Carlotta Paone

Gene expression response of *Chondrilla nucula* to a simulated marine heat wave event

UNIVERSIDADE DO ALGARVE

Faculdade de Ciencias e Tecnologias

2024

Supervised by Dr Elisavet Kaitetzidou and Dr Rita Castilho

Gene expression response of *Chondrilla nucula* to a simulated marine heat wave event

Declaration of authorship of work.

I declare I am the author of this work, which is original and unpublished. The sources consulted have been duly cited in the text and included in the list of references.

Candidate signature: _____

Abstract

Sessile marine organisms are particularly susceptible to stress conditions due to their reduced mobility. The identification of early biomarkers in response to climate change scenarios as indicators for stress thresholds can provide fundamental understanding in prevention of necrosis or symbiotic disruptions between and within species. Porifera have shown incredible evolutionary adaptability, and are integrated into diverse marine ecosystems, playing important roles in their functioning. *Chondrilla nucula* Shmidt 1862, a marine sponge, represents an abundant population within the benthic community of the Mediterranean Sea. mRNA sequencing and de novo transcriptome assembly techniques were implemented on this non-model species in order to investigate the gene expression dynamics of *Chondrilla nucula* in response to elevated temperatures, simulating a recent marine heat wave event. Differential expression and functional annotation analysis has induced the expression of genes involved in metabolic activity, binding processes, and regulation of biological functions. This has provided insight into the mechanisms of stress response and adaptation, which may contribute to better management and conservation of sponge ecosystems under future climate conditions.

RNA Sequencing, Chondrilla nucula, Porifera, Climate change

Resumo

É detectável uma mudança considerável nos padrões climáticos que afetam a vida aquática. De acordo com um estudo de 2023 do Painel Intergovernamental sobre Mudanças Climáticas, as últimas três décadas representam o período mais quente dos últimos 1400 anos para o Hemisfério Norte. A temperaturas da superfície do mar aqueceu 0,11°C por década e antecipa-se uma continuação deste aumento. Entre as consequências das alterações climáticas, os períodos discretos de aquecimento oceânico regional extremo, nomeadamente as ondas de calor marinho (OCMs), duplicaram em frequência entre 1982 e 2016, prevendo-se que se tornem mais frequentes, intensas e duradouras na segunda metade do século XXI. As OCMs têm sido associadas a perturbações em ecossistemas inteiros, causando alterações na abundância e distribuição de espécies, incluindo extinções locais, além de contrações de alcance e biodiversidade reduzida. Um estudo sobre as condições térmicas extremas de 2015-2019 no Mar Mediterrâneo encontrou uma correlação significativa entre eventos de mortalidade em massa e exposição a OCMs, afetando 50 taxa diferentes, incluindo 7,2% dos Porifera encontrados até 45 metros de profundidade.

Os Porifera, vulgo esponjas, abundantes e omnipresentes em comunidades temperadas, tropicais e polares, desempenham um papel crucial no funcionamento dos ecossistemas bentónicos. Porífera, também conhecidos como esponjas, são um filo do reino Animalia. São seres multicelulares que apresentam poros em todo o corpo, o que deu origem ao nome ”Porífera”, que significa ”portador de poros” em latim. As esponjas são organismos simples que se fixam em superfícies sólidas em ambientes aquáticos, tanto marinhos quanto de água doce. Possuem um conjunto notável de características entre as quais (1) não têm órgãos verdadeiros ou tecidos diferenciados como músculos ou nervos, ao invés, são compostos por células especializadas que desempenham funções específicas; (2) alimentam-se filtrando partículas suspensas na água, como bactérias e plâncton, puxando a água através dos seus poros com o auxílio de células flageladas chamadas coanócitos; (3) podem reproduzir-se de maneira sexuada, com a produção de gametas, e assexuada, por brotamento ou fragmentação e (4) possuem um esqueleto interno composto por espículas, que podem ser de sílica ou carbonato de cálcio, e/ou uma substância fibrosa chamada espongina. As suas funções incluem a estabilização, consolidação e regeneração do substrato; o acoplamento bentopelágico, incluindo a

reciclagem de nutrientes; e interações com outros organismos, facilitando assim a produção primária e secundária, fornecendo habitat e proteção contra predadores. Dada a importância ecológica dos poríferos como elementos da comunidade bentónica, é essencial investigar a sua resposta a cenários ambientais futuros para o desenvolvimento de estratégias de gestão e conservação.

A esponja escolhida para este estudo, Chondrilla nucula Schmidt 1862, um membro associado de recifes da classe Demospongiae, possui uma população abundante na comunidade bentónica sésseis do Mar Mediterrâneo, embora a sua distribuição também tenha sido reportada nos oceanos Atlântico, Pacífico e Índico. Para entender como C. nucula modifica seus padrões de expressão génica após exposição a temperaturas elevadas da água do mar, efectuou-se a sequenciação de RNA para avaliar o transcriptoma. A ausência de um genoma de referência para Chondrilla nucula torna necessária a construção de transcriptoma de novo.

Os perfis de expressão foram investigados em cinco indivíduos adultos de Chondrilla nucula, coletados na costa norte de Creta. Neste estudo, segmentamos dez partes de cada esponja que servem como réplicas biológicas. Estes fragmentos foram mantidos em tanques com água do mar natural para se aclimatarem. Utilizamos dois tipos de tanques experimentais: um grupo de controle que permaneceu a uma temperatura constante de 22 °C, enquanto o outro foi exposto a um aumento gradual de temperatura, simulando uma onda de calor, alcançando os 28°C ao longo de cinco dias. Este procedimento foi desenhado para imitar as condições reais observadas no Mar Egeu durante um período intenso de calor em julho e agosto de 2021. Após a aclimatação e novamente 12 horas após as temperaturas terem atingido o seu pico nos tanques, recolhemos cinco fragmentos de cada tanque para análise.

Neste trabalho foi extraído o RNA total dos tecidos dos fragmentos de Chondrilla nucula e medimos espectrofotometricamente a concentração e pureza e a qualidade do RNA foi verificada através de eletroforese em gel de agarose. Criámos 15 bibliotecas de RNA mensageiro (mRNA) seguindo um protocolo específico e sequenciamos essas bibliotecas usando uma tecnologia TruSeq Stranded da Illumina. O processamento destes dados de sequenciamento, incluindo a montagem de um transcriptoma de novo e a quantificação de expressão de genes, foi feito com recursos computacionais avançados. Para entender as diferenças na atividade dos genes sob as duas condições, utilizamos o pacote DESeq2 do Bioconductor, bem como várias ferramentas adicionais de análise visual incluindo Volcano e Pheatmap da linguagem R. A anotação funcional e análise dos conjuntos de dados genómicos foi então alcançada por meio da plataforma de bioinformática Blast2GO, com softwares como Blast, InterProScan, mapeamento GO e Anotação.

Identificamos um total de 367.754 genes transcritos únicos, e a análise mostrou que 92,2% destes estava completa. Detectámos 432 genes transcritos com mudanças significativas na expressão devido ao tratamento térmico; 322 tiveram um aumento na expressão e 100 diminuíram. No entanto, 87

destes genes não puderam ser identificados nas bases de dados do National Center for Biotechnology Information. Analisámos os dados para melhor compreender quais os processos celulares e funções biológicas estavam afectados, utilizando o Blast2GO (termos de ontologia incluíram processos celulares, regulação biológica, ligação, atividade dependente de ATP e entidade anatómica celular). Isto também nos permitiu identificar quais vias genéticas activadas em resposta ao stress térmico, incluindo aquelas envolvidas em processos metabólicos e de ligação.

A principal limitação deste estudo foi a ausência de um genoma de referência para a esponja, o que significa que não pudemos identificar todos os genes. Haverá muitos genes por anotar, pelo que as funções desses genes ficaram por saber. No entanto, este estudo oferece uma visão inicial de como a esponja C. nucula responde a aumentos de temperatura de curto prazo. Esta informação é vital para prever os efeitos das mudanças climáticas em espécies marinhas importantes e para apoiar a gestão e conservação destas espécies, bem como dos seus relacionamentos simbióticos essenciais. Num ambiente marinho que se torna cada vez mais desafiador, apenas as espécies mais adaptáveis irão prevalecer.

Contents

List of Figures

List of Tables

Abbreviations, Acronyms and Symbols

RIN: RNA integrity number
MHW: marine heat wave
qPCR: quantitative polymerase chain reaction
cDNA: complementary DNA
TPM: transcripts per kilobase million
FPKM: fragments per kilobase per million fragments mapped
LFC: log2 fold change
rlog: regularized logarithm
PCA: principal component analysis
GO: gene ontology
DE: differential expression
NGS: next-generation sequencing
RSEM: RNA-Seq by expectation-maximization

Chapter 1

Introduction

The impacts of human society on the world have been magnified with population growth, developments in technology, and capitalistic globalisation. Aside from the over-fishing, pollution and introduction of alien species currently plaguing fresh and marine ecosystems, there is also considerable change in climatic patterns affecting life in the water. According to a 2023 study by the IPCC, the last three decades have been the warmest period in 1400 years for the Northern Hemisphere. Sea surface temperatures have shown a warming of 0.11 °C per decade and are expected to continue increasing [1].

Porifera, abundant and ubiquitous throughout temperate, tropical and polar communities, hold an important role within benthic ecosystem functioning [2, 3]. Their functions include substrate stabilisation, consolidation and regeneration; bentho-pelagic coupling, including nutrient cycling; and relationships with other organisms thereby facilitating primary and secondary production, providing habitat and protection from predation [2, 4, 5, 6, 7]. Sponges, as sessile organisms inhabiting the dynamic marine benthic environment, are constantly susceptible to environmental fluctuations due to their limited physical capacity to avoid unfavorable conditions [8], including extreme temperature oscillations as a consequence of climate change. Impacts of ocean warming for species part of the marine benthic community have described reduced calcification rates and mass bleaching events in corals [9], as well as the disruption of reproduction and larval development [10]. Certain species of porifera appear to have a higher tolerance to ocean warming in comparison to other benthic groups [11, 12]. A 2016 study on the shallow water sponge, *Haliclona tubifera* suggests the activation of homeostatic-related genes, including heat shock proteins and antioxidants as a result of thermal-stress [13]. Another study on *Rhopaleoides orodabile* proposes differing mechanisms of response to stress with regards to the life-history stage of the sponge sample [14]. Sponge larval capacity reveals its capacity for withstanding temperatures several degrees Celcius higher than their adult counterparts, with molecular mechanisms that may provide means of dispersal into cooler

waters [14]. In 2011, Pantile and Webster identified strict thermal thresholds in *Rhopaloeides orodabile*, as well as a rapid down-regulation of stress-inducible genes within 24 hours of treatment, and subsequent activation of the heat shock response system. Koutsouveli et al.,[15] similarly tested sponge gene expression at elevated temperatures and revealed major shifts with signal transduction, inflammation and the apoptopic pathway. Prolonged exposure also upregulated cell regeneration and growth related genes, suggesting a capacity for resilience [15]. Goodwin et al. (2013) indicated that certain sponge species appear more vulnerable with respect to others, highlighting the issue that sensitivity to the changing climate conditions is species-specific, and therefore is expected to affect community composition [16].

Amongst the consequences of climate change, discrete periods of extreme regional ocean warming, namely, marine heatwaves (MHWs) have doubled in their frequency between the years of 1982 and 2016 [17, 18] and are projected to become more frequent, intense and longer lasting in the second half of the 21st century [19, 20]. Model and observational studies suggest a correlation between the observed intensification of extreme climate events and anthropogenic activities [19, 21, 22]. MHWs have been shown to disrupt entire ecosystems [20] by causing alterations in abundance and distribution of species [23, 24] including local extinctions in additional to range contractions [25, 26, 27] and depleted biodiversity [28, 29, 30, 31]. MHWs may additionally impact carbon sequestration [32, 24], nitrogen cycling and nutrient turnover by the loss of productivity in benthic habitats [33, 34, 32]. One study analysing the 2015–2019 extreme thermal conditions in the Mediterranean sea found significant correlation between mass mortality events and MHW exposure, affecting 50 different taxa including 7.2 % of Porifera found up to 45 meters in depth [35]. Given the ecological importance of Porifera as members of the benthic community, investigating their response to future environmental scenarios is essential for the development of management and conservation strategies [36].

The sponge chosen for this study, *Chondrilla nucula* Schmidt 1862, reef-associated member of the class Demospongiae, has an abundant presence within the sessile benthic community of the Mediterranean Sea. Reportedly, its preferred temperature range resides between 24.5 and 28.2 °C [37]. Despite being photophilous [38], *C. nucula* is sometimes also found in marginal systems such as caves, with an apparent change in pigmentation due to a loss of symbiotic cyanobacteria [39]. *C. nucula* is gonochoric and oviparous [40] where gamete formation occurs between July and October, presumably regulated by sea surface temperature (SST) [41, 42]. Additionally, *C. nucula* also reproduces asexually [43] and is thus characterized by its modular growth system via clonal production [44, 45]. Its morphology is relatively simple, with a skeleton formed by a single type of spicule [3].

The proposed experimental design aims to assess the molecular background of the thermal

response on *Chondrilla nucula*, a sessile marine invertebrate, organisms most affected by environmental pressures due to their reduced motility. The integrative approach of realistically simulating the impact of contemporary scenarios may allow molecular mechanisms that respond to imposed stress factors to be examined. The identification of early bio-markers in response to a MHW may further provide indicators for stress thresholds, and their variations, in prevention of necrosis or symbiotic disruptions between and within species. To understand how *C. nucula* modifies its gene expression patterns, RNA sequencing will be used to assess the transcriptome. The expression profiles, after exposure to elevated seawater temperatures are prepared for a single population across five distinct individuals, to investigate adaptations for stress response and tolerance.

RNA sequencing will be used to quantify genomic expression in a sample without the need to pre-define sequences of interest [46]. In contrast to other molecular approaches, such as qPCR or microarrays, RNA sequencing allows for sensitive and accurate detection of expression levels for a broader range of genes [47]in both model and non-model organisms. The lack of a reference genome for the sponge *Chondrilla nucula* makes the building of a de novo assembly necessary. However, it is often difficult to differentiate signals in symbiont-rich metazoan organisms, potentially introducing biases to interpretation [48, 49]. A 2019 study to adjust for microsymbionts in sponge transcriptome assembly did so by manually removing most prokaryotic sequences within the dataset [50]. They do point out, however, that this method does not filter non-sponge eukaryotic sequences that may be present, including fungi and dinoflagellate symbionts [50]. There are a number of considerations to adapt the RNA sequencing protocol to the targeted transcriptome. For the purpose of this study, stranded RNA sequencing is selected over non-stranded as it allows to more accurately quantify the expression levels for genes with overlapping genomic loci [51]. Paired-end reading of sequenced fragments is selected for the transcript quantitation to improve mapping specificity for the selected non-model organism.

Regardless of the gene expression method (RNASeq, qPCR, microarrays) to be followed, standardizing the RNA isolation process is crucial for subsequent data processing [52]. High quality RNA extraction of tissue, from porifera specimen, is a challenging step for subsequent cDNA library construction and gene expression analysis [53, 54]. Contamination by symbiotic communities or chemical compounds may inhibit or influence downstream applications. Prevalent RNA isolation methods include silica-gel based membranes or liquid to liquid extractions with acidic phenol-chloroform. In the former, RNA binds to the silica-gel membrane using ethanol, whose volume influences which transcripts are retained, depending on their size. The latter dissolves the cellular components into separate phases, typically followed by an alcohol precipitation to de-salt and concentrate the RNA, where efficiency may vary resulting in different RNA populations [46]. The quality of RNA extraction will be evaluated by measuring absorbance at 260, and 280 nm,

with an expected absorbance ratio (260/280 nm) of 2.0 +/- 0.15 for 'pure' RNA [53]. As RNA is highly susceptible to degradation, RNA integrity will be assessed using gel electrophoresis for the ribosomal RNA bands. Intact eukaryotic total RNA should yield clear 28s and 18s rRNA bands, with a 2:1 ratio. The interpretation of gel images will be done using a software algorithm to calculate an RNA Integrity Number (RIN) to facilitate comparability between samples [46].

To determine the molecular response of *C. nucula* to thermal stress, RNA sequencing will be performed to profile the expression levels for five individuals compared with the Control group. In particular, the focus will be on the differential expression genes involved in stress response and adaptation.

References

[1] Hoesung Lee et al. “IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.” In: (2023).

[2] James J Bell. “The functional roles of marine sponges”. In: *Estuarine, coastal and shelf science* 79.3 (2008), pp. 341–353.

[3] Michelle Klautau et al. “Does cosmopolitanism result from overconservative systematics? A case study using the marine sponge Chondrilla nucula”. In: *Evolution* 53.5 (1999), pp. 1414–1422.

[4] Klaus Rützler. “The role of burrowing sponges in bioerosion”. In: *Oecologia* 19.3 (1975), pp. 203–216.

[5] Klaus Rützler and Ian G Macintyre. “Siliceous sponge spicules in coral reef sediments”. In: *Marine Biology* 49.2 (1978), pp. 147–159.

[6] Klaus Ruetzler. “Sponges on coral reefs: a community shaped by competitive cooperation”. In: (2004).

[7] Janie L Wulff. “Ecological interactions of marine sponges”. In: *Canadian Journal of Zoology* 84.2 (2006), pp. 146–166.

[8] A Padua and M Klautau. “Regeneration in calcareous sponges (Porifera)”. In: *Journal of the Marine Biological Association of the United Kingdom* 96.2 (2016), pp. 553–558.

[9] Rebecca Albright et al. “Carbon dioxide addition to coral reef waters suppresses net community calcification”. In: *Nature* 555.7697 (2018), pp. 516–519.

[10] Rebecca Albright and Chris Langdon. “Ocean acidification impacts multiple early life history processes of the Caribbean coral Porites astreoides”. In: *Global change biology* 17.7 (2011), pp. 2478–2487.

[11] James J Bell et al. “Could some coral reefs become sponge reefs as our climate changes?” In: *Global change biology* 19.9 (2013), pp. 2613–2624.

[12] James J Bell et al. “Climate change alterations to ecosystem dominance: how might sponge-dominated reefs function?” In: *Ecology* 99.9 (2018), pp. 1920–1931.

[13] Christine Guzman and Cecilia Conaco. “Gene expression dynamics accompanying the sponge thermal stress response”. In: *PloS one* 11.10 (2016), e0165368.

[14] N Webster et al. “A complex life cycle in a warming planet: gene expression in thermally stressed sponges”. In: *Molecular Ecology* 22.7 (2013), pp. 1854–1868.

[15] Vasiliki Koutsouveli et al. “Gearing up for warmer times: transcriptomic response of Spongia officinalis to elevated temperatures reveals recruited mechanisms and potential for resilience”. In: *Frontiers in Marine Science* 6 (2020), p. 786.

[16] Eva Chatzinikolaou, Kleoniki Keklikoglou, and Panos Grigoriou. “Morphological properties of gastropod shells in a warmer and more acidic future ocean using 3D micro-computed tomography”. In: *Frontiers in Marine Science* 8 (2021), p. 645660.

[17] Thomas L Frölicher, Erich M Fischer, and Nicolas Gruber. “Marine heatwaves under global warming”. In: *Nature* 560.7718 (2018), pp. 360–364.

[18] Thomas Wernberg et al. “Climate change increases marine heatwaves harming marine ecosystems”. In: *ScienceBrief Crit. Issues Climate Change Sci* (2021).

[19] Gerald A Meehl and Claudia Tebaldi. “More intense, more frequent, and longer lasting heat waves in the 21st century”. In: *Science* 305.5686 (2004), pp. 994–997.

[20] Dan A Smale et al. “Marine heatwaves threaten global biodiversity and the provision of ecosystem services”. In: *Nature Climate Change* 9.4 (2019), pp. 306–312.

[21] Kevin E Trenberth, John T Fasullo, and Theodore G Shepherd. “Attribution of climate extreme events”. In: *Nature Climate Change* 5.8 (2015), pp. 725–730.

[22] Eric CJ Oliver et al. “The unprecedented 2015/16 Tasman Sea marine heatwave”. In: *Nature communications* 8.1 (2017), p. 16101.

[23] Katherine E Mills et al. “Fisheries management in a changing climate: lessons from the 2012 ocean heat wave in the Northwest Atlantic”. In: *Oceanography* 26.2 (2013), pp. 191–195.

[24] Miguel Ñiquen and Marilú Bouchon. “Impact of El Niño events on pelagic fisheries in Peruvian waters”. In: *Deep sea research part II: topical studies in oceanography* 51.6-9 (2004), pp. 563–574.

[25] Joaquim Garrabou et al. “Mass mortality in Northwestern Mediterranean rocky benthic communities: effects of the 2003 heat wave”. In: *Global change biology* 15.5 (2009), pp. 1090–1103.

[26] Dan A Smale and Thomas Wernberg. “Extreme climatic event drives range contraction of a habitat-forming species”. In: *Proceedings of the Royal Society B: Biological Sciences* 280.1754 (2013), p. 20122829.

[27] C Cerrano et al. “A catastrophic mass-mortality episode of gorgonians and other organisms in the Ligurian Sea (North-western Mediterranean), summer 1999”. In: *Ecology letters* 3.4 (2000), pp. 284–293.

[28] Thomas Wernberg et al. “Climate-driven regime shift of a temperate marine ecosystem”. In: *Science* 353.6295 (2016), pp. 169–172.

[29] Jordan A Thomson et al. “Extreme temperatures, foundation species, and abrupt ecosystem change: an example from an iconic seagrass ecosystem”. In: *Global Change Biology* 21.4 (2015), pp. 1463–1474.

[30] BE Brown. “Damage and recovery of coral reefs affected by El Niño related seawater warming in the Thousand Islands, Indonesia”. In: *Coral reefs* 8 (1990), pp. 163–170.

[31] Matthew S Edwards. “Estimating scale-dependency in disturbance impacts: El Niños and giant kelp forests in the northeast Pacific”. In: *Oecologia* 138 (2004), pp. 436–447.

[32] Núria Marbà and Carlos M Duarte. “Mediterranean warming triggers seagrass (Posidonia oceanica) shoot mortality”. In: *Global change biology* 16.8 (2010), pp. 2366–2375.

[33] Thomas Wernberg et al. “An extreme climatic event alters marine ecosystem structure in a global biodiversity hotspot”. In: *Nature Climate Change* 3.1 (2013), pp. 78–82.

[34] FP Chavez et al. “Biological and chemical consequences of the 1997–1998 El Niño in central California waters”. In: *Progress in Oceanography* 54.1-4 (2002), pp. 205–232.

[35] Joaquim Garrabou et al. “Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea”. In: *Global Change Biology* 28.19 (2022), pp. 5708–5725.

[36] Daniel Gómez-Gras et al. “Response diversity in Mediterranean coralligenous assemblages facing climate change: Insights from a multispecific thermotolerance experiment”. In: *Ecology and Evolution* 9.7 (2019), pp. 4168–4180.

[37] K Kaschner et al. “AquaMaps: Predicted range maps for aquatic species”. In: *World wide web electronic publication, www. aquamaps. org, Version* 8 (2016), p. 2016.

[38] Martina Milanese et al. “The marine sponge Chondrilla nucula Schmidt, 1862 as an elective candidate for bioremediation in integrated aquaculture”. In: *Biomolecular engineering* 20.4-6 (2003), pp. 363–368.

[39] Elda Gaino, Maurizio Pansini, and R Pronzato. “Aspetti dell’associazione fra Chondrilla nucula Schmidt (Demospongiae) e microorganismi simbionti (batteri e Cianoficee) in condizioni naturali e sperimentali”. In: *Cahiers de Biologie Marine* 18 (1977), pp. 303–310.

[40] L Scalera Liaci, M Sciscioli, and A Matarrese. “Sexual reproduction in some sponges: Chondrilla nucula OS and Chondrosia reniformis Nardo (Tetractinomorpha)”. In: *Rap Commis Internat Explor Sci Mer Méditerranée* 22 (1973), pp. 129–130.

[41] Elda Gaino et al. “Indagine ultrastrutturale sugli ovociti maturi di Chondrilla nucula Schmidt (Porifera, Demospongiae)”. In: *Cahiers de Biologie Marine* 21 (1980), pp. 11–22.

[42] L Scalera Liaci and MARGHERITA Sciscioli. “Sexual cycles of some marine Porifera”. In: *Pubblicazioni della Stazione Zoologica, Napoli (Italy)* (1975).

[43] Elda Gaino, Renata Manconi, and Roberto Pronzato. “Organizational plasticity as a successful conservative tactics in sponges”. In: *Animal Biology* 4 (1995), pp. 31–43.

[44] John NA Hooper and Rob WM Van Soest. “Systema Porifera. A guide to the classification of sponges”. In: *Systema Porifera*. Springer, 2002, pp. 1–7.

[45] F Brümmer et al. “Maintenance and growth of sponges in aquariums: fundamentals for in vitro cultivation approaches far from the sea”. In: *Bull Mus. Biol. Inst. Univ. Genova* 37 (2002), pp. 66–67.

[46] Friederike Dündar, Luce Skrabanek, and Paul Zumbo. “Introduction to differential gene expression analysis using RNA-seq”. In: *Applied Bioinformatics Core/Weill Cornell Medical College* (2015), pp. 1–67.

[47] V Cahais et al. “Reference-free transcriptome assembly in non-model animals from next-generation sequencing data”. In: *Molecular ecology resources* 12.5 (2012), pp. 834–845.

[48] Emma Cebrian Pujol et al. “Sponge Mass Mortalities in a Warming Mediterranean Sea: Are Cyanobacteria-Harboring Species Worse Off?” In: *PLoS ONE, 2011, vol. 6, núm. 6, p. e20211* (2011).

[49] Nicole S Webster, Rose E Cobb, and Andrew P Negri. “Temperature thresholds for bacterial symbiosis with a sponge”. In: *The ISME journal* 2.8 (2008), pp. 830–842.

[50] Tereza Manousaki et al. “A de novo transcriptome assembly for the bath sponge Spongia officinalis, adjusting for microsymbionts”. In: *BMC Research Notes* 12.1 (2019), pp. 1–3.

[51] Shanrong Zhao et al. “Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap”. In: *BMC genomics* 16.1 (2015), pp. 1–14.

[52] Marc Sultan et al. “Influence of RNA extraction methods and library selection schemes on RNA-seq data”. In: *BMC genomics* 15.1 (2014), pp. 1–13.

[53] RE Farrell Jr. “Electrophoresis of RNA”. In: *RNA Methodologies–a Laboratory Guide for Isolation and Characterization, edn* 2 (1998), pp. 174–177.

[54] ELS W MAAS. “COMPARISON OF RNA EXTRACTION METHODS FROM SPONGES”. In: *BMIB-Bollettino dei Musei e degli Istituti Biologici* 68 (2004).

Chapter 2

Gene expression response of *Chondrilla nucula* to a simulated marine heat wave event

Carlotta Paone

Supervised by Dr Elisavet Kaitetzidou and Dr Rita Castilho

Keywords: *RNA Sequencing, Chondrilla nucula, Porifera, Climate change*

Abstract

Sessile marine organisms are particularly susceptible to stress conditions due to their reduced mobility. The identification of early biomarkers in response to climate change scenarios as indicators for stress thresholds can provide fundamental understanding in prevention of necrosis or symbiotic disruptions between and within species. Porifera have shown incredible evolutionary adaptability, and are integrated into diverse marine ecosystems, playing important roles in their functioning. *Chondrilla nucula* Shmidt 1862, a marine sponge, represents an abundant population within the benthic community of the Mediterranean Sea. mRNA sequencing and de novo transcriptome assembly techniques were implemented on this non-model species in order to investigate the gene expression dynamics of *Chondrilla nucula* in response to elevated temperatures, simulating a recent marine heat wave event. Differential expression and functional annotation analysis has induced the expression of genes involved in metabolic activity, binding processes, and regulation of biological functions. This has provided insight into the mechanisms of stress response and adaptation, which may contribute to better management and conservation of sponge ecosystems under future climate conditions.

Introduction

The impacts of human society on the world have been magnified with population growth, developments in technology, and capitalistic globalisation. Aside from the over-fishing, pollution and introduction of alien species currently plaguing fresh and marine ecosystems, there is also considerable change in climatic patterns affecting life in the water. According to a 2023 study by the IPCC, the last three decades have been the warmest period in 1400 years for the Northern Hemisphere. Sea surface temperatures have shown a warming of 0.11 °C per decade and are expected to continue increasing [1].

Porifera, abundant and ubiquitous throughout temperate, tropical and polar communities, hold an important role within benthic ecosystem functioning [2, 3]. Their functions include substrate stabilisation, consolidation and regeneration; benthopelagic coupling, including nutrient cycling; and relationships with other organisms thereby facilitating primary and secondary production, providing habitat and protection from predation [2, 4, 5, 6, 7]. Sponges, as sessile organisms inhabiting the dynamic marine benthic environment, are constantly susceptible to environmental fluctuations due to their limited physical capacity to avoid unfavorable conditions [8], including extreme temperature oscillations as a consequence of climate change.

Amongst the consequences of climate change, discrete periods of extreme regional ocean warming, namely, marine heatwaves (MHWs) have doubled in their frequency between the years of 1982 and 2016 [9, 10] and are projected to become more frequent, intense and longer lasting in the second half of the 21st century [11, 12]. MHWs have been shown to disrupt entire ecosystems [12] by causing alter-

ations in abundance and distribution of species [13, 14] including local extinctions in additional to range contractions [15, 16, 17] and depleted biodiversity [18, 19, 20, 21]. One study analysing the 2015–2019 extreme thermal conditions in the Mediterranean sea found significant correlation between mass mortality events and MHW exposure, affecting 50 different taxa including 7.2 % of Porifera found up to 45 meters in depth [22]. Given the ecological importance of Porifera as members of the benthic community, investigating their response to future environmental scenarios is essential for the development of management and conservation strategies [23].

The sponge chosen for this study, *Chondrilla nucula* Schmidt 1862, reef-associated member of the class Demospongiae, has an abundant presence within the sessile benthic community of the Mediterranean Sea. The proposed experimental design aims to assess the molecular background of the thermal response on *Chondrilla nucula*, a sessile marine invertebrate, organisms most affected by environmental pressures due to their reduced motility. The integrative approach of realistically simulating the impact of contemporary scenarios may allow molecular mechanisms that respond to imposed stress factors to be examined. The identification of early bio-markers in response to a MHW may further provide indicators for stress thresholds, and their variations, in prevention of necrosis or symbiotic disruptions between and within species. To understand how *C. nucula* modifies its gene expression patterns, RNA sequencing will be used to assess the transcriptome. The expression profiles, after exposure to elevated seawater temperatures are prepared for a single population across five distinct individuals, to investigate adaptations for stress response and tolerance.

Materials and methods

2.1 Sample Preparation.

Five adult individuals of *Chondrilla nucula* Schmidt, 1862 were collected from the Northern coast of Crete, at the Bay of Malia in January 2023 via SCUBA diving (35°19'49.3"N 25°23'14.5"E). The temperature of the seawater upon collection was 18 °C. 10 fragments for each individual, representing biological replicates were reared in closed tanks, containing natural seawater, until acclimated. Acclimation commenced at 18 °C and was gradually increased over 28 days to 22 °C. Two experimental tanks were employed, one functioning as a control (C) and the other was exposed to a heat wave event over a 5 day time-period. Staring from 22 °C, the temperature in the experimental tank reached a maximum of 28 °C over 5 days. The control tank remained at a constant temperature of 22 °C throughout the experiment. The selected temperature elevation simulated true conditions as reported by Aegean sea surface temperatures during a prolonged and extreme heatwave event in June 2021 [24]. Five fragments from each tank for each individual specimen were sampled at

Figure 2.1: *Chondrilla nucula*
Image courtesy of Dr. Thanos Dailianis

.

each step; post-acclimation (T0) and 12 hours after reaching maximum temperature (22 and 28 °C for the control and experimental tanks, respectively) (T1) (fig. 2.2). Samples were selected to mitigate epibiont contamination and quickly flash frozen using dry ice.

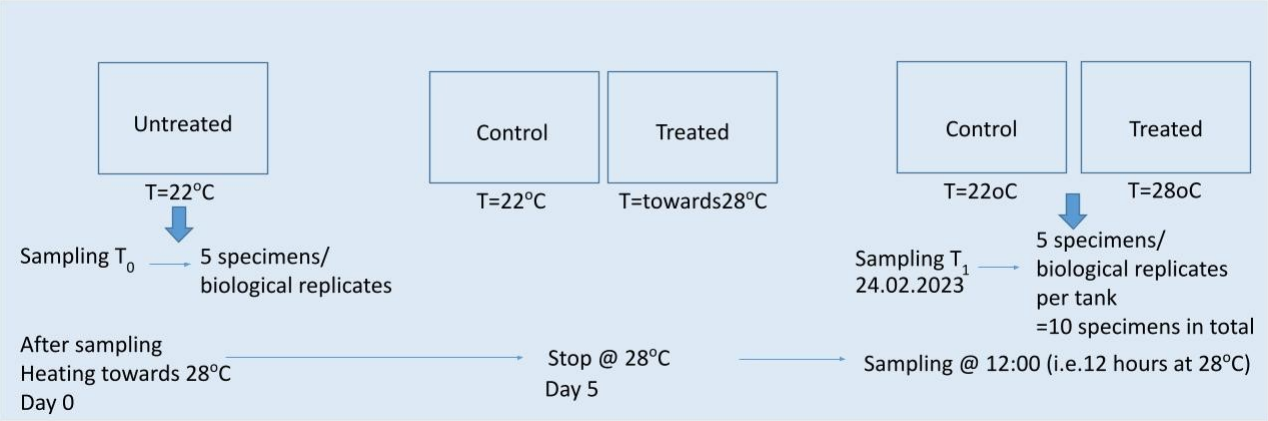


Figure 2.2: Description of the simulated MHW, showing temperature regulation during the experimental procedure for the control and treatment tanks. Samples of the specimen were taken in two phases, as labeled, before and after thermal stress treatment.

.

2.2 Wet bench research

2.2.1 RNA isolation.

Extraction of the total RNA was performed with the optimized protocol of Jordon-Thaden et al. (2015) using the cetyl trimethyl ammonium bromide (CTAB) kit and NucleoZOL (Macherey-Nagel, Düren, Germany). CTAB was originally used for the extraction of RNA from plant tissue, thus this combination enables for polysaccharides to be removed and nucleotides to be exclusively isolated. The wet weight of sponge tissue used for each extraction was between 0.1 - 0.2 g. Extractions were performed in duplicate and all methods were performed on the same day to reduce variability between samples and possible effects of storage. All

procedures were performed quickly using ice in order to avoid RNA degradation, prior to and post homogenization with liquid nitrogen.

In order to isolate total RNA from the tissue, the sample was first pulverized and lyophilized using a mortar and pestle in the presence of liquid nitrogen. The prepared CTAB extraction buffer was added to the lysate, vortexed and incubated at 55 °C to allow dissociation of the nucleoprotein complexes. Following CTAB standard protocol, samples were centrifuged, to separate the upper aqueous phase from the lower organic and inter-phase. The aqueous phase was transferred into a clean tube and choroform:iso-amyl-alcohol (IAA) mixture was added and centrifuged. Next, the aqueous phase was mixed with nucleoZOL and sarcosyl, centrifuged and the supernatant was transferred into clean tubes. The next steps followed the nucleoZOL standard procedure. Total RNA was precipitated with isopropanol and the pellet washed with ethanol (75 % EtOH). The RNA pellet was then re-dissolved in nuclease-free water, to approximately 1 microg/microL and stored at -20 °C overnight for optimal RNA self-hybridization.

2.2.2 Quantitative and qualitative assessment of the extracted RNA.

Sample concentration and purity were determined spectrophotometrically using the Nanodrop 1000 (Thermo Fisher Scientific) via OD measurements of 260/280 nm and 260/230 nm. Using the Beer-Lambert law, the 260/280 nm ratio should measure 2.0 for pure RNA, where a lower value may indicate the presence of proteins, phenols or other contaminants. The 260/230 nm ratio can be used as a secondary measure of nucleic acid purity, with purity values ranging between 2.0-2.2 [25]. The integrity of the extracted RNA was qualitatively evaluated via agarose gel electrophoresis (1.5 % agarose) (Sigma-Aldrich, Germany) (see A.1).

2.2.3 mRNA library preparation.

Efficient mRNA sequencing requires the exclusion of ribosomal RNA. As according to the Illumina (R) TruSeq (R) Stranded mRNA library preparation protocol, mRNA was isolated via their poly-A tail and the use of poly-T attached magnetic beads to complete this purification step. Subsequently, fragmentation of the mRNA to appropriate sizes for sequencing was performed using divalent cations under elevated temperatures.

First strand complementary DNA (cDNA) was synthesized from the RNA fragments using reverse transcriptase and random primers. Strand specificity was improved by adding Actinomycin D to the First Strand Synthesis (FSA Act D mix) to prevent spurious DNA-dependent synthesis.

Strand specificity was achieved in the Second Strand Marking Mix (SMM) by replacing dTTP with dUTP followed by second strand cDNA synthesis with DNA Polymerase I and RNase H. dUTP quenches the second strand synthesis during amplification.

The double-stranded cDNA fragments were further processed for hybridization onto a flow cell, following a standard preparation scheme which includes: end-polishing, A-tailing, adaptor ligation, and size selection using magnetic beads. The products were amplified with PCR and purified to generate the final cDNA library. Finally, 15 libraries, with one library per sample, were constructed.

2.2.4 RNA sequencing.

A pool of equal quantity of the prepared libraries was paired-end sequenced on an Illumina platform. Each library was estimated to generate 20 million reads.

2.3 NGS data processing.

All scripts used for the data analysis can be found in Appendix B.

2.3.1 de novo Transcriptome assembly

NGS data processing was accomplished using the high-performance computational (HPC) resources provided by IMBBC of the HCMR [26]. Bioinformatics analysis was performed using FastQC [27] for sequence quality check; Trimmomatic [28] for adaptor and quality trimming; TrinityRNASeq [29] for the de novo transcriptome assembly; RSEM [30] for transcript quantification; and Busco [31] for transcriptome completeness assessment.

As the organism of interest lacks a previously sequenced genome or transcriptome, the production of a de novo transcriptome was selected to assess gene expression. Quantification of the transcripts was achieved by first assembling sequence reads into contigs and subsequently mapping these contigs onto the generated transcriptome. This was achieved by de novo transcriptome assembly using Trinity, a process that builds transcript sequences from RNA-seq data without the need for a reference genome. RNA-seq reads are overlapped with each other to recreate the original transcripts. Trinity works in three stages: Inchworm assembles the reads into unique sequence transcripts; Chrysalis then clusters these transcripts and constructs de Bruijn graphs. The de Bruijn graphs are mathematical structures representing overlaps between sequences. Each node in the graph corresponds to a sequence of nucleotides, and the edges connect nodes with these overlapping sequences. Butterfly ultimately processes these graphs to tease out the full-length transcripts for each gene by finding a path through the graph that visits each edge exactly once [32].

2.3.2 Differential expression

As the approach uses two distinct *Chondrilla nucula* experimental groups, the RNA-Seq analysis will focus on identifying differential expression between transcripts. Differential gene expression analysis will be performed using original expression counts with the R [33] Bionconductor package DESeq2 [34]. Several additional

visual analysis tools were used with R, including Enhanced Volcano and Pheatmap [35, 36, 37, 38, 39].

2.3.3 Functional Annotation

The Blast2GO bio-informatics platform was used to perform functional annotation and analysis of genomic datasets. The gene ontology (GO) workflow uses Blast, InterProScan and GO Mapping and Annotation to obtain the most complete annotation labels for the sequences loaded [40]. For identifying protein products encoded by a nucleotide query, Blastx was run using the default target protein 'non-redundant' (nr) sequences database. Blast expectation value (E-value) had a threshold of 1.0E-3. Default values for protein sequences were kept: word size 3, HSP length cut-off 33 and HSP-hit coverage 0. Number of blast hits selected was 1. In parallel InterProScan performed protein domain analysis by comparing the searched sequences against the public EMBL-EBI InterPro web-service database of protein families, domains and functional sites. GO Mapping is a process of retrieving GO terms obtained by the blast search. Annotation is the process of selecting specific GO terms from the GO pool obtained by mapping and assigning them to query sequences. Default values in the annotation configuration were kept; cut-off 55, GO Weight 5, E-value-Hit-Filter $1.0E-6$, HSP-Hit Coverage Cutoff 0, Hit Filter 500.

Enrichment analysis is the statistical analysis to test the frequency in which GO terms are abundant or scare in a test sequence set as compared to a reference set. Enrichment analysis (Fisher's exact test), p-value filter was set as 0.05. As only sequences having finished annotation can be used in the Enrichment analysis, significantly up-regulated and down-regulated genes were compared to each other to perform this analysis. The down-regulated genes were selected as the 'test set' and run against all significant DEGs in the 'reference set'. The same was done for the up-regulated genes. Consequently, a one-sided test was selected over a two-

sided test; this tests only over-represented, abundant functions for the comparison. Fisher's Exact test uses a contigency table-based method to examine the associated between two kinds of classification.

Data analysis

2.4 RNA extraction and quality assessment

RNA extractions from porifera are a relatively novel concept, and thus optimisation of the protocol required establishing. RNA purity and integrity were assessed using UV spectroscopy and gel electrophoresis, respectively, as described above (see 2.2.1).

Absorbance ratios of the C. nucula replicates were all in range for pure RNA (table. A.1). Gel electrophoresis demonstrated relatively little smearing and a reasonable ratio between 28s and 18s bands (fig. A.1), where wells 4,5 and 6 are separate samples of *C. nucula*, wells 2 and 7 are ladders.

Preparation and sequencing was externally performed at Novogene's UK headquarters by technical staff for 15 mRNA libraries. To evaluate the quantity, integrity and purity requirements of samples, prior to library construction, Novogene performed supplementary quality control assessments using Nanodrop, Agilent and Agarose Gel Electrophoresis techniques. RIN was obtained via the Agilent 2100 Bioanalyzer system [41] reporting moderately degraded samples (table A.2).

2.5 De Novo Transcriptome Assembly

2.5.1 Trimming

The Trimmomatic command line tool was used to trim and crop the raw Illumina sequenced data, as well as to remove the adaptor sequences. Paired end mode was applied, along with the standard parameters supplied on the command line, which removes low quality leading and trailing bases (default 3); cutting average sequences when average quality per base drops below 15 (SLIDINGWINDOW: 4:15) and dropping reads below 36 bases long (see appendix B). All files were additionally run through Fastqc both prior to and post Trimming, for quality control. The percentages of sequences surviving the Trimming step are illustrated in Table A.3 below. All subsequent data elaboration and analysis were performed using trimmed and paired sequences.

Post-trimming FastQC resulted with no sequences flagged as poor quality; as exemplified by fig. A.2. All sequences were reported between 36 and 150 basepairs in length; the maximum read length was determined by the selected number of cycles for sequencing.

2.5.2 Trinity

De novo transcriptome assembly was achieved using the Trinity [29] suite v2.14.0, see appendix B. Time taken: 111h:39m:48s. Trimmed and paired-end transcript sequences from 15 biological replicates were processed using a minimum contig length of 200 bp and normalization by read set. The output was given as a fasta file used for all downstream analysis.

2.5.3 Transcriptome Assembly Quality Assessment: Basic Contig Statistics

The N50 statistic (fig. A.3) corresponds to the sequence length such that all contigs of at least that length compose at least 50 % of the assembly [42]. In transcriptome

assemblies, contigs represent transcripts, therefore the N50 metric can be misleading. To avoid this, the ExN50 [43] metric takes into account only the most highly expressed transcripts representing a percentage of the total normalized expression data (fig. A.4).

2.5.4 Quantitative Assessment: Busco

The BUSCO (Benchmarking Universal Single-Copy Orthologs) tool was used to quantitatively assess the completeness of the universal 'gene' matches of the Trinity de novo assembled transcriptome, both prior-to and post-filtering for the longest isoform per Trinity gene (figs. A.5 and 2.3). Busco works by comparing the 'genes' predicted in an assembly against a set of highly conserved, single-copy orthologous 'genes' that are expected to be present in nearly all members of a particular taxonomic group. Busco v5.4.6 was run on the Galaxy.org platform using the lineage dataset: $metazoa_odb10$ (Creation date: 2021-02-17, number of genomes: 65, number of BUSCOs: 954). A good quality assembly is considered as having completeness scores of $>$80 % BUSCO, indicating BUSCO gene matches in the transcriptome [43] with few gene matches missing or fragmented [44].

2.6 RSEM

De novo transcriptome assemblies typically result in more transcription sequences than expected based on the number of genes in the genome [43]. Thus, to eliminate technical biases such as chimera presence, sequencing depth, and gene length, the assembly was filtered using a minimum expression TPM (transcripts per kilobase million) threshold of 1. TPM was used instead of FPKM (fragments per kilobase per million fragments mapped) as a more accurate statistic for normalization of gene expression comparison across samples [45] (see appendix B).

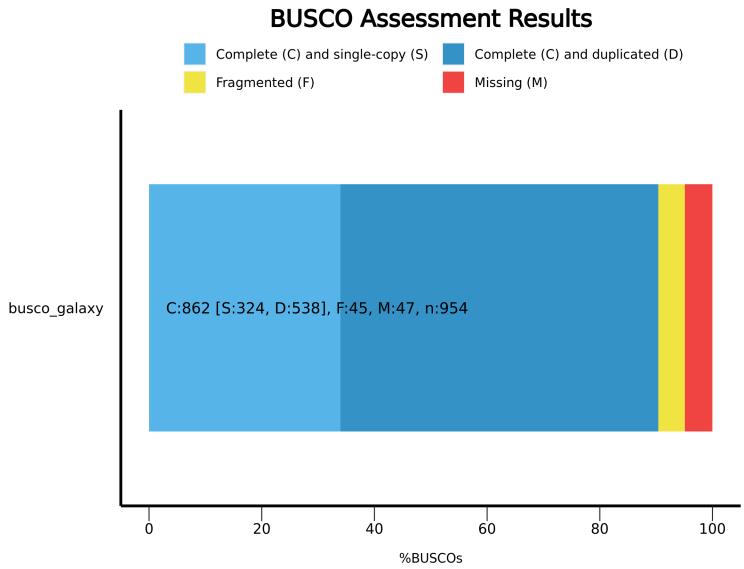


Figure 2.3: Busco results summary post-filtering for the longest isoform, against the metazoaodb10 lineage dataset. C:90.4 %[S:34.0 %,D:56.4 %],F:4.7 %,M:4.9 %,n:954. Where C marks complete BUSCOs, S: Complete and single-copy, D: complete and duplicated, F: fragmented, M: missing, n: total BUSCO groups searched.

Output of filtering: Retained 388731 / 477283 = 81.45% of total transcripts.

RSEM (RNA-Seq by Expectation-Maximization) software package was used to estimate gene and isoform expression levels from RNA-Seq data by aligning reads to the assembled transcriptome and then quantifying transcript abundance. It uses an expectation-maximization algorithm to accurately assign ambiguous reads and provide transcript expression levels [30]. This process determined which transcripts were either up or down-regulated in their gene expression, under the different treatment conditions. For the downstream analysis of this data, the RSEM transcript counts matrix output was used with DESeq2 in R (section 2.7).

2.7 Differential Expression Analysis

Differential expression was assessed through R and R studio with the DESeq2 package [46] using the not-normalized raw read count data matrix obtained from

RSEM as input. DESeq2 is suitable for handling data from non-model organisms that lack a fully annotated genome. DESeq2 normalizes the read counts to account for differences in library size and sequencing depth across samples. It then fits a negative binomial distribution model to the count data and uses a model to estimate the variance-mean dependence in the data, allowing for improved stability and accuracy of the differential expression estimates. DESeq2 then tests for differential expression by comparing conditions between samples using the Wald test, to provide log2 fold-changes and p-values for each gene [47].

Pre-filtering was performed prior to running DESeq, where transcripts having less than 10 reads total were removed, thus reducing the counts data from 477283 to 239392 observations across the 15 samples. The 'factor-level' for the differential expression analysis was set as the untreated control group (samples T1C6, T1C7, T1C8, T1C9 and T1C10). The 'intercept' against which differential expression was calculated was the T1S treatment group.

Summary of the DESeq output:

out of 239392 with nonzero total read count

adjusted $p - value < 0.1$

$LFC > 0(up) : 329, 0.14\%$

$LFC < 0(down) : 118, 0.049\%$

$outliers[1] : 25177, 11\%$

$lowcounts[2] : 118773, 50\%$

$(meancount < 2)$

Adjusting alpha as = 0.05

out of 239392 with nonzero total read count

adjusted $p - value < 0.05$

$LFC > 0(up) : 322, 0.13\%$

$LFC < 0(down) : 110, 0.046\%$
$outliers[1] : 25177, 11\%$
$lowcounts[2] : 154602, 65\%$
$(meancount < 7)$

Differential expression analysis thus detected a total of 432 unique transcripts exhibiting a significant change in expression (alpha = 0.05) attributed to treatment conditions. These results suggest that the thermal treatment triggers changes in gene expression that may influence cellular mechanisms. Noting that the results function of DESeq2 automatically performs independent filtering based on the mean of normalized counts for each gene, to optimize the number of transcripts given their adjusted p-value, according to the value of alpha [46].

To visualize the log2 fold change attributed to a given variable over the mean of normalized counts [46], the plotMA function is used, with alpha = 0.05, fig. A.6. The majority of the transcripts (grey) are not significantly differentially expressed as they cluster around the central line of log fold change = 0. The density of the points along the x-axis decreases as the mean of normalized counts increases, explained by the lower number of transcripts with very high expression levels. Consequently, the spread of variability (represented by the log fold change value) appeared higher for transcripts with lower mean expression levels; this is a common observation since measurement variability tends to be higher for lowly expressed transcripts [48].

A useful parameter for the visualization and ranking of transcripts is given by the shrinkage of effect size, known as LFC estimates. The log-fold-change shrink function, in which the significant transcripts based on fold-change were subset, was applied to stabilize variance estimates for transcripts with low counts to make significant changes more reliable. To compare shrinkage estimators, both 'normal' and 'apeglm' [49] type were applied and plotted (figs. A.7a and A.7b). Whilst all shrinkage methods account for "effect size" in ranking transcripts, using the log2fold change (LFC) across groups, the apeglm shrinkage method better preserves

variability in observed LFCs [34]. Similarly to the non-shrunken MA plot, the spread of the log fold changes seemed greater for transcripts with low counts, as expected. After the LFC shrinkage, type 'apeglm' seemed to have more transcripts being identified as significantly differentially expressed compared to the 'normal method', which appeared to produce a more conservative shrinkage. In addition, the spread of the significant transcripts appeared greater in the 'apeglm' plot, whereas in the 'normal' plot they were more centrally located. Where the 'normal' shrinkage is more conservative, the 'apeglm' method is sensitive to changes in expression, and is better suited for large log fold change estimations when dealing with low counts (see 2.7).

Plot counts are useful to examine the counts of reads for a single gene across the different treatment groups. The plot counts function normalizes the counts by the estimated size factors. The transcript with Trinity ID DN4310 c0 g2 i5 was identified as having the smallest p-adjusted value indicating that it is the most significant DE transcript between the conditions, fig. A.8. Log-fold change for this transcript is negative, indicating it was down-regulated under experimental conditions (see appendix A). In the plot count, the 'treated' condition showed a greater range of expression levels in comparison to the 'untreated' group, which has consistently low expression across the replicates. One possibility to explain such a response is that of a 'tank effect', where gene expression is influenced as a result of rearing conditions. When subsequently run through the transcriptome functional annotation software (section 2.8) the associated gene ontology was the biological process of protein phosphorylation (GO:0006468).

The transcript with the second-smallest p-value was additionally plotted by count read for comparison (fig. A.9). Here the 'control' group has consistently low counts, whereas both treated and untreated conditions convey greater variability in their expression. Log-fold change given for this transcript is positive, indicating its up-regulation under treatment conditions (see appendix A).

Transformed versions of the counts data are useful for visualization. The regu-

larized logarithm or rlog transformation produces data on the log2 scale which has been normalized, so as to stabilize the variance across the range of counts. This transformation method was chosen over the variance stabilizing transformation (VST) method due to the small sample size. The hierarchical clustering using the heat map (fig. 2.4) provides insight into the relationships between the transcripts and the samples. The rows, representing the clustering of the top 20 transcripts based on their mean expression level, suggest uniformity with the expression patterns across the samples.

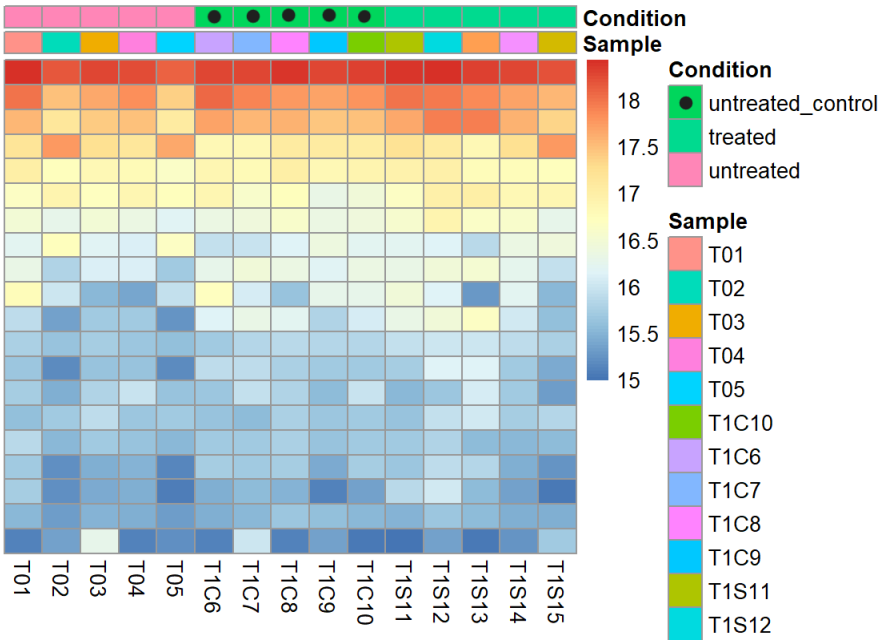


Figure 2.4: Heatmap of the rlog transformed counts matrix, showing the expression levels of genes across the different samples. Colors represent the range of rlog values, with red indicating higher expression and blue indicating lower expression. The x-axis represents the different samples and conditions; the y-axis represents the top 20 transcripts based on mean expression level, sorted according to mean expression values.

Further exploration of the transformed rlog data was achieved through sample clustering in order to visualize sample-to-sample distances. The following heatmap of the sample distance matrix gives an overview of the similarities and differences in gene expression profiles between samples and respective conditions according to their pairwise distances, fig. A.10. Both heatmaps seem to suggest inconsistent

responses with respect to the conditions according to the variability found within the clustering across the replicates. This could indicate technical artifacts or perhaps represent true biological variability.

The PCA plot is another useful analysis for visualizing the relationships between samples (fig. A.11) and identifying patterns between conditions(fig. 2.5). Default PCA plot settings use the 500 top features by variance. The x-axis, representing the first principal component, captures the greatest variance in the data, here given as 19 % of the total variance. Whilst this was substantial, it may suggest that there are additional factors that account for a significant portion of the variance in the data. The second principal component, or second-greatest variance in the data was given as 16 %. Given that the first two principal components summed explain 35 % of the variance, this was considered relatively low. There was a large amount of variance that might be explained by other components or was simply inherent noise in the data. Each data point representing a single sample was plotted according to its scores on the two principal components, thus providing a visual representation of the variability within and between conditions based on their gene expression profiles. The PCA plot by condition (fig. 2.5) shows the treated group (T1S, green dots) clusters tightly together in the upper left quadrant, suggesting that the treatment leads to a consistent change in gene expression across these samples. The untreated group (T0, blue dots) shows two samples clustering near the treated group, and one far away on the bottom left, which could indicate an outlier. In the PCA plot between samples (fig. A.11), this sample was identified as T03. The control group (T1C, red dots) has a greater spread, which could suggest greater variability in the gene expression within this group. One sample (T1C7, in the sample plot PCA, fig. A.11) was far along to the right from the others, suggesting it as different from the rest of the control group. The fact that the expression profiles of the sponges subjected to the thermal stress (treated) samples mostly cluster together away from the untreated samples suggests that the treatment had a significant effect on the gene expression profile. However, the untreated samples do not form a distinct cluster

separate from the treated samples, which may suggest that the treatment did not lead to a completely distinct gene expression profile or that other factors at play also have a strong influence on gene expression.

By sub-setting only the control and treated conditions (T1C and T1S) and plotting their PCA (fig. A.12) it can be seen that there is a greater observed variance in samples T1C7, T1S12, and T1S14, of the control and treated conditions. These are potential outliers or samples with a unique gene expression profile. Samples T1C6, T1C8, T1C9 and T1C10 of the untreated-control group have a tight cluster suggesting similar gene expression profiles and representing a consistent biological condition. Apart from the aforementioned outlier T1S13, the rest of the replicates of the treated condition (T1S) also appear to be grouped together consistently.

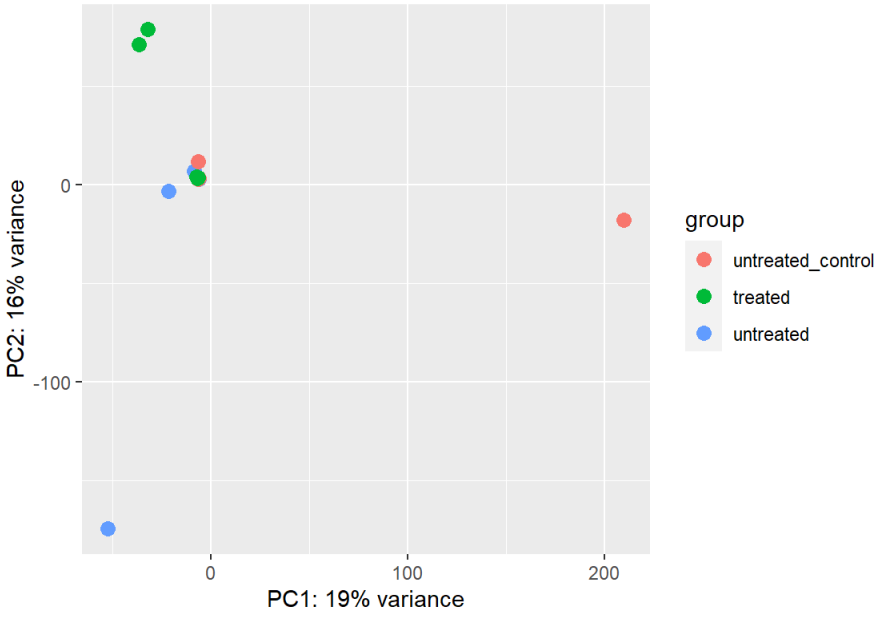


Figure 2.5: Principal Component Analysis (PCA) plot grouped by condition-type across samples of the top 500 features by variance. X and Y axes represent the variance in the data, where data points are plotted according to their variability within and between conditions based on their gene expression profile.

The Cook's distance function, calculated by DESeq, is a measure that assesses the influence of each sample on the model fit. It quantifies how much the fitted

values of the model would change if a particular observation were excluded. A Cook's distance greater than 1 indicates that the corresponding observation has a potentially large influence on the model fit. The boxplot (fig. A.13) of the Cook's distances for the sample data indicates that all samples have a Cook's distance less than 1, and does not highlight any one sample having a disproportionate influence on the model's results. Since all samples represented in the boxplot have relatively similar ranges, they suggest a consistent influence across samples, indicating the results of the analysis are robust and not driven by any singular outlier.

Plotting dispersion estimates is another useful diagnostic for visualizing variability by the mean of normalized counts (the average gene expression level) across all samples on a log scale (fig.A.14). The dispersion value on the y-axis quantifies how much the data varies around the expected mean, accounting for biological variability as well as technical noise. Each black point (gene-est) represents the estimated dispersion for an individual gene, which can be seen to decrease as the mean of normalized counts increases, which is typical because high-count transcripts tend to have more stable expression estimates and hence lower dispersion. The red trend line (fitted) shows the dispersion after fitting a model to the gene-wise estimates, to capture the general trend of dispersion across different expression levels. The blue points (final) represent the final estimates of dispersion used for the differential expression testing, accounting for factors such as gene-wise dispersion and other normalization factors. As seen in the left end of the dispersion plot (fig. A.14), there is a greater spread of dispersion estimates at the lower mean normalized counts, indicating that lowly expressed transcripts tend to have more variable expression. As the mean normalized counts increase, the dispersion estimates tend to decrease and converge, suggesting that highly expressed transcripts generally have more consistent expression across the samples. This is consistent with the previous MA plot (fig. A.6). The final shrinkage (blue dots) follows the trend line, indicating that the gene-wise dispersion has been adjusted suitably towards stable estimates based on the model. Thus indicating a good overall quality of the data and reliable

dispersion estimates.

Volcano plots provide a functional way to visualize the results of the differential expression analysis, including p-values, and log2 fold-changes [50]. In the volcano plot (fig. 2.6) the log2 fold change along the x-axis represents the down-regulated (negative, left) and up-regulated (positive, right) gene expressions between conditions. The y-axis represents the statistical significance of this differential expression, with higher values along the axis representing lower P-values. The red-colored points indicate transcripts that are both significantly differential expressed $p - values < 0.05$ and have fold changes above the threshold (1). Green-colored points represent transcripts that meet the threshold for fold-change, but not for statistical significance. Gray points represent non-significant transcripts that do not meet the threshold for fold-change or p-value.

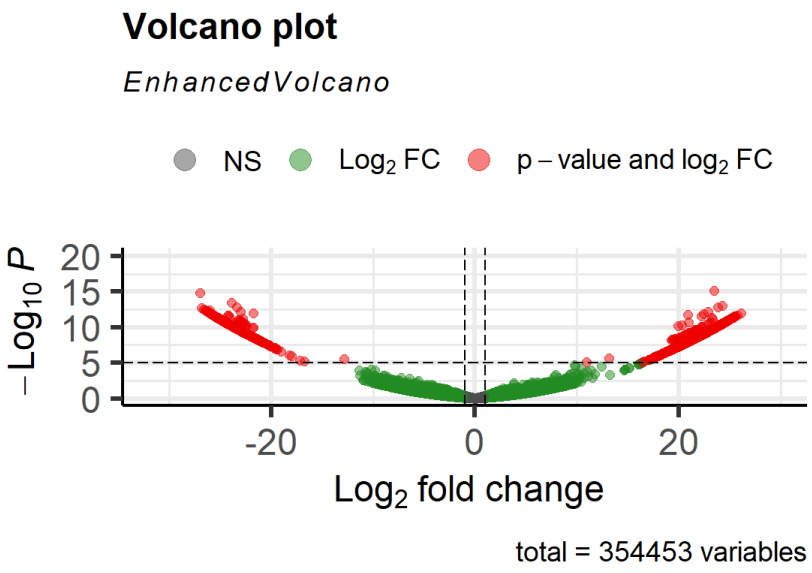


Figure 2.6: Volcano plot of the differential expression analysis data. The threshold for p-value is given as 0.05, and $\pm$ 1 for log2fold-change. Red-colored points indicate transcripts that meet both thresholds; green-colored points represent transcripts that meet the threshold for fold-change, but not for statistical significance; gray-points represent non-significant transcripts that do not meet the threshold for fold-change or p-value.

2.8 Transcriptome Functional Annotation

Significantly DE transcripts were identified using a threshold of 1 for Log-fold change and a $p - adjustedvalueof < 0.05$. Where $log2foldchange > 0$ was designated as up-regulated, and $log2foldchange < 0$ was down-regulated. A positive log2fold change of 1 indicates a doubling in expression, whereas a -1 log2fold change indicates a halving of expression value. The subset of these DE transcripts was used for the functional annotation, following the Blast2GO workflow [40] (see 2.3.3).

The following graphs give an overview of the Blast2GO project statistics (fig. A.15). The distribution of sequence lengths (fig. A.15a) indicates a distribution of varying sequence lengths, with the majority being small to medium-sized sequence lengths and a spare number of sequences above 3000 in length. This variability could reflect the biological diversity within the selected subset of transcripts. The data distribution plot (fig. A.15b) represents the number of sequences that reach each stage of the annotation process. Not all blasted sequences received GO annotations, which is common as not all terms meet the criteria used for confident annotations. Additionally, the number of sequences with no BLAST hits is notable. This could suggest poorly conserved transcripts, however, it is more likely a limitation in the reference database used since the selected sequences come from a non-model species [51].

The results table (fig. 2.7) summarizes the distribution of the top 20 GO terms by their category (BP: biological process; MF: molecular function; CC: cellular component). Direct GO count graphs were subsequently plotted to visualize all of the annotated GO terms by category. Enzyme code distribution was also graphed (fig. A.16).

The directed acyclic graph (DAG) generated by Blast2GO visualizes GO terms related to biological processes, molecular functions, and cellular components assigned to the selected set of sequences, integrating the structure of GO hierarchy

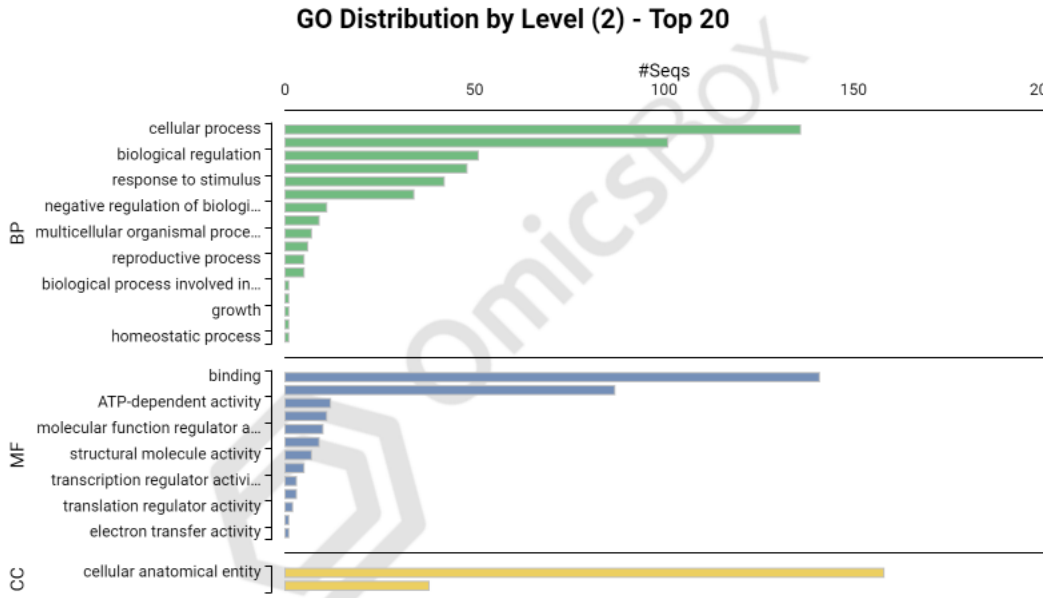


Figure 2.7: Summary of the top GO terms, grouped by category: BP (biological process), MF (molecular function), CC (cellular component), according to the number of sequences annotated for each term. Functional annotation was performed using only the differentially expressed transcripts, filtered according to thresholds (see 2.7).

into the annotation results (figs. A.18, A.19, A.20, A.21).

In addition to gene ontology descriptions, Blast2GO also characterizes the sequences, providing transcript-level descriptions based on shared similarity between the query and its sequence match. Certain transcript descriptions of the blasted sequences appeared to have a known role in stress response (tables A.6, A.7, A.8).

2.8.1 Enrichment Analysis (Fisher's Exact test)

Table of results from enrichment analysis using the down-regulated transcripts identified in DESeq2 (see 2.7) as the test sequences, table A.9. A bar chart for visual representation was also produced (fig. 2.8). Enrichment analysis identified significant activation of gene pathways involving binding, metabolic processes, and stress response. To be thorough, the same enrichment analysis was run using the up-regulated transcripts as the test sequences, with all annotated transcripts as the

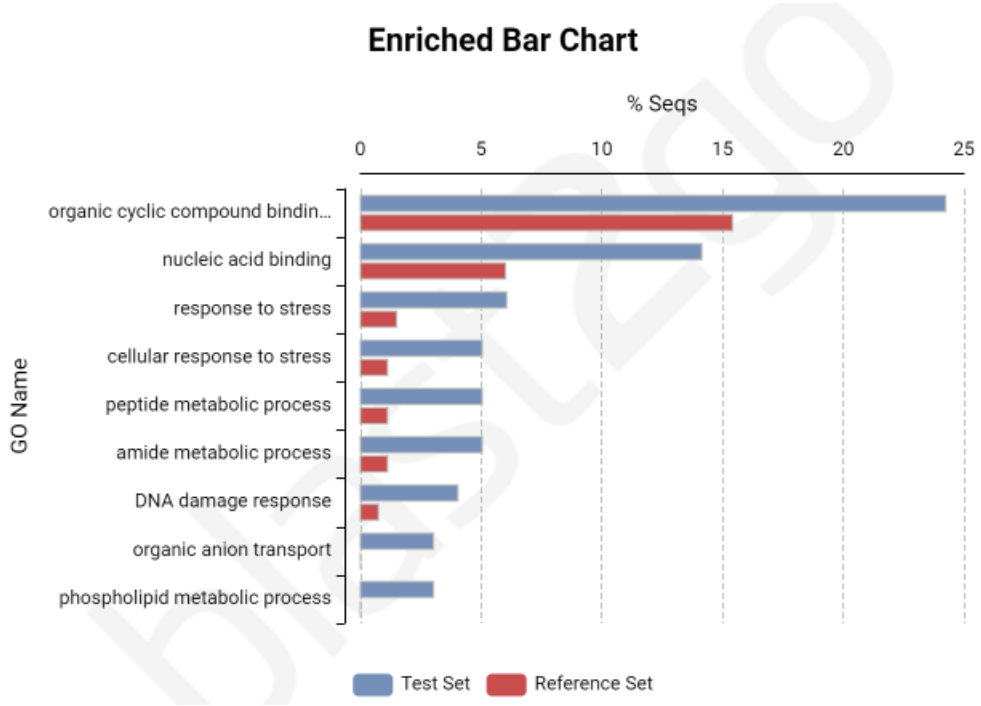


Figure 2.8: Enrichment analysis bar chart, using the down-regulated transcripts identified in DESeq2 as the test sequences. The reference set of sequences includes all of the sequences identified as differentially expressed (see 2.7).

reference set (table A.10 and fig A.22).

Discussion

2.9 Quality assessment of RNA extraction

Typical protocols used for the assessment of RNA quality and quantity in eukaryotes (e.g. Agilent Technologies Bioanalyzer2100) are established for vertebrates, and thus have a strong correlation with their corresponding RIN number. For many groups of invertebrates, the RIN number presents a limitation in characterizing the

RNA quality, as it has been shown in non-model organisms that the 28s rRNA band will break into smaller pieces, and possibly also form aggregations [52, 53]. As such, it was not expected that the RNA extraction of *C. nucula* represent clean 28s bands in its electrophoresis gel (fig. A.1). Furthermore, the total RNA extraction protocol was optimized to include the small RNAs, including the 5srRNA for its potential contribution to protein synthesis [54]. This would have produced an additional, unexpected peak if using the Agilent electrophoresis protocol, resulting in a ”poor” RIN number, as seen in the Novogene quality control results (fig. A.2). Having taken all of this into consideration, the RNA quality and quantity assessment, via UV spectroscopy and gel electrophoresis produced satisfactory results, and nonetheless received a passing score.

2.10 Quality assessment of the transcriptome assembly

The total number of unique ’transcripts’ identified in the Trinity assembly was given as 367,754 in fig. A.3. The standard metric N50 gives a length of 620 bases such that 50% of the assembled transcriptome was contained in continuous sequences that are at least this long. To avoid misrepresentation, ExN50 was plotted in figure A.4, indicating that the highly expressed transcripts tend to have a longer length, where the peak at which transcripts being considered include the longest contiguous sequences ($>$1000 bases long) contributing to the total assembled bases for that level of expression. Less expressed transcripts introduce shorter contigs and thus we see a decreasing ExN50 value to the right of the peak.

Busco results identify an acceptable percentage of 92.2 % complete Busco matches. The estimated percentage of duplicated transcripts as seen in the non-filtered transcriptome (fig. A.5), may be inflated by the presence of several isoforms of the same gene assembled by Trinity. Consequentially, the dataset was filtered so that only the longest isoform per gene was kept, which gives a reduction in this redundancy (fig. 2.3) [43]. Nonetheless, prior-to and post-filtering BUSCO results

both indicate relatively high levels of duplications, which could signal the presence of contaminant species or reflect a recent whole-genome duplication event. [31]. Another consideration is the selected lineage dataset that was used as a reference for comparison. In this case, the most specific dataset available was the metazoa (metazoa odb9) lineage; non-model, understudied organisms will lack sufficiently specific lineage datasets necessary for such analysis. This may result in inaccurate assessments of completeness, as the benchmark dataset may not capture the full diversity of transcripts present in the non-model species, *C. nucula*.

2.11 Differential expression analysis with DESeq2

Whilst the expression profiles of certain sample replicates showed larger variability, PCA analysis still observed changes across samples between conditions. Differential expression detected a total of 422 transcripts demonstrating a significant change in their expression across all treatments. With 322 transcripts as significantly up-regulated and 100 as down-regulated (see Appendix A).

Generally, the down-regulation of transcripts represents proteins that may be involved in maintaining homeostasis under normal conditions and thus could be energy intensive. Their down-regulation may be a mechanism for energy conservation. Up-regulated sequences in the context of thermal stress would most likely be attributed to stress-induced transcripts, consistent with several transcriptome studies [55, 56, 57]. Notably, thermal stress in *C. nucula* resulted in more up-regulation than down-regulation of transcripts (see 2.7), which is consistent with select transcriptome thermal studies [58, 59], and differing from others [60, 55]. The change in differential expression, quantified by the log2 fold change, indicated similar distributions between up and down-regulated transcripts, as visually illustrated by the volcano plot (see fig. 2.6, and appendix A.4).

In using plot counts to examine read counts across conditions for the most significantly expressed transcripts, cross-referencing with Blast2GO annotations

highlighted certain processes (fig. A.8). Protein phosphorylation was identified as the most significant and down-regulated biological process. Although it did not result as significantly enriched by Fisher's exact test, protein phosphorylation has been cited for its role in protein synthesis during thermal stress and adaptation [61]. Under heat stress conditions, protein phosphorylation is a critical component of the heat shock response pathways, leading to the activation of specific protein kinases, which in turn could phosphorylate downstream target proteins. Target proteins may include transcription factors, chaperone proteins, or other regulatory proteins involved in the heat shock response. The consequences of phosphorylation may include regulation of transcription factor activity; protein stabilization to prevent their degradation and maintain homeostasis; modulation of protein-protein interactions and cellular signaling pathways involved in the heat shock response; and regulation of enzyme activity, allowing for cells to adapt their metabolism in response to the stress [62].

The second most differentially expressed transcript, by p-value, was also plotted by its read counts across the experimental conditions (fig. A.9). TRINITY-DN940-c0-g1-i13 transcript was annotated as 'cytochrome P450' and involved in the molecular function of 'iron binding', 'steroid hydroxylase activity', 'oxidoreductase activity', and 'heme binding' [40]. Several papers have similarly observed changes in cytochrome P450 gene expression through exposure to heat stress conditions [63, 64]. In coral and cnidaria studies, cytochrome P450 has been cited for its involvement as a chemical defense mechanism, particularly in cell detoxification and protection from oxidative stress [64, 65].

2.12 Blast2GO functional annotation

Of the 422 transcripts identified as differentially expressed, 87 transcripts had no match in the Blast query. The top GO terms distributed across sequences and identified by the Blast2GO workflow included cellular processes (136 transcripts),

biological regulation (101 transcripts), binding (141 transcripts), ATP-dependent activity (87 transcripts), and cellular anatomical entity (158 transcripts) (fig. 2.7). The only directly associated Hsp (heat-shock protein), labeled as Hsp20 of the alpha crystallin family and attributed to TRINITY ID DN199443-c0-g1-i1 was found to be down-regulated. This finding is consistent with thermal studies performed on coral species and their symbionts [66, 67].

Of the transcript-level descriptions, 'fibronectin-like proteins' were frequently associated with molecular function 'protein binding' (table A.6). Fibronectin, an extracellular matrix glycoprotein has been proposed as a stress-responsive gene, binding directly to HSPs (heat-shock proteins) and being involved in cell adhesion and migration [68]. Ubiquitin carboxyl-terminal hydrolases were identified amongst the blasted transcripts; ubiquitination has been identified as being involved in the signaling of damaged proteins under stress [69] (table A.7). A cellular study on heat shock suggests the role of ubiquitination as an adaptive mechanism for recovery from thermal stress [70]. Ubiquitin proteins have been found important for organism survival in their regulation of cellular replication and proliferation [71]. Several cytochromes were identified; cytochromes participate in respiratory electron transport chains, as well as in reduction-oxidation regulation [72] (table A.8). Cytochrome b5 reductase specifically has revealed the induction of more efficient metabolic function, in response to energetic stress [73, 74]. Cytochrome c oxidase has also shown involvement in the heat-shock response (HSR), being activated in cells and tissues as a protective mechanism [75, 76].

Enzymatic activity

Transcript annotation also resulted in the identification of enzymatic activity (fig. A.16). Enzyme classes included lyases, oxidoreducutases, transferases, and hydrolases, distributed between 1 and 9 % of sequences. Lyases enzymes play a role in metabolic pathways involving the cleavage of bonds in molecules without the addi-

tion of water, a vital function in the breakdown and synthesis of various compounds. Similarly, hydrolases enzymes catalyse the breakdown of compounds containing compounds in the process of recycling or removing damaged molecules. Oxidoreductases participate in bioregulation synthesis and oxidative stress by controlling redox states within cells, preventing oxidative damage to tissues, consistent with other thermal stress studies on porifera [60] and coral [77, 78]. Transferases transfer functional groups between molecules, vital for modifying proteins or metabolites [79, 80, 81]. Collectively, the enzyme classes indicate an adjustment of cellular processes to mitigate the stress effects by ensuring stability and functionality.

It should be highlighted that the results of this study are based on transcriptomics data, which does not always correspond with the translated proteins [82, 83]. As such, enrichment analysis was performed to determine the probabilities (p-value) of observing joint values, in this case, of gene ontologies [84].

Fisher's Exact test

Enrichment analysis indicated 9 significantly enriched GO terms when comparing the down-regulated sequences (test-set, fig. 2.8) to all differentially expressed transcripts (reference set). Significantly enriched GO terms included 'organic cyclic compound binding', involving the interactions of proteins and organic cyclic compounds which may include various types of signaling molecules, hormones, or other metabolites. The ability of proteins to bind to these compounds is crucial for several cellular processes, including signal transduction, metabolic pathway regulation, and response to environmental stress. As such, under thermal stress, changes in this binding activity could be indicative of altered cellular signaling pathways, crucial in initiating the stress response [85, 86]. A recent transcriptome study on shrimp species *Palaemon caridean* under stress similarly found differential expression in organic cyclic compound binding functions [85]. 'Nucleic acid binding' refers to the molecular function where proteins interact specifically and

non-covalently with nucleic acids. This function is essential for a variety of cellular processes including transcription, translation, replication, and repair of nucleic acids, such as with heat-induced DNA or RNA damage. Proteins with nucleic acid binding activity play essential roles in regulating gene expression and maintaining genomic integrity by participating in the stabilization of nucleic acid structures [87]. Studies suggest the expression of nucleic acid binding as part of the immune response [88], which may be consistent with this analysis. Thus, 'nucleic acid binding' activity may be pivotal for cellular protection and stress adaptation, ensuring the integrity and functionality of genetic information [88].

'Response to stress' and 'cellular response to stress' are broad BP terms used in gene ontology to categorize the cellular, physiological, or systematic changes that occur as a reaction to stress, including temperature extremes. The response aim is to protect the cell or organism, restore homeostatic conditions, and often will involve alterations in gene expression, protein functions, and metabolic pathways. [89, 71, 90, 91, 58].

'Peptide metabolic process' encompasses all of the chemical reactions and pathways involving peptides, their modifications, and degradation. Peptide metabolism is crucial for numerous cellular functions including signaling, regulation, and as protein precursors [92]. Phospholipids play an important role in the regulation of biophysical properties, protein sorting, and cell signaling pathways of porifera [93]. 'Phospholipid metabolic process' encompasses the various biochemical activities involving phospholipids, components of cellular membranes. Their metabolism involves modifications to the lipid bilayer that may affect the fluidity and integrity of membranes. In the context of heat stress, down-regulation of the phospholipid metabolism may serve as a protective measure in adjusting the composition of their membranes, so as to maintain membrane integrity. This physical adjustment may additionally serve to conserve energy for critical functions, and possibly influence signaling pathways, altering the stress response [93, 94]. The 'amide metabolic process' also plays several important roles, including protein metabolism, nucleotide

synthesis, and signaling. Metabolism is significantly influenced by changes in temperature [95], even resulting in a metabolic system collapse under extreme stress scenarios [96]. One study on the metabolism of *C. nucula* suggests the strong influence of temperature above 26°C on respiration rates, whilst declining in food intake, possibly indicating a negative energy balance under stress exposure [97].

'Organic anion transport' is crucial for physiological functions, including the elimination of metabolic waste. In the context of heat stress, this could indicate a detoxification response. 'DNA damage response' refers to a complex network of cellular pathways that are activated in response to several types of DNA damage, including strand breaks, base modifications, and cross-links. This damage-response mechanism is necessary for maintaining genomic stability, preserving genetic information, and preventing mutations that could further damage the organism [60, 58]. The activation of this function could indicate that the sponge cells have already sustained damage.

Together, these significantly enriched terms suggest that upon exposure to heat stress, transcripts associated with cellular mechanisms involving lipid metabolism, transport processes, and DNA repair appear to make a coordinated effort in maintaining cellular integrity and function [92].

When instead testing the up-regulated transcripts against all of the significant DE transcripts (as the reference set), the following GO terms are identified as significantly enriched: 'regulation of catalytic activity', 'regulation of molecular function', and 'vesicle-mediated transport' as biological processes; and, 'cytoskeleton' as a cellular component (table A.10, fig. A.22). These terms suggest cellular adaptation to enhance or regulate metabolic processes, molecular interactions, and transport mechanisms as a response to environmental stress. The involvement of the cytoskeleton could allude to changes in cellular structure, a commonly observed strategy of reparative regeneration to isolate and recover damaged tissue [98, 99]. Having said this, no visible signs of tissue damage, or morphogenesis were observed during the course of the experiment.

2.13 Concluding remarks

The transcriptome analysis revealed insights into the mechanisms involved in the response of *C. nucula* to thermal stress conditions, mimicking temperatures in a recent Mediterranean MHW event. Related studies on thermal stress responses have demonstrated damaging effects such as oxidative stress, protein malfunction, and increased pathogen susceptibility [100, 77, 78, 101, 102]. Such reactions were observed in this study by the activation of 'response to stress' and 'DNA damage response' transcripts. The transcript expression of *C. nucula* indicated that several processes involved with biological regulation, particularly in binding and metabolism were triggered under exposure to stress. These responses may be considered short-term defensive responses to the temperature shift, having involved functions designed to prevent or minimize cellular damage. Although not significantly enriched, 5 DEGs exhibited 'calcium ion binding' (fig. A.17), and 2 expressed calcium ion import. Ca^{+2} has been indicated as an important ion in the thermosensibility of sponge *Spongia officinalis* [58]. Another consideration for the observedly subtle response of *C. nucula* to the simulated MHW may be that the temperature fluctuation was within the standard regime of the source population. As such, a temperature extreme was not induced, but rather the effect of rapid subjection to a temperature shift. Additionally, being a shallow water sponge, *C. nucula* may be adequately resilient to such rapid fluctuations and hence display a more controlled gene expression profile.

As previously mentioned, one of the limitations of this study was the lack of a specific reference genome. Functional annotation was performed using the default nr/nr (non-redundant protein sequence) database. Not all transcripts were successfully annotated; proteins that are specific to *C. nucula* may lack representation in the reference dataset, leading to an incomplete annotation of the transcriptome. Further studies would be necessary to illuminate the functions of these unannotated transcripts. This study provides a baseline of the gene expression response of *C.*

nucula to short-term temperature elevation. Future experiments addressing the effects of prolonged exposure to heat stress, alongside additional environmental stressors, such as acidity, could elucidate sponge genetic systems involved in recovery and survival. This is particularly relevant in the context of predicting future impacts of climate change on fundamental marine species, such as porifera. The comprehensive appraisal of the cellular, molecular, and biological processes that elicit sponge expression response mechanisms to environmental stimuli in the face of the changing climate will better support their management and conservation, as well as essential symbiotic relationships [103]. Considering the progressively stressful marine environment, only the most resilient of species can be expected to thrive.

Acknowldgements

First and foremost, my endless gratitude towards my mentor Elisavet Kaitetzidou, without whom none of this work would have been possible. For your patience, your wisdom and your friendship, I am beyond grateful. To all of the wonderful souls I met during my time in Crete, $\Sigma\epsilon\ \epsilon\upsilon\chi\alpha\rho\iota\sigma\tau\omega\ \pi\alpha\rho\alpha\ \rho o\lambda\upsilon$. You have all given me such colorful memories that I will forever cherish. To my family and my dogs, Pisello and Banana, I am so thankful to be surrounded by such wonderful souls that support me, even in the tougher moments. To Vittorio, if it were not for you, the adventure would not have started. I am humbled by the strength and kindness shared with me. I wish you every happiness. A special mention to Sci-Hub and all of its proxies, because ”knowledge belongs to all mankind”.

Endless gratitude goes to the members of the MACCIMO team who have contributed their time and wisdom for this project with me.

The research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers” (Project Number: 03280 MACCIMO - Multilevel Approaches to assess Climate Change Impact to Marine Organisms).

This research was supported in part through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI.

References

[1] Hoesung Lee et al. “IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.” In: (2023).

[2] James J Bell. “The functional roles of marine sponges”. In: *Estuarine, coastal and shelf science* 79.3 (2008), pp. 341–353.

[3] Michelle Klautau et al. “Does cosmopolitanism result from overconservative systematics? A case study using the marine sponge Chondrilla nucula”. In: *Evolution* 53.5 (1999), pp. 1414–1422.

[4] Klaus Rützler. “The role of burrowing sponges in bioerosion”. In: *Oecologia* 19.3 (1975), pp. 203–216.

[5] Klaus Rützler and Ian G Macintyre. “Siliceous sponge spicules in coral reef sediments”. In: *Marine Biology* 49.2 (1978), pp. 147–159.

[6] Klaus Ruetzler. “Sponges on coral reefs: a community shaped by competitive cooperation”. In: (2004).

[7] Janie L Wulff. “Ecological interactions of marine sponges”. In: *Canadian Journal of Zoology* 84.2 (2006), pp. 146–166.

[8] A Padua and M Klautau. “Regeneration in calcareous sponges (Porifera)”. In: *Journal of the Marine Biological Association of the United Kingdom* 96.2 (2016), pp. 553–558.

[9] Thomas L Frölicher, Erich M Fischer, and Nicolas Gruber. “Marine heatwaves under global warming”. In: *Nature* 560.7718 (2018), pp. 360–364.

[10] Thomas Wernberg et al. “Climate change increases marine heatwaves harming marine ecosystems”. In: *ScienceBrief Crit. Issues Climate Change Sci* (2021).

[11] Gerald A Meehl and Claudia Tebaldi. “More intense, more frequent, and longer lasting heat waves in the 21st century”. In: *Science* 305.5686 (2004), pp. 994–997.

[12] Dan A Smale et al. “Marine heatwaves threaten global biodiversity and the provision of ecosystem services”. In: *Nature Climate Change* 9.4 (2019), pp. 306–312.

[13] Katherine E Mills et al. “Fisheries management in a changing climate: lessons from the 2012 ocean heat wave in the Northwest Atlantic”. In: *Oceanography* 26.2 (2013), pp. 191–195.

[14] Miguel Ñiquen and Marilú Bouchon. “Impact of El Niño events on pelagic fisheries in Peruvian waters”. In: *Deep sea research part II: topical studies in oceanography* 51.6-9 (2004), pp. 563–574.

[15] Joaquim Garrabou et al. “Mass mortality in Northwestern Mediterranean rocky benthic communities: effects of the 2003 heat wave”. In: *Global change biology* 15.5 (2009), pp. 1090–1103.

[16] Dan A Smale and Thomas Wernberg. “Extreme climatic event drives range contraction of a habitat-forming species”. In: *Proceedings of the Royal Society B: Biological Sciences* 280.1754 (2013), p. 20122829.

[17] C Cerrano et al. “A catastrophic mass-mortality episode of gorgonians and other organisms in the Ligurian Sea (North-western Mediterranean), summer 1999”. In: *Ecology letters* 3.4 (2000), pp. 284–293.

[18] Thomas Wernberg et al. “Climate-driven regime shift of a temperate marine ecosystem”. In: *Science* 353.6295 (2016), pp. 169–172.

[19] Jordan A Thomson et al. “Extreme temperatures, foundation species, and abrupt ecosystem change: an example from an iconic seagrass ecosystem”. In: *Global Change Biology* 21.4 (2015), pp. 1463–1474.

[20] BE Brown. “Damage and recovery of coral reefs affected by El Niño related seawater warming in the Thousand Islands, Indonesia”. In: *Coral reefs* 8 (1990), pp. 163–170.

[21] Matthew S Edwards. “Estimating scale-dependency in disturbance impacts: El Niños and giant kelp forests in the northeast Pacific”. In: *Oecologia* 138 (2004), pp. 436–447.

[22] Joaquim Garrabou et al. “Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea”. In: *Global Change Biology* 28.19 (2022), pp. 5708–5725.

[23] Daniel Gómez-Gras et al. “Response diversity in Mediterranean coralligenous assemblages facing climate change: Insights from a multispecific thermotolerance experiment”. In: *Ecology and Evolution* 9.7 (2019), pp. 4168–4180.

[24] D. Denaxa and the POSEIDON scientific team. “Marine Heat Wave in the Aegean Sea in June 2021.” In: (2021).

[25] Thermo Scientific. *T042-Technical Bulletin NanoDrop Spectrophotometers, 260/280 and 260/230 Ratios*. 2010.

[26] Haris Zafeiropoulos et al. “0s and 1s in marine molecular research: A regional HPC perspective”. In: *GigaScience* 10.8 (2021), giab053.

[27] Simon Andrews et al. *FastQC*. Babraham Institute. Babraham, UK, Jan. 2012.

[28] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[29] Manfred G Grabherr et al. “Full-length transcriptome assembly from RNA-Seq data without a reference genome”. In: *Nature biotechnology* 29.7 (2011), pp. 644–652.

[30] Bo Li and Colin N Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12 (2011), pp. 1–16.

[31] Mosè Manni et al. "BUSCO: assessing genomic data quality and beyond". In: *Current Protocols* 1.12 (2021), e323.

[32] Kiran Gopinath Bankar et al. "Ameliorated de novo transcriptome assembly using Illumina paired end sequence data with Trinity Assembler". In: *Genomics data* 5 (2015), pp. 352–359.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[34] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12 2014), p. 550. DOI: 10.1186/s13059-014-0550-8.

[35] Martin Morgan and Marcel Ramos. *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.22. 2023. URL: https://CRAN.R-project.org/package=BiocManager.

[36] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[37] Kevin Blighe, Sharmila Rana, and Myles Lewis. *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*. R package version 1.20.0. 2023. DOI: 10.18129/B9.bioc.EnhancedVolcano. URL: https://bioconductor.org/packages/EnhancedVolcano.

[38] Raivo Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. 2019. URL: https://CRAN.R-project.org/package=pheatmap.

[39] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3. 2022. URL: https://CRAN.R-project.org/package=RColorBrewer.

[40] Stefan Götz et al. "High-throughput functional annotation and data mining with the Blast2GO suite". In: *Nucleic acids research* 36.10 (2008), pp. 3420–3435.

[41] Odilo Mueller, Samar Lightfoot, and Andreas Schroeder. "RNA integrity number (RIN)–standardization of RNA quality control". In: *Agilent application note, publication* 1 (2004), pp. 1–8.

[42] Marvin Mundry et al. "Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach". In: *PloS one* 7.2 (2012), e31410.

[43] Venket Raghavan et al. "A simple guide to de novo transcriptome assembly and annotation". In: *Briefings in bioinformatics* 23.2 (2022), bbab563.

[44] Felipe A Simão et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19 (2015), pp. 3210–3212.

[45] Marie-Agnès Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings in bioinformatics* 14.6 (2013), pp. 671–683.

[46] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), pp. 1–21.

[47] Michael Love, Simon Anders, and Wolfgang Huber. "Differential analysis of count data–the DESeq2 package". In: *Genome Biol* 15.550 (2014), pp. 10–1186.

[48] Charlotte Soneson, Michael I Love, and Mark D Robinson. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". In: *F1000Research* 4 (2015).

[49] Anqi Zhu, Joseph G Ibrahim, and Michael I Love. "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences". In: *Bioinformatics* 35.12 (2019), pp. 2084–2092.

[50] Adam McDermaid et al. "Interpretation of differential gene expression results of RNA-seq data: review and integration". In: *Briefings in bioinformatics* 20.6 (2019), pp. 2044–2054.

[51] Hamish McWilliam et al. "Analysis tool web services from the EMBL-EBI". In: *Nucleic acids research* 41.W1 (2013), W597–W600.

[52] Danielle M DeLeo et al. "RNA profile diversity across arthropoda: guidelines, methodological artifacts, and expected outcomes". In: *Biology Methods and Protocols* 3.1 (2018), bpy012.

[53] Ramiro Barcia, José Marıa Lopez-Garcıa, and Juan Ignacio Ramos-Martınez. "The 28S fraction of rRNA in molluscs displays electrophoretic behaviour different from that of mammal cells". In: *IUBMB Life* 42.6 (1997), pp. 1089–1092.

[54] Maciej Szymanski et al. "5S ribosomal RNA database". In: *Nucleic Acids Research* 30.1 (2002), pp. 176–178.

[55] A Moya et al. "The transcriptomic response to thermal stress is immediate, transient and potentiated by ultraviolet radiation in the sea anemone Anemonia viridis". In: *Molecular Ecology* 21.5 (2012), pp. 1158–1174.

[56] Jeremie Vidal-Dupiol et al. "Genes related to ion-transport and energy production are upregulated in response to CO2-driven pH decrease in corals: new insights from transcriptome analysis". In: *PloS one* 8.3 (2013), e58652.

[57] Xuelin Zhao et al. "Transcriptomic responses to salinity stress in the Pacific oyster Crassostrea gigas". In: (2012).

[58] Vasiliki Koutsouveli et al. "Gearing up for warmer times: transcriptomic response of Spongia officinalis to elevated temperatures reveals recruited mechanisms and potential for resilience". In: *Frontiers in Marine Science* 6 (2020), p. 786.

[59] Marcelo González-Aravena et al. "Warm temperatures, cool sponges: the effect of increased temperatures on the Antarctic sponge Isodictya sp." In: *PeerJ* 7 (2019), e8088.

[60] Christine Guzman and Cecilia Conaco. "Gene expression dynamics accompanying the sponge thermal stress response". In: *PloS one* 11.10 (2016), e0165368.

[61] Soyoung Park et al. "Modulation of protein synthesis by eIF2α phosphorylation protects cell from heat stress-mediated apoptosis". In: *Cells* 7.12 (2018), p. 254.

[62] Roger F Duncan and JW Hershey. "Protein synthesis and protein phosphorylation during heat stress, recovery, and adaptation." In: *The Journal of cell biology* 109.4 (1989), pp. 1467–1481.

[63] Yu-Cheng Wang et al. “High temperature stress induces expression of CYP450 genes and contributes to insecticide tolerance in Liriomyza trifolii”. In: *Pesticide Biochemistry and Physiology* 174 (2021), p. 104826.

[64] Nedeljka N Rosic et al. “Differential regulation by heat stress of novel cytochrome P450 genes from the dinoflagellate symbionts of reef-building corals”. In: *Applied and environmental microbiology* 76.9 (2010), pp. 2823–2829.

[65] Jared V Goldstone. “Environmental sensing and response genes in cnidaria: the chemical defensome in the sea anemone Nematostella vectensis”. In: *Cell biology and toxicology* 24 (2008), pp. 483–502.

[66] Sarah L Gierz, Sylvain Forêt, and William Leggat. “Transcriptomic analysis of thermally stressed Symbiodinium reveals differential expression of stress and metabolism genes”. In: *Frontiers in plant science* 8 (2017), p. 271.

[67] Nedeljka N Rosic et al. “Gene expression profiles of cytosolic heat shock proteins Hsp70 and Hsp90 from symbiotic dinoflagellates in response to thermal stress: possible implications for coral bleaching”. In: *Cell Stress and Chaperones* 16 (2011), pp. 69–80.

[68] Karim Colin Hassan Dhanani, William John Samson, and Adrienne Lesley Edkins. “Fibronectin is a stress responsive gene regulated by HSF1 in response to geldanamycin”. In: *Scientific reports* 7.1 (2017), p. 17617.

[69] H Shen et al. “Oxidative stress regulated expression of ubiquitin Carboxyl-terminal Hydrolase-L1: role in cell survival”. In: *Apoptosis* 11 (2006), pp. 1049–1059.

[70] Brian A Maxwell et al. “Ubiquitination is essential for recovery of cellular activities after heat shock”. In: *Science* 372.6549 (2021), eabc3593.

[71] Yvain Desplat et al. “Morphological and transcriptional effects of crude oil and dispersant exposure on the marine sponge Cinachyrella alloclada”. In: *Science of The Total Environment* 878 (2023), p. 162832.

[72] John B Schenkman and Ingela Jansson. “The many roles of cytochrome b5”. In: *Pharmacology & therapeutics* 97.2 (2003), pp. 139–152.

[73] Dong-Hoon Hyun and Ga-Hyun Lee. “Cytochrome b5 reductase, a plasma membrane redox enzyme, protects neuronal cells against metabolic and oxidative stress through maintaining redox state and bioenergetics”. In: *Age* 37 (2015), pp. 1–14.

[74] Robert Hall et al. “Cytochrome b5 reductases: Redox regulators of cell homeostasis”. In: *Journal of Biological Chemistry* (2022), p. 102654.

[75] Sebastian Vogt et al. “Heat shock protein expression and change of cytochrome c oxidase activity: presence of two phylogenic old systems to protect tissues in ischemia and reperfusion”. In: *Journal of bioenergetics and biomembranes* 43 (2011), pp. 425–435.

[76] Hsiang-Wen Chen et al. “Cytochrome c oxidase as the target of the heat shock protective effect in septic liver”. In: *International journal of experimental pathology* 85.5 (2004), pp. 249–256.

[77] MK DeSalvo et al. “Differential gene expression during thermal stress and bleaching in the Caribbean coral Montastraea faveolata”. In: *Molecular ecology* 17.17 (2008), pp. 3952–3971.

[78] Christian R Voolstra et al. “Effects of temperature on gene expression in embryos of the coral Montastraea faveolata”. In: *BMC genomics* 10 (2009), pp. 1–9.

[79] Thiang Yian Wong, Lori A Preston, and Neal L Schiller. “Alginate lyase: review of major sources and enzyme characteristics, structure-function analysis, biological roles, and applications”. In: *Annual Reviews in Microbiology* 54.1 (2000), pp. 289–340.

[80] Liquan Wu et al. “Five pectinase gene expressions highly responding to heat stress in rice floral organs revealed by RNA-seq analysis”. In: *Biochemical and Biophysical Research Communications* 463.3 (2015), pp. 407–413.

[81] Rui Zeng et al. “Study on differential protein expression in natural selenium-enriched and non-selenium-enriched rice based on iTRAQ quantitative proteomics”. In: *Biomolecules* 9.4 (2019), p. 130.

[82] Anderson B Mayfield et al. “The proteomic response of the reef coral Pocillopora acuta to experimentally elevated temperatures”. In: *PloS one* 13.1 (2018), e0192001.

[83] Xiaoyan Peng et al. “Integration of the proteome and transcriptome reveals multiple levels of gene regulation in the rice dl2 mutant”. In: *Frontiers in Plant Science* 6 (2015), p. 351.

[84] Lynne M Connelly. “Fisher’s exact test”. In: *MedSurg Nursing* 25.1 (2016), pp. 58–60.

[85] Amandine D Marie et al. “Transcriptomic response to thermal and salinity stress in introduced and native sympatric Palaemon caridean shrimps”. In: *Scientific Reports* 7.1 (2017), p. 13980.

[86] Elisa Zampieri et al. “Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles”. In: *Scientific Reports* 6.1 (2016), p. 25773.

[87] Martin Bartas et al. “Amino acid composition in various types of nucleic acid-binding proteins”. In: *International Journal of Molecular Sciences* 22.2 (2021), p. 922.

[88] Aldo Nicosia et al. “The nucleic acid-binding protein PcCNBP is transcriptionally regulated during the immune response in red swamp crayfish Procambarus clarkii”. In: *Cell Stress and Chaperones* 21 (2016), pp. 535–546.

[89] Ana Riesgo et al. “Transcriptomic analysis of differential host gene expression upon uptake of symbionts: a case study with Symbiodinium and the major bioeroding sponge Cliona varians”. In: *BMC genomics* 15.1 (2014), pp. 1–22.

[90] Yongguo Li, Kunyin Jiang, and Qi Li. “Comparative transcriptomic analyses reveal differences in the responses of diploid and triploid Pacific oysters (Crassostrea gigas) to thermal stress”. In: *Aquaculture* 555 (2022), p. 738219.

[91] Jose Maria Aguilar-Camacho and Grace P McCormack. “Molecular responses of sponges to climate change”. In: *Climate Change, Ocean Acidification and Sponges: Impacts Across Multiple Levels of Organization* (2017), pp. 79–104.

[92] Natasja Krog Noer et al. “Temporal regulation of temperature tolerances and gene expression in an arctic insect”. In: *Journal of Experimental Biology* 226.11 (2023), jeb245097.

[93] Emilie Genin et al. “New trends in phospholipid class composition of marine sponges”. In: *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 150.4 (2008), pp. 427–431.

[94] Carl Djerassi and Wai Kwan Lam. “Phospholipid studies of marine organisms. Part 25. Sponge phospholipids”. In: *Accounts of chemical research* 24.3 (1991), pp. 69–75.

[95] Ofir Levy et al. “Ontogeny constrains phenology: opportunities for activity and reproduction interact to dictate potential phenologies in a changing climate”. In: *Ecology Letters* 19.6 (2016), pp. 620–628.

[96] Raffaella Pantile and Nicole Webster. “Strict thermal threshold identified by quantitative PCR in the sponge Rhopaloeides odorabile”. In: *Marine Ecology Progress Series* 431 (2011), pp. 97–105.

[97] Mar Bosch-Belmar et al. “Effect of Acute Thermal Stress Exposure on Ecophysiological Traits of the Mediterranean Sponge Chondrilla nucula: Implications for Climate Change”. In: *Biology* 13.1 (2023), p. 9.

[98] Alexander V Ereskovsky et al. “Oscarella lobularis (Homoscleromorpha, Porifera) regeneration: epithelial morphogenesis and metaplasia”. In: *PloS one* 10.8 (2015), e0134566.

[99] Charlotte C Runzel. *Sponge physiology: the effects of temperature on the regeneration and reaggregation of sponges (Haliclona reniera)*. Tech. rep. PeerJ Preprints, 2016.

[100] Sophie Richier et al. “Oxidative stress and apoptotic events during thermal stress in the symbiotic sea anemone, Anemonia viridis”. In: *The FEBS journal* 273.18 (2006), pp. 4186–4198.

[101] Nicole S Webster, Rose E Cobb, and Andrew P Negri. “Temperature thresholds for bacterial symbiosis with a sponge”. In: *The ISME journal* 2.8 (2008), pp. 830–842.

[102] Emma Cebrian Pujol et al. “Sponge Mass Mortalities in a Warming Mediterranean Sea: Are Cyanobacteria-Harboring Species Worse Off?” In: *PLoS ONE, 2011, vol. 6, núm. 6, p. e20211* (2011).

[103] Malcolm Hill et al. “Sponge-specific bacterial symbionts in the Caribbean sponge, Chondrilla nucula (Demospongiae, Chondrosida)”. In: *Marine Biology* 148 (2006), pp. 1221–1230.

Appendix A

Tables and Figures

A.1 RNA extraction and quality assessment

Table A.1: Nanodrop 1000 spectrophotometer measurements of the *C. nucula* RNA extractions

Sample ID	Date	ng/ul	260/280	260/230	Constant	Cursor Pos.
T0_1	03/03/2023	655.96	1,95	0,86	40,00	230
T0_2	03/13/2023	388.19	1,95	1,32	40,00	230
T0_3	03/03/2023	815.61	1,84	1,40	40,00	230
T0_4	03/03/2023	1041.27	1,95	1,43	40,00	230
T0_5	03/13/2023	1281.72	2,02	1,91	40,00	230
T1C_6	03/03/2023	1499.85	1,96	0,79	40,00	230
T1C_7	03/03/2023	559.39	1,66	0,62	40,00	230
T1C_8	03/03/2023	720.22	1,79	0,93	40,00	230
T1C_9	03/03/2023	291.96	1,91	1,17	40,00	230
T1C_10	03/03/2023	1180.44	1,94	0,99	40,00	230
T1S_11	03/03/2023	881.62	1,98	1,87	40,00	230
T1S_12	03/03/2023	410.6	1,96	1,41	40,00	230
T1S_13	03/03/2023	895.49	1,79	0,81	40,00	230
T1S_14	03/03/2023	397.24	1,93	1,39	40,00	230
T1S_15	03/23/2023	1286.41	1,92	1,26	40,00	230

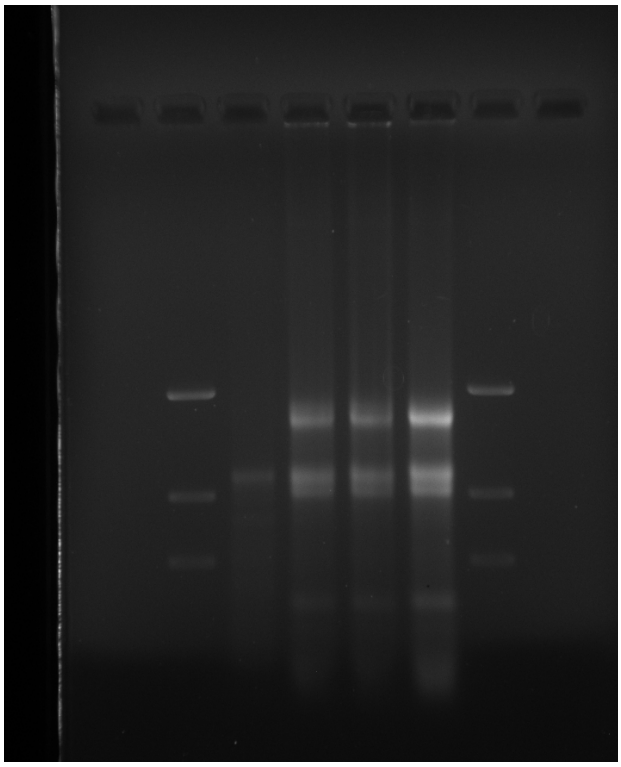


Figure A.1: Agarose gel electrophoresis imaging; wells 4,5 and 6 are separate samples of *C. nucula*, wells 2 and 7 are ladders

Table A.2: Novogene Quality Control Results Summary, to evaluate the quantity, integrity and purity requirements of *C. nucula* samples.

Sample ID	Nucleic Acid ID	Conc. (ng/ul)	Vol. (ul)	Amount(ug)	RIN	QC
T0_1	EKRN230021450-1A	39.52	16	0.63228	6.5	Pass
T0_2	EKRN230021451-1A	86.11	16	1.37783	7.2	Pass
T0_3	EKRN230021452-1A	43.77	16	0.70040	5.8	Pass
T0_4	EKRN230021453-1A	30.00	18	0.54004	5.5	Pass
T0_5	EKRN230021454-1A	106.61	17	1.81231	6.3	Pass
T1C_6	EKRN230021455-1A	63.20	17	1.07445	5.3	Pass
T1C_7	EKRN230021456-1A	47.47	17	0.80706	5.6	Pass
T1C_8	EKRN230021457-1A	89.26	17	1.51741	5.6	Pass
T1C_9	EKRN230021458-1A	54.73	16	0.87575	7.1	Pass

Table A.2: Novogene Quality Control Results Summary, to evaluate the quantity, integrity and purity requirements of *C. nucula* samples.

Sample ID	Nucleic Acid ID	Conc. (ng/ul)	Vol. (ul)	Amount(ug)	RIN	QC
T1C_10	EKRN230021459-1A	43.62	17	0.74149	5.7	Pass
T1S_11	EKRN230021460-1A	66.59	17	1.13203	6.3	Pass
T1S_12	EKRN230021461-1A	43.37	16	0.69396	6.2	Pass
T1S_13	EKRN230021462-1A	85.93	17	1.46081	5	Pass
T1S_14	EKRN230021463-1A	56.16	17	0.95472	6.5	Pass
T1S_15	EKRN230021464-1A	73.08	17	1.24241	6.5	Pass

A.2 De Novo Transcriptome Assembly

Basic Statistics

Measure	Value
Filename	T0_1_EKRN230021450-1A_HCK25DSX7_L3_2.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22505645
Sequences flagged as poor quality	0
Sequence length	150
%GC	46

Per base sequence quality

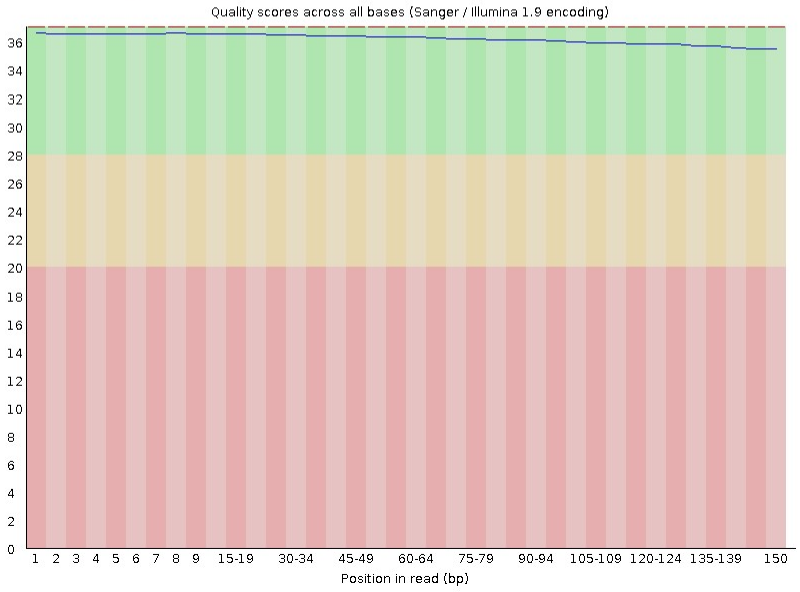


Figure A.2: FastQC quality control output summary for sample T01. Post-trimming FastQC resulted with no sequences flagged as poor quality; as exemplified. All sequences were reported between 36 and 150 basepairs in length; the maximum read length was determined by the selected number of cycles for sequencing. Graph illustrating per base sequence quality for sample T01

Table A.3: FastQC Results Summary illustrating the percentages of sequences surviving the Trimming step. All subsequent data elaboration and analysis were performed using only trimmed and paired sequences.

Condition	Replicates	Average total raw sequences	Average total trimmed-paired sequences	Average % of trimmed-paired sequences	Average % GC content of trimmed paired sequences
Initial Control	T01 T02 T03 T04 T05	24659283	24269608.8	98.42	46
Control Tank	T1C6 T1C7 T1C8 T1C9 T1C10	23293305	22906076	98.31	46
Experimental Tank	T1S11 T1S12 T1S13 T1S14 T1S15	22000740.2	21539419.6	97.91	46

A.2.1 Transcriptome Assembly QA: Basic Contig Statistics

```
################################
## Counts of transcripts, etc.
################################
Total trinity 'genes':  367754
Total trinity transcripts:      477283
Percent GC: 45.14

########################################
Stats based on ALL transcript contigs:
########################################

        Contig N10: 3230
        Contig N20: 2102
        Contig N30: 1466
        Contig N40: 984
        Contig N50: 620

        Median contig length: 271
        Average contig: 490.89
        Total assembled bases: 234291960


#####################################################
## Stats based on ONLY LONGEST ISOFORM per 'GENE':
#####################################################

        Contig N10: 2707
        Contig N20: 1595
        Contig N30: 875
        Contig N40: 498
        Contig N50: 352

        Median contig length: 259
        Average contig: 392.51
        Total assembled bases: 144346870
```

Figure A.3: Trinity assembly statistics where the N50 statistic corresponds to the sequence length such that all contigs of at least that length compose at least 50 % of the assembly [42].

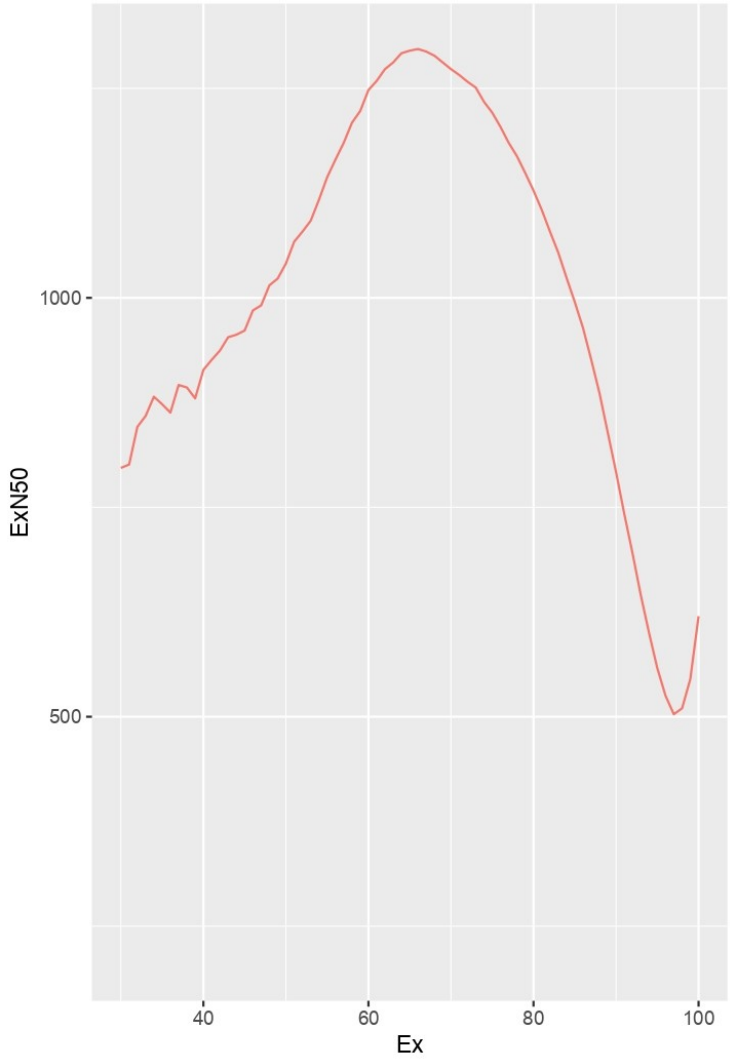


Figure A.4: Plot of Ex value against ExN50. The ExN50 [43] metric takes into account only the most highly expressed transcripts representing a percentage of the total normalized expression data.

A.2.2 Quantitative Assessment: Busco

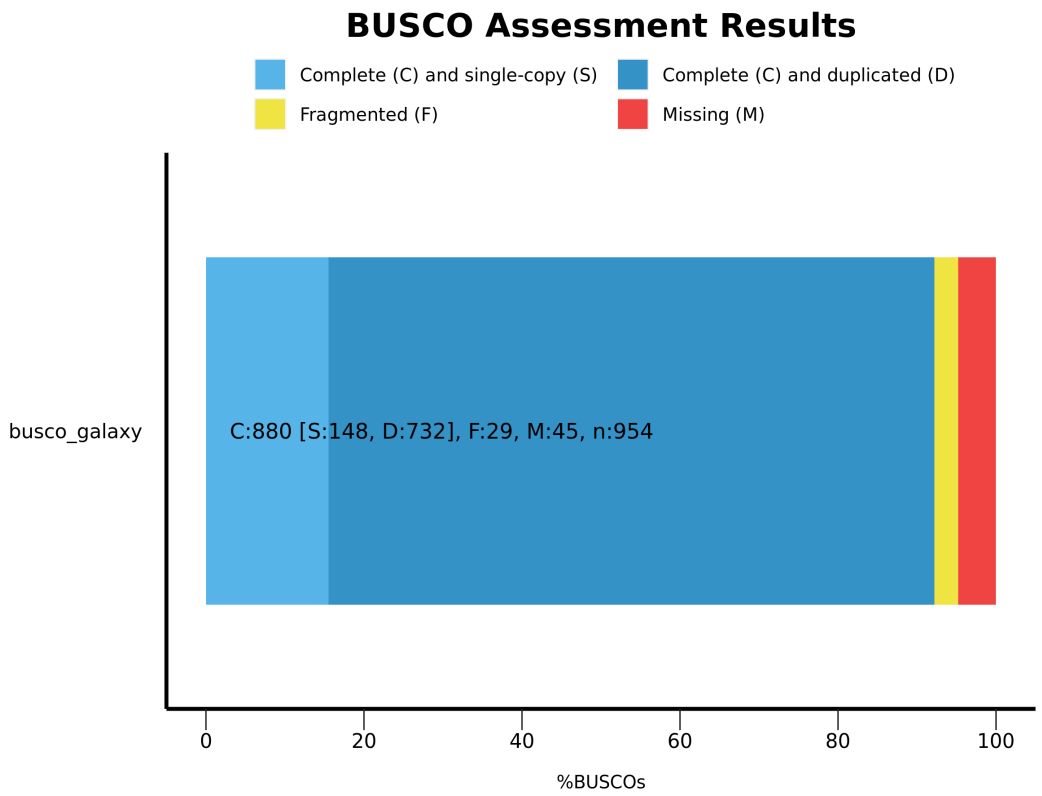


Figure A.5: Busco results summary prior to filtering, against the metazoaodb10 lineage dataset. C:92.2 % [S:15.5 %,D:76.7 %],F:3.0 %,M:4.8 %,n:954. Where C marks complete BUSCOs, S: Complete and single-copy, D: complete and duplicated, F: fragmented, M: missing, n: total BUSCO groups searched.

A.3 Differential Expression Analysis

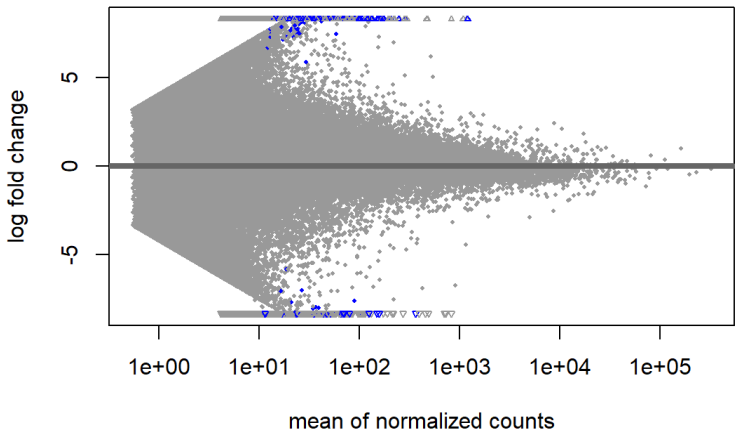
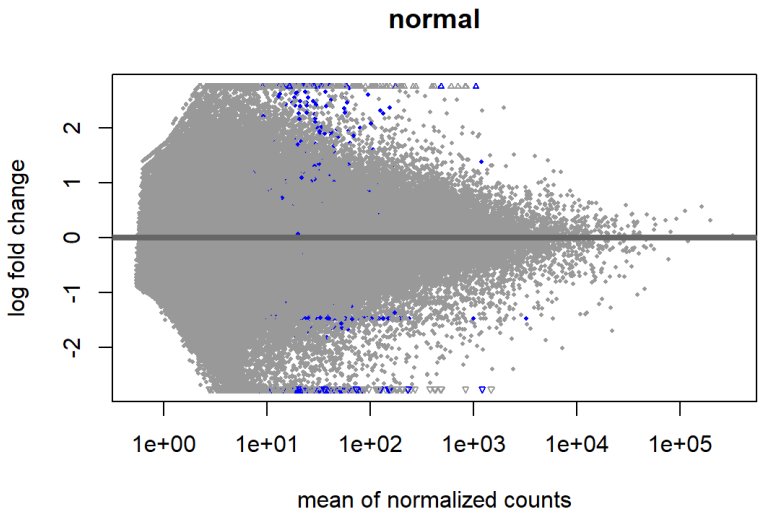
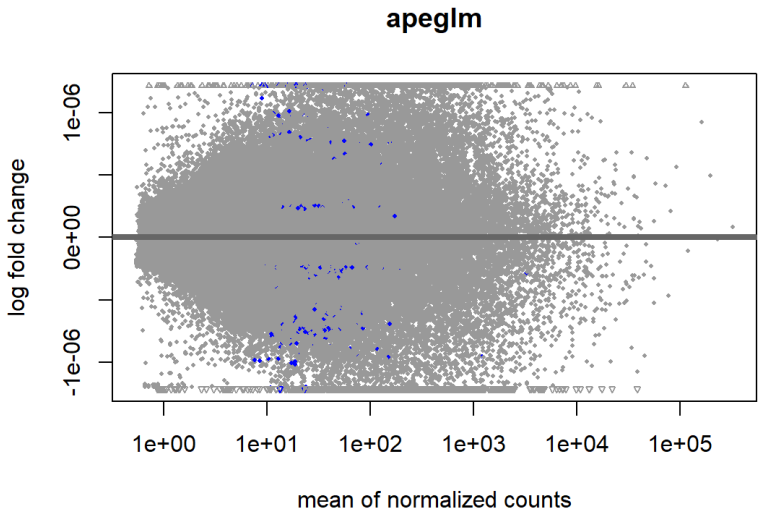


Figure A.6: MA-plot, showing the log2 fold changes from the treatment over the mean of normalized counts between treated and control conditions, where blue points above and below the threshold lines represent results that are significantly up or down-regulated, respectively, with statistical significance according to the given alpha = 0.05 value



(a) MA-plot with lfc shrinkage 'normal', showing the log2 fold changes, using 'normal' statistical shrinkage to improve LFC estimates, from the treatment over the mean of normalized counts between treated and control conditions. Blue points above and below the threshold lines represent results that are significantly up or down-regulated, respectively, with statistical significance according to the given alpha = 0.05 value



(b) MA-plot with lfc shrinkage 'apeglm', showing the log2 fold changes, using 'apeglm' statistical shrinkage to improve LFC estimates, from the treatment over the mean of normalized counts between treated and control conditions. Blue points above and below the threshold lines represent results that are significantly up or down-regulated, respectively, with statistical significance according to the given alpha = 0.05 value

Figure A.7: Shrinkage LFC estimates

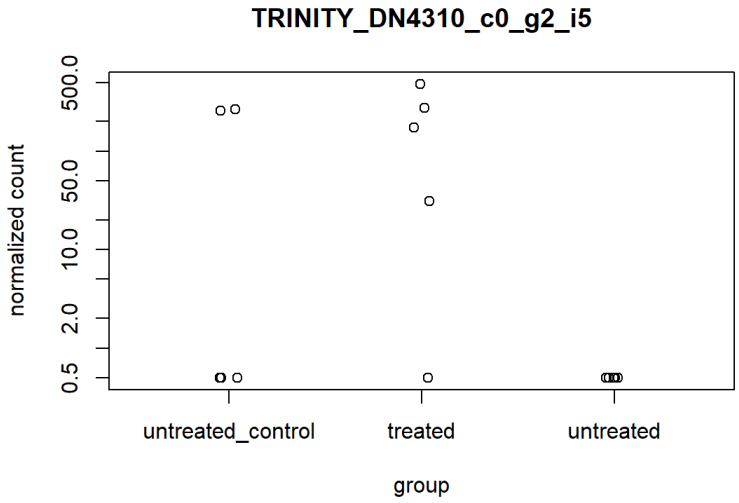


Figure A.8: Plot count of reads for the transcript with the smallest p-value, across the different conditions. Plot counts function normalizes the counts by the estimated size factors.

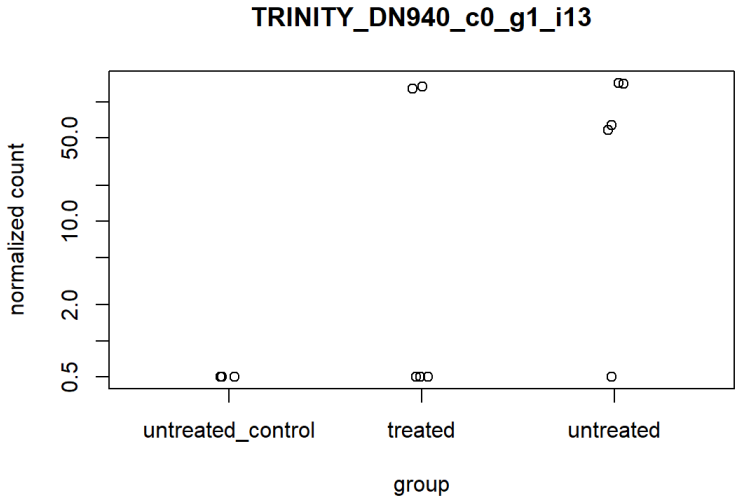


Figure A.9: Plot count of reads for the gene with the second smallest p-value, across the different conditions. Plot counts function normalizes the counts by the estimated size factors

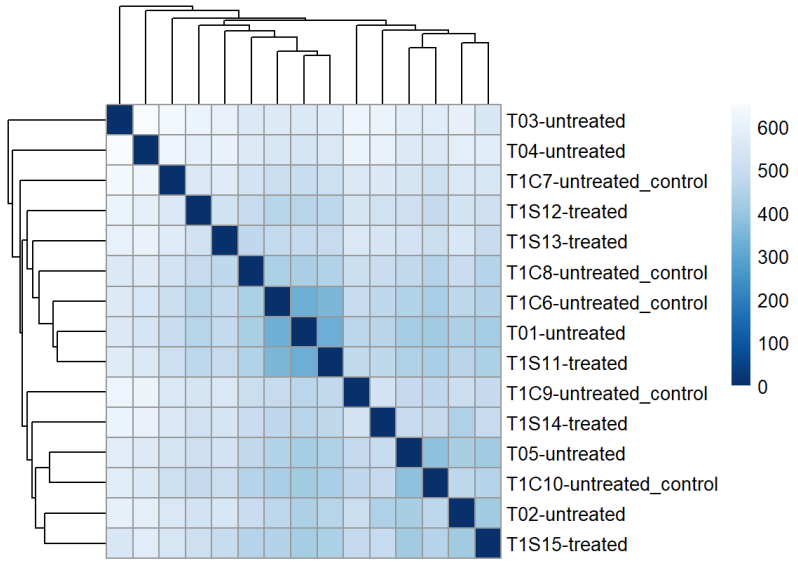


Figure A.10: Heatmap of the rlog transformed samples distance matrix, illustrating the similarity or dissimilarity between samples based on their gene expression profiles. The darker shades of blue represent smaller distances and thus greater similarity.

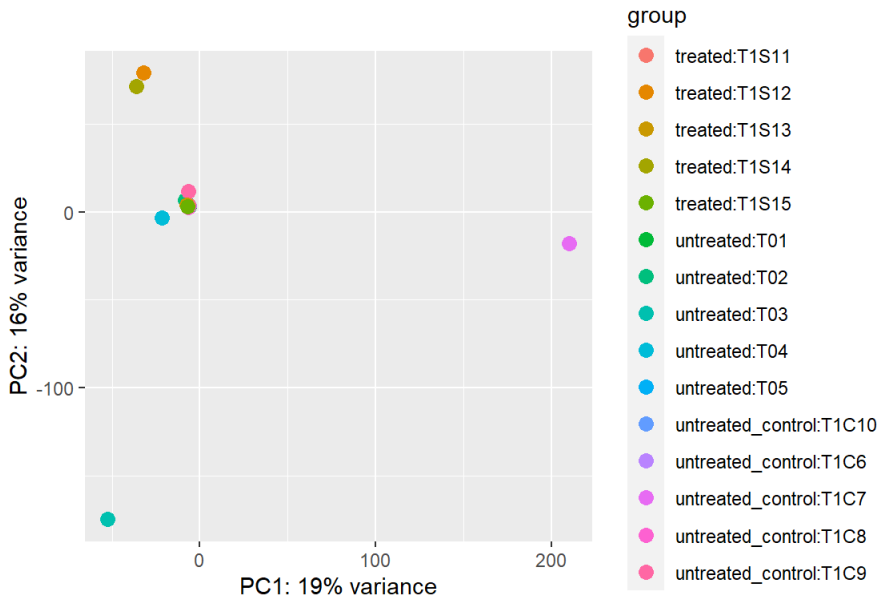


Figure A.11: PCA plot grouped by condition and sample-type using the top 500 features by variance. X and Y axes represent the variance in the data, where data points are plotted according to their variability within and between conditions based on their gene expression profile.

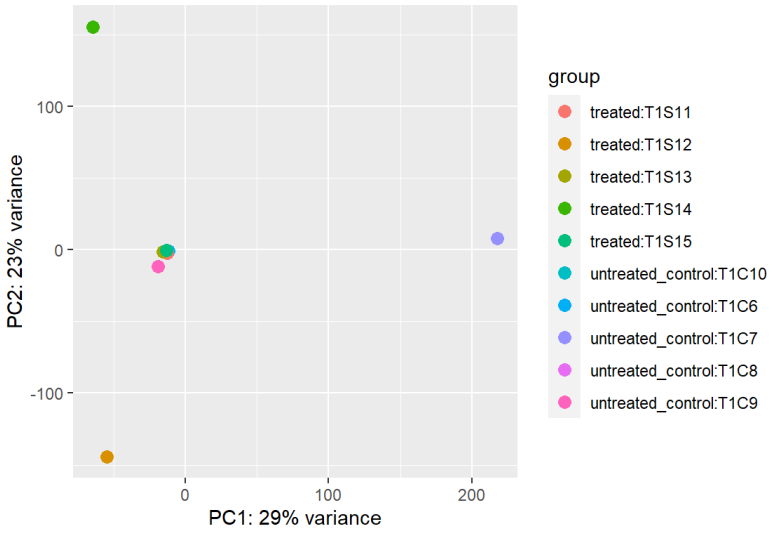


Figure A.12: PCA plot of only T1 samples, using the top 500 features by variance. X and Y axes represent the variance in the data, where data points are plotted according to their variability within and between conditions based on their gene expression profile.

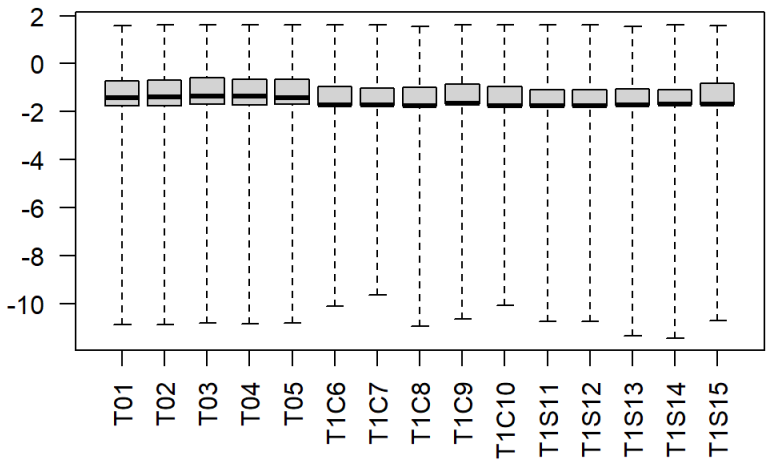


Figure A.13: Boxplot of the Cook's distances between samples assessing the influence of each sample on the model fit. A Cook's distance greater than 1 indicates that the corresponding observation potentially influences the model fit.

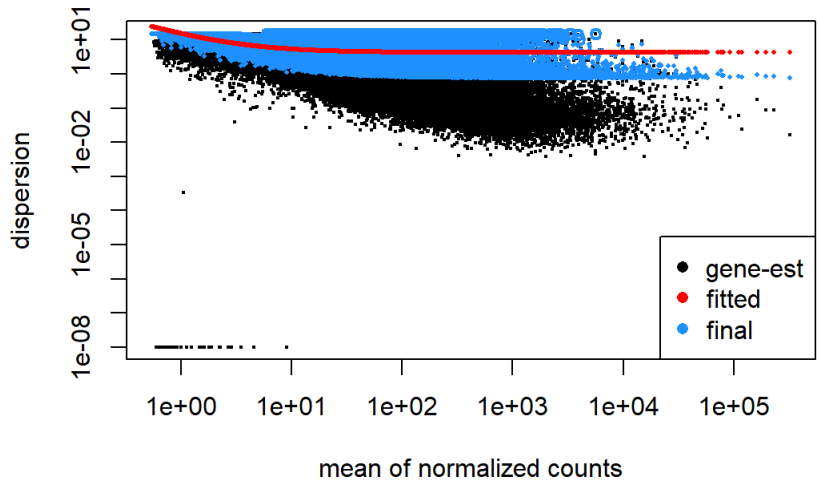


Figure A.14: Dispersion estimate plot, visualizing the variability by the mean of normalized counts (the average gene expression level) across all samples. Transcripts with low counts are towards the left of the x-axis, increasing towards the right. The y-axis represents the dispersion estimate for each gene on a log scale.

A.4 DESeq2 significant DE genes

Significantly DE transcripts were identified using a threshold of 1 for Log-fold change and a $p - adjustedvalueof < 0.05$. Where $log2foldchange > 0$ was designated as up-regulated, and $log2foldchange < 0$ was down-regulated.

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN940_c0_g1_i13	44.7710	22.2040	2.3813	9.3242	1.1185e-20	3.3337e-16
DN4662_c0_g1_i1	62.0725	22.0747	2.4089	9.1639	5.0054e-20	9.9463e-16
DN3474_c0_g1_i2	107.4187	23.1290	2.5441	9.0914	9.7784e-20	1.4573e-15
DN12701_c0_g1_i8	47.9665	21.8284	2.4231	9.0086	2.0869e-19	2.4881e-15
DN82348_c0_g1_i2	32.0153	21.4564	2.3899	8.9781	2.7550e-19	2.5078e-15
DN4584_c0_g1_i3	69.2354	22.3472	2.4911	8.9708	2.9448e-19	2.5078e-15

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN20042_c2_g1_i2	29.3393	20.7691	2.4065	8.6303	6.1176e-18	4.5586e-14
DN12472_c0_g1_i2	20.8024	21.2517	2.5697	8.2702	1.3368e-16	8.8544e-13
DN8925_c0_g1_i8	283.4272	24.1272	3.1175	7.7392	1.0007e-14	5.9657e-11
DN115878_c0_g1_i1	20.3122	19.7407	2.5563	7.7223	1.1429e-14	6.1937e-11
DN10002_c0_g1_i11	11.0182	19.4582	2.6868	7.2422	4.4162e-13	2.1939e-9
DN5213_c0_g1_i4	9.3029	18.4671	2.6338	7.0117	2.3546e-12	8.2567e-9
DN6355_c0_g2_i1	207.1960	25.0113	3.6498	6.8529	7.2392e-12	1.9616e-8
DN62786_c0_g1_i2	172.2007	24.9550	3.6498	6.8374	8.0651e-12	2.0904e-8
DN17754_c0_g1_i1	1199.8297	24.9203	3.6498	6.8279	8.6150e-12	2.1399e-8
DN16208_c0_g2_i17	200.5602	24.8037	3.6498	6.7959	1.0761e-11	2.4673e-8
DN785_c0_g1_i16	246.0253	24.7060	3.6498	6.7692	1.2953e-11	2.8599e-8
DN532_c0_g1_i20	154.0880	24.6281	3.6498	6.7478	1.5011e-11	3.0814e-8
DN140_c0_g1_i1	76.4247	24.5390	3.6499	6.7232	1.7781e-11	3.4192e-8
DN1218_c0_g1_i7	84.1559	24.3259	3.6499	6.6648	2.6498e-11	4.3398e-8
DN795_c1_g1_i2	54.7284	22.8396	3.4365	6.6461	3.0088e-11	4.5990e-8
DN6377_c0_g1_i1	89.1647	23.4553	3.5585	6.5914	4.3564e-11	6.4320e-8
DN1568_c0_g1_i14	57.6917	22.7573	3.4538	6.5891	4.4237e-11	6.4320e-8
DN6502_c1_g1_i1	46.7418	23.9929	3.6500	6.5733	4.9209e-11	6.9845e-8
DN4005_c0_g1_i7	75.3519	23.9373	3.6499	6.5583	5.4416e-11	7.5439e-8
DN2264_c0_g2_i2	133.9047	23.8681	3.6499	6.5394	6.1752e-11	8.1805e-8
DN60375_c0_g2_i3	132.3478	23.8244	3.6499	6.5274	6.6901e-11	8.3656e-8
DN4603_c0_g1_i1	59.9348	23.8212	3.6500	6.5264	6.7359e-11	8.3656e-8
DN26324_c0_g1_i2	77.8213	23.7492	3.6499	6.5068	7.6768e-11	9.1527e-8
DN454_c0_g1_i10	118.5879	23.6819	3.6499	6.4884	8.6775e-11	9.9480e-8
DN10921_c0_g1_i11	68.3063	23.6318	3.6499	6.4746	9.5072e-11	1.0305e-7
DN40500_c0_g2_i3	49.8559	22.4949	3.4793	6.4653	1.0111e-10	1.0531e-7

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN3132_c0_g1_i21	65.4801	23.5906	3.6499	6.4633	1.0246e-10	1.0531e-7
DN7962_c0_g2_i1	48.6632	23.5789	3.6500	6.4599	1.0477e-10	1.0586e-7
DN15707_c0_g1_i1	42.6390	23.5270	3.6501	6.4456	1.1515e-10	1.1440e-7
DN33249_c0_g1_i1	38.7349	23.4615	3.6501	6.4276	1.2964e-10	1.1930e-7
DN10969_c0_g1_i2	27.9386	21.4956	3.3443	6.4275	1.2971e-10	1.1930e-7
DN56178_c0_g1_i1	47.4048	23.4409	3.6500	6.4221	1.3442e-10	1.1930e-7
DN35457_c0_g6_i1	33.9329	23.3326	3.6502	6.3922	1.6355e-10	1.3356e-7
DN2610_c0_g1_i11	38.8648	23.0779	3.6502	6.3224	2.5755e-10	1.7393e-7
DN20717_c1_g1_i1	34.0743	23.0626	3.6502	6.3182	2.6462e-10	1.7393e-7
DN6961_c0_g2_i7	58.6077	23.0615	3.6500	6.3182	2.6469e-10	1.7393e-7
DN5595_c0_g1_i1	150.2126	23.0597	3.6500	6.3177	2.6550e-10	1.7393e-7
DN3_c1_g1_i7	43.1627	23.0148	3.6501	6.3053	2.8766e-10	1.8367e-7
DN3890_c0_g3_i2	30.8318	22.9650	3.6502	6.2914	3.1465e-10	1.9096e-7
DN310174_c0_g1_i1	37.9084	22.9542	3.6501	6.2886	3.2034e-10	1.9096e-7
DN563_c0_g2_i3	17.4441	20.7257	3.2980	6.2843	3.2932e-10	1.9136e-7
DN9385_c0_g1_i1	25.6511	22.9371	3.6503	6.2835	3.3096e-10	1.9136e-7
DN16599_c0_g2_i7	60.3182	22.9300	3.6501	6.2821	3.3408e-10	1.9136e-7
DN3_c0_g1_i23	55.5805	22.9245	3.6500	6.2807	3.3705e-10	1.9136e-7
DN15218_c0_g1_i1	25.7768	22.9038	3.6503	6.2744	3.5094e-10	1.9268e-7
DN32018_c0_g2_i1	38.3618	22.8807	3.6501	6.2685	3.6461e-10	1.9268e-7
DN1869_c0_g1_i4	40.0761	21.7983	3.4788	6.2660	3.7044e-10	1.9268e-7
DN14583_c0_g1_i3	65.5982	22.8709	3.6501	6.2659	3.7068e-10	1.9268e-7
DN26333_c0_g1_i1	32.3916	22.8703	3.6502	6.2655	3.7170e-10	1.9268e-7
DN3187_c1_g1_i1	25.0431	22.8635	3.6504	6.2634	3.7677e-10	1.9360e-7
DN20915_c0_g1_i2	21.7435	22.8471	3.6505	6.2587	3.8827e-10	1.9451e-7
DN11175_c0_g1_i3	122.2937	22.8130	3.6501	6.2500	4.1042e-10	1.9972e-7

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN43212_c0_g2_i1	51.5277	22.8118	3.6500	6.2498	4.1106e-10	1.9972e-7
DN3094_c0_g1_i1	37.8150	22.8111	3.6501	6.2494	4.1209e-10	1.9972e-7
DN11175_c0_g1_i1	98.6394	22.7869	3.6501	6.2428	4.2968e-10	2.0657e-7
DN7322_c0_g1_i8	41.8118	22.7615	3.6501	6.2358	4.4947e-10	2.0976e-7
DN1033_c0_g1_i18	94.4964	22.7606	3.6501	6.2356	4.4994e-10	2.0976e-7
DN5637_c0_g1_i8	27.7235	20.8972	3.3513	6.2355	4.5026e-10	2.0976e-7
DN4522_c0_g1_i2	63.3065	22.7601	3.6501	6.2355	4.5040e-10	2.0976e-7
DN1494_c0_g1_i3	40.4640	21.5435	3.4616	6.2235	4.8609e-10	2.2290e-7
DN24085_c1_g1_i4	84.1524	22.7033	3.6501	6.2199	4.9750e-10	2.2639e-7
DN162_c0_g2_i2	30.7529	21.4477	3.4578	6.2026	5.5536e-10	2.5081e-7
DN23733_c1_g1_i6	67.6429	22.6332	3.6501	6.2007	5.6229e-10	2.5203e-7
DN126918_c0_g3_i3	24.0190	22.6194	3.6504	6.1964	5.7756e-10	2.5694e-7
DN24932_c0_g2_i2	46.0045	22.6120	3.6502	6.1948	5.8356e-10	2.5769e-7
DN37266_c1_g1_i7	47.6006	22.5963	3.6502	6.1905	5.9981e-10	2.6291e-7
DN41506_c0_g2_i1	32.4196	22.5819	3.6502	6.1865	6.1530e-10	2.6774e-7
DN11402_c1_g1_i1	22.5566	22.5570	3.6504	6.1793	6.4399e-10	2.7819e-7
DN64488_c0_g1_i1	20.7810	22.5334	3.6505	6.1727	6.7143e-10	2.8743e-7
DN312446_c0_g1_i1	18.2976	22.5310	3.6506	6.1718	6.7502e-10	2.8743e-7
DN309908_c0_g1_i1	25.9235	22.5231	3.6503	6.1701	6.8237e-10	2.8850e-7
DN237732_c0_g1_i1	24.2295	22.5179	3.6504	6.1686	6.8883e-10	2.8918e-7
DN3599_c0_g2_i6	32.3129	22.4966	3.6502	6.1631	7.1348e-10	2.9743e-7
DN275884_c0_g1_i1	23.1427	22.4923	3.6504	6.1616	7.2030e-10	2.9819e-7
DN3387_c2_g1_i4	162.6840	22.4849	3.6502	6.1599	7.2769e-10	2.9917e-7
DN1613_c0_g1_i19	19.2024	22.4610	3.6506	6.1527	7.6162e-10	3.0886e-7
DN62786_c0_g2_i1	17.1692	22.4293	3.6507	6.1439	8.0528e-10	3.2436e-7
DN7204_c0_g1_i3	99.0022	22.3946	3.6502	6.1351	8.5087e-10	3.4042e-7

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN32533_c0_g2_i2	19.8455	22.3878	3.6505	6.1327	8.6380e-10	3.4168e-7
DN4913_c0_g3_i2	18.5798	22.3870	3.6506	6.1324	8.6548e-10	3.4168e-7
DN13606_c0_g2_i4	17.6937	22.3663	3.6506	6.1267	8.9743e-10	3.5196e-7
DN63241_c0_g1_i1	18.3266	22.3523	3.6506	6.1229	9.1901e-10	3.5807e-7
DN3426_c0_g1_i1	17.7913	22.3431	3.6506	6.1203	9.3380e-10	3.6147e-7
DN9026_c0_g1_i2	21.3512	22.3379	3.6505	6.1192	9.4074e-10	3.6181e-7
DN4002_c0_g1_i5	59.1546	21.7644	3.5592	6.1150	9.6533e-10	3.6574e-7
DN1645_c0_g1_i18	18.8643	22.3211	3.6506	6.1144	9.6936e-10	3.6574e-7
DN98643_c0_g1_i1	18.4317	22.3012	3.6506	6.1089	1.0032e-9	3.7613e-7
DN8502_c0_g1_i2	37.3606	21.5461	3.5280	6.1072	1.0139e-9	3.7774e-7
DN8664_c0_g1_i1	34.3696	22.2130	3.6503	6.0852	1.1633e-9	4.2412e-7
DN17110_c0_g1_i1	34.9771	22.2076	3.6503	6.0837	1.1741e-9	4.2412e-7
DN636_c0_g1_i3	33.7217	22.2066	3.6503	6.0834	1.1762e-9	4.2412e-7
DN8233_c1_g1_i7	32.0546	22.2042	3.6503	6.0828	1.1810e-9	4.2412e-7
DN15308_c0_g2_i1	18.2053	22.1994	3.6506	6.0810	1.1944e-9	4.2636e-7
DN6355_c1_g1_i1	15.3779	22.1876	3.6508	6.0775	1.2208e-9	4.3064e-7
DN6722_c0_g1_i1	15.8353	22.1634	3.6508	6.0709	1.2720e-9	4.4344e-7
DN6506_c0_g3_i2	15.7932	22.1517	3.6508	6.0677	1.2976e-9	4.4838e-7
DN18418_c0_g1_i3	41.6934	22.1475	3.6503	6.0672	1.3012e-9	4.4838e-7
DN3641_c0_g1_i1	16.7495	22.1342	3.6507	6.0630	1.3361e-9	4.5774e-7
DN237269_c0_g1_i1	16.0491	22.1310	3.6507	6.0621	1.3439e-9	4.5780e-7
DN4354_c0_g1_i1	15.5205	22.1251	3.6508	6.0604	1.3581e-9	4.6000e-7
DN5148_c0_g2_i1	29.4680	22.0998	3.6504	6.0541	1.4120e-9	4.7288e-7
DN164103_c0_g1_i1	14.8870	22.0952	3.6508	6.0521	1.4296e-9	4.7348e-7
DN9241_c0_g1_i1	84.4869	22.0815	3.6504	6.0492	1.4561e-9	4.7957e-7
DN65457_c0_g1_i3	70.7309	22.0663	3.6504	6.0450	1.4942e-9	4.8716e-7

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN4239_c0_g1_i2	13.2133	22.0696	3.6510	6.0449	1.4955e-9	4.8716e-7
DN39722_c0_g1_i2	54.8075	22.0497	3.6504	6.0404	1.5374e-9	4.9539e-7
DN528_c3_g1_i4	35.2320	22.0274	3.6504	6.0342	1.5972e-9	5.1189e-7
DN13088_c0_g2_i3	28.5962	22.0229	3.6504	6.0330	1.6096e-9	5.1312e-7
DN14547_c0_g2_i8	42.0624	22.0142	3.6504	6.0306	1.6332e-9	5.1786e-7
DN80940_c0_g2_i3	23.3767	22.0052	3.6505	6.0281	1.6593e-9	5.2336e-7
DN316_c0_g1_i2	42.0723	21.9846	3.6504	6.0225	1.7176e-9	5.3328e-7
DN9109_c3_g1_i13	29.8216	21.9779	3.6504	6.0206	1.7375e-9	5.3667e-7
DN540_c0_g1_i26	41.9424	21.9689	3.6503	6.0184	1.7615e-9	5.3930e-7
DN25229_c0_g1_i17	14.2557	21.9716	3.6509	6.0182	1.7641e-9	5.3930e-7
DN3538_c0_g1_i9	32.2378	21.9642	3.6504	6.0170	1.7768e-9	5.4042e-7
DN1887_c1_g1_i1	24.7806	21.9594	3.6505	6.0155	1.7932e-9	5.4263e-7
DN89668_c1_g1_i1	23.0405	21.9344	3.6505	6.0086	1.8709e-9	5.5890e-7
DN7538_c0_g1_i8	38.3479	21.8949	3.6505	5.9978	1.9996e-9	5.9011e-7
DN11857_c0_g1_i1	19.0836	21.8666	3.6506	5.9899	2.0997e-9	6.1659e-7
DN35427_c0_g3_i1	12.8609	21.8606	3.6510	5.9876	2.1301e-9	6.2246e-7
DN10244_c0_g1_i47	28.6537	21.8299	3.6505	5.9800	2.2320e-9	6.4905e-7
DN22311_c0_g1_i3	21.7369	21.7796	3.6506	5.9661	2.4303e-9	7.0329e-7
DN3263_c0_g1_i10	28.8214	21.7460	3.6505	5.9569	2.5702e-9	7.3310e-7
DN3234_c0_g2_i4	46.0984	21.7162	3.6506	5.9487	2.7022e-9	7.6708e-7
DN3419_c0_g1_i9	34.3665	21.6625	3.6506	5.9339	2.9575e-9	8.1622e-7
DN5329_c0_g2_i4	25.0926	21.6187	3.6507	5.9219	3.1832e-9	8.6711e-7
DN48348_c0_g1_i1	11.0446	21.6216	3.6512	5.9217	3.1855e-9	8.6711e-7
DN4852_c0_g1_i8	25.4580	21.6071	3.6506	5.9187	3.2447e-9	8.7853e-7
DN13860_c0_g2_i1	27.3891	21.6049	3.6507	5.9181	3.2569e-9	8.7853e-7
DN12250_c0_g1_i1	24.9946	21.5842	3.6507	5.9124	3.3720e-9	9.0141e-7

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN15919_c0_g2_i1	15.3615	21.5642	3.6508	5.9067	3.4898e-9	9.2525e-7
DN2264_c0_g1_i2	20.0503	21.5633	3.6507	5.9066	3.4922e-9	9.2525e-7
DN17405_c0_g5_i1	37.2839	21.5580	3.6507	5.9052	3.5219e-9	9.2900e-7
DN7359_c1_g1_i2	31.6638	21.5542	3.6506	5.9043	3.5418e-9	9.3013e-7
DN6961_c0_g2_i1	24.2572	21.5340	3.6507	5.8986	3.6663e-9	9.5070e-7
DN709_c0_g1_i10	25.6403	21.5337	3.6507	5.8985	3.6680e-9	9.5070e-7
DN120_c0_g1_i2	16.0538	21.5236	3.6508	5.8956	3.7331e-9	9.6337e-7
DN1914_c0_g2_i1	16.4496	21.4813	3.6508	5.8840	4.0051e-9	1.0291e-6
DN180_c0_g1_i32	39.2314	21.4604	3.6507	5.8784	4.1433e-9	1.0581e-6
DN19008_c1_g1_i5	26.7493	21.4522	3.6508	5.8761	4.2009e-9	1.0657e-6
DN130646_c0_g1_i2	26.5940	21.4423	3.6507	5.8734	4.2693e-9	1.0784e-6
DN4811_c0_g2_i1	35.4459	21.4205	3.6508	5.8674	4.4272e-9	1.1136e-6
DN1144_c0_g1_i9	50.7250	21.3771	3.6508	5.8555	4.7563e-9	1.1863e-6
DN67185_c0_g2_i2	18.1065	21.3733	3.6508	5.8543	4.7889e-9	1.1895e-6
DN93760_c0_g1_i16	35.5316	21.3618	3.6508	5.8512	4.8794e-9	1.2069e-6
DN4996_c1_g1_i2	37.9026	21.3508	3.6501	5.8493	4.9364e-9	1.2160e-6
DN8629_c0_g1_i13	18.6232	21.3451	3.6509	5.8466	5.0184e-9	1.2311e-6
DN12997_c0_g1_i2	43.9504	21.3390	3.6508	5.8450	5.0663e-9	1.2378e-6
DN8355_c0_g1_i4	15.6417	21.3211	3.6509	5.8399	5.2231e-9	1.2709e-6
DN739_c0_g1_i11	25.9722	20.9730	3.5923	5.8382	5.2757e-9	1.2785e-6
DN1417_c0_g2_i4	23.8740	21.2983	3.6508	5.8340	5.4129e-9	1.3064e-6
DN3636_c0_g1_i1	176.7324	10.6563	1.8270	5.8327	5.4523e-9	1.3106e-6
DN30971_c0_g2_i2	25.1065	21.2729	3.6509	5.8267	5.6521e-9	1.3532e-6
DN2556_c0_g1_i14	19.3524	21.2614	3.6509	5.8236	5.7606e-9	1.3736e-6
DN3524_c0_g1_i4	31.7556	21.2534	3.6508	5.8215	5.8312e-9	1.3849e-6
DN33769_c0_g3_i6	46.2634	21.1895	3.6510	5.8038	6.4820e-9	1.5213e-6

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN2265_c0_g1_i8	26.0762	21.1174	3.6510	5.7839	7.2974e-9	1.6941e-6
DN9534_c0_g1_i10	18.8974	21.1170	3.6511	5.7838	7.3036e-9	1.6941e-6
DN2440_c0_g1_i3	25.4141	21.1075	3.6511	5.7812	7.4164e-9	1.7136e-6
DN63381_c0_g3_i4	34.6603	21.0953	3.6511	5.7779	7.5649e-9	1.7412e-6
DN90904_c0_g2_i2	18.1313	21.0859	3.6511	5.7753	7.6836e-9	1.7617e-6
DN1253_c0_g1_i3	61.4681	21.0784	3.6510	5.7732	7.7758e-9	1.7760e-6
DN4875_c0_g2_i2	22.2342	21.0373	3.6511	5.7619	8.3189e-9	1.8928e-6
DN143028_c1_g1_i2	19.6834	21.0072	3.6512	5.7536	8.7384e-9	1.9734e-6
DN103_c1_g1_i8	22.7963	21.0071	3.6512	5.7535	8.7393e-9	1.9734e-6
DN4314_c0_g1_i4	18.4969	20.9745	3.6512	5.7445	9.2176e-9	2.0736e-6
DN11172_c0_g2_i2	17.7856	20.9506	3.6512	5.7379	9.5844e-9	2.1479e-6
DN4536_c0_g1_i1	40.8617	20.9381	3.6512	5.7346	9.7759e-9	2.1827e-6
DN5935_c0_g1_i3	20.3324	20.8913	3.6513	5.7216	1.0552e-8	2.3472e-6
DN51273_c0_g1_i90	16.0941	20.8820	3.6513	5.7190	1.0715e-8	2.3745e-6
DN2753_c0_g1_i4	18.5634	20.8606	3.6513	5.7132	1.1089e-8	2.4483e-6
DN20997_c0_g1_i2	17.3003	20.8504	3.6514	5.7103	1.1275e-8	2.4739e-6
DN2424_c0_g1_i5	18.1196	20.8495	3.6513	5.7101	1.1292e-8	2.4739e-6
DN11123_c0_g1_i4	18.0290	20.8464	3.6514	5.7092	1.1350e-8	2.4739e-6
DN40014_c0_g1_i5	19.6770	20.8452	3.6513	5.7089	1.1371e-8	2.4739e-6
DN4827_c0_g1_i2	21.3103	20.8004	3.6514	5.6966	1.2225e-8	2.6500e-6
DN2119_c0_g2_i9	29.6967	20.7841	3.6511	5.6926	1.2513e-8	2.7026e-6
DN726_c0_g1_i21	18.1546	20.7033	3.6514	5.6699	1.4286e-8	3.0745e-6
DN111_c0_g1_i11	15.8915	20.7016	3.6515	5.6693	1.4340e-8	3.0750e-6
DN216_c0_g1_i16	29.2183	20.6975	3.6515	5.6682	1.4429e-8	3.0829e-6
DN3406_c1_g2_i12	11.8767	20.6352	3.6517	5.6509	1.5961e-8	3.3981e-6
DN4956_c1_g1_i21	13.5943	20.5998	3.6517	5.6412	1.6891e-8	3.5706e-6

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN12231_c0_g1_i7	13.0672	20.5959	3.6517	5.6401	1.6998e-8	3.5806e-6
DN2916_c0_g1_i6	17.3323	20.5881	3.6517	5.6379	1.7209e-8	3.6123e-6
DN3649_c0_g1_i8	14.2134	20.4935	3.6518	5.6118	2.0020e-8	4.1876e-6
DN593_c0_g1_i3	26.6860	20.4443	3.6519	5.5983	2.1643e-8	4.5113e-6
DN21703_c0_g3_i5	27.0697	20.4401	3.6516	5.5977	2.1727e-8	4.5130e-6
DN29795_c0_g2_i3	65.8157	20.4211	3.6518	5.5920	2.2446e-8	4.6369e-6
DN147168_c0_g4_i1	9.2901	20.4212	3.6520	5.5917	2.2479e-8	4.6369e-6
DN1109_c0_g1_i8	23.9566	20.4005	3.6519	5.5862	2.3206e-8	4.7703e-6
DN15070_c0_g1_i2	12.4142	20.3963	3.6520	5.5850	2.3374e-8	4.7882e-6
DN3414_c0_g1_i6	13.8277	20.3901	3.6520	5.5832	2.3608e-8	4.8197e-6
DN3447_c1_g2_i3	22.7582	20.3805	3.6520	5.5807	2.3957e-8	4.8743e-6
DN16129_c0_g2_i2	31.0018	20.3587	3.6520	5.5747	2.4801e-8	5.0287e-6
DN39253_c0_g1_i10	13.9876	20.3426	3.6521	5.5702	2.5451e-8	5.1324e-6
DN53950_c0_g1_i3	13.7181	20.3419	3.6521	5.5699	2.5484e-8	5.1324e-6
DN4310_c0_g2_i1	16.8885	20.3325	3.6521	5.5674	2.5861e-8	5.1907e-6
DN8893_c0_g1_i2	18.7911	20.3101	3.6521	5.5612	2.6792e-8	5.3546e-6
DN929_c0_g1_i3	24.8599	20.3082	3.6520	5.5608	2.6857e-8	5.3546e-6
DN1427_c0_g2_i1	19.1084	20.2448	3.6522	5.5432	2.9706e-8	5.9028e-6
DN9095_c0_g1_i1	10.6377	20.2421	3.6523	5.5423	2.9855e-8	5.9127e-6
DN126606_c0_g1_i2	21.7514	20.2252	3.6522	5.5378	3.0638e-8	6.0477e-6
DN13201_c0_g1_i3	10.2223	20.2199	3.6523	5.5362	3.0914e-8	6.0785e-6
DN2134_c0_g1_i10	26.2840	20.2176	3.6522	5.5357	3.0998e-8	6.0785e-6
DN76088_c1_g1_i14	17.7760	20.2116	3.6523	5.5340	3.1306e-8	6.1188e-6
DN568_c0_g1_i16	14.1813	20.1729	3.6524	5.5233	3.3276e-8	6.4827e-6
DN6553_c0_g1_i7	32.0718	20.1260	3.6524	5.5104	3.5811e-8	6.9537e-6
DN75172_c0_g1_i2	14.0304	20.0982	3.6525	5.5026	3.7434e-8	7.2452e-6

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN4754_c1_g2_i5	12.8953	20.0929	3.6525	5.5011	3.7748e-8	7.2825e-6
DN384_c0_g1_i10	8.0241	20.0840	3.6526	5.4985	3.8309e-8	7.3668e-6
DN736_c0_g2_i1	35.9580	20.0614	3.6525	5.4925	3.9633e-8	7.5969e-6
DN17161_c0_g2_i2	45.9235	20.0308	3.6525	5.4841	4.1567e-8	7.9405e-6
DN4635_c0_g1_i10	10.0644	20.0289	3.6527	5.4833	4.1748e-8	7.9405e-6
DN89110_c1_g2_i2	29.1207	20.0263	3.6525	5.4830	4.1825e-8	7.9405e-6
DN3993_c1_g1_i11	14.4143	20.0104	3.6526	5.4783	4.2932e-8	8.1247e-6
DN28734_c0_g1_i5	22.3985	20.0001	3.6527	5.4755	4.3634e-8	8.2316e-6
DN3204_c0_g1_i1	8.1103	19.9972	3.6528	5.4745	4.3886e-8	8.2529e-6
DN9798_c0_g1_i14	7.7879	19.9797	3.6529	5.4696	4.5103e-8	8.4551e-6
DN453_c0_g1_i25	11.4747	19.9691	3.6528	5.4668	4.5834e-8	8.5653e-6
DN14653_c0_g1_i6	10.8366	19.9298	3.6525	5.4565	4.8556e-8	9.0454e-6
DN98996_c0_g1_i10	7.2365	19.8934	3.6531	5.4457	5.1611e-8	9.5757e-6
DN26109_c0_g1_i6	7.7630	19.8857	3.6531	5.4436	5.2214e-8	9.6367e-6
DN5674_c0_g1_i5	18.6479	19.8742	3.6530	5.4406	5.3113e-8	9.7723e-6
DN10277_c1_g1_i11	15.1319	19.8671	3.6530	5.4386	5.3702e-8	9.8503e-6
DN2564_c0_g1_i14	6.9513	19.8643	3.6531	5.4376	5.4002e-8	9.8750e-6
DN9082_c1_g2_i4	9.2540	19.8419	3.6531	5.4315	5.5872e-8	1.0186e-5
DN384_c0_g1_i11	8.9278	19.8244	3.6532	5.4266	5.7452e-8	1.0442e-5
DN668_c0_g1_i6	35.7114	19.8175	3.6530	5.4249	5.7975e-8	1.0505e-5
DN3618_c1_g1_i9	11.2258	19.7286	3.6534	5.4001	6.6619e-8	1.2034e-5
DN3867_c2_g1_i6	7.0941	19.7064	3.6535	5.3939	6.8959e-8	1.2420e-5
DN19036_c0_g2_i3	11.1795	19.7008	3.6535	5.3924	6.9541e-8	1.2487e-5
DN119225_c0_g1_i8	10.9705	19.6914	3.6535	5.3898	7.0556e-8	1.2631e-5
DN373_c0_g1_i13	11.2515	19.6788	3.6535	5.3863	7.1928e-8	1.2838e-5
DN5714_c0_g1_i2	18.4072	19.6625	3.6535	5.3818	7.3731e-8	1.3113e-5

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN63614_c0_g3_i1	10.2346	19.6614	3.6536	5.3814	7.3908e-8	1.3113e-5
DN1196_c1_g1_i27	16.0030	19.6268	3.6536	5.3719	7.7893e-8	1.3779e-5
DN35694_c0_g2_i4	14.9500	19.5723	3.6538	5.3568	8.4730e-8	1.4944e-5
DN8261_c0_g1_i9	10.5114	19.5535	3.6529	5.3529	8.6541e-8	1.5218e-5
DN385_c0_g1_i5	10.1711	19.5199	3.6535	5.3428	9.1516e-8	1.6046e-5
DN9257_c0_g1_i13	8.1935	19.4846	3.6531	5.3337	9.6214e-8	1.6820e-5
DN851_c0_g1_i2	10.4486	19.4460	3.6542	5.3215	1.0289e-7	1.7935e-5
DN152_c0_g1_i1	7.7939	19.3791	3.6525	5.3056	1.1229e-7	1.9459e-5
DN8180_c0_g1_i11	10.1968	19.3645	3.6545	5.2988	1.1657e-7	2.0142e-5
DN2911_c0_g1_i2	10.1689	19.2833	3.6548	5.2762	1.3189e-7	2.2723e-5
DN39985_c0_g1_i9	8.6620	19.2271	3.6550	5.2605	1.4368e-7	2.4683e-5
DN228_c1_g1_i19	10.1577	19.1982	3.6550	5.2525	1.5003e-7	2.5700e-5
DN14197_c0_g1_i3	48.5930	19.1937	3.6548	5.2516	1.5080e-7	2.5759e-5
DN52594_c0_g2_i2	23.2092	19.0252	3.6556	5.2044	1.9457e-7	3.3140e-5
DN55238_c0_g1_i7	8.5469	18.9000	3.6564	5.1691	2.3524e-7	3.9953e-5
DN229_c1_g1_i1	11.3850	18.8773	3.6564	5.1628	2.4323e-7	4.1192e-5
DN14609_c0_g2_i4	8.7502	18.8747	3.6565	5.1619	2.4441e-7	4.1275e-5
DN5260_c0_g3_i4	7.1224	18.8058	3.6569	5.1426	2.7098e-7	4.5633e-5
DN11567_c0_g1_i1	7.9849	18.7299	3.6564	5.1225	3.0155e-7	5.0637e-5
DN4631_c0_g3_i1	7.3771	18.5039	3.6586	5.0577	4.2438e-7	7.1063e-5
DN5339_c0_g2_i2	94.9650	9.9992	1.9790	5.0526	4.3587e-7	7.2783e-5
DN18352_c0_g2_i7	8.2388	18.4361	3.6590	5.0386	4.6898e-7	7.8094e-5
DN1109_c0_g1_i12	168.8378	11.4490	2.2736	5.0356	4.7632e-7	7.9094e-5
DN9057_c2_g1_i1	20.0886	18.3778	3.6591	5.0225	5.0999e-7	8.4450e-5
DN3067_c0_g1_i1	11.6381	18.3413	3.6592	5.0124	5.3761e-7	8.8776e-5
DN3064_c0_g2_i2	21.4167	18.3279	3.6594	5.0084	5.4877e-7	9.0369e-5

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN1721_c0_g1_i16	19.0539	18.0400	3.6616	4.9267	8.3618e-7	1.3732e-4
DN21069_c0_g2_i1	17.5820	17.9885	3.6622	4.9119	9.0183e-7	1.4769e-4
DN65268_c1_g1_i3	28.9969	17.8214	3.6635	4.8645	1.1472e-6	1.8737e-4
DN454_c0_g1_i2	29.3353	17.4212	3.6676	4.7500	2.0344e-6	3.3136e-4
DN1259_c0_g1_i2	11.6692	17.2689	3.6699	4.7055	2.5323e-6	4.1133e-4
DN403_c0_g1_i11	12.6805	7.2657	1.6208	4.4827	7.3692e-6	1.1905e-3
DN15717_c0_g1_i5	102.2993	9.9381	2.2216	4.4733	7.7004e-6	1.2407e-3
DN5122_c0_g1_i1	133.6010	10.0529	2.2477	4.4726	7.7288e-6	1.2419e-3
DN647_c0_g1_i24	23.0689	7.7099	1.7457	4.4166	1.0026e-5	1.6067e-3
DN1370_c1_g1_i8	79.6245	9.6786	2.2136	4.3724	1.2288e-5	1.9638e-3
DN1600_c0_g1_i4	24.1392	7.7735	1.8183	4.2752	1.9094e-5	3.0354e-3
DN24194_c0_g2_i6	25.5408	7.5413	1.7985	4.1931	2.7517e-5	4.3628e-3
DN4880_c1_g1_i2	34.4189	8.9463	2.1433	4.1741	2.9911e-5	4.7296e-3
DN7993_c0_g1_i3	37.5680	8.5444	2.0640	4.1397	3.4782e-5	5.4854e-3
DN5093_c0_g1_i3	62.4104	15.4661	3.7497	4.1247	3.7129e-5	5.8246e-3
DN307_c0_g1_i12	25.0442	8.0364	1.9638	4.0924	4.2701e-5	6.6637e-3
DN84236_c0_g1_i3	35.9935	9.2192	2.2637	4.0727	4.6472e-5	7.2229e-3
DN4608_c0_g4_i2	36.5841	8.7431	2.1469	4.0724	4.6526e-5	7.2229e-3
DN3931_c0_g1_i4	55.6306	8.8540	2.1790	4.0632	4.8397e-5	7.4937e-3
DN25859_c0_g1_i2	29.5509	5.9089	1.4633	4.0379	5.3923e-5	8.3141e-3
DN2088_c0_g1_i33	41.6764	8.8360	2.1884	4.0377	5.3974e-5	8.3141e-3
DN79840_c1_g1_i1	33.5443	8.5621	2.1264	4.0266	5.6597e-5	8.6956e-3
DN14852_c0_g1_i2	9.4686	6.6622	1.6807	3.9640	7.3689e-5	1.1264e-2
DN29672_c1_g5_i1	59.1149	7.4683	1.9081	3.9140	9.0788e-5	1.3807e-2
DN4890_c0_g4_i1	16.8512	6.4221	1.6483	3.8962	9.7703e-5	1.4795e-2
DN73720_c0_g2_i6	18.7815	14.6512	3.7620	3.8945	9.8396e-5	1.4850e-2

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN4956_c1_g2_i2	13.4095	7.5069	1.9310	3.8876	1.0126e-4	1.5244e-2
DN67795_c1_g1_i8	28.8458	8.8998	2.2917	3.8834	1.0301e-4	1.5468e-2
DN49200_c0_g1_i4	15.0921	7.4807	1.9314	3.8732	1.0742e-4	1.6089e-2
DN24946_c0_g2_i1	53.7761	8.4501	2.1969	3.8464	1.1985e-4	1.7907e-2
DN3684_c0_g2_i1	15.7555	7.5293	1.9596	3.8422	1.2195e-4	1.8174e-2
DN653_c0_g1_i8	52.7790	9.4409	2.4695	3.8230	1.3185e-4	1.9552e-2
DN70366_c0_g2_i8	24.4700	7.4745	1.9561	3.8212	1.3282e-4	1.9648e-2
DN5440_c1_g1_i8	16.6787	7.8665	2.0776	3.7864	1.5285e-4	2.2499e-2
DN1538_c0_g1_i16	11.9663	6.6637	1.7677	3.7698	1.6340e-4	2.3992e-2
DN6486_c1_g1_i10	28.8813	8.0750	2.1439	3.7665	1.6553e-4	2.4245e-2
DN2459_c0_g1_i15	17.6818	7.4768	1.9940	3.7496	1.7710e-4	2.5876e-2
DN63897_c0_g1_i19	12.9389	7.7433	2.0670	3.7462	1.7952e-4	2.6165e-2
DN2980_c0_g1_i3	41.6526	8.1831	2.1855	3.7442	1.8096e-4	2.6312e-2
DN2495_c0_g2_i1	26.8944	8.1634	2.1821	3.7410	1.8329e-4	2.6585e-2
DN5952_c2_g1_i9	21.9337	7.3295	1.9686	3.7232	1.9673e-4	2.8466e-2
DN90904_c0_g2_i6	124.9256	9.1650	2.4663	3.7161	2.0233e-4	2.9134e-2
DN3358_c0_g2_i2	32.0616	8.6662	2.3441	3.6970	2.1813e-4	3.1333e-2
DN30462_c0_g1_i8	8.8930	6.4871	1.7682	3.6687	2.4378e-4	3.4934e-2
DN1585_c0_g1_i11	63.3119	9.0735	2.4751	3.6658	2.4652e-4	3.5242e-2
DN10179_c0_g1_i6	22.7387	7.9632	2.1741	3.6628	2.4945e-4	3.5576e-2
DN3060_c1_g1_i13	18.3708	7.3791	2.0235	3.6466	2.6568e-4	3.7800e-2
DN12936_c0_g2_i3	17.2326	7.1763	1.9755	3.6327	2.8045e-4	3.9691e-2
DN350_c0_g1_i19	21.3815	7.8077	2.1496	3.6322	2.8097e-4	3.9691e-2
DN4248_c0_g1_i15	17.8009	8.2033	2.2829	3.5934	3.2640e-4	4.6000e-2
DN52991_c0_g2_i3	20.7355	7.6804	2.1382	3.5920	3.2814e-4	4.6026e-2
DN3272_c0_g1_i2	41.8877	8.5093	2.3737	3.5849	3.3722e-4	4.6995e-2

Table A.4: significant up-regulated genes, as identified by DESeq2

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN29080_c0_g1_i7	24.3776	7.7422	2.1598	3.5847	3.3740e-4	4.6995e-2
DN30602_c0_g1_i11	27.2355	8.3685	2.3378	3.5796	3.4409e-4	4.7814e-2
DN6201_c0_g1_i8	22.8028	8.2736	2.3146	3.5745	3.5092e-4	4.8553e-2
DN1174_c0_g1_i11	24.3999	7.6542	2.1414	3.5744	3.5104e-4	4.8553e-2

Table A.5: significant down-regulated genes, as identified by DESeq2.

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN4310_c0_g2_i5	99.0393	-24.0131	2.4542	-9.7846	1.3105e-22	7.8124e-18
DN13559_c0_g2_i3	363.4516	-26.3513	3.6498	-7.2200	5.2005e-13	2.3848e-9
DN103_c1_g1_i2	157.3805	-26.1064	3.6498	-7.1528	8.5021e-13	3.6203e-9
DN1885_c0_g2_i6	149.6177	-25.9985	3.6498	-7.1232	1.0543e-12	4.1901e-9
DN6967_c0_g1_i18	25.2217	-23.2046	3.2988	-7.0342	2.0043e-12	7.4678e-9
DN10247_c2_g1_i17	99.9742	-25.4327	3.6499	-6.9681	3.2128e-12	1.0640e-8
DN1135_c0_g1_i18	105.5994	-25.3730	3.6499	-6.9517	3.6090e-12	1.1323e-8
DN829_c0_g1_i4	79.5884	-25.1033	3.6499	-6.8777	6.0811e-12	1.8018e-8
DN33769_c0_g3_i5	124.9191	-25.0810	3.6499	-6.8716	6.3474e-12	1.8018e-8
DN3165_c0_g1_i5	40.5046	-23.2307	3.4183	-6.7960	1.0757e-11	2.4673e-8
DN1011_c0_g1_i2	65.4881	-24.6432	3.6500	-6.7515	1.4633e-11	3.0814e-8
DN3227_c0_g2_i2	49.5807	-24.6127	3.6501	-6.7431	1.5507e-11	3.0814e-8
DN2749_c0_g1_i1	47.1774	-24.4574	3.6501	-6.7005	2.0773e-11	3.8698e-8
DN27965_c0_g1_i1	16.8519	-22.0666	3.3051	-6.6766	2.4458e-11	4.3398e-8
DN2450_c0_g1_i2	44.7695	-24.3441	3.6501	-6.6694	2.5683e-11	4.3398e-8
DN21161_c1_g1_i1	50.3645	-24.3404	3.6501	-6.6685	2.5847e-11	4.3398e-8
DN16320_c0_g1_i5	55.6867	-23.3973	3.5118	-6.6624	2.6936e-11	4.3398e-8
DN3227_c0_g2_i3	48.9641	-24.3005	3.6501	-6.6575	2.7845e-11	4.3682e-8
DN2582_c1_g1_i10	35.3628	-23.8714	3.6503	-6.5397	6.1661e-11	8.1805e-8
DN23733_c1_g1_i10	31.0168	-23.8506	3.6503	-6.5339	6.4090e-11	8.3056e-8
DN340_c0_g1_i1	28.8484	-23.7806	3.6503	-6.5146	7.2880e-11	8.8665e-8
DN25055_c0_g3_i7	26.1803	-23.7138	3.6504	-6.4962	8.2373e-11	9.6285e-8
DN840_c0_g1_i11	70.4053	-23.6664	3.6504	-6.4833	8.9757e-11	1.0096e-7
DN261_c0_g1_i6	68.7444	-23.6504	3.6504	-6.4789	9.2423e-11	1.0203e-7
DN1829_c0_g1_i6	36.9238	-23.6242	3.6504	-6.4717	9.6925e-11	1.0318e-7
DN1350_c0_g2_i7	52.9527	-23.4953	3.6504	-6.4364	1.2235e-10	1.1930e-7

Table A.5: significant down-regulated genes, as identified by DESeq2.

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN7183_c0_g2_i1	23.6701	-23.4836	3.6505	-6.4330	1.2511e-10	1.1930e-7
DN12342_c0_g3_i3	26.3154	-23.4493	3.6504	-6.4237	1.3297e-10	1.1930e-7
DN9431_c0_g2_i1	22.7800	-23.4430	3.6505	-6.4218	1.3465e-10	1.1930e-7
DN2142_c0_g1_i2	25.0399	-23.4370	3.6505	-6.4202	1.3609e-10	1.1930e-7
DN2676_c0_g1_i44	46.8542	-23.4204	3.6505	-6.4156	1.4024e-10	1.2116e-7
DN1434_c0_g2_i2	54.7457	-23.3862	3.6505	-6.4062	1.4917e-10	1.2703e-7
DN1713_c0_g3_i6	28.4399	-23.3684	3.6506	-6.4013	1.5404e-10	1.2933e-7
DN311812_c0_g1_i1	21.5933	-23.3579	3.6506	-6.3985	1.5696e-10	1.2996e-7
DN3179_c0_g1_i5	21.7019	-23.3225	3.6506	-6.3887	1.6730e-10	1.3478e-7
DN8466_c0_g1_i17	23.6859	-23.2720	3.6506	-6.3748	1.8319e-10	1.4420e-7
DN16760_c0_g1_i4	19.7829	-23.2702	3.6506	-6.3743	1.8384e-10	1.4420e-7
DN7462_c0_g2_i6	19.7452	-23.2538	3.6506	-6.3698	1.8929e-10	1.4614e-7
DN748_c0_g1_i3	19.4613	-23.2483	3.6507	-6.3682	1.9122e-10	1.4614e-7
DN9442_c0_g1_i2	28.0686	-23.1872	3.6507	-6.3515	2.1326e-10	1.6092e-7
DN13054_c0_g1_i2	18.8500	-23.1637	3.6507	-6.3450	2.2240e-10	1.6573e-7
DN6626_c0_g2_i5	23.9505	-23.1526	3.6505	-6.3423	2.2629e-10	1.6654e-7
DN310517_c0_g1_i1	23.3144	-23.1172	3.6505	-6.3326	2.4105e-10	1.7249e-7
DN5504_c1_g2_i3	20.0463	-23.1183	3.6507	-6.3325	2.4119e-10	1.7249e-7
DN18129_c0_g1_i5	19.6552	-23.1139	3.6507	-6.3313	2.4305e-10	1.7249e-7
DN840_c0_g1_i9	28.9788	-23.0977	3.6507	-6.3269	2.5013e-10	1.7393e-7
DN16112_c0_g3_i3	28.3580	-23.0853	3.6508	-6.3234	2.5583e-10	1.7393e-7
DN5540_c0_g1_i1	17.4128	-23.0824	3.6508	-6.3226	2.5719e-10	1.7393e-7
DN3860_c1_g1_i8	24.7051	-23.0324	3.6508	-6.3089	2.8110e-10	1.8214e-7
DN7105_c0_g2_i7	26.0656	-23.0117	3.6507	-6.3033	2.9133e-10	1.8367e-7
DN6641_c0_g1_i3	26.0431	-23.0095	3.6508	-6.3026	2.9270e-10	1.8367e-7
DN11112_c0_g2_i3	18.3163	-22.9942	3.6507	-6.2985	3.0048e-10	1.8659e-7

Table A.5: significant down-regulated genes, as identified by DESeq2.

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN36288_c0_g1_i2	16.2214	-22.9759	3.6509	-6.2933	3.1077e-10	1.9096e-7
DN1337_c0_g1_i7	25.4475	-22.9610	3.6508	-6.2893	3.1882e-10	1.9096e-7
DN12127_c0_g5_i1	15.8011	-22.9369	3.6509	-6.2826	3.3304e-10	1.9136e-7
DN3710_c0_g2_i4	33.9752	-22.9230	3.6509	-6.2787	3.4136e-10	1.9198e-7
DN2741_c0_g1_i8	15.3064	-22.9009	3.6509	-6.2726	3.5498e-10	1.9268e-7
DN29438_c1_g1_i2	16.3256	-22.8984	3.6508	-6.2721	3.5626e-10	1.9268e-7
DN2783_c1_g1_i7	24.9662	-22.8926	3.6509	-6.2704	3.6016e-10	1.9268e-7
DN1911_c0_g1_i15	16.2639	-22.8773	3.6508	-6.2663	3.6976e-10	1.9268e-7
DN2299_c0_g1_i2	42.0657	-22.8611	3.6507	-6.2620	3.7997e-10	1.9360e-7
DN23346_c0_g1_i2	15.6460	-22.8525	3.6509	-6.2594	3.8655e-10	1.9451e-7
DN82723_c1_g2_i9	14.7183	-22.8374	3.6510	-6.2552	3.9709e-10	1.9727e-7
DN21224_c0_g3_i1	15.2149	-22.7426	3.6510	-6.2291	4.6905e-10	2.1676e-7
DN85322_c0_g1_i1	13.5492	-22.4669	3.6511	-6.1535	7.5790e-10	3.0886e-7
DN6035_c0_g1_i5	18.6116	-22.3294	3.6515	-6.1151	9.6486e-10	3.6574e-7
DN8628_c0_g2_i6	35.2348	-22.2599	3.6516	-6.0959	1.0884e-9	4.0300e-7
DN12057_c0_g1_i7	22.0420	-22.2379	3.6517	-6.0898	1.1308e-9	4.1613e-7
DN7167_c2_g3_i4	14.1935	-22.1985	3.6517	-6.0789	1.2103e-9	4.2945e-7
DN6658_c0_g1_i3	20.0883	-22.1853	3.6516	-6.0755	1.2364e-9	4.3356e-7
DN4182_c0_g1_i11	55.0801	-22.1262	3.6518	-6.0589	1.3703e-9	4.6150e-7
DN4985_c0_g1_i6	24.4701	-22.1035	3.6519	-6.0526	1.4251e-9	4.7348e-7
DN119225_c0_g2_i6	14.1481	-22.0699	3.6519	-6.0435	1.5083e-9	4.8865e-7
DN3983_c0_g2_i1	15.3288	-22.0068	3.6521	-6.0259	1.6821e-9	5.2776e-7
DN7219_c0_g2_i3	16.4187	-21.9974	3.6521	-6.0233	1.7095e-9	5.3328e-7
DN5267_c0_g1_i1	28.4244	-21.9440	3.6519	-6.0089	1.8679e-9	5.5890e-7
DN1334_c0_g1_i15	20.1007	-21.9424	3.6520	-6.0083	1.8751e-9	5.5890e-7
DN15229_c0_g1_i1	11.1186	-21.9341	3.6522	-6.0058	1.9043e-9	5.6478e-7

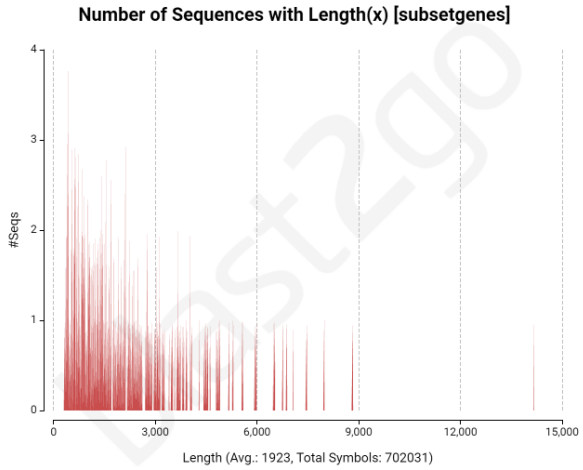
Table A.5: significant down-regulated genes, as identified by DESeq2.

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN32928_c0_g2_i2	11.0491	-21.7853	3.6525	-5.9646	2.4529e-9	7.0640e-7
DN4928_c0_g1_i3	23.6458	-21.7791	3.6525	-5.9628	2.4792e-9	7.1054e-7
DN126918_c0_g3_i1	36.7837	-21.7091	3.6507	-5.9465	2.7398e-9	7.7303e-7
DN2600_c0_g1_i1	19.2209	-21.7176	3.6525	-5.9459	2.7491e-9	7.7303e-7
DN3886_c0_g1_i4	10.0320	-21.7042	3.6527	-5.9420	2.8152e-9	7.8790e-7
DN5157_c0_g2_i3	7.9291	-21.6893	3.6526	-5.9380	2.8852e-9	8.0371e-7
DN1705_c0_g1_i4	16.5070	-21.6858	3.6526	-5.9371	2.9018e-9	8.0457e-7
DN7145_c0_g1_i2	20.1797	-21.6376	3.6527	-5.9237	3.1476e-9	8.6470e-7
DN5014_c0_g1_i8	17.4495	-21.6000	3.6528	-5.9132	3.3554e-9	9.0103e-7
DN371_c0_g1_i14	11.4397	-21.5558	3.6529	-5.9010	3.6132e-9	9.4472e-7
DN56204_c0_g1_i5	16.8437	-21.4735	3.6532	-5.8780	4.1535e-9	1.0581e-6
DN2522_c1_g1_i1	18.6974	-21.4184	3.6527	-5.8637	4.5268e-9	1.1338e-6
DN990_c1_g1_i3	11.9259	-21.2552	3.6530	-5.8186	5.9338e-9	1.4037e-6
DN5350_c0_g1_i1	12.0453	-21.2484	3.6532	-5.8164	6.0125e-9	1.4167e-6
DN8469_c0_g1_i3	8.2290	-21.1837	3.6540	-5.7974	6.7343e-9	1.5743e-6
DN5728_c0_g1_i1	11.5185	-20.6275	3.6562	-5.6418	1.6832e-8	3.5706e-6
DN5938_c0_g1_i3	36.5365	-19.9198	3.6582	-5.4453	5.1723e-8	9.5757e-6
DN11085_c0_g1_i2	8.0601	-19.4481	3.6591	-5.3150	1.0666e-7	1.8538e-5
DN1236_c1_g1_i4	36.7047	-7.9804	1.7208	-4.6376	3.5254e-6	5.7108e-4
DN1013_c0_g1_i7	37.4461	-8.5042	1.9725	-4.3114	1.6224e-5	2.5860e-3
DN990_c0_g2_i1	47.6874	-9.2779	2.2489	-4.1255	3.6998e-5	5.8194e-3
DN12408_c0_g1_i8	26.8555	-7.0274	1.7105	-4.1084	3.9845e-5	6.2344e-3
DN16599_c0_g3_i7	26.1170	-8.5759	2.1615	-3.9675	7.2639e-5	1.1132e-2
DN11115_c0_g2_i3	43.2807	-9.3639	2.3853	-3.9257	8.6480e-5	1.3185e-2
DN1246_c0_g2_i2	21.1382	-8.2844	2.1264	-3.8960	9.7782e-5	1.4795e-2
DN1571_c1_g1_i1	18.8957	-5.8314	1.5234	-3.8279	1.2922e-4	1.9210e-2

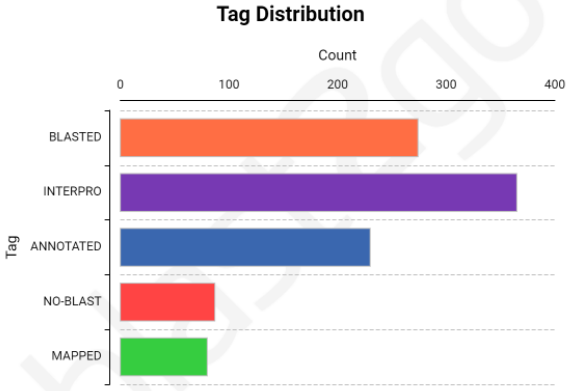
Table A.5: significant down-regulated genes, as identified by DESeq2.

	baseMean	LFC	lfcSE	stat	pvalue	padj
DN2112_c0_g1_i3	16.6323	-7.0472	1.8600	-3.7888	1.5140e-4	2.2340e-2
DN199443_c0_g1_i1	89.5835	-7.6328	2.0524	-3.7190	1.9999e-4	2.8866e-2
DN5209_c0_g1_i1	34.4800	-8.0726	2.2198	-3.6366	2.7623e-4	3.9207e-2
DN6136_c0_g1_i16	39.0467	-8.0206	2.2329	-3.5921	3.2805e-4	4.6026e-2
DN56144_c0_g2_i7	12.4254	-7.4085	2.0658	-3.5863	3.3541e-4	4.6936e-2
DN64328_c1_g1_i3	21.0986	-7.6983	2.1578	-3.5677	3.6019e-4	4.9703e-2

A.5 Blast2GO



(a) Blast2go sequence length distribution. Functional annotation was performed using only the differentially expressed transcripts, filtered according to thresholds (see 2.7).



(b) Blast2go data distribution represents the number of sequences that reach each stage of the annotation process. Not all blasted sequences received GO annotations. Functional annotation was performed using only the differentially expressed transcripts, filtered according to thresholds (see 2.7).

Figure A.15: Overview of the Blast2GO project statistics

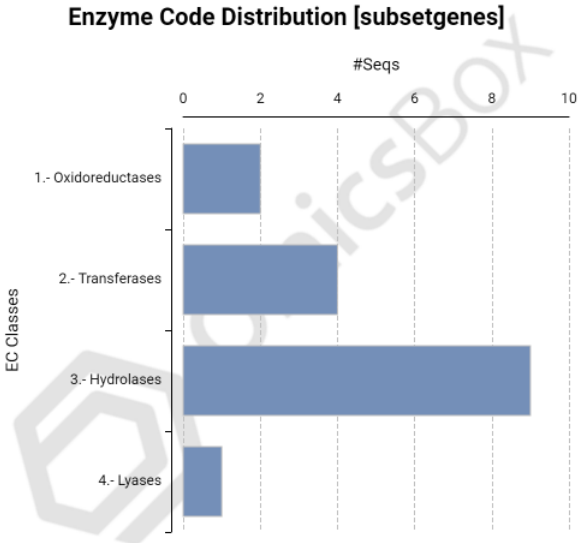
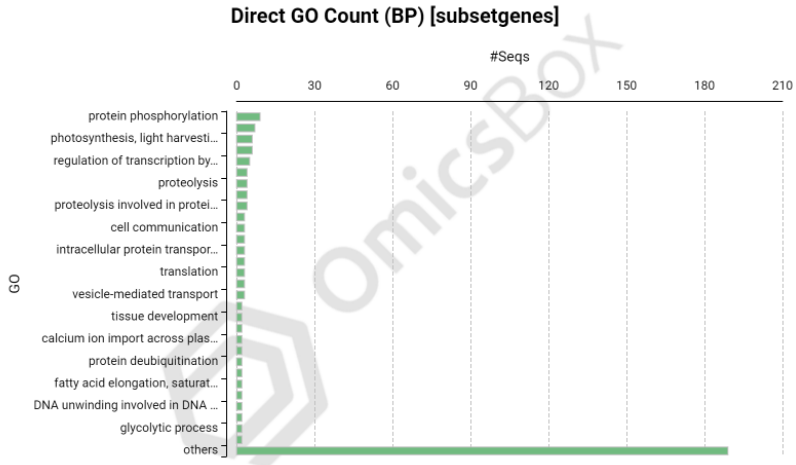
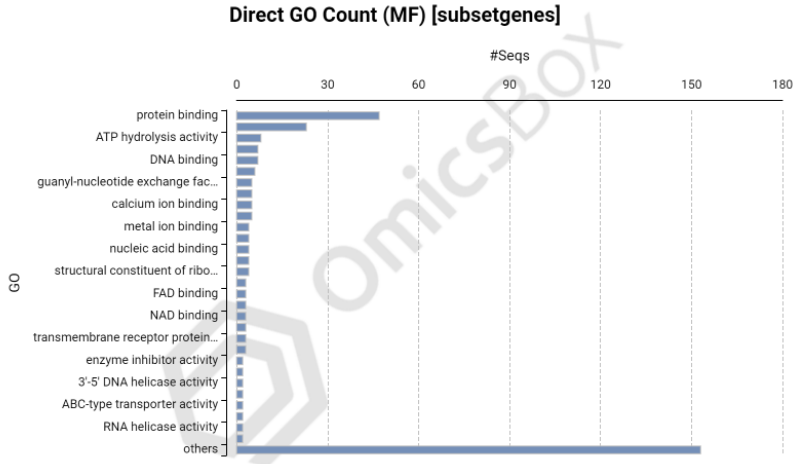


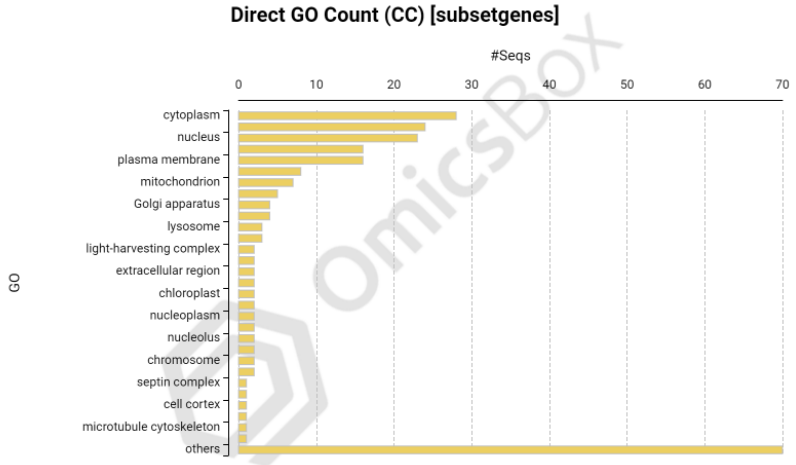
Figure A.16: Summary of the enzyme code distribution, according to the number of sequences annotated for each term. Functional annotation was performed using only the differentially expressed transcripts, filtered according to thresholds (see 2.7).



(a) biological processes



(b) molecular functions



(c) cellular components

Figure A.17: Direct GO count graphs, according to the number of sequences, for all annotated terms by category. Functional annotation was performed using only the differentially expressed transcripts, filtered according to thresholds (see 2.7).

A.5.1 Combined graphs: representation of GO terms

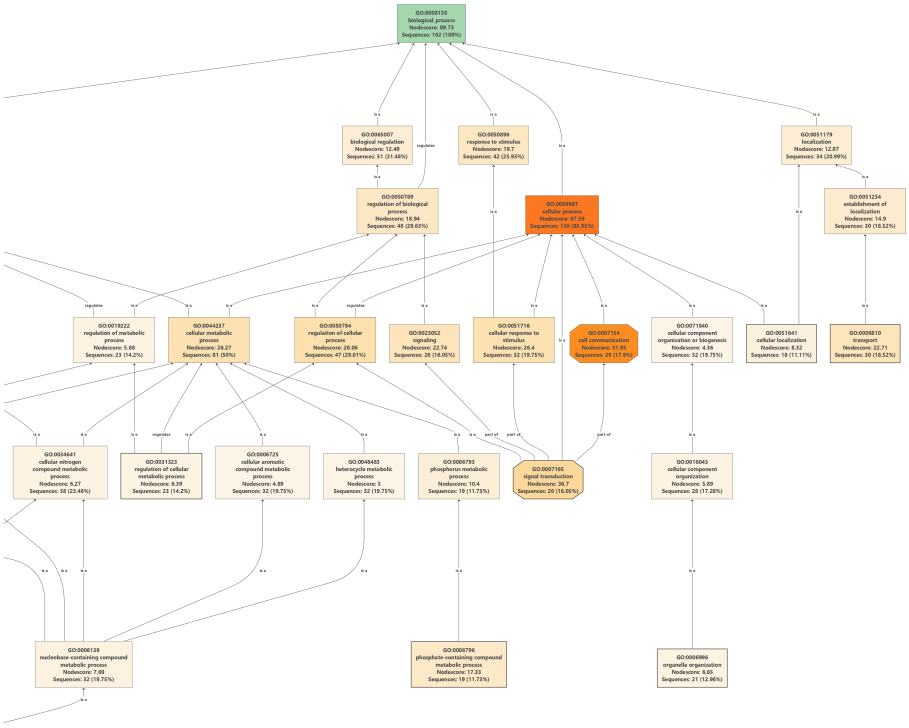


Figure A.18: DAG (directed acyclic graph) gene ontology graph for biological processes, (right-half), integrating the structure of GO hierarchy. Each node represents a GO term, with the increasing orange color intensity correlating with the number of associated sequences.

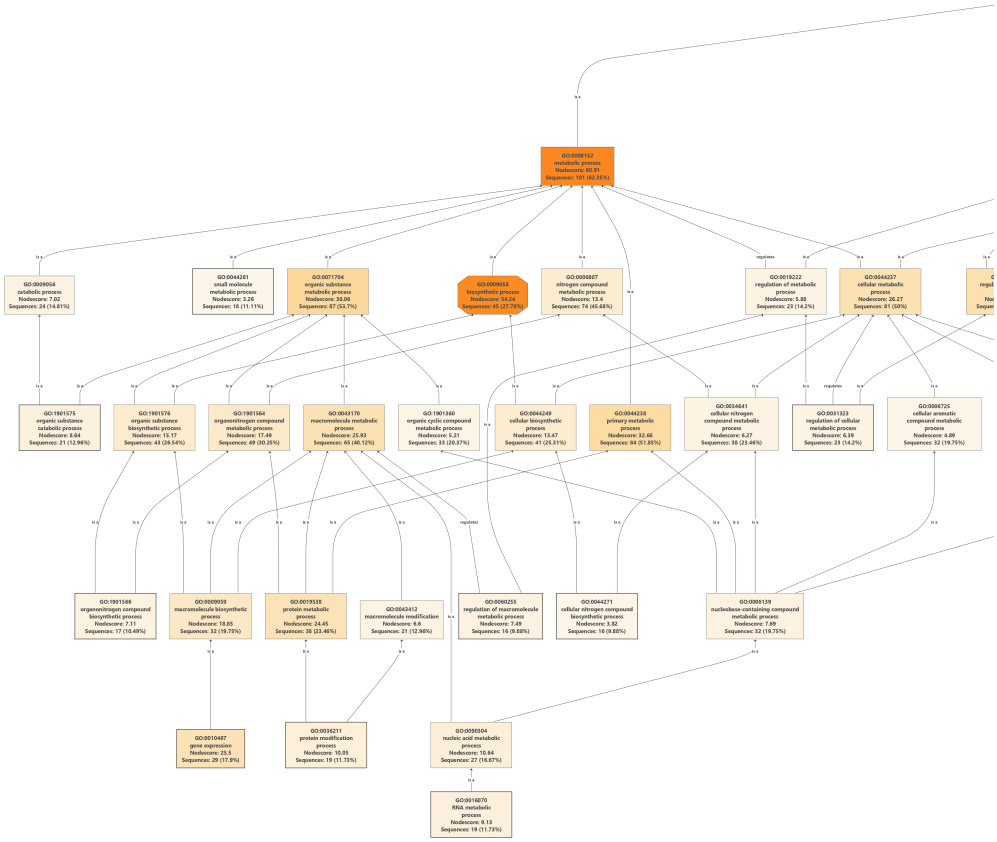


Figure A.19: DAG (directed acyclic graph) gene ontology graph for biological processes, (left-half), integrating the structure of GO hierarchy. Each node represents a GO term, with the increasing orange color intensity correlating with the number of associated sequences.

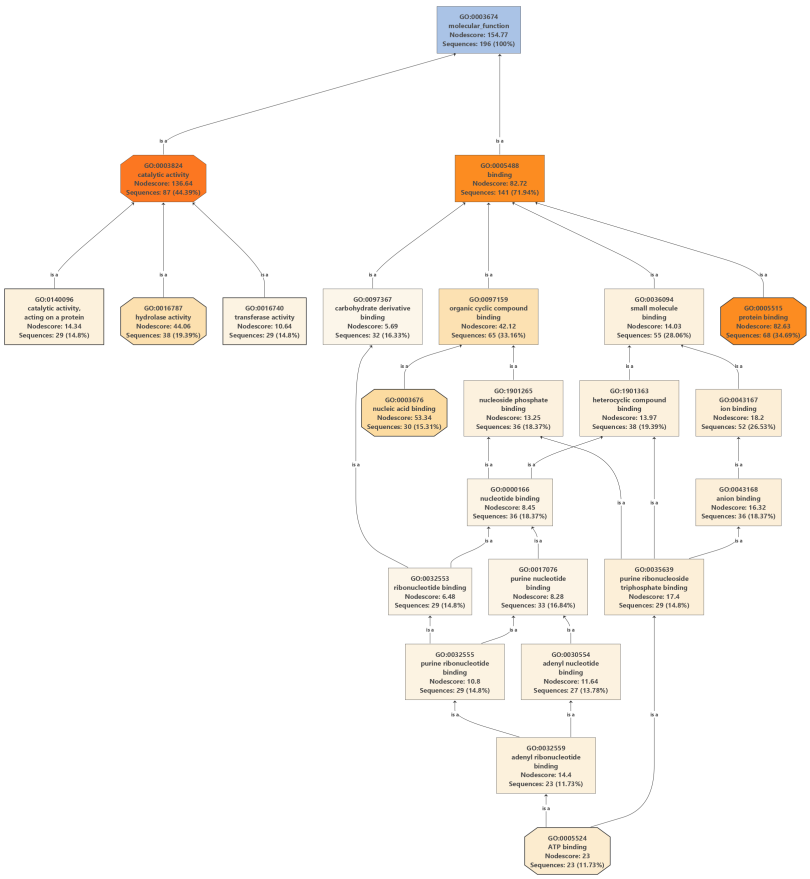


Figure A.20: DAG (directed acyclic graph) gene ontology graph for molecular functions, integrating the structure of GO hierarchy. Each node represents a GO term, with the increasing orange color intensity correlating with the number of associated sequences.

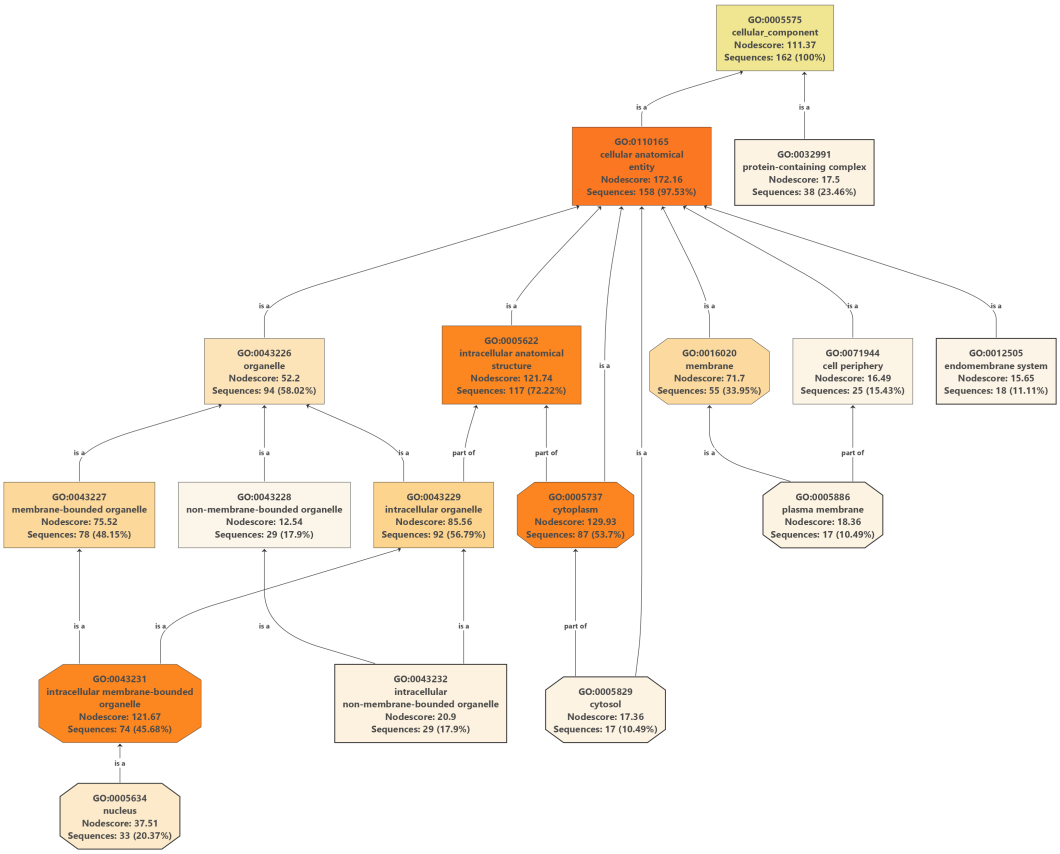


Figure A.21: DAG (directed acyclic graph) gene ontology graph for cellular components, integrating the structure of GO hierarchy. Each node represents a GO term, with the increasing orange color intensity correlating with the number of associated sequences.

Table A.6: Blast transcript-level description: Fibronectin. Summarized table of BLAST2Go functional annotation results.

SeqName	Description	e-Value	GO Names
DN316_c0_g1_i2	hypothetical protein GBAR_LOCUS25894	2.57165E-7	F:protein binding
DN653_c0_g1_i8			F:protein binding
DN829_c0_g1_i4	fibronectin type III domain-containing protein	1.66508E-4	F:protein binding
DN929_c0_g1_i3	serine/threonine-protein kinase svkA-like	1.0861E-11	F:protein binding
DN1494_c0_g1_i3	hypothetical protein	9.61725E-38	F:protein binding
DN3447_c1_g2_i3	—NA—		F:protein binding
DN4522_c0_g1_i2	multiple PDZ domain protein-like	4.68931E-58	F:protein binding
DN5148_c0_g2_i1	hypothetical protein	4.3457E-18	F:protein binding
DN5209_c0_g1_i1	—NA—		F:protein binding
DN5329_c0_g2_i4	SH3 domain-containing protein 19	1.98441E-98	F:protein binding
DN6967_c0_g1_i18	fibronectin-like protein	1.35969E-9	F:protein binding
DN6961_c0_g2_i7	DISP complex protein LRCH3	1.09168E-64	F:protein binding
DN12472_c0_g1_i2	fibronectin type III domain-containing protein	1.46022E-12	F:protein binding
DN25055_c0_g3_i7	—NA—		F:protein binding
DN39253_c0_g1_i10	—NA—		F:protein binding
DN40500_c0_g2_i3	MICAL-like protein 2	8.75068E-47	F:protein binding
DN115878_c0_g1_i1	hypothetical protein J437_LFUL014779	3.89177E-17	F:protein binding
DN2134_c0_g1_i10	UBX domain-containing protein 6-like isoform X2	9.9362E-91	F:protein binding; C:cytoplasm
DN6641_c0_g1_i3	F-box/LRR-repeat protein 2-like isoform X1	4.14107E-113	F:protein binding; C:cytoplasm
DN21703_c0_g3_i5	MIOS	8.7041E-134	F:protein binding; C:cytoplasm
DN5595_c0_g1_i1	proline-serine-threonine phosphatase-interacting protein 1-like	8.39515E-140	F:protein binding; C:cytoplasm; C:plasma membrane
DN16208_c0_g2_i17	mCG15322, partial	5.02671E-67	F:protein binding; C:cytosol
DN9241_c0_g1_i1	angiopoietin-1 receptor-like	6.16087E-23	F:protein binding; C:extracellular space; C:extracellular matrix
DN52594_c0_g2_i2	deleted in malignant brain tumors 1 protein-like	2.43666E-7	F:protein binding; C:membrane
DN1538_c0_g1_i16	fibronectin type protein	1.98542E-6	F:protein binding; C:plasma membrane

Table A.7: Blast transcript-level description: Ubiquitin. Summarized table of BLAST2Go functional annotation results

SeqName	Description	e-Value	GO Names
DN2676_c0_g1_i44	—NA—		P:proteasome-mediated ubiquitin-dependent protein catabolic process; F:ATP binding
DN4985_c0_g1_i6	ubiquitin carboxyl-terminal hydrolase 19-like	1,79086E-83	P:protein deubiquitination; F:cysteine-type deubiquitinase activity
DN28734_c0_g1_i5	ubiquitin carboxyl-terminal hydrolase 2-like isoform X2	2,63733E-94	P:protein deubiquitination; F:cysteine-type deubiquitinase activity; C:cytoplasm

Table A.8: Blast transcript-level description: Cytochrome. Summarized table of BLAST2Go functional annotation results

SeqName	Description	e-Value	GO Names
DN16599_c0_g3_i7	cytochrome b5 reductase 4-like	1,2395E-59	F:oxidoreductase activity; F:FAD binding
DN11085_c0_g1_i2	cytochrome c oxidase subunit III	3,5234E-79	P:respiratory electron transport chain; F:cytochrome-c oxidase activity; C:membrane
DN940_c0_g1_i13	cytochrome P450 3A41-like	1,05888E-73	F:iron ion binding; F:steroid hydroxylase activity; F:oxidoreductase activity

A.5.2 Enrichment Analysis

Table A.9: Results of Fisher's Exact test, test set down-regulated DE transcripts. P-value filter 0.05 as the chosen multiple test correction method. #Test is the number of sequences that are annotated with the GO and are in the test set. #NotAnnotTest is the number of sequences not annotated with that GO in the test set.

Tags	GO ID	GO Name	GO Category	FDR	P-Value	Nr Test	Nr Reference	Non Annot Test	Non Annot Reference
[OVER]	GO:0006644	phospholipid metabolic process	BIOLOGICAL_PROCESS	1.0	0.019513363348983927	3	0	96	266
[OVER]	GO:0043603	amide metabolic process	BIOLOGICAL_PROCESS	1.0	0.036704469095914065	5	3	94	263
[OVER]	GO:0003676	nucleic acid binding	MOLECULAR_FUNCTION	1.0	0.01331653500670121	14	16	85	250
[OVER]	GO:0033554	cellular response to stress	BIOLOGICAL_PROCESS	1.0	0.036704469095914065	5	3	94	263
[OVER]	GO:0097159	organic cyclic compound binding	MOLECULAR_FUNCTION	1.0	0.03765537043386585	24	41	75	225
[OVER]	GO:0015711	organic anion transport	BIOLOGICAL_PROCESS	1.0	0.019513363348983927	3	0	96	266
[OVER]	GO:0006974	DNA damage response	BIOLOGICAL_PROCESS	1.0	0.04849497298324462	4	2	95	264
[OVER]	GO:0006950	response to stress	BIOLOGICAL_PROCESS	1.0	0.0275326376011748	6	4	93	262
[OVER]	GO:0006518	peptide metabolic process	BIOLOGICAL_PROCESS	1.0	0.036704469095914065	5	3	94	263

Table A.10: Results of Fisher's Exact test, test set up-regulated DE transcripts. P-value filter 0.05 as the chosen multiple test correction method. #Test is the number of sequences that are annotated with the GO and are in the test set. #NotAnnotTest is the number of sequences not annotated with that GO in the test set.

Tags	GO ID	GO Name	GO Category	FDR	P-Value	Nr Test	Nr Reference	Non Annot Test	Non Annot Reference
[OVER]	GO:0050790	regulation of catalytic activity	BIOLOGICAL_PROCESS	1.0	0.040321816630833035	10	0	256	99
[OVER]	GO:0005856	cytoskeleton	CELLULAR_COMPONENT	1.0	0.040321816630833035	10	0	256	99
[OVER]	GO:0065009	regulation of molecular function	BIOLOGICAL_PROCESS	1.0	0.040321816630833035	10	0	256	99
[OVER]	GO:0016192	vesicle-mediated transport	BIOLOGICAL_PROCESS	1.0	0.040321816630833035	10	0	256	99

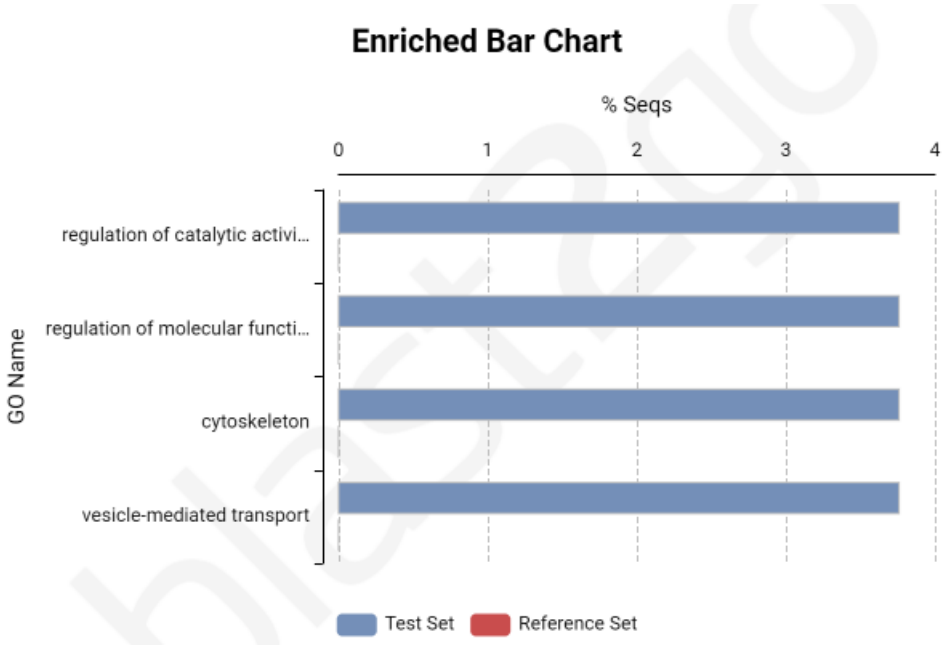


Figure A.22: Enrichment analysis bar chart using the up-regulated transcripts identified in DESeq2 as the test sequences. The reference set of sequences includes all of the sequences identified as differentially expressed (see 2.7).

Appendix B

Scripts

B.1 Trimmomatic

$\#!/bin/bash \#SBATCH --partition = batch$
$\#SBATCH --nodes = 1$
$\#SBATCH --ntasks - per - node = 20$
$\#SBATCH --mem =$
$\#SBATCH --job - name = "trim - T*"$
$\#SBATCH --output = trimT*.output$
$\#SBATCH --mail - user = carlottapaone26@gmail.com$
$\#SBATCH --mail - type = ALL$
$\#SBATCH --requeue$

$adapters = /home1/carlotta/illumina - 3adaptors.fasta$

$java - jar/mnt/big/NGS/Trimmomatic - 0.39/trimmomatic - 0.39.jar$
$PE - threads20 - phred33/home1/carlotta/T* - R1.fq.gz$
$/home1/carlotta/T* - R2.fq.gzoutput - T* - R1 - trim - paired.fastq.gz$
$output - T* - R2 - trim - paired.fastq.gzoutput - T* - R1 - trim -$
$unpaired.fastq.gz$

$output - T * -R2 - trim.unpaired.fastq.gz$
$ILLUMINACLIP : \$adapters : 3 : 30 : 10LEADING : 3TRAILING :$
$3SLIDINGWINDOW : 4 : 15MINLEN : 36$

B.2 FastQC

$\#!/bin/bash\#SBATCH--partition = batch\#SBATCH--nodes = 1\#SBATCH-$
$--ntasks - per - node = 1$
$\#SBATCH - -mem - per - cpu = 4000$
$/mnt/big/NGS/FastQC-0.11.9/fastqc/home1/carlotta/fastq/output-$
$T * .fastq.gz$

B.3 Trinity

$\#!/bin/bash-l\#SBATCH--partition = hugemem\#SBATCH--nodes =$
$1\#SBATCH - -ntasks - per - node = 20$
$modulepurge$
$moduleloadsingularity/3.7.1$

$\#wgethttps : //data.broadinstitute.org/Trinity/TRINITY -$
$SINGULARITY/trinityrnaseq.v2.14.0.simg$
$trinity - image = /home1/carlotta/trinityrnaseq.v2.14.0.simg$

$echo"RunningTrinityv2.14.0fornon - strand - specificrnareads.$
$Needsthefilesamples.txtinthecurrentdirectory."echo"nToreproducerun : "$
$echo"singularityexec-e\$trinity-imageTrinity--seqTypefq--SS-lib-$
$typeRF - -max - memory640G - -CPU20 - -min - contig - length200 -$
$--jaccard-clip--outputtrinity--full-cleanup--normalize-by-read-$
$set - -left/home1/carlotta/output - T * -R1 - trim - paired.fastq.gz -$

$-right/home1/carlotta/output - T * -R2 - trim - paired.fastq.gz$
$- -output/home1/carlotta/Cnucula - trinity - all - out$"
$mvtrinity.Trinity.fastaTrinity.fasta$
$singularityrun\$trinity - image/usr/local/bin/util/TrinityStats.pl$
$Trinity.fastaTrinity - stats$

B.4 BUSCO

$\#!/bin/bash - l\#SBATCH - -partition = batch$
$busco - i/home1/carlotta/Cnucula - trinity - all - out.Trinity.fasta -$
$oCnucula-busco-all-out-l/mnt/big/Assembly/busco-3.0.2b/datasets/metazoa-$
$odb9 - mtranscriptome - cpu20$

B.5 ExN50 Statistic

$\#!/bin/bash - l\#SBATCH - -partition = batch$
$modulepurge$
$moduleloadsingularity/3.7.1$
$trinity - image = /home1/carlotta/trinityrnaseq.v2.14.0.simg$
$\#Trinitycommandsingularityexec - e\$trinity - image$
$/home1/carlotta/trinityrnaseq-v2.15.1/util/misc/plot-ExN50-statistic.Rscript$
$/home1/carlotta/ExN50.statsxpdfExN50.stats.plot.pdf$
$moduleunloadsingularity/3.7.1$
exit 0

B.6 Filter low expression transcripts

$Parameters : singularityexec - e\$trinity - image$
$/home1/carlotta/trinityrnaseq-v2.15.1/util/filter-low-expr-transcripts.pl-$

m
$/home1/carlotta/RSEM-transcripts/RSEM.isoform.TPM.not-cross-norm-t$
$/home1/carlotta/trinity-fasta/Cnucula-trinity-all-out.Trinity.fasta--min-expr-any1--gene-to-trans-map/home1/carlotta/trinity-fasta/Cnucula-trinity-all-out.Trinity.fasta.gene-trans-map>$
$/home1/carlotta/outputfilteredB.fasta$

Output: Retained 388731 / 477283 = 81.45% of total transcripts.

B.7 RSEM: RNA-Seq by Expectation-Maximization

$\#!/bin/bash-l\#SBATCH--partition=batch$
$modulepurge$
$moduleloadsingularity/3.7.1$
$trinity-image=/home1/carlotta/trinityrnaseq.v2.14.0.simg$
$\#Trinitycommandsingularityexec-e\$trinity-image/usr/local/bin/util/align-and-estimate-abundance.pl--transcripts$
$/home1/carlotta/Cnucula-trinity-all-out.Trinity.fasta--SS-lib-typeRF--seqTypefq--samples-fileCnucula-samples.txt--est-methodRSEM--aln-methodbowtie2--trinity-mode--prep-reference--thread-count20--output-dir/home1/carlotta/Cnucula-output.RSEM$
$moduleunloadsingularity/3.7.1$
exit 0

B.8 DESeq2 with R

script to perform differential gene expression analysis using DESeq2 package
$library(DESeq2)$
$library(tidyverse)$

Step 1: preparing count data

$counts - data <- read.matrix(RSEM.gene.counts.matrix)$

$colData <- read.table('colData')$ #Sample information

row names in colData need to match the column names in counts-data and be in the same order

$all(colnames(counts - data)\%in\%rownames(colData))$

$all(colnames(counts - data) == rownames(colData))$

Step 2: constructing a DESeqDataSet object

$dds <- DESeqDataSetFromMatrix(countData = counts-data, colData = colData,$

$design = Condition)$

#conditions as untreated, untreated-control and treated

pre-filtering: removing rows with low gene counts # keeping rows that have at least 10 reads total $keep <- rowSums(counts(dds)) >= 10 dds <- dds[keep,]$

setting the factor level

$dds\$Condition <- relevel(dds\$Condition, ref = "untreated - control")$

Step 3: Run DESeq

$dds <- DESeq(dds)$

$res <- results(dds)$

$summary(res)$

#alpha is automatically set to 0.1; for downstream analysis I have set alpha to 0.05

B.8.1 MAplot and lfc shrinkage

#MAplot of the unshrunken data results from DESeq2 plotMA(res)

#Performing the lfc shrinkage, type "apeglm", and then running the MAplot on the transformed data

$res_lfcShrink <- lfcShrink(dds, coef = "Condition_treated_vs_untreated_control", type = "apeglm")$

$plotMA(res - lfcShrink)$

B.8.2 Plot count

#Plot count grouped by treatment variable ('Condition'), specifying the gene with the smallest p-adjusted value from the results

$plotCounts(dds, gene = which.min(res\$padj), intgroup = "Condition")$

B.8.3 Rlog transformation function

#The blind was set to false as it was not necessary to re-estimate the dispersion values (as DESeq2 had already been run)

$rld <- rlog(dds, blind = FALSE)$

#The assay function is used to extract the matrix of normalized values

$head(assay(rld)$

#Heatmap of the rlog transformed data

$library("pheatmap")select <- order(rowMeans(counts(dds, normalized = TRUE)),$
$decreasing = TRUE)[1:20]$
$df <- as.data.frame(colData(dds)[, c("Condition","Sample")])$
$pheatmap(assay(rld)[select,], cluster_rows = FALSE, show_rownames = FALSE,$
$cluster_cols = FALSE, annotation_col = df)$

B.8.4 Principal component analysis plot

#PCA plot of the rlog transformed data, grouped by condition-type

$plotPCA(rld, intgroup = c("Condition"))$

#PCA plot of the rlog transformed data, grouped by condition and sample

$plotPCA(rld, intgroup = c("Condition","Sample"))$

B.8.5 Cook's distances boxplot

#Cook's distances are stored as a matrix available in

$assays(dds)[["cooks"]]$

#Constructing a boxplot of the Cook's distances across samples

$par(mar = c(8,5,2,2))boxplot(log10(assays(dds)[["cooks"]]), range = 0, las = 2)$

B.8.6 Dispersion plot

#Plotting of the dispersion estimates

$plotDispEsts(dds)$

B.8.7 Volcano plot

#Default cut-off for p value for the basic volcano plot is given as 10e-6; cut-off for log2FC is $> |2|$

$BiocManager::install('EnhancedVolcano')$

$library(EnhancedVolcano)$

$EnhancedVolcano(res, lab = rownames(res), x =' log2FoldChange', y =' pvalue')$

B.8.8 Subset significant DE genes

#Subsetting the significant up-regulated and down-regulated genes given log2FOldChange and p-value thresholds

$resSig <- subset(res, padj < 0.05)$

$upReg <- subset(resSig, log2FoldChange > 0)$

$downReg <- subset(resSig, log2FoldChange < 0)$