables users to analyse biological replicate clustering and identify key sources of variation across samples.
- **Differential Analysis** – identification of DEGs between experimental conditions, presented in volcano plots and interactive tables to facilitate rapid inspection.
- **Heatmap** – visualization of DEG expression profiles across treatments and generations, enabling users to detect condition-specific and shared expression trends.

- **Transgenerational Comparison** – identification of genes uniquely or consistently affected across generations, visualized through Venn diagrams and linked interactive tables.
- **Functional Enrichment** – integration of GO terms, KEGG pathways and eggNOG annotations to link transcriptional changes with biological functions and pathways, with enrichment results displayed as barplots and complemented by interactive, gene-level annotated tables.

To perform the analyses in the PCA, Differential Analysis, Heatmap and Transgenerational Comparison tabs, users first need to provide the metadata and counts tables containing protein identifiers, as described in the Methods section 3.2.4. For the Functional Enrichment tab, the functional annotation table must also be supplied, following the procedure outlined in the Methods section 3.2.6.

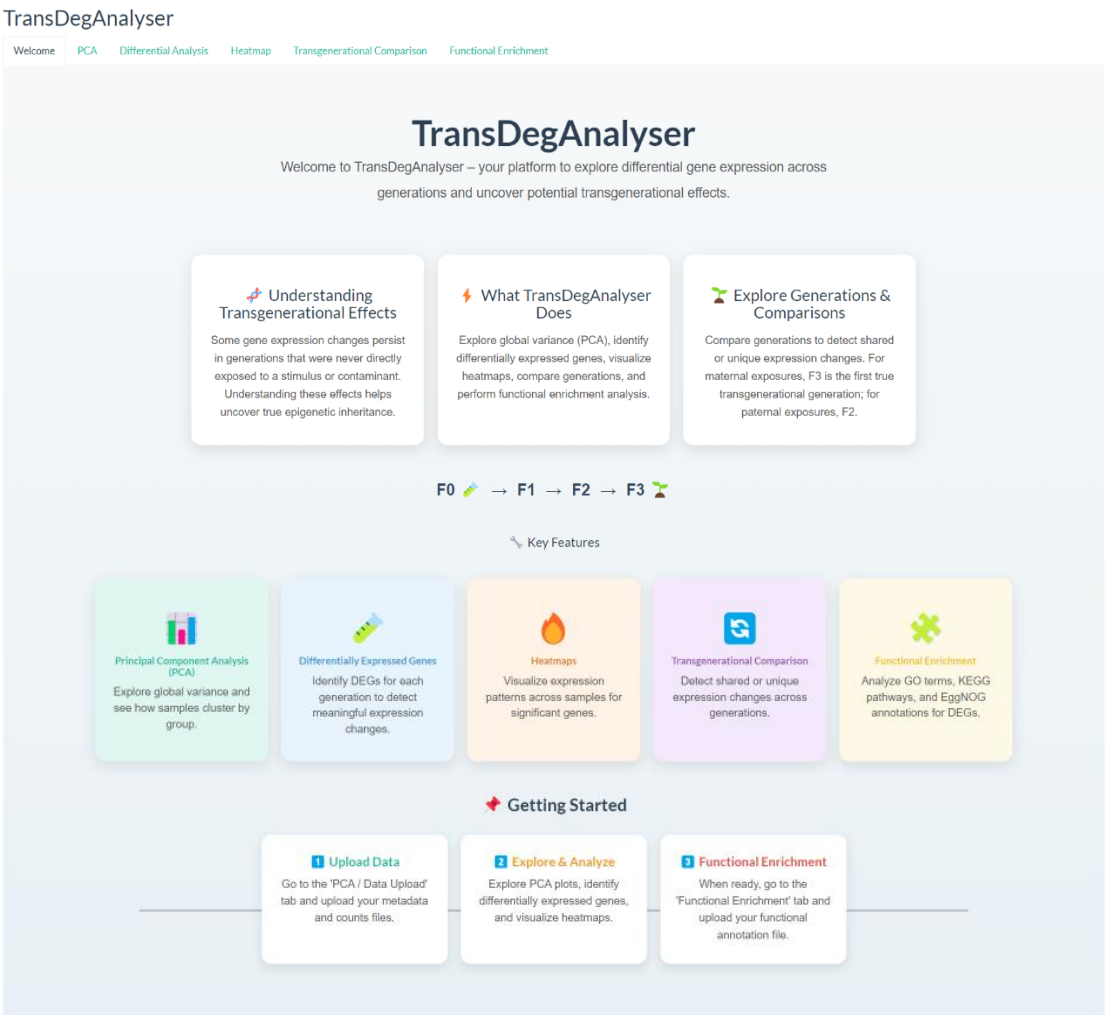


Figure 9 - Welcome tab of the TransDegAnalyser Shiny application for transcriptomic analysis. The app allows interactive exploration of global gene expression patterns, identification of differentially expressed genes, transgenerational comparisons and functional enrichment analysis in *G. locusta* male RNA-seq data.

By consolidating these functionalities, the platform allows efficient assessment of data quality, reproducibility among replicates, detection of genes uniquely or consistently affected across generations and exploration of biological pathways associated with observed direct exposure and transcriptional changes. The application thus represents a key result of this thesis, demonstrating how high-dimensional RNA-seq datasets can be transformed into interpretable results in an interactive and reproducible manner.

In the following subsections, the platform was employed to explore global expression patterns, differential expression, transgenerational effects and functional enrichment, providing insights into both direct exposure and transgenerational responses to 320 ng/L SIM treatment in *G. locusta* males.

4.8. PCA

Initial data analysis was conducted to investigate global patterns of gene expression across generations and treatment conditions in *G. locusta* males. PCA was performed on variance-stabilized counts obtained from DESeq2 and visualized using the custom Shiny application developed for this study. This approach allowed assessment of sample clustering and evaluation of major sources of variation within the dataset. **Figure 10a illustrates the PCA tab of the app and Figure 10b illustrates the PCA plot-only for a better visualization analysis.**

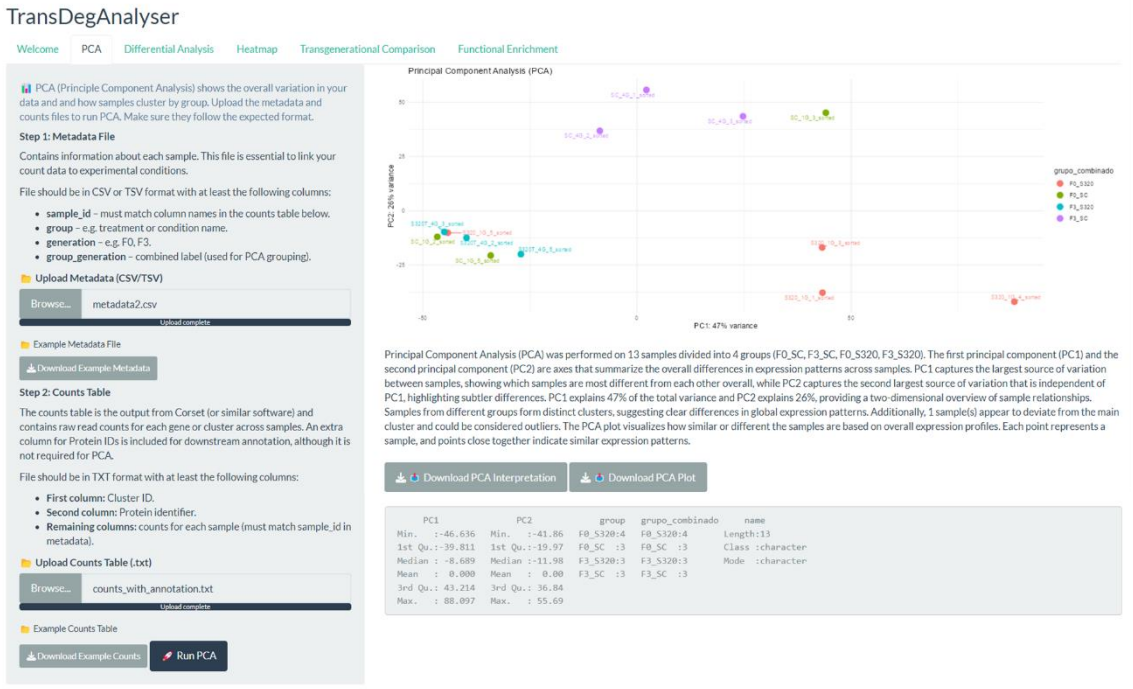


Figure 10a - PCA tab panel of TransDegAnalyser app. The tab shows the interface of this tab, the structure of the input files required to run the PCA plot and PCA plot of gene expression data with a description of the results. Samples are colored by combined group (F0_SC, F0_S320, F3_SC, F3_S320) to illustrate clustering by generation and treatment.

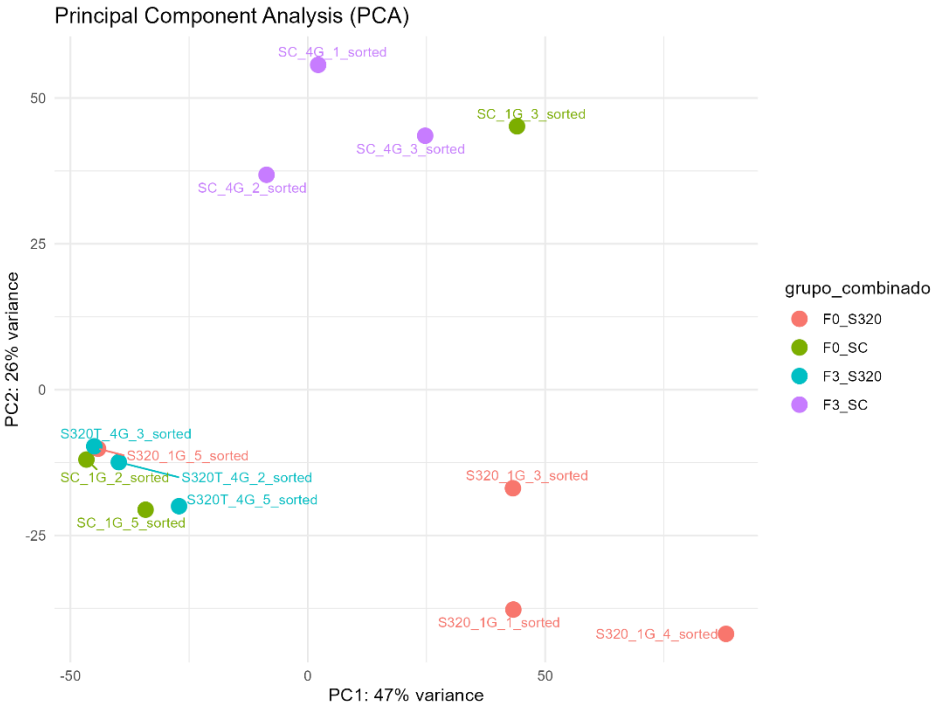


Figure 10b – Shows a detailed view of the PCA plot shown in Figure 10a. Samples are colored by combined group (F0_SC in green, F0_S320 in orange, F3_SC in purple, F3_S320 in blue) to illustrate clustering by generation and treatment.

To specifically explore potential direct exposure and transgenerational effects, we first performed a PCA on variance-stabilized counts from all expressed genes in male *G. locusta* samples. This exploratory approach enables visualization of complex datasets in less dimensions, revealing patterns, clusters and differences between samples (Ivosev et al., 2008). Specifically, we aimed to determine whether F3 transgenerational samples cluster with F0 exposed samples, which would indicate that exposure effects experienced by the parental generation (F0) had been transmitted to unexposed transgenerational F3 descendants.

The PCA provided an overview of the main sources of variation in the dataset. The first principal component (PC1, x-axis) explained 47% of the total variance, while the second principal component (PC2, y-axis) explained 26%, capturing secondary differences (Figure 10b). This means that nearly half of the differences in global gene expression profiles between samples can be described along the first axis (PC1), with a further quarter explained by the second axis (PC2). In practical terms, PC1 represents the principal source of expression variability, separating experimental groups most clearly, while PC2 explains additional, less pronounced differences.

In the PCA plot, **F0 exposed samples, orange points in PCA S320_1G** (*G. locusta* males exposed to 320 ng/L of SIM in the F0 generation) **are separated along PC1 from the other groups: F0 control, green points in PCA SC_1G** (*G. locusta* males in the control F0 generation), **F3 control, violet points in PCA SC_4G** (*G. locusta* males in the F3 control generation) **and F3 transgenerational, blue points in PCA S320_4G** (*G. locusta* males truly unexposed to SIM in the F3 generation) (Figure 10b).

In this analysis, **no clear clustering was observed between F0 direct exposed and F3 transgenerational samples**, which would have indicated strong transgenerational inheritance of SIM exposure effects at the global transcriptome level. Instead, F3 transgenerational samples clustered entirely with F0 controls and partially with F3 controls along PC1. This suggests that the majority of variance in gene expression (47%, captured by PC1) is dominated by direct SIM exposure effects in F0.

These results indicate that transgenerational effects are not immediately evident throughout the entire transcriptome. Although this provides an overview of global expression patterns, the fact that PCA includes all expressed genes may mask more subtle and biologically meaningful signals, which might could be observed by using DEGs in posterior analysis. This is what we will see later, where analyses focusing on DEGs revealed clearer associations between F0 exposed and F3 transgenerational samples. These associations are illustrated in the following sections through heatmaps

of DEGs, suggesting that transgenerational effects may be detectable at the level of DEGs rather than the entire transcriptome.

4.9. Differential Expression Analysis

Differential expression analysis was performed using DESeq2 to compare SIM-treatment group versus control group within each generation, using the control group as reference to know if a gene in up- or down-regulated in F3 generation. Genes with low counts (<10) were filtered before running the analysis and significance thresholds were set in the TransDegAnalyser interface at log2 fold change (log2FC) ≥ 1 and adjusted p-value (padj) ≤ 0.05, as explained in Methods section 3.2.7.5 (Figure 11a).

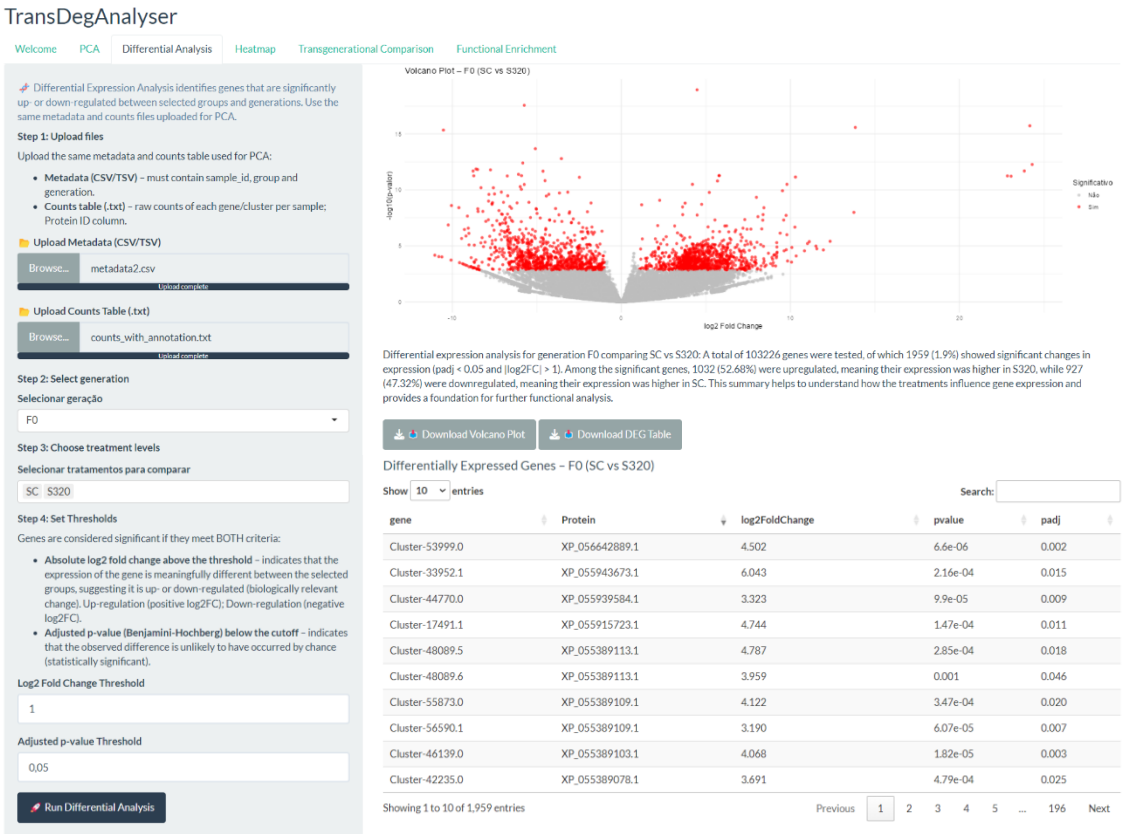


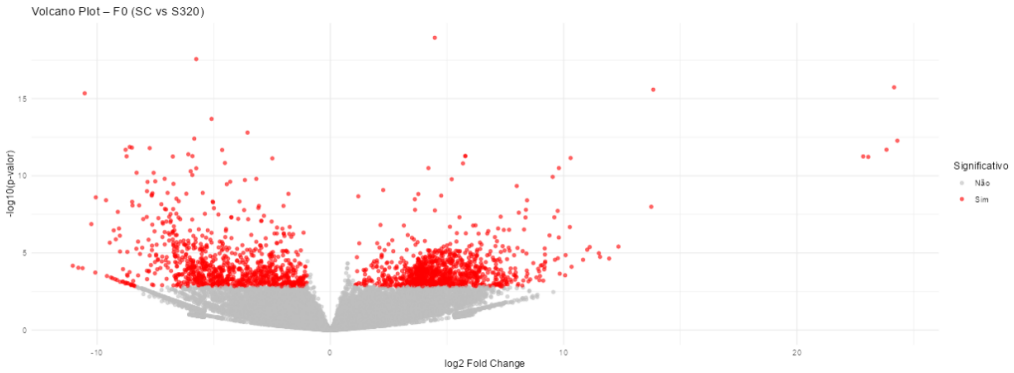
Figure 11a - Differential Analysis tab panel of the TransDegAnalyser app. The sidebar shows the input files needed for differential gene expression analysis, allows the user to choose the generation and treatment groups to compare and provides options to set thresholds for log2 fold change and adjusted p-value. The main panel shows a volcano plot highlighting significantly differentially expressed genes (red dots) for the selected treatment comparison (control vs exposed) based on log2 fold change and adjusted p-value thresholds, along with an interactive table providing detailed information for each identified gene.

From the **initial set of 136,740 transcript clusters**, **103,226 clusters** with more than 10 counts in the **F0 generation were retained for DEGs analysis**. Similarly, **96,161 clusters** with more than 10 counts in the **F3 generation were retained for DEGs analysis** (Table 8). This filtering step ensured that only reliably expressed clusters were included in the analysis. By excluding very low-abundance genes, it helped minimize transcriptional noise that is unlikely to provide meaningful biological insights.

Table 8 – Summary of differential expression analysis for F0 and F3 generations

Generation	F0	F3
Clusters >10 counts	103,226 (75.49%)	96,161 (70,32%)
Clusters dif. expressed	1,959 (1.9%)	5,390 (5.61%)
Clusters Up-regulated	1,032 (52.68%)	1,195 (22.17%)
Clusters Down-regulated	927 (47.32%)	4,195 (77.83%)
Annotated with NCBI nr	125 (6.38%)	259 (4.81%)

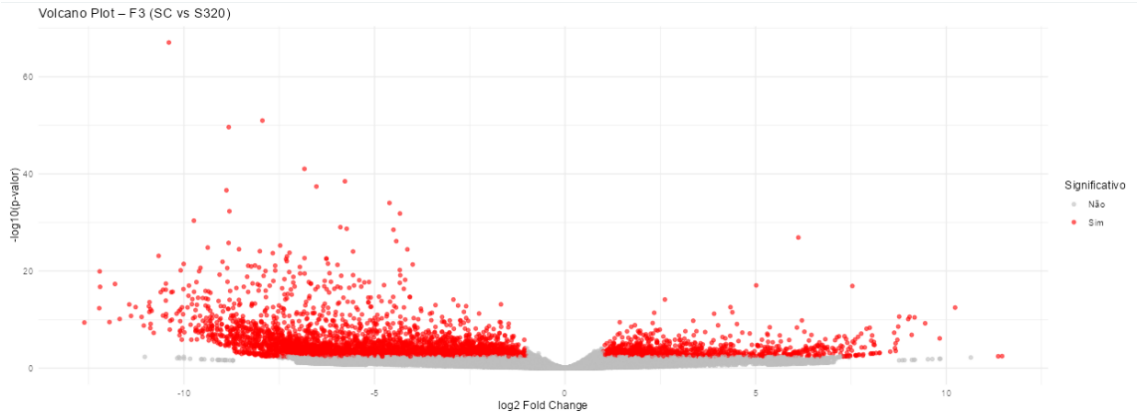
In the **parental exposed generation (F0)** (Figure 11b), from the 103,226 clusters with >10 counts from the V1 transcriptome, a total of **1,959 clusters** (1.9%) were identified **as significantly differentially expressed** for a padj < 0.05 and a log2fc ≥ 1 (Table 8). Among these DEGs, **1,032 clusters (52.7%) were up-regulated** and **927 clusters (47.3%) were down-regulated**, showing that the treatment produced a similar effect across the transcriptome. Only a small fraction of DEGs **(125 clusters, 6.4%) could be functionally annotated** against the NCBI NR database. This low annotation rate is expected in non-model species such as *G. locusta*, for which no reference genome is available, and is further compounded by the limited representation of invertebrate genomes in public databases. Nevertheless, the annotated DEGs give us clues about which molecular pathways the treatment might affect and highlight genes that are worth studying in more detail in further functional investigation.



Differential expression analysis for generation F0 comparing SC vs S320: A total of 103226 genes were tested, of which 1959 (1.9%) showed significant changes in expression (padj < 0.05 and |log2FC| > 1). Among the significant genes, 1032 (52.68%) were upregulated, meaning their expression was higher in S320, while 927 (47.32%) were downregulated, meaning their expression was higher in SC. This summary helps to understand how the treatments influence gene expression and provides a foundation for further functional analysis.

Figure 11b – Differentially expressed gene clusters in the F0 generation of male *G. locusta*. From 103,226 clusters, 1,959 clusters (1.9%) were significantly differentially expressed between treatment groups (padj < 0.05, log2FC ≥ 1) and 1,032 clusters (52.7%) were up-regulated and 927 clusters (47.3%) were down-regulated. Only 125 clusters (6.4%) had functional annotations in the NCBI NR database. The figure illustrates the distribution of up- and down-regulated DEGs.

In the F3 transgenerational lineage **(Figure 11c)**, from the **96,161 clusters (70.32%)** with >10 counts in the V1 transcriptome, **5,390 clusters (5.61%)** were identified as significantly differentially expressed (padj < 0.05, log2FC ≥ 1) (Table 8). From these DEGs, **1,195 clusters (22.2%) were up-regulated**, while **4,195 clusters (77.8%) were down-regulated**, showing that down-regulation was more common in the transgenerational response. . Only **259 DEGs (4.8%)** could be functionally annotated against the NCBI NR database, with the majority remaining unannotated, which is expected for a non-model species such as *G. locusta*.



Differential expression analysis for generation F3 comparing SC vs S320: A total of 96161 genes were tested, of which 5390 (5.61%) showed significant changes in expression (padj < 0.05 and |log2FC| > 1). Among the significant genes, 1195 (22.17%) were upregulated, meaning their expression was higher in S320, while 4195 (77.83%) were downregulated, meaning their expression was higher in SC. This summary helps to understand how the treatments influence gene expression and provides a foundation for further functional analysis.

Figure 11c - Volcano plot of DEGs in the F3 generation of male *G. locusta*. Each dot represents a transcript cluster and red dots indicate significant DEGs (padj < 0.05, log2FC ≥ 1). In total, 5,390 DEGs were identified, including 1,195 up-regulated and 4,195 down-regulated.

We observed a marked discrepancy between generations: whereas F0 showed 1,959 DEGs with a relatively balanced number of up- and down-regulated genes (1,032 up; 927 down), the transgenerational F3 generation displayed 5,390 DEGs, with a pronounced predominance of down-regulated genes (4,195 down vs 1,195 up). This pattern may reflect biological differences between generations, but it could also indicate that the number of DEGs increases across generations due to the progressive action of epigenetic mechanisms. **Over successive generations, molecular effects may accumulate or amplify, resulting in a broader transcriptomic response in F3**. In

other words, the higher number of DEGs observed in F3 may reflect **epigenetic inheritance and transgenerational reprogramming**, revealing regulatory mechanisms that fully manifest only in subsequent generations following direct parental exposure.

To further investigate whether these patterns reflect genuine transgenerational effects, follow-up analyses were performed: (i) heatmap clustering of DEGs to visualize whether F0 exposed and F3 transgenerational samples show similar expression patterns, potentially indicating transgenerational transcriptional alterations; and (ii) overlap analysis of DEGs between F0 and F3 to assess shared DEGs across generations.

4.10. Heatmap Visualization

To explore potential direct and transgenerational effects at the level of specific gene subsets, heatmap clustering was performed on DEGs identified in direct exposed generation (F0) and transgenerational lineage (F3). This approach made it easy to see gene expression patterns across samples and to check if F3 transgenerational samples show similar patterns to F0 exposed samples, which could suggest inherited effects. The heatmap (Figure 12a,b) shows gene expression for parental (F0, directly exposed to SIM) and transgenerational (F3, unexposed) samples, along with their controls (SC). Hierarchical clustering groups similar expression patterns, with colors showing levels of expression (red = upregulated, blue = downregulated). Group annotations indicate treatment (represented in blue and purple in Figure 12) and generation annotations indicate generational origin of the samples (represented in green and orange in Figure 12). This visualization is crucial for validating differential expression results and for identifying gene co-expression patterns, which can reveal shared molecular responses and underlying transcriptional networks (Eisen et al., 1998).

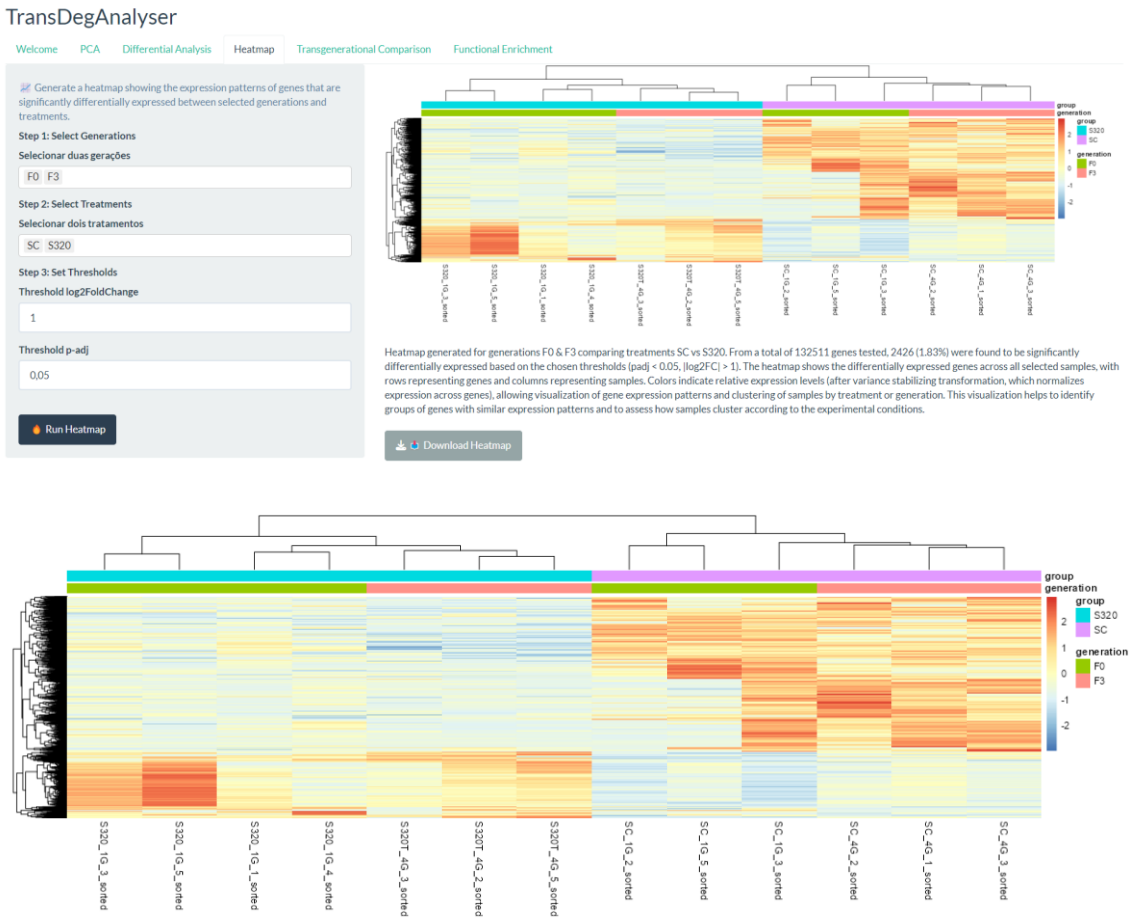


Figure 12a,b – Heatmap tab of the TransDegAnalyser app and respective detailed visualization . Heatmap of differentially expressed genes (log2fc >1 and padj<0.05) across generations (F0 and F3) and treatments (S320 vs. SC). Hierarchical clustering groups similar expression patterns, with colors showing levels of expression (red = upregulated, blue = downregulated). Group annotations indicate treatment and generational origin of the samples.

In the heatmap, it is possible to observe a clear separation between control (violet) and exposed/transgenerational groups (blue) within each generation (F0 - green; F3 - orange). Genes that are upregulated in one treatment group (e.g., F0 exposed) are generally downregulated or show no differential expression in the corresponding control group, and a similar pattern is evident in F3. It is also possible to observe that F0 exposed samples cluster together with F3 transgenerational samples, while F0 and F3 controls also cluster together. This indicates that transcriptional changes induced by direct SIM exposure in F0 are largely maintained in the unexposed F3 generation, suggesting that potential transgenerational effects are concentrated in DEGs rather than across the entire transcriptome.

This DEG-based clustering gives a clearer view than the previous PCA, where no clear separation between control and exposed/transgenerational samples was observed. By

focusing on genes that show significant differential expression, the heatmap shows the subset of genes where transgenerational effects are more present.

To interprete these patterns, we compared the male data with previously published female results using similar experimental conditions (Neuparth et al., 2020). Female heatmaps also showed clear separation between exposed and control groups, indicating that SIM induces sex-independent molecular effects. Importantly, the clustering of F0 exposed with F3 transgenerational males aligns with female patterns and supports the conclusion that transcriptional effects observed in F0 can persist in F3, reinforcing the biological relevance of transgenerational responses.

Overall, the heatmap analysis demonstrates that:

1. DEGs effectively separate exposed/transgenerational and control groups within each generation.
2. Similarity between F0 exposed and F3 transgenerational samples indicates the preservation of transcriptional changes across generations.
3. These patterns, consistent with both male and female data, suggest that heritable molecular alterations may be transmitted through transgenerational mechanisms, requiring further investigation of underlying epigenetic processes.

4.11. Transgenerational Comparison of DEGs

The analysis of DEGs between the F0 and F3 male generations revealed both shared and generation-specific transcriptional changes, highlighting transgenerational effects. TransDegAnalyser tab for transgenerational comparisons contains a Venn diagram (Figure 13) illustrating these patterns, showing 1,959 (1,759 F0 + 200 shared) DEGs in F0 and 5,390 (5,190 F3+ 200 shared) DEGs in F3, with 200 genes consistently altered across both generations. These shared genes represent a core molecular signature of SIM response, which persists in the absence of direct exposure and potentially reflect heritable transcriptional by epigenetic modifications (Vandegehuchte et al., 2008; Nikinmaa, 2014).

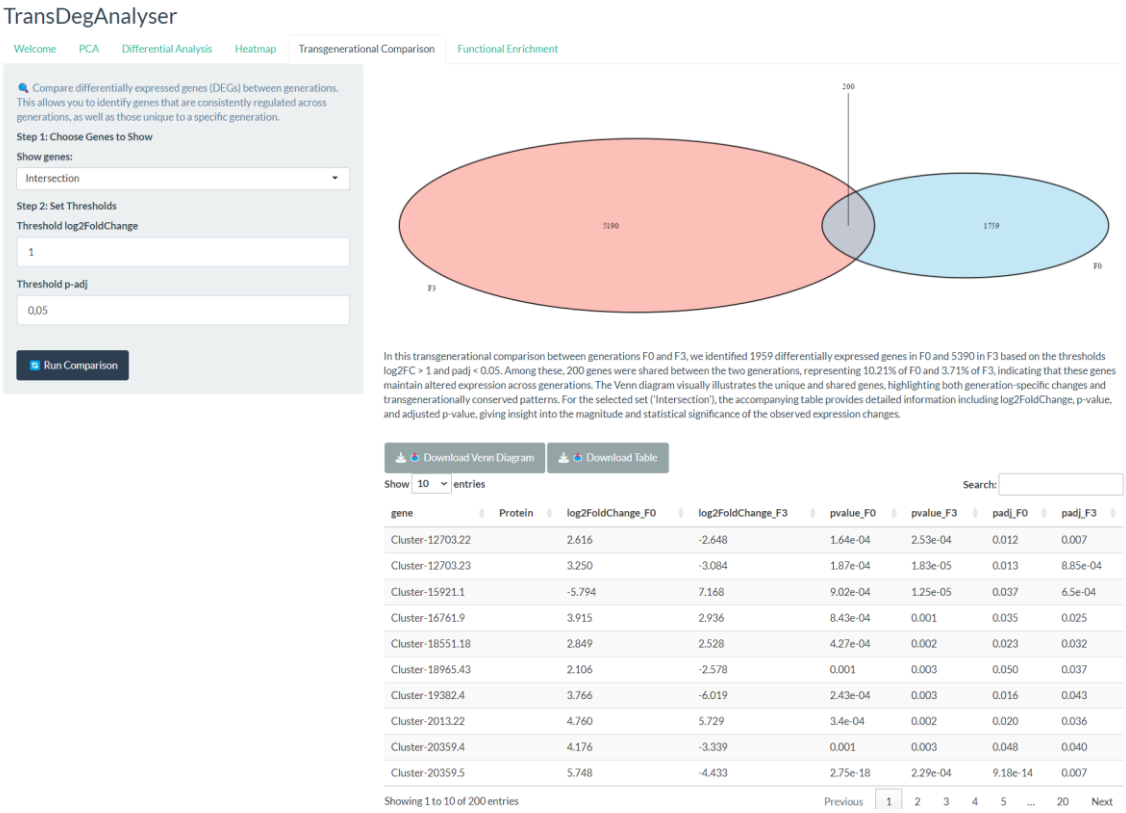


Figure 13 – Transgenerational Comparison tab from TransDegAnalyser. Comparison of DEGs in F0 and F3 generations, showing which genes are unique or shared based on the chosen log2FoldChange and adjusted p-value thresholds. A Venn diagram shows the distribution, and a table lists the genes found in both generations with their protein IDs, fold changes, and statistical values.

4.12. Functional Enrichment Analysis

To understand the biological meaning of the changes in gene expression, we performed functional enrichment analysis on the DEGs found in male *G. locusta.* This approach allows us to determine which biological processes, molecular functions, cellular components (GO terms), signalling pathways (KEGG) and functional categories (EggNOG) are overrepresented among the DEGs.

DEGs were analysed in two distinct groups:

- **Shared DEGs (F0 + F3)** - genes differentially expressed in both the directly exposed F0 generation and the transgenerational F3 generation. These genes represent **persistent transgenerational effects**, showing the molecular changes caused by SIM exposure in F0 that continue in later generations.

- **F3-specific DEGs** - genes differentially expressed **only in the F3 generation**. These genes may reflect **secondary or compensatory responses** arising in the

transgenerational generation, potentially modulating additional pathways beyond the original F0 effects by epigenetic reprogramming mechanisms.

The following subsections present the enrichment results for these groups, providing insights into the potential molecular mechanisms underlying direct and transgenerational effects of SIM in *G. locusta* males.

4.12.1. Enrichment Analysis for the Set of Shared DEGs (F0+F3)

For this analysis, the annotation file described in Methods 3.2.6 was uploaded to the TransDegAnalyser app. Although the annotation file contained multiple columns, only the EggNOG description, GO terms and KEGG pathways were used for enrichment analysis. After uploading the file and selecting the generation 'Intersection', significance thresholds, and the number of top terms to display, bar plots were generated.

To gain insights of how many DEGs mapped to NCBI, Eggnog, GO and KEGG pathways databases, we looked at the description of the barplot available in the app (Figure 14a), it was possible to see the identification of a core set of 200 shared DEGs, of which 39 (19.4%) had functional annotation in NCBI. A smaller fraction was annotated through EggNOG, with 8 (3.98%) genes linked to functional descriptions, 5 (2.49%) assigned to GO terms and 8 (3.98%) mapped to KEGG pathways. This limited annotation reflects the challenges of functional interpretation in non-model organisms such as *G. locusta*.

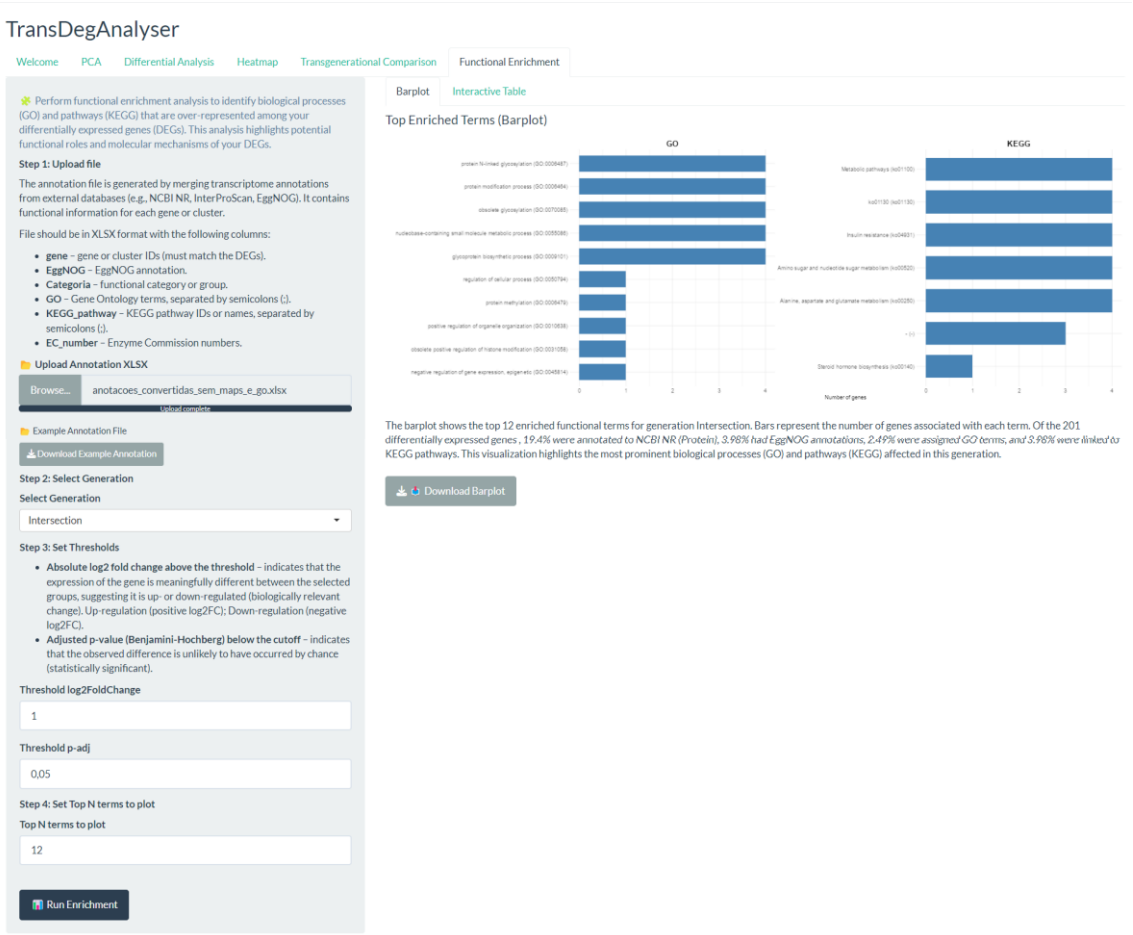


Figure 14 a- Functional Enrichment tab of the TransDegAnalyser. This tab requires the input of an annotations file with functional annotation of the transcript clusters. The selection of the generation (intersection, f0 or f3) and significant thresholds is also required. The main panel enables the visualization of barplots representing the number of the most expressed GO terms and Kegg Pathways amongst the transcripts of the selected generation and descriptions for how many clusters mapped to functional annotations.

To further explore the functional patterns within the gene subset, we looked at the enrichment barplots generated in TransDegAnalyser (Figure 14 a,b). In Figure 14b the enrichment annotations present in the plots are more perceptible. These plots show the most common GO terms and KEGG pathways among the annotated DEGs. They give a visual overview of the main biological processes and pathways that might be affected across generations.

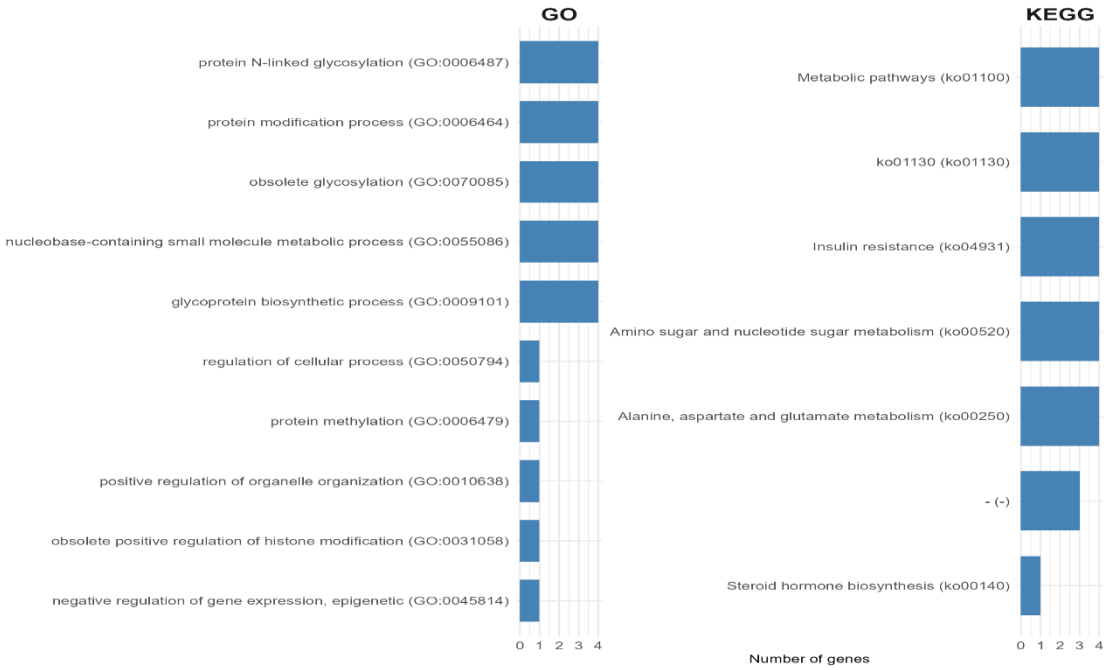


Figure 14b -Barplots showing the top 10 GO terms and KEGG pathways for the Intersection generation, which corresponds to the merge of the DEGs for both F0 and F3 generations, helping to identify biological process and pathways of the genes that were affected due to transgenerational inheritance.

The 10 most significantly enriched GO terms and KEGG pathways were analysed and are summarized below.

Key GO Terms Enrichment (F0/F3 Intersection)

The functional enrichment analysis of the core set of DEGs shared between F0 and F3 generations highlighted several biologically relevant processes in male *G. locusta*.

- **Protein modification and glycosylation:**
  Most significant enriched GO terms included *protein N-linked glycosylation* (GO:0006487), *protein modification process* (GO:0006464), *glycoprotein biosynthetic process* (GO:0009101), and *obsolete glycosylation* (GO:0070085). These processes are important for protein folding, stability and function and may influence key factors such as signalling proteins and hormone receptors involved in male reproductive systems (Willey, 1999; Campo et al., 2019; Huang et al., 2023).

- **Metabolic processes:**
  GO terms such as *nucleobase-containing small molecule metabolic process* (GO:0055086) indicate adjustments in nucleotide and small molecule

metabolism, potentially supporting energy allocation, detoxification or other metabolic demands in males exposed to SIM (Rodrigues et al., 2024; EMBL-EBI, 2025a).

- **Epigenetic and transcriptional regulation:**
  Enriched terms associated with epigenetic mechanisms included *negative regulation of gene expression, epigenetic* (GO:0045814), *obsolete positive regulation of histone modification* (GO:0031058), *protein methylation* (GO:0006479), *regulation of cellular process* (GO:0050794) and *positive regulation of organelle organization* (GO:0010638). These processes suggest that transgenerational inheritance could be mediated via epigenetic reprogramming mechanisms, potentially influencing male-specific physiological responses to SIM (Neuparth et al., 2020b; Alves et al., 2021).

Key KEEG Pathways (F0/F3 Intersection)

- **Metabolic and hormonal pathways (KEGG):**
  KEGG pathway enrichment highlighted metabolism-related pathways, including *Amino sugar and nucleotide sugar metabolism* (ko00520) and *Alanine, aspartate and glutamate metabolism* (ko00250), as well as broader *Metabolic pathways* (ko01100) and *Insulin resistance* (ko04931), reflecting potential systemic metabolic adjustments. Additionally, *Steroid hormone biosynthesis* (ko00140) was enriched, which is particularly relevant in *G. locusta*, as crustaceans rely on tightly regulated hormonal pathways for male reproductive function. Changes in steroid hormone production can therefore directly affect fertility, mating behavior or sex-related traits (Hosamani et al., 2017; Knigge et al., 2021).

To identify which clusters were associated with enriched functions, we consulted the Interactive Table within the Functional Enrichment tab and retrieved the corresponding clusters (Figure 15).

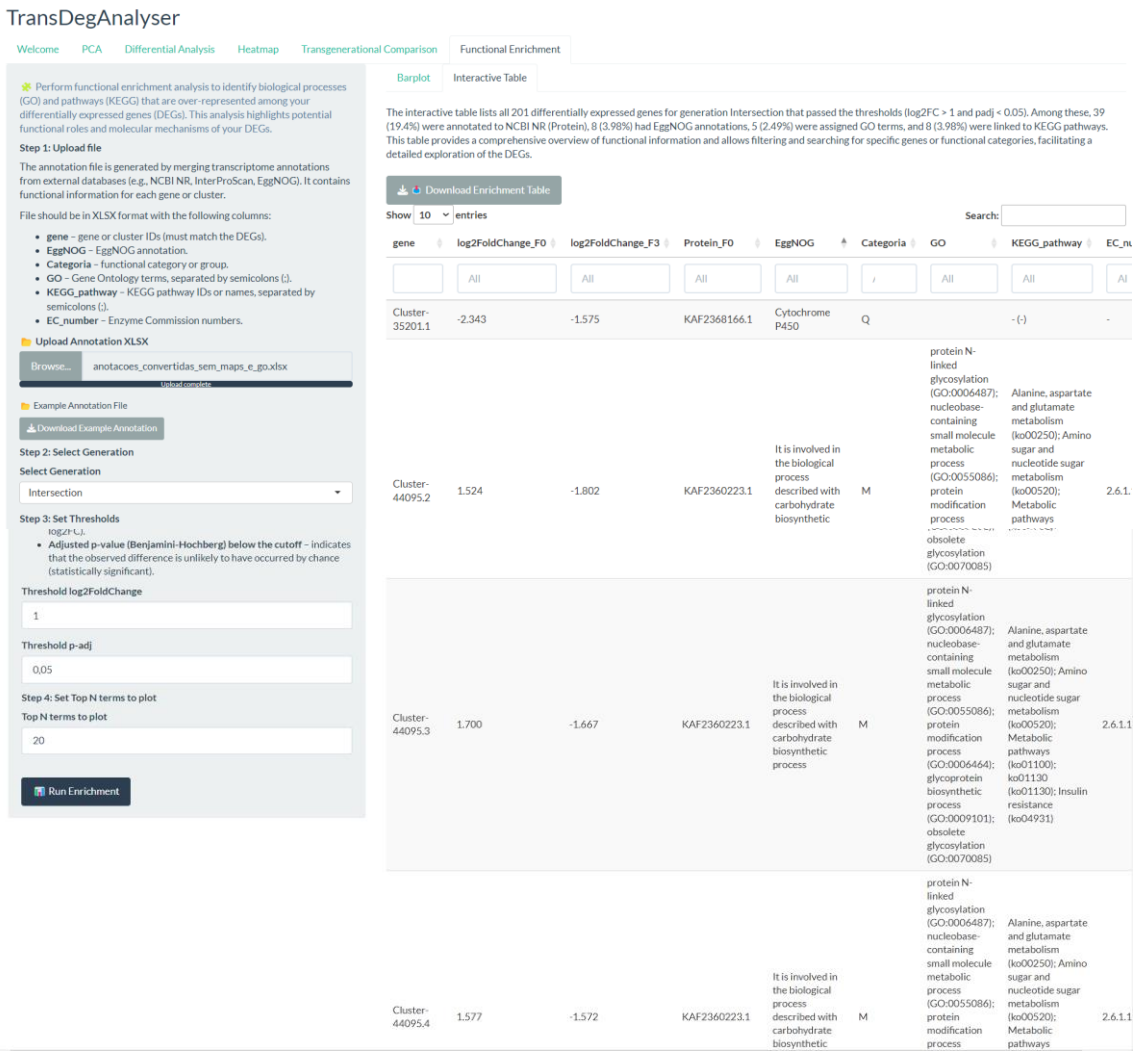


Figure 15 – Interactive table from the Functional Enrichment tab of TransDegAnalyser. It is required to upload an annotations file, select the generation to analyse and the significant thresholds. The interactive table presents a description for the % of annotated genes for each database and the respective table allows the exploration of DGEs and its respective enrichment analysis. This table provides a comprehensive overview of functional information and allows filtering and searching for specific genes or functional categories, facilitating a detailed exploration of the DEGs.

Clusters associated with enriched functions (F0/F3 Intersection)

Individual transcript clusters reveal how these functional shifts manifest at the gene level. For the intersection of the F0 and F3 generations, the annotated clusters obtained were:

- **Cluster-35201.1 (Cytochrome P450):**
  This gene is associated with detoxification processes (GO not assigned) (Lu et al., 2021). Its expression is downregulated in both F0 (log2FC = -2.34) and F3 (log2FC = -1.57), suggesting that the effect of SIM exposure on detoxification and possibly hormone metabolism persists across generations.

- **Cluster-44095.2/3/4 (Protein involved in carbohydrate biosynthesis):**
  These genes are linked to glycoprotein biosynthesis and protein modification (GO:0006487, GO:0006464). Interestingly, they increase in F0 (log2FC = 1.52–1.70) but decrease in F3 (log2FC = -1.57–1.80). This switch suggests that energy metabolism and protein modification may adjust or adapt across generations.

- **Cluster-44271.12 (Sulfotransferase):**
  This gene participates in steroid hormone biosynthesis (ko00140) and shows a similar flip, being slightly upregulated in F0 (log2FC = 1.22) but downregulated in F3 (log2FC = -1.33). Such changes could indicate endocrine adjustments for example in reproduction in males as a long-term effect of parental exposure.

- **Cluster-41387.2 (WD domain, G-beta repeat):**
  Associated with epigenetic regulation (GO:0045814, GO:0006479) and regulation of histone modification (GO:0031058), this gene is up in F0 (log2FC = 2.68) and down in F3 (log2FC = -2.34). This pattern suggest that SIM exposure initially activated epigenetic pathways in F0, likely involving changes in DNA methylation and histone modifications that regulate gene expression. In F3, the downregulation suggests that these epigenetic changes are either reset or compensated, reflecting transgenerational regulatory mechanisms that maintain cellular homeostasis and prevent persistent overactivation. Such modulation may have functional consequences for reproductive physiology or stress responses in the descendants, highlighting the dynamic nature of epigenetic inheritance (Alves et al., 2021; Rodrigues et al., 2024)

4.12.2. Enrichment Analysis of The Transgenerational DEGs (F3) with Apical Observations Correlations

For this analysis, the same methodological procedure as for the shared genes was applied in the TransDegAnalyser App, except for the generational selection which instead of choosing ‘Intersection’, we chose F3. The functional enrichment analysis of the core set of DEGs in F3 generations highlighted the transgenerational effects observed on several biologically relevant processes in male *G. locusta*. The barplot showed for all annotated DEGs, which were 5,390 for generation F3 that passed the thresholds (log2FC > 1 and padj < 0.05), the top 20 most expressed GO terms and KEGG pathways. Among these, 1,306 (24.23%) were annotated to NCBI NR (Protein), 494

(9.16%) had EggNOG annotations, 282 (5.23%) were assigned GO terms and 494 (9.16%) were linked to KEGG pathways (Figure 16).

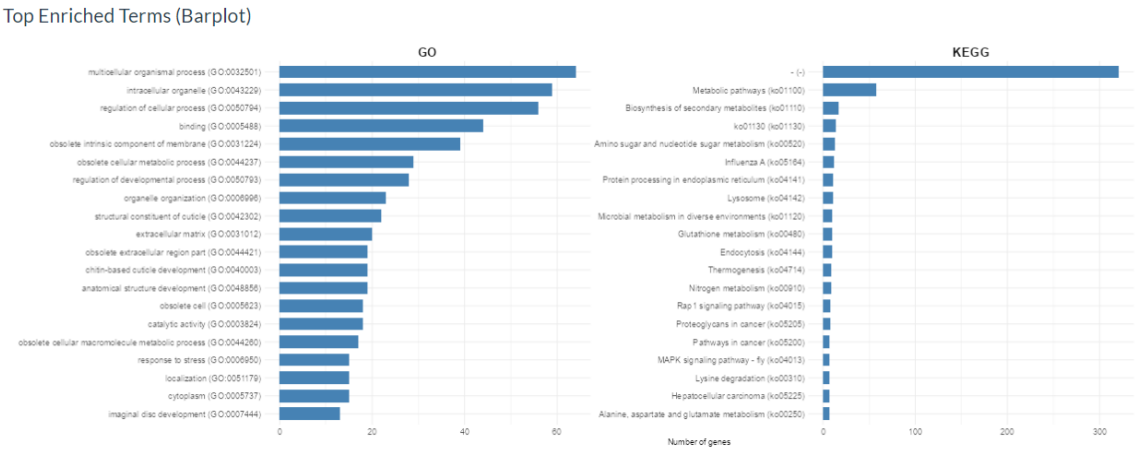


Figure 16 - The barplots show the top 20 enriched functional terms for generation F3. Bars represent the number of genes associated with each term. Of the 5390 differentially expressed genes, 24.23% were annotated to NCBI NR (Protein), 9.16% had EggNOG annotations, 5.23% were assigned GO terms, and 9.16% were linked to KEGG pathways. This visualization highlights the most prominent biological processes (GO) and pathways (KEGG) affected in this generation.

GO terms and KEGG pathways enrichment in F3:

Functional enrichment analysis of F3 males revealed that SIM exposure induced broad transcriptional changes spanning development, structural organization, metabolism and stress-response pathways (Figure 16).

GO enrichment:

GO enrichment highlighted multicellular organismal and developmental processes (e.g., GO:0032501, GO:0050793, GO:0048856), cellular and organelle organization (GO:0006996, GO:0043229), structural components of the cuticle and extracellular matrix (GO:0042302, GO:0031012), as well as metabolic and stress-response processes (GO:0044237, GO:0006950). These terms point to molecular shifts that can underlie the observed apical effects in *G. locusta*, including reduced body size, altered vacuole abundance and diameter and histological modifications in the tubular epithelium (Neuparth et al., 2014; 2020).

KEGG pathways enrichment:

KEGG enrichment further supported these findings, with significant terms related to general metabolic pathways (ko01100), amino sugar and nucleotide sugar metabolism (ko00520), glutathione metabolism (ko00480), and protein processing in the endoplasmic reticulum (ko04141). These pathways show that metabolism and cell maintenance change across generations. Signaling pathways like MAPK (ko04013) and Rap1 (ko04015) might be involved in the regulation of cell growth, differentiation and stress responses.

Clusters associated with enriched functions in F3 Transgenerational with Apical Correlations

Individual transcript clusters reveal how these functional shifts manifest at the gene level in F3 transgenerational males. Clusters associated with enriched functions are summarized in the Table below (Table 9).

Table 9 - Functional enrichment of F3 transgenerational male *G. locusta* genes.

Functional Category/ Cluster	EggNOG**Description**	**GO term / K**EGG	**log2FC**	**Function**
Multicellular & developmental processes				
38347.1	Zona Pellucida, ZP domain	GO:0045995	+7.865	Embryonic development; may influence gamete formation
2426.6	Insect cuticle protein	GO:0042302, GO:0032501	+8.583	Cuticle structure, reduced body size, structural remodeling
53297.0	PI3/PI4-kinase family	GO:0050793	+4.228	Tissue differentiation, growth regulation
33993.0	Microtubule motor activity	GO:0050793	+3.895	May relate to tubular epithelium thickening
55647.0	Kinesin family	GO:0000578	+4.594	Embryonic axis specification, vesicle transport
2663.31	Actin family	GO:0042692	-1.166	Muscle differentiation, cytoskeletal changes
49305.0	Rho GTPase	GO:0048814, GO:0050793	+3.145	Dendrite morphogenesis, developmental regulation
Cellular & organelle organization				
58460.0	Sequence-specific DNA binding	GO:0003677	+5.997	Transcription factor, TGF-beta signaling, cellular remodeling
34348.4	Ion channel regulatory protein UNC-93	GO:0006811	+5.449	Muscle system regulation, tubular epithelium adjustments
45971.0	UDP-glycosyltransferase	ko00140	+4.846	Steroid hormone biosynthesis, detoxification
17891.2	Protein secretion	ko04141	+4.044	ER function, protein processing
14477.4	Calcium ion binding	GO:0005509	+3.051	Golgi organization, ion homeostasis
43773.0	Mitochondria-eating protein	GO:0005739	-2.882	Organelle organization, energy availability
2663.28	Actin	GO:0043231	-2.172	Cytoskeleton and organelle organization
43244.2	Ribosomal RNA maturation	GO:0006364	-1.706	Protein synthesis capacity
37622.11	Immunoglobulin	GO:0043231	+2.522	Organelle organization, cell differentiation
Structural & extracellular matrix				
42284.3	Insect cuticle protein	GO:0042302	+36.970	Molting, cuticle biosynthesis
38347.1	ZP domain	GO:0045995	+7.865	ECM formation and organelle coordination
29955.3/4/5 & 44251.0	Chitinases / Glycosyl hydrolase	GO:0005975	+3.049–3.711	Cuticle degradation during molting
31093.14	Insect cuticle protein	GO:0042302	-3.387	Old cuticle component downregulation
Metabolic & stress-response processes				
48685.0 / 38765.x	Alpha-Anhydrase Carbonic	GO:0004089	+8.855	Calcification/sclerotization of cuticle
12688.2 / 6866.0	Metaloaminopeptidase M1	GO:0008233	+7.352	Protein recycling, molting
28774.9	Alpha-lactalbumin / Lysozyme C	GO:0006952	+6.243	Immune function, ECM degradation
55199.0	Serine/Threonine kinase – MAPK	GO:0004674	+4.353	Stress-response and development
49225.0	Glutathione S-transferase	GO:0004364	+2.493	Oxidative stress response
38659.9 / 35883.2	Pyruvate dehydrogenase / Cytochrome c oxidase	GO:0006090	-1.425 to -1.968	Reduced energy metabolism, trade-off for structural processes
Localization & binding functions				
3067.0	Chitin-binding / Chitin deacetylase	GO:0008061	+5.161	Processes old cuticle material
22641.0	MYND finger	GO:0046872	+8.819	Metal ion binding, supports calcification

49549.x / 42347.0	Lectin C-type domains	GO:0005515	+5.099	ECM organization, immune protection
37434.0	Syntaxin binding / Neurotransmitter transport	GO:0005488	+6.304	Vesicle transport, vacuole formation
45453.2	BTG family	GO:0008285	+4.397	Negative regulation of proliferation, focus on structural protein synthesis
58460.0	Sequence-specific DNA binding	GO:0003677	+5.997	Master regulator of molting, structural remodeling, stress-response genes

To further explore these enriched functions, individual transcript clusters associated with these GO terms and KEGG pathways were analysed for their potential impact on apical outcomes in F3 transgenerational males.

- **Structural and developmental regulation:**

Clusters related to cuticle proteins (e.g., Cluster-2426.6, Cluster-42284.3) and chitin-binding proteins (Cluster-3067.0) were strongly upregulated in F3 males. Since chitin is the main component of the exoskeleton in *G. locusta*, these transcriptional changes suggest active remodeling of the exoskeleton, encompassing both synthesis and degradation of cuticle and chitin during growth and molting. Such regulation likely helps build the extracellular matrix, supports tissue differentiation and keeps structural integrity.These molecular patterns are consistent with the reduced body size observed in both F0 and F3 males (Neuparth et al., 2014), supporting a mechanistic link between transcriptional remodeling of structural components and apical growth effects.

- **Cellular organization:**

Upregulation of ion channel regulators (Cluster-34348.4) and calcium-binding proteins (Cluster-14477.4) in F3 males suggests adjustments in ion transport and homeostasis, potentially underpinning the increased tubular epithelium thickness and the elevated number and diameter of non-digestive vacuoles observed in F0 exposed males (Neuparth et al., 2014). Additionally, vesicle transport–related genes (Cluster-37434.0) were enriched, indicating coordinated intracellular trafficking and extracellular matrix assembly. These transcriptional patterns are consistent with the F0 histological observations by Neuparth et al. (2014) and suggest that similar cellular machinery remains transcriptionally primed in F3 descendants.

- **Metabolic and stress responses:**

Glutathione S-transferase (Cluster-49225.0) and carbonic anhydrases (Clusters-48685.0/38765.x) were upregulated, pointing to enhanced oxidative stress defense and calcification/pH regulation. Conversely, pyruvate dehydrogenase and cytochrome c oxidase clusters (Clusters-38659.9 / 35883.2) were downregulated, indicating a reduction in basal energy metabolism. Upregulation of metal ion-binding

proteins (Cluster-22641.0) suggests that interactions with essential metals, like Fe, Zn, Cu, and Ca are altered. Because many enzymes that manage oxidation need Fe or Cu, these changes may show oxidative stress and activation of detox pathways (Lushchak, 2011). This shift in metabolism likely moves energy from growth to stress responses, which fits with the smaller body size observed.

Overall, these findings indicate that transgenerational SIM exposure in males primarily impacts multiple biological functions that are directly linked to observed apical outcomes. Altered clusters related to **structural and developmental regulation** (e.g., cuticle and chitin-binding proteins) correlate with reduced body size and active exoskeleton remodeling. Changes in **cellular organization** genes (e.g., ion channel regulators, calcium-binding proteins, vesicle transport genes) align with increased tubular epithelium thickness and the higher number and diameter of non-digestive vacuoles. Modulation of **metabolic and stress-response pathways** (e.g., glutathione S-transferases, carbonic anhydrases, pyruvate dehydrogenase) reflects enhanced oxidative stress defense and energy reallocation, possibly supporting structural maintenance at the cost of growth. Taken together, the persistence of these transcriptional programs across F0 and F3 suggests epigenetic mechanisms maintain molecular responses even in the absence of direct exposure, providing a mechanistic link between molecular signatures and transgenerational phenotypic effects.

These molecular alterations are consistent with previously documented phenotypic effects, namely reduced body size and impaired reproductive output in both the parental and transgenerational lineages following SIM exposure in the F0 generation. Such changes have the potential to disrupt ecosystem functioning. Reduced growth decreases the energetic value of *G. locusta* as prey, meaning predators may need to consume a greater number of individuals to meet their metabolic demands. At the same time, impaired reproduction can lead to lower population densities of *G. locusta*, reducing prey availability. Together, these effects may propagate through the food web, compromising energy transfer across trophic levels and potentially affecting higher-level consumers, including commercially relevant species that humans rely on. Thus, transgenerational impacts of SIM have the potential to induce broader ecological consequences that extend beyond the directly exposed organisms.

4.13. Limitations and Future Perspectives

This study has some limitations. First, the analysis used three input files, two of which came from intermediate RNA-Seq steps (FeatureCounts count matrix and functional annotation files). Creating these files required extra bash scripts for the construction of the count matrix and Python scripts for annotation, which had to be run outside the main tool. This approach allowed us to study gene expression and functional enrichment, but it added variability and made reproducing the analysis more complex.

Second, we focused on transcriptomic data. This shows gene expression, but not protein levels or activity. So, our interpretations of biological pathways are indirect. Similarly, although we found genes linked to epigenetic regulation, we did not measure DNA methylation or histone modifications. Evidence for epigenetic effects is therefore indirect and needs more study.

For future studies, a more integrated and user-friendly pipeline could help. Automating annotation, normalization, and preprocessing in a single workflow would reduce variability and simplify reproducibility. Using containerized environments (e.g., Docker or Conda) or workflow systems (e.g., Snakemake or Nextflow) would also improve reproducibility. Adding quality control metrics, visualization tools, and standardized output formats would strengthen confidence in analyses and make results easier to interpret.

Finally, integrating multi-omics approaches combining RNA-Seq with proteomics, metabolomics and epigenetics would give a fuller picture of how gene expression affects physiology. The workflow could also be applied to study other environmental contaminants, helping us understand how pollution affects inheritance and adaptation in natural populations.

5. Conclusions

This study provides a comprehensive evaluation of the direct and transgenerational effects of SIM in *G. locusta* males, integrating high-resolution transcriptomic profiling with functional interpretation of phenotypic effects. By examining both the directly exposed parental generation (F0) and the transgenerational unexposed third-generation males (F3), this study identified persistent molecular alterations that likely underpin SIM-induced transgenerational phenotypic effects, including reduced body size.

The custom Shiny-based tool TransDegAnalyser, developed as part of this study allows interactive and reproducible exploration of male *G. locusta* transcriptome. TransDegAnalyser provides dynamic visualization of PCA, volcano plots, heatmaps and functional enrichment, allowing researchers - even those without bioinformatics expertise - to easily perform analysis, explore and interpret RNA-Seq data. While designed for *G. locusta*, TransDegAnalyser can be readily adapted for transcriptomic analyses in other species, offering a user-friendly framework for high-dimensional data exploration.Overall, SIM exposure induces heritable transgenerational effects in males that are potentially related with epigenetic programming, with persistent transcriptional changes spanning structural, metabolic and stress-response pathways. The integration of a male-specific transcriptome, functional annotation and interactive analysis via TransDegAnalyser provides a robust framework to link molecular signatures to phenotypic effects and offers valuable resources for future ecotoxicological studies in non-model aquatic species.

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Alves, N., Neuparth, T., Barros, S., & Santos, M. M. (2021). The anti-lipidemic drug simvastatin modifies epigenetic biomarkers in the amphipod Gammarus locusta. *Ecotoxicology and Environmental Safety*, *209*, 111849. https://doi.org/10.1016/j.ecoenv.2020.111849

Andrews, S. (2010). *FastQC a quality control tool for high throughput sequence data*. Babraham.ac.uk. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrrano, J. A., Tietge, J. E., & Villeneuve, D. L. (2010). Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, *29*(3), 730–741. https://doi.org/10.1002/etc.34

Ankley, G. T., Daston, G. P., Degitz, S. J., Denslow, N. D., Hoke, R. A., Kennedy, S. W., Miracle, A. L., Perkins, E. J., Snape, J., Tillitt, D. E., Tyler, C. R., & Versteeg, D. (2006). Toxicogenomics in Regulatory Ecotoxicology. *Environmental Science & Technology*, *40*(13), 4055–4065. https://doi.org/10.1021/es0630184

Anway, M. D., & Skinner, M. K. (2006). Epigenetic Transgenerational Actions of Endocrine Disruptors. *Endocrinology*, *147*(6), s43–s49. https://doi.org/10.1210/en.2005-1058

Attardo, S., Musumeci, O., Velardo, D., & Toscano, A. (2022). Statins Neuromuscular Adverse Effects. *International Journal of Molecular Sciences*, *23*(15), 8364. https://doi.org/10.3390/ijms23158364

Aus der Beek, T., Weber, F.-A., Bergmann, A., Hickmann, S., Ebert, I., Hein, A., & Küster, A. (2016). Pharmaceuticals in the environment-Global occurrences and perspectives. *Environmental Toxicology and Chemistry*, *35*(4), 823–835. https://doi.org/10.1002/etc.3339

Baker, T. R., King-Heiden, T. C., Peterson, R. E., & Heideman, W. (2014). Dioxin induction of transgenerational inheritance of disease in zebrafish. *Molecular and Cellular Endocrinology*, *398*(1-2), 36–41. https://doi.org/10.1016/j.mce.2014.08.011

Bellés, X., Martín, D., & Piulachs, M.-D. (2005). THE MEVALONATE PATHWAY AND THE SYNTHESIS OF JUVENILE HORMONE IN INSECTS. *Annual Review of Entomology*, *50*(1), 181–199. https://doi.org/10.1146/annurev.ento.50.071803.130356

Beverly, Brandiese E. J., Lambright, C. S., Furr, J. R., Sampson, H., Wilson, V. S., McIntyre, B. S., Foster, Paul M. D., Travlos, G., & Gray, L. E., Jr. (2014). Simvastatin and Dipentyl Phthalate Lower Ex Vivo Testicular Testosterone Production and Exhibit Additive Effects on Testicular Testosterone and Gene Expression Via Distinct Mechanistic Pathways in the Fetal Rat. *Toxicological Sciences*, *141*(2), 524–537. https://doi.org/10.1093/toxsci/kfu149

Bhandari, R. K., vom Saal, F. S., & Tillitt, D. E. (2015). Transgenerational effects from early developmental exposures to bisphenol A or 17α-ethinylestradiol in medaka, Oryzias latipes. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep09303

Bioinformatics at the National Cancer Institute. (2025). *Lesson5: Visualizing clusters with heatmap and dendrogram - Data Visualization with R*. Bioinformatics.ccr.cancer.gov. https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/Lesson5_intro_to_ggplot/

Blalock, B. J., Robinson, W. E., Loguinov, A., Vulpe, C. D., Krick, K. S., & Poynton, H. C. (2018). Transcriptomic and Network Analyses Reveal Mechanistic-Based Biomarkers of Endocrine Disruption in the Marine Mussel, Mytilus edulis. *Environmental Science & Technology*, *52*(16), 9419–9430. https://doi.org/10.1021/acs.est.8b01604

Blonç, M., Lima, J., Joan Carles Balasch, Tort, L., Gravato, C., & Teles, M. (2023). Elucidating the Effects of the Lipids Regulators Fibrates and Statins on the Health Status of Finfish Species: A Review. *Animals*, *13*(5), 792–792. https://doi.org/10.3390/ani13050792

Boleda, M. R., Galceran, M. T., & Ventura, F. (2011). Behavior of pharmaceuticals and drugs of abuse in a drinking water treatment plant (DWTP) using combined conventional and ultrafiltration and reverse osmosis (UF/RO) treatments. *Environmental Pollution*, *159*(6), 1584–1591. https://doi.org/10.1016/j.envpol.2011.02.051

Bonduriansky, R., & Day, T. (2009). Nongenetic Inheritance and Its Evolutionary Implications. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 103–125. https://doi.org/10.1146/annurev.ecolsys.39.110707.173441

Bossus, M. C., Guler, Y. Z., Short, S. J., Morrison, E. R., & Ford, A. T. (2014). Behavioural and transcriptional changes in the amphipod Echinogammarus marinus exposed to two antidepressants, fluoxetine and sertraline. *Aquatic Toxicology*, *151*, 46–56. https://doi.org/10.1016/j.aquatox.2013.11.025

Boxall, A. B. A., Rudd, M. A., Brooks, B. W., Caldwell, D. J., Choi, K., Hickmann, S., Innes, E., Ostapyk, K., Staveley, J. P., Verslycke, T., Ankley, G. T., Beazley, K.

F., Belanger, S. E., Berninger, J. P., Carriquiriborde, P., Coors, A., DeLeo, P. C., Dyer, S. D., Ericson, J. F., & Gagné, F. (2012). Pharmaceuticals and Personal Care Products in the Environment: What Are the Big Questions? *Environmental Health Perspectives*, *120*(9), 1221–1229. https://doi.org/10.1289/ehp.1104477

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Buesen, R., Chorley, B. N., da Silva Lima, B., Daston, G., Deferme, L., Ebbels, T., Gant, T. W., Goetz, A., Greally, J., Gribaldo, L., Hackermüller, J., Hubesch, B., Jennen, D., Johnson, K., Kanno, J., Kauffmann, H.-M., Laffont, M., McMullen, P., Meehan, R., & Pemberton, M. (2017). Applying 'omics technologies in chemicals risk assessment: Report of an ECETOC workshop. *Regulatory Toxicology and Pharmacology*, *91*, S3–S13. https://doi.org/10.1016/j.yrtph.2017.09.002

Burggren, W. (2016). Epigenetic Inheritance and Its Role in Evolutionary Biology: Re-Evaluation and New Perspectives. *Biology*, *5*(2), 24. https://doi.org/10.3390/biology5020024

Burns, J. A., Daniels, J., Becker, K. P., Casagrande, D., Roberts, P., Orenstein, E., Vogt, D. M., Teoh, Z. E., Wood, R., Yin, A. H., Genot, B., Wood, R. J., Katija, K., Phillips, B. T., & Gruber, D. F. (2024). Transcriptome sequencing of seven deep marine invertebrates. *Scientific Data*, *11*(1), 679. https://doi.org/10.1038/s41597-024-03533-4

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., & Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of

individual B cells. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms16027

Camilo Escobar-Sierra, Hassan, S., Weichert, F. G., Henrik Aronsson, Lampert, K. P., Henner Hollert, Backhaus, T., & Inostroza, P. A. (2025). De novo transcriptome assembly and annotation of the common freshwater amphipod (Gammarus pulex) a valuable resource for ecotoxicogenomics. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-05872-2

Camilo Escobar-Sierra, & Lampert, K. P. (2024). Field application of de novo transcriptomic analysis to evaluate the effects of sublethal freshwater salinization on Gasterosteus aculeatus in urban streams. *PLoS ONE*, *19*(3), e0298213–e0298213. https://doi.org/10.1371/journal.pone.0298213

Campo, S., Andreone, L., Ambao, V., Urrutia, M., Calandra, R. S., & Rulli, S. B. (2019). Hormonal Regulation of Follicle-Stimulating Hormone Glycosylation in Males. *Frontiers in Endocrinology*, *10*. https://doi.org/10.3389/fendo.2019.00017

Carvalho, I. T., & Santos, L. (2016). Antibiotics in the aquatic environments: A review of the European scenario. *Environment International*, *94*, 736–757. https://doi.org/10.1016/j.envint.2016.06.025

Chen, J., Zha, L., Xu, J., & Chen, G. (2024). Prolonged treatment of compactin, atorvastatin, lovastatin or simvastatin decreased glucose consumption and expressions of proteins for glucose metabolism and insulin signaling in L6 skeletal muscle cells. *Food, Nutrition and Health*, *1*(1). https://doi.org/10.1007/s44403-024-00001-0

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0881-8

Costa, E., Johnson, K. J., Walker, C. A., & O'Brien, J. M. (2024). Transcriptomic point of departure determination: a comparison of distribution-based and gene set-based approaches. *Frontiers in Genetics*, *15*. https://doi.org/10.3389/fgene.2024.1374791

Costa, F. O., Correia, A. D., & Costa, M. H. (1998). Acute Marine Sediment Toxicity: A Potential New Test with the AmphipodGammarus locusta. *Ecotoxicology and Environmental Safety*, *40*(1-2), 81–87. https://doi.org/10.1006/eesa.1998.1646

Costa, F. O., & Costa, M. H. (1999). Life history of the amphipod Gammarus locusta in the Sado estuary (Portugal). *Acta Oecologica*, *20*(4), 305–314. https://doi.org/10.1016/s1146-609x(99)00136-8

Costa, F. O., Neuparth, T., Correia, A. D., & Helena Costa, M. (2005). Multi-level assessment of chronic toxicity of estuarine sediments with the amphipod Gammarus locusta: II. Organism and population-level endpoints. *Marine Environmental Research*, *60*(1), 93–110. https://doi.org/10.1016/j.marenvres.2004.08.005

Costa, F., Neuparth, T., Theodorakis, C., Costa, M., & Shugart, L. (2004). RAPD analysis of southern populations of Gammarus locusta: comparison with allozyme data and ecological inferences. *Marine Ecology Progress Series*, *277*, 197–207. https://doi.org/10.3354/meps277197

Costa, R., Ferreira, C., Alves, A., Nunes, S., Reis, F., Malva, J., & Viana, S. (2025). Lipid-lowering statins and polyphenol-based supplementation: a scoping review on drug-food interaction potential. *Frontiers in Pharmacology*, *16*. https://doi.org/10.3389/fphar.2025.1541871

Cunha, V., Santos, M. M., P. Moradas-Ferreira, & Ferreira, M. (2016). Simvastatin effects on detoxification mechanisms in Danio rerio embryos. *Environmental Science and Pollution Research*, *23*(11), 10615–10629. https://doi.org/10.1007/s11356-016-6547-y

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Davies, J. T., Delfino, S. F., Feinberg, C. E., Johnson, M. F., Nappi, V. L., Olinger, J. T., Schwab, A. P., & Swanson, H. I. (2016). Current and Emerging Uses of Statins in Clinical Therapeutics: A Review. *Lipid Insights*, *9*, LPI.S37450. https://doi.org/10.4137/lpi.s37450

Dhangar, K., & Kumar, M. (2020). Tricks and tracks in removal of emerging contaminants from the wastewater through hybrid treatment systems: A review. *Science of the Total Environment*, *738*, 140320. https://doi.org/10.1016/j.scitotenv.2020.140320

Donelson, J. M., Salinas, S., Munday, P. L., & Shama, L. N. S. (2017). Transgenerational plasticity and climate change experiments: Where do we go from here? *Global Change Biology*, *24*(1), 13–34. https://doi.org/10.1111/gcb.13903

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Ellesat, K. S., Tollefsen, K.-E., Åsberg, A., Thomas, K. V., & Hylland, K. (2010). Cytotoxicity of atorvastatin and simvastatin on primary rainbow trout (Oncorhynchus mykiss) hepatocytes. *Toxicology in Vitro*, *24*(6), 1610–1618. https://doi.org/10.1016/j.tiv.2010.06.006

EMBL-EBI. (2025a). *QuickGO*. Ebi.ac.uk. https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0055086

EMBL-EBI. (2025b). *QuickGO API EMBL-EBI*. Ebi.ac.uk. https://www.ebi.ac.uk/QuickGO/api/index.html

Endo, A. (2010). A historical perspective on the discovery of statins. *Proceedings of the Japan Academy, Series B*, *86*(5), 484–493. https://doi.org/10.2183/pjab.86.484

ENVIRONMENT DIRECTORATE CHEMICALS AND BIOTECHNOLOGY COMMITTEE OECD Omics Reporting Framework (OORF): Guidance on reporting elements for the regulatory use of omics data from laboratory-based toxicology studies Series on Testing and Assessment No. 390. (2023). https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/11/oecd-omics-reporting-framework-oorf-guidance-on-reporting-elements-for-the-regulatory-use-of-omics-data-from-laboratory-based-toxicology-studies_9f10a387/6bb2e6ce-en.pdf

Ericsson, J., & Gustav Dallner. (1993). Distribution, Biosynthesis, and Function of Mevalonate Pathway Lipids. *Sub-Cellular Biochemistry/Subcellular Biochemistry*, 229–272. https://doi.org/10.1007/978-1-4615-2912-5_11

Fakhri, M. S., Nasrin Ghassemi Barghi, Mahdis Moradnia Mehdikhanmahaleh, Mahdi, M., Mousavi, T., Ramin Rezaee, Mojtaba Daghighi, & Abdollahi, M. (2024). Pharmaceutical wastewater toxicity: An ignored threat to the public health. *Sustainable Environment*, *10*(1). https://doi.org/10.1080/27658511.2024.2322821

Garcia-Reyero, N., & Perkins, E. J. (2010). Systems biology: Leading the revolution in ecotoxicology. *Environmental Toxicology and Chemistry*, *30*(2), 265–273. https://doi.org/10.1002/etc.401

Geffard, O., Xuereb, B., Chaumot, A., Geffard, A., Biagianti, S., Noël, C., Abbaci, K., Garric, J., Charmantier, G., & Charmantier-Daures, M. (2010). Ovarian cycle and embryonic development in Gammarus fossarum: Application for reproductive

toxicity assessment. *Environmental Toxicology and Chemistry*, *29*(10), 2249–2259. https://doi.org/10.1002/etc.268

Gerhardt, A., Bloor, M., & Mills, C. L. (2011). Gammarus: Important Taxon in Freshwater and Marine Changing Environments. *International Journal of Zoology*, *2011*, 1–2. https://doi.org/10.1155/2011/524276

Ghosh, M., & Corneliu Duca, R. (2021). Biomarkers of Exposure, Effect and Susceptibility to Environmental and Occupational Chemicals. In *Frontiers research topics*. Frontiers Media. https://doi.org/10.3389/978-2-88971-834-4

Gil, A., Garcia, A., Fernandez, M., Vicente, M. A., González, B., Rives, V., & S.A. Korili. (2017). Effect of dopants on the structure of titanium oxide used as a photocatalyst for the removal of emergent contaminants. *Journal of Industrial and Engineering Chemistry*, *53*, 183–191. https://doi.org/10.1016/j.jiec.2017.04.024

Gilchrist, M. J., Sobral, D., Khoueiry, P., Fabrice Daian, Laporte, B., Ilya Patrushev, Matsumoto, J., Dewar, K., Hastings, K. E. M., Yutaka Satou, Lemaire, P., & Rothbächer, U. (2015). A pipeline for the systematic identification of non-redundant full-ORF cDNAs for polymorphic and evolutionary divergent genomes: Application to the ascidian Ciona intestinalis. *Developmental Biology*, *404*(2), 149–163. https://doi.org/10.1016/j.ydbio.2015.05.014

GitHub. (2025). *GitHub*. GitHub. https://github.com/

Goldstein, J. L., & Brown, M. S. (1990). Regulation of the mevalonate pathway. *Nature*, *343*(6257), 425–430. https://doi.org/10.1038/343425a0

Goodman, W. G., & Cusson, M. (1976). *The Juvenile Hormones*. 310–365. https://doi.org/10.1016/b978-0-12-384749-2.10008-1

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., & Friedman, N. (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

Graham van Aggelen, Ankley, G. T., Baldwin, W. M., Bearden, D. W., Benson, W. E., Chipman, J. K., Collette, T., Craft, J. A., Denslow, N. D., Embry, M. R., Falciani, F., George, S. L., Helbing, C. C., Hoekstra, P. F., Iguchi, T., Yoshi Kagami, Ioanna Katsiadaki, Kille, P., Liu, L., & Lord, P. (2010). Integrating Omic Technologies into Aquatic Ecological Risk Assessment and Environmental Monitoring: Hurdles, Achievements, and Future Outlook. *Environmental Health Perspectives*, *118*(1), 1–5. https://doi.org/10.1289/ehp.0900985

Grenni, P., Ancona, V., & Barra Caracciolo, A. (2018). Ecological effects of antibiotics on natural ecosystems: A review. *Microchemical Journal*, *136*, 25–39. https://doi.org/10.1016/j.microc.2017.02.006

Guadamuz, J. S., Shooshtari, A., & Qato, D. M. (2022). Global, regional and national trends in statin utilisation in high-income and low/middle-income countries, 2015–2020. *BMJ Open*, *12*(9), e061350. https://doi.org/10.1136/bmjopen-2022-061350

Guerra, B., Recio, C., Aranda-Tavío, H., Guerra-Rodríguez, M., García-Castellano, J. M., & Fernández-Pérez, L. (2021). The Mevalonate Pathway, a Metabolic Target in Cancer Therapy. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.626971

Haas et al. (2013). *Page Redirection*. Github.io. https://transdecoder.github.io/

Harney, E., Paterson, S., Collin, H., Chan, B. H. K., Bennett, D., & Plaistow, S. J. (2022). Pollution induces epigenetic effects that are stably transmitted across multiple generations. *Evolution Letters*. https://doi.org/10.1002/evl3.273

Head, J. A. (2014). Patterns of DNA Methylation in Animals: An Ecotoxicological Perspective. *Integrative and Comparative Biology*, *54*(1), 77–86. https://doi.org/10.1093/icb/icu025

Head, J. A., Ewald, J. D., & Basu, N. (2024). The DIKW of Transcriptomics in Ecotoxicology: Extracting Information, Knowledge, and Wisdom From Big Data. *Environmental Toxicology and Chemistry*. https://doi.org/10.1002/etc.5954

Heard, E., & Martienssen, Robert A. (2014). Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell*, *157*(1), 95–109. https://doi.org/10.1016/j.cell.2014.02.045

Helenius, A., & Aebi, M. (2004). Roles of N-Linked Glycans in the Endoplasmic Reticulum. *Annual Review of Biochemistry*, *73*(1), 1019–1049. https://doi.org/10.1146/annurev.biochem.73.011303.073752

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

HERNANDO, M., MEZCUA, M., FERNANDEZALBA, A., & BARCELO, D. (2006). Environmental risk assessment of pharmaceutical residues in wastewater effluents, surface waters and sediments. *Talanta*, *69*(2), 334–342. https://doi.org/10.1016/j.talanta.2005.09.037

Hime, G. R., Stonehouse, S. L., & Pang, T. Y. (2021). Alternative models for transgenerational epigenetic inheritance: Molecular psychiatry beyond mice and man. *World Journal of Psychiatry*, *11*(10), 711–735. https://doi.org/10.5498/wjp.v11.i10.711

Hiroki Itokawa, Hashimoto, T., & Murakami, T. (2003). Characterization of Municipal Wastewater Organic Components for Activated Sludge Process Modelling. *Environmental Engineering Research*, *40*, 41–52. https://doi.org/10.11532/proes1992.40.41

Hosamani, N., Reddy B, S., & Reddy P, R. (2017). Crustacean Molting: Regulation and Effects of Environmental Toxicants. *Journal of Marine Science: Research & Development*, *07*(05). https://doi.org/10.4172/2155-9910.1000236

Hsu, P.-C., Jhong, J.-Y., Huang, L.-P., Lee, K.-H., Chen, H.-P., & Guo, Y.-L. (2021). Transgenerational Effects of Di(2-Ethylhexyl) Phthalate on Anogenital Distance, Sperm Functions and DNA Methylation in Rat Offspring. *International Journal of Molecular Sciences*, *22*(8), 4131. https://doi.org/10.3390/ijms22084131

Huang, Z., Lai, P. F., Alexander, Haslam, S. M., Dell, A., Hugh, & Johnson, M. R. (2023). Roles of N-linked glycosylation and glycan-binding proteins in placentation: trophoblast infiltration, immunomodulation, angiogenesis, and pathophysiology. *Biochemical Society Transactions*, *51*(2), 639–653. https://doi.org/10.1042/bst20221406

Huerlimann, R., Wade, N. M., Gordon, L., Montenegro, J. D., Goodall, J., McWilliam, S., Tinning, M., Siemering, K., Giardina, E., Donovan, D., Sellars, M. J., Cowley, J. A., Condon, K., Coman, G. J., Khatkar, M. S., Raadsma, H. W., Maes, G. E., Zenger, K. R., & Jerry, D. R. (2018). De novo assembly, characterization, functional annotation and expression patterns of the black tiger shrimp (Penaeus monodon) transcriptome. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-31148-4

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Istvan, E. S. (2001). Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. *Science*, *292*(5519), 1160–1164. https://doi.org/10.1126/science.1059344

Ivosev, G., Burton, L., & Bonner, R. (2008). Dimensionality Reduction and Visualization in Principal Component Analysis. *Analytical Chemistry*, *80*(13), 4933–4944. https://doi.org/10.1021/ac800110w

J.D. Groopman. (2016). Biomarkers of Exposure, Effect, and Susceptibility. *Elsevier EBooks*, 188–201. https://doi.org/10.1016/b978-0-12-801238-3.64104-1

Jjemba, P. K. (2006). Excretion and ecotoxicity of pharmaceutical and personal care products in the environment. *Ecotoxicology and Environmental Safety*, *63*(1), 113–130. https://doi.org/10.1016/j.ecoenv.2004.11.011

Joshua Niklas Ebner. (2021). Trends in the Application of “Omics” to Ecotoxicology and Stress Ecology. *Genes*, *12*(10), 1481–1481. https://doi.org/10.3390/genes12101481

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, *12*(3). https://doi.org/10.1002/ctm2.694

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2018). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, *47*(D1), D590–D595. https://doi.org/10.1093/nar/gky962

Kasprzak, P., Mitchell, L., Kravchuk, O., & Timmins, A. (2020). Six Years of Shiny in Research - Collaborative Development of Web Tools in R. *R Journal*, *12*(2), 155–155. https://doi.org/10.32614/rj-2021-004

Kasprzyk-Hordern, B., Dinsdale, R. M., & Guwy, A. J. (2009). The removal of pharmaceuticals, personal care products, endocrine disruptors and illicit drugs during wastewater treatment and its impact on the quality of receiving waters. *Water Research*, *43*(2), 363–380. https://doi.org/10.1016/j.watres.2008.10.047

KEGG API. (2024). Www.kegg.jp. https://www.kegg.jp/kegg/rest/keggapi.html

Keiichi Kakui, Fleming, J. F., Mori, M., Fujiwara, Y., & Arakawa, K. (2021). Comprehensive Transcriptome Sequencing of Tanaidacea with Proteomic Evidences for Their Silk. *Genome Biology and Evolution*, *13*(12). https://doi.org/10.1093/gbe/evab281

Kizilaslan, M., Braz, C. U., Townsend, J., Taylor, T., Crenshaw, T. D., & Khatib, H. (2025). Transgenerational Epigenetic and Phenotypic Inheritance Across Five Generations in Sheep. *International Journal of Molecular Sciences*, *26*(13), 6412. https://doi.org/10.3390/ijms26136412

Knigge, T., LeBlanc, G. A., & Ford, A. T. (2021). A Crab Is Not a Fish: Unique Aspects of the Crustacean Endocrine System and Considerations for Endocrine Toxicology. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.587608

Landers, T. F., Cohen, B., Wittum, T. E., & Larson, E. L. (2012). A Review of Antibiotic Use in Food Animals: Perspective, Policy, and Potential. *Public Health Reports*, *127*(1), 4–22. https://doi.org/10.1177/003335491212700103

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1). https://doi.org/10.1186/1471-2105-9-559

Le-Minh, N., Khan, S. J., Drewes, J. E., & Stuetz, R. M. (2010). Fate of antibiotics during municipal water recycling treatment processes. *Water Research*, *44*(15), 4295–4323. https://doi.org/10.1016/j.watres.2010.06.020

Li, B., Zhang, Y., Du, J., Liu, C., Zhou, G., Li, M., & Yan, Z. (2025). Application of Multi-Omics Techniques in Aquatic Ecotoxicology: A Review. *Toxics*, *13*(8), 653–653. https://doi.org/10.3390/toxics13080653

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, S., Wagner, C. A., Friesen, J. A., & Borst, D. W. (2003). *3-Hydroxy-3-methylglutaryl-coenzyme A reductase in the lobster mandibular organ: regulation by the eyestalk*. *134*(2), 147–155. https://doi.org/10.1016/s0016-6480(03)00246-6

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lind, M. I., & Spagopoulou, F. (2018). Evolutionary consequences of epigenetic inheritance. *Heredity*, *121*(3), 205–209. https://doi.org/10.1038/s41437-018-0113-y

Littarru, G. P., & Langsjoen, P. (2007). Coenzyme Q10 and statins: Biochemical and clinical implications. *Mitochondrion*, *7*, S168–S174. https://doi.org/10.1016/j.mito.2007.03.002

Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature*, *405*(6788), 827–836. https://doi.org/10.1038/35015701

Lombard, J., & Moreira, D. (2010). Origins and Early Evolution of the Mevalonate Pathway of Isoprenoid Biosynthesis in the Three Domains of Life. *Molecular Biology and Evolution*, *28*(1), 87–99. https://doi.org/10.1093/molbev/msq177

López-Landavery, E. A., Galindo-Sánchez, C. E., López-Galindo, L., Ramírez-Álvarez, N., Anaid Saavedra-Flores, Amador-Cano, G., Ventura-López, C., San, P., & Hernández-Herrera, R. I. (2023). Hydrocarbons occurrence and transcriptomic response of oyster Crassostrea virginica from lagoons of the Southern Gulf of Mexico. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1085858

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, K., Song, Y., & Zeng, R. (2021). The role of cytochrome P450-mediated detoxification in insect adaptation to xenobiotics. *Current Opinion in Insect Science*, *43*, 103–107. https://doi.org/10.1016/j.cois.2020.11.004

Lushchak, V. I. (2011). Environmentally induced oxidative stress in aquatic animals. *Aquatic Toxicology*, *101*(1), 13–30. https://doi.org/10.1016/j.aquatox.2010.10.006

Lv, M., Sun, Q., Hu, A., Hou, L., Li, J., Cai, X., & Yu, C.-P. (2014). Pharmaceuticals and personal care products in a mesoscale subtropical watershed and their application as sewage markers. *Journal of Hazardous Materials*, *280*, 696–705. https://doi.org/10.1016/j.jhazmat.2014.08.054

Machado, A. M., Fernández-Boo, S., Nande, M., Pinto, R., Costas, B., & Castro, L. F. C. (2022). The male and female gonad transcriptome of the edible sea urchin, Paracentrotus lividus: Identification of sex-related and lipid biosynthesis genes. *Aquaculture Reports*, *22*, 100936. https://doi.org/10.1016/j.aqrep.2021.100936

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00013

MacManes, M. D., & Eisen, M. B. (2013). Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ*, *1*, e113. https://doi.org/10.7717/peerj.113

Manni, M., Berkeley, M. R., Mathieu Seppey, Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2106.11799

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. https://doi.org/10.1101/gr.079558.108

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682. https://doi.org/10.1038/nrg3068

McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2018). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics*, *20*(6), 2044–2054. https://doi.org/10.1093/bib/bby067

Menezes, C., Cavalcante, K., Silva, Lúcia, A., Duncan, W. P., Silva, J. C., Lopes, F. M., Artoni, R. F., & Matoso, D. A. (2025). Transcriptomic Analysis of Tambaqui (Colossoma macropomum) Exposed to Trichlorfon-Induced Toxicity. *Animals*, *15*(12), 1807–1807. https://doi.org/10.3390/ani15121807

MEYER, E., AGLYAMOVA, G. V., & MATZ, M. V. (2011). Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Molecular Ecology*, no-no. https://doi.org/10.1111/j.1365-294x.2011.05205.x

Miao, X.-S., & Metcalfe, C. D. (2003). Determination of cholesterol-lowering statin drugs in aqueous samples using liquid chromatography–electrospray ionization tandem mass spectrometry. *Journal of Chromatography A*, *998*(1-2), 133–141. https://doi.org/10.1016/s0021-9673(03)00645-9

Michael, I., Rizzo, L., McArdell, C. S., Manaia, C. M., Merlin, C., Schwartz, T., Dagot, C., & Fatta-Kassinos, D. (2013). Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: A review. *Water Research*, *47*(3), 957–995. https://doi.org/10.1016/j.watres.2012.11.027

Mirbahai, L., & Chipman, J. K. (2014). Epigenetic memory of environmental organisms: A reflection of lifetime stressor exposures. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *764-765*, 10–17. https://doi.org/10.1016/j.mrgentox.2013.10.003

Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, *62*(1), R21–R45. https://doi.org/10.1530/jme-18-0055

Mitchell, J., Pabon, N. A., Collier, Z. A., Egeghy, P. P., Cohen-Hubal, E., Igor Linkov, & Vallero, D. A. (2013). *A Decision Analytic Approach to Exposure-Based Chemical Prioritization*. *8*(8), e70911–e70911. https://doi.org/10.1371/journal.pone.0070911

Miziorko, H. M. (2011). Enzymes of the mevalonate pathway of isoprenoid biosynthesis. *Archives of Biochemistry and Biophysics*, *505*(2), 131–143. https://doi.org/10.1016/j.abb.2010.09.028

Moreno-Santillán, D. D., Machain-Williams, C., Hernández-Montes, G., & Ortega, J. (2019). De Novo Transcriptome Assembly and Functional Annotation in Five Species of Bats. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-42560-9

Morteza Sheikh-Assadi, Naderi, R., Seyed Alireza Salami, Mohsen Kafi, Reza Fatahi, Vahid Shariati, Martinelli, F., Cicatelli, A., Triassi, M., Guarino, F., Improta, G., & Manuel Gonzalo Claros. (2022). Normalized Workflow to Optimize Hybrid De Novo Transcriptome Assembly for Non-Model Species: A Case Study in Lilium ledebourii (Baker) Boiss. *Plants*, *11*(18), 2365–2365. https://doi.org/10.3390/plants11182365

Moyosore Joseph Adegbeye, Babatunde Oluwafemi Adetuyi, Anem Igirigi, Abosede Adisa, Valiollah Palangi, Aiyedun, S., Edwin Rafael Alvarado-Ramírez, Mona M.M.Y. Elghandour, Ofelia Márquez Molina, Oladipo, A. A., & Abdelfattah Z.M. Salem. (2024). Comprehensive insights into antibiotic residues in livestock products: Distribution, factors, challenges, opportunities, and implications for food safety and public health. *Food Control*, 110545–110545. https://doi.org/10.1016/j.foodcont.2024.110545

Nam, S.-E., Cheon, S., Do, S. D., Lee, S., Bae, D.-Y., Jeong, T.-Y., Han, D., & Rhee, J.-S. (2025). Multi-omics integration reveals underlying toxicological mechanisms of triclosan in the freshwater fish model Zacco platypus. *Aquatic Toxicology*, *287*, 107522–107522. https://doi.org/10.1016/j.aquatox.2025.107522

Neuparth, T., Alves, N., Machado, A. M., Pinheiro, M., Montes, R., Rodil, R., Barros, S., Ruivo, R., Castro, L. F. C., Quintana, J. B., & Santos, M. M. (2022). Neuroendocrine pathways at risk? Simvastatin induces inter and transgenerational disruption in the keystone amphipod Gammarus locusta. *Aquatic Toxicology*, *244*, 106095. https://doi.org/10.1016/j.aquatox.2022.106095

Neuparth, T., Costa, F. O., & Costa, M. H. (2002). Effects of Temperature and Salinity on Life History of the Marine Amphipod Gammarus locusta. Implications for Ecotoxicological Testing. *Ecotoxicology*, *11*(1), 61–73. https://doi.org/10.1023/a:1013797130740

Neuparth, T., Machado, A. M., Montes, R., Rodil, R., Barros, S., Alves, N., Ruivo, R., Castro, L. F. C., Quintana, J. B., & Santos, M. M. (2020a). Transcriptomic data on the transgenerational exposure of the keystone amphipod Gammarus locusta to simvastatin. *Data in Brief*, *32*, 106248. https://doi.org/10.1016/j.dib.2020.106248

Neuparth, T., Machado, A. M., Montes, R., Rodil, R., Barros, S., Alves, N., Ruivo, R., Castro, L. F. C., Quintana, J. B., & Santos, M. M. (2020b). Transgenerational inheritance of chemical-induced signature: A case study with simvastatin. *Environment International*, *144*, 106020. https://doi.org/10.1016/j.envint.2020.106020

Neuparth, T., Martins, C., Santos, C. B. de los, Costa, M. H., Martins, I., Costa, P. M., & Santos, M. M. (2014). Hypocholesterolaemic pharmaceutical simvastatin disrupts reproduction and population growth of the amphipod Gammarus locusta at the

ng/L range. *Aquatic Toxicology*, *155*, 337–347. https://doi.org/10.1016/j.aquatox.2014.07.009

Newman, M. C., & Newman, M. C. (2014). *Fundamentals of Ecotoxicology*. CRC Press. https://doi.org/10.1201/b17658

NIH HPC. (2015). *Rcorrector: Error correction for Illumina RNA-seq reads*. Nih.gov. https://hpc.nih.gov/apps/Rcorrector.html

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, *40*(20), 10084–10097. https://doi.org/10.1093/nar/gks804

Nunes, J. P. L. (2017). Statins and the cholesterol mortality paradox. *Scottish Medical Journal*, *62*(1), 19–23. https://doi.org/10.1177/0036933016681913

OECD Omics Reporting Framework (OORF): Guidance on reporting elements for the regulatory use of omics data from laboratory-based toxicology studies. (2025). OECD. https://doi.org/10.1787/6bb2e6ce-en

Ojeda, D. I., Mattila, T. M., Ruttink, T., Kujala, S. T., Kärkkäinen, K., Verta, J.-P., & Pyhäjärvi, T. (2019). Utilization of Tissue Ploidy Level Variation in de Novo Transcriptome Assembly of Pinus sylvestris. *G3: Genes|Genomes|Genetics*, *9*(10), 3409–3421. https://doi.org/10.1534/g3.119.400357

Ola Reibo. (2016, January 21). *TangloppeTorsdag: Gammarus locusta (Linnaeus, 1758) | Evertebratsamlingen*. W.uib.no. https://evertebrat.w.uib.no/2016/01/21/tangloppetorsdag-gammarus-locusta-linnaeus-1758/

Orsini, L., Gilbert, D., Podicheti, R., Jansen, M., Brown, J. B., Solari, O. S., Spanier, K. I., Colbourne, J. K., Rusch, D. B., Decaestecker, E., Asselman, J., De Schamphelaere, K. A. C., Ebert, D., Haag, C. R., Kvist, J., Laforsch, C., Petrusek,

A., Beckerman, A. P., Little, T. J., & Chaturvedi, A. (2016). Daphnia magna transcriptome by RNA-Seq across 12 environmental stressors. *Scientific Data*, *3*(1), 160030. https://doi.org/10.1038/sdata.2016.30

Ortúzar, M., Esterhuizen, M., Olicón-Hernández, D. R., González-López, J., & Aranda, E. (2022). Pharmaceutical Pollution in Aquatic Environments: A Concise Review of Environmental Impacts and Bioremediation Systems. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.869332

Ottmar, K. J., Colosi, L. M., & Smith, J. A. (2012). Fate and transport of atorvastatin and simvastatin drugs during conventional wastewater treatment. *Chemosphere*, *88*(10), 1184–1189. https://doi.org/10.1016/j.chemosphere.2012.03.066

Ozsolak, F., & Milos, P. M. (2010). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Panchal, N. K., & Prince Sabina, E. (2023). Non-steroidal anti-inflammatory drugs (NSAIDs): A current insight into its molecular mechanism eliciting organ toxicities. *Food and Chemical Toxicology*, *172*, 113598. https://doi.org/10.1016/j.fct.2022.113598

Papageorgiou, M., Kosma, C., & Lambropoulou, D. (2016). Seasonal occurrence, removal, mass loading and environmental risk assessment of 55 pharmaceuticals and personal care products in a municipal wastewater treatment plant in Central Greece. *Science of the Total Environment*, *543*, 547–569. https://doi.org/10.1016/j.scitotenv.2015.11.047

Patel, M., Kumar, R., Kishor, K., Mlsna, T., Pittman, C. U., & Mohan, D. (2019). Pharmaceuticals of Emerging Concern in Aquatic Systems: Chemistry, Occurrence, Effects, and Removal Methods. *Chemical Reviews*, *119*(6), 3510–3673. https://doi.org/10.1021/acs.chemrev.8b00299

Paut Kusturica, M., Jevtic, M., & Ristovski, J. T. (2022). Minimizing the environmental impact of unused pharmaceuticals: Review focused on prevention. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.1077974

Péden, R., Poupin, P., Sohm, B., Flayac, J., Giambérini, L., Klopp, C., Louis, F., Pain-Devin, S., Potet, M., Serre, R.-F., & Devin, S. (2019). Environmental transcriptomes of invasive dreissena, a model species in ecotoxicology and invasion biology. *Scientific Data*, *6*(1). https://doi.org/10.1038/s41597-019-0252-x

Pereira, A. M. P. T., Silva, L. J. G., Meisel, L. M., Lino, C. M., & Pena, A. (2015). Environmental impact of pharmaceuticals from Portuguese wastewaters: geographical and seasonal occurrence, removal and risk assessment. *Environmental Research*, *136*, 108–119. https://doi.org/10.1016/j.envres.2014.09.041

Perez, M. F., & Lehner, B. (2019). Intergenerational and transgenerational epigenetic inheritance in animals. *Nature Cell Biology*, *21*(2), 143–151. https://doi.org/10.1038/s41556-018-0242-9

Peschke, K., Geburzi, J., Ko¨hler, Heinz-R., Wurm, K., & Triebskorn, R. (2014). Invertebrates as indicators for chemical stress in sewage-influenced stream systems: Toxic and endocrine effects in gammarids and reactions at the community level in two tributaries of Lake Constance, Schussen and Argen. *Ecotoxicology and Environmental Safety*, *106*, 115–125. https://doi.org/10.1016/j.ecoenv.2014.04.011

Petrie, B., Barden, R., & Kasprzyk-Hordern, B. (2015). A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. *Water Research*, *72*, 3–27. https://doi.org/10.1016/j.watres.2014.08.053

Pisanti, S., Rimondi, E., Pozza, E., Melloni, E., Enrico Zauli, Bifulco, M., Martinelli, R., & Marcuzzi, A. (2022). Prenylation Defects and Oxidative Stress Trigger the Main Consequences of Neuroinflammation Linked to Mevalonate Pathway Deregulation. *International Journal of Environmental Research and Public Health*, *19*(15), 9061–9061. https://doi.org/10.3390/ijerph19159061

Pola-Sánchez, E., Karen Magdalena Hernández-Martínez, Pérez-Estrada, R., Sélem-Mójica, N., Simpson, J., María Jazmín Abraham-Juárez, Herrera-Estrella, A., & José Manuel Villalobos-Escobedo. (2024). RNA-Seq Data Analysis: A Practical Guide for Model and Non-Model Organisms. *Current Protocols*, *4*(5). https://doi.org/10.1002/cpz1.1054

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., & Finn, R. D. (2011). The Pfam protein families database. *Nucleic Acids Research*, *40*(D1), D290–D301. https://doi.org/10.1093/nar/gkr1065

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, *33*(suppl_2), W116–W120. https://doi.org/10.1093/nar/gki442

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rauthan, M., & Pilon, M. (2011). The mevalonate pathway in C. elegans. *Lipids in Health and Disease*, *10*(1), 243. https://doi.org/10.1186/1476-511x-10-243

Ré, A., Freitas, R., Sampaio, L., Rodrigues, A. M., & Quintino, V. (2009). Estuarine sediment acute toxicity testing with the European amphipod Corophium multisetosum Stock, 1952. *Chemosphere*, *76*(10), 1323–1333. https://doi.org/10.1016/j.chemosphere.2009.06.041

Rebelo, D., Correia, A. T., & Nunes, B. (2021). Acute and chronic effects of environmental realistic concentrations of simvastatin in danio rerio: evidences of oxidative alterations and endocrine disruptive activity. *Environmental Toxicology and Pharmacology*, *81*, 103522. https://doi.org/10.1016/j.etap.2020.103522

Richardson, S. D., & Ternes, T. A. (2021). Water Analysis: Emerging Contaminants and Current Issues. *Analytical Chemistry*, *94*(1), 382–416. https://doi.org/10.1021/acs.analchem.1c04640

Rinske Loeffen, Kristien Winckers, Ford, I., J. Wouter Jukema, Robertson, M., Stott, D. J., Spronk, H. M., Hugo ten Cate, & Lowe, G. D. (2014). Associations Between Thrombin Generation and the Risk of Cardiovascular Disease in Elderly Patients: Results From the PROSPER Study. *The Journals of Gerontology Series A*, *70*(8), 982–988. https://doi.org/10.1093/gerona/glu228

Rissman, E. F., & Adli, M. (2014). Minireview: Transgenerational Epigenetic Inheritance: Focus on Endocrine Disrupting Compounds. *Endocrinology*, *155*(8), 2770–2780. https://doi.org/10.1210/en.2014-1123

Ritter, C. J., & Bourne, D. G. (2024). Marine amphipods as integral members of global ocean ecosystems. *Journal of Experimental Marine Biology and Ecology*, *572*, 151985–151985. https://doi.org/10.1016/j.jembe.2023.151985

Rodrigues, J. A., Chaves, R. S., Santos, M. M., Neuparth, T., & Gil, A. M. (2024). Direct and transgenerational effects of simvastatin on the metabolism of the amphipod Gammarus locusta. *Aquatic Toxicology*, *279*, 107221–107221. https://doi.org/10.1016/j.aquatox.2024.107221

Russo, E., Lauritano, C., Giuliana d'Ippolito, Fontana, A., Sarno, D., Eric von Elert, Ianora, A., & Carotenuto, Y. (2020). RNA-Seq and differential gene expression analysis in Temora stylifera copepod females with contrasting non-feeding nauplii survival rates: an environmental transcriptomics study. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07112-w

Saha, S., Cooksey, A. M., Childers, A. K., Poelchau, M. F., & McCarthy, F. M. (2021). Workflows for Rapid Functional Annotation of Diverse Arthropod Genomes. *Insects*, *12*(8), 748. https://doi.org/10.3390/insects12080748

Salgado, R., Noronha, J. P., Oehmen, A., Carvalho, G., & Reis, M. A. M. (2010). Analysis of 65 pharmaceuticals and personal care products in 5 wastewater treatment plants in Portugal using a simplified analytical methodology. *Water Science and Technology*, *62*(12), 2862–2871. https://doi.org/10.2166/wst.2010.985

Santos, M. M., Ruivo, R., Lopes-Marques, M., Torres, T., de los Santos, C. B., Castro, L. F. C., & Neuparth, T. (2016). Statins: An undesirable class of aquatic contaminants? *Aquatic Toxicology*, *174*, 1–9. https://doi.org/10.1016/j.aquatox.2016.02.001

Sauer, U. G., Deferme, L., Gribaldo, L., Hackermüller, J., Tralau, T., van Ravenzwaay, B., Yauk, C., Poole, A., Tong, W., & Gant, T. W. (2017). The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook. *Regulatory Toxicology and Pharmacology*, *91*, S14–S26. https://doi.org/10.1016/j.yrtph.2017.09.020

Sha, Y., Phan, J. H., & Wang, M. D. (2015). Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2015*, 6461–6464. https://doi.org/10.1109/EMBC.2015.7319872

Shields, P. G. (2023). Role of untargeted omics biomarkers of exposure and effect for tobacco research. *Addiction Neuroscience*, *7*, 100098. https://doi.org/10.1016/j.addicn.2023.100098

Shyamal, S., Das, S., Guruacharya, A., Mykles, D. L., & Durica, D. S. (2018). Transcriptomic analysis of crustacean molting gland (Y-organ) regulation via the

mTOR signaling pathway. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-25368-x

Sirtori, C. R. (2014). The Pharmacology of Statins. *Pharmacological Research*, *88*, 3–11. https://doi.org/10.1016/j.phrs.2014.03.002

Skinner, M. K. (2008). What is an epigenetic transgenerational phenotype? *Reproductive Toxicology*, *25*(1), 2–6. https://doi.org/10.1016/j.reprotox.2007.09.001

Skinner, M. K. (2014). Environmental stress and epigenetic transgenerational inheritance. *BMC Medicine*, *12*(1). https://doi.org/10.1186/s12916-014-0153-y

Smith, K. F., Pochon, X., Melvin, S. D., Wheeler, T. T., & Tremblay, L. A. (2025). Integration of “Omics”-Based Approaches in Environmental Risk Assessment to Establish Cause and Effect Relationships: A Review. *Toxics*, *13*(9), 714–714. https://doi.org/10.3390/toxics13090714

Snape, J. R., Maund, S. J., Pickford, D. B., & Hutchinson, T. H. (2004). Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology. *Aquatic Toxicology*, *67*(2), 143–154. https://doi.org/10.1016/j.aquatox.2003.11.011

Song, Y., Zheng, K., Dag Anders Brede, Gomes, T., Xie, L., Kassaye, Y., Brit Salbu, & Knut Erik Tollefsen. (2023). Multiomics Point of Departure (moPOD) Modeling Supports an Adverse Outcome Pathway Network for Ionizing Radiation. *Environmental Science & Technology*, *57*(8), 3198–3205. https://doi.org/10.1021/acs.est.2c04917

Stossel, T. P. (2008). The Discovery of Statins. *Cell*, *134*(6), 903–905. https://doi.org/10.1016/j.cell.2008.09.008

Sui, Q., Cao, X., Lu, S., Zhao, W., Qiu, Z., & Yu, G. (2015). Occurrence, sources and fate of pharmaceuticals and personal care products in the groundwater: A review. *Emerging Contaminants*, *1*(1), 14–24. https://doi.org/10.1016/j.emcon.2015.07.001

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604. https://doi.org/10.1038/nprot.2017.149

Tan, H., Gao, P., Luo, Y., Gou, X., Xia, P., Wang, P., Yan, L., Zhang, S., Guo, J., Zhang, X., Yu, H., & Shi, W. (2023). Are New Phthalate Ester Substitutes Safer than Traditional DBP and DiBP? Comparative Endocrine-Disrupting Analyses on Zebrafish Using *In Vivo*, Transcriptome, and *In Silico* Approaches. *Environmental Science & Technology*, *57*(37), 13744–13756. https://doi.org/10.1021/acs.est.3c03282

Tan, K., Xu, P., Lim, L.-S., Nie, C., Tan, K., Peng, Y., Cai, X., Yan, X., Xu, Y., & Kwan, K. Y. (2023). Iso-seq and RNA-seq profiling of Fenneropenaeus penicillatus: Unravelling the signaling pathway and genes involved in the ovarian development. *Aquaculture Reports*, *32*, 101731. https://doi.org/10.1016/j.aqrep.2023.101731

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tete, V. S., Nyoni, H., Mamba, B. B., & Msagati, T. A. M. (2020). Occurrence and spatial distribution of statins, fibrates and their metabolites in aquatic environments. *Arabian Journal of Chemistry*, *13*(2), 4358–4373. https://doi.org/10.1016/j.arabjc.2019.08.003

The Gene Ontology Consortium. (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Thomas, R. S., Philbert, M. A., Auerbach, S. M., Wetmore, B. A., DeVito, M. J., Cote, I., J. Craig Rowlands, Whelan, M., Hays, S. M., Andersen, M. E., Meek, M. E.,

Reiter, L. T., Lambert, J. C., Clewell, H. J., Stephens, M. L., Q. Jay Zhao, Wesselkamper, S. C., Flowers, L., Carney, E. W., & Pastoor, T. P. (2013). Incorporating New Technologies Into Toxicity Testing and Risk Assessment: Moving From 21st Century Vision to a Data-Driven Framework. *Toxicological Sciences*, *136*(1), 4–18. https://doi.org/10.1093/toxsci/kft178

Tian, Y., Fu, Y., Yu, G., Zhang, B., & Wang, Y. (2014). A Statistical Approach to Identifying Significant Transgenerational Methylation Changes. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1409.7827

Tollefsen, K. E., Scholz, S., Cronin, M. T., Edwards, S. W., de Knecht, J., Crofton, K., Garcia-Reyero, N., Hartung, T., Worth, A., & Patlewicz, G. (2014). Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). *Regulatory Toxicology and Pharmacology*, *70*(3), 629–640. https://doi.org/10.1016/j.yrtph.2014.09.009

Vaklavas, C., Chatzizisis, Y. S., Ziakas, A., Zamboulis, C., & Giannoglou, G. D. (2009). Molecular basis of statin-associated myopathy. *Atherosclerosis*, *202*(1), 18–28. https://doi.org/10.1016/j.atherosclerosis.2008.05.021

Valdez-Carrillo, M., Abrell, L., Ramírez-Hernández, J., Jaime Alonso Reyes-López, & Concepción Carreón-Diazconti. (2020). *Pharmaceuticals as emerging contaminants in the aquatic environment of Latin America: a review*. *27*(36), 44863–44891. https://doi.org/10.1007/s11356-020-10842-9

Van Boeckel, T. P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B. T., Levin, S. A., & Laxminarayan, R. (2014). Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *The Lancet Infectious Diseases*, *14*(8), 742–750. https://doi.org/10.1016/s1473-3099(14)70780-7

van der Oost, R., Beyer, J., & Vermeulen, N. P. E. (2003). Fish bioaccumulation and biomarkers in environmental risk assessment: a review. *Environmental*

Toxicology and Pharmacology, *13*(2), 57–149. https://doi.org/10.1016/s1382-6689(02)00126-6

Vandegehuchte, M. B., Filip Lemière, Vanhaecke, L., Wim Vanden Berghe, & Janssen, C. R. (2010). Direct and transgenerational impact on Daphnia magna of chemicals with a known effect on DNA methylation. *Comparative Biochemistry and Physiology C-Toxicology & Pharmacology*, *151*(3), 278–285. https://doi.org/10.1016/j.cbpc.2009.11.007

Vandegehuchte, M. B., & Janssen, C. R. (2011). Epigenetics and its implications for ecotoxicology. *Ecotoxicology*, *20*(3), 607–624. https://doi.org/10.1007/s10646-011-0634-0

Vellinger, C., Felten, V., Sornom, P., Rousselle, P., Beisel, J.-N., & Usseglio-Polatera, P. (2012). Behavioural and Physiological Responses of Gammarus pulex Exposed to Cadmium and Arsenate at Three Temperatures: Individual and Combined Effects. *PLoS ONE*, *7*(6), e39153. https://doi.org/10.1371/journal.pone.0039153

Verlicchi, P., Al Aukidy, M., & Zambello, E. (2012). Occurrence of pharmaceutical compounds in urban wastewater: Removal, mass load and environmental risk after a secondary treatment—A review. *Science of the Total Environment*, *429*, 123–155. https://doi.org/10.1016/j.scitotenv.2012.04.028

Villeneuve, D. L., Blackwell, B. R., Bush, K., Harrill, J., Harris, F., Hazemi, M., Le, M., Stacy, E., & Flynn, K. M. (2024). Transcriptomics-Based Points of Departure for *Daphnia magna* Exposed to 18 Per- and Polyfluoroalkyl Substances. *Environmental Toxicology and Chemistry*. https://doi.org/10.1002/etc.5838

Villeneuve, D. L., & Garcia-Reyero, N. (2010). Vision & strategy: Predictive ecotoxicology in the 21st century. *Environmental Toxicology and Chemistry*, *30*(1), 1–8. https://doi.org/10.1002/etc.396

Vysakh, V. G., Sandhya Sukumaran, Sebastian, W., & Gopalakrishnan, A. (2025). The transcriptomic footprint of Mytella strigata: de novo transcriptome assembly of a major invasive species. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-04559-y

Walker, D. M., Zhou, X., Cunningham, A. M., Lipschultz, A. P., Ramakrishnan, A., Cates, H. M., Bagot, R. C., Shen, L., Zhang, B., & Nestler, E. J. (2021). Sex-Specific Transcriptional Changes in Response to Adolescent Social Stress in the Brain's Reward Circuitry. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2021.02.964

Walley, T., Folino-Gallo, P., Stephens, P., & Van Ganse, E. (2005). Trends in prescribing and utilization of statins and other lipid lowering drugs across Europe 1997-2003. *British Journal of Clinical Pharmacology*, *60*(5), 543–551. https://doi.org/10.1111/j.1365-2125.2005.02478.x

Wang, Q., Xu, Z., Wang, Y., Huo, G., Zhang, X., Li, J., Hua, C., Li, S., & Zhou, F. (2023). Transcriptomics Analysis of the Toxicological Impact of Enrofloxacin in an Aquatic Environment on the Chinese Mitten Crab (Eriocheir sinensis). *International Journal of Environmental Research and Public Health*, *20*(3), 1836–1836. https://doi.org/10.3390/ijerph20031836

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Web Application Framework for R [R package shiny version 1.6.0]. (2021, January 25). Cran.r-Project.org. https://cran.r-project.org/package=shiny

Wei, X., Zhang, Z., Gu, Y., Zhang, R., Huang, J., Li, F., He, Y., Lu, S., Wu, Y., Zeng, W., Liu, X., Liu, C., Liu, J., Ao, L., Shi, F., Chen, Q., Lin, Y., Du, J., Jin, G., & Xia, Y. (2024). Inter- and trans-generational impacts of real-world PM2.5 exposure on male-specific primary hypogonadism. *Cell Discovery*, *10*(1). https://doi.org/10.1038/s41421-024-00657-0

Whiteley, N. M., Samuel, Lunt, D. H., & Rock, J. (2011). Latitudinal variations in the physiology of marine gammarid amphipods. *Journal of Experimental Marine Biology and Ecology*, *400*(1-2), 70–77. https://doi.org/10.1016/j.jembe.2011.02.027

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. In *Use R!* Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Willey, K. (1999). An elusive role for glycosylation in the structure and function of reproductive hormones. *Human Reproduction Update*, *5*(4), 330–355. https://doi.org/10.1093/humupd/5.4.330

Yang, Y., & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, *14*(1), 328. https://doi.org/10.1186/1471-2164-14-328

Zapata, R., Martín, D., Maria-Dolors Piulachs, & Bellés, X. (2002). Effects of hypocholesterolaemic agents on the expression and activity of 3-hydroxy-3-methylglutaryl-CoA reductase in the fat body of the German cockroach. *Archives of Insect Biochemistry and Physiology*, *49*(4), 177–186. https://doi.org/10.1002/arch.10018

Zapata, R., Piulachs, M.-D. ., & Bellés, X. (2002). Ovarian 3-hydroxy-3-methylglutaryl-CoA reductase in Blattella germanica (L.): pattern of expression and critical role in embryogenesis. *Journal of Insect Physiology*, *48*(7), 675–681. https://doi.org/10.1016/s0022-1910(02)00086-0

Zhang, D., Hu, Q., Liu, X., Zou, K., Sarkodie, E. K., Liu, X., & Gao, F. (2020). AllEnricher: a comprehensive gene set function enrichment tool for both model and non-model species. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-3408-y

Zhang, L., Liu, Y., Chen, H., & Cai, W. (2022). Transcriptome analysis reveals sex-specific alterations in gonads of green mussel exposed to organophosphorus insecticide triazophos. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, *257*, 109333. https://doi.org/10.1016/j.cbpc.2022.109333

Zhang, X., Xia, P., Wang, P., Yang, J., & Baird, D. J. (2018). Omics Advances in Ecotoxicology. *Environmental Science & Technology*, *52*(7), 3842–3851. https://doi.org/10.1021/acs.est.7b06494

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

Zhou, J., Zhang, X., Wang, Y., Liang, H., Yang, Y., Huang, X., & Deng, J. (2024). Contamination Survey of Insect Genomic and Transcriptomic Data. *Animals*, *14*(23), 3432–3432. https://doi.org/10.3390/ani14233432

Zhou, Q., & Liao, J. (2009). Statins and Cardiovascular Diseases: From Cholesterol Lowering to Pleiotropy. *Current Pharmaceutical Design*, *15*(5), 467–478. https://doi.org/10.2174/138161209787315684

Ziylan, A., & Ince, N. H. (2011). The occurrence and fate of anti-inflammatory and analgesic pharmaceuticals in sewage and fresh water: Treatability by conventional and non-conventional processes. *Journal of Hazardous Materials*, *187*(1-3), 24–36. https://doi.org/10.1016/j.jhazmat.2011.01.057

Attachments

The bioinformatics pipelines developed in this work, including the code for the *TransDegAnalyser* application, are publicly available on GitHub: https://github.com/abiliaframboesa/Thesis_RNASeq