

Earthworm system immunity and its modulation by nanoparticles

Thesis for Doctor of Philosophy

Szabolcs Balazs Hernadi

Cardiff University

2020



Supervisory group:

Prof. P. Kille¹, Dr C. Svendsen², Prof P. Borri¹.

¹ Cardiff School of Biosciences, BIOSI 1, University of Cardiff, P.O. Box 915, Cardiff, CF10 3TL, UK.

² Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Wallingford, Oxfordshire OX10 8BB, UK.

Abstract

The particulate nature of nanoparticles (NPs) dictates a preferential interaction with cells of the immune system assigned to recognition and elimination of foreign particulates. Probing safety of nano-objects by defining immune responses of environmental organisms is therefore key to environmental nanosafety. Earthworms represent major immunosafety models representing keystone ecosystem engineers and being in intimate contact with soils ensuring exposure to terrestrial NPs. Innate immunity represents the first line of defence against pathogens in invertebrates and despite extensive description of cell populations involved a dearth of information about the associated molecular components exists. This thesis aimed to use genomics approaches to generate a comprehensive systems immunity description of earthworm prior to exploiting transcriptomics to explore the interaction of NPs and earthworms.

A tissue-specific transcriptomic atlas has been established for *Eisenia fetida* and *Eisenia andrei* representing six tissues from each species. The resultant comparative transcriptomic resource represented an innate immunity database containing immune-related genes from major immune signalling pathways. To refine the tissue-specific database we generated a *de novo* genome for *E. fetida* with high contiguity and completeness. Finally, to enhance insight into the different immune functions of the individual types of coelomocytes we generated transcriptomic datasets from eleocytes, hyaline and granular amoebocytes.

The interaction between NPs and the earthworm immune system was then explored by combining the direct introduction of copper oxide NPs with bacterial challenge and following the transcript changes within individual coelomocytes cell populations. This resolved the spatial-temporal impact on the immune system into three distinct phases: direct, systemic and differentiation responses. A complementary soil-based exposure using a range of CuNPs, Silicon-CuNPs and copper ion doses explored the comparative response after soil-based biotransformation. This revealed Si-CuNPs to elicit a negligible response whilst differentially regulated genes under high CuNPs were distinct from equivalent copper ion exposure the pathways impacted intersected substantially.

Acknowledgements

First, of all I would like to express my truest gratitude to Prof. Pete Kille for his timeless guidance, patience, friendship and source of morning coffees/hot chocolates, spiced with the most inspiring scientific theories during this PhD. He always provided a warm atmosphere for me and my family and many times he had the confidence in me to push me towards things that I never expected I could do.

Special thanks goes to Dr Stephen Short for his selfless help, provided insights and friendship through this endeavor.

I would like to thank to my parents and my brother for the constant support during my studies as well as for keeping me on track even from thousands of miles away.

Thank for my partner Barbara who not only sustained, motivated, tolerated and kept my sanity during the whole journey, but also involuntarily learnt a lot of things outside of her particular interest such as the challenges of HMW DNA extraction, earthworms immunology or complex genome assemblies.

Huge credit should go to Dr Alexander G. Robinson, Dr Amaia Green Etxabe, Dr Claus Svendsen, Prof. Dave Spurgeon, and Elmer Swart for both the constant supply of earthworms and for generating data for bioinformatic analysis.

For their help and support further thanks must go to Prof. Laszlo Molnar, Angela Marchbank, Daniel Pass, Georgina Smethurst, Iain Alexander Perry, Dr Luis Cunha, Oliver Rimington.

This PhD was funded in the framework of H2020 Marie Skłodowska- Curie ITN programme. (Grant ID: 671881, Grant title: 'Probing safety of nano-objects by defining immune responses of environmental organisms' (PANDORA)

Table of Contents

1	General Introduction	1
1.1	Immunity	2
1.2	Invertebrate Immune system.....	2
1.2.1	Innate immunity - conserved but diverse	2
1.2.2	Invertebrate immunity in general.....	3
1.3	Aspects of earthworm immunity.....	5
1.3.1	Basics about earthworm anatomy	5
1.3.2	Physical barrier.....	Error! Bookmark not defined.
1.3.3	Cellular immunity	7
1.3.4	Humoral immunity	8
1.4	Immune system modulated by Nanoparticles	10
1.4.1	Definition and origin of Nanoparticles.....	10
1.4.2	Applications and environmental risks.....	14
1.4.3	Nanomaterials and Innate Immunity.....	14
1.4.4	Nanoparticles and earthworm immunity	15
1.5	The Aims of this Thesis	17
1.5.1	Chapter 2 - Transcriptomic Tissue Atlas of Earthworm Immunity.....	17
1.5.2	Chapter 3 – Genomic template to innate immunity.....	17
1.5.3	Chapter 4 – Characterisation of the coelomic fluid	17
1.5.4	Chapter 5 – Spatio-temporal response to challenge	18
1.5.5	Chapter 6 - Effects of indirect MNPs exposure.....	18
2	Transcriptomic Tissue Atlas of Earthworm Immunity	19
2.1	Introduction.....	20
2.2	Materials and Methods	21

2.2.1	Biological material.....	21
2.2.2	Genotyping.....	21
2.2.3	Experimental design and sample preparation.....	22
2.2.4	Transcriptomic analysis.....	23
2.2.5	Tissue differential expression analysis.....	23
2.2.6	Annotation	26
2.2.7	Gene enrichment analysis.....	26
2.3	Results	27
2.3.1	Transcriptome assembly	27
2.3.2	Transcriptome annotation	29
2.3.3	Tissue functional profiling.....	33
2.3.4	Immune gene identification.....	37
2.4	Discussion	42
2.4.1	Transcriptome assembly and annotation	42
2.4.2	Tissue functional profiling.....	43
2.4.3	Immune gene identification.....	43
2.4.4	Limitations of the non-model <i>de novo</i> pipeline and future developments	44
3	Genomic template for Innate Immunity.....	47
3.1	Introduction.....	48
3.2	Materials and Methods	51
3.2.1	DNA extraction	51
3.2.2	DNA quality and quantity assessment	55
3.2.3	Nanopore library preparation.....	55
3.2.4	Library loading and flow-cell priming	56
3.2.5	Nanopore sequencing and base calling	56

3.2.6	Illumina Short Read Genomic Data Generation.....	59
3.2.7	Quality control and error correction of the sequencing data	59
3.2.8	Initial <i>de novo</i> assembly	59
3.2.9	Scaffolding using transcriptomic data.....	60
3.2.10	Scaffolding with Nanochrome.....	60
3.2.11	Assembly QC and filtering	62
3.2.12	Repeat masking	65
3.2.13	Gene prediction.....	65
3.2.14	Phylogenetic analysis	66
3.3	Results	67
3.3.1	DNA extraction and long read sequencing	67
3.3.2	Paired-end and Chromium 10X short read sequencing.....	70
3.3.3	Initial long read assembly	70
3.3.4	Scaffolding with Nanochrome.....	70
3.3.5	Contiguity and completeness of the final assembly	72
3.3.6	Identification of repeats and low complexity regions	75
3.3.7	Genome annotation	76
3.3.8	Identification of Toll-like receptor genes.....	77
3.4	Discussion	81
3.4.1	HMW DNA extraction for Nanopore sequencing	81
3.4.2	Nanopore long-read sequencing.....	82
3.4.3	Assembly and 10X Chromium scaffolding.....	82
3.4.4	Genome annotation	84
3.4.5	Identification of Toll-like receptor genes.....	87
3.4.6	Concluding key results and possible future developments.....	88
4	Characterisation of the Coelomic fluid	89

4.1	Introduction.....	90
4.2	Materials and Methods	92
4.2.1	Experimental design.....	92
4.2.2	Cell preparation for sorting.....	92
4.2.3	Cell sorting.....	93
4.2.4	Library preparation and sequencing	94
4.2.5	Data analysis	94
4.2.6	Functional annotation and enrichment	94
4.2.7	Analysis overview	95
4.3	Results	97
4.3.1	Coelomocyte preparation and sorting.....	97
4.3.2	Sequencing and read processing	99
4.3.3	Cell type-specific marker genes	100
4.3.4	Gene enrichment analysis of cell type-specific markers	101
4.3.5	Expression profile of Toll-like receptors	105
4.3.6	Expression profile of immune related genes	107
4.4	Discussion	108
4.4.1	Coelomocyte preparation for cell sorting.....	108
4.4.2	Identification of cell specific marker genes	108
4.4.3	Functional enrichment analysis	108
4.4.4	Cell-specific expression patterns of TLRs and other target genes of earthworm immunology.....	110
4.4.5	Conclusion	112
5	Temporal response to challenge	113
5.1	Introduction.....	114
5.2	Materials and Methods	118

5.2.1	Biological material.....	118
5.2.2	Experimental design.....	118
5.2.3	Nanoparticle exposure.....	120
5.2.4	Bacterial challenge.....	121
5.2.5	Coelomocyte preparation and sorting.....	121
5.2.6	RNA extraction and library preparation.....	122
5.2.7	Data generation and QC.....	122
5.2.8	Mapping and differential gene expression analysis.....	123
5.2.9	Temporal Gene clustering.....	123
5.2.10	Functional enrichment analysis.....	123
5.2.11	Analysis overview.....	124
5.3	Results.....	125
5.3.1	Quality control and signal processing.....	125
5.3.2	Differential gene expression analysis using cubic spline regression model 130	
5.3.3	Gene Ontology enrichment on the spatial data.....	130
5.3.4	Pathway enrichment on the spatial data.....	131
5.3.5	Temporal response to combined bacterial and NPs challenge.....	138
5.3.6	Combined impact of bacterial and NPs on copper metabolism.....	142
5.3.7	Combined impact of bacterial and NPs effect on the Toll-like pathway	146
5.4	Discussion.....	150
5.4.1	Experimental design.....	150
5.4.2	Differential gene expression analysis.....	151
5.4.3	Time independent, spatial analysis of DEGs.....	151
5.4.4	Temporal gene enrichment analysis.....	153
5.4.5	Metal stress response.....	154

5.4.6	Conclusion	155
6	Comparing NP effect within direct and Indirect exposures	156
6.1	Introduction.....	157
6.1.1	Environmental effects of engineered metal NPs	157
6.1.2	Experimental design.....	158
6.2	Materials and Methods	159
6.2.1	Experimental design.....	159
6.2.2	RNA isolation	160
6.2.3	Sequence generation	160
6.2.4	Read pre-processing and QC.....	160
6.2.5	Mapping and read counting.....	161
6.2.6	Differential gene expression analysis	161
6.2.7	Functional enrichment	161
6.3	Results	163
6.3.1	Sequencing and quality trimming	163
6.3.2	Read mapping	163
6.3.3	Exploration of holistic data	165
6.4	Overlaps in the differentially expressed gene profiles	172
6.5	Gene Ontology enrichment analysis	172
6.6	Pathway enrichment analysis.....	178
6.7	Discussion	181
6.7.1	Differential gene expression analysis	181
6.7.2	Bioavailability of copper ions, released from NPs	181
6.7.3	Gene ontology and pathway enrichment analysis.....	182
6.7.4	Shared effects in the high concentration of Cu ⁺ and Cu-NP exposures.....	183
6.7.5	Conclusion	185

7	Final discussion	186
7.1	Background of the study	187
7.2	Challenges.....	187
7.3	Summary and brief interpretation of key findings.....	188
7.3.1	Genomic and transcriptomic framework of earthworm immunity	188
7.3.2	Spatio-Temporal characterisation of the direct NP effect.....	190
7.3.3	Comparing direct NP impact with ecotoxicological relevance	191
7.4	Limitations	191
7.4.1	The genomic framework of earthworm immunity	191
7.4.2	Spatio-temporal immune effect of CuNP	193
7.4.3	Indirect Copper exposure.....	193
7.5	Future developments	194
7.5.1	Metal metabolism in earthworm – a cell specific affair	194
7.5.2	Differential transcripts usage and non-polyadenylated RNAs.....	194
7.5.3	Origin and maturation of coelomocytes	195
7.5.4	Coelomocyte classification based on transcriptomic fingerprints.....	196
7.5.5	Conclusion	196
8	References	198
9	Appendices.....	221

1 General Introduction

1.1 Immunity

According to our best estimation, the first primitive anaerobic life forms appeared around 3700 million years ago (MYA) followed by simple photosynthesising unicellular organisms, not long after the Earth was formed (Björn and Govindjee 2008). Ever since then, it is a necessity for all the living creatures at any level of organisation to try to preserve their integrity and survive against the attack of different pathogens and parasites. The simple meaning of the word "Immunity" is "released from a burden" which perfectly describes the critical abilities required even for a primitive microorganism to protect itself against an infectious agent (Buchmann, 2018). First, the microorganism needs to be able to create contact with the possible pathogen, followed by the recognition processes and a response with either phagocytosis or rejection. From the multicellular organism, one of the most important improvements was the capability of self and non-self-recognition which has allowed the development of the cell-cell signalling, immunological memory, humoral or cellular effector mechanisms, and created the possibility of developing tissues by allowing the formation of cell junctions .

1.2 Invertebrate Immune system

1.2.1 Innate immunity - conserved but diverse

Study of the invertebrate immune system has a comprehensive and rich history (Cooper Edwin et al. 1992). Several vital experiments in modern immunology were conducted not in humans or different vertebrate model organisms, but in sea star larvae and other invertebrates (Dheilly et al. 2014). Furthermore, the recognition of major primary immune defence mechanisms, such as phagocytosis and encapsulation, became possible due to the dedicated observations of invertebrate species (Metchnikoff 1968). Although, invertebrate species represent roughly 90% of the animal species across the whole planet and embraces a massive diversity of anatomical structures and lifestyles. For obvious reasons, with attention directed at medical and veterinary purposes, most of the immunological studies focuses on human or different well-known vertebrate model organisms (Canesi et al. 2016). Aside from a few exceptions, our knowledge about the immune system of the individual invertebrate taxa is often rather incomplete in detail. The lack of detailed data on immune responses in the case of non-mammalian

species resulted in an inaccurate interpretation which suggested that invertebrates have a rather simple and rudimentary immune system. However, during the last few decades, comparative immunology demonstrated that in many cases the immune system of lower order vertebrates and invertebrates is much more complex than we assumed previously (Loker et al. 2004, Rowley and Powell 2007).

1.2.2 Invertebrate immunity in general

In the absence of the conventional adaptive immune system, invertebrate species rely on an alternative system for biological host defence, called innate immunity (Figure 1) (Iwanaga and Lee 2005, Humphreys and Reinherz 1994). Innate immunity is a component of both invertebrate and vertebrate systems forming the first line of defence against different types of pathogens such as bacteria, viruses, fungi or other eukaryotic parasites (Figure 1). Due to the lack of an adaptive immune system, invertebrates do not have a diverse group of immunoglobulins but have alternative mechanisms to recognise and react to different germ-line encoded microbial surface antigens (Rämet et al. 2003, Rämet et al. 2001). Although, these are not antigen-specific mechanisms they are easily able to recognise several typical pathogenic molecules such as lipoproteins, peptidoglycan, lipopolysaccharide, lipoteichoic acid, and β -D-glucans (Medzhitov 2001). In general the invertebrate innate immune system consists of two different elements which are known as cellular and humoral immunity (Ratcliffe et al. 1982). The humoral immunity is based on physical barriers and different macromolecules such as antimicrobial peptides, cytokines, components of the complement system, lysozyme, pattern recognition receptors, toxic oxygen, and nitrogen metabolites. While in most invertebrate species the cellular defence is achieved by the relatively quick responses of specialised immune cell subpopulations. These can neutralise and remove the pathogens, using distinct cellular immune processes such as encapsulation phagocytosis or degranulation, a process that releases antimicrobial molecules from secretory vesicles.

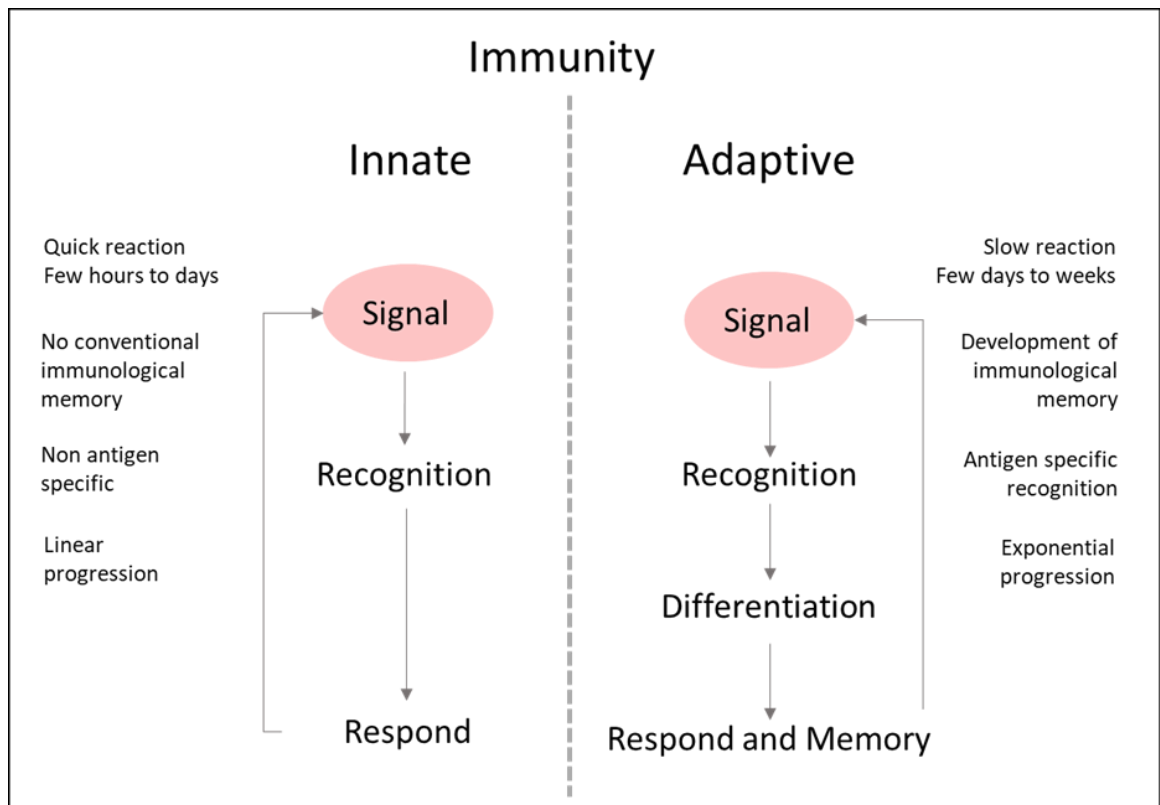


Figure 1: General characteristics comparison of the Innate and Adaptive immune response.

One of the most impressive observations of modern immunology was the recognition of conservativeness in the fundamental immune mechanism across the members of the whole animal kingdom. Several components of the pathogen recognition and signalling processes including the large group of pattern recognition receptors (PRRs) and small signalling molecules, proved to be well conserved not only across the animal species but in plants as well (Zhang et al. 2010, Zipfel 2014).

In contrast to the observed similarities, invertebrate immunity provides a fascinating divergence of the self-protective molecular mechanisms within distinct phyla and classes (Schulenburg et al. 2009). Due to the fact Invertebrates represent phyla with long independent evolutionary backgrounds, the variation within their anatomic structures and lifestyles is considerable. Many of these species live in an infectious agent rich environment, and have persisted long enough with only marginal changes in their habitats, to drive the formation of a lineage-specific group of pathogens. The results of the last few years in the field of comparative immunology pointed out that in many cases even species from the same order has developed rather different molecular

mechanisms to reserve their integrity against a group of coevolved pathogens or parasites (Ghosh et al. 2011, Coates and Nairn 2014).

1.3 Aspects of earthworm immunity

1.3.1 Basics about earthworm anatomy

Earthworms represents the largest family (Lumbricidae) within the Annelida phylum. Their habitat shows great variety, they can be found from terrestrial to freshwater environments. They have a great impact on the ecosystem due to their role in decomposing the organic layer of the soil (Jouquet et al. 2006). They are protostomian animals developing a true coelom derived from the mesenchyme. The coelomic cavity is segmented by transversal septa and filled with the coelomic fluid (Figure 2.). The coelomic fluid is in contact with all the organs and it is open to the environment by a pair of dorsal pore and nephridia at each segment (Lów et al. 2016). Therefore the coelomic cavity always contains a relatively high number and diverse group of microorganisms (Dvořák et al. 2016). Due to their habitat earthworms live in an extremely antigenic environment in contrast to most other invertebrate taxa (Drake and Horn 2007). To survive in such an immunologically challenging environment combined with the openness of the coelomic cavity, they require an extremely effective immune system. To provide sufficient defence from the constantly invading pathogens the coelomic fluid of the earthworms contains a high number of free-floating cells called coelomocytes, a heterologous population of cells that have been shown to be a fundamental component of the earthworm immune response (Stein et al. 1977, J Diogène 1997).

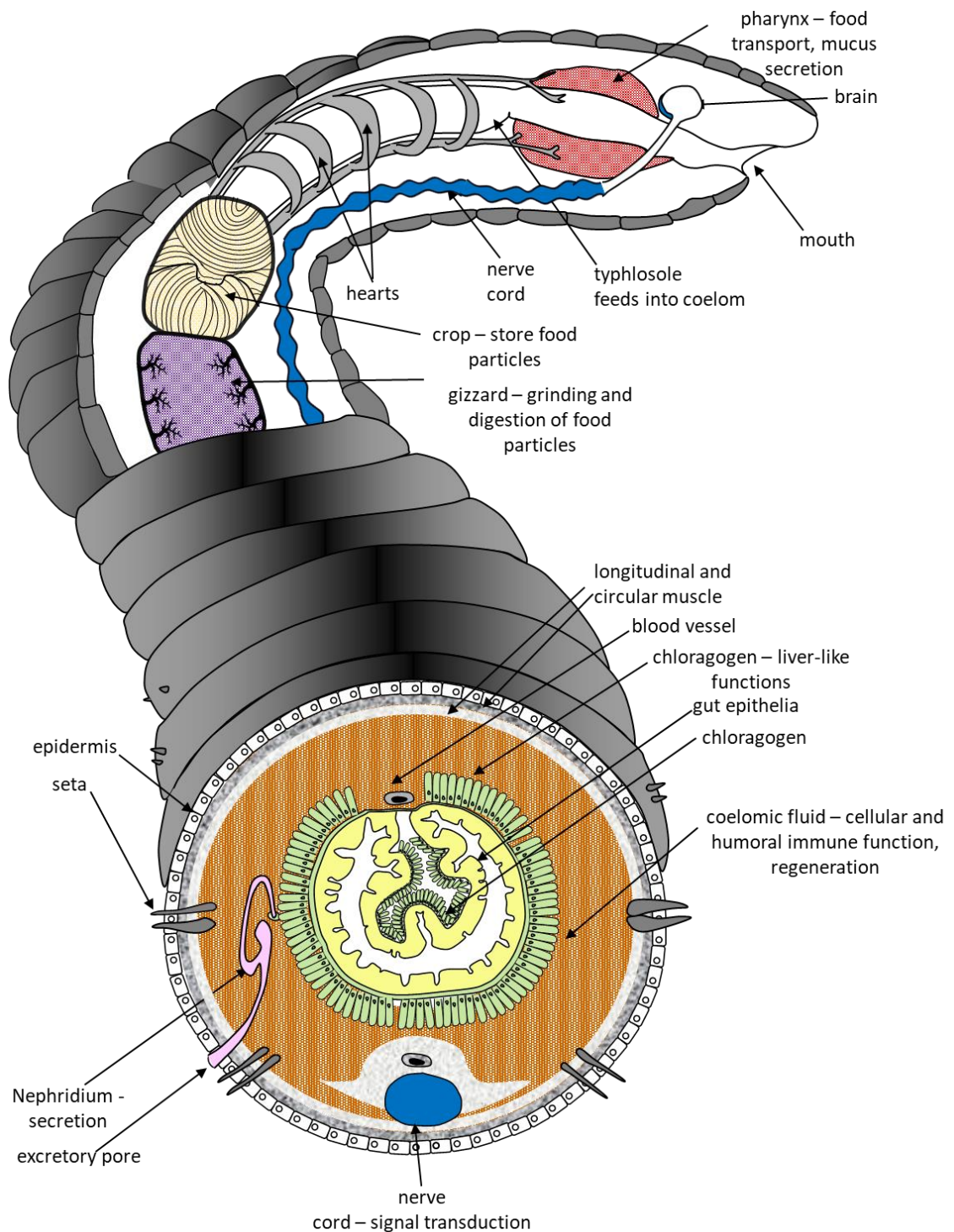


Figure 2: Schematic diagram illustrating the different anatomical structures of an earthworm, as well as showing some of their characteristic physiological functions.

The first line of immune defence in earthworms is provided by the physicochemical properties of the skin (Rahemtulla and Løvtrup 1974). Due to the conductivity of the dorsal pores, the epidermis only supplies partial but not an aseptic separation between the coelomic fluid and the external environment. However, the thin cuticle layer on the

surface of the epithelium is rich in different mucopolysaccharides which act as an effective antimicrobial barrier (Bilej et al. 2013).

1.3.2 Cellular immunity

The second line of defence is provided by the coelomocytes, as they represent the major part of cellular immunity (Hostetter and Cooper 1974). Similar to many invertebrate species where cellular immunity is provided by a heterologous population of free-floating cells, according to their morphological, histochemical, and functional characteristics earthworm coelomocytes can be divided between three major distinct cell populations (Figure 3). Hyaline and granular amoebocytes (Figure 3), these both have important roles in the elimination of the invading pathogens due to their ability to conduct phagocytosis and encapsulation. These cells are both originated from the mesenchyme and they were named according to their specific histologic characteristics (Mácsik et al. 2015). Although the third cell population named eleocytes (Figure 3) is lacking the ability to perform phagocytosis, it still has an important role in the cellular immune defence. Eleocytes are large cells with a high number of special fluorescent granules (chloragosomes) in their cytoplasm (Figure 3), derived from the chloragogen tissue. Earlier, by their functional similarities, the chloragogen was referred to as a liver like tissue, however the results of the last few years in the field of earthworm histochemistry highlighted its possible myelo-erythroid nature (Fischer 1993). Although free-floating eleocytes and the chlorogogenous tissue share their basic functional characteristics, beyond the eleocytes nutritive functions they also play a major role in the degranulation process as well as producing several antimicrobial peptides and bioactive molecules. According to a recent study eleocytes also have similar properties to neutrophil granulocytes in higher-order vertebrates, which manifests in the capability of creating extracellular traps (NETs) (Homa 2018). These structures can protect against extracellular pathogens by regulating their spread in the coelomic cavity and disarm them with the deployment of different antimicrobial molecules (Homa et al. 2016).

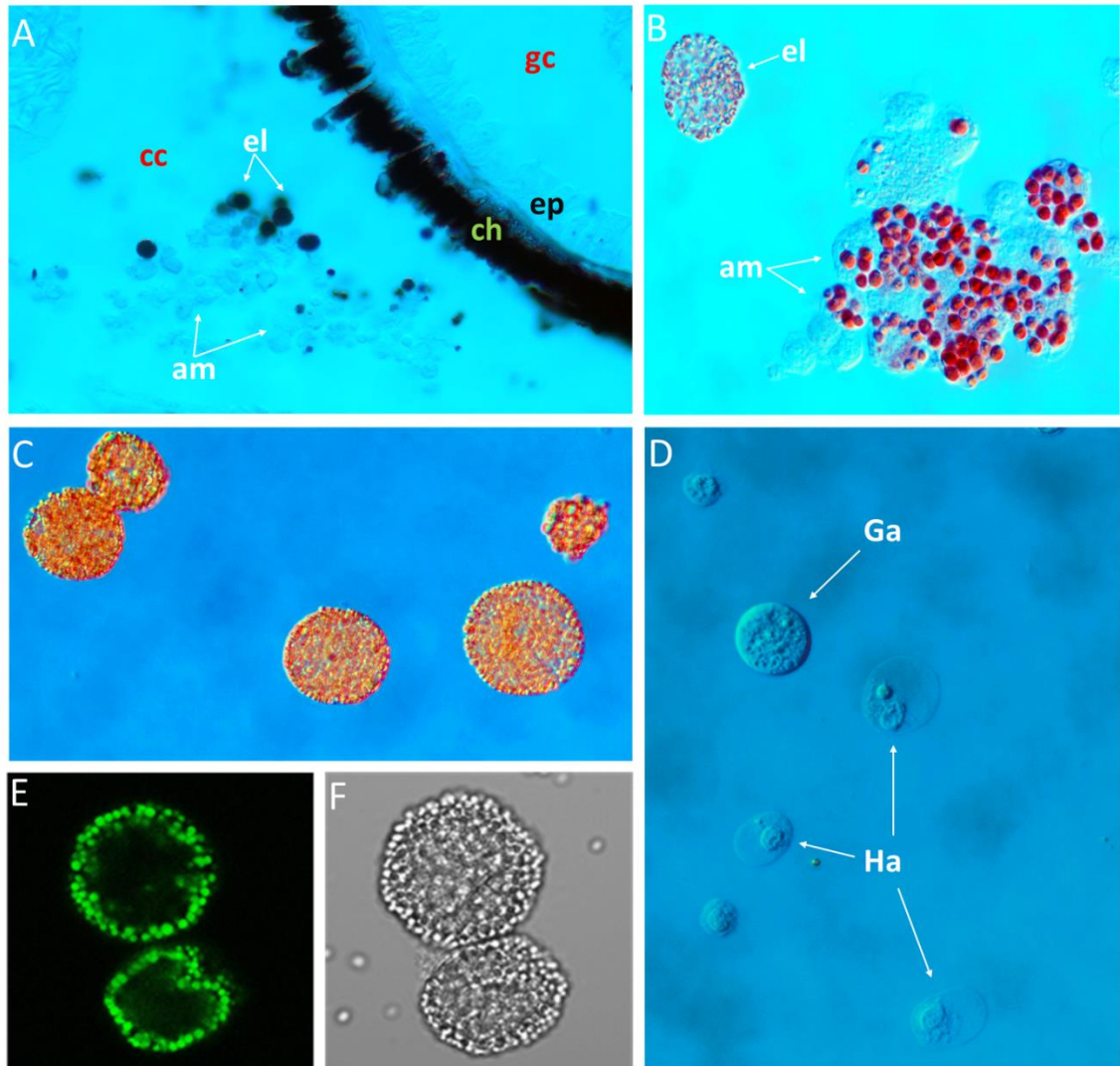


Figure 3: Light microscopic pictures of the three major free-floating cell types of the *Eisenia fetida* coelomocyte population. Panel (A) represent a cross-section of *Eisenia fetida* stained with Perl's reaction (cc: coelomic cavity, el: eleocytes, am: amoebocytes, ch: chloragogen tissue, ep: midgut epithelium, gc: gut cavity). eleocytes shown by panel (C, E, F) are easily recognisable by their morphological characteristics as well as their autofluorescence (E) caused by the high riboflavin content of their granules. Phagocytic, but non-autofluorescence amoebocytes are showed in panel (B) while panel (D) represents an example of Hyaline (Ha) and Granular (Ga) amoebocytes.

1.3.3 Humoral immunity

1.3.3.1 Pattern recognition receptors

One of the most important elements of the earthworms' humoral immune system is the germ line-encoded pattern recognition receptors (PRRs) which are capable of initiating the immune system responses (Prochazkova et al. 2019b). These receptors are able to recognise the different conserved pathogen-associated molecular patterns (PAMPs)

from both endogenous and exogenous sources (Škanta et al. 2013). Although a relatively low number of PRRs have been described in the Annelida phylum, according to their structural properties and functional domains these receptors can be subdivided into different groups: Toll-like receptors (TLRs) with a high number of leucine rich repeats (LRR), Peptidoglycan recognition receptors (PGRPs), C-type lectins, Nucleotide-binding oligomerisation domain (NOD)-like receptors (NLRs), and different scavenger receptors (SRs) (Engelmann et al. 2016b).

In contrast to other invertebrate groups, PRRs are described relatively poorly in Annelids and this is especially true in the case of earthworms. Coelomic cytolytic factor (CCF) was the first unique type of PRR which was observed in *Eisenia fetida* (Bilej et al. 2013). CCF protein contains several binding domains such as peptidoglycan, β -1,3-glucan and lipopolysaccharide (Bilej et al. 1998). Due to these attributes, CCF can recognise a wide range of microorganisms. Later other CCF homologous pattern recognition receptors have been described for some other annelid species, for example, *Lumbricus terrestris* and *Lumbricus rubellus* (Šilerová et al. 2006).

One of the most researched PRR groups (especially in the case of Annelids) is the Toll-like family (Johnson et al. 2003). First, the toll receptor protein was observed in *Drosophila*, in which it has been proved to participate in the embryogenesis and play a key role during the innate immune response in the case of fungal infection by enhancing the synthesis of antimicrobial peptides (Belvin and Anderson 1996). Following the early identification of the protein immune function, a whole family of Toll-like receptors were characterised from a wide spectrum of organisms from plants to humans (Roach et al. 2005). Toll-like receptors are well conserved transmembrane proteins containing three important functional domains. The extracellular part responsible for the ligand-binding using leucine rich repeats (LRR), the importance of the short transmembrane domain is to anchor the receptor to the cell or intracellular membrane. While the intracellular part always contains a Toll/IL-1 domain which is responsible for the initiation of the downstream signalling pathways of inflammatory cytokines when the ligand is bound (Jin and Lee 2008). The first evidence of the presence of TLR in Annelids appeared in *Capitella* and *Helobdella* during genome studies (Davidson et al. 2008). Later, a European

research group successfully cloned diverse coding sequences of TLRs from *Eisenia andrei* coelomic cells (Figure 4) (Škanta et al. 2013).

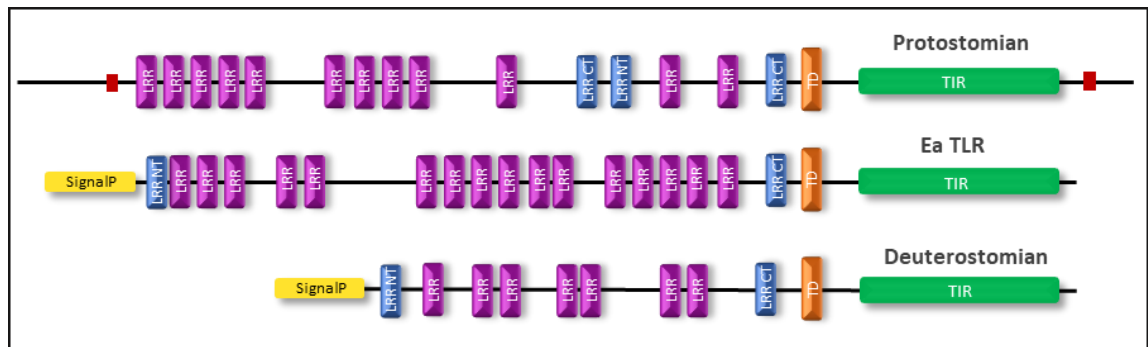


Figure 4: Simplified structure of the protostomian and deuterostomian Toll-like receptors, compared to the *Eisenia andrei* toll-like receptor (Ea TLR). SignalP: signal peptide, LRR: leucine-rich repeat, TIR: Toll/Interleukin-1 receptor homology domain, TM: transmembrane domain, LRR-NT/LRR-CT: leucine-rich repeat N/C- terminal domain, red bars: low-complexity regions (Škanta et al. 2013).

1.3.3.2 Sphingomyelin binding protein family

As mentioned before the coelomic fluid is rich in diverse bioactive molecules. Nonetheless the knowledge of the exact nature and functions of these molecules is still only partly understood. In earthworms, one of the most typical group of proteins which have an important function in the immune process is the Sphingomyelin binding protein family (Shakor et al. 2003). This family consists of fetidin, lysenin, and lysenin-related proteins. These molecules share strong molecular homology and it is known they have hemolytic, cytotoxic and smooth muscle contraction activities. Some studies managed to prove the bacterial challenge of coelomocytes could trigger the expression of lysenin (Czuryło et al. 2008). However, only infection with gram-positive bacteria caused increased expression of this family of molecules (Swiderska et al. 2016).

1.4 Immune system modulated by Nanoparticles

1.4.1 Definition and origin of Nanoparticles

Naturally occurring nanoparticles (NPs) have been surrounding us throughout human history. However, due to the lack of applicable detecting technologies, their scientific discovery is only considered to be part of the twentieth century (Heiligtag and Niederberger 2013). A nanoparticle can be defined as a wide ranging particle of matter

which has at least one dimension with less than 100 nm, associated with specific physicochemical characteristics which are not shared with non-nano scale particles created from material with the same chemical composition (Auffan et al. 2009, Strambeanu et al. 2015b). In general, they are highly reactive and physiochemically dynamic molecules due to their high surface-to-volume ratio (Auffan et al. 2009). NPs can be classified according many different criteria such as their origin (anthropogenic, natural), chemical composition (inorganic, organic) and size. Naturally occurring nanoparticles can be originated from several natural sources such as volcanic eruptions, wildfires, interstellar and desert dust, and from diverse microbial processes (Figure 5). Anthropogenic nanoparticles can be produced in different ways. Engineered nanoparticles are intentionally produced by modern laboratory synthesis, in general, they have precisely selected shapes, sizes, and chemical compositions (Shang et al. 2014b). Based on their chemical compositions engineered NPs can be classified within three major categories, which are the organic, inorganic and carbon based NPs. These types can be subdivided into several subgroups based on their physicochemical characteristics (Mauricio et al. 2018, Ealia and Saravanakumar 2017). A simplified classification of manufactured NPs alongside their typical size ranges and example applications are shown in Table 1. In contrast, "incidental" NPs are the side effects of various human activities (Figure 5). In contrast to engineered NPs these have highly variable chemical and physical properties (Strambeanu et al. 2015a). Three different layers can be distinguish in engineered nanoparticles, the surface layer, shell layer and the core of the nanoparticle (Khan et al. 2019).

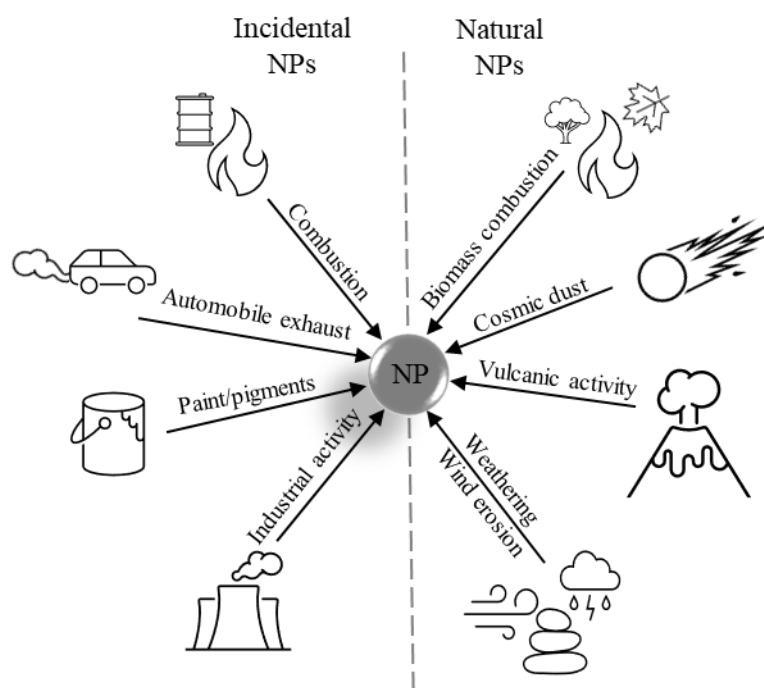
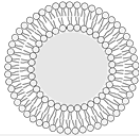
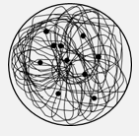
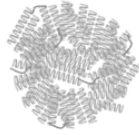
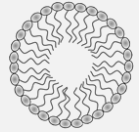




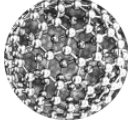
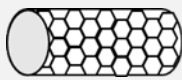
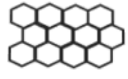


Figure 5. Schematic illustration of processes leading to the formation of some of the main groups of incidental (induced by different human activities) and natural nanoparticles (Sharma 2015).

Table 1: Classification and example applications of engineered nanoparticles. Based on their chemical composition nanoparticles can be divided between three major group (Organic, Inorganic, Carbon based) which can be subdivided to different types.

Material	Type	Typical size range [nm]	Example applications	Schematic Illustrations
Organic	Liposomes	50-1000	carriers, cosmetics, biomedicine	
	Polymers	20-250	drug delivery	
	Ferritin NPs	12nm	drug delivery, bioassay, molecular imaging	
	Micelles	5-100	solubilize poorly soluble drugs, gene delivery,	
Inorganic	Metallic	1-100	anticancer agents, imaging techniques	
	Metal oxide	10-100	sunscreen, antimicrobial agents, electronics	
	Ceramic	<50	bone repair, electronics	
	Quantum dots	2-8	solar cells, LEDs, medical imaging	
Carbon based	Fullerenes	100-400	lubricants, drug delivery	
	Carbon nanotubes	1-2	reinforcing structural components, energy storage, molecular electronic	
	Graphene	<50	batteries, biomedical sensors, solar cells	

1.4.2 Applications and environmental risks

Applications of nanotechnology are resulting in major innovations in biology, medicine, and industry (Salata 2004). The diversity of the nanomaterials being exploited, both in size as well as chemical composition, is extensive and when combined with surface derivatisation gives rise to an exponentially growing portfolio of nanoparticles. Classification and characterisation approaches of these synthetic particles are now well established (Nagarajan 2008). Despite, the potential extensive use, the diversity of nanomaterials and the associated environmental release, the effect of NPs on ecosystems is only modestly understood. Even though the number of publications related to NPs has been increasing, a major issue stems from safety concerns associated with the dearth of information in relation to nanoparticles' toxicity as this related to production process safety, safe use and environmental fate. One of the most critical areas that require investigation is the interaction between the components of the immune system with different types of NPs (Zolnik et al. 2010). The recent increase in the field of industrial nanomaterial production caused relatively high environmental release of these particles (intentionally or unintentionally) as well as their derivatives (Simonet and Valcárcel 2008). Due to the intimate relationship with their habitats, most invertebrate species are potentially vulnerable to the environmental release of NPs (Tourinho et al. 2012, Baun et al. 2008).

1.4.3 Nanomaterials and Innate Immunity

To understand the potential health risks of nanomaterials (NMs) it is essential to know more about their possible interaction with the innate immune system. Observing immunity can be exploited as an indicator to monitor the overall well-being of the organism under sublethal exposure to NPs (Hayashi and Engelmann 2013). It is also important to understand how nanoparticles modulate the inflammatory processes at different levels of organisation, this understanding provides new opportunities in the field of engineered nanomedical products (Kim et al. 2010). Geochemically derived nanoparticles are an intrinsic element of our environment and have been present throughout the evolutionary development of the immune system. Due to the almost continuous exposure to NPs the immune system has learned to recognise and eliminate these foreign substances. Engineered NPs do not differ substantially from naturally

occurring counterparts, however, they still could pose a potential immune risk by challenging these elegantly evolved mechanisms by delivering new chemical ‘types’ of NPs representing a novel combination of shape, size, chemical composition, and surface charge (Boraschi et al. 2017).

Innate immunity represents the first line of defence for vertebrates whilst invertebrates rely exclusively on this response to defend against pathogen attack, it is therefore concerning that innate immunity has the potential to be highly affected by NPs. The type of interaction between the immune system and NPs can be highly dependent on the physicochemical characteristics of the NPs (Liu et al. 2017). It is known that different PRRs, like the Toll-like receptor family, can be activated by a range of metal nanoparticles (MNPs) initiating inflammation processes (Smith et al. 2013). Other NPs may enter immune cells through hijacking the endocytic system that is a central component of the innate immune response (Shang et al. 2014a). This appropriation can have a particular effect on the downstream immunological cell-cell signalling processes. Due to the surface chemistry of many NPs they tend to agglomerate and create aggregates. First, when NPs come in contact with different body fluids they can react with protein and other biomolecules to form different biomolecule-nanoparticle complexes (Lundqvist 2013). It is known that an array of different types of biomolecules could bind to nanoparticles spontaneously which result in the formation of a biomolecular corona (Casals and Puntès, Lundqvist et al. 2008). This process is an active area of bionanoscience research but most studies observe the interaction between NPs and different mammalian plasma proteins (Tenzer et al. 2013). Far fewer studies have investigated the interaction of invertebrate immune proteins and NPs.

1.4.4 Nanoparticles and earthworm immunity

The *Eisenia fetida* celomic secretory protein lysenin, a multifunctional component of the innate immune system, has been revealed to perform yet another role in immunity forming a biomolecular corona around nanosized silver particles (Hayashi and Engelmann 2013). Subsequent to this observation additional research groups have endeavoured to clarify the underlying mechanism of biomolecule-nanoparticle complex formation and the contribution of lysenin to this process (Engelmann et al. 2016b). The results of these experiments show that lysenin is not absorbing to silver nanoparticles

purely by its relative abundance but rather by an unidentified property that makes the lysenin more suitable to adsorb to the Ag-NPs surface (Hayashi et al. 2013). Currently the exact mechanism behind this selective binding is unresolved (Bodó et al. 2020).

Earthworms can opsonise different bacterial surfaces utilizing the extracellular receptor protein coelomic cytolytic factor (CCF) (Bilej et al. 1998). Additional interplay between innate immune system of NPs was revealed when CCF was also observed to be deposited onto the NPs surface. The adhesion capability - which was shown by CCF - was an early example of pattern-independent opsonisation of nonbiological particles (Engelmann, Hayashi et al. 2016). This study also reported that coronas formed by secretory proteins from coelomocytes on silver NPs caused significantly more NP accumulation in coelomocytes than when the coronas were formed by FBS (fetal bovine serum proteins). Furthermore, the accumulation was even greater in the case of eleocytes compared to amoebocytes when they used recombinant lysenin for corona forming. Based on these findings they suggested that lysenin is involved in opsonisation-induced cellular interactions and which could show differences in the case of amoebocytes and eleocytes (Engelmann et al. 2016b).

1.5 The Aims of this Thesis

The thesis was designed to enrich our understanding of the interaction between MNPs and the innate immune system, using earthworms as a model organism. It was clear that it was necessary to describe the components of the innate immunity related pathways in our target organism, develop resources, and methodology that would allow us to dissect the interaction between MNPs and the Innate Immune system (IIS).

1.5.1 Chapter 2 - Transcriptomic Tissue Atlas of Earthworm Immunity

Since each tissue can play a distinct role in the immune response process, this chapter aimed to identify immune-related genes in our model organisms and determine a tissue-specific expression profile for innate immune genes. Furthermore, deriving a transcriptome directly from the coelomic fluid, which contains mixed populations of immune cells, would result in better coverage of immune genes than could be derived from whole-body transcriptomes. The tissue-specific approach also allowed us to observe the expression of the immune components between different tissues, and it could deliver an excellent reference for the later experiments focussed on earthworm immunity and its modulation by MNPs exposure.

1.5.2 Chapter 3 – Genomic template to innate immunity

Establishment of a reference genome was required to further polish the results of the tissue-specific transcriptomes. A reference genome can be utilise to resolve problems associated with *de novo* transcriptomics pipelines, such as the recovery of gene isoforms in the case of expanded gene families associated with innate immune system, reduce the number transcripts created by assembly artefacts and to be able to explore splice variants. Therefore the aim of this chapter was the create a new, highly contiguous and complete *de novo* reference genome for *E. fetida*, using the advantages of the recently developed long read sequencing platforms

1.5.3 Chapter 4 – Characterisation of the coelomic fluid

Since cell-cell interactions play a key role in immune system processes, this chapter aimed to provide a coelomocyte population-specific frame for the main components of the earthworm innate immune system, on a transcriptomic level. This aim involves two steps: First, to achieve a well-dispersed, single-cell suspension from the extracted

coelomocytes, a development of a new cell preparation protocol was required, which allowed a clear separation between the three major coelomocyte populations. The coelomocytes then can be separated based on the differences in their light scattering characteristics (FACS). Finally, after solving the methodological challenge, the analysis of RNA-Seq data generated from the individual cell-types could help to identify the specific immunological roles associated with the different coelomocyte populations.

1.5.4 Chapter 5 – Spatio-temporal response to challenge

The aim of this chapter was to trace the direct impact of the MNPs exposure on the earthworm immune cells, in both a spatial and temporal manner. To deliver a spatially and temporally resolved cellular response the resources generated in earlier chapters were used to design an experiment where the coelomocyte-specific temporal response of *E. fetida* was followed in both presence and absence of CuNPs in their coelomic cavity.

1.5.5 Chapter 6 - Effects of indirect MNPs exposure

Since the chosen experiment design in Chapter 5 only allowed us to follow the direct effect of the injected NPs, the aim of this chapter was to compare these results to a ecologically more realistic scenario. For this reason, earthworms were exposed to Cu ions and CuNPs using soil as an exposure medium. This not only created the possibility to compare the effects caused by the direct and indirect exposure, but identify the possible differences in the biological response to the ionic copper and CuNPs.

2 Transcriptomic Tissue Atlas of Earthworm Immunity

2.1 Introduction

One of the major challenges in studying the immune system of invertebrate species is the limitation in the number of known immune-related genes. This scarcity of information has two profound consequences: the incomplete description of the pathways involved restricts interpretation of immune response, and a lack of quantitative methods and functional assays to explore the inflammatory processes. Although *Eisenia fetida* is one of the keystone model species within the *Lumbricidae* family, the limited transcriptomic and genomic resources mean molecular immunology research has focused on a handful of genes assumed to have immune roles, usually based on functional extrapolations from higher-order vertebrates. Consequently, generating a more complete picture of the earthworm immune system has significant potential to more fully describe the immune components with their phylogenetic conservation and identify any pathways.

Over the past decade, the evolution of high-throughput technologies such as RNA-sequencing has enabled the cost effective sequencing of hundreds of thousands of transcripts simultaneously (Kukurba and Montgomery 2015). Parallel improvements in the field of *de novo* transcriptome assembly pipelines has almost eliminated the necessity for a reference genome to reconstruct transcriptomes *in silico* and also allows the measurement of transcript abundance differences between biological samples.

Despite the availability of RNA-Seq technology, there is still a relatively high number of poorly represented taxonomic groups. Although several species from the Annelida phylum are routinely used as ecotoxicological sentinels, their genomic and transcriptomic resources are less well established compared to other invertebrate phyla, such as the Arthropods or Molluscs. This tendency for under-representation is also observable by analysing the number of available NCBI Entrez records belonging to specific invertebrate phyla. Currently (December 2020) there are around 1,300,000 nucleotides and 140,000 protein sequences available for the whole Annelida phylum, from which only approximately 10,000 nucleotides and 2,400 protein sequences belong to the *Eisenia* genus. Although the expansion of isoforms and splice variants represents an important aspect of innate immunity by increasing the diversity of the proteome and providing the possibility to create a wide set of transcript isoforms even from a single

gene (Giblin et al. 2020, Anderson 2000), the knowledge about these mechanisms also highly limited in many invertebrate, non-model organism.

2.2 Materials and Methods

2.2.1 Biological material

To reduce the intraspecific genetic variability, earthworms were supplied from the monophyletic laboratory cultures from the UK Centre of Ecology and Hydrology (Wallingford, UK). The sacrificed individuals were genotyped to ensure that only earthworms from the same lineage were further processed.

2.2.2 Genotyping

Tissue (25 mg) was collected from earthworms by cutting off the last few tail sections, then DNA samples were extracted using the DNAeasy Blood and Tissue Kit (QIAGEN Ltd, Manchester, UK) and following the manufacturer's instructions. A fragment of the mitochondrial Cytochrome C Oxidase subunit (COI or COXI) gene was amplified, amplicon purified (QIAquick PCR Purification Kit, QIAGEN Ltd) and sequenced in both forward and reverse directions by Sanger sequencing performed by Eurofins Genomics (Germany). Amplification was performed on a VeritiPro Thermal Cycler (ThermoFisher Scientific Inc, UK) using Promega GoTaq PCR Mix (Promega UK). Primer sequences (Folmer et al. 1994) are given in Table 2A, composition of the PCR reaction is provided in Table 2B and amplification conditions summarised in Table 2C.

Table 2: Primers and PCR conditions used for genotyping earthworm species. Primer sequences are given in Panel 2A, composition of reaction components are provided in Panel 2B and amplification conditions are given in Table 2C.

Table 2A

Name	Sequence	Origin
Universal COI F primer	5' - GGTCACAAATCATAAAGATATTGG -3'	(Folmer, Black, Hoeh, et al. 1994)
Universal COI R primer	5'- TAAACTTCAGGGTGACCAAAAAATCA - 3'	(Folmer, Black, Hoeh, et al. 1994)

Table 2B

Buffer		10 µl
MgCl ₂		3 µl
dNTPs		0.5 µl
Forward primer (10 uM)		1 µl
Reverse primer (10 uM)		1 µl
Taq polymerase (5 U)		0.25 µl
Molecular grade H ₂ O		32.25 µl
Template		1 µl
Total		50 µl

Table 2C

95°C	2 minutes	
94°C	30 seconds	35 Cycles
45°C	30 seconds	
72°C	60 seconds	
72°C	10 minutes	

2.2.3 Experimental design and sample preparation

Tissue-specific transcriptomes were generated for six tissues (Cf: Coelomic fluid, Cr: Crop, GCh: Gut+Chloragogen, Gi: Gizzard, Nc: Ventral nerve-cord, Ph: Pharynx) by pooling RNA extracted from six genotyped individuals of *E. fetida* and separately *E. andrei*. After precise dissection tissues were homogenised in chilled TRIzol reagent (Zymo Research, CA, USA) individually and stored at -80°C. The RNA was extracted using

Direct-zol RNA Miniprep procedure (Zymo Research, CA, USA) according to the manufacturer's protocol. The quality of the RNA samples was measured used by 4200 TapeStation System (Agilent Technologies LDA UK Limited, Stockport, UK) with RNA ScreenTape (Agilent), then the quantities were verified by fluorometric quantification using Qubit Flex Fluorometer using high sensitivity Qubit reagents (ThermoFisher Scientific Inc, UK). cDNA libraries were generated from approximately 0.4 µg RNA for each library, using the TruSeq Stranded mRNA Sample Preparation LS kit (Illumina, CA, USA) according to the manufacturer's instructions. Transcriptome sequencing was conducted on an Illumina NextSeq 500 system using both a medium output and high output flow cell. Operation of the sequencing platform was performed by Ms Angela Marchbank, Cardiff School of Biosciences Genomics Hub, Cardiff University.

2.2.4 Transcriptomic analysis

Quality filtering and adaptor removal was performed with Trimmomatic (v0.39) (Bolger et al. 2014). Paired-End (PE) reads from the different tissue samples were individually assembled into transcriptomic contigs using the non-genome guided *de novo* assembly pipeline and the Trinity software package (v2.5.0) using the default settings. (Grabherr et al. 2011b). A combined assembly was generated using the combined data from all tissue sampled from a single species. Thereafter, the composite assembly and assemblies from different tissues were concatenated together and used as an input to the EvidentialGene:tr2aacds (v20mar15) (Nakasugi et al. 2013) mRNA Transcript Assembly Software to remove transcript redundancy and quality filter the final transcriptome (Gilbert 2013) . We used BUSCO (v4) (Waterhouse et al. 2018) software with the Metazoan dataset (metazoa_odb10) to quantify transcriptome completeness based on evolutionary-informed expectations of gene content from near-universal single-copy orthologs (Simao et al. 2015).

2.2.5 Tissue differential expression analysis

Trimmed short reads were mapped to the transcriptome assembly using the built-in RSEM-Bowtie2 module of the Trinity (v2.5.0) software package (Li and Dewey 2011). In the case of 75 bp read-length data, an increased multimapping rate was observed (Figure 6). For this reason only the data with 150bp average read length was utilised for counting purposes. To calculate the relative expression profiles counts were normalised

using weighted trimmed mean of M-values (TMM). Differential gene expression was determined between tissue samples using normalised counts using the NOIseq module of the Omicsbox (Biobam) software package (Tarazona et al. 2015, BioBam 2019). NOIseq analysis was chosen due to its capability to analyse the dataset without the presence of true technical replication by simulating technical replicates, an approach that is valid since each library was made from a pool of individuals. To exclude genes with low counts across the different libraries an additional count-per-million (CPM) based filtering was conducted with an applied cut-off value of 1 CPM. Based on this method we removed any transcripts from the analysis which showed lower expression than 1 CPM in all of the libraries. The number of simulated replicates were set to 5 with a size of 20% of the original datasets. Only transcripts with at least 0.9 probability were considered as differentially expressed. Characteristic (“tissue-specific”) transcripts of the different tissues were identified by filtering for differentially expressed transcripts that are upregulated in certain tissue relative to others.

The conducted bioinformatic pipeline from the raw-read pre-processing to differential expression analysis is summarised on Figure 6.

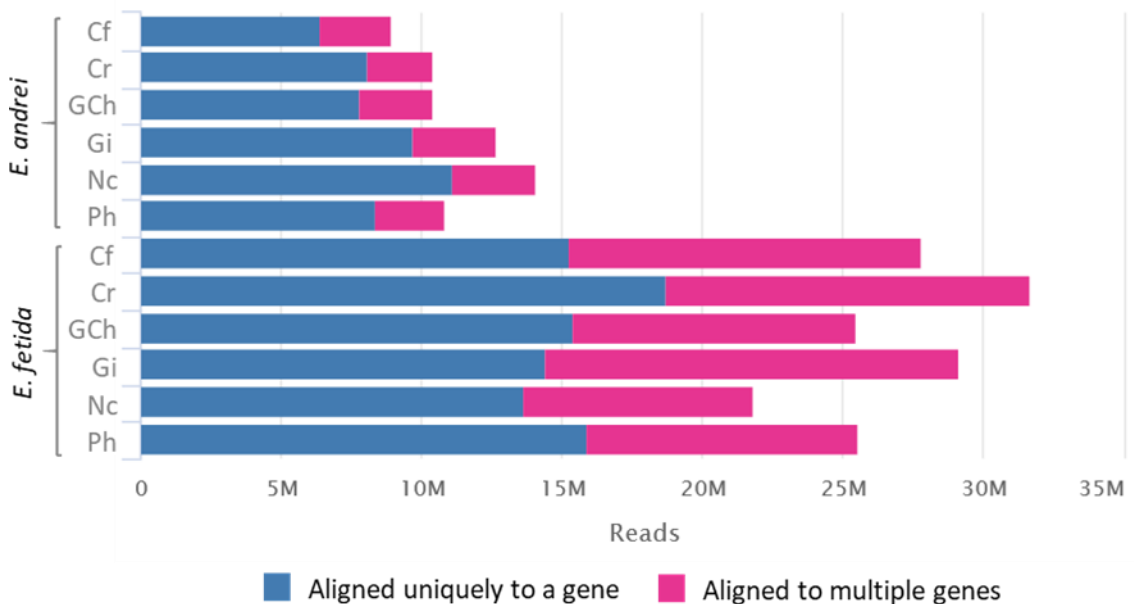


Figure 6: Number of uniquely and non-uniquely mapped reads in the case of different earthworm tissues (Cf: Coelomic fluid, Cr: Crop, GCh: Gut+Chloragogen, Gi: Gizzard, Nc: Ventral nerve-cord, Ph: Pharynx).

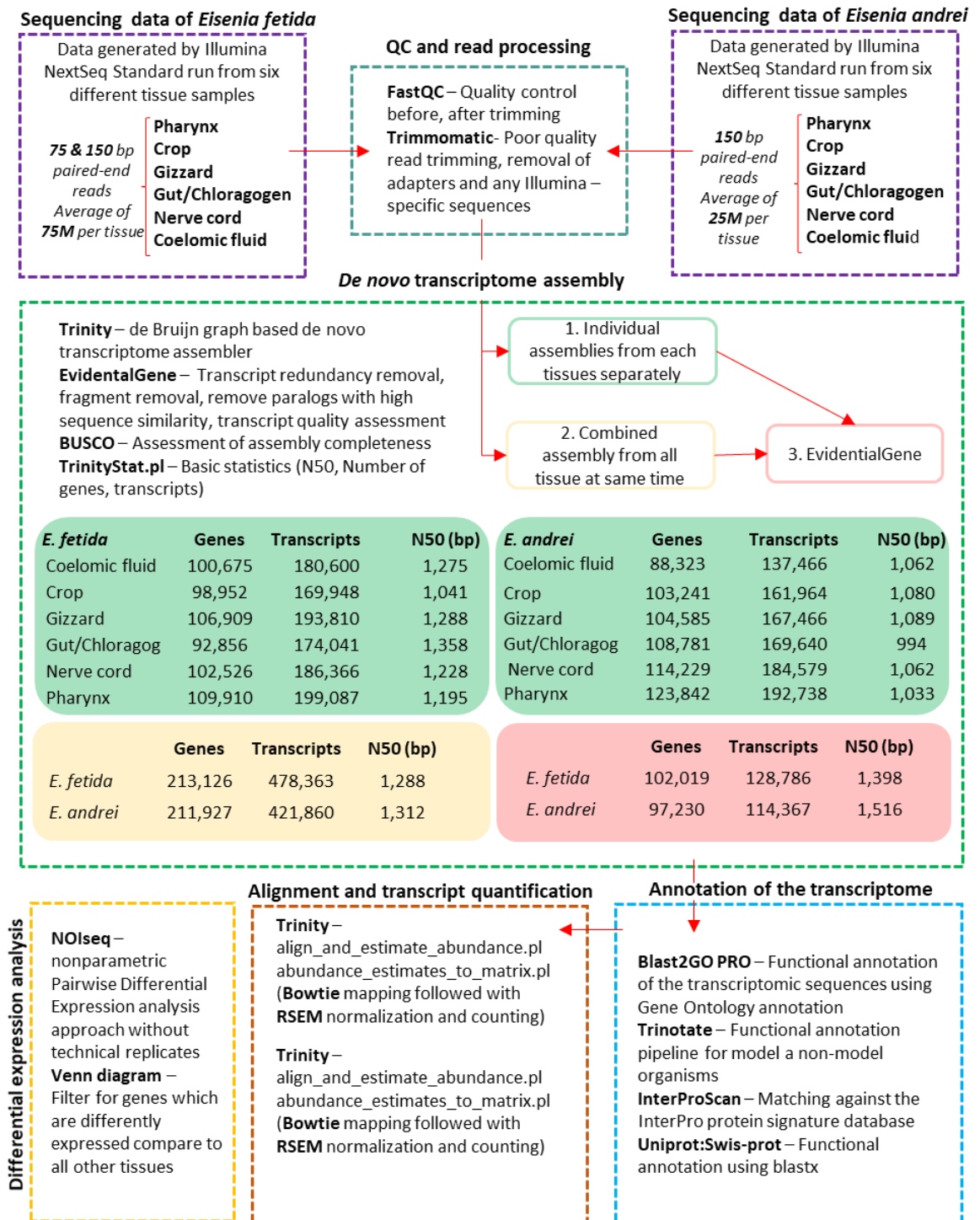


Figure 7: Bioinformatic workflow employed for transcriptome assembly and analysis. Major components of the analysis bounded by uniquely coloured, dashed lines.

2.2.6 Annotation

To annotate the transcripts and identify the putative immune genes, we used the Functional Annotation option of the Blast2GO and Trinotate programmes (Götz et al. 2008) (Bryant et al. 2017). To identify further genes, an additional annotation was derived based on homology analysis with BLASTx against the Uniprot:Swis-prot database (UniProt 2019). BLASTx analyses were performed using the BLOSUM 62 matrix with a gap cost of 11 and word size of 3 (McGinnis and Madden 2004) using the *Homo sapiens* (proteome:UP000005640), *Mus musculus* (proteome:UP000000589), *Drosophila melanogaster* (proteome:UP000000803), *Caenorhabditis elegans* (proteome:UP000001940), *Saccharomyces cerevisiae* (proteome:UP000002311) datasets. Only hits with an E-value lower than 1E-10 and a bit score higher than 30 were approved for annotation.

2.2.7 Gene enrichment analysis

Over-representation analysis was performed on the identified tissue-specific gene lists individually. Enrichment analysis was performed using the g:GOST functional profiling module of the gProfiler software (Raudvere et al. 2019), utilising the Gene Ontology (GO) (The Gene Ontology Consortium 2018), KEGG (Kanehisa et al. 2015), WikiPathways (Slenter et al. 2017) and Reactome (Jassal et al. 2020) databases. Significance cut-off was set to 0.05 and p-values were corrected using the g:SCS method. In order to compute functional enrichments of tissue-specific gene lists, annotated genes from *E. fetida* and *E. andrei* were used as a background gene set. Following the enrichment analysis summarisation and redundancy filtering of the GO terms was conducted using the REVIGO software (Supek et al. 2011).

2.3 Results

2.3.1 Transcriptome assembly

In total, approximately 650 million Illumina PE reads were generated using six different tissues from *E. fetida* and *E. andrei* (Figure 8). In the case of both earthworms species the coelomic fluid (Cf), pharynx (Ph), crop (Cr), gizzard (Gi), ventral nerve cord (Nc), and gut/chloragogen (GCh) tissues were processed for sequencing. After removing the adaptor sequences and reads with low quality, in the case of *E. fetida* we successfully assembled more than 147,000 transcripts with an of N50: 1,386 bp while the *E. andrei* results more than 146,000 transcripts with an N50: 1,169 bp (Figure 9). To determine the 'completeness' of the *de novo* assemblies we assessed the presence of evolutionarily conserved single-copy orthologs using the BUSCO pipeline (Waterhouse et al. 2018). Each tissue provided a representation of between 60-80% of the conserved suite of genes. However, combining and removing redundant transcripts generated a composite transcriptome representing 94% complete (95% when including fragmented genes) for *E. fetida* and 96% complete (97% when including fragmented genes) for *E. andrei* (Figure 10).

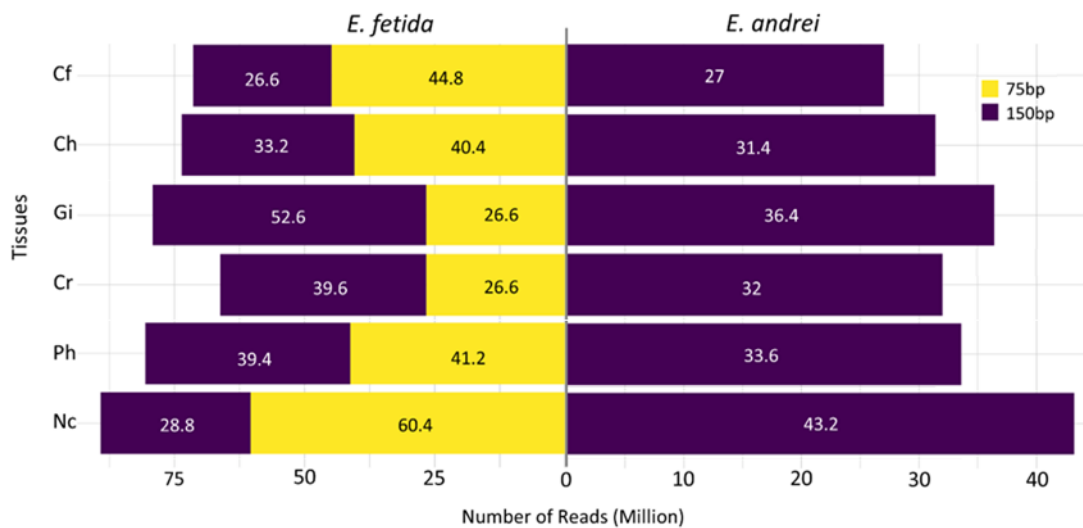


Figure 8: Sequencing depth for libraries used to derive tissue transcriptome atlases for *E. fetida* & *E. andrei*.

<i>Eisenia andrei</i>	Genes	Transcripts	N50 of Genes	N50 of Transcripts
Tissues combined + Evidential gene	138,979	146,524	1,169	1,123
Tissues combined	643,001	1,013,853	1,052	873
Coelomic fluid	88,323	137,466	1,062	823
Gut/Chloragog	108,781	169,640	994	887
Gizzard	104,585	167,466	1,089	913
Crop	103,241	161,964	1,080	881
Pharynx	123,842	192,738	1,033	901
Nerve cord	114,229	184,579	1,062	841
<i>Eisenia fetida</i>				
Tissues combined + Evidential gene	131,251	147,067	1,386	1,425
Tissues combined	611,828	1,103,852	1,234	1,036
Coelomic fluid	92,856	174,041	1,358	1,130
Gut/Chloragog	98,952	169,948	1,041	854
Gizzard	102,526	186,366	1,228	1,027
Crop	100,675	180,600	1,275	1,069
Pharynx	106,909	193,810	1,288	1,138
Nerve cord	109,910	199,087	1,195	1,000

Figure 9: Statistical metrics of de novo assemblies for tissue-specific and composite transcriptomes for *E. fetida* and *E. andrei*.

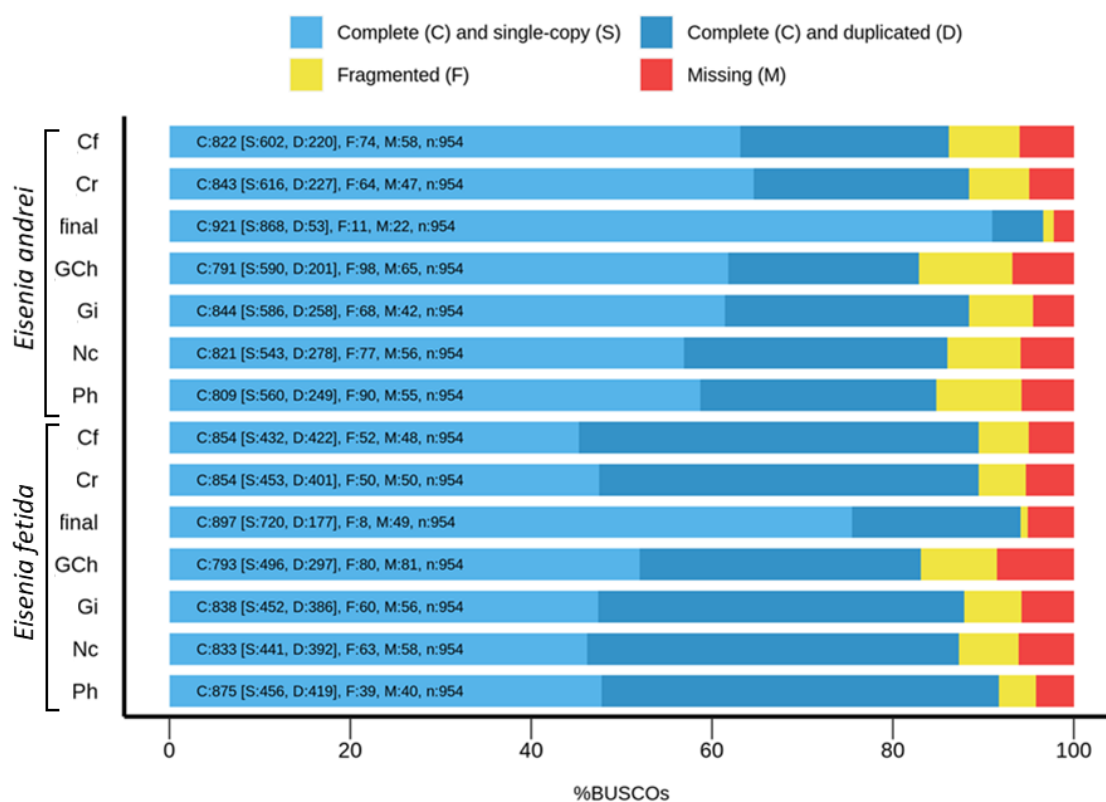


Figure 10: Completeness analysis of *de novo* assemblies of tissue-specific and composite transcriptomes from *E. fetida* and *E. andrei*. Single-copy orthologs (BUSCO v4) based transcriptome completeness benchmarking for the different *E. andrei* and *E. fetida* transcriptomes. Two/three-letter codes represent assemblies using single-tissue assemblies (Cf- Coelomic fluid, Cr – Crop, Gi – Gizzard, Nc – Nerve cord, Ph – Pharynx, GCh – Gut/Chloragogen) while “final” represent the final composite and redundancy filtered transcriptomes.

2.3.2 Transcriptome annotation

To annotate the *de-novo* reference transcriptome with a high functional annotation success rate, different annotation methods were employed. Initially, a simple, homology analysis based annotation (BLASTx) was performed using the Uniprot:Swis-prot database (UniProt 2019) In total, datasets from five different species, where the complete proteome was available, were analysed which were *Homo sapiens* (proteome:UP000005640), *Mus musculus* (proteome:UP000000589), *Drosophila melanogaster* (proteome:UP000000803), *Caenorhabditis elegans* (proteome:UP000001940), *Saccharomyces cerevisiae* (proteome:UP000002311). The highest BLAST success rate was observed when sequences from the *H. sapiens* UniProt

proteome was used as query set. Although using the other four species generated lower overall numbers each of them resulted in several unique annotations. Following the creation of a non-repetitive annotation set by combining blast results from the different reference species, we could annotate more than 30,000 transcripts in the case of *E. fetida*. However, the equivalent number was only slightly in excess of 22,000 when analysing the *E. andrei* transcriptome. More details about the number of unique annotation hits and about the identified overlap between them are reported in Figure 11 and Figure 12.

A secondary approach was conducted using a GUI based bioinformatics platform called Blast2GO (Conesa et al. 2005) The Blast2GO based annotation exploited a composite blastx result from the top 5 hits from the NCBI non-redundant database of proteins from all eukaryotes (threshold of 1E-10) combined with a function domain analysis based on InterProScan (v5.0) (Hunter et al. 2009). This approach yielded the highest number of successful annotations identifying more than 38,000 hits for *E. andrei* and approximately 48,000 hits when analysing the *E. fetida* transcriptome (Figure 13). Although Blast2GO was able to retrieve the highest number of successful blast hits, the output format highly limited the possible options for down-stream analysis. For this reason, all of the annotation-based further analyses were conducted using the UniProt database based annotations.

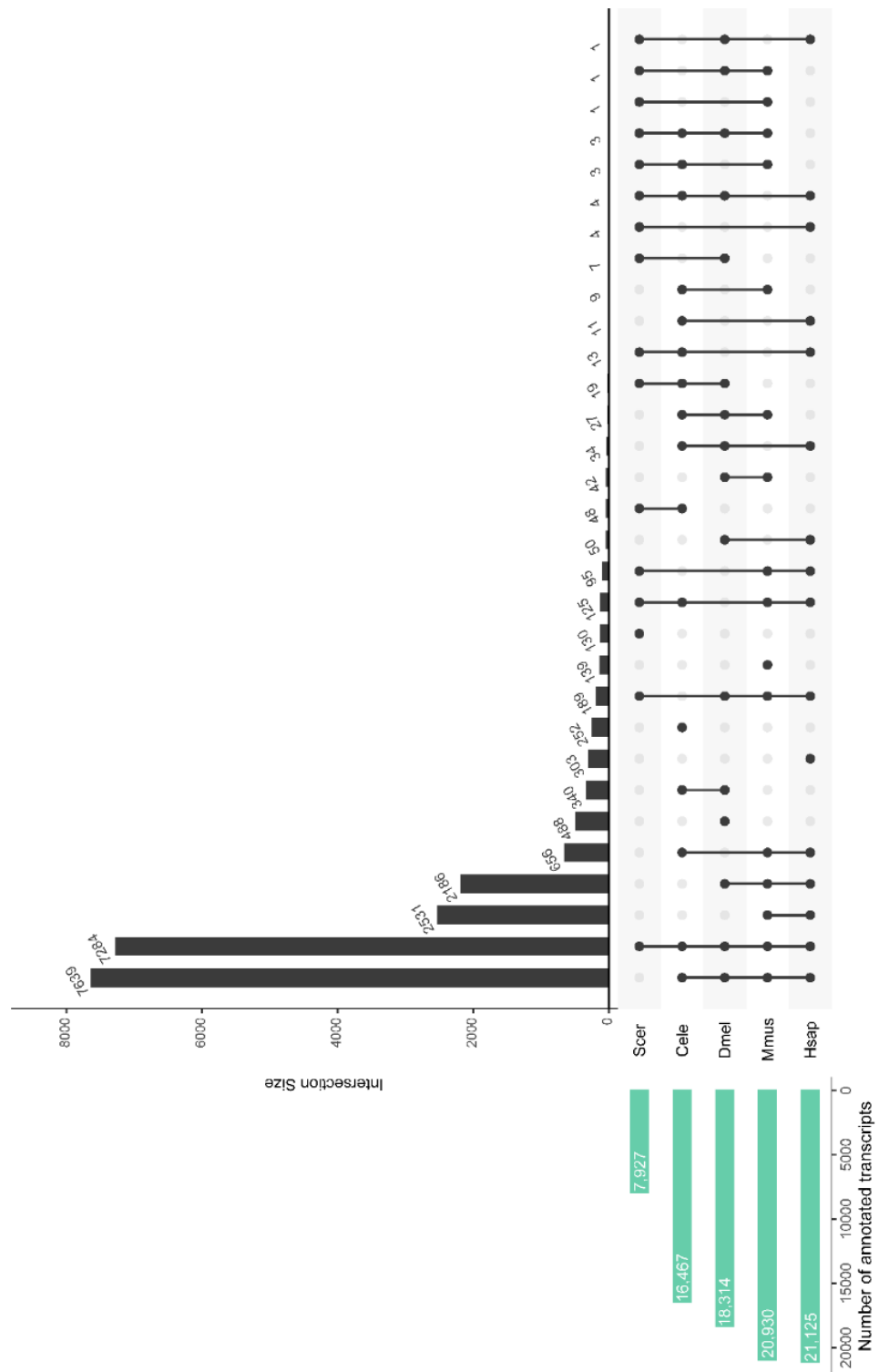


Figure 11: Evolutionary comparative proteome annotation of Annelida transcriptomes. An upset plot was generated where green coloured bars show the number of successful BLAST hits when the *E. andrei* transcriptome was used as reference set against five different species proteomes. The bars with grey colour represent the overlaps between the retrieved BLAST hits using different Uniprot datasets. Hsap: *H. sapiens* (UP000005640), Mmus: *M. musculus* (UP000000589), *D. melanogaster* (UP000000803), Cele: *C. elegans* (UP000001940), Scer: *S. cerevisiae* (UP000002311).

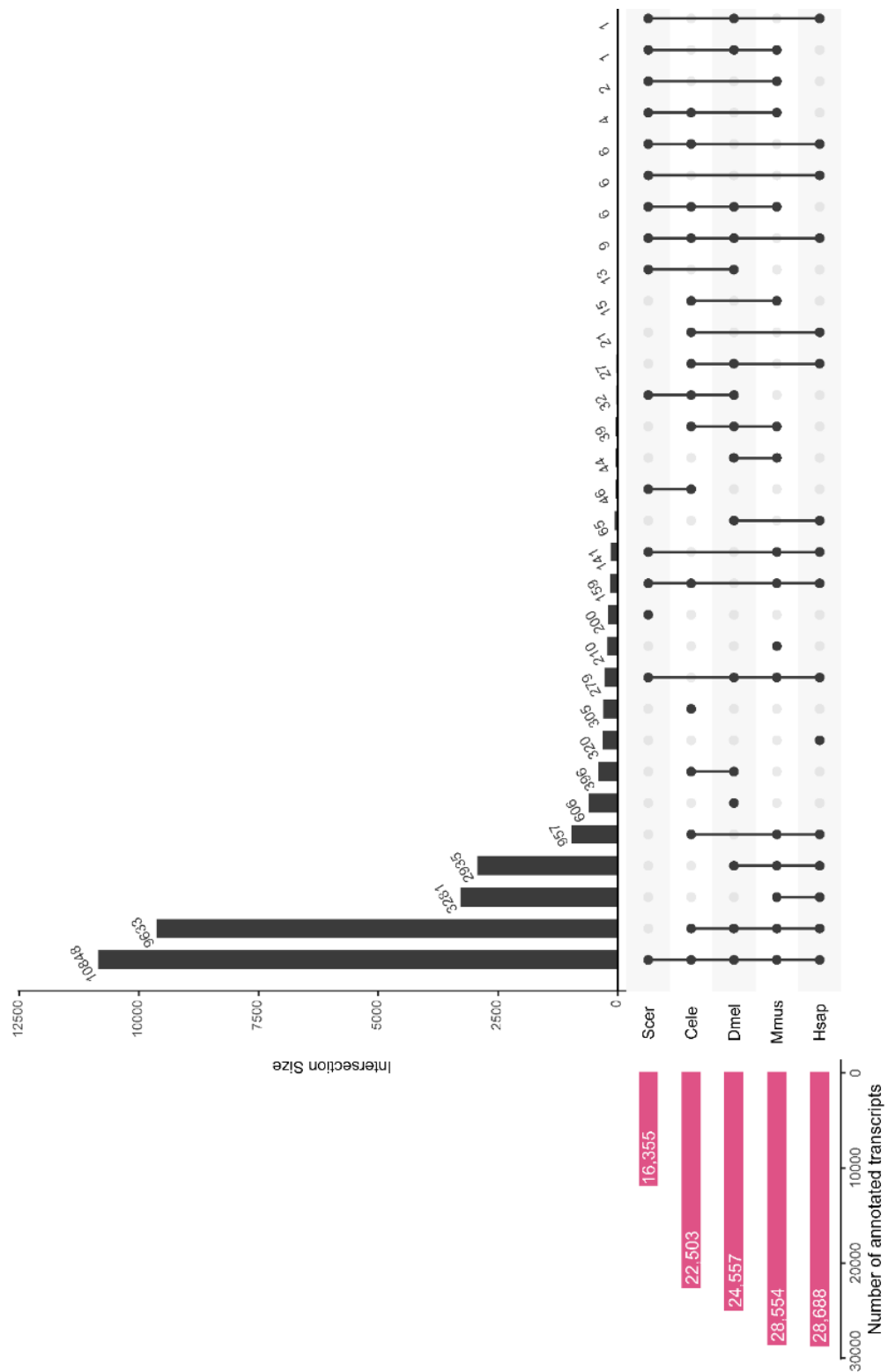


Figure 12: Evolutionary comparative proteome annotation of Annelida transcriptomes. An upset plot was generated where pink coloured bars show the number of successful BLAST hits where the *E. fetida* transcriptome was used as reference set against five different species proteomes. The bars with grey colour represent the overlaps between the retrieved BLAST hits using different Uniprot datasets. Hsap: *H. sapiens* (UP000005640), Mmus: *M. musculus* (UP000000589), *D. melanogaster* (UP000000803), Cele: *C. elegans* (UP000001940), Scer: *S. cerevisiae* (UP000002311).

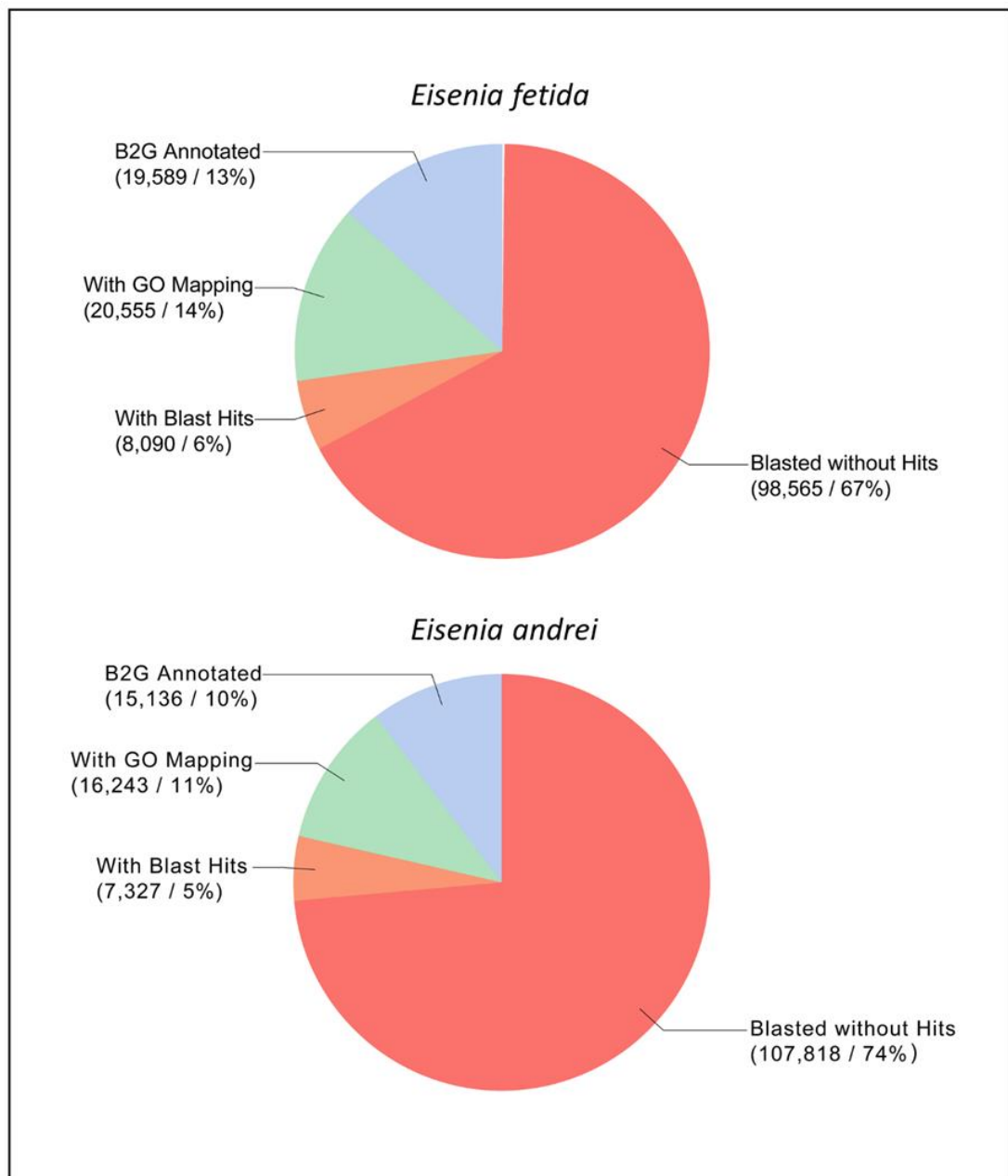


Figure 13: Annotation success rate of *E. fetida* and *E. andrei* transcriptomes using the Blast2GO annotation pipeline. (Mapped: pool GO terms could be associated with the hits earlier identified by BLAST analysis, Annotated: GO terms were successfully selected from the GO pool identified by the mapping step.)

2.3.3 Tissue functional profiling

Transcript expression values were normalised and measured in all six tissues using RSEM (Li and Dewey 2011) Next, normalised counts were used to perform differential expression analysis between the tissue samples and identify "tissue-specific" genes that were upregulated in specific tissues when comparing them to all others. The number of tissue-specific transcripts ranged between 747 and 4703, where the crop tissue

produced the smallest number, and the highest number of “specifically” expressed transcripts were observed in the nerve cord sample (Table 3). The identified tissue-specific transcripts then were used to characterise the fundamental functions of the corresponding tissues on the transcriptomic level, using functional gene enrichment analysis. We performed over-representation analysis retrieving data from the Gene Ontology (GO) (The Gene Ontology Consortium 2018), KEGG (Kanehisa et al. 2015), WikiPathways (Slenter et al. 2017), and Reactome databases (Jassal et al. 2020).

The most significant GO-BP terms were observed in the case of nerve cord specific genes. Top terms were mostly oriented around synaptic signalling such as “*synaptic signal*” or “*chemical synaptic transmission*” and ion transports (“*cation transport*”, “*ion transmembrane transport*”). In terms of significance, the nerve cord was followed by the gut/chloragogen and crop tissues. While in the case of gut/chloragogen the most significant terms appeared to be associated with protein targeting to membrane (“*protein targeting to ER*”, “*cotranslational protein targeting to membrane*”, crop appeared to be involved in different oxidation-reduction and metabolic processes (“*oxidation-reduction processes*”, “*fatty acid metabolic processes*”, “*small molecule metabolic processes*”). Coelomic fluid-specific genes showed most significant enrichment for “*response for stimulus*”, “*multicellular organism processes*” and for “*cell communication*”. Enriched GO-BP terms with lowest significance were identified in the case of gizzard a pharynx specific transcripts. The top gizzard specific terms were the “*actin filament-based processes*”, “*sarcomere organisation*” and “*actin cytoskeleton organisation*” which seemed to correlate with the high muscle content of the tissue. Pharynx specific genes resulted most significant enrichment for cilium related processes such as “*cilium organisation*”, “*cilium assembly*”, “*microtubule-based processes*”, “*cell projection organisation*”. The top ten GO term corresponding to Biological Processes (BP) are showed in Figure 14.

Table 3: Number of identified tissue-specific transcripts in the case of each used tissue as well as their annotation success rate. Tissue-specific transcripts were identified based on differential expression analysis using the NIOseq module of the Omicsbox software package.

<i>Eisenia fetida</i>	Tissue-specific transcripts	Annotated transcripts	Percentage of annotation
Coelomic fluid	3,133	856	27%
Gut/Chloragog	2,030	948	46%
Gizzard	747	255	34%
Crop	819	340	41%
Pharynx	3,574	1,376	38%
Nerve cord	4,703	1,469	31%

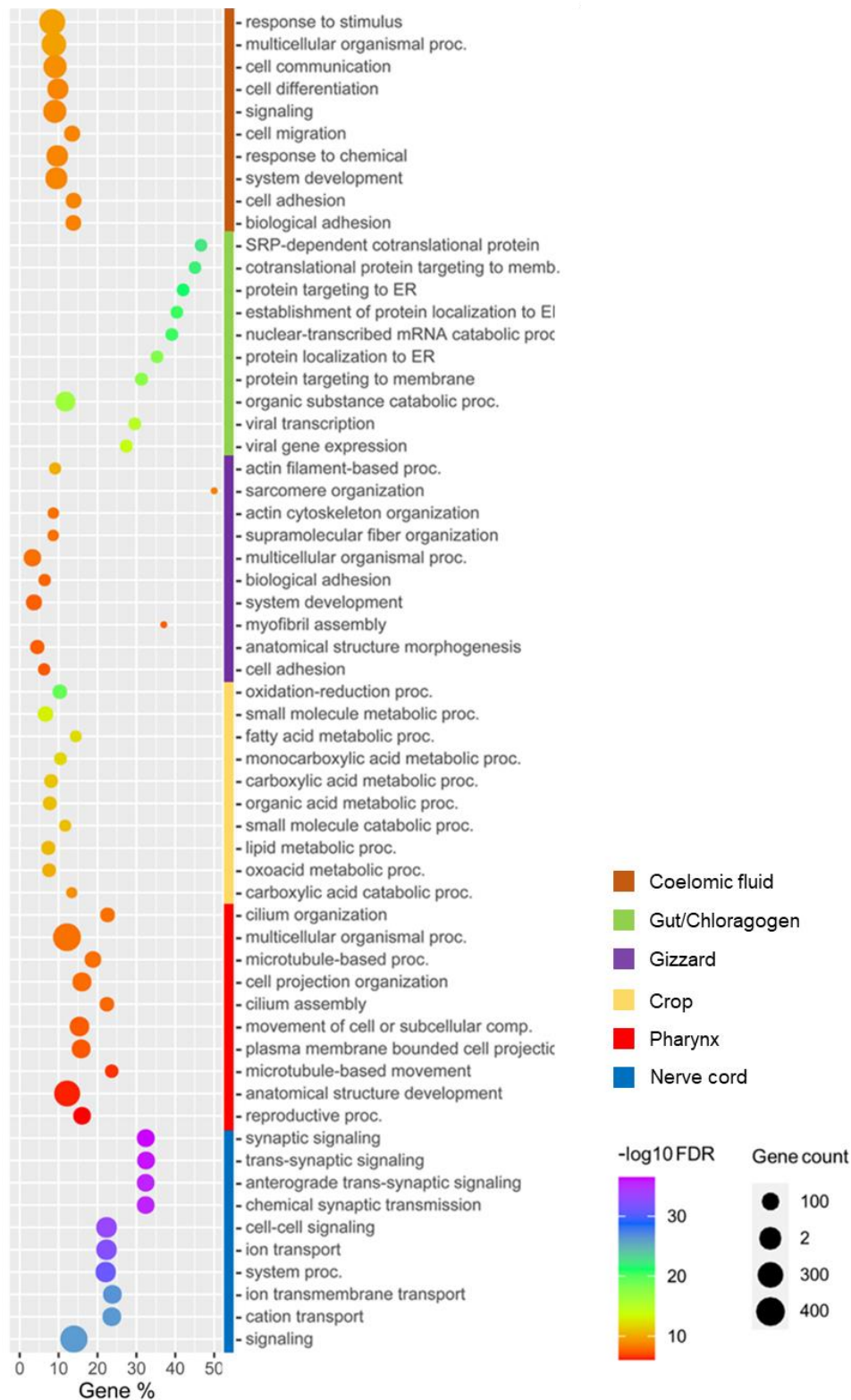


Figure 14: Top ten most significantly enriched Gene Ontology Biological Processes (GO-BP). Tissue-specific genes from *E. fetida* and *E. andrei* were used as the input with the total annotated gene set used as the population to generate enrichment using gProfiler. The statistical significance of the identified terms represented by their colours while the size of the dots illustrates the number of genes associated with certain terms. Enrichment results using specific genes from different tissues separated by row-side colours.

2.3.4 Immune gene identification

Using different homology-based annotation techniques (BLAST), in total 3,692 putative immune system processes (GO:0002376) related transcripts were identified in *E. andrei* while this number was 3,699 in *E. fetida*. From the 3,692 and 3,699 immune system-related transcripts approximately 24% appeared to be unique genes (960). Following the identification of the putative immune-related *E. fetida* transcripts, their tissue expression profile was also characterised (Figure 15). The highest number of immune-related genes appeared in the Pharynx tissue (296) followed by the coelomic fluid (204) and the nerve cord (139), while the lowest number was observed in the gizzard tissue (49). Tissue-specific immune-related transcripts could be identified in all the six tissues where the highest number of tissue-specific genes were counted in the Pharynx (66) closely followed by the nerve cord (63) and coelomic fluid (57).

To gain more detail about the distribution of these immune system-related transcripts between the different functional pathways associated with innate immunity, these contigs were classified using pathways from the InnateDB. Using the data of the tissue-specific atlas we successfully associated transcripts with innate immune pathways such as the chemokine signalling, complement cascade, cytosolic DNA-sensing, Jak-STAT signalling, MAPK signalling, mTOR signalling, Natural killer cell-mediated cytotoxicity, NOD-like receptor signalling, Regulation of autophagy, RIG-I-like receptor signalling and Toll-like receptor signalling. In total more than 1860 transcripts were assigned to at least one of the above-mentioned Innate immune pathways (Figure 16). The filtering for only unique annotations resulted in 494 annotations altogether, from which the highest number of unique genes were assigned to the MAPK signalling pathway while the highest pathway coverage appeared in the case of the mTOR pathway (65%). The lowest pathway representation was observed in the case of Jak-STAT signalling related genes with around 23% successful identification rate (Figure 17). The components of the Toll-like receptor signalling pathway were also visualised using the KEGG pathway mapper tool (Kanehisa and Sato 2020), in total 48 genes associated with 'Toll-like receptor signalling' could be mapped, as shown in Figure 18.

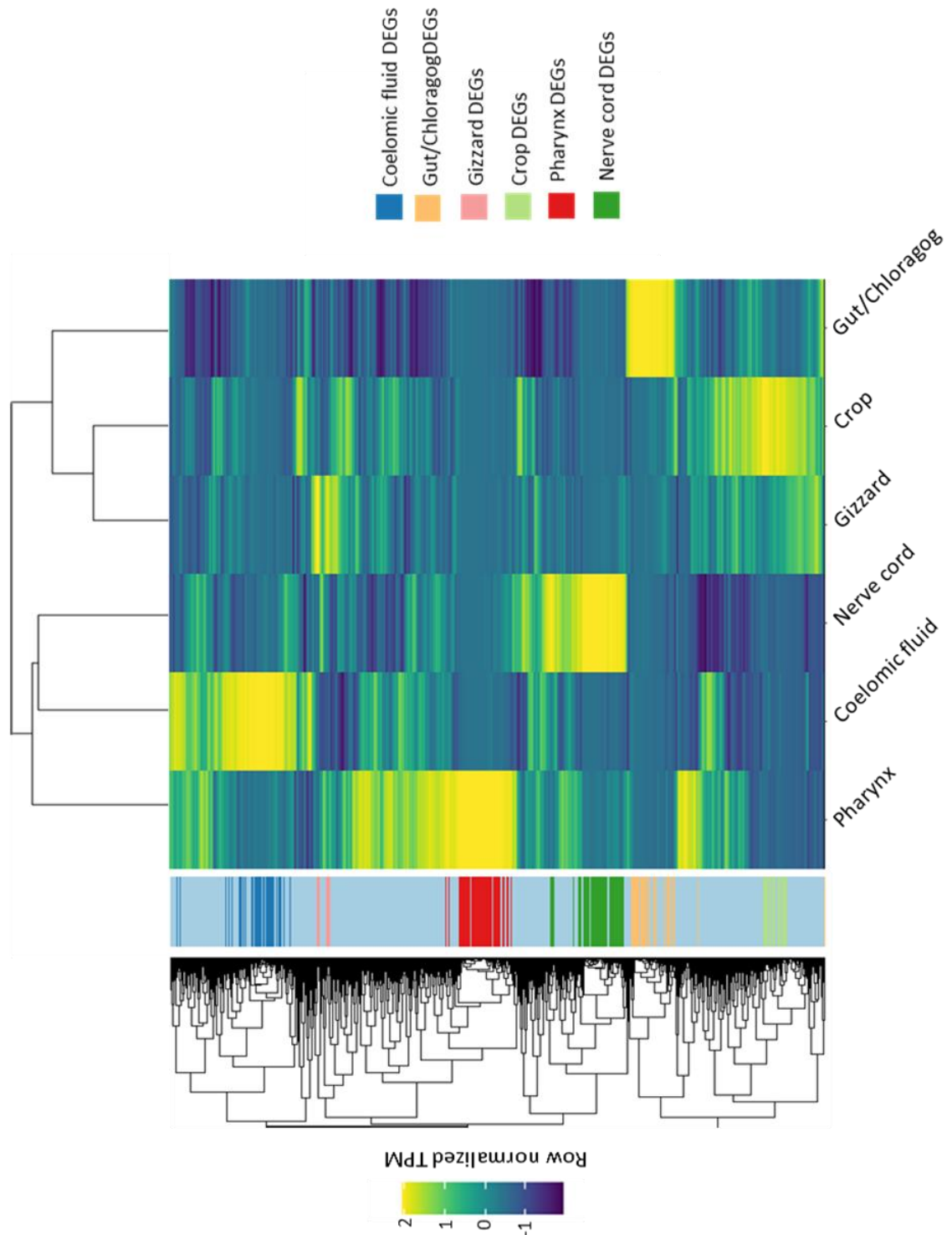


Figure 15: Immune-related genes based on their tissue expression profiles in the case of *E. fetida*. Heatmap showing the hierarchical clustering of immune-related genes with row-side colours representing the tissue-specificity of the genes. Expression values were plotted based on normalised Transcript per Million (TPM) values extracted from the output of the RSEM software. In total 3,699 immune-related transcripts were identified from which 494 had unique functional annotation.

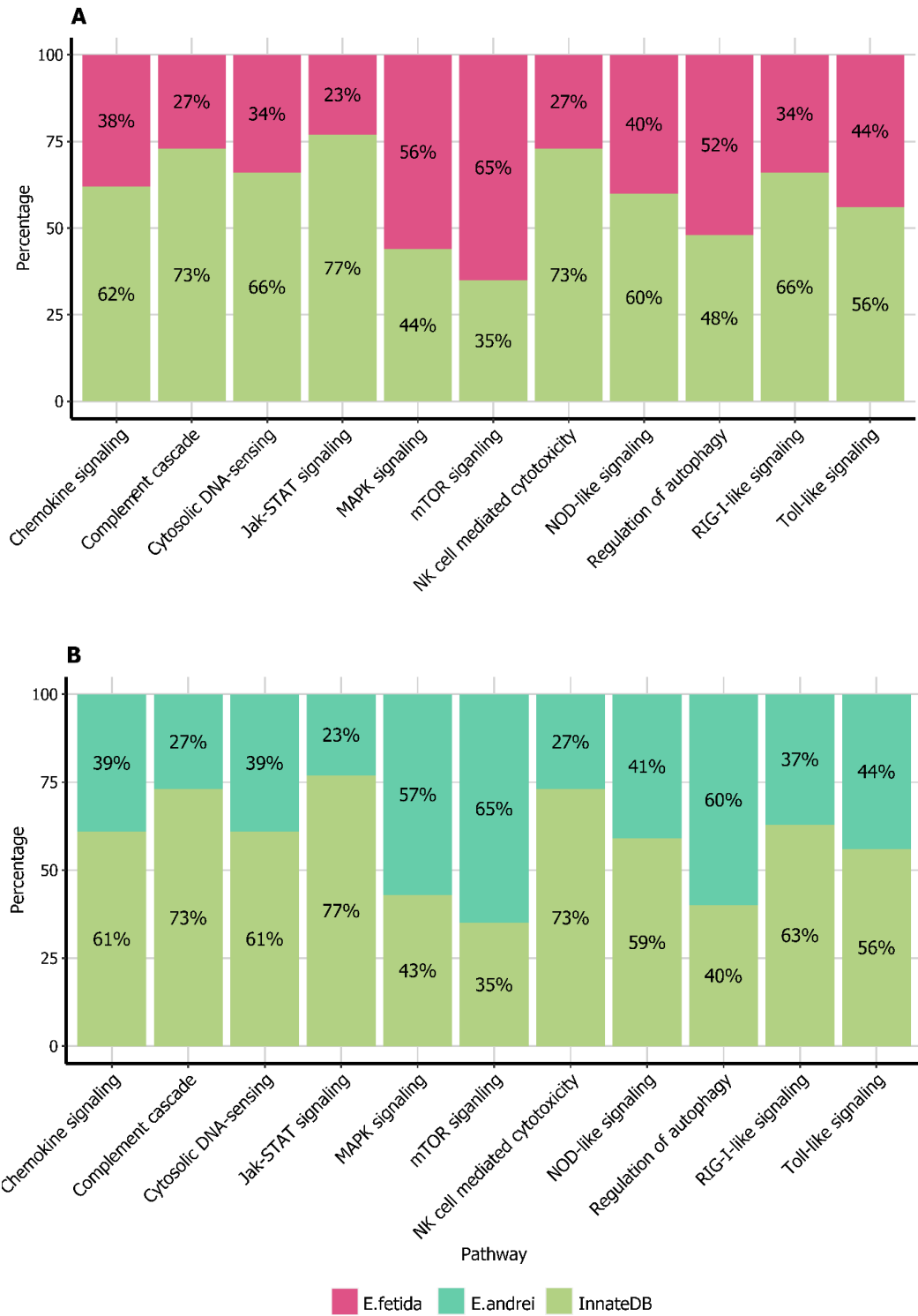


Figure 17: Identification success rate (percentage) for components the major Innate Immune pathways. Data is presented for *E. fetida* (Panel A) and *E. andrei* (Panel B) transcriptomes. The innate immune genes were divided between the different pathways based on the InnateDB (green), using it as a pathways reference set.

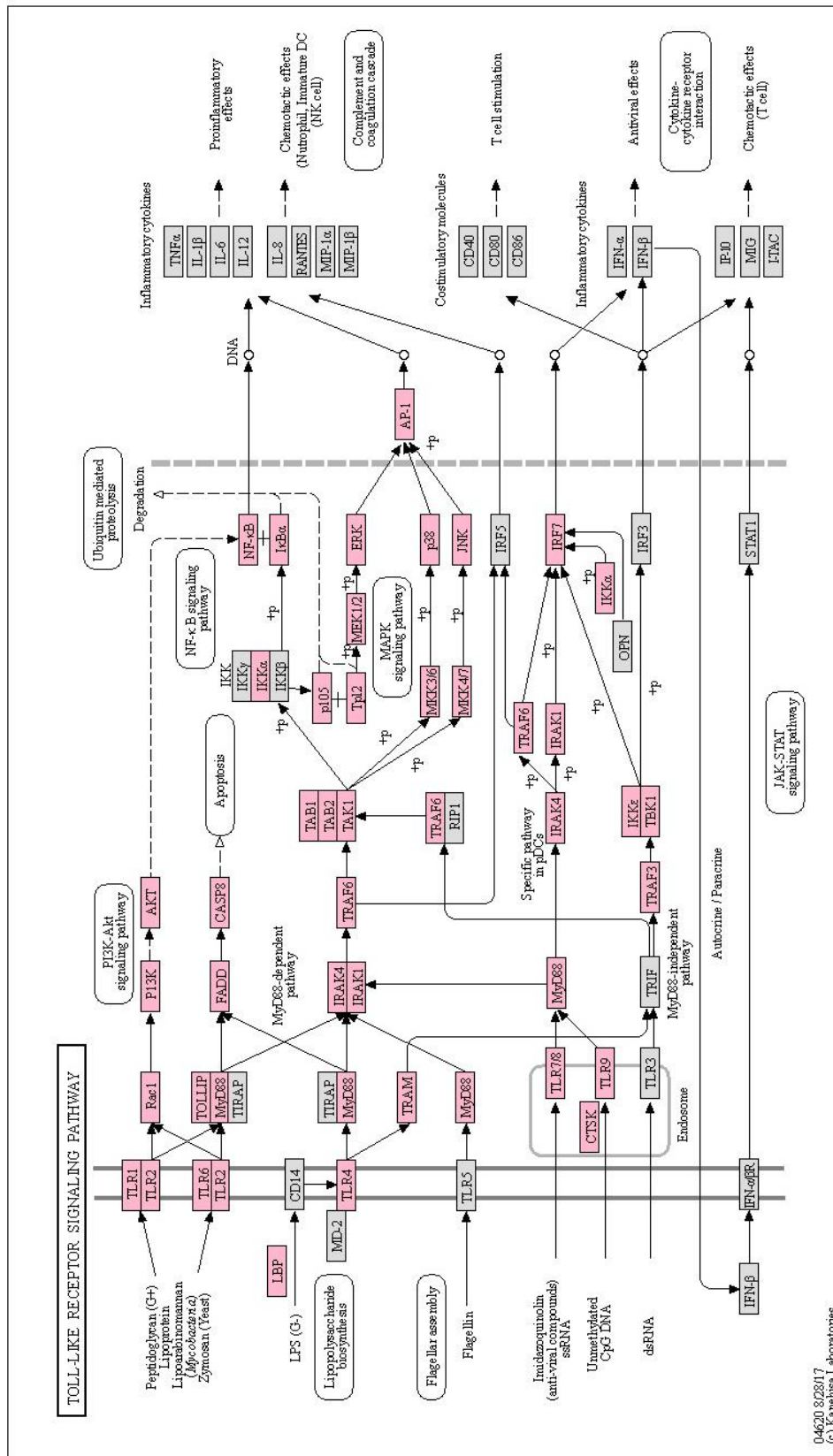


Figure 18: Immune system pathway mapping. KEGG Toll-Like Receptor Signalling pathway is shown where light-red colour shows the genes which were identified in both of our *Eisenia* transcriptomes. First gene symbols were converted to KEGG Gene IDs then they were mapped using the online KEGG pathway Mapper tool.

2.4 Discussion

2.4.1 Transcriptome assembly and annotation

Transcriptomes generated using whole-body sequencing data often contain rather biased results from the view of completeness. By using the whole body of the earthworms as starting material, tissues could be represented with a rather different sequencing depths. It is obvious that in most cases the distinct types of tissue provides significantly different mass to total body weight ratio due to their variant abundance in the sample, both when using the whole body or only a cross-section of it. The high percentage of muscle content in the body samples can monopolise most of the sequencing data while providing a relatively low sequencing coverage for other, less abundant tissues, such as the coelomic fluid or the nervous system. This could be a problem especially in the case of genes that are expressed at a relatively low level. Due to these poorly represented transcripts, even the modern *de novo* transcriptome assembly pipelines can suffer a significant loss in the terms of both assembly completeness and contiguity. By starting the RNA-seq library preparation from nearly the same quantity of total RNA, extracted from precisely dissected tissue samples individually, the sequencing depth was nearly equalised between the different tissues. By utilizing this method, we were able to generate a reference transcriptome for the two closely related *Eisenia* species (*E. fetida* and *E. andrei*) with outstanding completeness and continuity.

In many cases, another challenging perspective of the *de novo* transcriptome assembly pipeline is the usage of genetically highly diverse individuals within the same samples. It is known that in earthworms, especially the *Eisenia* genus, a relatively high genetic diversity can be observed among the different lineages. With a lack of gene objects provided by a reference genome, combining this diversity with *de novo* transcriptome assemblers results in an artificially expanded number of transcripts, many of which represent only different allelic variants of the same transcript. The distinction between true gene isoforms and the allelic variants makes the functional characterizations of the gene families extremely challenging, as well as they can greatly reduce the efficiency of the differential gene expression methods by increasing count loss due to multi mapping. The generation of a tissue-specific atlas using only individuals from the same

phylogenetic clade also contributed to the high contiguity and completeness of the final transcriptomes.

2.4.2 Tissue functional profiling

By analysing the tissue-specific dataset we successfully identified several tissue-specific transcripts in the case of each tissue sample. This has allowed us to characterise the baseline biological functions belonging to certain tissues of the earthworm at the transcriptomic level. In general, the results of over-representation analysis applied on the lists of tissue-specific genes showed a high analogy with their expected main biological functions. This provides evidence that the generated dataset is high quality and the subsequent analysis of the tissue-specific immune aspects of the data is highly reliable. For example, the high number of synaptic signalling related GO terms enriched in the case of nerve cord-specific genes well represent the basic signal transmission function of the ventral nerve cord (Hess 1925). Similarly GO terms such as “*sarcomere organisation*” and “*actin filament based processes*” enriched in the gizzard can be easily associated with its high muscle content and its food particle grinding function (Carlhoff and D'Haese 1987, Peters and Walldorf 1986). Although, in the pharynx and crop, basic tissue functions were also easily recognisable in the top GO-BP terms, coelomic fluid and gut/chloragogenous tissues resulted in more general GO terms. In the case of these tissues, pathway enrichment results based on the KEGG and Reactome databases were more suitable in representing the tissue function on the transcriptomic level (table showing pathway enrichment results provided in Appendix 2.1). The low appearance of tissue-specific genes in the case of crop and gizzard could be a result of their close-*in vivo* location and the lack of clearly visible border between the tissues. Despite the conducted precise dissection, this could slightly decrease the purity of the crop and gizzard samples which may result in a higher background noise in the case of these samples.

2.4.3 Immune gene identification

To understand the normal function of the earthworm's innate immune system, first, it is essential to determine its main components. Deriving a transcriptome directly from coelomic fluid, that contain mixed populations of immune cells, resulted in better coverage of immune genes with higher transcript contiguity than could be derived from

whole-body transcriptomes. Based on sequence homology analysis (functional annotation) we were not only able to successfully identify a high number of transcripts with possible immune-related function in both *E. fetida* and *E. andrei*, but also separated them between the major innate immune pathways. It was also important to describe the tissue-specific expression profile of the identified putative innate immune genes, since it is well known that different tissues often play distinct roles in the immune process (Dvořák, 2016, Prochazkova, 2019). Even in earthworms, several recent immunological studies pointed out the immunological importance of the different parts of the intestine (hindgut, foregut, gut), which are some of the most highly exposed anatomical structures to different microbial invasions (Fiołka, 2012). An interesting example of the immunological importance of the intestinal tract was its lysozyme activity, which showed the highest expression in the gut tissue. Another antimicrobial peptide appeared to be most highly expressed in the pharynx tissue described recently by Bodo et al. (Bodó, 2019). Although the majority of earthworm immunological studies still exclusively target the coelomic fluid, our results also suggest that other tissues, such as the pharynx, may play an important role in the earthworm innate immune surveillance. Performing the tissue-specific analysis was justified given the extent of tissue-specific immune gene expression identified in the analysed tissues. Although in earthworms the coelomic fluid is considered as the tissue with the highest immunological importance (Hostetter and Cooper 1974), our transcriptomic analysis retrieved a higher number of pharynx, and nerve-cord-specific immune-related transcripts.

2.4.4 Limitations of the non-model *de novo* pipeline and future developments

Although the *de novo* tissue-specific transcriptomic approach allowed us to observe the expression of the immune components between different tissues, and delivered an excellent reference for the later experiments focused on earthworm immunity, it still has its own limitations. During the last few years, *de novo* transcriptomic pipelines evolved greatly to be able to reconstruct the greater number of transcripts with the highest possible structural completeness and assembly accuracy, without the necessity of a genomic reference (Grabherr et al. 2011b). However, when applying these methods on non-model species with high allelic variance the recovery of multigene families, high

number of paralogs, and the estimation of allelic diversity is still problematic (Ungaro et al. 2017). Although the experimental design tried to eliminate these limitations by decreasing the allelic variation within the samples by using a small number of inbred individuals, we retrieved transcripts with a high number of gene paralogs with reasonable transcript length reconstruction. This problem was easily recognisable in the example of Toll-like receptor family. As it is shown in Figure 19 the short-read-based coverage showed high variation along the length of the Toll-like reference transcript. This suggests the possible existence of expanded number of Toll-like gene paralogs in the *Eisenia* genome. The coverage histogram also suggests that the intracellular Toll/interleukin-1 receptor (TIR) domain highly conserved so multi mapping induced coverage loss appeared to be less significant in this region, while in the less conserved extracellular regions the coverage drops below the level where the *de novo* pipeline cannot reconstruct the different paralogs with full-length. This problem well represents the challenge of the *de novo* assembly pipeline when the high gene paralog number combined with high partial sequence homology.

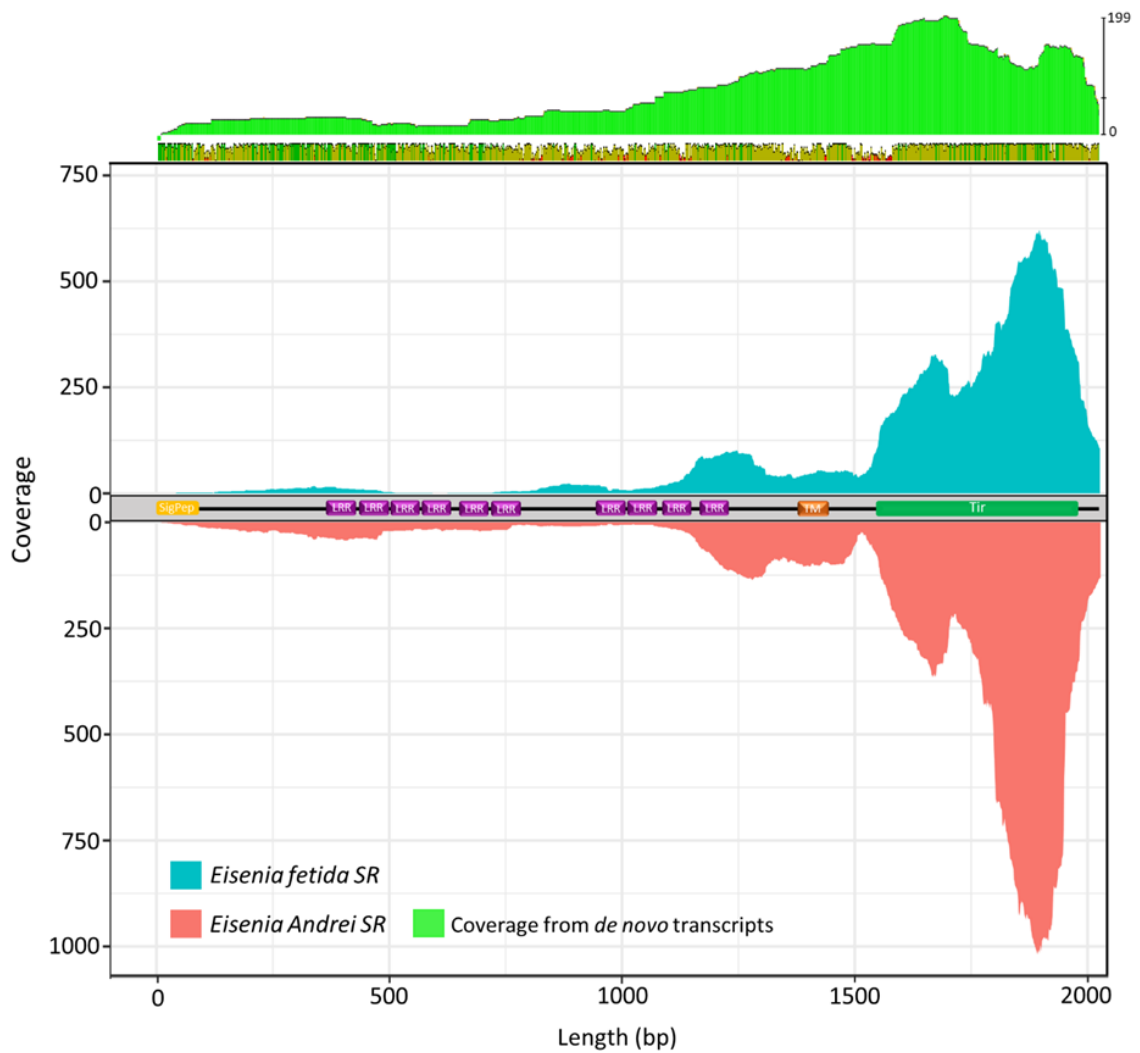


Figure 19: Read coverage of Toll-like receptor. Adapter trimmed reads derived from *E. andrei* (red) and *E. fetida* (blue) short-read data are mapped against a Toll-like receptor isoform (described in *E. andrei*) as a reference sequence. The position of the main functional groups of the Toll-like receptor is shown in the middle of the plot (Tir - intracellular Tir domain, TM – transmembrane domain, LRR- Leucin rich repeats, SigPep- extracellular signal peptide). The top histogram (green) represents the coverage retrieved from the assembled contigs extracted from the *E. andrei* and *E. fetida* transcriptomes.

As a result of the earlier mentioned challenges *de novo* assembly pipelines can generate a high number of fragmented transcripts, where different allelic variants and paralogs can be mixed within the same contig. To overcome this challenge and possibly resolve these closely related paralog transcript sequences, one solution is the utilization of a highly contiguous genomic reference to act as a template for transcript reconstruction (Huang, 2016).

3 Genomic template for Innate Immunity

3.1 Introduction

Comparative immunology has provided novel insights into the evolution of the fundamental interactions between pathogen and host, from recognition and signalling to the downstream immune processes (Cooper 2003). However, cellular and molecular immunology and mechanistic studies have focused on vertebrates. This focus is reflected in the availability of molecular data, including existing genome data that provides the most comprehensive picture of components contributing to immune processes. The resulting paucity of data on immune responses in non-mammalian species has skewed our understanding; however, access to new genomic resources is starting to reveal the true diversity and complexity of immune pathways (Dheilly et al. 2014). Genomic data provides the lens by which the full complexity of an organism's immune system can be appreciated; by combining genomic data with a tissue atlas we can start to map out the 'what' and 'where' of the earthworms immune system.

First-generation sequencing started with the introduction of Sanger sequencing, which was developed by Frederic Sanger and his colleagues in 1977. It was the most widely used sequencing technique during the last 40 years and it was based on *in vitro* DNA replication using chain-terminating dideoxynucleotides and DNA polymerase (Sanger et al. 1977). This technique was able to generate individual reads up to one kilobase in length. Although the first human genome was assembled relying exclusively on data generated with Sanger sequencing, due to the high cost and long processing time, the genome sequencing of non-model organisms was still problematic (Schuster 2008). Therefore, before the last decade, comparative immunology, regarded from the viewpoint of genomics and transcriptomics, was mainly based on the use of a limited number of reference genomes which almost entirely originated from vertebrates including mammals, birds, amphibians and fishes, with representatives from invertebrate phyla including one insect and nematode species. Furthermore, molecular immunological studies relied mainly on techniques such as quantitative polymerase chain reaction (qPCR) with some groups performing limited comparative *de novo* transcriptomic studies (Schultz and Adema 2017, He et al. 2020).

As Next Generation Sequencing (NGS) technologies become established the cost of sequencing has dropped substantially (Sboner et al. 2011), and in the same time

improvements in the field of bioinformatics managed to drastically reduce the labour-intensive nature of the genome assembly pipelines (Schmidt and Hildebrandt 2017). Although during the last few years this resulted in a high increase in the number of available *de novo* reference genomes, NGS-based assembly methods still have some major drawbacks mainly due to the relatively short length of the used reads (~75-500bp). The assembly of genomes with high allelic variability and repeat sequence content (Rimmington 2018), as in the case of earthworm species, remained challenging, even with the use of extremely high depth short-read (SR) sequencing. In many cases, the length of the different repetitive regions is much longer than the average insert size of the SR library, which can lead to misassemblies and gaps in the assembled genomic contigs. Most importantly it can result in a rather fragmented *de novo* assembly. Another problematic aspect of the NGS is its dependency on PCR reactions, which are known to have difficulties amplifying genomic regions with extreme GC content. This can lead to missing or underrepresented regions in the sequencing data.

Just after NGS started to be widely used, third generation sequencing (TGS) techniques were developed that began to overcome the earlier genomic sequencing challenges. TGS utilise different single-molecule sequencing (SMS) methods which are based on the real-time sequencing of the DNA or RNA molecules without the necessity of PCR-based amplification. The first official TGS method was released by a company called Pacific Biosciences (PacBio) in early 2011 (Rhoads and Au 2015), shortly followed by Oxford Nanopore Technologies (ONT) (Brown and Clarke 2016). Both of these sequencing methods are capable of producing several kilobases (kb) or, in some cases, megabases (Mb) long sequences originated from a single molecule. Even with relatively low coverage, these long-read sequencing datasets could greatly improve the contiguity of the assemblies, especially when genomes are highly heterozygous, complex, and repetitive.

Usage of the newest available genomic and transcriptomic resources during the last decade has started to expand our knowledge of comparative immunology and is providing a very complex picture about the evolution of the immune system across the whole animal kingdom (Rast and Messier-Solek 2008). New evidence derived from NGS based comparative genomics suggested that several gene families which play an

important role in different innate immune mechanism, such as pattern recognition receptors, are highly conserved and can be identified even in sponges and cnidarians (Srivastava et al. 2010, Putnam et al. 2007). Toll-like receptors (TLRs) are one of the most intensely studied receptors in both vertebrate and invertebrate immunology (Takeda and Akira 2015, Takeda et al. 2003). However, they present extreme divergence between animal groups. For example, genome analysis of the purple sea urchin revealed 253 different TLRs (Buckley and Rast 2012a), while the average number of TLRs in vertebrates is around 10 (Roach et al. 2005, Areal et al. 2011) and only a few annelid TLR sequences have been published (Davidson et al. 2008, Škanta et al. 2013).

Furthermore, several research groups find evidence, not only for receptors like TLRs and different antimicrobial molecules associated with the innate immune system but also hypervariable recognition molecules from different invertebrate species (Cerenius and Söderhäll 2013). However, although these highly variable molecules appear in many different invertebrate taxa such as molluscs (Fibrinogen-Related Proteins - FREPs), insects (Down syndrome cell adhesion molecule - Dscam), and crustaceans (Adema 2015, Ng and Kurtz 2020), we are still far from fully understanding their role. These molecules show characteristics analogous to vertebrate antibodies and were the first signs of an alternative immune memory-like phenomenon in invertebrates (Ottaviani 2011). It was important to recognise that despite the conserved nature of some components of the innate immunity, different invertebrate groups appear to have evolved several phyla-specific molecular mechanisms for mounting a defence against pathogens

At present, although more and more NGS data has become available from invertebrate organisms, there is only limited knowledge about the molecular biology underlying the annelid immune response, and this is especially true in the case of earthworms. The contiguity and completeness of the available reference genomes are not sufficient to make a detailed exploration of key components in the earthworm immune system, whether well-conserved or taxa-specific.

3.2 Materials and Methods

3.2.1 DNA extraction

An essential requirement for the exploitation of Third-Generation Sequencing (TGS) to support genome assembly is the production of high quality starting material. Since TGS technologies have the capability of sequencing long fragments of DNA or RNA molecules as one segment, in some cases without any amplification or the necessity of fragmentation, these sequencing methods require DNA with High Molecular Weight (HMW) and outstanding chemical purity. However, despite the recent expansion of Long-Read Sequencing (LR), the methodologies behind these techniques are still in a rather developmental stage compared to Short-Read Sequencing (SR). Routine protocols for extracting DNA, with such high quality, are mostly available only in the case of laboratory microorganisms and tissues/cells of well-studied model species, such as human and mice. Due to the dearth of information, the extraction of HMW DNA from various biological materials in many cases requires individual, species, and sometimes even tissue-specific DNA extraction methods to be developed. To meet the above-mentioned quality and quantity expectations of the Nanopore and Chromium (10X) sequencing, several different HMW DNA extraction protocols were tested and modified. Initially different column-based extraction methods from two well-known suppliers were used (DNeasy Blood and Tissue kit, Qiagen Ltd, Quick-DNA HMW MagBead Kit, Zymo Research, Tustin, CA), one of which was designed to achieve a relatively long fragment length (Figure 20). Subsequent approaches compared the resulting DNA from these commercial kits with more classical Phenol-Chloroform extraction methods (Wood 1983), followed by different approaches to precipitation and salt removal DNA washing (Figure 21). Then finally, to achieve the best chemical purity with a large fragment size, we combined the Phenol-Chloroform extraction method with dialysing. Details of this extensive optimisation are only provided in brief so as not to detract from the final data analysis, however, a comprehensive description of the final procedure is provided below.

DNA extraction protocol was as follows: High Molecular Weight (HMW) DNA was extracted from skin and pharynx muscle tissue of a freshly prepared (washed) single individual. Following the precise dissection, the tissue sample was digested in a

premixed tissue digestion buffer (720 μ l ATL buffer mixed with 80 μ l proteinase K supplied (Qiagen Ltd) together with 20 μ g/ml DNase free RNase (Promega, WI, USA). Tissue digestion was conducted at 56°C for 4-6 hours. To speed up the digestion process, a gentle mixing was performed by rotating the tube slowly 5-10 times after every 60 minutes. The fully digested tissue sample was transferred into a new 2 ml centrifuge tube and mixed with an equal volume of Phenol:Chloroform:Isoamyl Alcohol (25:24:1, pH 7.7-8.3) mixture. Then the sample was mixed on a rotation mixer on the lowest speed settings available until the fine emulsion was created (20-40 min.). To separate the aqueous phase from the organic phase and Interphases, the samples were centrifuged at 14,000 rpm (Eppendorf 5417C, Eppendorf Ltd, Stevenage, UK) for 20 min. The top layer was collected with a wide bore pipette and transferred into a new 2 ml tube after which the Phenol:Chloroform:Isoamyl Alcohol extraction was repeated. To remove any residual phenol, the sample was mixed with 0.5x of the volume of Chloroform and incubated on a hula mixer (ThermoFisher, MA USA) for 10 minutes, followed by 10 minutes of high-speed centrifugation at 14,000 rpm. The top layer was collected once again with a wide bore tip, and DNA was precipitated overnight using 0.3x volume Ammonium-acetate (7.5 M) and 0.6x volumes of ethanol. The 'cloudy' HMW DNA was collected with a sterile glass hook and washed in 50 ml 70% ethanol then left it to air dry before eluting in preheated (60°C) Tris-EDTA buffer (Wood 1983). Subsequently, RNase treatment was performed using RNase solution supplied from Promega (Promega, WI, USA) with a final concentration of 50 μ g/ml. The sample was then dialysed for 48 hours using a Genomic Tube-O-DIALYSER provided within a Mega Long TM kit (G-Biosciences, MO, USA), this allowed removal of leftover ribonucleotides, small fragments of DNA and any residual chemical contamination.

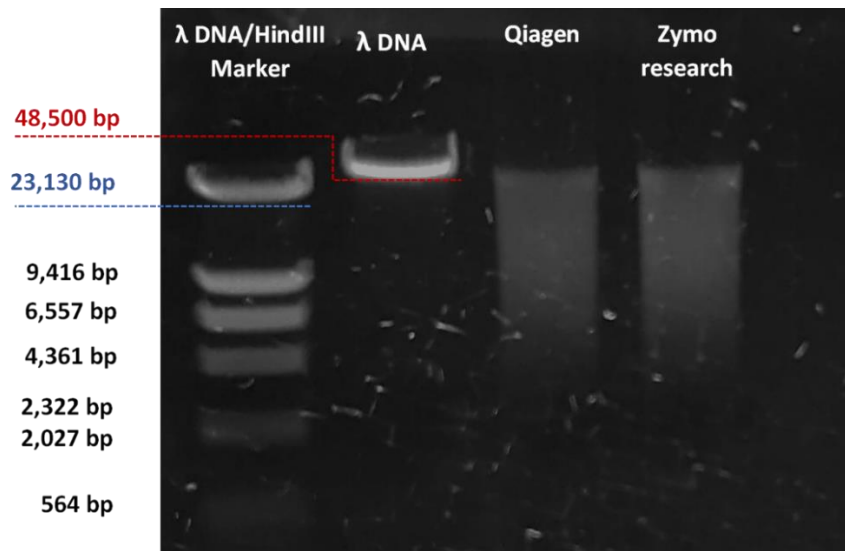


Figure 20: Size distribution evaluation of routinely used commercial “column” based DNA purification methods (Qiagen: DNeasy Blood & Tissue Kit, Zymo research Genomic DNA Clean & Concentrator). DNA was extracted from *E. fetida* muscle tissue and 100 ng of the extracted DNA sample was analysed on a 0.4% ultra-pure agarose gel (Sigma-Aldrich). Extracted DNA samples were compared against 100 ng of both undigested λ DNA and λ DNA digested with HindIII (New England Biolab, UK).

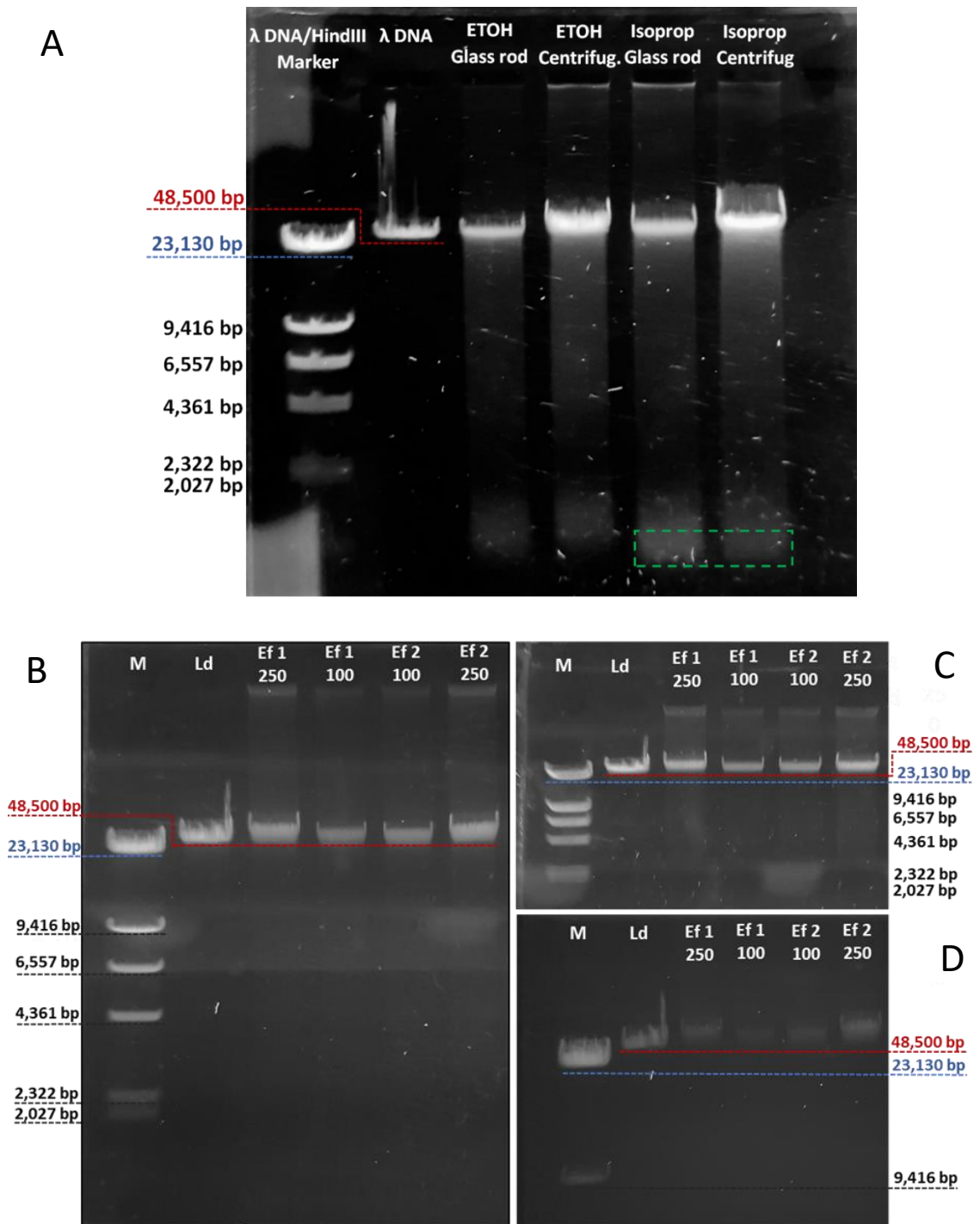


Figure 21: Methodological optimisation of HMW DNA sample using size distribution analysis. DNA was extracted from *E. fetida* muscle tissue and then 100 and 250 ng of the extracted DNA sample on was analysed by electrophoretic separation on a 0.4% Ultra-pure agarose gel. DNA samples were extracted using Phenol:Chloroform method followed by ethanol or isopropanol precipitation and collected with a glass hook or centrifugation. Final chemical purification was conducted using DNA dialysis (48 hour). Panel A shows the comparison between the different precipitation and precipitated collection methods. The B panel shows the final samples after only 2 h electrophoresis

while C and D represent the same samples after 4h and 6h running (M: λ DNA digested with HindIII, Ld: against undigested λ DNA - New England Biolab, UK).

3.2.2 DNA quality and quantity assessment

The fragment size of the DNA was assessed by running approximately 100 ng of sample on an Agilent Genomic DNA screen tape. Then quality was also measured by running 100 and 250 ng extracted DNA on a low percentage, ultra-pure agarose gel (0.4%) at 60 V for 6 hours. To assess the chemical purity of the sample, we measured the 260/280 and 260/230 ratios using a NanoDrop 1000 spectrophotometer (ThermoFisher, MA USA). The concentration of the sample was evaluated after vortexing for 5 s to facilitate physical fragmentation and measurement against appropriate standards using fluorometry (Qubit 4, ThermoFisher, MA USA).

3.2.3 Nanopore library preparation

MinION (Oxford Nanopore Technologies Ltd, Oxford, UK – ONT Ltd) sequencing libraries were prepared according to the manufacture's protocol used by the ligation method (1D gDNA – SQK-LSK109) with minor modifications. These revisions included use of 5 μ g of HMW DNA prepared as described in 4.2.1 which was fragmented to a mean size of 20 kb using physical fragmentation exploiting a Covaris g-TUBE as prescribed by the manufacture to yield the appropriate fragment size distribution (Covaris Inc, MA, USA). Subsequently, small DNA fragments (<1,500 bp) were eliminated using SPRI beads (Beckman Coulter Life Sciences, IN, USA) at a ratio of 0.4 beads to sample. The fragmented and size-selected DNA (~50 ng in 1 μ g) was assessed using an Agilent Genomic DNA ScreenTape (Agilent 4200 TapeStation, Agilent Technologies, CA, USA). Following fragmentation, 2 μ g DNA was simultaneously repaired and end-tailed using NEBNext FFPE DNA Repair Mix combined with the NEBNext End repair / dA-tailing Module (NEB). At this point, an AMPure XP bead clean-up was performed to remove unwanted enzymes and buffer constituents (Beckman Coulter Life Sciences, IN, USA). Adapter ligation was performed using the T4 Ligase from the NEBNext Quick Ligation Module (E6056, NEB). Enrichment of long DNA fragments (>3 kb) was achieved during the post ligation bead clean-up by using the L Fragment Buffer from the ligation kit, with the final ligated DNA being eluted in 15 μ l of the ligation kit elution buffer (EB).

3.2.4 Library loading and flow-cell priming

The final concentration of the library was measured with a Qubit 4 fluorimeter (4.2.2), and the molarity was using considering the earlier results from the Agilent Genomic DNA ScreenTape (4.2.3). Approximately 40 fmol of ligated DNA was transferred from the library to a clean 1.5 ml low DNA binding centrifuge tube where it was diluted to 12 μ l with buffer EB from the SQK-LSK109 ligation kit. After mixing the library with Sequencing Buffer (SQB) and Loading beads (LB), the total volume of 75 μ l sample was loaded into a previously primed MinION flow cell.

3.2.5 Nanopore sequencing and base calling

The sequencing data was generated using the MiniKNOW software (ONT Ltd) with the live base-calling option selected. In total, approximately 33 Gb of long read sequences were generated using three 1D flow-cells (FLO-MIN106D, ONT Ltd). These data considered the available C-value based genome size estimation of ~ 0.7 equivalent to around 47X coverage for the whole genome of *Eisenia fetida* (Vitturi et al. 2000). During all of the three individual runs, a relatively quick (nearly logarithmical) decrease was observable in the number of available pores during the sequencing run. Compared to the typical lifetime expectations based on the suggestions from ONT Ltd, the yield was around 50% of the sequencing capacity of the flow-cells due to the establishment of the rapid pore unavailability (Figure 22, 23). The flow-cells used by this project generated around 50% of the final sequencing yield within the first 12 hours and almost completely stopped sequencing following the first 40 hours, producing 90% of the final output. To our best knowledge, the reduction in both the sequencing time and yield was present due to the specific 3D structures of the earthworm DNA. As it already known, the DNA isolated from some organisms, especially bird (chicken) species, contains a high number of specific repeat regions which can cause the formation of different 3D structures of the DNA. Due to these 3D DNA conformations, the pores can become rapidly blocked causing inhibition of sequencing, early pore unavailability and a low yield.

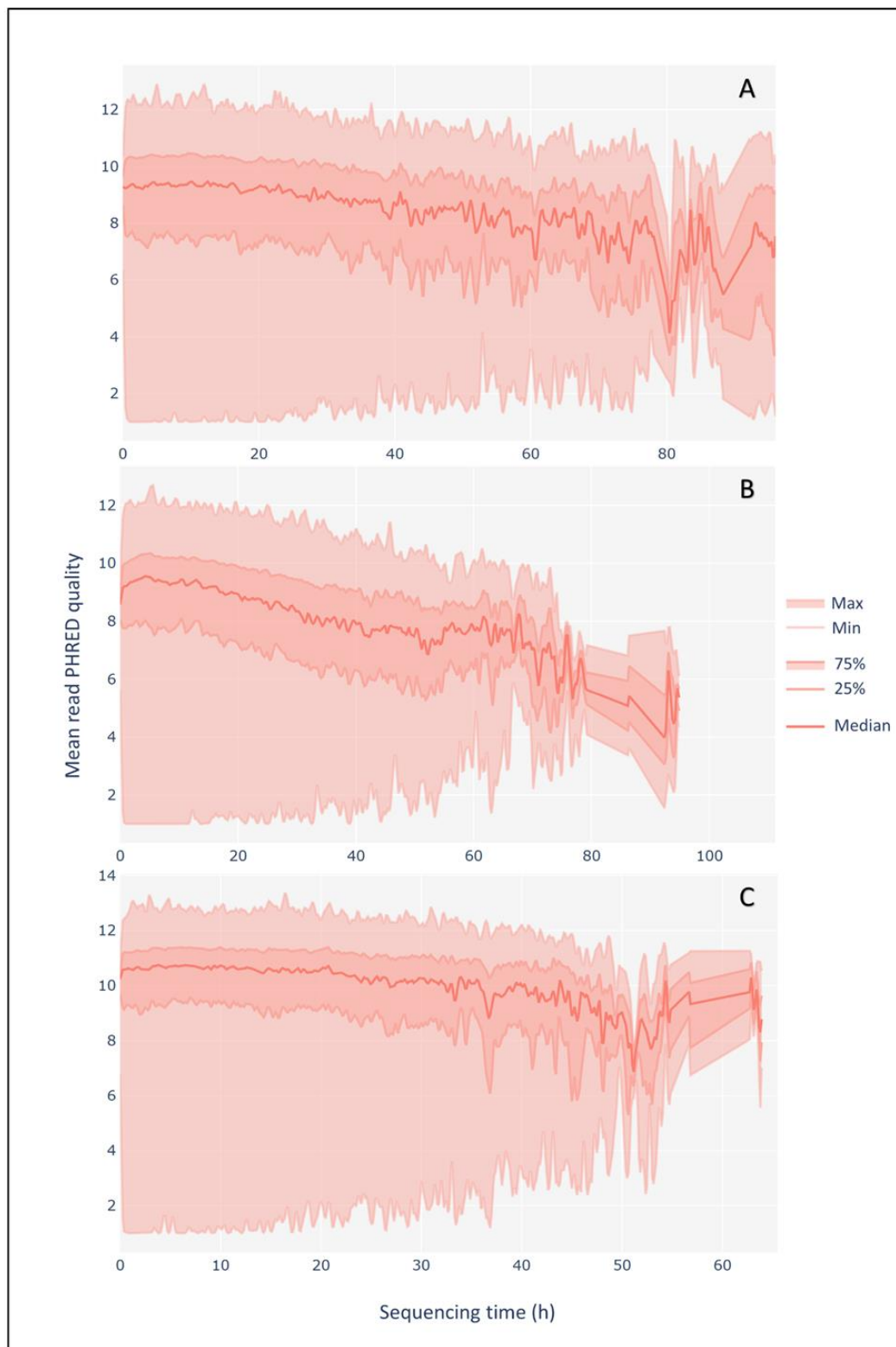


Figure 22: Change in base quality over time. Plots were generated with PycoQC and represent how the median read quality (PHRED score) changed during each of the three Nanopore minION flow cell runs (A-C). The highest median read quality was observed during the final (C) Nanopore sequencing run. The slight differences between the runs can be a result of subtle differences in flow cell or library quality.

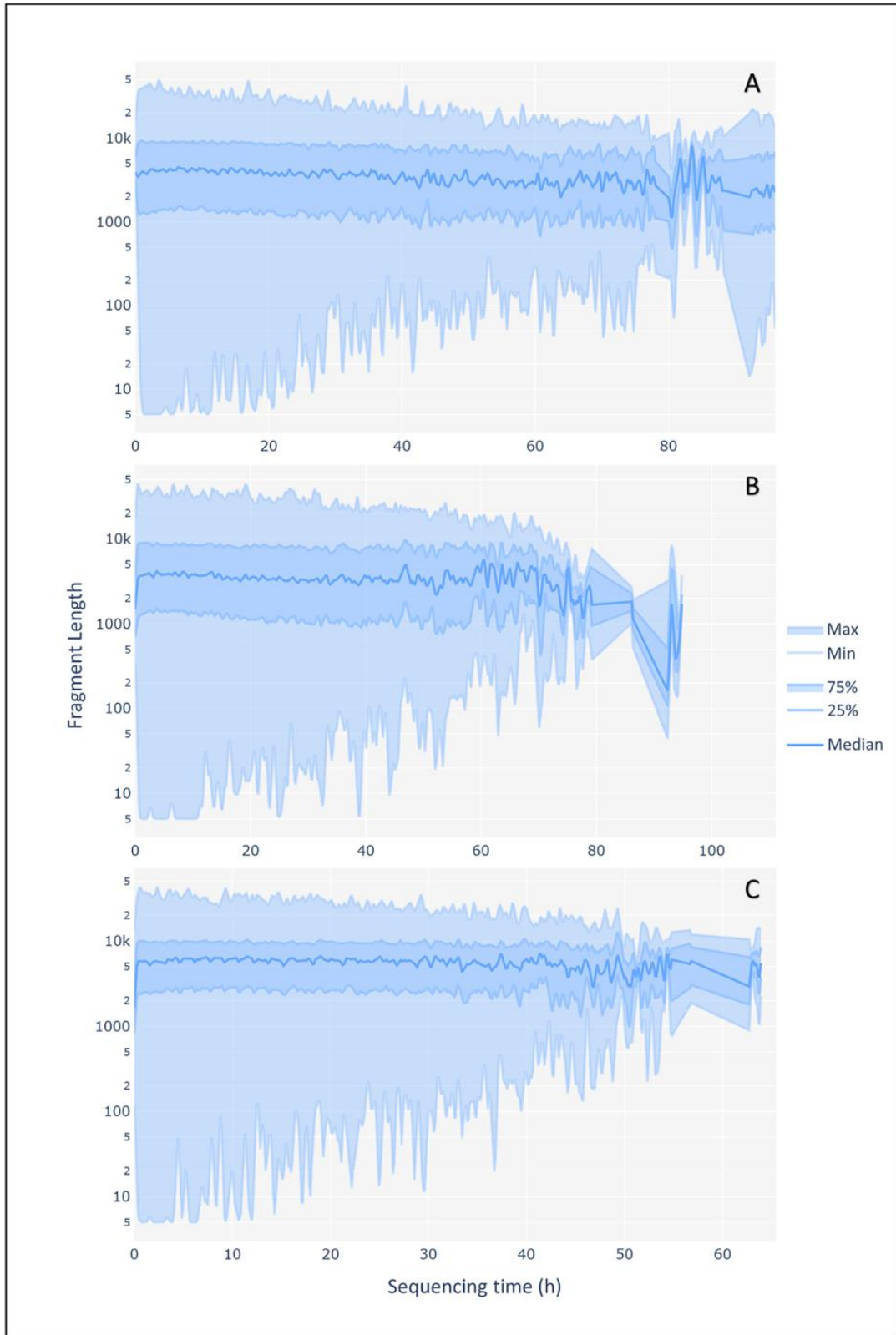


Figure 23: Change in fragment length over time. Plots were generated with PycoQC and represent how fragment length changed during each of the three Nanopore minION flow cell runs (A-C) . The highest median read quality was observed during the final (C) Nanopore sequencing run. The slight differences between the runs can be a result of subtle differences in flow cell or library quality.

3.2.6 Illumina Short Read Genomic Data Generation

In parallel to the long read sequencing, a small subsample of the DNA used for the MinION library preparation was used to generate complementary short reads. The library was prepared using the TrueSeq DNA (PCR-Free) whole-genome sequencing kit as recommended by the manufacturer, creating a library with insert size of ~450 bp (Illumina Inc, CA, USA). Sequencing was performed on a NextSeq 500 system using high output flow cells. Operation of the sequencing platform was performed by Ms Angela Marchbank, Cardiff School of Biosciences Genomics Hub, Cardiff University.

3.2.7 Quality control and error correction of the sequencing data

Currently, the main disadvantage of using Nanopore sequencing is the generation of long but error-prone reads (5-20% error content, depending on the type of library preparation and other factors such as the type of the molecule). To overcome this issue, error correction of the long and 'noisy' reads was needed. Since the pre-assembly correction of the raw reads would simplify and speed up the overlap and layout based assembly process, a hybrid correction method was applied. This hybrid correction pipeline uses short read sequencing data to build an FM-index using multi-string Burrows-Wheeler Transformation (Wang et al., 2018). For this reason, the Nanopore raw reads from all sequencing runs were concatenated into a single FASTQ file before conversion to FASTA format. Then approximately 50X coverage of Illumina short reads (with an insert size of 150 bp) (see 4.2.6) was generated from the same individual used to generate the long read sequencing data.

3.2.8 Initial *de novo* assembly

To achieve the best possible assembly contiguity along with high completeness, a number of different genome assembler pipelines were tested. Some of which exclusively depend on error corrected or raw long reads: Flye (Kolmogorov et al. 2019), Miniasm (Li 2016), Canu (Koren et al. 2017) and Wtdgb2 (Ruan and Li 2019), while other hybrid pipelines such as MaSuRCA (Zimin et al. 2013) and SPAdes (Bankevich et al. 2012) attempts to benefit from using the short and long read sequences during the same assembly process. After comprehensive testing, the assembly with the highest contiguity was created by a long read assembler based on a fuzzy Bruijn graph approach called Wtdgb2 (Ruan and Li, 2019). To reduce missassemblies and possible repetition,

the contigs assembled by Wtdgb2 (Ruan and Li 2019) were corrected using two rounds of consensus calling using Racon (Vaser et al. 2017).

3.2.9 Scaffolding using transcriptomic data

The first round of genome assembly scaffolding was performed using L_RNA_scaffolder (Xue et al. 2013). This program has allowed us to use the previously generated highly contiguous and complete tissue specific transcriptomic assemblies to scaffold the initial genome assembly. L_RNA_scaffolder works based on a similar theory as scaffolding with mate-pair data. However, in this case rather than using short read sequence pairs to identify possible joints, the program tries to identify genomic contigs that share different exons of the same gene based on long read RNA or cDNA sequences. After scaffolding with the tissue-specific transcriptomic atlas, the N50 of the genome assembly increased from 176 to 209 kb.

3.2.10 Scaffolding with Nanochrome

Chromium linked reads from 10X Genomics is a relatively new sequencing technology which exploits the advantages of a microfluidics platform to barcode short read sequences originated from the same HMW DNA molecule. When sufficient sequencing depth is available, the structure original HMW molecule can be resolved. To achieve higher contiguity by scaffolding using the Chromium method approximately 35 Giga-base of 10x paired-end sequencing data has been generated. Linked reads were generated from the same HMW DNA samples used for the Nanopore sequencing. The sequencing and library preparation was conducted by Novogene (Novogene, Co. Ltd, Beijing, China). Libraries were generated from the identical material used for Nanopore and short read sequence generation. Libraries were constructed using 10X Chromium Genome library products following the manufacturer's instructions and the sequence was generated on an Illumina Novaseq Platform (Illumina Inc, CA, USA).

Although, this technique works extremely well in the case of most well studied model organisms, the existing pipelines such as Supernova (Weisenfeld et al. 2017) and ARKS (Coombe et al. 2018) showed much less efficiency assembling allelically divergent earthworm species. To address this issue the second round of scaffolding was prepared by a novel program called Nanochrome. Nanochrome is able to use the scaffolding

information from the linked reads database and combine it with long reads from Nanopore into a single scaffolding process, relying on the strength of each technique (Figure 24). First, the chromium library is used to separate collections of short contigs and create allelic groups based on the number or shared chromium barcodes. Then the validation and orientation of the joinable candidates occurs by finding support for the joints with the alignments of the corrected nanopore reads. To reach a megabase size N50, we performed three round of scaffolding with Nanocrome. Between each of the scaffolding runs, a tiling path-based long read, assembly gap closer (LR_Gapcloser) was performed to fill the relatively long gaps which were generated during the scaffolding process (Xu et al. 2019). The final assembly polishing was conducted with two rounds of consensus calling by Racon (Vaser et al. 2017) followed with an additional run of Pilon (Walker et al. 2014) to eliminate any misassembly and remove repetitive scaffolds.

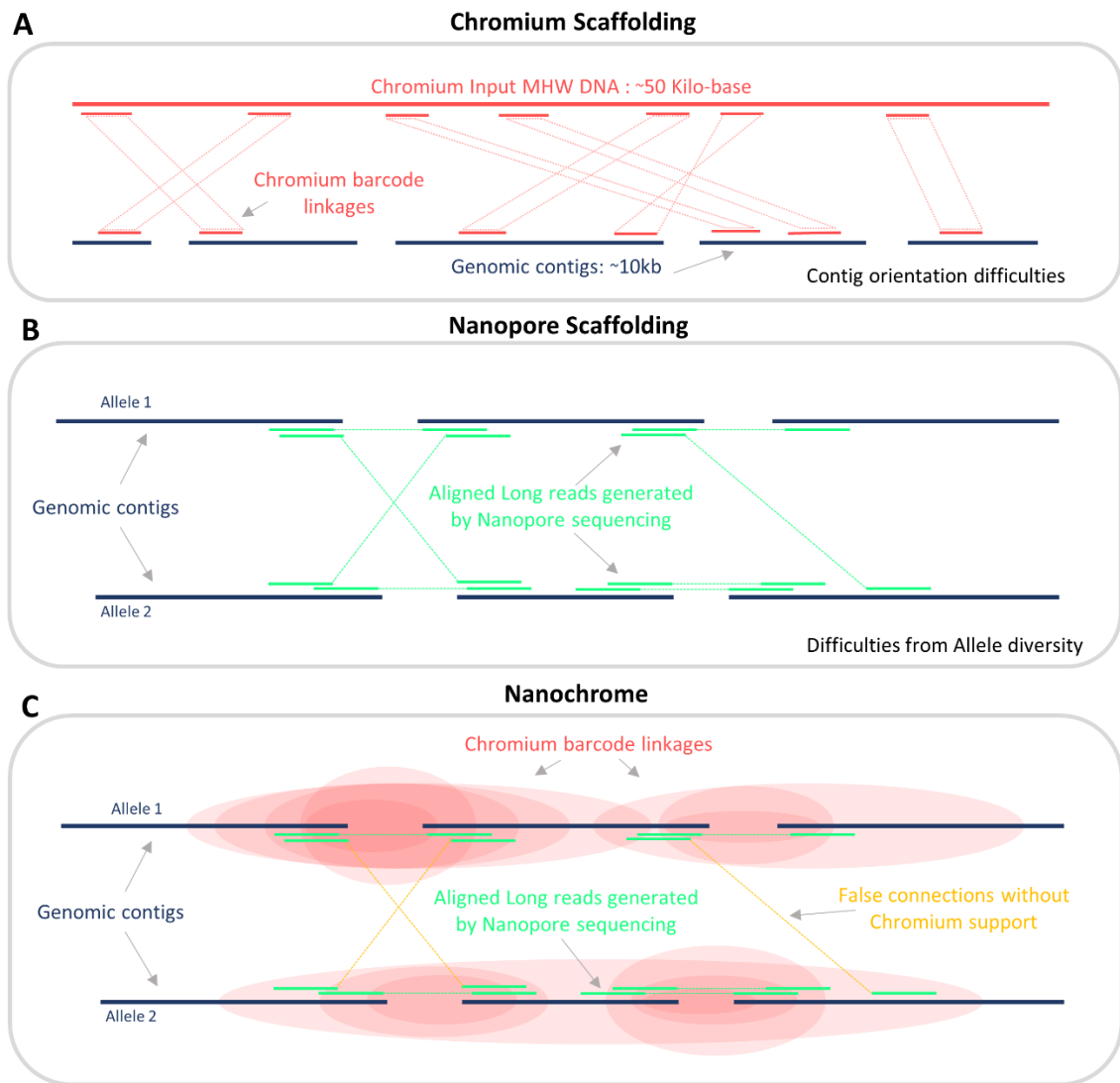


Figure 24: Schematic representation of the limitation factors of individual sequencing strategies. Nanopore long read (B) and Chromium (A) linked reads data based scaffolding are shown to be optimal for allelically highly variant genome assemblies. Panel C shows how Nanochrome utilise the benefits of both sequencing techniques to and achieve better overall scaffolding.

3.2.11 Assembly QC and filtering

The completeness of the assembly was estimated based on the number of near-universal single-copy orthologs identified in the assembly used by the pipeline of BUSCO v4. software (Seppey et al. 2019, Waterhouse et al. 2018), used to estimate completeness of the assembly based on the number of near-universal single-copy orthologs (metazoa_odb10). This was performed in genome mode with the Auto-lineage option selected. To separate any possible microbiome or other organism associated sequences and remove contigs with short read length (smaller than 8 kb) from the

genome assembly, a modular, command-line based quality control and taxonomic partitioning was prepared using the BlobTools2 (Challis et al. 2020) pipeline (Figure 25). The individual bioinformatic steps of the genome assembly pipeline from raw reads to the polished reference genome summarised on Figure 26.

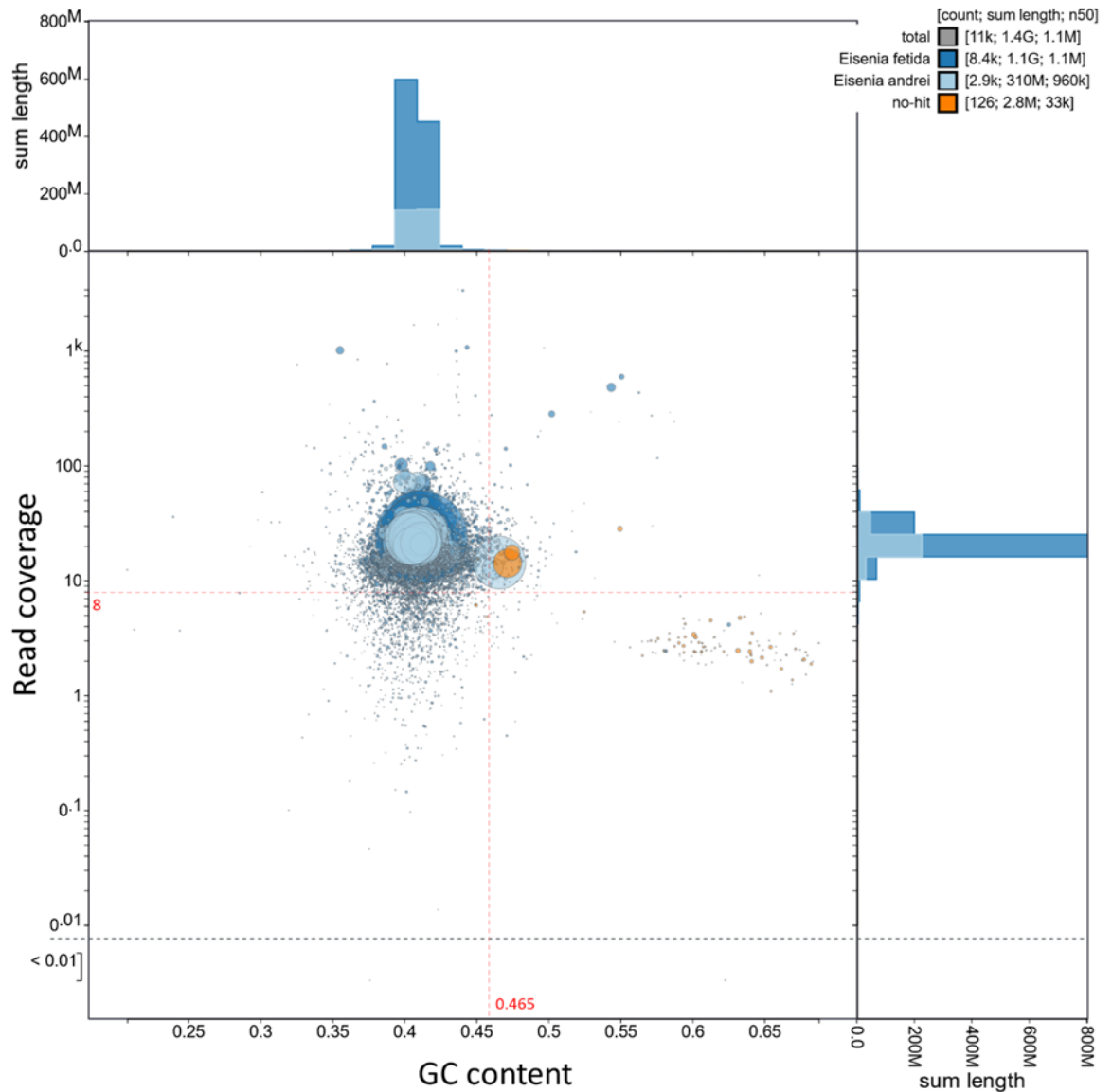


Figure 25: Evaluation of Genome assembly. Bubble plot was generated using Blobtools2. Each contig represented by a circle where the size of the contig is indicative of circle size. Contigs also represented by their read coverage from short reads and their GC content. The two shades of blue colour of the bubbles shows the identity of the contig's best blast hit, using the *E. andrei* and *E. fetida* tissue-specific transcriptomes as local database. Red dashed lines shows the cut-off values which were implemented during the assembly filtering step.

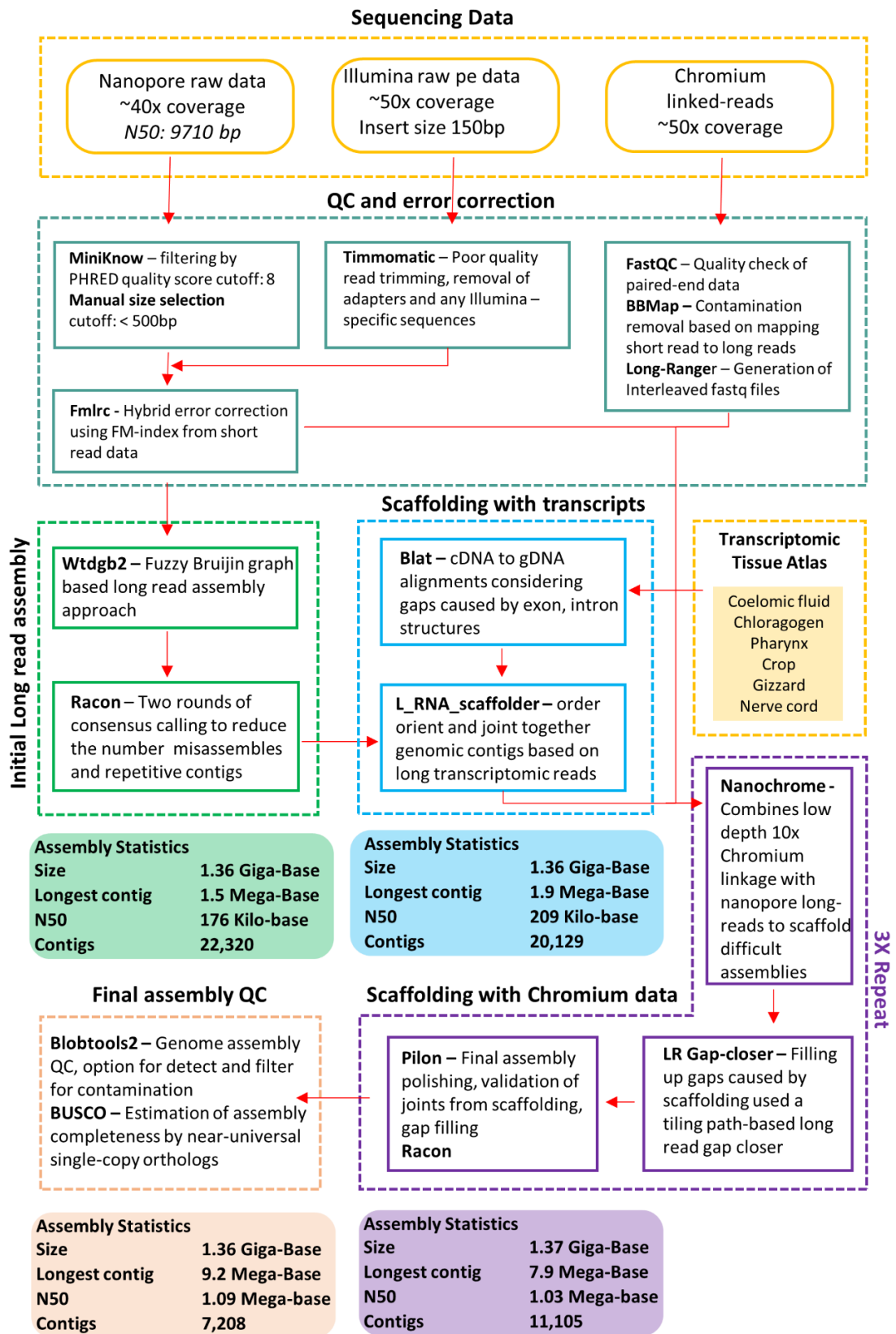


Figure 26: Workflow showing the basic details of the applied bioinformatic analysis used to generate the final *E. fetida* genome assembly. The major components of the analysis bounded by uniquely coloured dashed lines, while different background colours shows basic statistical metrics corresponding to the different assembly stages.

3.2.12 Repeat masking

To reduce the false positive rate of gene predictions by the *Ab-initio* genome annotation methods, the masking of the interspersed repeats and low complexity DNA regions was required. The masking of the genomic repeats were performed in to different steps. First, RepeatMasker (Tarailo-Graovac and Chen 2009) was used with the already available, curated online repeat libraries such as Repbase (Jurka et al. 2005) and Dfam (Hubley et al. 2015). Then a *de novo* repeat library was made using a transposable element and family identification modelling package called RepeatModeler (Smit AFA 2008). An additional RepeatMasker run was prepared to hard mask the newly identified, *de novo* repeat elements of the *E. fetida* genome.

3.2.13 Gene prediction

To achieve a genome annotation with high accuracy and completeness, several gene prediction pipelines were tested. Some of which are based on *Ab-initio* gene identification algorithms such as AUGUSTUS (Stanke et al. 2006), GeneMark (Brúna et al. 2020, Auffan et al. 2009), SNAP (Korf 2004) and others, using Assemble Spliced Alignments (PASA) (Haas et al. 2003). Based on the manual curation of a group of well-known genes, the best overall results were provided by the *Ab-Initio* gene-finding algorithm AUGUSTUS, which uses a protein family knowledge-based gene prediction method. AUGUSTUS benefits greatly from transcriptomic hints provided by RNA sequencing data in a short-read alignment form. For this reason we used approximately 220 M paired-end reads from the tissue-specific transcriptomic database (including: Pharynx, Crop, Gizzard, Gut/Chloragog, Coelomic fluid, Nerve cord tissues – see Chapter 2) to generate an alignment file which could be supplied to the AUGUSTUS pipeline. The gene prediction process was run as part of the OmicsBox software package environment. One of the downsides of most *Ab-initio* gene prediction pipelines is the difficulty in identifying short genes and the lack of validation in the form of experimental evidence (when there is no available evidential support for a gene). For this reason an Assembly Spliced Alignments based program called PASA was used in conjunction with a transcriptome generated using a Trinity Genome Guided assembly pipeline (Grabherr et al. 2011a).

3.2.14 Phylogenetic analysis

Phylogenetic analysis of gene families was performed using MEGA software (Kumar et al. 2018). Maximum likelihood models evaluated based on the associated BIC scores (Bayesian Information Criterion), which were computed using the standard “Find Best DNA/Protein Models” function of the MEGA X software. Models resulted with the lowest BIC score were considered to describe the substitution model the best (Thomas 2001). The lowest BIC score was identified in the case of the Generalised Time Reversible (GTR) substitution model when combined with Gamma distribution (+G) to account for substitution rate heterogeneity over the different alignment sites and with the assumption that a proportion of invariant sites are evolutionary invariable (+I). Therefore the GTR+G+I model was selected to conduct the phylogenetic analysis with a bootstrapping value set to 1,000. The generated trees then were exported from MEGAX and visualised using the Interactive Tree of Life (iTOL) online software, where nodes with lower than 0.5 (50% recovery) bootstrap values were collapsed (Kumar et al. 2018).

3.3 Results

3.3.1 DNA extraction and long read sequencing

Following the optimisation of the high molecular weight DNA (HMW DNA) extraction protocol, the final DNA sample was extracted from the body wall of a single adult *E. fetida* individual. Using the optimised protocol, a total more than 40 µg HMW DNA was purified with a fragment length higher than 50 KB. Spectrophotometric analysis confirmed that the dialysis based DNA decontamination method removed all the significant chemical contaminants from the samples.

To achieve the highest possible output with reasonable average read length, both the rapid (tagmentase based) and the ligation based nanopore libraries were compared. Since the output from the rapid library test run produced a significantly lower sequencing yield, final libraries were prepared using the ligation method. Excluding the preliminary test runs, a total three different genomic DNA library were prepared from the same HMW DNA sample and sequenced using three Nanopore MinION flow cells. The first run produced 8.8 gb of data with a raw N50 of 10 kb, while the second run produced approximately 10 gb of data with an N50 of 9.8 kb. Finally, the last run generated 9.5 gb of sequence data with an N50 of 9.9 kb. This means that in total around 28.5 gb of Nanopore long-read sequencing data was generated with an overall N50 of 9.9 kb and a median PHRED score of 9.8 (Figure 27, 28). Based on available *E. fetida* C-value measurement this amount of data correspond to approximately 40X coverage for the estimated total genome size.

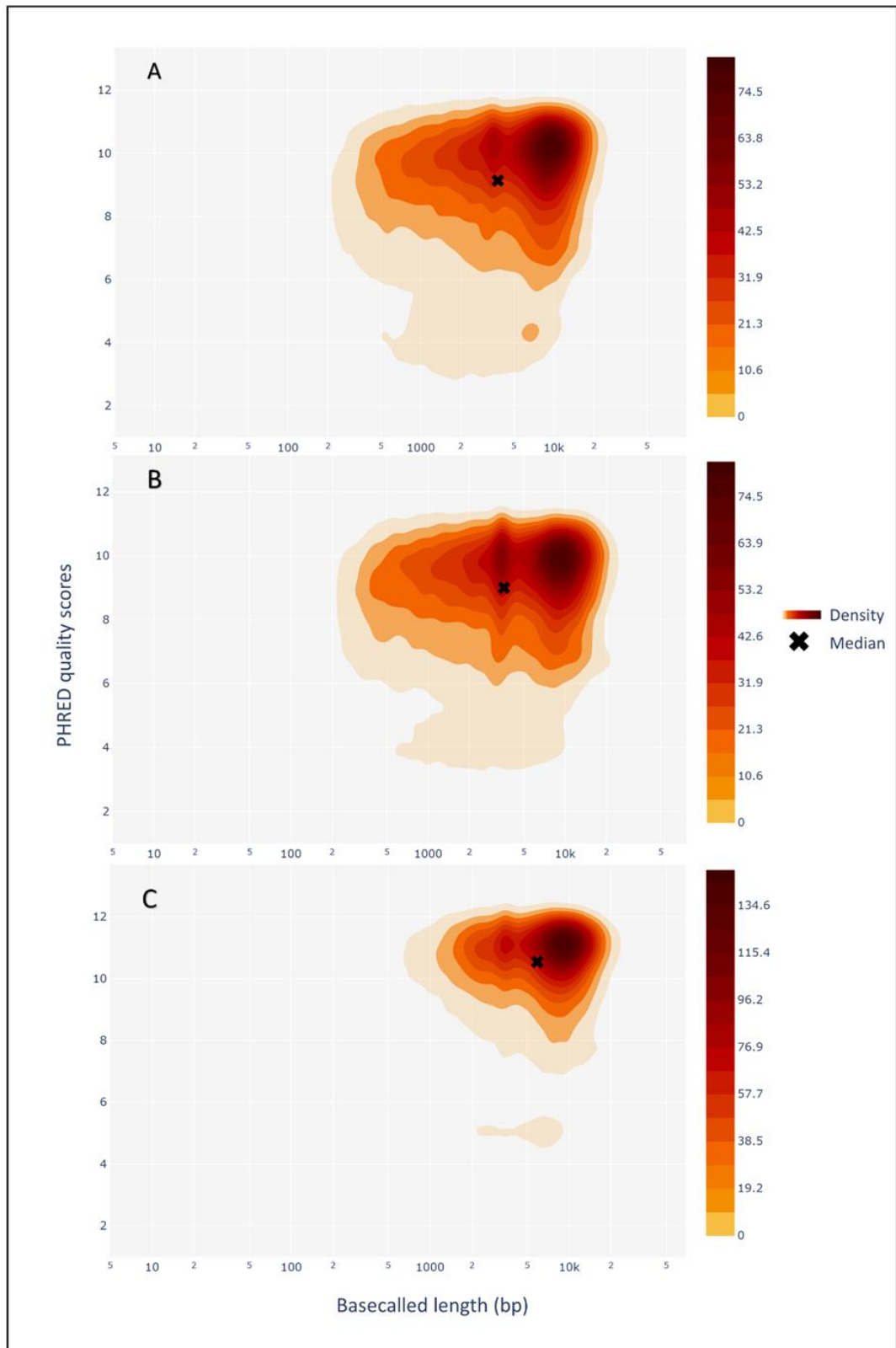


Figure 27: Evaluation of read length and base quality for Nanopore sequencing reads. 2D density plot was produced with PycoQC, where sequenced Nanopore reads were plotted based on read length (bp) and read quality (PHRED score). The first (A) and second (B) run produced fairly similar results while sequencing data generated by the third (C) flow cell resulted in a slightly higher median read length and read quality.

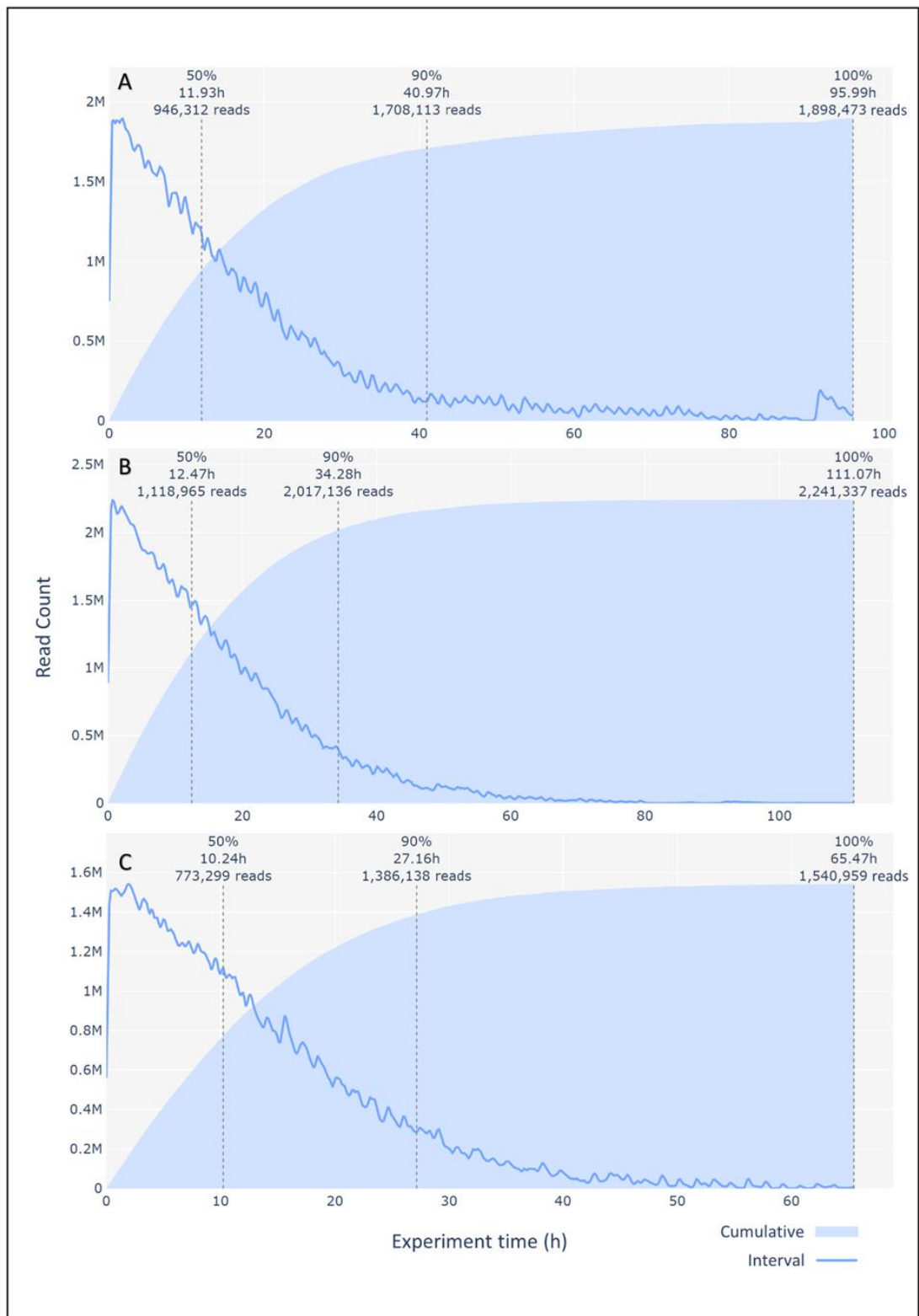


Figure 28: Temporal analysis of sequencing for Nanopore sequencing. Read count histogram generated by PycoQC illustrating the increase in the counts of sequenced Nanopore long reads over experiment time. Both cumulative and interval yields are shown. A, B and C panels represent the first, second, and third sequencing run respectively. Highest output was generated by the second flow-cell both in terms of read numbers and total base content.

3.3.2 Paired-end and Chromium 10X short read sequencing

To be able to correct the long and error prone long reads sequenced with Nanopore, additional 50X coverage worth of paired-end short read data was generated using the genomic DNA from same individual, with an average insert size of 150bp. Nanopore reads were then successfully corrected with short reads using the FMLRC software (Wang et al. 2018). In addition, another 35Gb (approximately 50X coverage) Chromium 10X data was successfully produced using HMW DNA from the same individual as a starting material, for assembly scaffolding purposes.

3.3.3 Initial long read assembly

To achieve a long read based initial *de novo* *E. fetida* genome assembly with high contiguity, several different command line based long read assemblers were tried with multiple parameters. Both in terms of contiguity and completeness the best result was generated by a fuzzy Bruijn graph (FBG) based genome assembler called Wtdgb2 (v2) (Ruan and Li 2019). The initial assembly contained around 22,000 contig with a total size of approximately 1.3 Gb. The N50 of the initial assembly was 175,9 Kb which was improved to 205 Kb following the transcriptome based scaffolding using the L_RNA scaffolder (Xue et al. 2013).

3.3.4 Scaffolding with Nanochrome

The initial assembly was scaffolded using three iterations of Nanochrome scaffolding using the corrected long reads and the chromium paired-end data as input. Since Nanochrome produced a substantial amount of gap in the contigs, following each round semi-scaffolded assemblies were gap closed using a long read based gap closer approach (Xu et al. 2019). The first round of Nanochrome scaffolding resulted assembly had an N50 of 609 kb while this was increased to 999 kb after the second round and reached 1.068 mb after the third scaffolding run was conducted (Figure 29). The number of non-ATGC characters (gaps introduced by the scaffolding) was increased to 1.3 Million following the three rounds of scaffolding and gap closing.

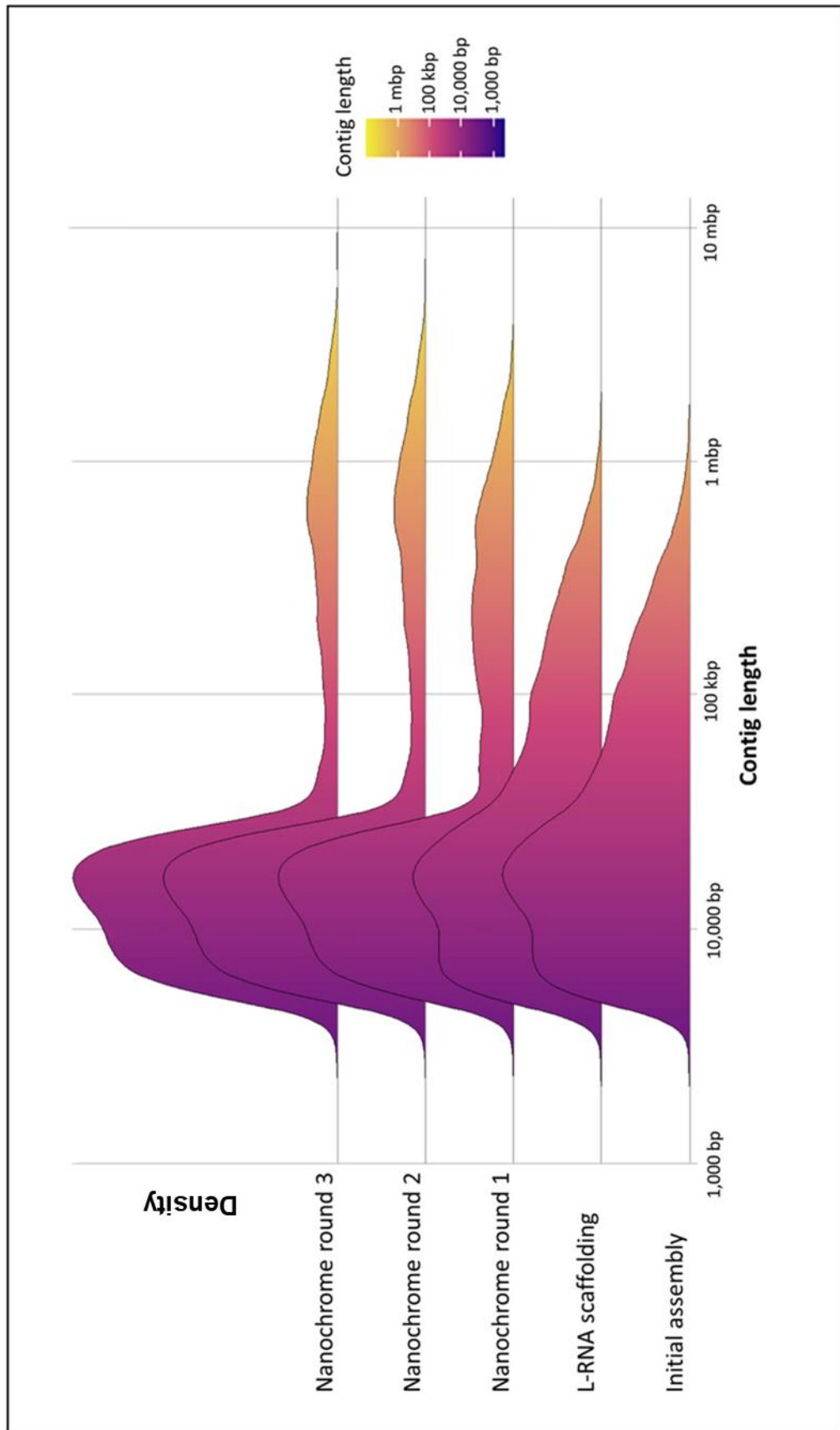


Figure 29: Genome assembly contiguity changes at the different stages of the scaffolding.

3.3.5 Contiguity and completeness of the final assembly

Following the filtering of low coverage and bacterial symbiont related contigs (Figure 30), the final scaffolded and polished assembly contained 7,208 scaffolds with an N50 of 1.09 mb and with the length of the longest scaffold more than 9.2 mb. The mean contig length was 189.2 kb with an 40.8% overall GC content. The completeness of the final assembly was estimated with BUSCO (v.4) using the metazoan database (metazoa_odb10) which successfully identified 906 complete BUSCOs, which corresponds to around 95% overall completeness (Figure 31). From the 906 complete BUSCOs, 862 (90.4%) appeared as single-copy and 44 (4.6%) showed duplication while only 8 (0.8%) BUSCOs were fragmented and 40 (4.2%) were missing. To represent the basic continuity related metrics of the newly generated *E. fetida* genome, a snail plot was generated using the Blobtools2 tool (Figure 32).

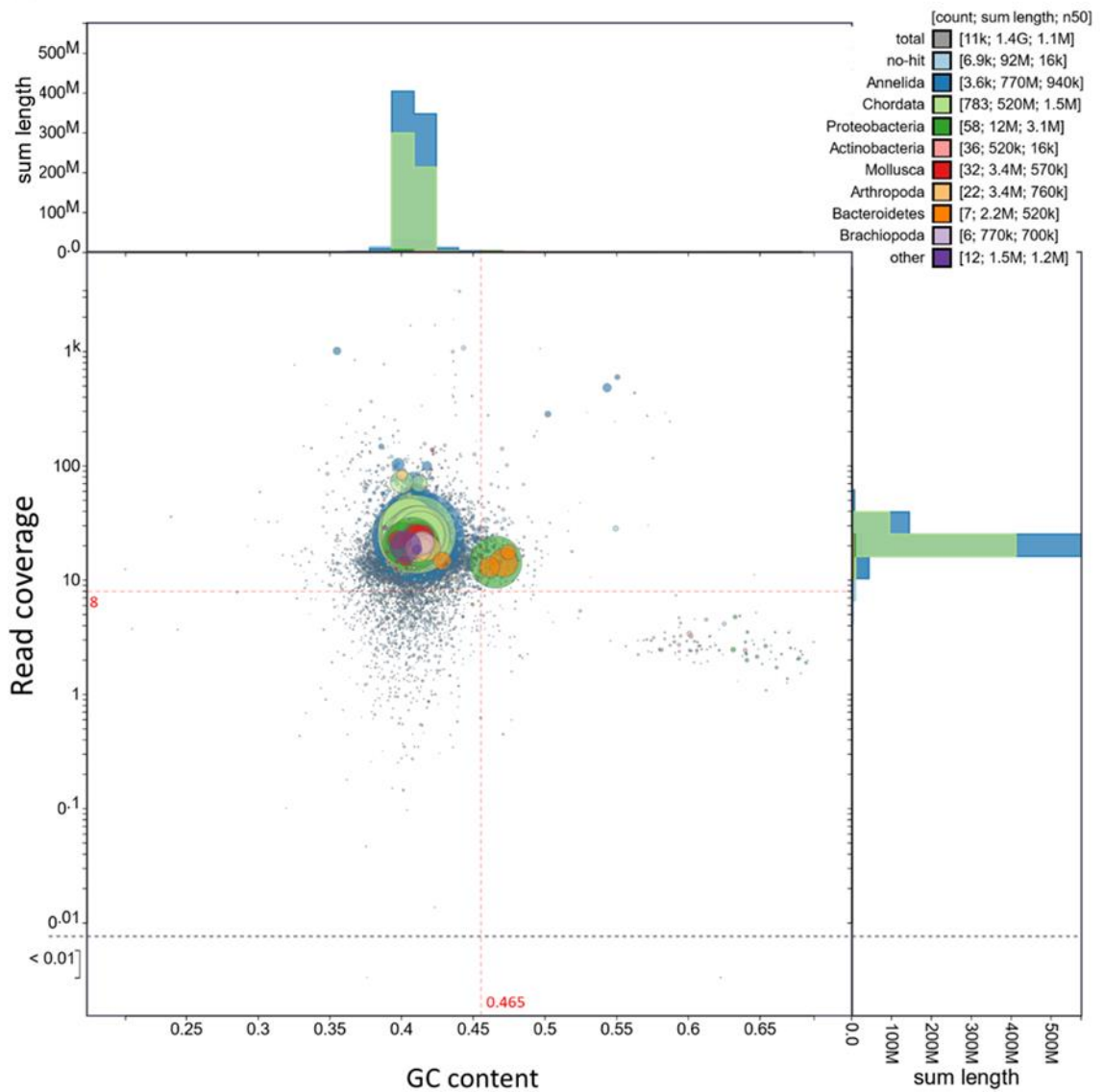


Figure 30: Evaluation of genome assembly. Bubble plot generated by Blobtools2. Each contig is represented by a circle where the size of the contig is indicative of circle size. Contigs also represented by their read coverage from short reads and their GC content. By blasting the contigs against the nr database (NCBI) each contig was associated with a taxonomic group based on the best hit. The different taxonomic groups are represented in the colour of the circles. Red dashed lines shows the cut-off values which were implemented during the assembly filtering step.

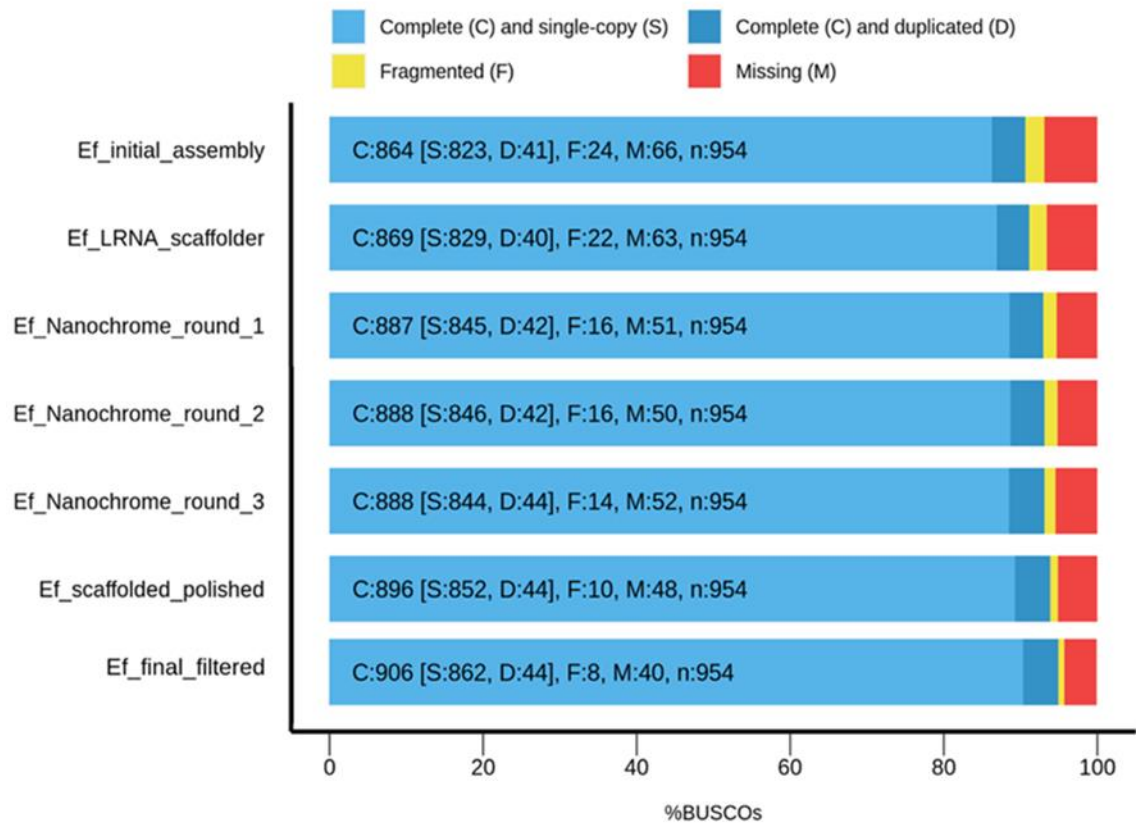


Figure 31: Comparison of the assembly completeness at the different stages of the genome scaffolding workflow. Completeness determined by benchmarking universal single-copy orthologs (BUSCO v.4). Complete (blue) fragmented (yellow), missing (red) proportions of the genome are represented with different colours. Analysis was conducted using the metazoan database (metazoa_odb10).

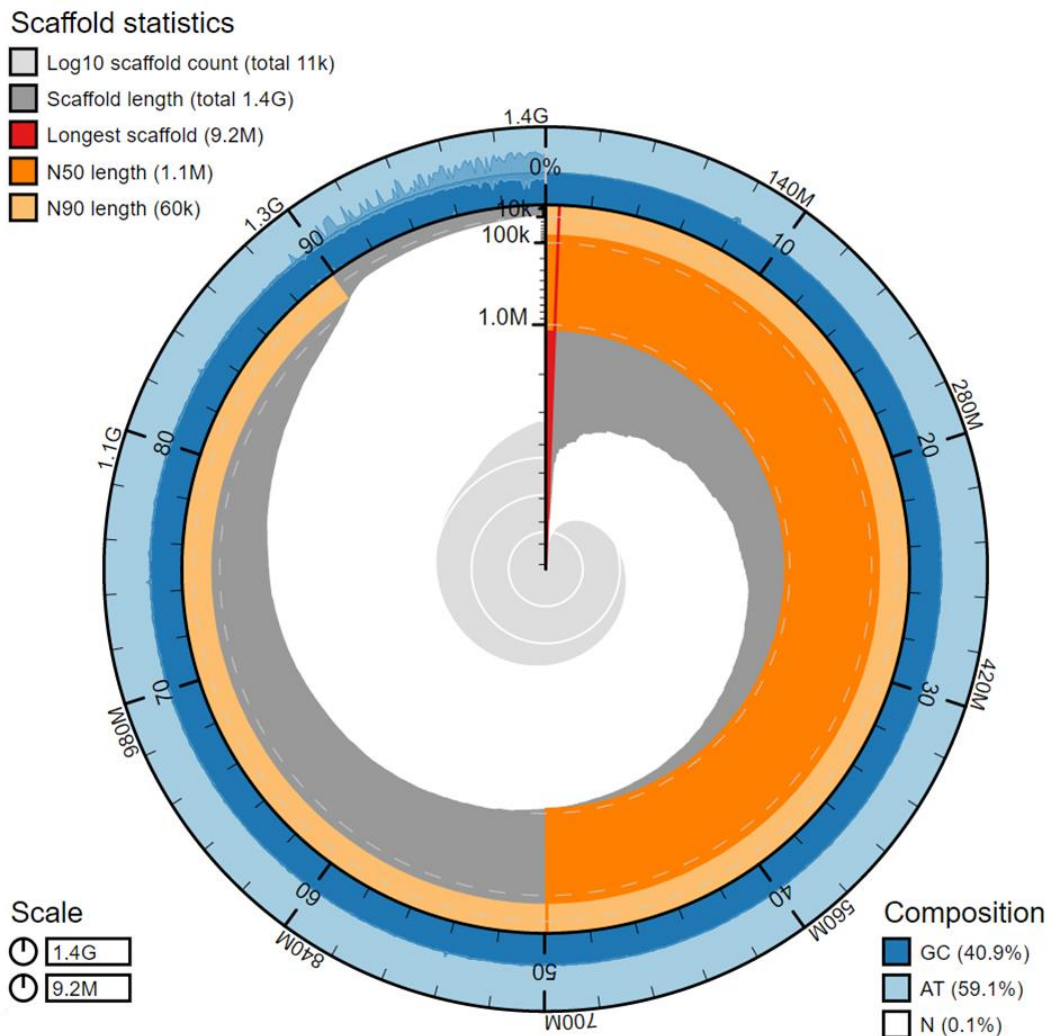


Figure 32: Basic contiguity related metrics of the final, scaffolded and filtered *E. fetida* genome assembly. Data is represented using a Snail plot generated by Blobtools v2. The longest scaffold appeared to be longer than 9.2 Mb. The total number of non-ATGC characters only accounted for approximately 0.1% of the final genome assembly.

3.3.6 Identification of repeats and low complexity regions

To make the downstream *ab-initio* gene prediction analysis more accurate highly repetitive and low complexity regions of the final genome were identified and masked. The identification was conducted in two steps. First, genome assembly was analysed with Repeat Masker using the annotated RepBase repeat database which identified only 5% of the total assembly size as repetitive region (Figure 33). As a second step repeats which more specific to *E. fetida* were annotated using the *de novo* repeat identification repeat Modeller pipeline. The *de novo* repeat identification pipeline showed that approximately 58% of the *E. fetida* genome represents a repetitive and low complexity region from which 33% could not be classified (Figure 33).

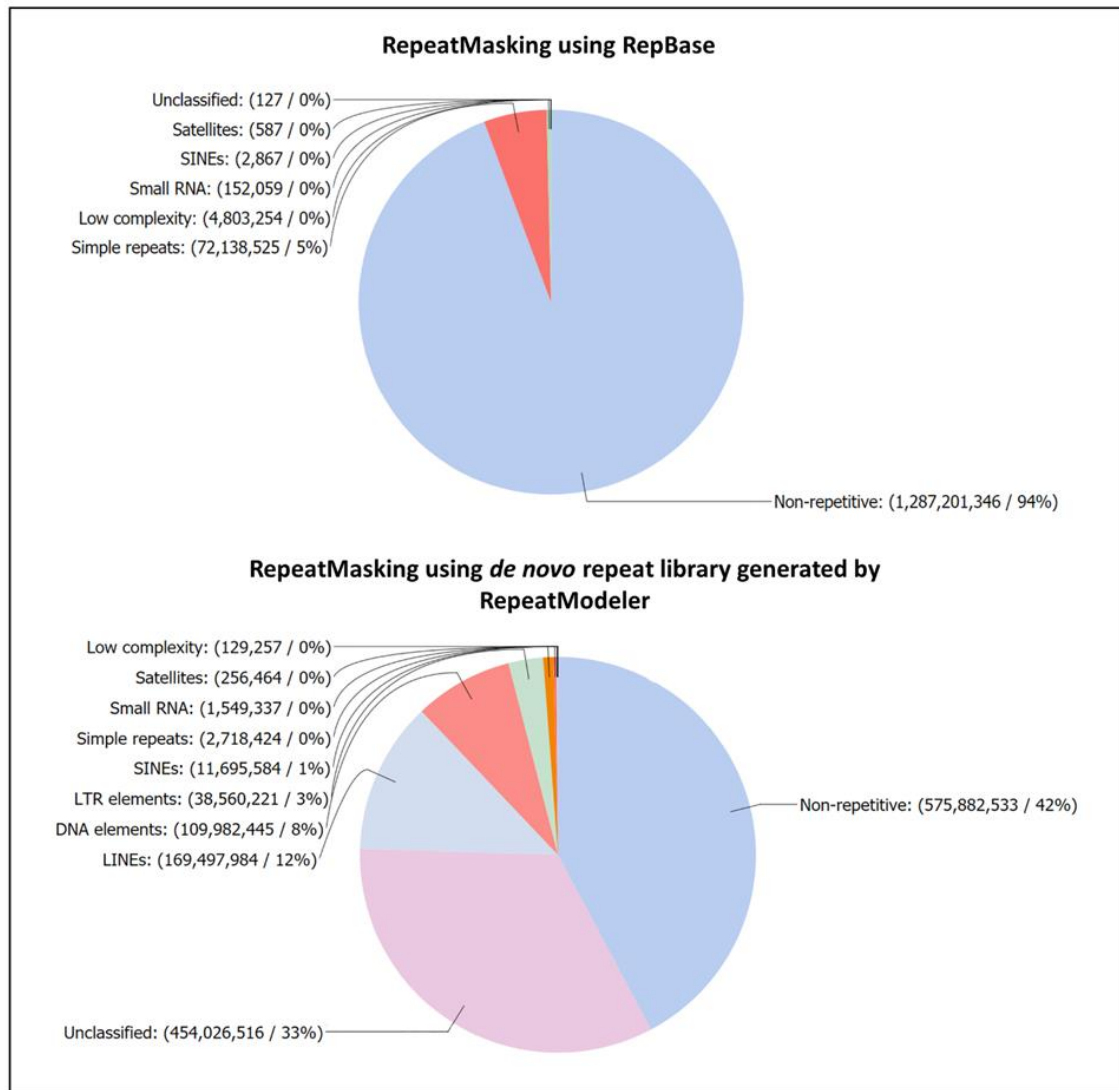


Figure 33: Distribution of the identified repetitive and low complexity elements of the genome, identified based both on the RepBase database and by the *de novo* RepeatModeler pipeline (Smit AFA 2008).

3.3.7 Genome annotation

To annotate the assembled *de novo* reference genome with high completeness different gene prediction bioinformatics methods were trialled. Although utilising the results of the tissue-specific reference transcriptome (Chapter 2) in the assembly of spliced alignments, using PASA, based gene prediction result, the overall best gene annotation performance was achieved with the *ab-initio* based AUGUSTUS gene prediction software. Using the previously generated tissue-specific transcriptomic datasets as a guide for the exon prediction, AUGUSTUS could identify 129,581 transcript sequences on the repeat masked *E. fetida* reference genome. In terms of contiguity, the

new genome based transcriptome reached an N50 of 2163 bases while its completeness measured by BUSCO was more than 90%, with 867 successfully identified BUSCOs. From complete BUSCOs 78% (788) were identified in as single-copy and 12.9% appeared to be duplicated within the transcriptome. The number of missing BUSCOs represented only 3.9% (37). The functional annotation of the identified genes was conducted based on the Uniprot:Swis-prot database (UniProt 2019) using the BLASTx based method, described earlier in Section 2.3.2.

3.3.8 Identification of Toll-like receptor genes

Using the new, highly contiguous, annotated reference genome we could identify 39 putative Toll-like receptors in *E. fetida*, each of which were recovered with full-length coding regions. Following the translation of cDNA sequences to protein, InterProScan was able to identify their characteristic functional domains such as the TIR domain, transmembrane region and leucine-rich repeats. The identified toll like genes then were clustered by their tissue-specific expression profiles based on transcriptomic data from Chapter 2. In general, most tissues showed high specificity in their Toll-like receptor tissue expression with 5 highly expressed in the coelomic fluid, 7 in the pharynx, 1 in the crop, 2, in the gizzard, 3 in the nerve cord and 8 in the gut/chloragogen tissue. The highest number of Toll-like receptors appeared to be expressed in the Gut/Chloragogen tissue, while in the crop and gizzard only a few Toll-like receptors were highly expressed compared to other tissues (Figure 34).

To gain more information about the functional characteristics, the identified Toll-like genes were phylogenetically analysed (4.2.14). The phylogenetic analysis was conducted both using sequences only from the TIR domain (Figure 35) region as well as utilising the whole coding region of the receptors (Figure 36). In both cases, phylogenetic analysis revealed four distinct clad of Toll-like receptors.

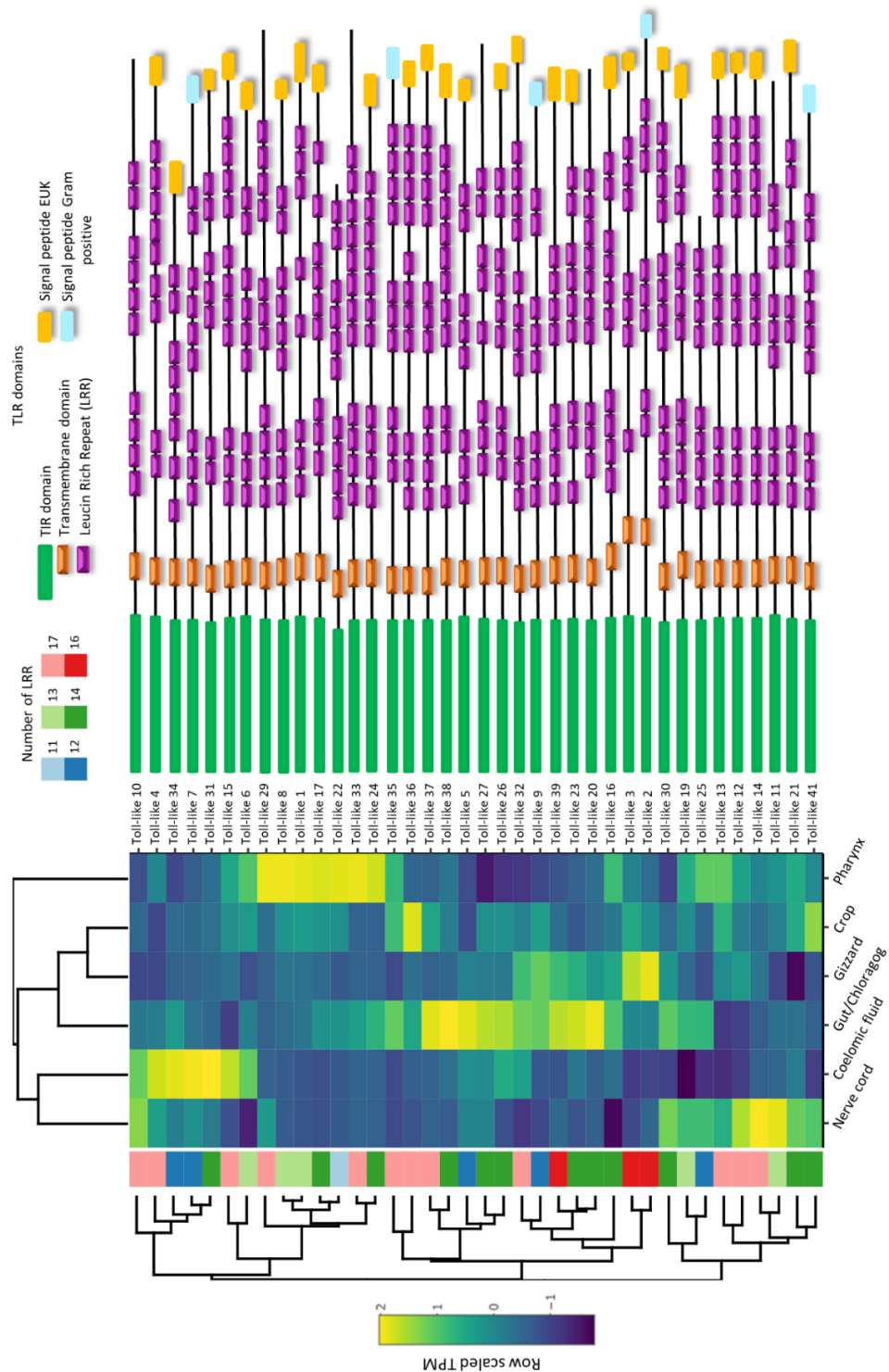


Figure 34: Tissue-specific expression profile of the identified Toll-like receptors. Hierarchical clustering was conducted based on the normalised tissue expression values (TPM), while row-side annotation represents the number of identified leucine-rich repeats (LRR) in the case of each TLR. The main functional domains (Toll/interleukin-1 receptor homologous region - TIR, LRR, Transmembrane domain – TM, signal peptide) of the receptors were identified by the InterProScan module of the Geneious Prime software (v. 2020.1). The relative location of the described domains are represented on a true scale.

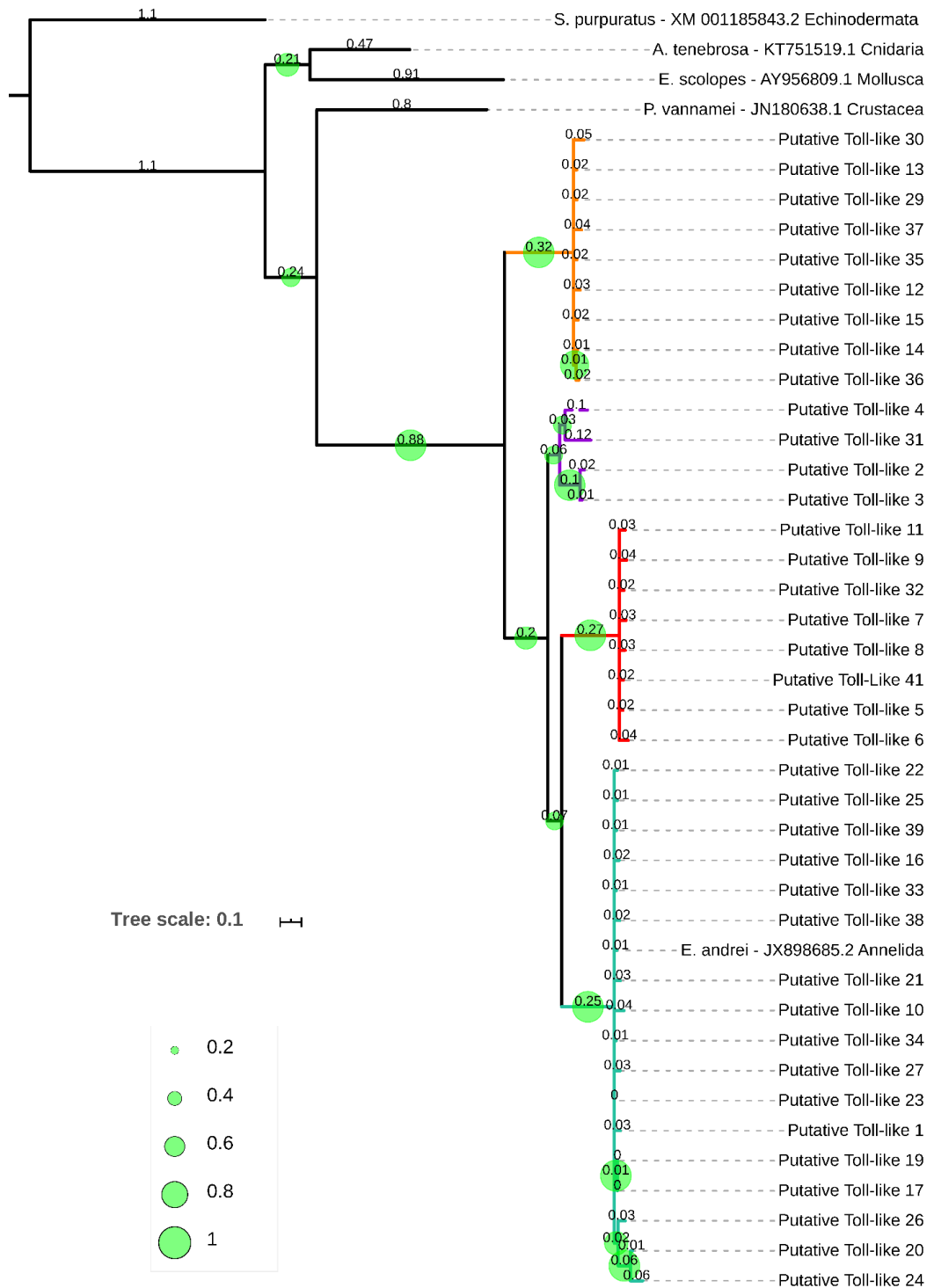


Figure 35: Phylogenetic relationship between the TIR domains of *E. fetida* Toll-like receptor genes. TIR regions of a few invertebrate species (Echinodermata: XM_001185843.2, Cnidaria: KT751519.1, Mollusca: AY956809.1, Crustacea: JN180638.1) are incorporated for comparison. Tree was visualised using the Interactive Tree of Life (iTOL) software (Letunic and Bork 2019). Nodes with lower than 0.5 (50% recovery) bootstrap values were collapsed. An example of TIR domain extracted originated *S. purpuratus* (Echinodermata) was used as root. Numbers represent the branch length while bootstrap values are illustrated by the size of the green circles.

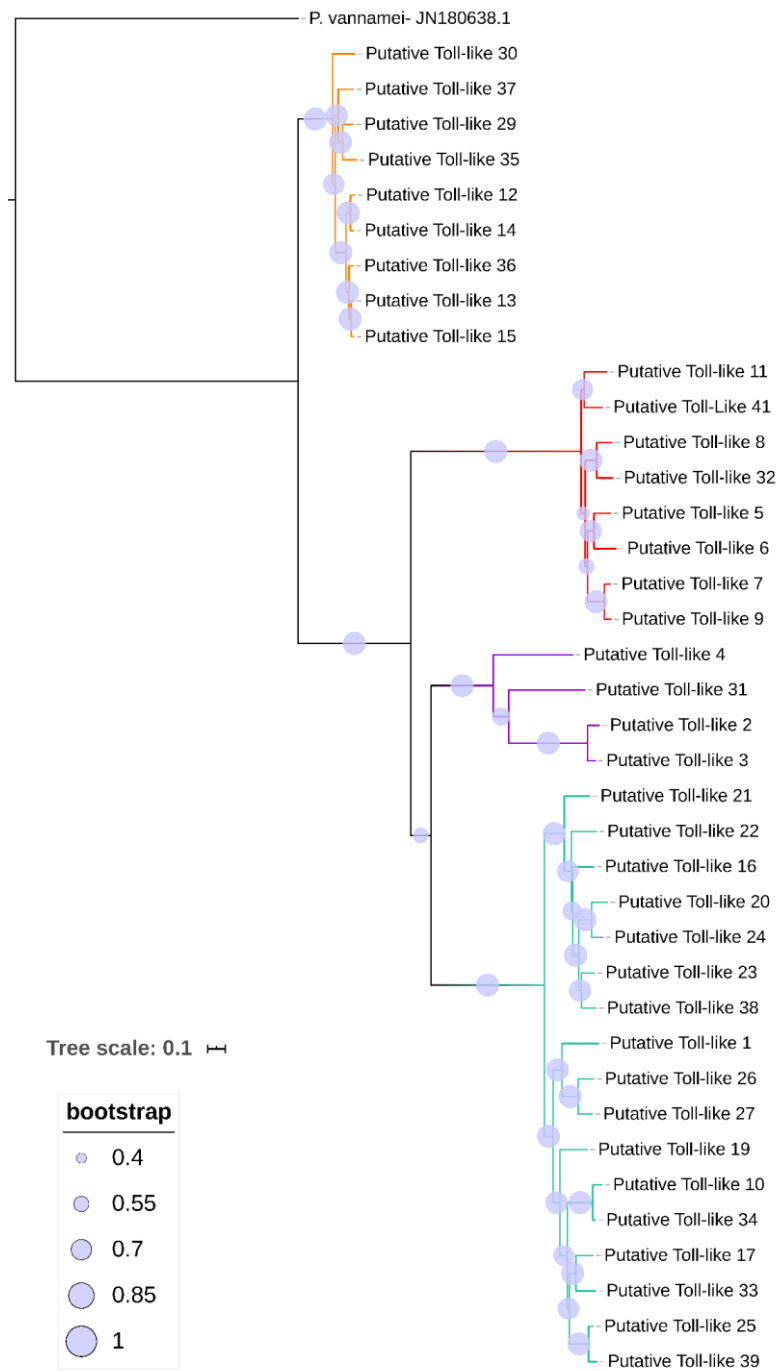


Figure 36: Phylogenetic tree based on the full-length coding region of the identified *E. fetida* TLRs. The tree was rooted using the full-length coding sequence of the *P. vannamei* (Accession ID: JN180638.1) TLR. Bootstrap values are represented by the size of the purple circle. The phylogenetic analysis was conducted using the MEGA software with the General Time Reversible model with Gamma distribution and factoring several sites as Evolutionarily Invariable (GTR+G+I) substitution model and the bootstrapping value set to 1,000. The substitution model was selected based on MEGA predictions (4.2.14). Tree was visualised using the Interactive Tree of Life (iTOL) software (Letunic and Bork 2019). Nodes with lower than 0.5 (50% recovery) bootstrap values were collapsed.

3.4 Discussion

3.4.1 HMW DNA extraction for Nanopore sequencing

During the last decade the advantages of using the new long-read sequencing techniques in *de novo* genome assemblies have become almost unquestionable (Amarasinghe et al. 2020, van Dijk et al. 2018). However, their input material requirements and library preparation protocols still remained in a less well-established stage, compared to the widely used short-read based sequencing techniques. While recently long-read sequencing workflows became more and more standardised for the more generally used model organisms, in the case of non-model organisms the problem of protocol optimisation remained a crucial challenge to solve. Although “spin column” based nucleic acid purification protocols have revolutionized the DNA extraction methods both in terms of yield, hands-on time, and scalability, the quality of the extracted DNA in many cases inadequate for most long-read sequencing techniques. The average length of column purified genomic DNA is usually well below the long-read library requirements, containing a high amount of rather short (only a few hundred bp) length fragments. Another limitation of the column-based technique from the perspective of long read library preparation is the frequent chemical contamination (mainly salts) that originates from washing buffer residues, which is challenging to avoid without significant yield loss.

In contrast to the spin column-based DNA purification techniques, the Phenol-Chloroform based extraction methods provide a viable option - with less physical force caused fragmentation included during the process -to purify genomic DNA with high molecular weight (HMW DNA). Although a relatively high number of phenol-chloroform based DNA extraction protocols are available (Wood 1983), most of them optimised for cell cultures or a relatively high amount of muscle or soft-tissues. To be able to extract HMW DNA from a single earthworm based on this technique, in a quantity that suits the non-PCR based long-read DNA sequencing library preparation protocols, several optimisation and small modification steps were required. By changing the physical tissue homogenizing options to longer but more “gentle” multistage proteinase K based digestion method, the physical force caused fragmentation was minimised. Ethanol and isopropanol based precipitation combined with the manual, glass rod-based precipitate

collection and washing, provided a great alternative to avoid DNA shredding due to centrifugal forces which normally occurs when pelleting with high-speed centrifugation. Although the glass rod-based HMW DNA washing and resuspension method was highly affected by chemical contaminants (ETOH, C₂H₇NO₂, Chloroform), the utilised column-based TE dialysis provided a “gentle” way to both chemically purify and, at the same time, reduce the number of short fragments in the HMW DNA samples. These small technical modifications finally allowed us to produce HMW DNA samples in sufficient quantity and with the required quality.

3.4.2 Nanopore long-read sequencing

Compared to short-read sequencing methods such as Illumina, both the quality and quantity of the output of Nanopore genome sequencing shows great variety and it is highly dependent on multiple factors. The yield of the sequencing run not only dependent on the input quality (both DNA fragment size and chemical composition) but it is also highly dependent on the molarity of the DNA library loaded to the flow cell as well as on the species-specific characteristics (such as repeat content) of the DNA sample. For this reason, several test runs were performed using *E. fetida* HMW DNA samples to be able to optimize the library preparation and the loading of the flow-cells, for long read length with sufficient read quantity. The output of the preliminary sequencing runs was limited to between 3-5 gb, likely due to the high repeat content of the earthworm DNA causing a high rate of pore blockage, resulting in early pore unavailability and low sequencing yields. Using the new updated rev-D flow cells, this problem became less significant and after optimising the molarity of the load library, our final three flow cell generated more than 8 gB of data each.

3.4.3 Assembly and 10X Chromium scaffolding

The use of long-read sequencing combined with chromium linked read scaffolding has allowed us to assemble a high-quality reference genome for *E. fetida*. The combination of the long-read sequencing and the newly established scaffolding method based on Chromium linked reads (Nanochrome) means, compared to the recently published *E. fetida* genome assemblies (Bhambri et al. 2018, Zwarycz et al. 2015), we significantly increased both the contiguity and completeness while decreasing the non ATGC content of the assembly (Table 4). The reference genome described in this chapter shows a more

than hundredfold increase in scaffold N50 and an increase from completed BUSCOs from 14% to 95%, when compared to the best published *E. fetida* genome assembly (Figure 37) (Zwarycz et al. 2015, Bhambri et al. 2018).

Table 4: Assembly statistics summarising table, comparing the two most recent, short read based *E. fetida* genome assembly with the newly generated long read based reference genome.

Parameter	Zwarycz et al. 2015	Bhambri et al. 2017	Eisenia LR final
Assembly size (Gb)	1	1.4	1.3
Total gap (nonATGC in contigs)	314,083,977	1,016,501,040	1,350,056
Contig N50 (Kb)	0.2	1	205
Scaffold N50 (Kb)	1	9	1092
Number of scaffold	1,659,527	399,006	7,208

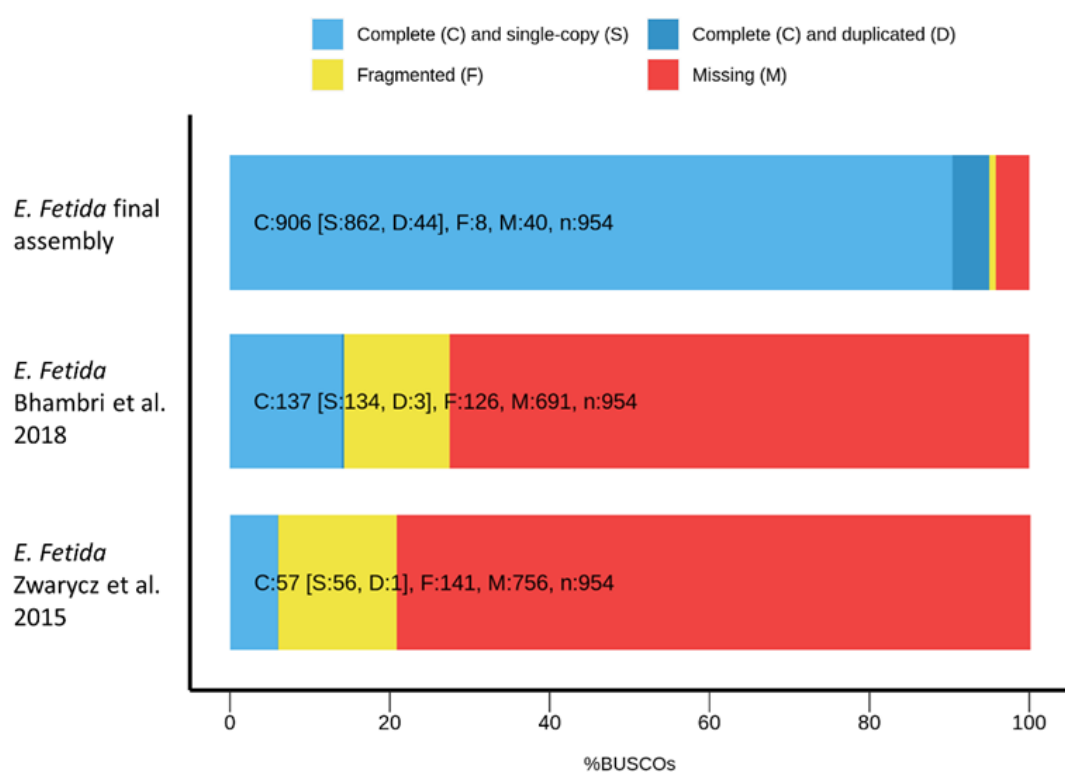


Figure 37: Comparison of the assembly completeness of the two recently published *E. fetida* genomes with our scaffolded and filtered, long-read based *E. fetida* assembly. The different colours represent the number of complete (blue), fragmented (yellow), and missing (red) BUSCOs, where BUSCOs were identified using the metazoan database.

3.4.4 Genome annotation

Using our tissue-specific transcriptomic datasets, we managed to predict approximately 130,000 mRNA sequences in the genome, from which approximately 30% were functionally annotated. This created the opportunity to refine our reference transcriptome described in chapter 2. Although the number of identified transcripts were in a similar range to the de-novo transcriptomic results described in chapter 2, as illustrated in Figure 38, the genome-based transcriptome showed a huge (1.5X) improvement in terms of overall transcript contiguity compared to the earlier described reference transcriptomes (Chapter 2). This suggests that the genome-based methods recovered transcripts in full length with a much higher success rate compared to the *de novo* transcriptomic approach. Although the number of functionally annotated transcripts appeared to be in a similar range between the genome-based (~38,000) and tissue-specific reference assemblies (*E. fetida*: ~30,000 *E. andrei*: ~22,000), the results of the transcript coding region prediction pointed out an important difference. While in RNASeq based transcriptomes only 44% of the transcripts appeared to be in a complete form (both start and stop codons were detected), this number showed a high increase in the genome originated transcriptome where it reached 83% (Figure 39). This seems to support that the increase in the N50 was the result of a better and more complete transcript length recovery, rather than results of a possible gene prediction inaccuracy.

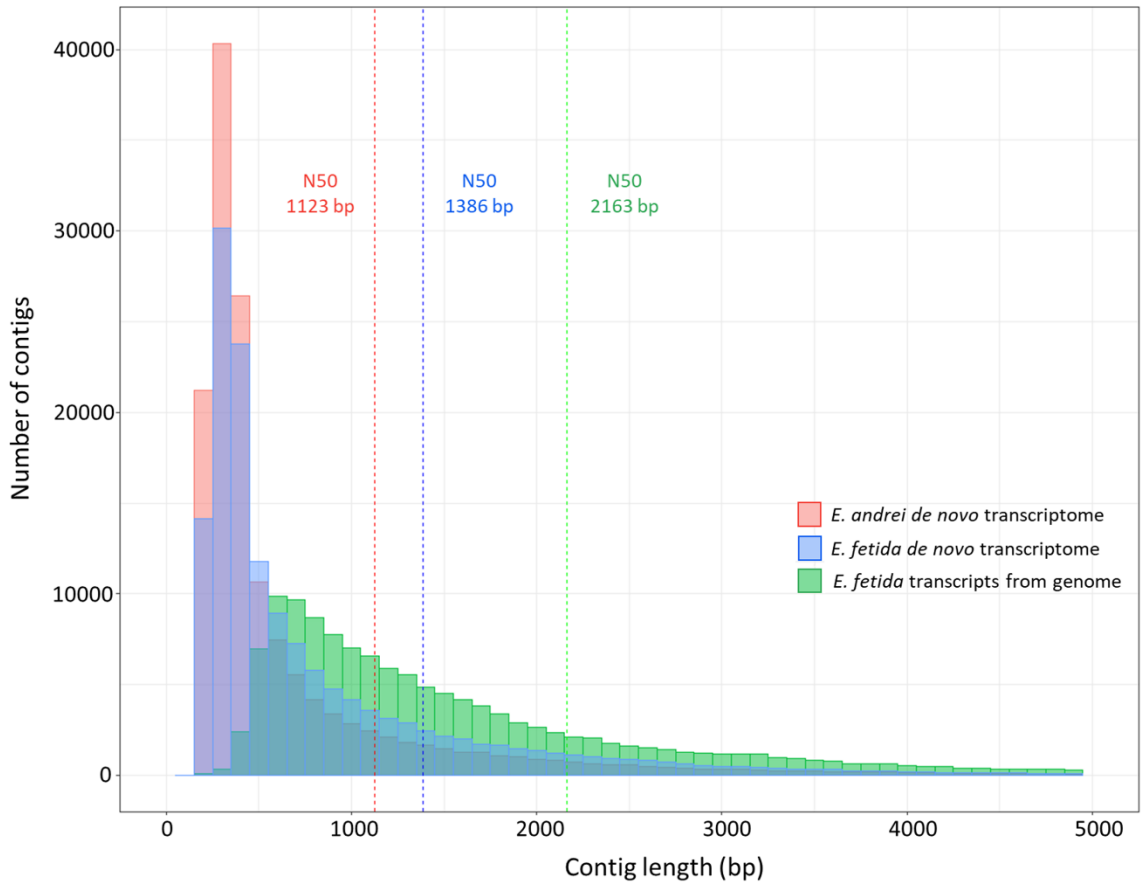


Figure 38: Size distribution of the *in silico* identified transcriptome using the annotated reference genome, compared to the *E. fetida* and *E. andrei de novo* tissue-specific transcriptome described in Chapter 2. Transcriptome generated based on the genome assembly resulted higher N50 with significantly less number of small (100 -400 bp) length transcripts.

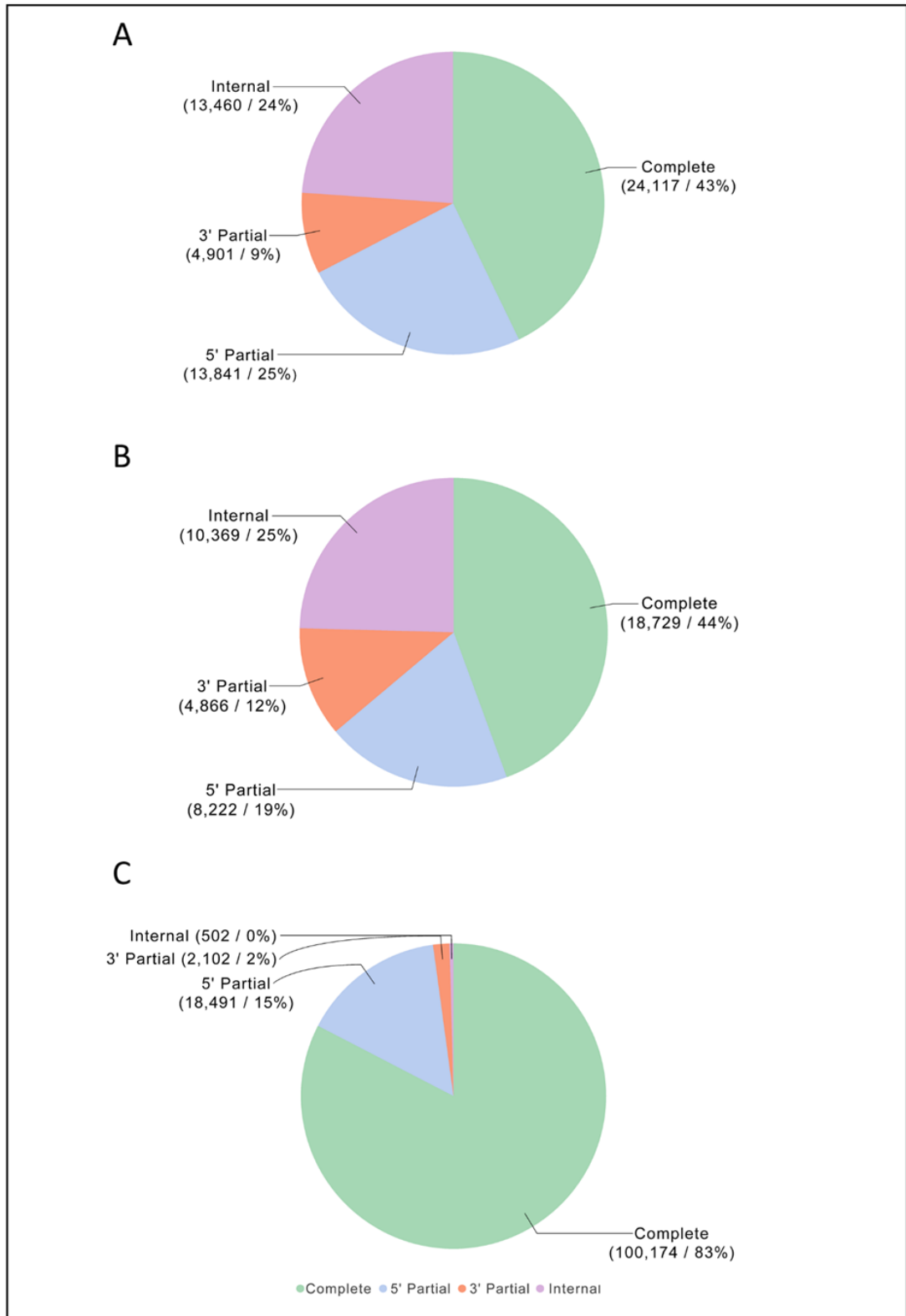


Figure 39: Transcript structural completeness, measured by The TransDecoder module of the OmicsBox software package (v. 1.3). The predicted completeness of the genome-based transcriptome (A) was compared with both the *E. andrei* (A) and *E. fetida* (B) tissue-specific transcriptomes. A transcript was defined as complete when both START and STOP codon could be identified.

3.4.5 Identification of Toll-like receptor genes

Toll-like receptors are one of the main pillars of innate immunity conserved across a wide phylogenetic range. This receptor super-family were successfully characterised from several low order animals such as species of the Porifera (Wiens et al. 2007) or Cnidaria phylum (Brennan and Gilmore 2018) to high order vertebrates (Roach et al. 2005). The last few years of invertebrate immunology research have revealed that many invertebrate species have an increased number of Toll-like receptors compared to most vertebrate species. A particularly interesting example found the sea urchin genome appeared to possess hundreds of TLR candidates (Rast et al. 2006). A relatively high number of TLR receptor sequences were also identified in the Annelida phylum, for example, based on the bioinformatic analysis of the *Capitellate capitata* and *Helobdella robusta* genome, 105 and 16 putative TLR receptor were identified in these species respectively (Davidson et al. 2008). Although the first earthworm TLR was identified in *E. andrei* and there was a possibility that the *E. andrei* genome contained a relatively high number of TLRs (Prochazkova et al. 2019a), until now only two TLR sequences were described in the *Eisenia* genus (Škanta et al. 2013).

Although the large number of TLRs was already partially represented in our tissue-specific reference transcriptomes, their applications in read counts based functional studies appeared to be extremely difficult. The *de novo* transcriptomic limitations described in Chapter 2, resulted in a large number of highly fragmented TLR contigs, where differentiating between gene paralogs, splice variants, and assembly artifacts would be an extreme challenge. The new, highly complete, contiguous and annotated *E. fetida* genome described in this chapter has helped overcome the mentioned limitations almost completely, revealing the coding regions of 39 Toll-like receptors with high completeness (i.e. from the start to stop codon).

The combination of the tissue-specific transcriptomic resources with the newly identified TLR genes created the first insight into how the different earthworm tissues rely on different sets of TLRs. The different groups of TLRs showed a clear, distinct tissue-specific expression profile. While the phylogenetic analysis expanding the gene expression-based functional classification with evolutionary relationships between the identified clades of *E. fetida* TLRs. Due to their ability of recognise different molecular

patterns on a nanometric scale and initiate an inflammatory processes, previously TLRs were described as possible targets of the NPs induced immune response (Roy et al. 2014, Chen et al. 2013). Although changes in TLR expression due to metal nanoparticle exposure was observed in earthworms earlier (Bodó et al. 2020), the biological processes behind this is still yet to be described. The newly identified TLR genes can be utilised to achieve a precise differentiation within the different TLRs, which can help understand the exact biological mechanisms behind the interaction of the earthworm TLRs and engineered NPs.

3.4.6 Concluding key results and possible future developments

Although this chapter aimed to provide a high-quality genome for *E. fetida* to overcome the limitations of *de novo* reference transcriptome described in Chapter 2, it not only helped to generate a refined and more contiguous reference transcriptome but also created the possibility of characterising one of the cornerstones of innate immunity-related gene family (TLRs). This study created the first opportunity to get an insight into the tissue-specific expression and phylogeny of a large number of intraspecies gene variants in the *E. fetida* TLR family. The generated highly contiguous and complete reference genome opens new possibilities. By greatly increasing the number of known target genes, it will be possible to not only expand the field of earthworm immunology but all molecular biology-based research in earthworms. Due to the completeness of the newly identified gene objects, the results of this chapter could provide a more sophisticated template to conduct the gene expression analysis described in Chapter 4 and 5.

4 Characterisation of the Coelomic fluid

4.1 Introduction

For the majority of multicellular invertebrate species, a significant part of immune system processes is based on cellular responses and host-pathogen cell-cell interactions. It is also evident that certain cell-types are responsible for different inflammation-related functional mechanisms. For this reason, to achieve a comprehensive and detailed understanding of the earthworm immune system, it is essential to gain more knowledge about the roles of the different free-floating coelomocyte populations in the context of inflammatory processes. In the case of earthworms, information about the immune role played by specific cell types is rather limited (Loker et al. 2004). During the last decade, cytochemical and histological studies have provided relatively detailed knowledge about the morphological and histochemical characteristics of the three major coelomocyte types, which has given us a descriptive insight into their functional roles. However, since most molecular-immunology studies of the coelomocytes use a mixed cell population model, our knowledge of the distinct immunological characteristics of the specific cell types on the transcriptomic, proteomic and metabolic level is limited. Therefore, our understanding about which cellular recognition and effector related functional pathways are associated with certain coelomocytes populations is restricted. We have insufficient information about which cell types are involved in the effector processes or responsible for the recognition of the different challenging agents. We know even less about their origin, differentiation, interactions and signalling between the different cell types. There is currently a dearth of research aimed at identifying the functional fingerprints of the morphologically distinct coelomocyte populations at the transcriptomic and genomic level. Some of the earliest information about the transcriptomic characteristics of the individual coelomocyte populations was provided by Bodó *et al.* (2018), revealing gene expression level differences between the two major coelomocyte populations (eleocytes, amoebocytes) by measuring the expression of several immune-related genes. Although this publication generated a useful starting point for understanding the immunological functions of the different cell-types, it was restricted to the study of just a handful of immunity-related genes using semiquantitative RT-PCR, and therefore provided a relatively narrow insight into the specific functions of immune cell types. Another limitation of this study was that

although it used a light scatter-based cell sorting approach to distinguish between coelomocyte populations, it only focused to separate the mixed population of amoebocytes (both containing hyaline and granular) from the eleocytes. For this reason, we still do not have information about the transcriptomic differences between the two cytochemistry distinct amoebocyte subpopulation.

The current chapter describes the transcriptomic fingerprints of the three major coelomocytes populations in *E. fetida*, which were separated by their physical characteristics (size and granulation) using a fluorescence-activated cell sorting (FACS) method. A new coelomic cell washing method was developed to improve the production of single-cell suspension using whole coelomic fluid samples as starting material, performed without applying any centrifugation-based method that can cause cell damage and sample degradation. This chapter focuses on innate immunity-related molecular functions and biological pathways associated with the certain subpopulations of the earthworm coelomocytes. Using the distinct transcriptomes, we can identify transcripts presenting high expression in one cell-type relative to the other coelomocyte cell populations, as well comparing expression across different tissues using the tissue-specific dataset described in Chapter 2. This allows us to identify the key cell-type-specific immunological markers and analyse the main immunological characteristics of the different coelomocyte populations.

4.2 Materials and Methods

4.2.1 Experimental design

The earthworms sacrificed within this study (*E. fetida*) were supplied from a genotyped laboratory culture maintained at the UK Centre of Ecology and Hydrology (Wallingford, UK). Prior to cell extraction earthworms were placed in a petri dish containing moist filter paper for depuration purposes. Following the depuration, earthworms were quickly submerged and washed using 1X phosphate-buffered saline (PBS) to remove any left-over soil contamination and reduce the mucus content on the body surface. To collect the coelomic fluid earthworms were placed into individual petri dishes containing 1 ml of 1xPBS supplemented with 5 mM EDTA solution. To induce the extrusion of the coelomic fluid, earthworms were stimulated using current from a fresh 9V battery in the case of each individual. The extruded coelomic fluid then was well mixed with the PBS solution using a wide-bore pipette tip and immediately placed into ice in a pre-chilled 1.5 ml centrifuge tube.

4.2.2 Cell preparation for sorting

To increase the efficiency and purity of the cell separation, three different cell preparation methods were tested to achieve a well-distributed single coelomocyte suspension. During the initial optimisation, centrifugation based cell washing steps, described by Škanta *at al.* (2016), were employed. However, even short-time low speed centrifugation (~100 g) resulted in a discernible and relatively high percentage of cell loss, especially in the case of the eleocyte population. Alternative methods employing gentle centrifugation (~100 g) prior to washing generated a solid cell pellet that resulted in a huge decrease in the number of viable eleocyte, even when applying gentle resuspension using a wide bore pipette tip. Washing the coelomocytes using centrifugation also lead to a considerable number of cell aggregates. To address these problems, several cell washing methods were tested that did not rely on centrifugation, such as sedimentation and filtering. The best results were achieved with optimised filtering using different sized plurStrainer cell strainers (plurSelect, Germany). The extruded cell suspension (~1 ml in volume) was initially passed through a strainer with 100 µm mesh size to eliminate any soil content contamination alongside with the mucus released by the earthworm due to the electric stimulus. The collected flowthrough was

then transferred to a pre-moistened new strainer with a mesh size of 1 μm , which had been attached to a 50 ml conical tube. After transferring the cell suspension to the filter, the conical tubes were placed on a tabletop shaker and agitated gently to avoid clogging the strainer. During this process, the cell-free flowthrough was collected in the conical tube while the coelomocytes suspension was concentrated at the top of the strainer. Cell washing was achieved after filtration by adding ice-cold cell extraction buffer, approximately eighty percent of the original volume ($\sim 800 \mu\text{l}$), to the top of the filter. The washing process was repeated until the yellowish colour of the flow-through had completely disappeared, indicating the increased purity of the cell suspension. Finally, the concentrated coelomocyte suspension was collected from the top of the filter, and the volume was adjusted to 1 ml with fresh, ice-cold cell extraction buffer.

4.2.3 Cell sorting

Freshly collected coelomic cells were separated using flow cytometry into the three major coelomocyte subpopulations (granular amoebocytes, hyaline amoebocytes, eleocytes). Coelomocytes were sorted on a BD FACS Aria Fusion cell sorter (Miltenyi Biotec Ltd, Bergisch Gladbach, Germany) according to their forward scatter/side scatter (488 nm Blue laser), which represents their cellular size and granular characteristics. Photomultiplier tubes (PMT) voltages were adjusted to ensure the best separation. Based on microscopic inspection of the sorted amoebocytes populations (haemocytometer), the purity of the isolated subpopulation was $>97\%$ for granular and $>98\%$ for hyaline amoebocytes. As the post sorting visual examination of eleocytes was not efficient due to the cell death caused by the physical impact, a bioinformatic approach was used to assess the efficiency of the separation. The gene expression profiles of the amoebocyte-specific markers (Bodó et al. 2018) allowed us to estimate the purity of the isolated eleocyte subpopulation, which was $>99\%$. To avoid non-self-recognition processes, coelomocyte samples from different individuals were treated separately before cell-sorting. In the case of granular and hyaline amoebocytes, around 150,000 cells were collected to generate an RNA sample using three different individual coelomocyte samples (50,000 cells per animal). This number was doubled for eleocytes due to their lower RNA content. Isolated cells from all cell population types were placed directly into cell lysis buffer and frozen at $-80 \text{ }^\circ\text{C}$.

4.2.4 Library preparation and sequencing

To increase the yield of the RNA extractions, samples (sorted cells in lysis buffer) were first passed through a QIAshredder column (Qiagen, Hilden, Germany). Then RNA extraction was carried out using the RNeasy Plus Mini kit from (Qiagen) following the manufacturer's instructions (Qiagen, Hilden, Germany). The quality and quantity of the RNA samples were measured by running a High Sensitivity RNA ScreenTape (Agilent 2200 TapeStation, Agilent). Following the RNA extractions, three replicate RNA samples (originated from the same cell population) were combined for library generation for each cell type. The cDNA libraries were generated from approximately 100 ng of RNA for each sample, used by the KAPA mRNA HyperPrep kit (Roche Ltd, UK) according to the manufacturer's instructions. Transcriptome sequencing was conducted on an Illumina NextSeq 500 system using both a medium and a high output flow cells. Operation of the sequencing platform was performed by Ms Angela Marchbank, Cardiff School of Biosciences Genomics Hub, Cardiff University.

4.2.5 Data analysis

After the library preparation and sequencing, the ~30 M (75 bp) paired-end reads generated per sample were subsequently quality filtered and trimmed (Trimmomatic) (Bolger et al. 2014). The cell-type specific expression profiles were analysed after count normalisation with RSEM (v. 1.3.3) using our annotated genome as a reference for the mapping (Li and Dewey 2011). To study the transcriptomic characteristics of the separated cell populations, TMM normalised counts were used to identify transcripts that showed minimum 10X higher expression in certain cell populations when compared to other cell populations and tissues, using the transcriptomic data from chapter 2.

4.2.6 Functional annotation and enrichment

Cell-type "specific" transcripts were then filtered, and only those which had functional annotation were further processed for Gene Ontology (GO) term and pathway enrichment analysis. GO term enrichment analysis was conducted with gProfiler (Reimand et al. 2019). The list of annotated genes from our earthworm tissue atlas was used as background and P-values were corrected using the Benjamini-Hochberg method (Haynes 2013). To reduce the redundancy and to be able to summarise the high number of GO terms, any terms associated with biological processes were functionally grouped

on their term–term similarity statistics using Revigo (Supek et al. 2011) and final graphical presentation prepared using ggplot2 in R (Wickham 2016). Biological pathway enrichment was further refined using ClueGo engine in Cytoscape environment with a statistical filter of 0.05 FDR (Bindea et al. 2009). Pathway enrichment analysis was conducted using the KEGG (Kanehisa et al. 2015), WikiPathways (Slenter et al. 2017), and Reactome databases (Jassal et al. 2020) following similar term grouping methods between the different pathway databases.

4.2.7 Analysis overview

A summary figure (Figure 40) was generated to provide an overview about both the key wetlab methodological steps as well as highlight the utilised bioinformatic tools, which were used during the *in-silico* data analysis.

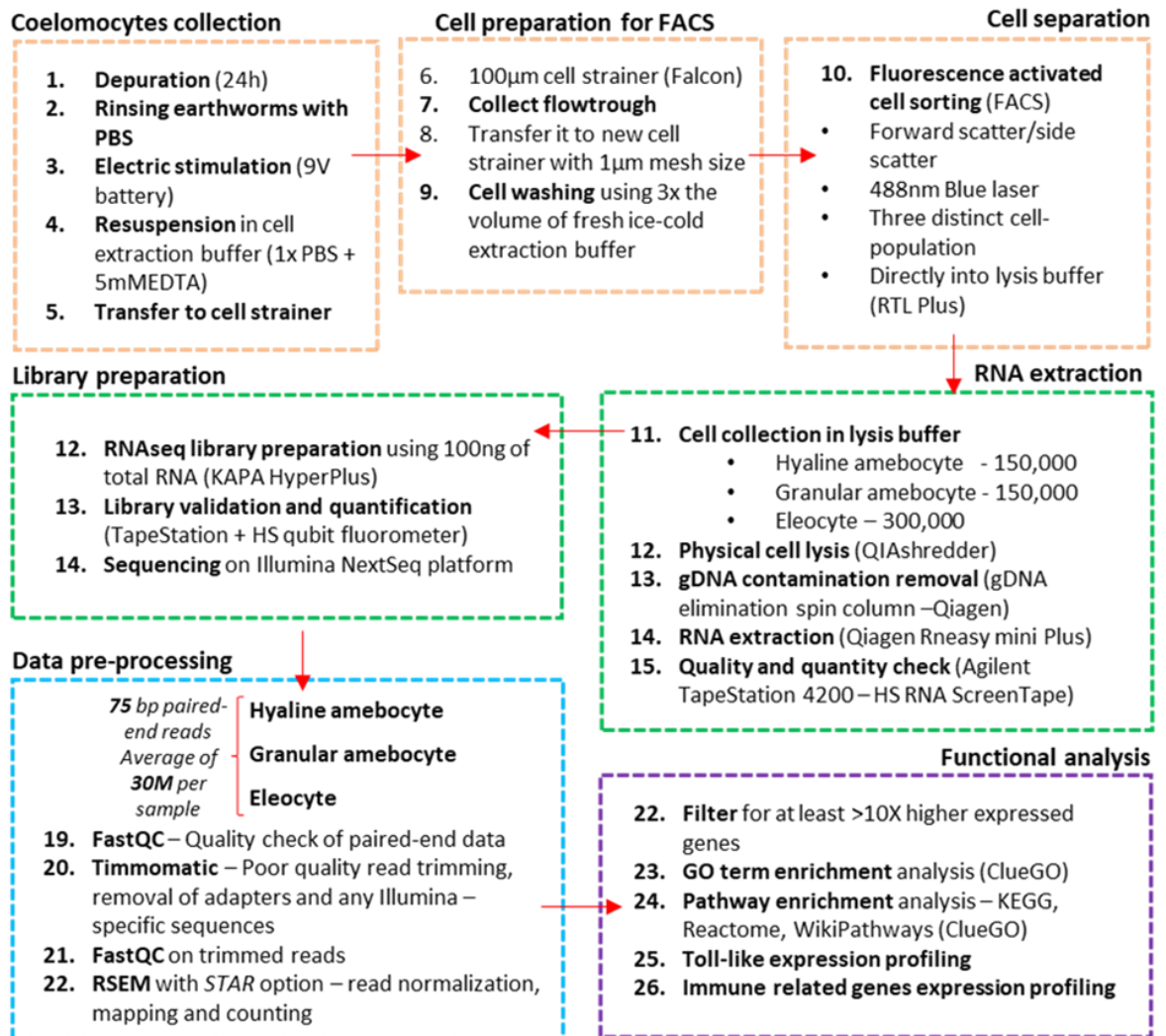


Figure 40: Workflow of the conducted Wet-lab sample preparation steps and bioinformatic analysis.

4.3 Results

4.3.1 Coelomocyte preparation and sorting

The optimised cell strainer based coelomocyte preparation protocol achieved a clear separation between the different cell populations. All three coelomocyte populations (hyaline amoebocytes, granular amoebocytes, eleocytes) could reproducibly separate based on their physical characteristics and light scattering properties (Figure 41). Although the ratio between the different cell populations showed a slight variation between cell samples from different individuals, average values taken over 10 test cell separation runs (conducted using cells from single individuals) reveal eleocytes represented the largest proportion of the detected events (~43%) while hyaline amoebocytes (~20%) and granular amoebocytes (16%) were represented in smaller but nearly equal proportions (Figure 42).

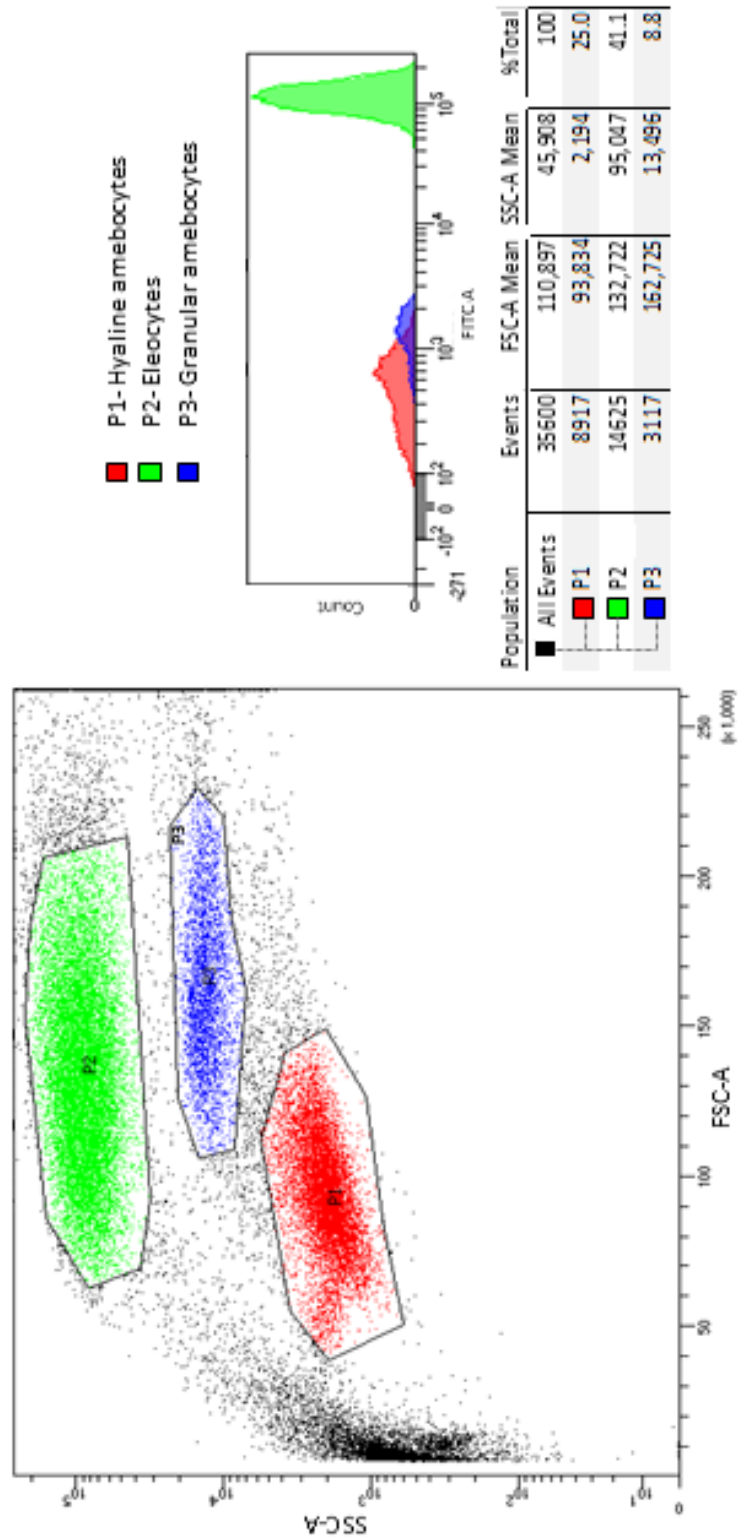


Figure 41: Example of *E. fetida* coelomocyte FACS sorting. Different colours represent the distinct free-floating coelomocytes populations, P1 (red) Hyaline amoebocytes, P2 (green) eleocytes, and P3 (blue) Granular amoebocytes. Coelomocytes were separated by the forward (FSC-A, relates to cell size) and side scatter (SSC-A, relates to granularity). Tables illustrate the basic statistical values regarding the distribution of the detected events.

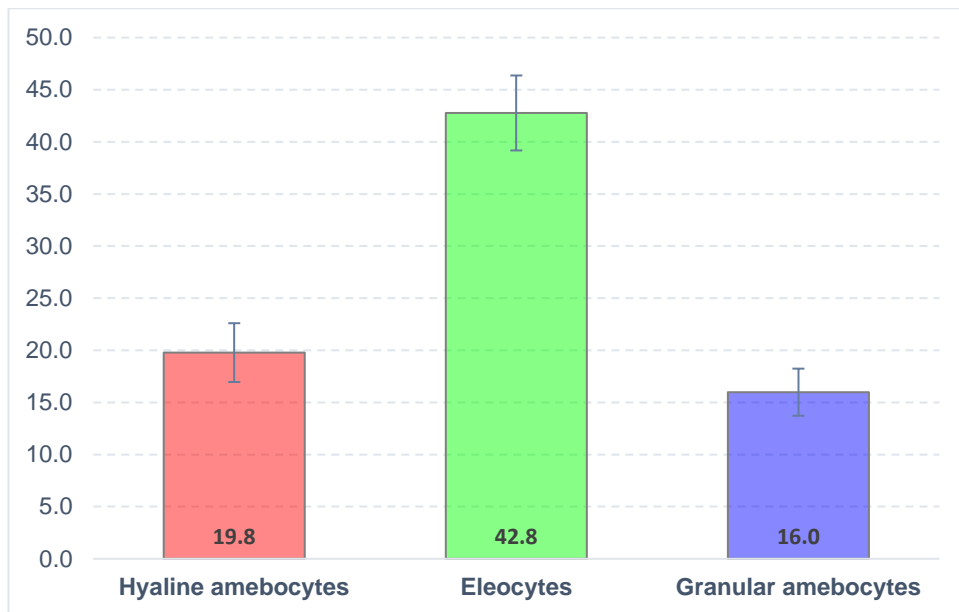


Figure 42: Average proportion of the different coelomocytes from the total detected events. Average values represent the sorting statistics of separation of ten independent cell separations, where error bars represents standard deviation for each cell population.

4.3.2 Sequencing and read processing

In total around 30 M paired-end reads were generated in the case of each cell-type, with an average insert size of 75 bp. Following the removal of the adaptor related and low quality sequences, on average approximately 5% read loss was observed (Figure 43A). Mapping the trimmed reads to the annotated reference genome described in Chapter 3 (3.3.5), resulted in an overall unique mapping score of 78%. On average 16% of the reads could not be mapped to the reference genome, most of which were discarded due to their short alignment (Figure 43B).

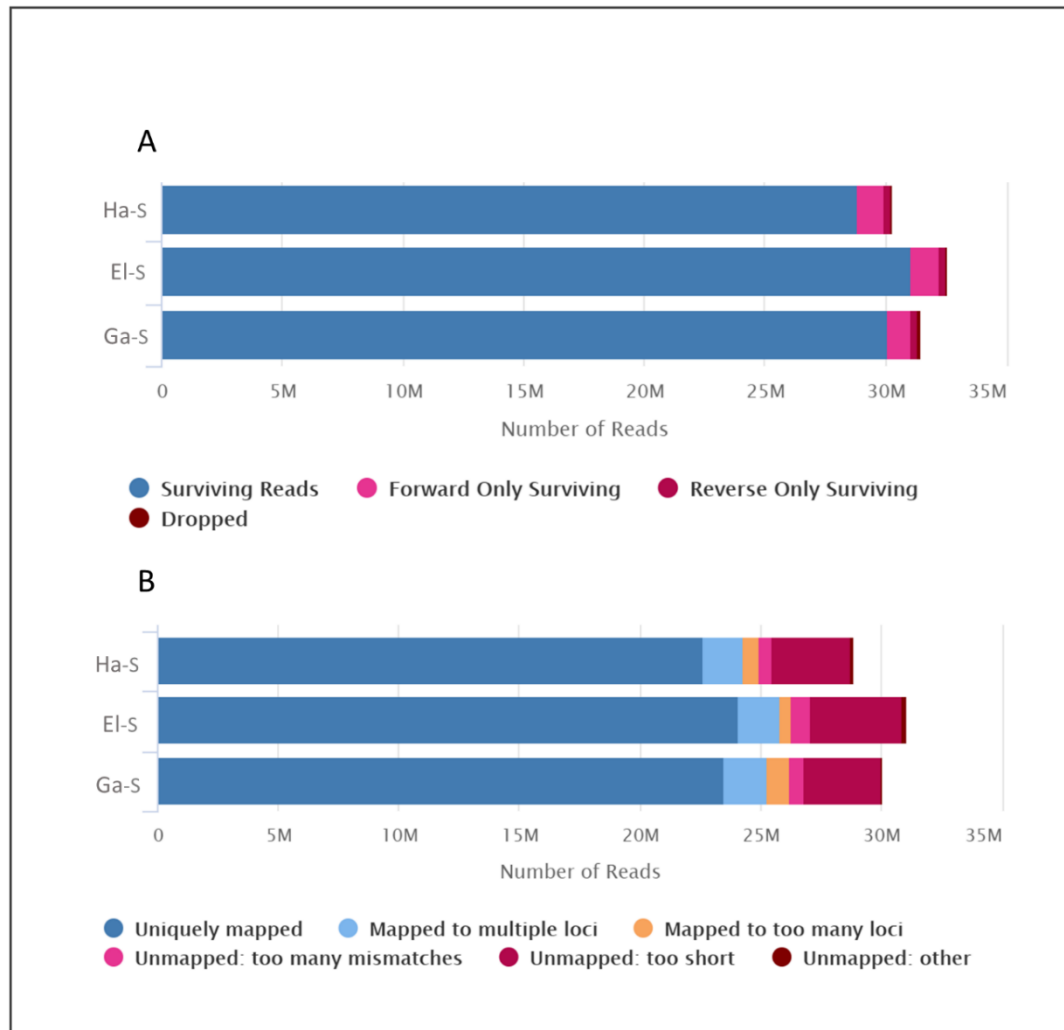


Figure 43: Statistics of the read processing pipeline. The different panels displays the results of the trimming (A) conducted by the Trimmomatic software and the metrics of the read mapping (B) generated by the RSEM package using the STAR module (v. 1.3.3).

4.3.3 Cell type-specific marker genes

To identify ‘cell-type specific’ genes, gene objects were identified that display at least a 10-fold higher expression in a given cell-type, when compared to the other two sorted cell populations and discrete tissues (described in chapter 2). The number of identified ‘cell-type specific genes’ showed considerable variation, with the lowest number found in granular amoebocytes, that present 279 ‘specific’ genes. Hyaline amoebocytes contained 530 ‘specific’ genes, with the highest number, 819, observed in the eleocytes (Figure 44A). The cell-type specific’ gene lists from each cell population was annotated using the Blast2GO pipeline (Figure 44B) and resulted in relatively similar annotation levels for the three gene lists. In the case of hyaline amoebocytes, 386 individual

transcripts were annotated (~73%), while 646 (~79%) and 200 (72%) were annotated for the eleocytes and granular amoebocytes respectively. Duplicated annotations within certain cell-types were subsequently removed, as were overlapping annotations between cell-types following identification using a Venn diagram (Figure 44C). This approach determined cell-specific marker genes. The highest number of cell-specific marker genes were observed in the case of eleocytes (377), with the lowest number were found in the granular amoebocytes (118).

4.3.4 Gene enrichment analysis of cell type-specific markers

To characterise the transcriptomic fingerprints of the different coelomocyte populations especially from the view of their immune function, the identified cell type-specific genes were analysed using both Gene Ontology (GO) and Pathway (KEGG (Kanehisa et al. 2015) and Reactome databases (Jassal et al. 2020)) enrichment.

In the case of eleocytes, the most significant biological processes were linked with different metabolic associated terms, such as “lipid metabolic processes”, “fatty acid derivative metabolic processes”, “biological oxidation” “acylglycerol metabolic processes” and to chemical homeostasis, with terms such as “response to xenobiotic stimulus” and “response to drug”. Although innate immunity related GO biological processes could not be identified in the case of eleocytes, the results of pathway enrichment analysis provided a few, interesting, pathways related to immune system function, such as “synthesis of leukotrienes and eoxins” and “Retinol metabolism”. When analysing hyaline (HA) and granular amoebocytes (GA) a small number of more general immune-related GO terms appeared. Both HA and GA-specific gene lists were enriched for “humoral immune response”, while GA additionally appeared to be significantly over-represented for genes involved in “defense response” and “cytokine activity”. More information about the most significant enriched GO terms is provided in Figure 45. The pathway enrichment analysis provided a slightly higher resolution for both GA and HA samples to describe the possible cell-specific immune responses (Figure 46). In case of the HA samples enriched pathways appeared to be organised around “leukocyte transendothelial migration” and “complement system”, while in GAs pathways such as “melanogenesis” and “antimicrobial peptides” were significantly enriched.

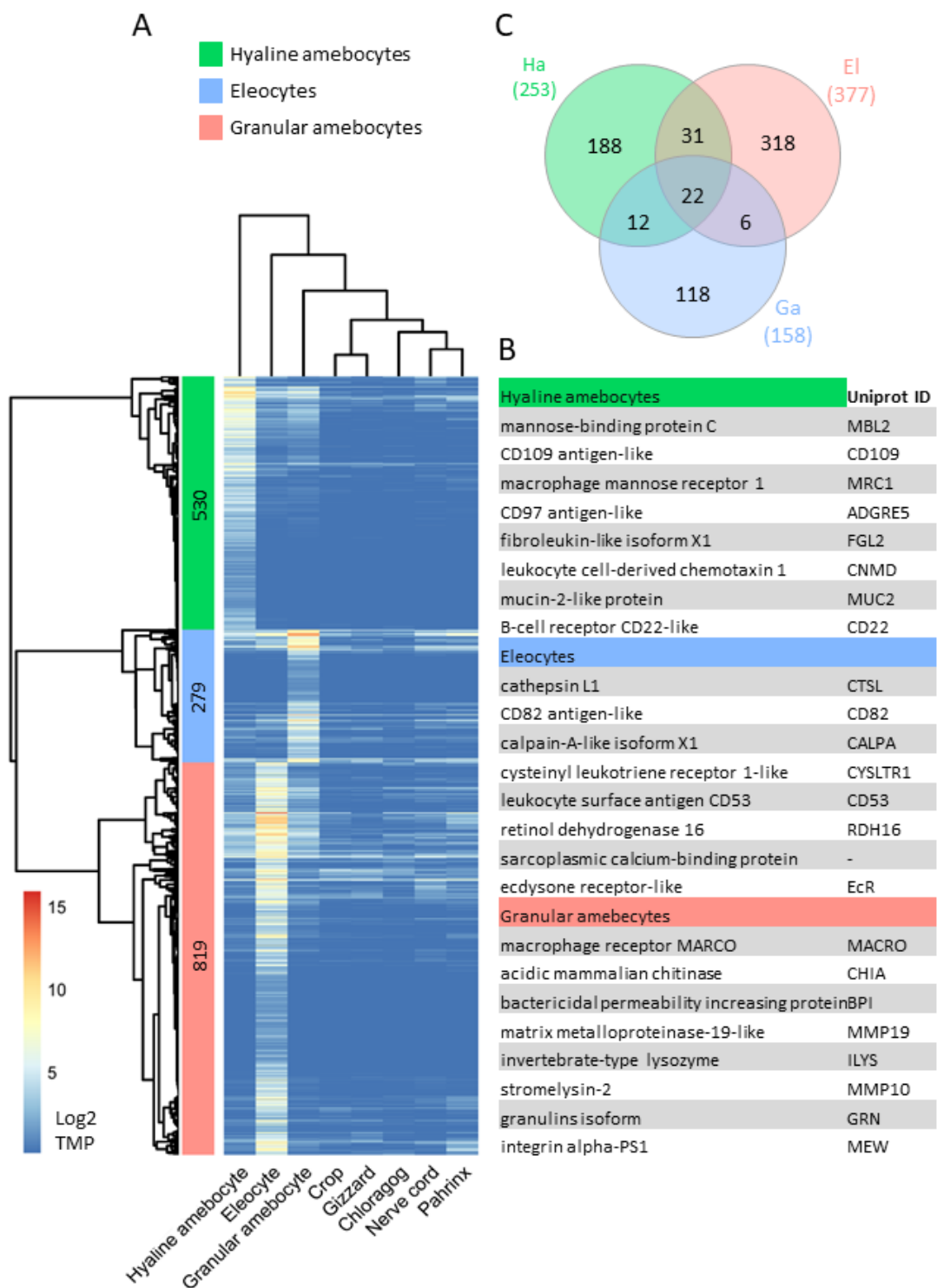


Figure 44: Identification and annotation of cell type-specific genes. Heatmap (Panel A) represents the expression profile of the identified in the different cell populations and tissues of *E. fetida*. Cell type-specific marker genes were annotated using the Blast2GO software from which a selection is shown on panel “B”. Venn diagram (C) shows the number of Unique annotations and their overlaps between the different cell types. To avoid confusing different transcript variants, overlapping sequences were removed from the cell specific gene “marker” list.

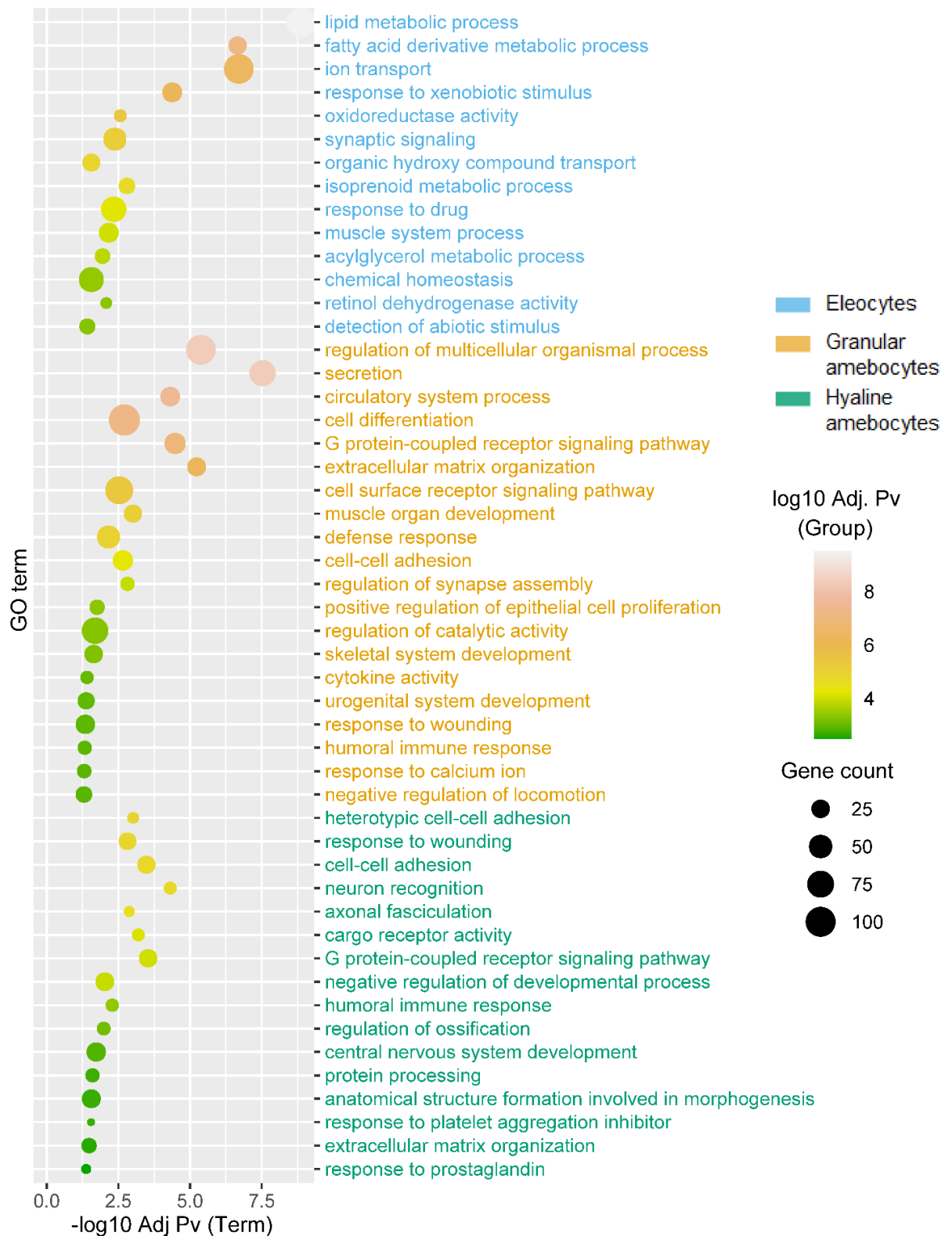


Figure 45: Significantly enriched Gene Ontology (GO) terms between separated coelomocyte populations. The colour of the dots shows the level of significance based on Benjamini-Hochberg adjusted p -values (FDR), while sizes represent the number of genes that are associated with the given terms.

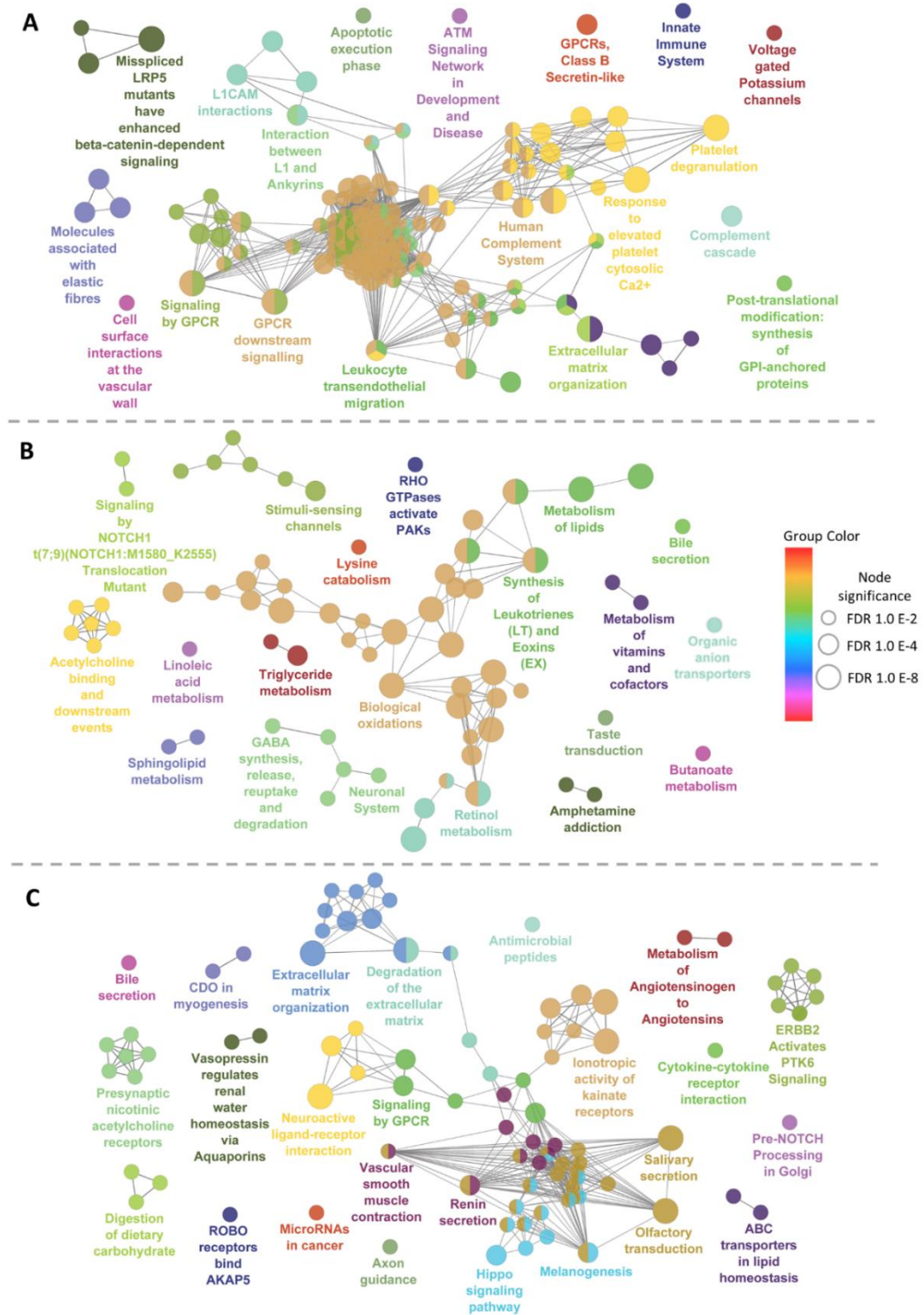


Figure 46: Pathway enrichment networks for separated coelomocyte populations. Networks are shown for Hyaline amoebocyte (A) Eleocyte (B) and Granular amoebocyte (C) specific marker genes. The size of the circles represents the level of significance based on Benjamini-Hochberg adjusted p -values. Nodes were coloured according to their group identity, based on related functions. When multiple terms are associated with the same group, only the pathway with the highest significance were labelled.

4.3.5 Expression profile of Toll-like receptors

Cell type-specific expression profiles were determined for the Toll-like receptors identified previously using the genomic sequences (chapter 3). Although the highest number of highly expressed TLR genes were identified in gut/chloragogen tissue, some TLRs presented their highest expression in the different cell types. From the total of 39 TLRs, six and seven TLRs showed their highest expression in granular and hyaline amoebocytes respectively, with the least TLR genes (3) presenting their highest expression in the eleocytes (Figure 47).

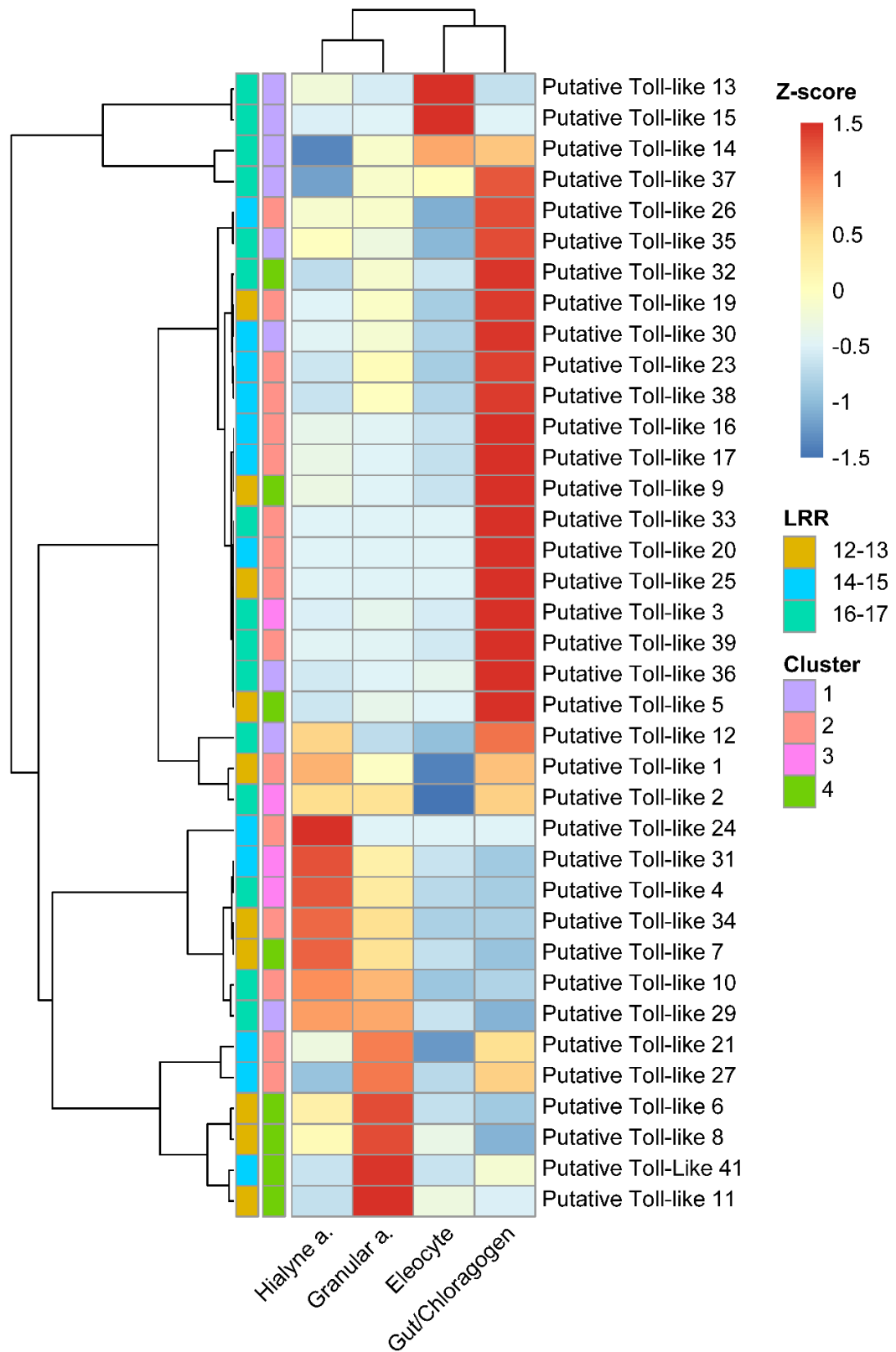


Figure 47: Hierarchical clustering of the Toll-like receptors based on their gut/chloragogen and cell-specific expression profiles. Row side colours represent the number of annotated Leucine-rich repeats (LRR) and their clade identity based on the analysis conducted in Chapter 3. Number of LRRs were identified using the LRRfinder online tool (Offord et al. 2010).

4.3.6 Expression profile of immune related genes

Following the characterisation of TLR expression profiles, the cell-specific expression pattern of some of the most well studied immune-related genes were also analysed and compared to the gut/chloragogen tissue. The immune gene expression profile based hierarchical clustering showed the highest correlation between the two amoebocyte populations. Although most of the analysed immune-related and general stress response associated genes showed highest expression in the amoebocyte populations, the immune associated gene lysenin appeared to be “specifically” expressed in eleocytes. Furthermore, the coelomic cytolytic factor was identified in the gut/chloragogen, as well as the granular, but not hyaline, amoebocytes. The detailed information about the expression of the analysed immune-related genes is shown in Figure 48.

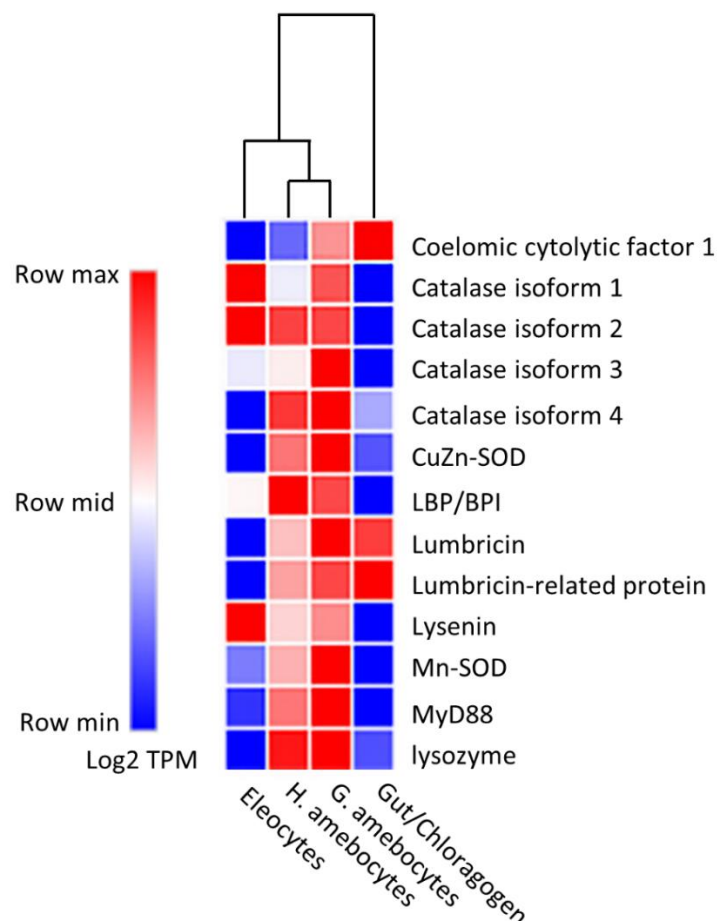


Figure 48: Cell type-specific expression profiles of selected earthworm immune-related and general stress-response genes.

4.4 Discussion

4.4.1 Coelomocyte preparation for cell sorting

By replacing the traditional centrifugation method with the new cell-strainer based coelomocytes washing protocol, we managed to prepare well-dispersed coelomocyte solutions from coelomic fluid. This has allowed us to wash the coelomocytes without creating a high rate of cellular mortality by decreasing the cell disruption normally caused by physical force, that would occur during the centrifugation and resuspension of the cells. This application also greatly contributed to the high efficiency and purity of the cell separation and helped to reduce the possible background noise of the RNAseq experiment by decreasing the number of cell duplets, triplets or bigger cell aggregates.

4.4.2 Identification of cell specific marker genes

The generation of sequencing data from the separate cell populations allowed the identification of a large number of 'cell-specific' marker genes. Following the utilisation of the tissue-specific data described in the second chapter we were able to characterise many of the coelomic fluid-specific genes identified earlier, now on a higher cell type-specific level. The high-throughput analysis of genes provides the first, transcriptomic based insight of the possible more unique biological functions of the three major coelomocyte populations present in the earthworm's coelomic cavity.

4.4.3 Functional enrichment analysis

4.4.3.1 Eleocyte specific genes

The functional enrichment analysis of the identified 'cell-specific' marker genes allowed the characterisation of the main functional characteristics of the different cell-types. In the case of eleocytes, both the pathway and enrichment analysis resulted highly significant terms associated with nutritive functions, which seems to support the earlier hypothesis which suggeststing a chlogogenous tissue-related origin of these cells (Jamieson 1981, Affar et al. 1998, Valembois et al. 1985). Despite the largely abundant metabolic processes, a few immune system relevant pathway were also identified in eleocytes. Although, due to their lack of phagocytic activity, eleocytes are routinely excluded from the majority of earthworm immunological studies, their importance in immunological defence mechanisms has recently gained more interest following

identification of their neutrophil NETs like functions (Homa 2018). The appearance of leukotriene synthesis during the pathway enrichment analysis also raises their possible role in regulating innate immunity, since leukotrienes are known to play an important role in the regulation of the innate immune system (Peters-Golden et al. 2005).

Although the Gene Ontology analysis of the HA and GA samples only resulted in more generic biological processes related to innate immunity, the conducted pathway enrichment analysis provided a the first comprehensive insight into the different immune-related functions represented in the distinct amoebocyte populations.

4.4.3.2 Hyaline amoebocytes specific genes

Opsonisation is a well-known immunologically important process that is based on the coating of the engulfed particles by different humoral factors, resulting in a modulation in phagocytic activity (Griffin 1977). Invertebrate components of the opsonisation complement, such as C3b, are known to play an important role in these processes. Previous studies not only identified ancient alternatives of the complement cascade in many invertebrate species, but the introduction of mammalian opsonins (C3b fragment) induced phagocytic increase was described in earthworm (*L. terrestris*) coelomocytes (Laulan et al. 1988, Nonaka 2011). The abundance of complement system associated terms in hyaline amoebocytes not only supports the existence of previously observed complement-like mechanisms (Laulan et al. 1988) but also clearly associates these processes with hyaline amoebocytes.

4.4.3.3 Granular amoebocytes specific genes

Melanisation is one of the cornerstone defence mechanisms of the innate immune system (Dudzic et al. 2015), during which a layer of melanin is produced and deposited on the surface of invading pathogenic microbes which results in the physical encapsulation of the pathogen (Cerenius et al. 2008, Götz 1986). In earthworms, melanisation is modulated by the prophenoloxidase (pro-PO) cascade through the formation of brown bodies (Procházková et al. 2006). Previous studies provided evidence for both the existence of a prophenoloxidase cascade at the protein level (Procházková et al. 2006), as well as histochemical and ultrastructural evidence of melanin production in *Eisenia* coelomocytes (Valembois et al. 1994). The results of the pathway enrichment analysis not only supported these observations on a transcriptomic

level, by identifying over-represented pathways such as “melanogenesis” when analysing ‘specifically expressed’ genes from GA, but also successfully revealed the specific importance of granular amoebocytes in such a basic innate immune function as encapsulation.

4.4.4 Cell-specific expression patterns of TLRs and often target genes of earthworm immunology

Using the genome as a reference, 39 novel Toll-like receptors with a full-coding sequence were identified (chapter 3.3.8). Based on the results of the generated ‘cell-type specific’ transcriptomic dataset, it was now possible to consider their ‘cell-specific’ expression profiles. Although the highest number of toll-like variability was presented in the gut/chloragogen, it is unsurprising when considering the high exposure levels of gut tissue to a wide range of microorganisms. Clustering the TLR genes by expression characteristics revealed insights into the functions of the different coelomocyte populations. The expression in eleocytes, where a low number of unique TLRs could be identified, suggests the existence of a highly specific target recognised by this select group of cells. Insights into the immunological functions of different cell populations was also discernible where a similar TLR expression patterns were observed. For example, granular and hyaline amoebocytes showed high overall expression similarities, although subtle differences in the expression of some genes were observed. Overall, expression pattern comparisons suggest eleocytes represent a more functionally distinct population. In general, the analysis of the immune-related genes showed agreement with the results described by Bodo et al. (Bodó et al. 2018), which not only validated the success of the described transcriptomic method here, but also improved the characterisation of the two broad cell types (eleocytes and amoebocytes), and resolved differences between the two amoebocyte populations at the whole transcriptomic scale (Figure 49)

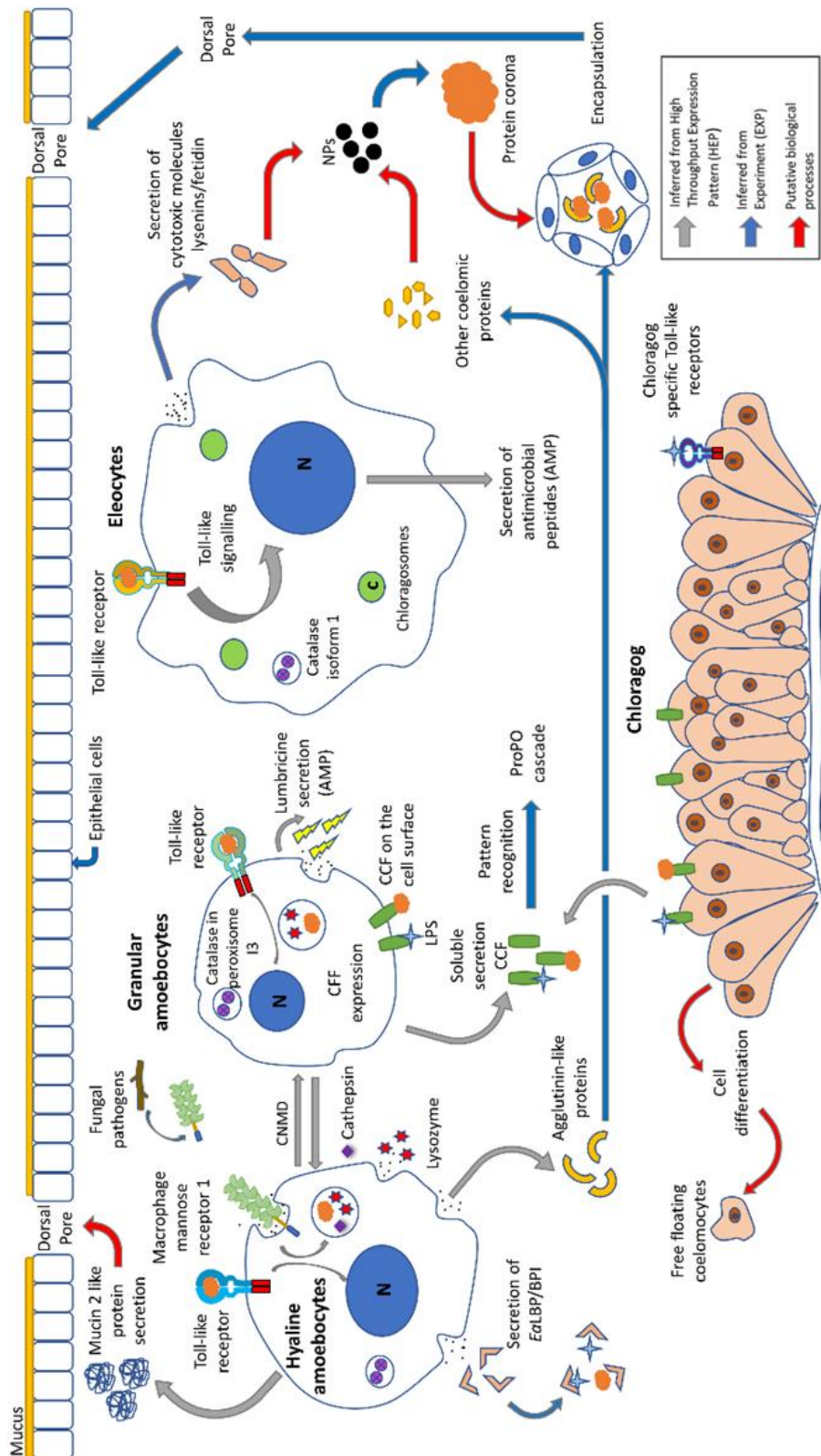


Figure 49: An idealized model of immune signalling. Intended to represent possible features of an Earthworm's innate immune signalling pathway acknowledging their cell-type-specific representation. The different arrow colours represent the the source of the information, where grey arrows accounts for information inferred from the cell-type specific high-throughput analysis, described in Chapter 4.

4.4.5 Conclusion

The creation of a 'cell-specific' expression atlas has revealed the transcriptomic fingerprints of the individual cell types found in the coelomic cavity. This facilitates a better understanding of the function of the different coelomocyte cell populations and provides the first transcriptomic insight into the unique immunological functions of the cell types, such as melanisation and opsonisation. This also helps to trace the immune-related pathways which can be affected by NPs and provides a great base for future experiments that attempt to understand earthworm innate immunity at the cell-type level.

5 Temporal response to challenge

5.1 Introduction

To effectively defend against infectious diseases but at the same time avoid any significant tissue damage due to the excessive pro-inflammatory processes, the innate immune response requires constant and precise self-regulation. To achieve this level of control, most inflammatory processes employ different positive and negative self-regulatory feedback loops on numerous levels (cell-specific, signal-specific, gene-specific) (Liu and Cao 2016, Wang and Liu 2007). Since in many cases, the components of these regulatory processes operate on different time scales the effects on the innate immune response can be separated into different temporal phases. However, the transition between these phases are less definite, and there exist substantial overlaps between them, in the case of vertebrates, we can divide most of the innate inflammatory processes around three temporal stages (Foley and O'Farrell 2004). The early phase requires a quick response to the pathogen-associated molecular pattern (PAMP) challenge, which is mainly realised by utilizing both soluble and membrane-bound pattern recognition receptors (PRR) (Akira et al. 2001, Franchi et al. 2009, Kawai and Akira 2008). The activation of different PRRs is normally followed by a rapid increase in the expression of cytosolic transcription factors (TF) such as the the nuclear factor kappa-light-chain-enhancer of activated B cells (NFκB) which has an important role regulateing the second phase by stimulating the production of different cytokines and cell-signalling molecules (Dev et al. 2011, Wang and Liu 2007). In the second phase, both positive and negative transcription and post-transcriptional regulation factors play an essential, antagonistic role to keep the cellular and humoral responses on an appropriate level (Piccirillo et al. 2014, Carpenter et al. 2014). The late phase has a key role in alternating the fates of newly generated macrophages by producing signalling molecules that are involved in cell differentiation processes (Nappi 1973) as well as helping the recovery processes and the creation of the new homeostatic state (Figure 50).

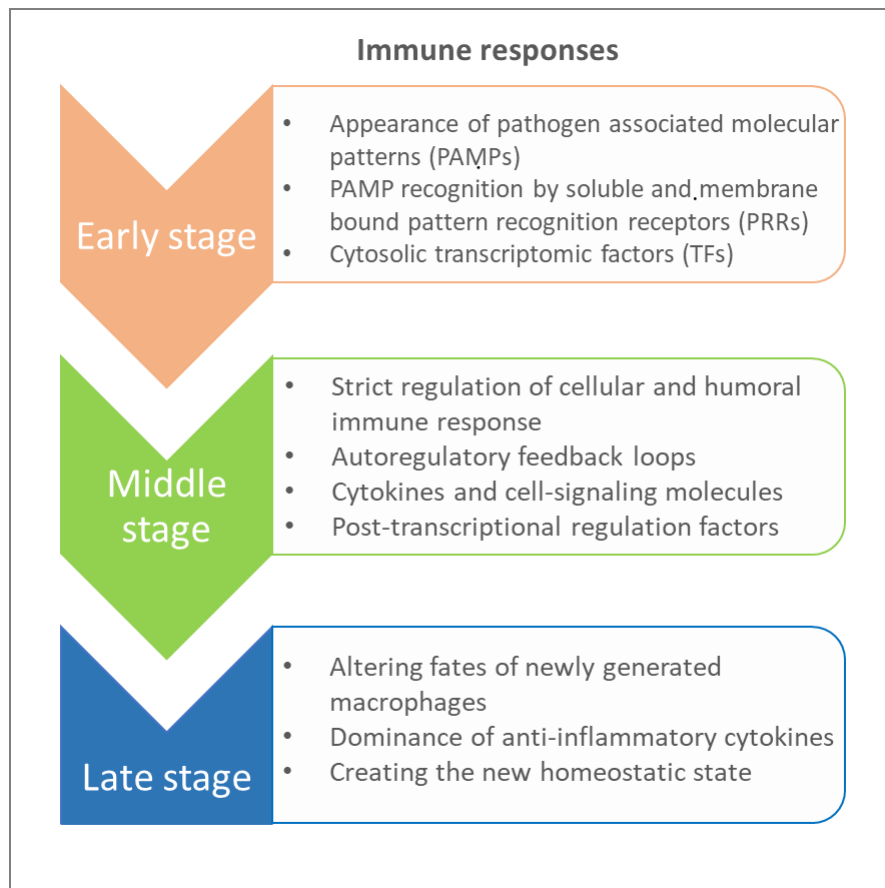


Figure 50: Flowchart illustrating some of the characteristic biological processes of the different tetemporal stages associated with an innate immune response (Dev et al. 2011, Piccirillo et al. 2014, Wang and Liu 2007).

Although understanding the spatial (cell-specific) and temporal aspects of the innate immune response is essential for building a comprehensive, vertebrate like, understanding of the immune mechanisms of the earthworms, there has been limited research into this topic to date. In the case of earthworms, there are several relatively well studied immune-related genes with temporal information, such as some components of the Toll-like signalling pathway (Škanta et al. 2013), with most of these studies focusing on individual gene expression changes using mixed coelomocytes cell populations (Francis et al. 2007). Substantially less data is available about the spatial expression of these genes and almost nothing is known about how the different cellular and humoral components interact in time when challenged with infectious disease.

This chapter aims to give insight into the spatiotemporal characteristics of the earthworm immune response induced by bacterial challenge both in the presence and absence of copper nanoparticles (CuNPs). This was achieved by utilising the

coelomocyte separation method described in Chapter 4 (4.3.1), while temporal information was provided by sampling at nine different timepoints (0h, 3h, 6h, 12h, 18h, 24h, 48h, 72h, 96h) during the immune challenge. Based on the analogy with the earlier described vertebrate terminology, our objective was to resolve at least three different phases of the earthworm immune response. An early phase comprising immediate and direct response, such as pathogen recognition, a middle phase with systemic responses directed by cell-cell communications, and a late phase which modulates the new coelomocyte differentiation processes, assists recovery and re-establishes homeostasis.

To specifically circumvent environmental transformation of the NPs and therefore observe the direct impact of the CuNPs on the earthworm immune system, nanoparticles were directly injected into the coelomic cavity 24h prior to the initiation of the bacterial dose. The injection method also allowed us to use a non-lethal dose and observe the NP effect on the immune system rather than completely masking the immune modulation effects caused by a high external dose that induces metal cytotoxicity. The immune response was induced by dermal challenge of the earthworms with a Gram-positive bacteria (*Bacillus subtilis*) (Dvořák et al. 2016, Josková et al. 2009). Since the high rate of mortality could compromise the results, it was important to maintain a high survival rate throughout the full exposure period (96h). To achieve the non-lethal bacterial effect, different bacterial loads were tested prior to the experiment and the final concentration was set just below the level where mortality started to occur.

To achieve a comprehensive spatiotemporal picture the experiment was needed to provide a relatively high number of sampling points balanced whilst recognising the high resource demand of transcriptomic analysis (cost of library preparation/sequencing). The final design incorporated a total of 81 RNA-seq libraries representing the three different coelomocyte populations harvested at nine different timepoints, under three experimental conditions (control, bacterial challenge, NPs exposure followed with bacterial challenge). To be able to cover the temporal aspect without substantially increasing the cost, biological replication was addressed through organism pooling. This allowed us to show temporal frequency rather than replication of specific points by applying spline regression model based statistics (Michna et al. 2016) to detect differential gene expression across the conditions.

5.2 Materials and Methods

5.2.1 Biological material

All *Eisenia fetida* were adults with developed clitella, supplied from a genotyped laboratory culture maintained at the UK Centre of Ecology and Hydrology (Wallingford, UK). Before the experiment the earthworms were acclimatised under standard laboratory conditions for four months in a mixture of compost and topsoil (3:1 volume ratio) at a constant temperature of 22°C, using, using 12 h light/dark cycles. The cultures were regularly fed using horse manure collected from horses without any recent medical treatment. The horse manure was kept at -80°C prior to application, to avoid any earthworm cross-contamination and reduce the possibility of introducing eukaryotic parasites into the earthworm cultures.

5.2.2 Experimental design

To reduce the soil content in the gut, earthworms were depurated by placing them on moist filter paper (1 x PBS buffer) 12 hours before the experiment was started. Following the depuration, earthworms were randomly sorted between the three different treatment groups. First, as a pre-treatment, earthworms were individually injected with either 5 µl copper-oxide nanoparticles resuspended in PBS solution (CuNPs) or 5 µl pure PBS as a negative control. To avoid coelomic fluid extrusion due to manual stimulus earthworms were anesthetized, by individually submerging them in freshly prepared carbonated water until the lack of body movement was discernible (around 45 s). After being anesthetized earthworms were individually transferred into a sterile multichannel pipet reagent reservoir containing a 5 mls of carbonated water. The injections were conducted using a 10 µl Hamilton syringe attached with a sharp, curved needle. Earthworms were injected 10 - 15 segments posterior to the clitellum towards the caudal part of the body by puncturing the dermo-muscular tube while avoiding penetration of the gut (Figure 51). This precise manoeuvre was conducted using a stereo binocular microscope. Following the injections, earthworms were quickly transferred into individual Petri dishes containing wet filter paper before incubation for 24 hours at a constant temperature of 22°C. After the 24 h incubation earthworms were transferred individually into a new pre-prepared Petri dish containing paper granules wetted using either *Bacillus subtilis* liquid culture or PBS solution according to the corresponding

treatment group. Thereafter, the Petri dishes containing the earthworms were sealed using parafilm tape and placed into a 22°C incubator for the appropriate experimental period (3h, 6h, 12h, 18h, 24h, 48h, 72h, 96h). A summary of the overall experimental design shown on Figure 52.

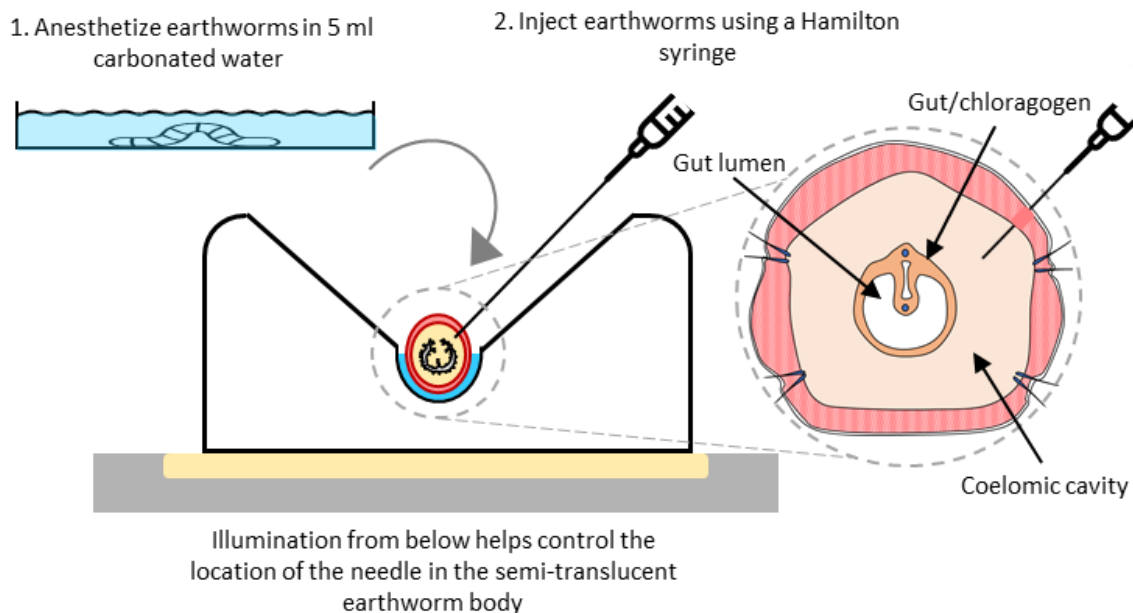


Figure 51: Schematic representation of the earthworm injection protocol. Earthworms first were first anesthetized using fresh carbonated water then injected into with 10 μ l PBS or copper-oxide nanoparticles based on the experimental condition. To avoid injection into the gut lumen, this precise manoeuvre was conducted under a binocular microscope with the base illumination turned on.

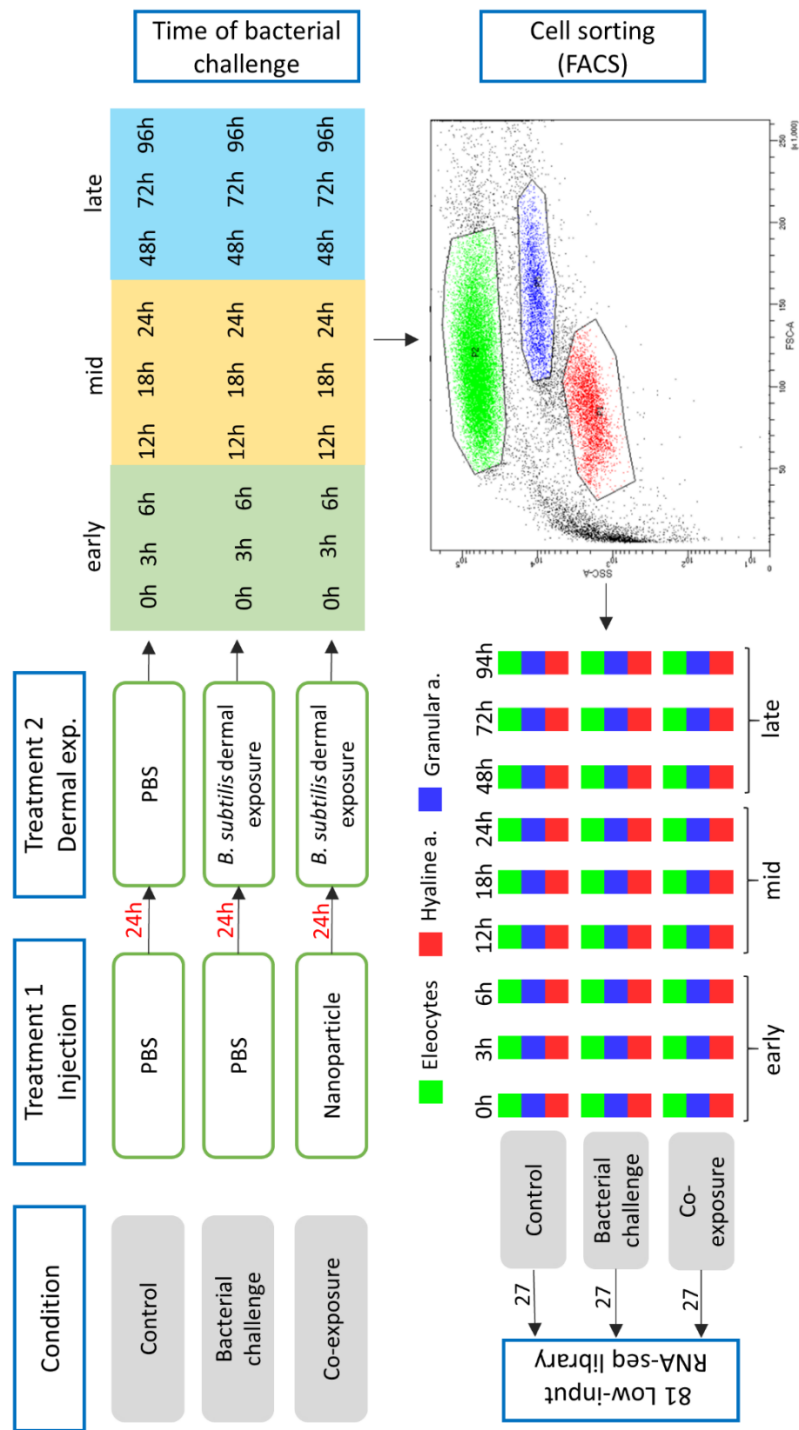


Figure 52: Flowchart illustrating the main aspects of the implemented experimental design.

5.2.3 Nanoparticle exposure

The copper oxide nanoparticles (CuNPs) used during the exposure were supplied by Applied Nanoparticles (Barcelona, Spain). The nanoparticles were spherically shaped with an average diameter of 40nm and were dispersed in 10 mM TMAOH dispersing

media (25% w/v tetramethylammonium hydroxide) to magnify their chemical and colloidal stability. Crystal size and monodispersity of the particles were characterised using transmission electron microscopy (TEM) while CuO content was measured using ICP-MS. To avoid any pro-inflammatory processes caused by possible endotoxin content, CuONPs were prepared and preserved under endotoxin-free conditions. Endotoxin concentration was determined using the limulus amoebocyte lysate (LAL) test, to determine that the sample had an endotoxin content of <0.25 Eu/mL. To avoid any toxic effect caused by the dispersion solution, nanoparticles were resuspended in 1x PBS just prior the injections. In total 0.05 mg CuNPs were injected into each earthworm in a total 5 µl volume.

5.2.4 Bacterial challenge

To challenge their immune system, *E. fetida* earthworms were exposed to a gram-positive spore-forming bacteria species, specifically the W23 strain of *Bacillus subtilis* (Czech Collection of Microorganisms, Brno, Czech Republic). The microorganisms were cultured in LB broth at 37°C with constant shaking at 300 RPM. The liquid culture of bacteria was removed from the incubator and used to stimulate the earthworms in the logarithmic growth phase (OD₆₀₀ reached 0.8). Midlog bacteria were quickly pelleted by centrifugation, and resuspended in PBS. In total, approximately 1×10^{10} Colony Forming Units (CFUs) of *B. subtilis*, in a volume of 3 mL, were transferred into individual Petri dishes containing dried paper towel pellets.

5.2.5 Coelomocyte preparation and sorting

Following the various incubation times in the presence of *B. subtilis*, or PBS in the case of negative controls, earthworms were individually transferred to a sterile Petri dish containing 1 mL of ice-cold cell extrusion buffer (1 X PBS supplemented with 5 mM EDTA). To extrude coelomocytes earthworms were exposed to gentle electrical stimulation using a fresh 9V battery. Extruded coelomocytes were then suspended in the extrusion buffer which was collected with a wide bore pipette tip. The coelomic fluid suspension then was passed through a 100 µm mesh sized cell strainer to remove any mucus content, soil or large cell aggregates. Finally, coelomic fluid was collected and transferred to a non-stick 1.5 ml centrifuge tube and placed on ice until cell-sorting. Coelomocytes were sorted into three different populations (eleocytes, hyaline

amoebocytes, granular amoebocytes) using the same method described in Chapter 4. In this case, 150,000 cells were collected for both amoebocyte populations, while this number was doubled for the eleocytes. Cells were sorted directly into cell lysis buffer (RTL Plus, Qiagen Ltd), snap-frozen in liquid nitrogen then stored at -80°C until the total RNA extraction.

5.2.6 RNA extraction and library preparation

The RNA extraction was carried out in the same way as described in Chapter 4. The quality and quantity of the extracted RNA samples were measured by evaluating 2 µl of the sample using a Agilent High Sensitivity RNA ScreenTape system (Agilent TapeStation 4200, Agilent). The RNA-seq library preparation was carried out using the NEBNext Single Cell/Low Input RNA Library Kit (New England BioLabs Inc., MA, USA – NEB) with 2 ng of total RNA as an input, following the manufacturer's instructions. The quality of the prepared libraries was individually validated by evaluation on a D1000 ScreenTape assay (Agilent TapeStation 4200) while quantities were measured using a high sensitivity fluorometer (Qubit 4 see 4.2.2).

5.2.7 Data generation and QC

The final libraries were sequenced on the Illumina NextSeq platform using high output flow cells. Operation of the sequencing platform was performed by Ms Angela Marchbank, Cardiff School of Biosciences Genomics Hub, Cardiff University. In total, approximately 775 M single-end 75 bp reads were generated with an average insert size of 240 bp. On average, around 9.6 M reads were sequenced per sample using three high output flow cells. To remove any adapter sequences and low quality bases, high-throughput sequencing data was initially filtered using the Flexbar pipeline using the settings provided by NEB (Roehr et al. 2017). The efficiency of the trimming was verified by running FastQC (Andrews 2010) and analysing the metrics provided by the Flexbar pipeline. Since RNA-Seq libraries were generated using a relatively low amount of input material (2 ng), a low-input library preparation method was applied which exploited several cycles of amplification using PCR. For this reason, additional care was taken to examine the duplication content of the individual samples. For this reason, additional care was taken to examine the duplication content of the individual samples. Specifically, duplicated reads were marked using the MarkDuplicates tool from the

Picard package (Broad Institute 2018). Following this step, real duplicates originating from highly expressed transcripts were distinguished from artefacts produced by PCR over amplification using the dupRadar R package (Sayols et al. 2016). After analysing the plots generated by dupRadar, there was no indication of a high technical duplication rate (data not shown).

5.2.8 Mapping and differential gene expression analysis

Trimmed and quality-filtered reads were mapped to the annotated genome as described in Chapter 3 using RSEM with the STAR mapping module (Dobin et al. 2013). Thereafter, normalised counts were used to identify differential expression between the three experimental conditions using the gene expression time-course data. To conduct the differential expression analysis, normalised counts were first log₂ transformed then a natural cubic spline regression model-based method was used (spline TimerR) (Michna et al. 2016). This method first fits a natural cubic spline regression model to each gene within each condition, then the difference in the values of the coefficients of the fitted splines between the different experimental conditions used to detect the gene expression differences in time. The significance value cut-off was set to 0.05 corrected using the Benjamini-Hochberg method (FDR) while the degrees of freedom were defined as three.

5.2.9 Temporal Gene clustering

Normalised counts were log₂ transformed then an optimal cluster number was determined by running the data through the Clust automatic clustering pipeline using the default setting (Abu-Jamous and Kelly 2018). The identified optimal number of clusters were to an R package, Mfuzz, which uses soft clustering (fuzzy C-means) to analyse the temporal nature of the DEGs (Kumar and Futschik 2007). To have a mean value of zero and a standard deviation of one, normalised counts first were standardized (z-score). To avoid random data clustering, the fuzzifier value was estimated using the built-in “mestimate” command of the package. For extracting cluster cores, gene membership value cut-off was set to 0.5 (Michna et al. 2016).

5.2.10 Functional enrichment analysis

Functional and pathway enrichment analysis was performed as described in 4 (4.2.6).

5.2.11 Analysis overview

To provide a clear, graphical overview, the steps of the conducted bioinformatic analysis alongside with the used softwares were illustrated on Figure 53.

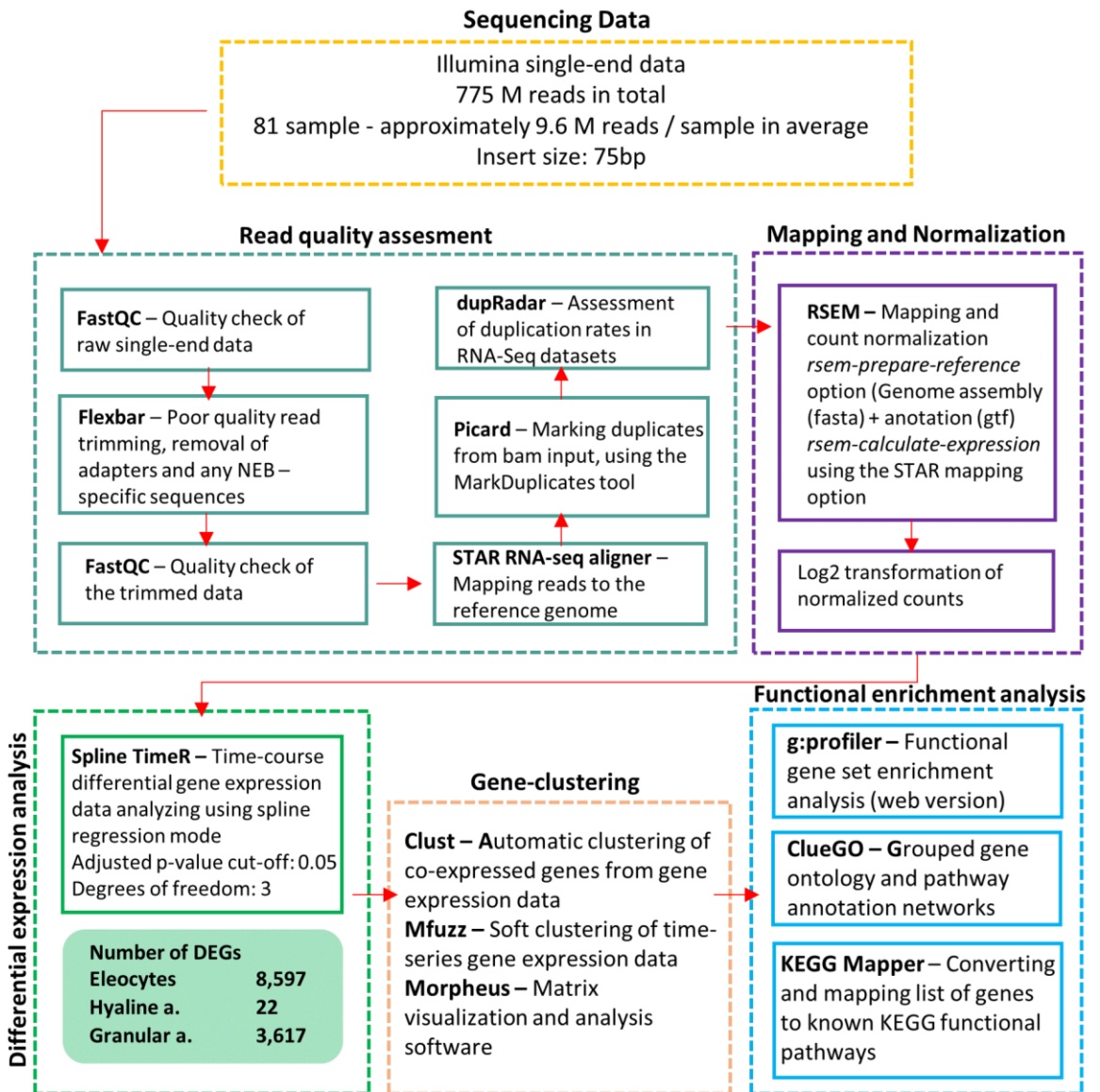


Figure 53: Workflow of the conducted bioinformatic analysis.

5.3 Results

5.3.1 Quality control and signal processing

After merging and quality control (QC) the data from the three Illumina high-capacity flow cells yielded approximately 767 M single-end reads passing the quality filter corresponding to a 98.9% survival rate. The quality score across the length of the sequence reveals a Phred score of >28 for 95% of derived sequences (Figure 54). The number of total read counts used in the RNAseq mapping step is shown in Figure 55. Overall, the sequencing yielded an average of 9.4 M reads which passed QC filtering per sample, these reads were subsequently used for short-read mapping. Analysis of the uniquely mapped reads indicated 70-80% of uniquely mapped reads (Figure 56). The overall average duplication rate across the samples was 39%, however, a detailed examination of duplication characteristics using Picard's 'MarkDuplicates' pipeline followed by dupRadar testing suggested that this level of duplication resulted from highly expressed transcripts rather than technical artifacts, such as PCR duplication. After scrutiny of the plots generated using dupRadar, no unusual sequencing bias was observed in the data, either due to the low input library preparation method (increased number of PCR cycles) or caused by the sequencing (Figure 57).

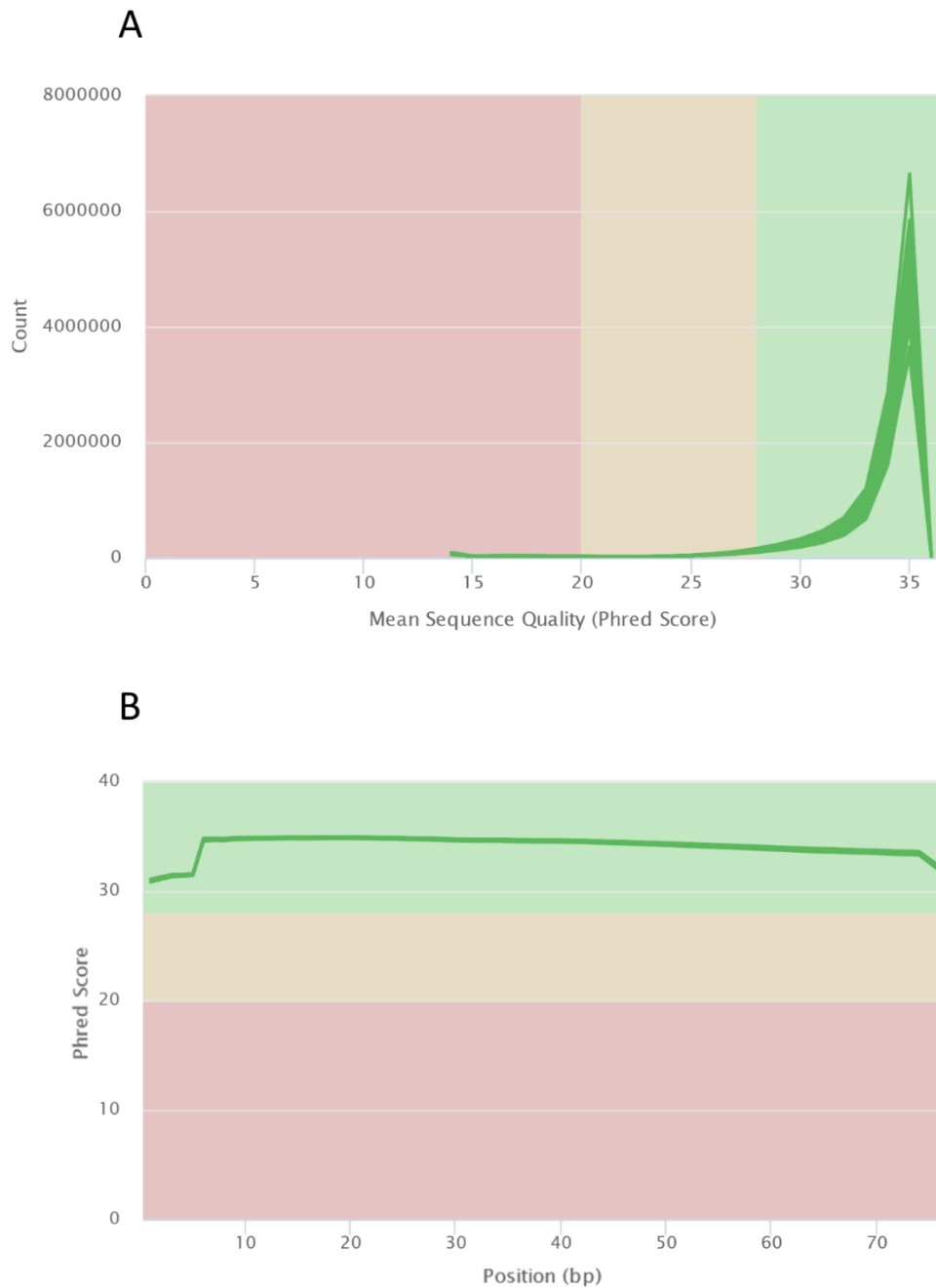


Figure 54: Quality metric for raw sequence reads. The average Phred score for each read (panel A) was analysed together with the Phred score for each position along the length of the sequence (panel B). Analysis was performed within Flexbar (Roehr et al. 2017).

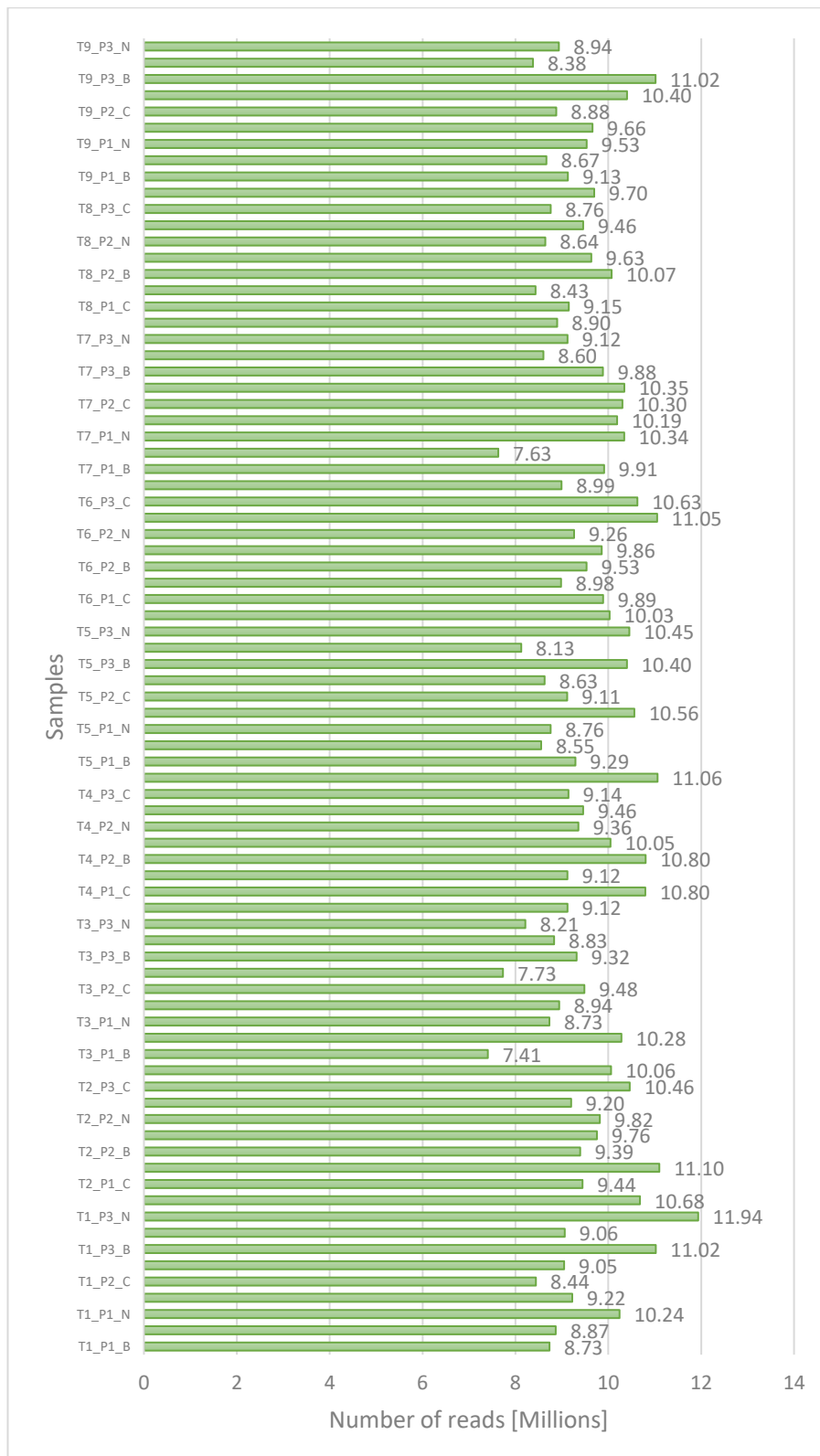


Figure 55: Number of reads successfully mapped to the annotated genome. Mapping was performed using STAR (Dobin et al. 2013).

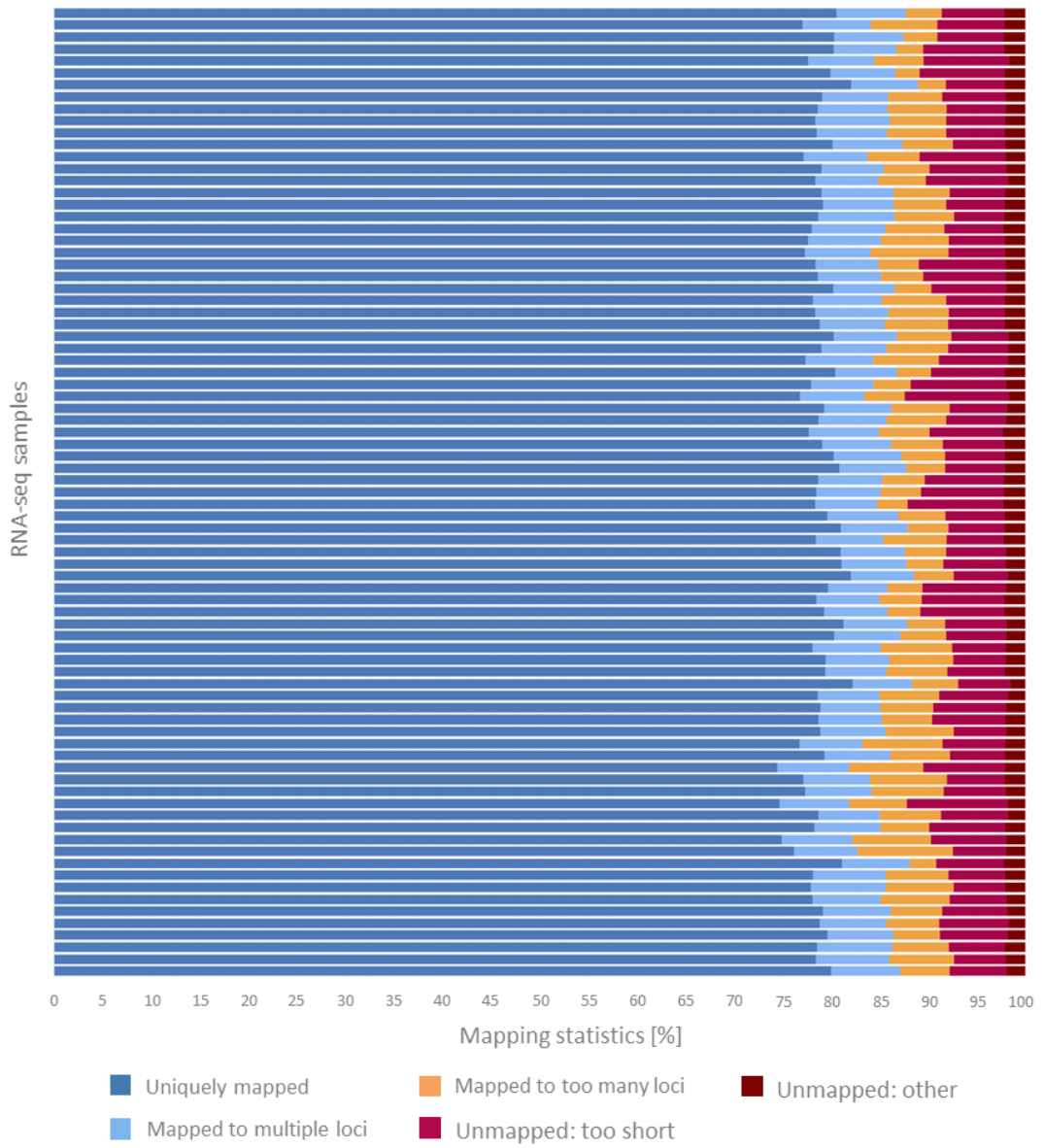


Figure 56: Proportion of reads that mapped to the annotated genome yielding unique and multiple-mapping. Analysis was performed using Picard (Broad Institute 2018) and multiQC (Ewels et al. 2016).

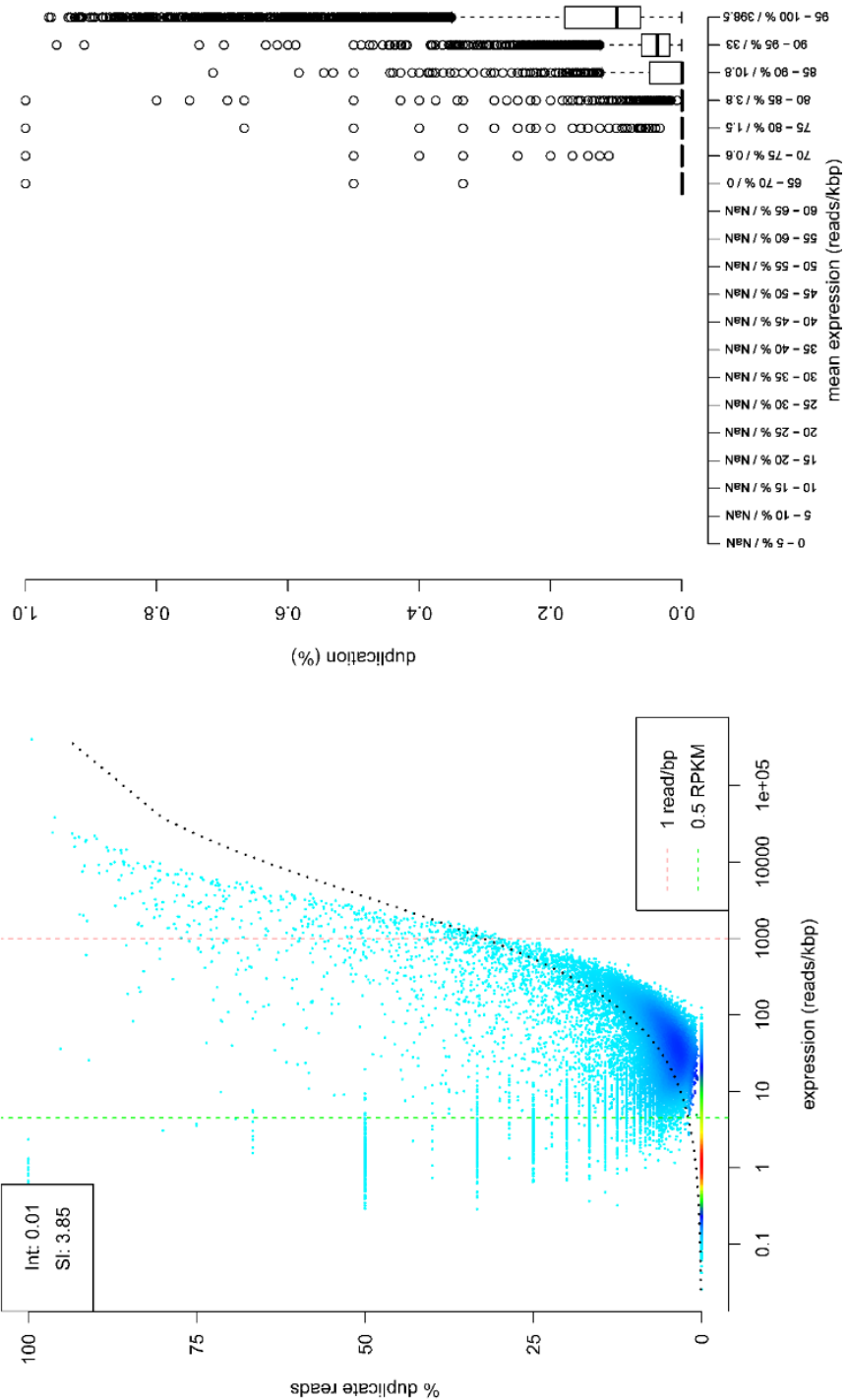


Figure 57: Sequence duplication analysis where the relationship between the duplication rate and gene expression level (read per kilobase - RPK) is shown. Normally, RPK values are expected to be proportional to the abundance of the transcripts assuming that highly expressed genes are responsible for the majority of the read duplication. By fitting a logistic regression curve onto the data, duplication resulting from technical error (e.g. library preparation) can be separated from natural duplication that results from highly expressed genes. Both the 2d density scatterplot (Panel A) and boxplot (Panel B) showed a low duplication rate at low RPK values which rises rapidly at higher RPK values a finding consistent with the observed read duplication having a natural origin.

5.3.2 Differential gene expression analysis using cubic spline regression model

The time-course data were used to identify genes presenting differential expression between the experimental conditions across the three major coelomocyte cell populations (hyaline amoebocytes, granular amoebocytes, eleocytes). Differential expression analysis was conducted using the splineTimeR package (Michna et al. 2016) which is based on the implementation of natural cubic spline regression models. This model uses a two-way experimental design, comparing one control with a specific treatment group using time as a continuous variable. The analyses of the control versus *B. subtilis* challenged individuals did not result in any significant gene expression changes. The lack of significant gene expression changes in the animals challenged solely with *B. subtilis* presumably occurred as a relatively low level of bacterial exposure was used to maintain the high survival rate throughout the experimental timecourse. However, in the case of CuNP primed animals which were subsequently immune challenged, differentially expressed genes (DEGs) could be identified in all the three cell populations. The number of DEGs identified in the three coelomocyte populations showed relatively high diversity. The lowest number of DEGs were identified in hyaline amoebocytes, where only twenty-two affected genes could be identified. This was followed by granular amoebocytes, that presented more than three thousand gene objects displaying significant expression changes. Finally, the highest number of DEGs was observed in the Eleocyte population, revealing approximately eight thousand differentially expressed genes.

5.3.3 Gene Ontology enrichment on the spatial data

To resolve the overall spatial (cell-type specific) aspect of the data, a functional enrichment analysis was performed using all the identified DEGs from the different coelomocyte types across all nine sampling points without considering changes in their temporal expression pattern. The identification of the over-represented Biological Processes was conducted using the Gene Ontology (GO) database.

5.3.3.1 Hyaline amoebocytes

Hyaline amoebocytes yielded the least number of enriched Biological Processes (Figure 58A). However, the small number of differentially regulated genes, affected by the combination of NP and *B.subtilis* treatment, are significantly associated with the

relatively specific terms of “Cellular response to thyroxine stimulus” and “Glutathione metabolic processes”.

5.3.3.2 Granular amebocytes

In contrast to the hyaline amebocytes, GO analysis of DEGs present in granular amebocytes reveals several immune and cell signalling related response terms (Figure 58B). However, highly affected biological processes were organised mostly around “RNA metabolic process” and “RNA splicing” supplemented with several very specific immune response-related terms, such as “NIK/NF-kappaB signalling”, “antigen processing and presentation of exogenous peptide antigen by MHC class I” and “interspecies interaction between organisms”.

5.3.3.3 Eelocytes

The eelocyte population showed most significant enrichment for terms associated with “Immune response” followed by several terms linked to “Metabolic processes” (Figure 58C). Although, immune-related terms were more determinant than in the other two cells population, interestingly these were less specific. Biological processes associated with cell death and apoptotic process were also detected within this cell lineage.

5.3.4 Pathway enrichment on the spatial data

Cell type specific roles were further explored by examining DEG enrichment associated with over-represented biological pathways. The analysis was conducted by enrichment analysis of pathway defined within the Reactome database, this yielded higher functional resolution separating general cytotoxic response caused by the CuNPs from enhancement in immune response. Additionally, changes in cell signalling between the different cell types were identified, that may be caused by the combined effect of the NP and bacterial treatment. Pathway analysis not only allowed the partial separation of three major cellular responses, metal-induced cytotoxicity, immune response enhancement and changes in intercellular signalling, caused by the different aspects of the combined treatments but also provided a much more detailed insight into the affected biological processes by dividing them between more specific biological pathways.

Pathway analysis was conducted in the case of all three cell populations, however, hyaline amoebocytes only resulted in one significant node, namely 'Glutathione conjugation'. For this reason, networks were only generated using the functional enrichment results of the other two cell population (Figure 59A-B). 'Glutathione conjugation' was also present in granular amoebocytes. Furthermore, there was evidence of copper cytotoxicity associated with DNA damage in eleocytes, as several DNA repair related molecular mechanisms are present. It was interesting to observe that the combined treatment had a relatively high impact on many of the innate immunity related pathways. Pathogen recognition associated with pathways such as the 'Toll-like receptor cascade' seemed to be affected in both Granular amoebocytes and eleocytes. While in eleocytes, most immune response related terms were associated with cell signalling, some effector functions like 'Neutrophil degranulation' also appeared in granular amoebocytes. Different cell signalling pathways also showed high variety between the different cell types. In granular amoebocytes, cell-signalling was mainly mediated by tumor necrosis factor receptor (TNFR2) through NF- κ B, while in the case of eleocytes different interleukins (IL) seemed to play a key role.

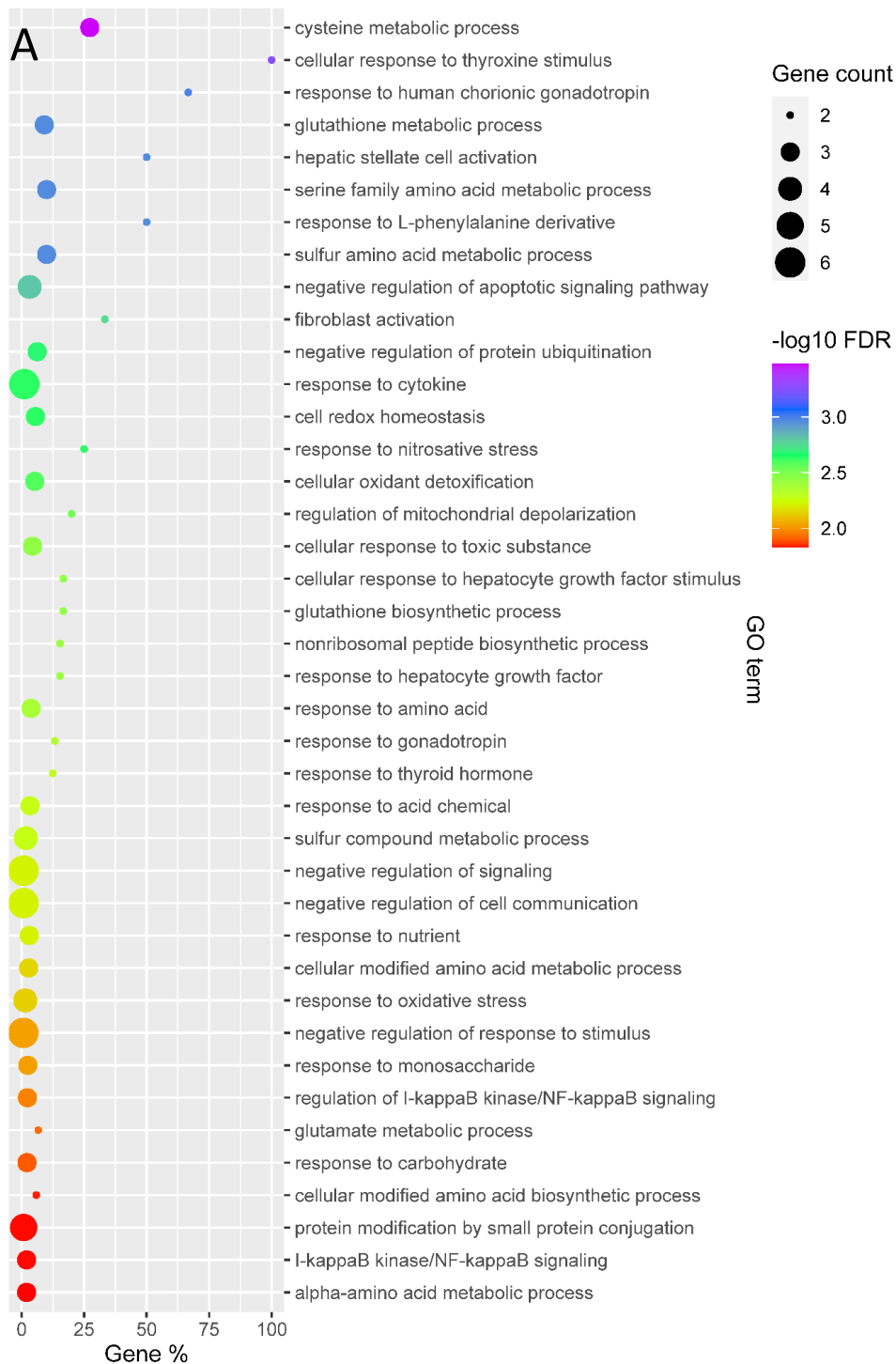


Figure 58: Biological Process GO enrichment analysis for differentially expressed genes (DEGs) observed in specific coelomic cell lineages under combined challenge of bacterial pathogen and CuNPs. Panel A described the enriched groups observed for Hyaline amoebocytes (P3). Enrichment analysis was performed using gProfiler (Reimand et al. 2019) followed with GO term redundancy filtering based on term-term similarity statistics conducted with the REVIGO software (Supek et al. 2011). Final enrichment results were plotted using the ggplot2 R package (Wickham 2016).

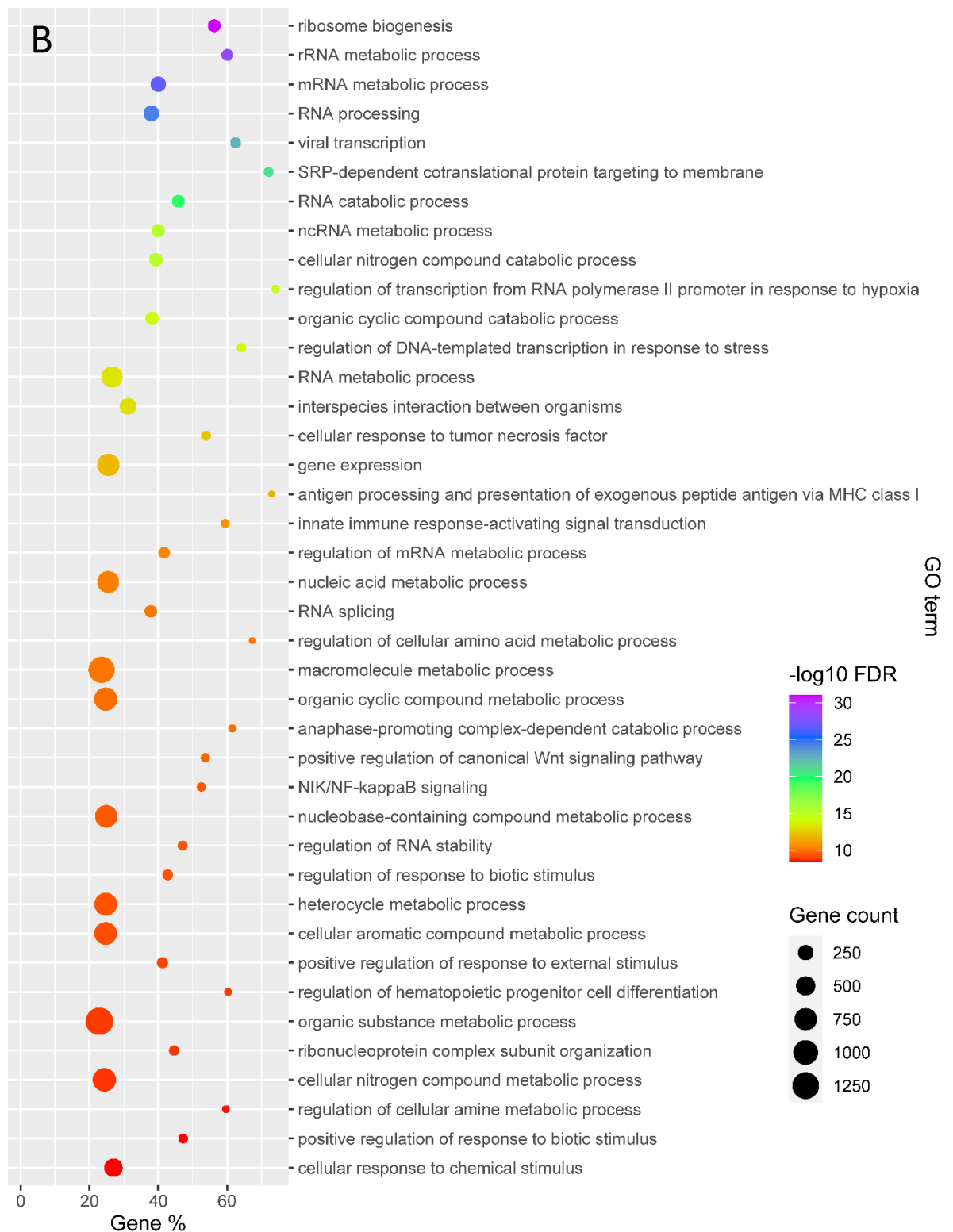


Figure 58: Biological Process GO enrichment analysis for differentially expressed genes (DEGs) observed in specific coelomic cell lineages under combined challenge of bacterial pathogen and CuNPs. Panel B described the enriched groups observed for for granular amoebocytes (P1). Enrichment analysis was performed using gProfiler (Reimand et al. 2019) followed with GO term redundancy filtering based on term-term similarity statistics conducted with the REVIGO software (Supek et al. 2011). Final enrichment results were plotted using the ggplot2 R package (Wickham 2016).

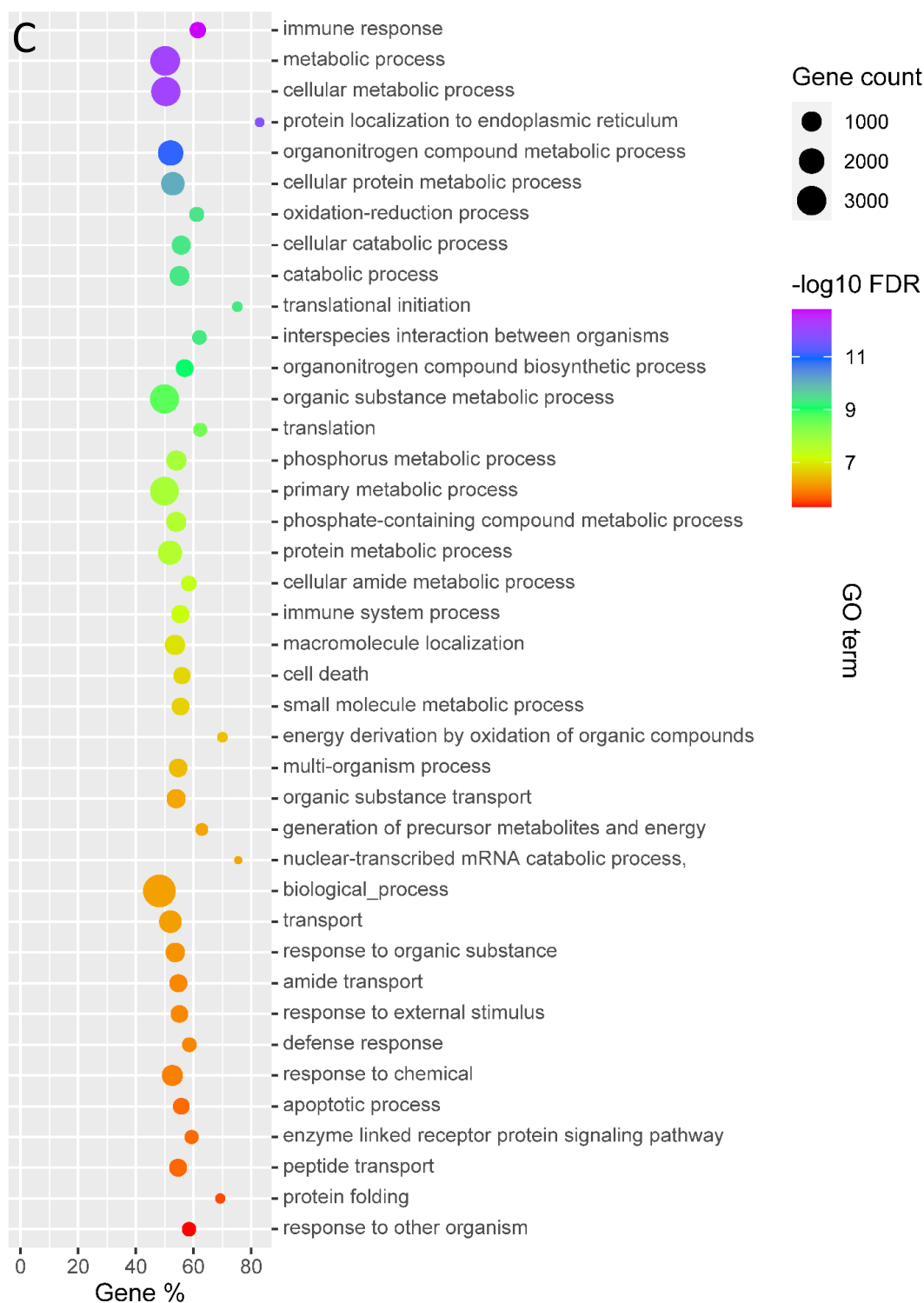


Figure 58: Biological Process GO enrichment analysis for differentially expressed genes (DEGs) observed in specific coelomic cell lineages under combined challenge of bacterial pathogen and CuNPs. Panel C A described the enriched groups observed for for eleocytes (P2). Enrichment analysis was performed using gProfiler (Reimand et al. 2019) followed with GO term redundancy filtering based on term-term similarity statistics conducted with the REVIGO software (Supek et al. 2011). Final enrichment results were plotted using the ggplot2 R package (Wickham 2016).

A

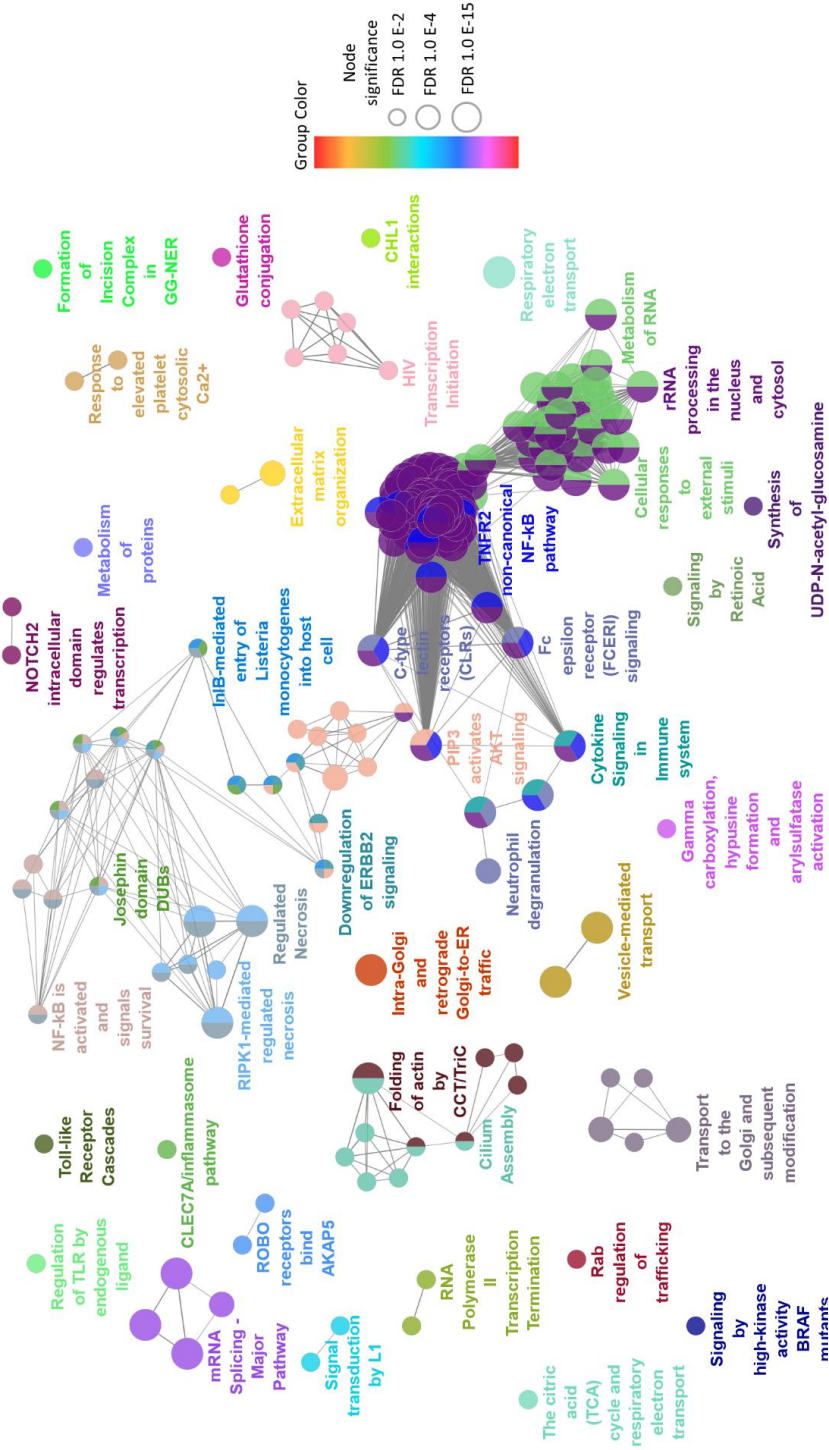


Figure 59: Network representation of pathway enrichment analysis for differentially expressed genes (DEGs) observed in specific coelomic cell lineages under a combined challenge of bacterial pathogen and CuNPs. Panel A describes the enriched pathways observed for granular amoebocytes (P1) and Panel B for eelocytes (P2). Nodes represent significantly enriched KEGG (Kanehisa et al. 2015) and, Reactome (Jassal et al. 2020) pathways. To reduce redundancy, pathways were functionally grouped and colours based on gene membership scores (kappa score), where only terms with the highest significance are shown. The edges represent the statistical association (overlaps) between the enriched terms (Bindea et al. 2009).

B

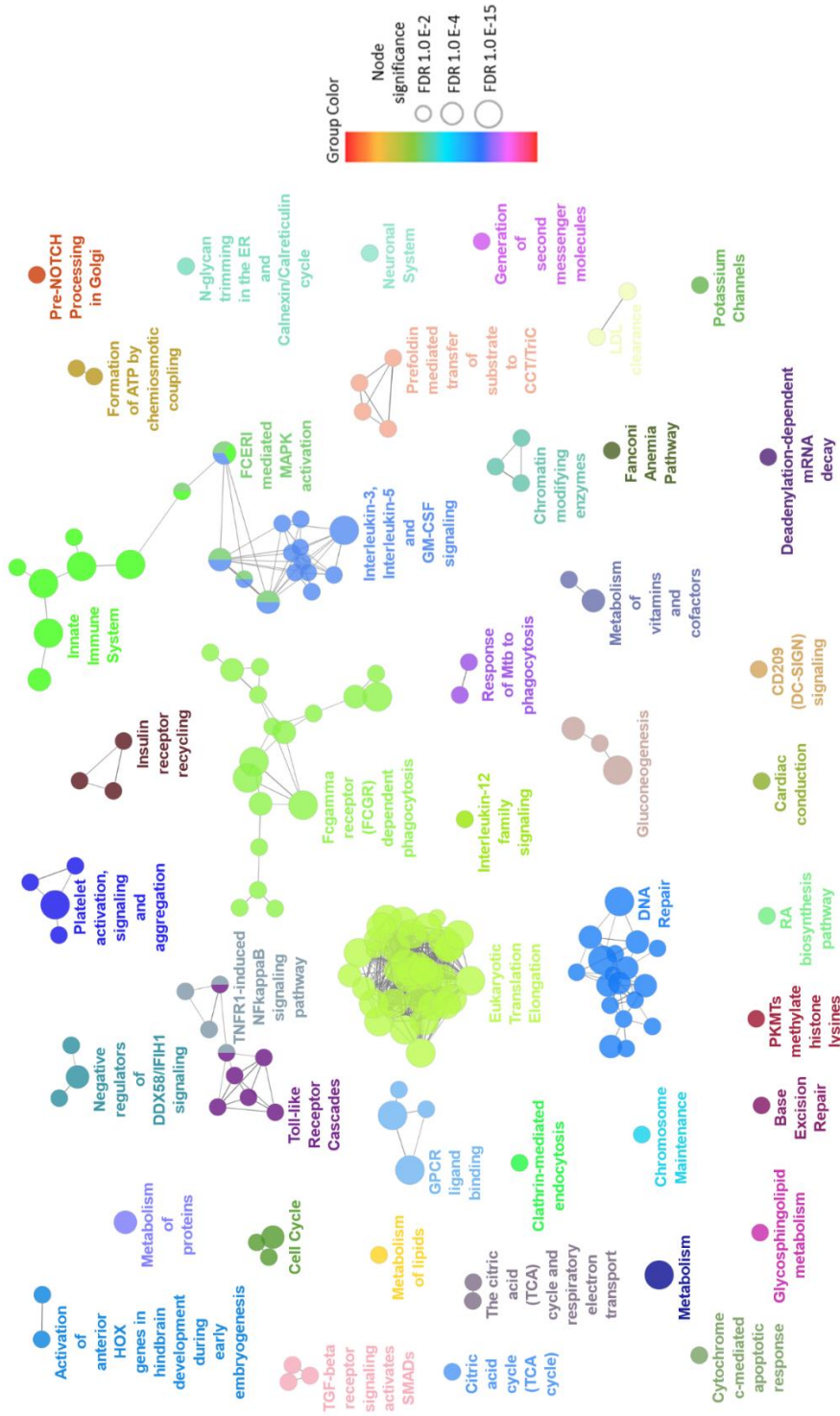


Figure 59: Network representation of pathway enrichment analysis for differentially expressed genes (DEGs) observed in specific coelomic cell lineages under a combined challenge of bacterial pathogen and CuNPs. Panel A describes the enriched pathways observed for granular amoebocytes (P1) and Panel B for eleocytes (P2). Nodes represent significantly enriched KEGG (Kanehisa et al. 2015) and, Reactome (Jassal et al. 2020) pathways. To reduce redundancy, pathways were functionally grouped and colours based on gene membership scores (κ score), where only terms with the highest significance are shown. The edges represent the statistical association (overlaps) between the enriched terms (Bindea et al. 2009).

5.3.5 Temporal response to combined bacterial and NPs challenge

Analysing DEGs from the different cell-types without separating the different time-points allowed us to identify main characteristics of the cytotoxic and immune response enhancing effect of the CuNP pre-treatment, however, this did not allow us to dissect the phasing of these events over time. To achieve a better understand about how the immune response was affected by the CuNP in time, a different analytic approach was employed. DEGs were initially clustered by their expression profile across the different time points. To enable this, a Fuzzy C-means based soft clustering method was used to identify co-regulated clusters of genes, followed by the extraction of core clusters (genes with cluster membership value > 0.5). The generated clusters were divided between Early (0-12h), Mid (12-24h), and Late (24-96h) response groups by the time of their peak expression changes (Figure 60). The genes clusters associated with the same temporal phase were used to create combined gene lists. Finally, these gene lists were used to perform pathway enrichment analysis, this time by considering the time-dependent nature of the data. Overrepresented pathways were successfully identified in all the three populations of the free-floating coelomocytes. Using this information, a functional network was generated to explore the overall cellular responses at the three different temporal phases, while at the same time associating these with the different coelomocyte populations (Figure 61).

The early phase of the cellular response was dominated by immune processes for both eleocytes and granular amoebocyte. In contrast, DEGs in hyalin amoebocytes produced only one cluster associated with the glutathione metabolism, in which all genes showed a continuous decrease in their expression profile over time (Figure 61). The peak of the expression of these genes was between the first 0-6 h measured from the start of the bacterial challenge, while they started to decrease at 18 h then normalised around 24 h following *B. subtilis* treatment. In the case of these granular ameobocytes, most significantly Toll-like receptor-associated processes were affected. Although, in eleocytes several components of Toll-like receptor signalling were also identified, the majority of the enriched pathways were associated with Interleukin-1 (IL-1) signalling. IL-1 is a known major proinflammatory mediator of the innate immune system can rapidly cause an increase in the mRNA expressions of thousands of genes (Weber et al.

2010). In the Mid phase of the immune response, metabolism-related pathways were mainly affected in all the three cell populations. While in the Late phase, DEGs in leucocytes produced more generic terms, in granular amoebocytes 'Glutathione biosynthetic processes' were affected, suggesting CuNPs induces cytotoxic effects and oxidative stress even after 24 hours.

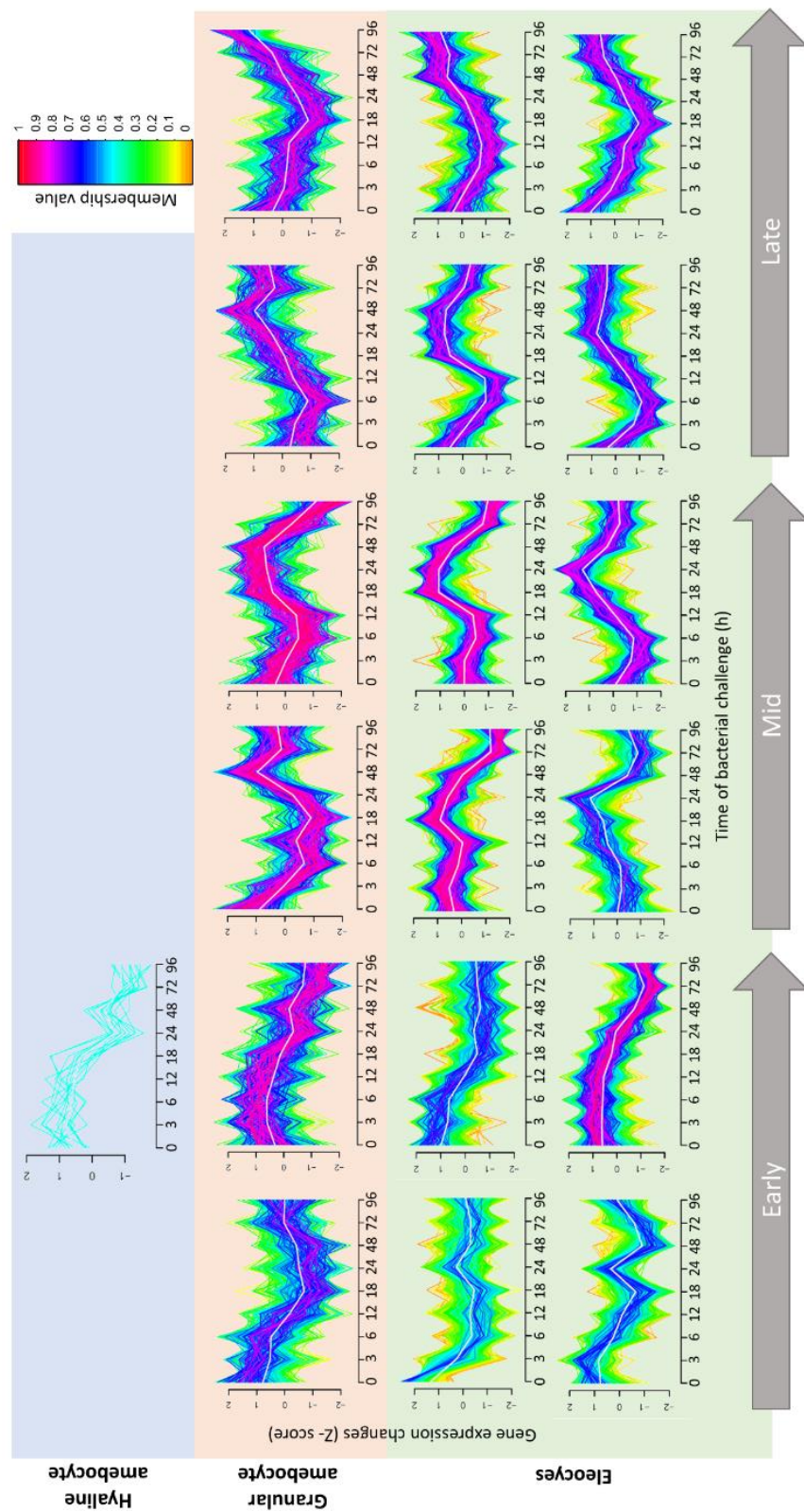


Figure 60: Cell type-specific temporal resolved gene clusters. Fuzzy C-means soft clustering was used to identify co-regulated genes, followed by the extraction of core clusters cores (genes with cluster membership value > 0.5). The generated clusters were divided between Early (0-12h), Mid (12-24h), and Late (24-96h) response groups by the time of their peak expression changes.

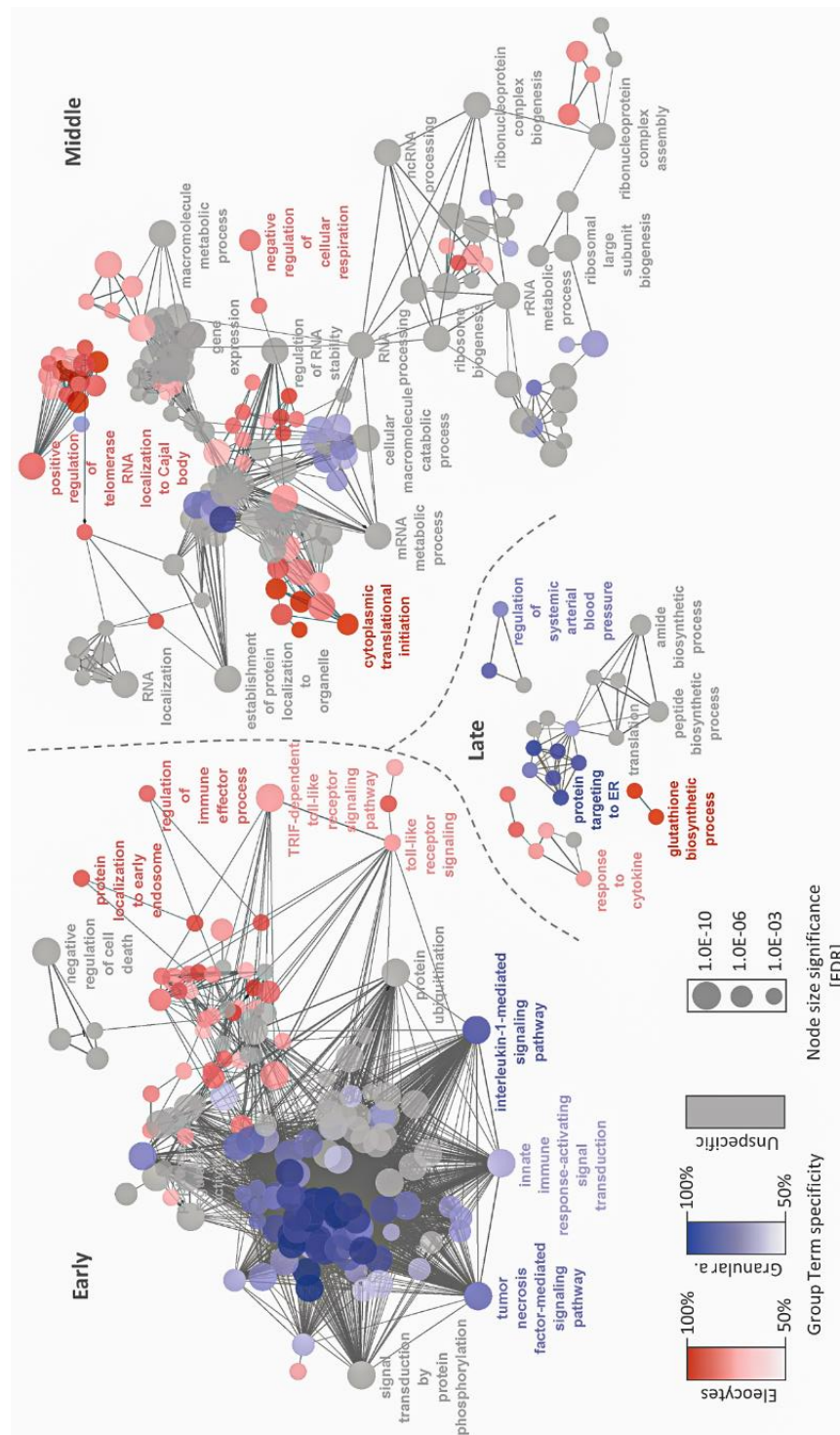


Figure 61: Eleocyte functional networking of temporal phased DEGs. Fuzzy C-means soft clustering of DEGs from eleocytes were divided between Early (0-12h), Mid (12-24h), and Late (24-96h) response groups by the time of their peak expression changes. Nodes represent significantly enriched KEGG (Kanehisa et al. 2015) and, Reactome (Jassal et al. 2020) pathways. To reduce redundancy, pathways were functionally grouped (colours) based on gene membership scores (kappa score) where only terms with the highest significance are shown. The edges represent the statistical association (overlaps) between the enriched terms (Bindea et al. 2009).

5.3.6 Combined impact of bacterial and NPs on copper metabolism

To temporally resolve the copper ion driven impact of CuNPs on the different subpopulations of coelomocytes, genes associated with the critical components of the copper homeostasis pathway including; uptake, storage, and detoxification were manually re-analysed (Table 5). Intriguingly, equivalents of the cell surface Ferric/cupric reductase, homologs to either human of STEAP or yeast FRE1 proteins could not be identified. This protein plays an essential role in performing reduction of Cu^{2+} to Cu^+ before uptake through the CTR1 high-affinity channels. Its absence may suggest either that the function is being performed by an equivalent reductase, as yet unidentified, or the reduction may be exploiting the earthworm gut microbiome. In contrast, we believe the absence of earthworm ATXO1 homolog is a technical issue associated with genome annotation, since it is present in the transcript assemblies. The failure to identify a ATXO1 genomic loci is probably due to the small size of the transcript, the coding region being distributed between numerous exons and the limited size of the conserved functional motif.

Analysis of expression across the time course in the subpopulations of coelomocytes (Figure 62), revealed that eleocytes were the most responsive to CuNP challenge whilst hyaline and granular amebocytes displayed mainly small non-significant expression changes in copper trafficking genes. Comparing generalist chaperones known to respond to cytosolic Cu (Table 5) suggests that in eleocytes this role is played by CutC, a known copper-specific chaperone, whilst granular amebocytes express both phytochelatin synthase (PCS) isoforms at different times together with a lower elevation of CutC and metallothionein isoforms. Interestingly, at 24 hours in hyaline amebocytes, where PCS_1 was upregulated, we observe a decrease in MT_2 and CutC_1, suggesting a compensatory role between the different generalist metallo-chaperones.

Detailed analysis of eleocytes revealed downregulation of copper transporters, at the plasma (CTR1), mitochondrial (COX11) and vesicular (ATP7A) membranes. Furthermore, chaperones involved in the delivery of copper to Cytochrome C Oxidase 1 (SCO1/COX11) and Superoxide Dismutase (CCS), were also strongly downregulated at each of the measured time-points. Combining these observations with the elevation of MTF1 and strong induction of the CutC suggests that, when exposed to significant excess

copper, eelocytes attempt to induce mechanisms to prevent the internal copper system being overwhelmed. These include reducing uptake, lowering targeted chaperone-mediated transport whilst increasing copper specific binding by CutC.

Table 5: Components of the Copper trafficking pathway.

Symbols		Description	Process	Compartment
Human	Earthworm			
STEAP	Not found	Copper Reductase	$\text{Cu}^{2+} \rightarrow \text{Cu}^+$	Plasma membrane
CTR1 (SLC31A1)	CRT1_1, CTR1_2	High-affinity copper transport	Copper transport	Plasma membrane
MT	MT_1, MT_2	Metallothionin	Metallo-chaperone	Cytosol
Not present	PCS_1, PCS_2	Phytochelatin Synthase	Metallo-chaperone synthesis	Cytosol
CutC	CutC_1, CutC_2	Copper Homeostasis Protein CutC	Metallo-chaperone (Cu specific)	Cytosol
ATOX1	Not identified	Antioxidant 1 Copper Chaperone	Metallo-chaperone to ATP7A	Cytosol
CCS	CCS	Copper chaperone for superoxide dismutase	Metallo-chaperone (delivery to SOD)	Cytosol
COX17	COX17	Cytochrome C Oxidase Copper Chaperone COX17	Metallo-chaperone (delivery to Mitochondrion)	Cytosol
COX11	Cox11	Cytochrome C Oxidase Copper Chaperone COX11	Mitochondrial Metallo-chaperone	Mitochondria
SCO1	Sco1	Cytochrome C Oxidase Assembly Protein	Recipient of copper	Mitochondria
SCO2	Sco2	Synthesis Of Cytochrome C Oxidase 2	Recipient of copper	Mitochondria
ATP7A/7B	ATP7A	ATPase Copper Transporting protein Alpha	Transmembrane Copper transporter	Vesicular (golgi) membrane
MTF1	MTF1	Metal Transcription Factor 1	Transcription Factor	Transcriptional regulation of MT and other metal responsive genes

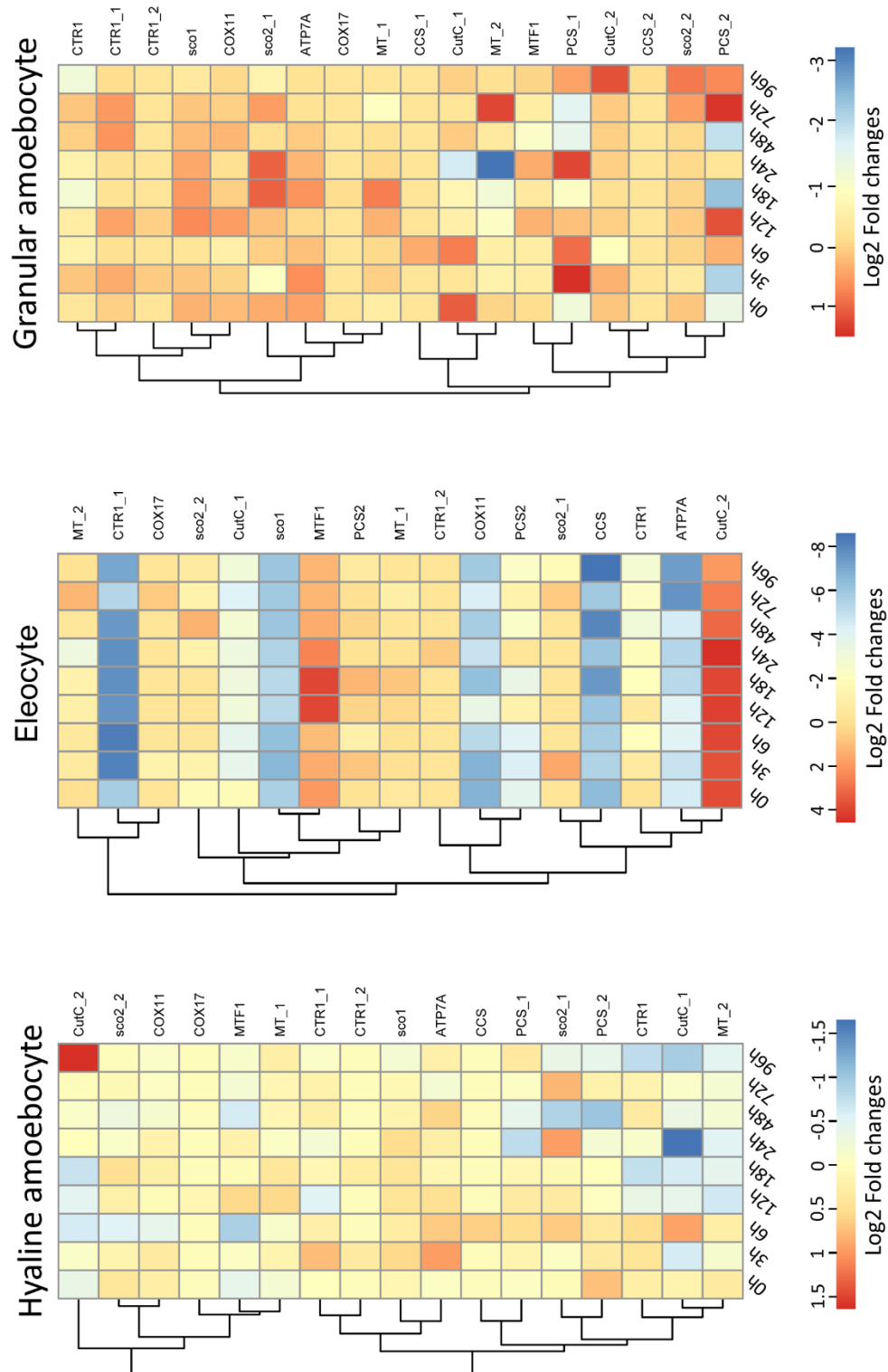


Figure 62: Coelomocyte cell-type specific temporal analysis of expression changes within copper trafficking genes under the combined bacterial (*B. subtilis*) and CuNP challenge. Changes in genes associated with Cu metabolism were analysed across the full-time course. Changes in expression were calculated against time-matched controls (no CuNPs or bacterial challenge). Hierarchical clustering was performed in R using pheatmap (Kolde 2012) using Pearson correlation as distance measure for row clustering.

5.3.7 Combined impact of bacterial and NPs effect on the Toll-like pathway

To understand the temporal impact of the combination of CuNPs treatment and *B. subtilis* challenge on innate immunity, we investigated the affected genes associated with the “Toll-like receptor signalling pathway”. Initially we identified *E. fetida* homologs corresponding to the elements linked to this KEGG pathway and subsequently mapped on those genes displaying differential expression (DE) for each of our coelomic cell types. Hyaline amebocytes showed no DE genes associated with the toll-like pathway at any point within the time course exposure, whilst granular amebocytes show a small number (20 genes). Eleocytes, however, showed an extensive range of DEGs associated with this pathway (57 genes), these were extracted and mapped, using their gene symbols via associated uniprot ID, to link them to the human KEGG homologs (mapping table is provided in Appendix 5.1). Those genes identified and DEGs are displayed in Figure 63.

Differentially expressed genes from granular amebocytes and eleocytes were subsequently clustered by their expression change over time; then analysed by applying the earlier mentioned three response phase-based methodology (GA: Appendix 5.2, Eleocytes: Figure 64). The interpretation of the data is confounded by a number of factors including complexities of nomenclature (synonyms and database mapping), the presence of multiple earthworm isoforms, as well as the independent/combined action of copper and the bacterial challenge. We have included an annotation mapping table (Appendix 5.1), however, for clarity it should be noted that the key DE players within the toll pathway identified multiple times within the temporal analysis (e.g. MAK14K and RAC1) represent multiple independent isoforms identified from the *E. fetida* genome. Despite these complexities, clustering within Toll-like receptor pathway components clearly reveals a temporal pattern showing with highest number of Toll-like receptor signalling associated genes are present in two waves which were located between 0-3 hours and 18-24 hours (Figure 64).

Closer examination of the data suggests induction of the MAPK signalling pathway (MAPK14/MAP3K7), NF-KB signalling (NFKB1/NFKB1A) and cFOS (FOS) at time zero, independent of bacterial challenge. At time zero the cells have been exposed to CuNPs for 24 h and the upregulation of the copper chaperone CutC would indicate elevated intracellular copper. Elevated copper is a known activator MAPK, NFKB and cFOS

pathway (VanHook 2012, Tchounwou et al. 2008), therefore we would postulate that this upregulation is a direct result of the CuNPs exposure within the initial 24 h before bacterial exposure. After introduction of the bacterial challenge, we observe a number of cycles of the PI3K-Akt pathway being upregulated at multiple time points, with RAC1 (3h & 24 h), PI3K (PIK3CD 24 h, PIK3CB 18/24 h), and AKT(1/3) (6 & 24 h). The lipopeptide of *B. subtilis* is well established as an activator of PI3K-Akt (Zhao et al. 2017), illustrating that the CuNP exposure has sensitised the earthworm's innate immune response since equivalent stimulation was not seen in the absence of NPs.

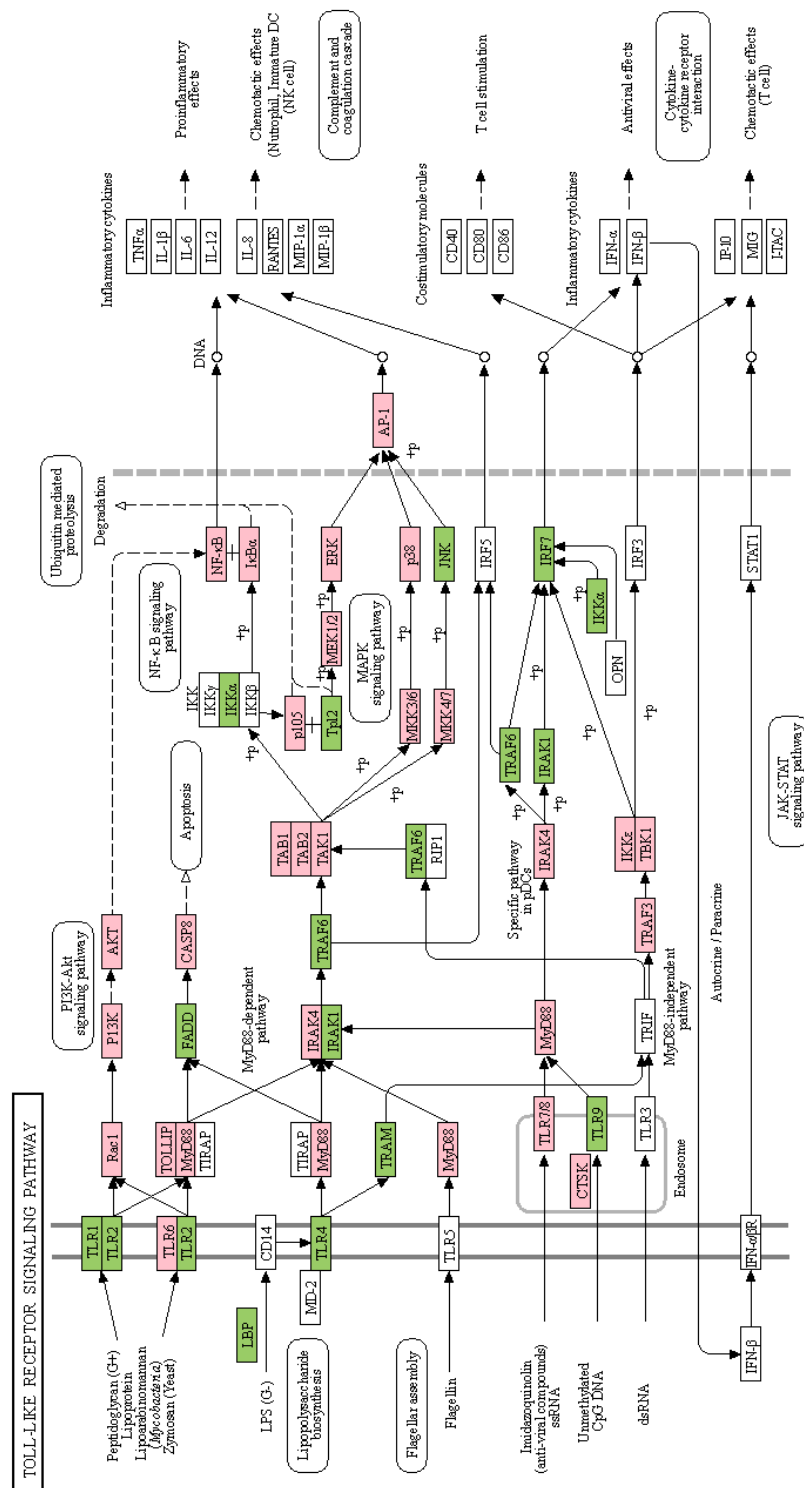


Figure 63. Identification of earthworm homologs for components of the toll-like signalling pathway. Differentially expressed genes associated with the “Toll-like receptor signalling” KEGG pathway. *E. fetida* homologs are indicated by coloured shading of pathway components (green or red) whilst no homologs could be identified for genes shown with a white background. The genes shaded red are differentially regulated (up or down) in eleocytes at any point during the experimental time course.

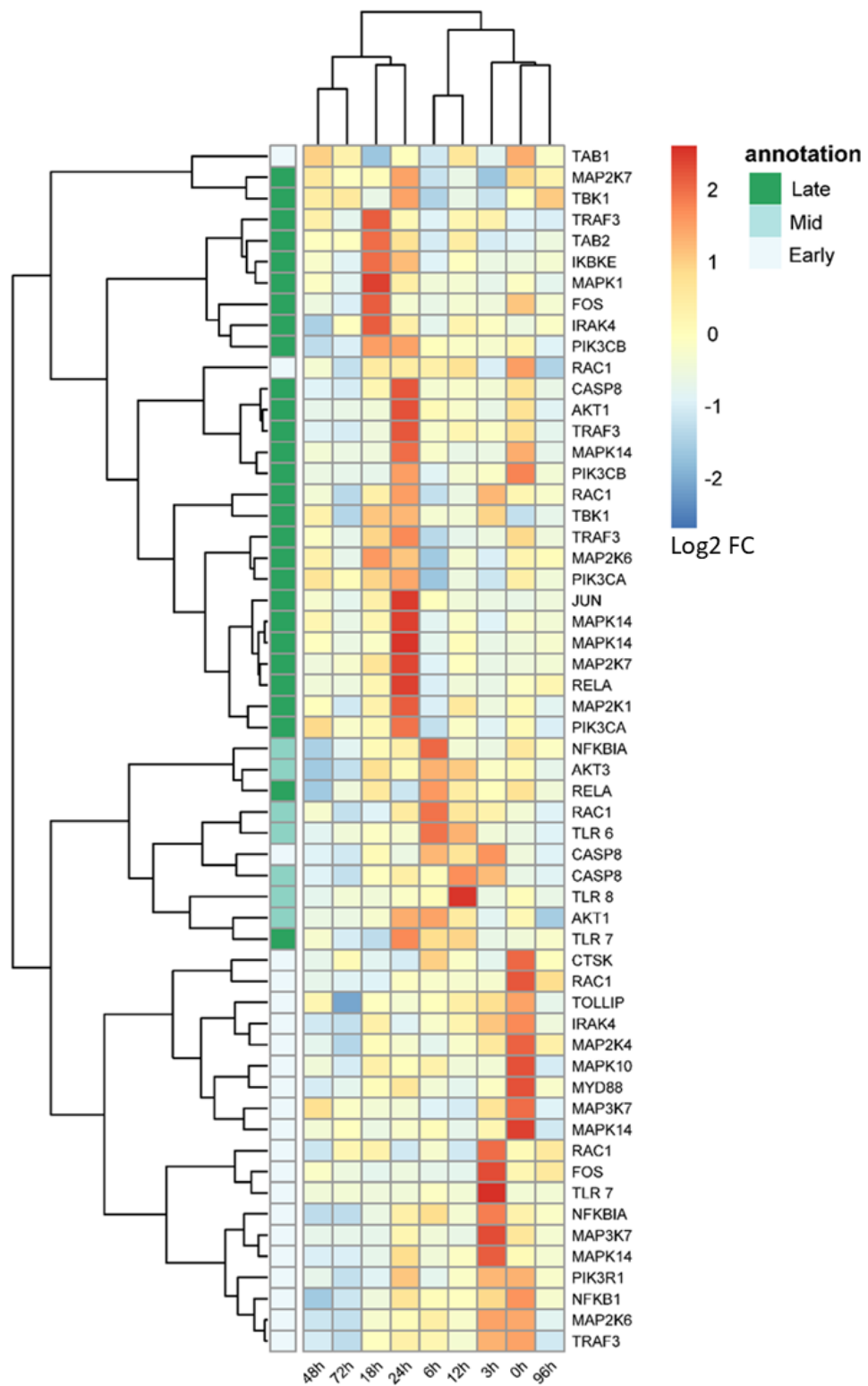


Figure 64. Temporal analysis expression changes in eleocytes under the combined bacterial (*B. subtilis*) and CuNP challenge. Differentially expressed genes associated with the “Toll-like receptor signalling” KEGG pathway were evaluated across the full-time course of analysis. Changes in expression were calculated against time-matched controls (no CuNPs or bacterial challenge). Hierarchical clustering was performed in R using pheatmap (Kolde 2012) with the Pearson correlation based distance measure option.

5.4 Discussion

5.4.1 Experimental design

Selecting an experimental design that uses time as a continuous variable and fitting a spline model on the individual genes allowed the examination of temporal changes under multiple challenges in a cost-efficient manner. The applied design resulted in a balance between the cost of each analysis and the high number of sampling points required to provide temporal resolution for each of the three coelomocyte populations. The approach has allowed the identification of differentially expressed genes between the different treatments, without the necessity of using technical replication at each timepoint. By directly injecting NPs into the earthworms coelomic cavity, we were able to examine how nanoparticles can modulate the earthworm homeostasis in their true nanometric form, especially from the view of the innate immune system.

Typically, one of the biggest technical challenges of earthworm ecotoxicological studies that use soil as an exposure medium is the problem of maintaining a reproducible exposure between the individuals. During a soil exposure, dose could be highly dependent on the activity (feeding behaviour, location in the soil) of the individual earthworms, resulting in increased background noise in the experiment. The injection method which was applied in this experiment not only helped to avoid the environmental effect caused by transformation of the nanoparticles, but also ensured that each individual was exposed to a precise amount of NPs independently of their feeding behaviour.

The elimination of the long cell-washing procedures prior to cell sorting (FACS) helped to reduce handling time, potentially reducing any artificial transcriptomic changes in the coelomocytes. This rapid cell preparation method also promoted cell viability and reduced cell-cell adhesion, resulting in a better separation between the different coelomocyte populations. Also, the avoidance of centrifugation based cell sedimentation followed by cell resuspension resulted in lower lysis of coelomocytes during preparation. The approach also likely decreased the free RNA content in the coelomic fluid suspension, RNA that could be carried over during the cell sorting procedure and lead to background noise in the cell-type specific sequencing data. By

choosing a low-input RNA library preparation kit, it was possible to successfully construct RNA-seq libraries from relatively small number of cells. This greatly reduced both the time of the cell-separation process, as well as the volume of the created population-specific cell-suspension later employed in the RNA extraction step. Bioinformatic quality control methods showed that the two-step amplification method, necessary for the the low-input RNA library protocol, did not result in a detectable increase in the number of technically duplicated (artificial) reads, yielding a high proportion of quality reads suitable for subsequent RNA-seq analysis.

5.4.2 Differential gene expression analysis

Likely due to the application of only a mild bacterial challenge, the differential expression analysis could not detect any significant gene expression difference between the control and the *B. subtilis* challenged animals in any of the examined cell populations. However, when earthworms were pre-injected with CuNPs and then challenged with bacteria, a strong transcriptomic response was detectable between the control and co-treated organisms. Although the lack of significant expression changes in the *B. subtilis* treated animals prevented us observing an immune response stimulated from bacteria alone, it supported the reliability of the chosen statistical method to identify DEGs. The absence of a significant number of DEGs passing our statistical threshold when analysing *B. subtilis* immune challenged animals, suggest that the spline regression-based method was robust enough and it was not too sensitive to the possible noise caused by slight differences in the cell-sorting efficiency between the different samples.

5.4.3 Time independent, spatial analysis of DEGs

5.4.3.1 Hyaline ameobocytes

In the case of hyaline ameobocytes, the most determinant transcriptomic response was observed related to 'thyroxine caused cellular stimulus'. It is known that the components of the thyrotropin-releasing hormone (TRH) neuropeptide-receptor pathway are highly conserved across different phyla. During the last few years thyroid hormone receptors (TRs) have been successfully identified in many invertebrate families such as annelids and molluscs. Furthermore, TRH-like neuropeptide precursors were even described in nematodes species (Van Sinay et al. 2017). Evidence for the presence

of thyroglobulin-like molecules (thyroid hormone precursor) in earthworm (*Eisenia fetida*) neural tissues were also documented as early as in the 1980s (Marcheggiano et al. 1985). While it is known that thyroid hormones (THs) play a key role in vertebrate development, growth, and metabolism, their physiological function is much less well understood in the case of invertebrate species (Mourouzis et al. 2020, Taylor and Heyland 2017). It is also known that in higher-order vertebrates the serum copper level is strongly regulated by thyroid hormone (Mittag et al. 2012). Our results suggest maybe similar biological processes can regulate the copper content of the coelomic fluid, this could result in altered thyroxin homeostasis following increased copper load via the injection of CuO nanoparticles. The general toxic effect of CuNPs on the population of hyaline amebocytes seemed to be supported by the observed change in glutathione metabolism. Glutathione is known to play a key role in protecting cells from oxidative damage caused by reactive transition metals, such as copper, and thereby contributing to maintaining the redox homeostasis of the organism (Bhattacharjee et al. 2017, Mwaanga et al. 2014).

5.4.3.2 Granular amebocytes

Compared to hyaline amebocytes, where the affected biological processes were mainly associated with a general toxic response possibly induced by CuNPs exposure, granular amebocytes presented a higher number of affected immune-related processes. The activation of the Toll-like pathway suggested that granular amebocytes were highly involved in the launch of innate immune responses. The upregulation of the different RNA metabolism-related genes are also known to modulate several immune system processes, since several post-transcriptional RNA modification-related processes, such as RNA splicing, have important roles in regulating the innate immune response (Li et al. 2012, Heward and Lindsay 2014, Carpenter et al. 2014).

5.4.3.3 Eleocytes

Due to the lack of phagocytic activity, most of the earlier earthworm immunology studies neglected the examination of eleocytes. Early research suggested eleocyte function was mostly limited to nutrient metabolism (Šíma 1994). However, these cells are also known to contain different immunologically important molecules, such as riboflavin (Plytycz and Morgan 2011, Plytycz et al. 2006, Mazur et al. 2011). Within our

study, out of the three different examined coelomocyte populations, the highest number of innate immune response-related terms were associated DEGs identified in eleocytes. This raises the possibility that eleocytes could play a more important role in the earthworm's innate immune response than expected. The combined treatment of NPs and *B. subtilis* caused expression changes in several immune response initiation related pathway, such as Toll-like receptor and IL-1-receptor, with the 'Platelet activation, signalling and aggregation" pathway was also affected. The appearance of the platelet related terms might be a reflection of the eleocytes recently described neutrophil extracellular trap (NET) like nature (Homa 2018)

5.4.4 Temporal gene enrichment analysis

By analysing the gene expression changes over time, then grouping DEGs using a well-characterised vertebrate like nomenclature (three-phase response), we were able to characterise the earlier described cell-specific responses in a temporal manner.

Although the clustering of hyaline amebocyte associated DEGs only resulted in a single cluster, it was clear how the NPs cytotoxic effect changed during the exposure time. Glutathione metabolism and thyroxine stimulus-related genes showed peak expression within the first 6 hours of sampling, then started to decrease after around 18 hours and returned to normal after 24 hours. This suggests that CuNPs induced heavy metal stress was minimalised between 18-24 hours after the start of bacterial challenge.

The early response of the Toll-like receptor signalling pathway in the granular amoebocytes was less striking, since it is well documented that Toll-like receptors play a crucial role during the initiation of innate immune response by recognising different nanometric and pathogen-associated molecular patterns (PAMPs). Furthermore, recently PAMPs have also been successfully connected with the possible immune modulation effect of certain metal nanoparticles due to their ability to recognise PAMP like, artificial nanometric patterns (Du et al. 2017). It was intriguing to observe that, concurrent with the PAMP response, eleocytes seemed to be strongly affected by Interleukin-1. Toll-like receptors and IL-1 receptor family have highly similar functions, both triggering innate inflammation. Since both receptor families contain a highly analogous Toll-IL-1Receptor (TIR) domain, their structural similarity is also unquestionable. One of the significant differences between the two types of receptors

is that whilst Toll-receptors are usually activated by different PAMPs originating from pathogen stimulation, the activation of IL-1 receptors are caused through the recognition of IL-1 cytokines (Kuno and Matsushima 1994). These results suggests that highly mobile granular amoebocytes - which are known to be an effective antigen scavenger - were stimulated through their Toll-like receptors during the early response stages. The activation of these cells might have started cell-signalling processes and, by doing this, activate eleocytes via their IL-1 receptors. This hypothesis is supported by the fact that the broadest set of Toll-like receptors were expressed in granular amoebocytes while this number is much lower in eleocytes (chapter 4).

5.4.5 Metal stress response

Using the newly generated, highly complete *E. fetida* genome annotation (chapter 3) we could identify many full-length homologs of the known Copper trafficking genes, including multiple isoforms for many of these components. Using these genes as a mapping template for the analysis of the spatio-temporal data, we were able to follow how the different cell-types were affected by the CuNPs and how it changed through the 96 h experimental period. The results of this analysis highlight that eleocytes were most responsive to CuNP challenge, indicating that this population could play a key role in the regulation of copper homeostasis within the coelomic cavity. The origin of eleocytes is yet to be fully described, although their morphological and histochemical properties has led some to suggest the chloragogen tissue as the source (Homa 2018, Liebmann 1942). During the last decade, a high number of ecotoxicological experiments have shown that the chloragogen tissue plays a significant part in the outstanding resilience of earthworms to most heavy metals (Cancio et al. 1995, Świątek et al. 2020). The results of this study suggests that *E. fetida* eleocytes not only have histochemical similarities to chloragocytes but, in similar manner to the chloragogen tissue they also seem highly involved in heavy metal metabolic processes. While in eleocytes a strong downregulation of the copper ion importer (CTR1) gene was observed during the whole experimental period, it reached a negative peak between 27-30 h (3-6 h of *B.subtilis* challenge) after the CuNPs were injected. This suggests that copper ion uptake into eleocytes is mainly regulated by the expression of CTR1 gene and the these cells reached their copper ion uptake capacity after 27 hours from the start of the exposure. The

strong downregulation of the two copper chaperones (CCS, ATP7A) suggest eleocytes try to avoid the transport of Cu(I) ions to the Nucleus and Golgi-apparatus. The saturation of cells with copper ion was also represented by the upregulation of the CutC gene, which is known to have a great importance in efflux trafficking of the cuprous ion. These results suggest that the integrity of eleocytes against the copper stress was achieved by cutting the main copper ion transport lines and at the same time pumping the excess Cu(I) ions back from the cytoplasm to the coelomic fluid. In contrast to eleocytes in granular amoebocytes the highest copper response was provided by a heavy metal detoxification related gene named phytochelatin synthase (PCS). It was interesting to see that different isoforms of the genes were activated at distinct time points. This may result from the different isoform preferences between earthworm individuals rather than differences due to temporal changes.

5.4.6 Conclusion

Results here presented provided the first broad spatiotemporal insight into how copper nanoparticles can modulate the innate immune system in *E. fetida*. Studying the different cell populations separately allowed us to observe how the individual coelomocyte populations respond to the CuNP caused oxidative stress as well as identify their most affected copper metabolism-related pathways. Since coelomocytes represent the key participants of both the cellular and humoral innate immune response, the impact of nanoparticle induced metal challenge modified their immunological functions as well. Analysing the cell-specific dataset indicates eleocytes play a major role in the copper detoxification process, likely resulting from their chloragocyte-like characteristics. Overall, it appears that eleocytes may have a critical role maintaining the heavy metal homeostasis of the coelomic fluid.

6 Comparing NP effect within direct and Indirect exposures

6.1 Introduction

6.1.1 Environmental effects of engineered metal NPs

By decomposing a high amount of organic materials originated from plants, and homogenising it with the inorganic compounds (minerals) from soil, earthworms are one of the most crucial terrestrial ecosystem engineers. Newly developed industrial and agricultural applications of the different metallic nanomaterials (MNM) such as the use of AgNPs as nanopesticides or nanofertilizers, resulted in a significant rise in the release of these particles into the environment (Dubey and Mailapalli 2016). This developed an increased concern about the environmental fate of the NMs and their effects on both the terrestrial and freshwater ecosystems.

Due to their high importance in keeping the continental ecosystem sustainable, earthworms are commonly used as terrestrial invertebrate toxicological sentinel species (Bundy et al. 2008). Although a relatively detailed knowledge was built on the ecotoxicological effect of the different metal NPs using earthworms as a model organism, their possible innate immune system modulation effects have not been studied extensively. The majority of these immunity focused studies used *in vitro* experimental design that does not represent the real ecological interaction between the NPs and the used model organisms (Ploeg et al. 2014). Due to the high organic and inorganic content of the soil, combined with the high surface reactivity of most nanoparticles in a soil environment, these particles can be physiochemically transformed in a short period of time (Tourinho et al. 2012). In most cases these transformations result in nanoparticles losing their nano-specific physiochemical characteristics and create complex aggregates with the colloidal components of soil (El Hadri et al. 2018). However, the environmental transformation of metallic NPs can manifest in the creation of different ionic compounds, a number of studies have shown that ionic compounds originated from engineered nanomaterials (ENM) can have significantly different effect on the ecosystem than their natural representatives (Novo et al. 2020).

6.1.2 Experimental design

The results presented in Chapter 5 gave an overview of the direct impact of CuNPs on the immunological modulation of earthworm coelomocyte populations. Since that experiment directly injected nanoparticles into the coelomic cavity, it provided very little to no information about the real ecotoxicological risks of the copper-oxide nanoparticles through environmental exposure. The aim of this experiment was to understand if environmentally transformed metallic NPs exposed via the soil can modulate the earthworm's innate immune system in similar way to the directly injected nanometric form. In this chapter we implemented a more ecotoxicologically realistic experimental design in which earthworms were exposed to copper in two different nanoparticulate forms together with exposure to the ionic form. To follow the environmental effect of copper related treatments, soil was used as a carrier media. In the case of ionic copper three different doses were used. The purpose of this treatment was to allow the separation of the ionic copper effect from the nanoparticle specific biological responses.

6.2 Materials and Methods

6.2.1 Experimental design

All tests were conducted in a standard agricultural soil (Lufa 2.2) (LUFAspeyer, Germany), a commonly used natural standard soil in earthworm chronic toxicity tests following OECD (2000) guidelines. The selected soil has a representative pH of 6 and organic matter content of 4.2 %. The engineered nanomaterials (ENMs) used for the experiment were 3 nm CuO, porous SiO₂-CuO hybrid NMs (Hadrup et al. 2020) and ionic copper (in the form CuCl₂) alongside appropriate controls (Figure 65). The test concentrations used were 0, 31.37, 65.6, 287 mg Cu/kg (dry weight soil) for the copper ions and 0, 65.6, 287 CuO ENMs and 65.6, 287 mg Cu/kg for the porous SiO₂-CuO hybrid ENMs. The exposure concentrations employed were based on EC values based on the response to ionic copper. For each material, the concentrations are based on the amount of copper present, the total mass of either copper chloride for the ionic, or the measured amount of copper in the NPs. Four replicates were used for all treatments. Since all exposures were run concurrently, effects could be benchmarked against a universal control treatment. This comprised eight separate replicates of 4.2 Lufa soil without Zn or Ag amendment. The nanomaterials were dried prior to an appropriate quantity being mixed into 100 g of exposure substrate (soil), which was then mixed into the remainder of the soil to achieve the required dose. This method was chosen to ensure a consistent exposure between replicates and to avoid transformation in the water phase (Waalewijn-Kool et al. 2012). Each replicate comprised a total of 700 g per container wetted to 50% of water holding capacity and containing 10 individual worms. The toxicity test procedure followed OECD guidelines. Ten adult, fully-clitellated earthworms were added to the soil surface. All containers were then placed in a controlled temperature room (20 +/- 1 °C in a 12:12 hour light:dark cycle) for a total of 56 days. After 14 days, the containers were sorted and the numbers of earthworms alive in each counted, weighed and returned to the soil for a further 14 days. At the end of the full 28 days, the earthworms were again sorted from the soil and weighed. All test soils were returned to the controlled temperature facility for a further 28 days to allow laid cocoons to hatch. For counting, containers were placed in a water bath at 60°C

for 15 minutes to force juveniles to the soil surface to allow counting. Exposures were performed at UKCEH by Dr Alexander G. Robinson (UKCEH).

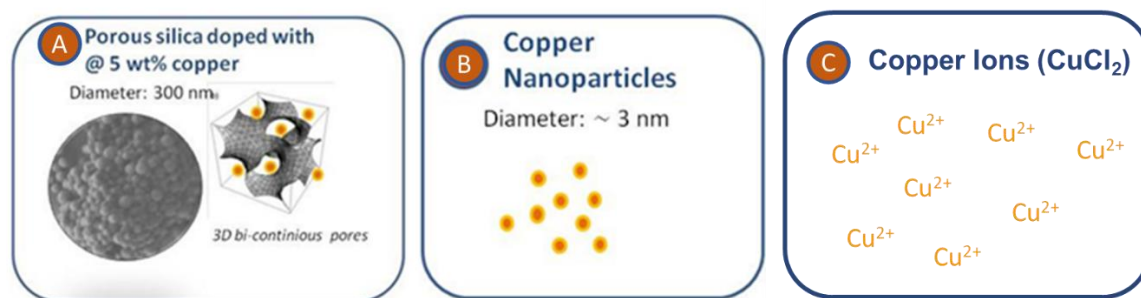


Figure 65. Exposure materials included silica doped with copper (CuSi: panel A), copper oxide NPs (CuNPs: Panel B) and Copper Ions (Cu⁺: Panel C).

6.2.2 RNA isolation

RNA was generated using approximately 20 segments of gut/chlorogen tissue isolated 15 segments distal to clitellum using the Zymo Quick-RNA Miniprep (Zymo Research, CA, USA). RNA extractions were performed by Dr Green Etxabe (UKCEH).

6.2.3 Sequence generation

RNA quality control and sequence generation were performed by Edinburgh Genomics using TruSeq Standard mRNA library kits (Illumina Inc, CA, USA) with pair end 100 bp sequence generation being performed on a NanoSeq platform (Illumina Inc, CA, USA).

6.2.4 Read pre-processing and QC

The basic quality scores of the generated paired-end data were inspected by analysing the raw reads using the FastQC (Andrews 2010) quality control tool. Following the primary quality check, potential adaptor and index related sequences were trimmed off before reads with low quality scores were removed using the Trimmomatic (v0.39) (Bolger et al. 2014) read pre-processing tool. To make sure that the data pre-processing was sufficient, the quality of the trimmed reads were verified by running FastQC on the trimmed data. Sequence duplication was further investigated using the “MarkDuplicates” command within the Picard analysis suite (Broad Institute 2018).

6.2.5 Mapping and read counting

Pre-processed reads were mapped to the annotated reference genome (Chapter 3) by supplementing the predicted genes in a GTF file format to STAR RNA-Seq aligner (Dobin et al. 2013), using the default stringency and mapping parameters. Raw gene counts then were generated from the bam files by utilising the built in “GeneCounts” quantification option of STAR aligner.

6.2.6 Differential gene expression analysis

The RNA-Seq differential gene expression analysis was conducted by using the Sartools R package in an RStudio environment (Varet et al. 2016). However, Sartools has the option to run both the edgeR and DESeq statistical module, for this analysis differentially expressed genes (DEGs) were identified by running the DESeq2 (Love et al. 2014) based pipeline of the package. To enable cross-sample comparisons the raw read counts first were normalised for systematic technical biases such as sequencing depth and RNA composition with the built in normalisation methods of the DESeq package (median of ratios). The P-value adjusting method was set to Benjamini-Hochberg option with a threshold of statistical significance of 0.05. To analyse and visualise the possible intersections between the lists of differentially expressed genes in the case of the distinct doses and conditions, DEGs were plotted using the DiVenn web based visualisation tool (Sun et al. 2019). Principal component analysis (PCA) was conducted using the interactive PCAExplorer R package (Marini and Binder 2019).

6.2.7 Functional enrichment

Following DEG identification, the lists of differentially expressed genes were extracted and used to conduct functional enrichment analysis in the case of both up and down regulated genes. Gene Ontology (GO) over-representation analysis was performed using the g:GOST module of the gProfiler web server (Raudvere et al. 2019). Significance values were corrected using the Benjamini-Hochberg (FDR) method, with the threshold set to 0.05. For more accurate statistical results, the list of background genes was limited to genes which were successfully annotated in the case of the *E. fetida* reference genome (Chapter 3). To get a better idea about the effected biological processes, the long list of successfully enriched functional GO terms were summarised and redundancy filtered buy supplementing them into the REViGO web interface (Supek et al. 2011).

Final enrichment results were plotted using custom scripts within the ggplot2 R package. Pathway enrichment results were conducted using the ClueGO plugin (Bindea et al. 2009) of the Cytoscape software, with a significance cut-off of 0.05 corrected with BH (FDR) and based on the KEGG (Kanehisa et al. 2015) and Reactome databases (Jassal et al. 2020) .

6.3 Results

6.3.1 Sequencing and quality trimming

Trimming, and removal of adapter and low-quality sequences yielded more than 4,600 M, a 98.3% overall read survival representing ~29 M reads per sample. When analysing the samples using FastQC tools, a slight increase in the number of duplicated reads was observed with an average of ~73% read duplication across the samples. A high read duplication rate in RNA-Seq data could occur due to both the natural over sequencing of highly expressed genes, as well as resulting from multiple technical errors (during library construction or sequencing). To gain more information about the nature of the duplicated reads, samples were also analysed with Picard using the “MarkDuplicates” function. This function not only measured overall duplication rate across the samples revealing around 20% duplication but also assigned approximately 2.5% of the duplicated reads as optical (bridge PCR based) read duplicates. This result is indicative of flow-cell overloading, a sub-optimal technical error leading to read duplication. The low variation of read duplication between the different samples also suggested that the increased duplication rate resulted from a significant number of highly expressed genes, a finding common in most RNA-Seq data (Parekh et al. 2016).

6.3.2 Read mapping

To assess the genetic divergence between the individuals and the reference genome (chapter 3), mapping statistics were analysed based on the outputs of STAR-aligner. The percentage of successfully aligned reads showed relatively low variation across the different samples, with an average of 75% uniquely mapped reads per sample (Figure 66). These results suggest, that the genetic variation between the sacrificed individuals is likely to have little to no effect on the differential analysis between the various experimental conditions.

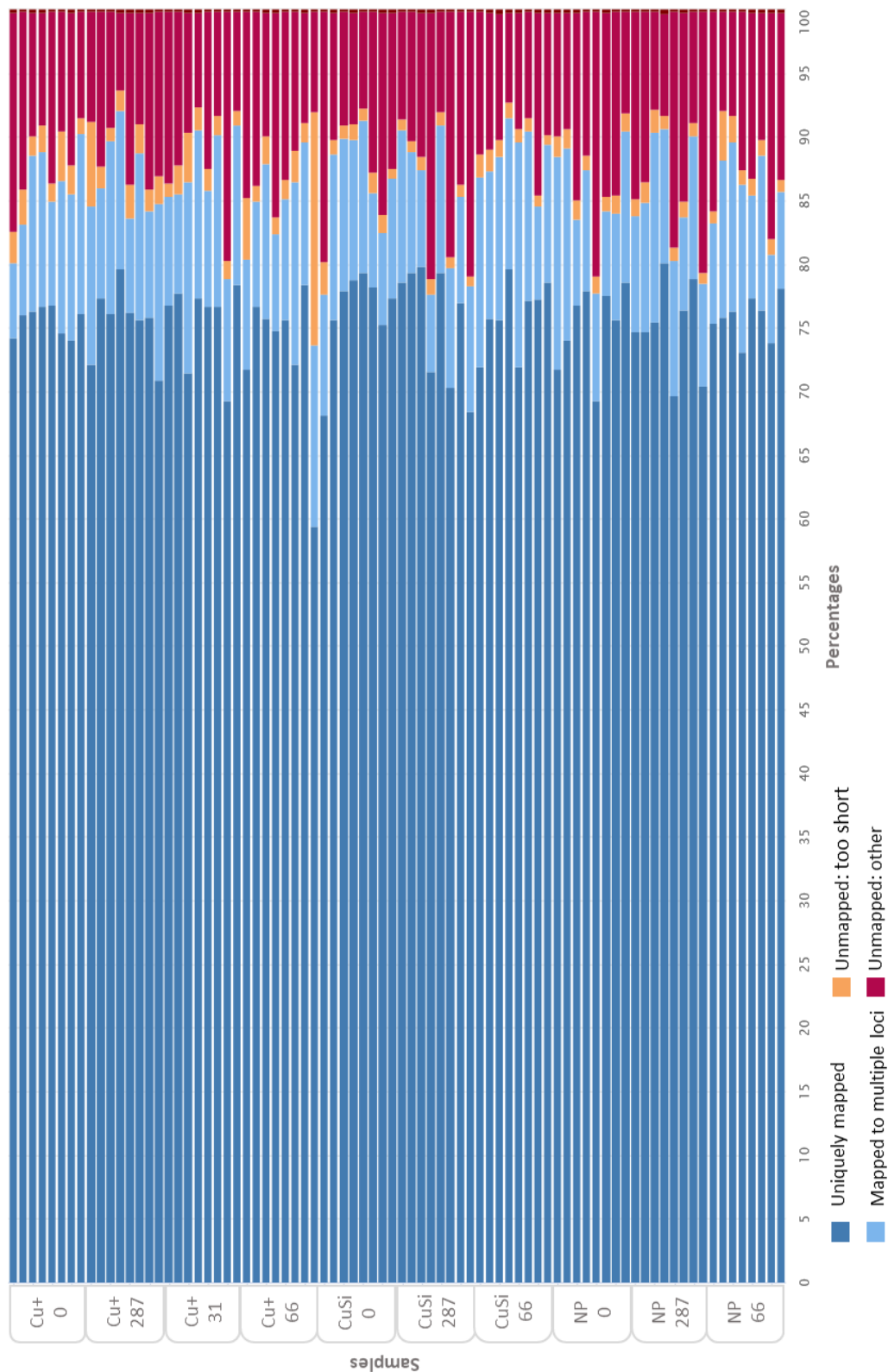


Figure 66: Mapping percentages of the sequenced samples based on statistics outputted by the aligner (STAR). The percentage of unique mapping was relatively even, indicating low individual genetic variation between the earthworms within the used in the experiment.

6.3.3 Exploration of holistic data

Initial analysis focused on comparing the different doses of the same compound exposure (Cu⁺, CuSi, CuNP) using the Sartools pipeline. Assessing the number of mapped reads associated with individual samples revealed low variation when comparing read counts within and between the different experimental conditions (Figure 67), with an average of 15 M reads associated with each library. Presumably the low amount of variation was the consequence of only a slight difference in rRNA content of the total RNA samples, or was due to negligible imbalance within the RNA concentrations of the sequenced libraries.

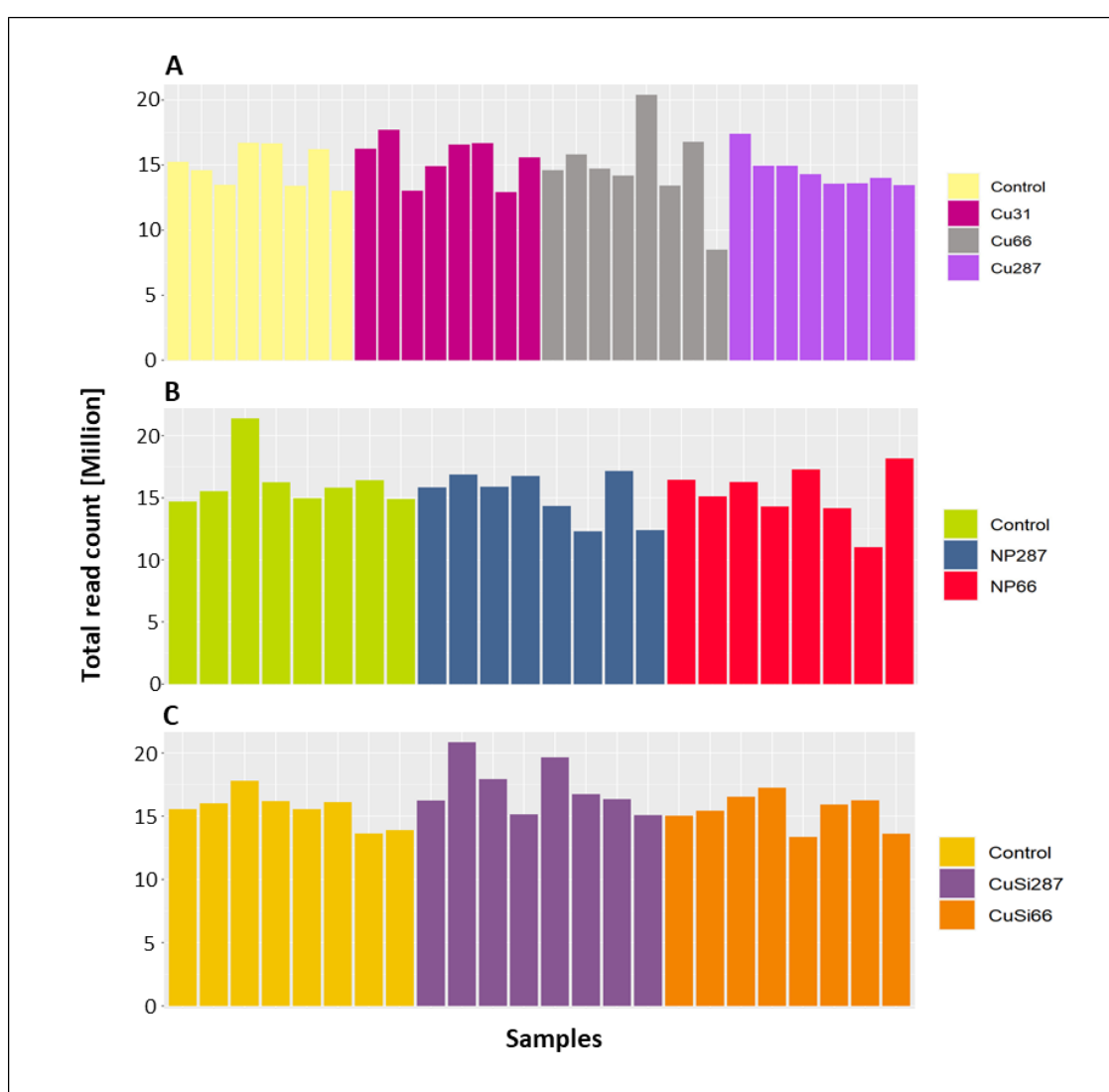


Figure 67: Total mapped read count used for DEG assessment in the case of CuCl (A), Cu-NP (B) and CuSi-NP exposures. The particle type and the exposure concentrations used are represented by different colours.

Features with null counts in the samples were identified in each sample and, although the features were left in the data, they were not included in the further differential expression analysis (DESeq2). The percentages of null read counts showed a similar pattern between the replicates and the different conditions. In the case of copper ion samples, null read counts accounted for around 13% of all the features while this number was 15% in the CuNP samples and 14% in CuSi samples (Figure 68).

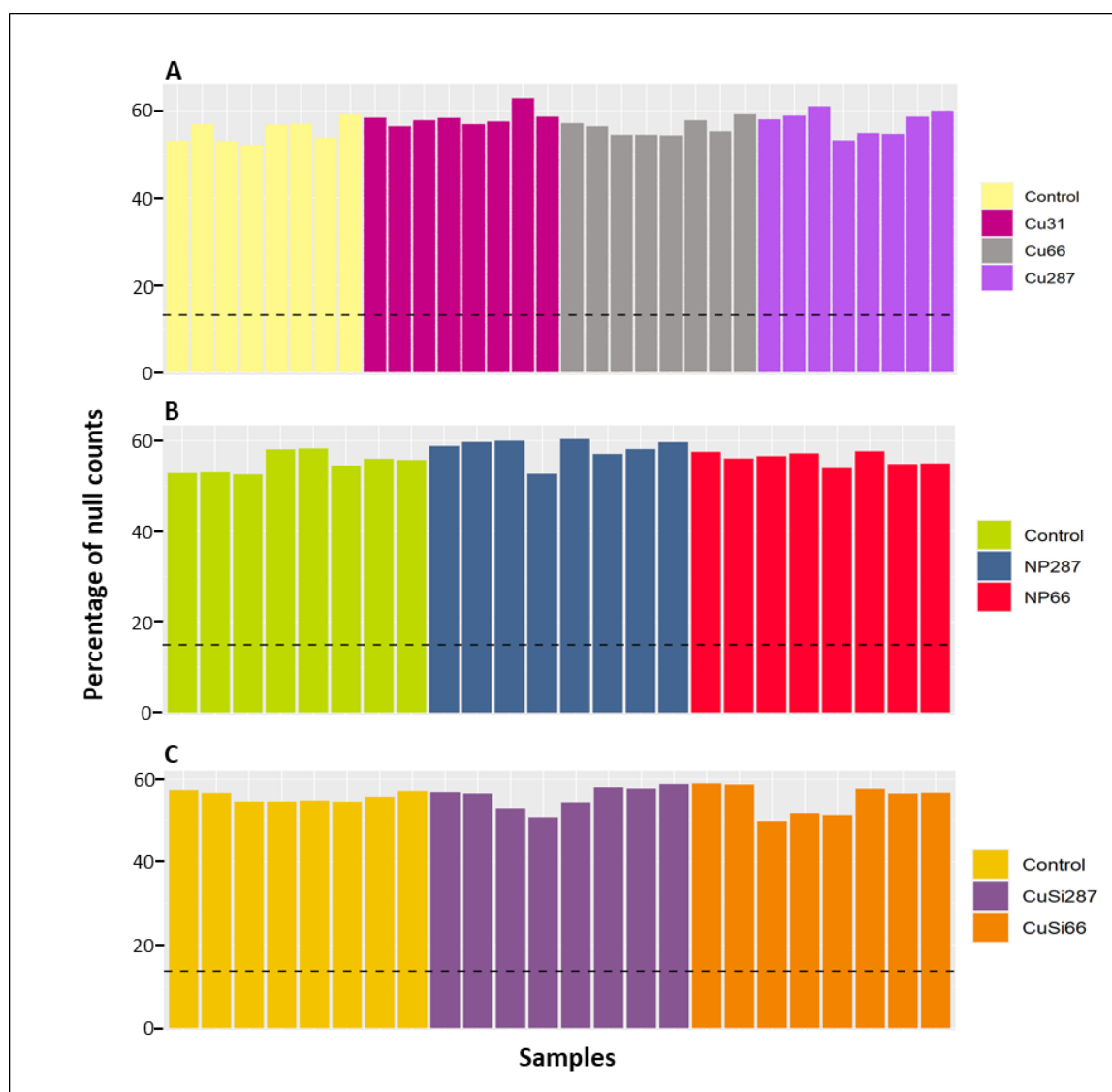


Figure 68: The proportion of annotated genome features with null counts across the different samples. Percentage of null counts show low variation within the different samples and experimental conditions with approximately 60% maximal value. Features with no read counts in all the 32/24 samples (dashed line) were treated as NA values and were excluded from later analysis. Panels (A) represent samples treated with Cu+, (B) shows Cu-NP treatments while (C) illustrates the null-counts associated with CuSi-NP samples

To enable comparison across different samples, in addition to accounting for systemic technical biases, counts were normalised prior to differential expression analysis. Read normalisation was conducted using the built in DESeq2 method with “lfcfunc” option set to “median”. The quality of the normalisation process is represented in Figure 69, which shows that count distribution was successfully stabilised across the different samples.

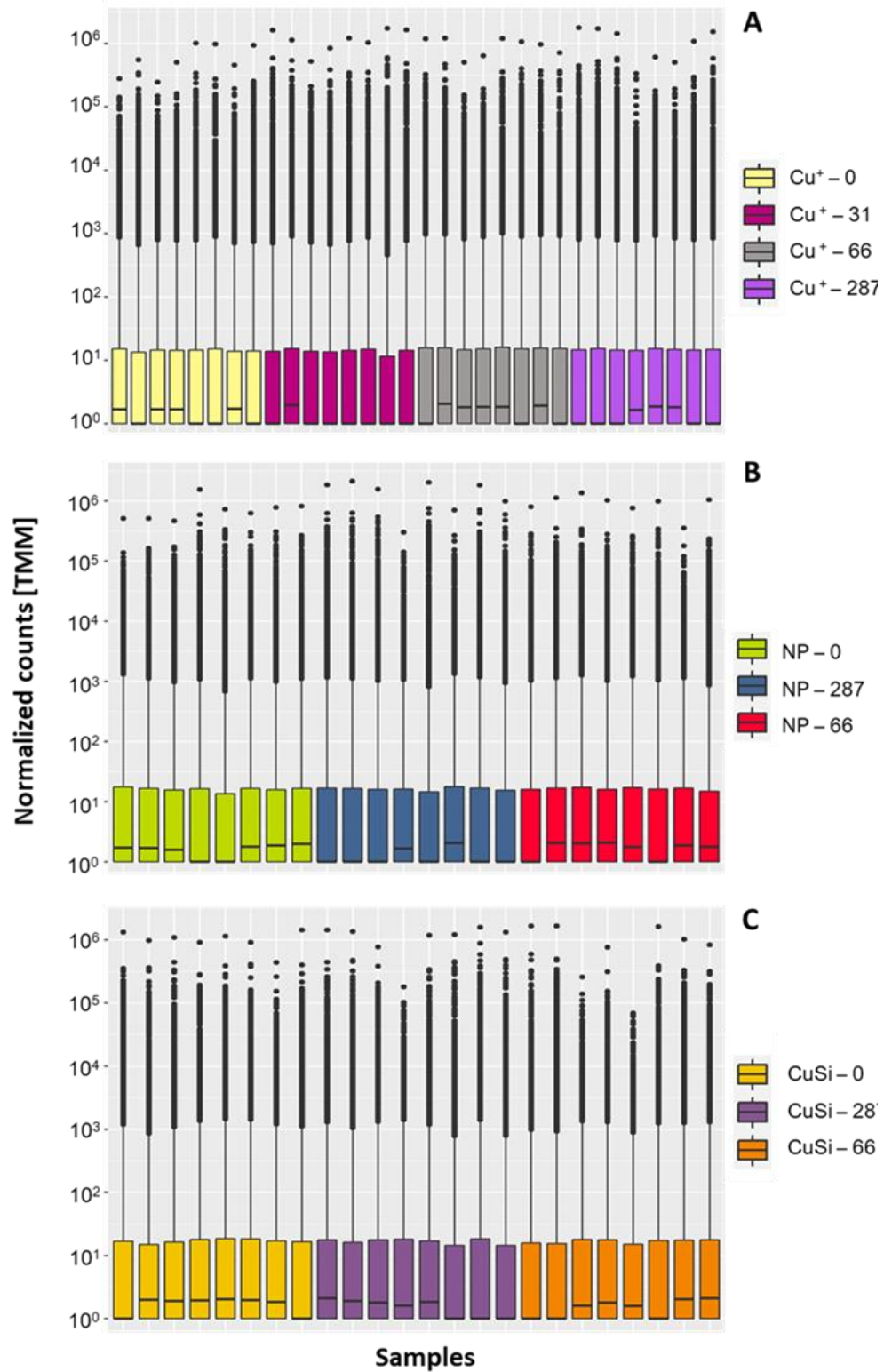


Figure 69: Distribution of normalised read-counts in the case of samples exposed to Cu+ (A), Cu-NP (B) and CuSi-NP (C).

Multi-variate analysis of the relationship between the expression profiles generated for each exposure group was explored using Principal Component Analysis conducted for all three treatments and different doses separately using the “PCAexplorer” (Marini and

Binder 2019) interactive R package. To identify the components associated with the biological variance between the control and exposed animals, components were manually selected for each PCA plot. In the case of high (Cu+287) and low (Cu+31) copper ion exposures, as well as the high copper nanoparticle exposure (CuNP-287), the best separation linked to exposure dose was observed by performing separations based on the first Principal Component (PC1). However, in the case of middle copper ion doses (Cu+66) samples were mainly separated using PC4 while low dose NPs (CuNP-66) could be only partially distinguish when PC2 was chosen. Significant separations are shown in Figure 70.

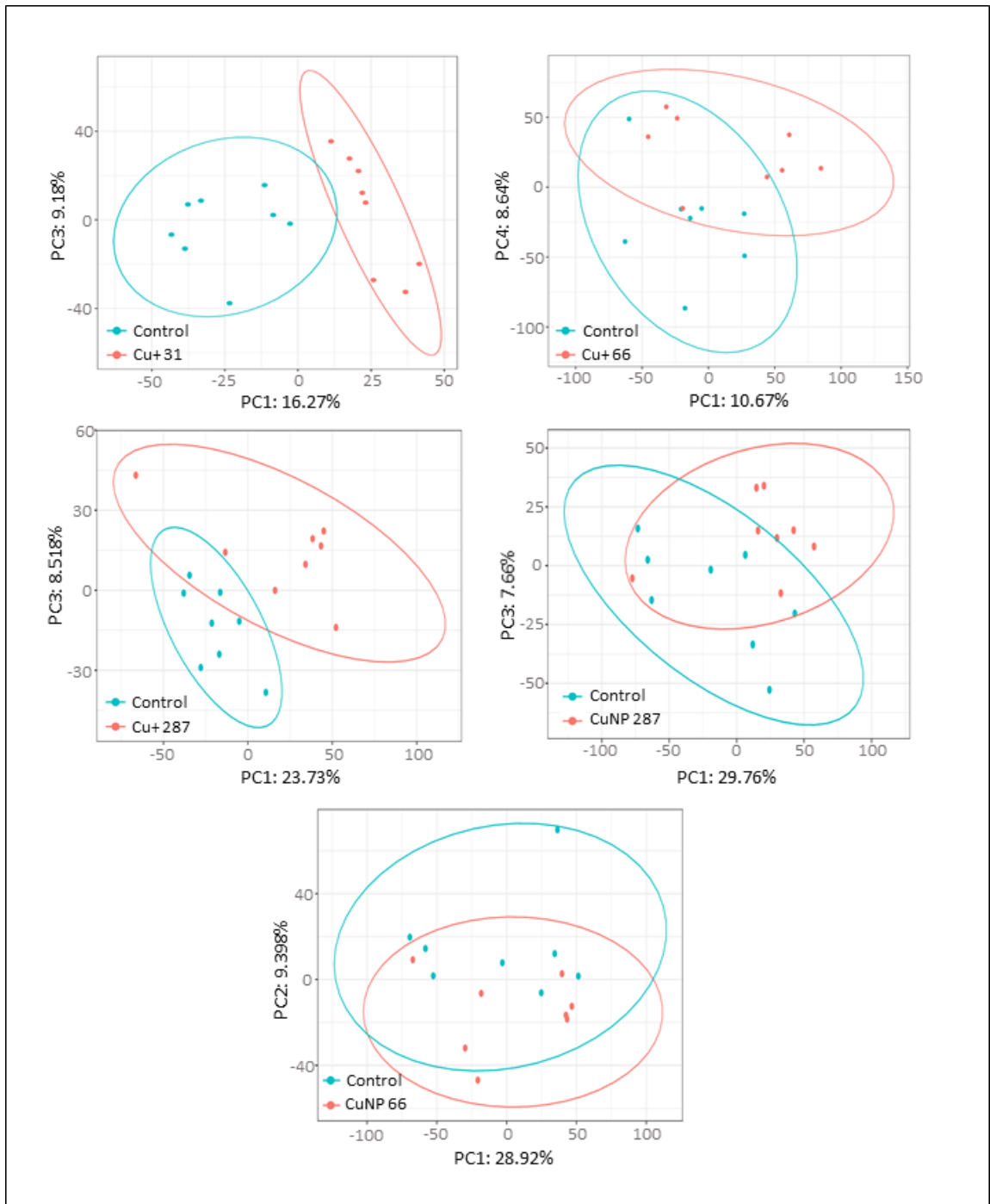


Figure 70: 2D Principle Component Analysis (PCA) of gene expression profiles from selected exposures against control organisms. Control (blue) and challenged (red) indicate the samples used for the different treatments selected. To identify variance between the samples, the 3000 genes with highest variance were analysed. The ellipses around the groups were drawn with a confidence interval of 0.95.

Although, when comparing control animals to exposed individuals, DESeq analysis was able to identify differentially expressed genes in the case of each treatment and dose,

the number of DEG showed high degree of variation. The highest number of DEGs were observed between control and the low copper ion dose (Cu+31), this was closely followed by the high copper ion exposure (Cu+287) (Table 6). However, both CuSi and low dose of CuNP exposure resulted in only marginal biological changes between the control and treatment groups, which was represented in the low number of detected differentially expressed features (Table 6).

Functional enrichment analysis (both Gene Ontology and Pathway database based methods) is a routinely used method to gain a broad overview about the exposure triggered changes in biological processes. In this case, functional enrichment analysis was conducted using the lists of up and down-regulated DEGs. The number of DEGs observed for CuSi-NP and low dose Cu-NP samples were too small to allow gene set enrichment analysis (GSEA) (Table 6).

Table 6: Number of Differential Expressed Genes (DEGs) and indication of analysis performed. Green colour indicated that analysis was performed while red means analysis was not conducted.

Treatment	Differentially expressed genes			Conducted analysis	
	Upregulation	Downregulation	Total	DGE analysis	Functional enrichment analysis
Cu+ 31	465	713	1,177	✓	✓
Cu+ 66	76	197	273	✓	✓
Cu+ 287	350	474	824	✓	✓
CuSi 66	7	17	24	✓	✗
CuSi 287	4	15	19	✓	✗
CuNP 66	4	12	16	✓	✗
CuNP 287	287	440	727	✓	✓

6.4 Overlaps in the differentially expressed gene profiles

To estimate the similarity of the biological effects caused by the different experimental conditions, the overlap between DEG were analysed both between applied soil concentrations for a single compound, as well as comparing different forms of copper exposure (select overlaps are shown in Figure 71. In the case of low (Cu+31), medium (Cu+66) and high (Cu+287) copper ion exposures only 56 common genes were identified, of which 6 showed differential direction of expression change. Interestingly, when comparing between low (Cu+31) and high (Cu+287) copper ion exposure, the majority of the DEGs observed were unique to a specific exposure condition with 474 only appearing in Cu+31 and 316 in Cu+287 whilst only 100 were shared with 17 of these being differentially regulated. This result suggests a very distinct response at the gene level to these two copper exposure doses. Genes from the middle concentration (Cu+66) showed high proportion of DEGs overlapping with both the Cu+31 (79 – 36%) and Cu+287 (45 – 21%) samples, suggesting this condition reflects characteristics of both the high and low exposures. When DEGs from the Cu+287, Cu+31, and Cu-NP287 were analysed, the venn diagram analysis showed that only 130 genes were shared between high ion (Cu+287) and high CuNP (CuNP-287) (27% of Cu-NP genes) conditions, with 80 shared between Cu+31 and Cu-NP287 – and over a third (36%) of this later overlap showing contrary changes to the expression direction. This suggests that high exposure to high dose CuNPs more closely aligns to the high dose ion exposure than it does the low dose. However, it is also true that a considerable number of specific genes are upregulated in CuNP and Cu ion exposed animals.

6.5 Gene Ontology enrichment analysis

Performing Gene Ontology (GO) based enrichment analysis using the low dose copper ion exposed animals identified several enriched processes using both the up and down-regulated genes. In general, most biological processes associated with down-regulated expression were linked to different transport mechanisms, such as “vesicle-mediated transport”, “cytosolic transport”, “transferrin transport” and “endosomal transport”, while a few innate immunity-related terms were also identified (“neutrophil degranulation”, “myeloid leukocyte activation”). Analysis of up-regulated genes within

the low concentration copper DEGs indicated a general association with metabolic processes, such as “small molecule metabolic processes” “alpha-amino acid metabolic processes” “carboxylic acid metabolic processes”, “vitamin D metabolic processes” while a few abiotic stress-related terms were also observed, such as “oxidation-reduction processes”, “response to toxic substance”, “response to chemical” (Figure 72).

Differentially expressed genes from the high dose ionic copper exposure showed down-regulation for a large number of cell cycle and DNA replication related biological processes; including: ‘cell cycle DNA replication’, “nuclear DNA replication”, “nucleic acid metabolic processes” and “mitotic cell cycle phase transition”. Although, in the case of up-regulated genes, terms related to angiogenesis, including: “negative regulation of blood vessel morphogenesis”, tissue development”, “blood circularisation” and “tissue development” could be identified, the “regulation of iron transport” and dendrite regeneration were also impacted (Figure 73).

When analysing Gene Ontology-based Biological processes in the case of Cu-NP treatment (Figure 74), the list of up-regulated terms showed a large overlap with the above described high dose copper ion effect. However, the GO analysis of the CuNP treatment resulted in terms associated with immune response, including the appearance of “antigen processing and presentation” and “antigen processing and presentation of exogenous peptide antigen”, which could not be identified in the corresponding response to high dose copper ion exposure. In contrast to the similar overall effect in down-regulated genes, the up-regulated genes originating from high Cu-NP (CuNP-287) treatment resulted in a very distinct functional enrichment profile compared to high copper ion exposure (Cu+287 samples). In this case, terms with highest significance were organised around muscle tissue related biological processes, including “muscle system processes”, “muscle contraction”, “muscle organ development” and “actin-filament based movement”. Furthermore, a few immune-related biological processes, such as the “antimicrobial humoral response” and “negative regulation of NK cell differentiation involved in immune response” were also identified.

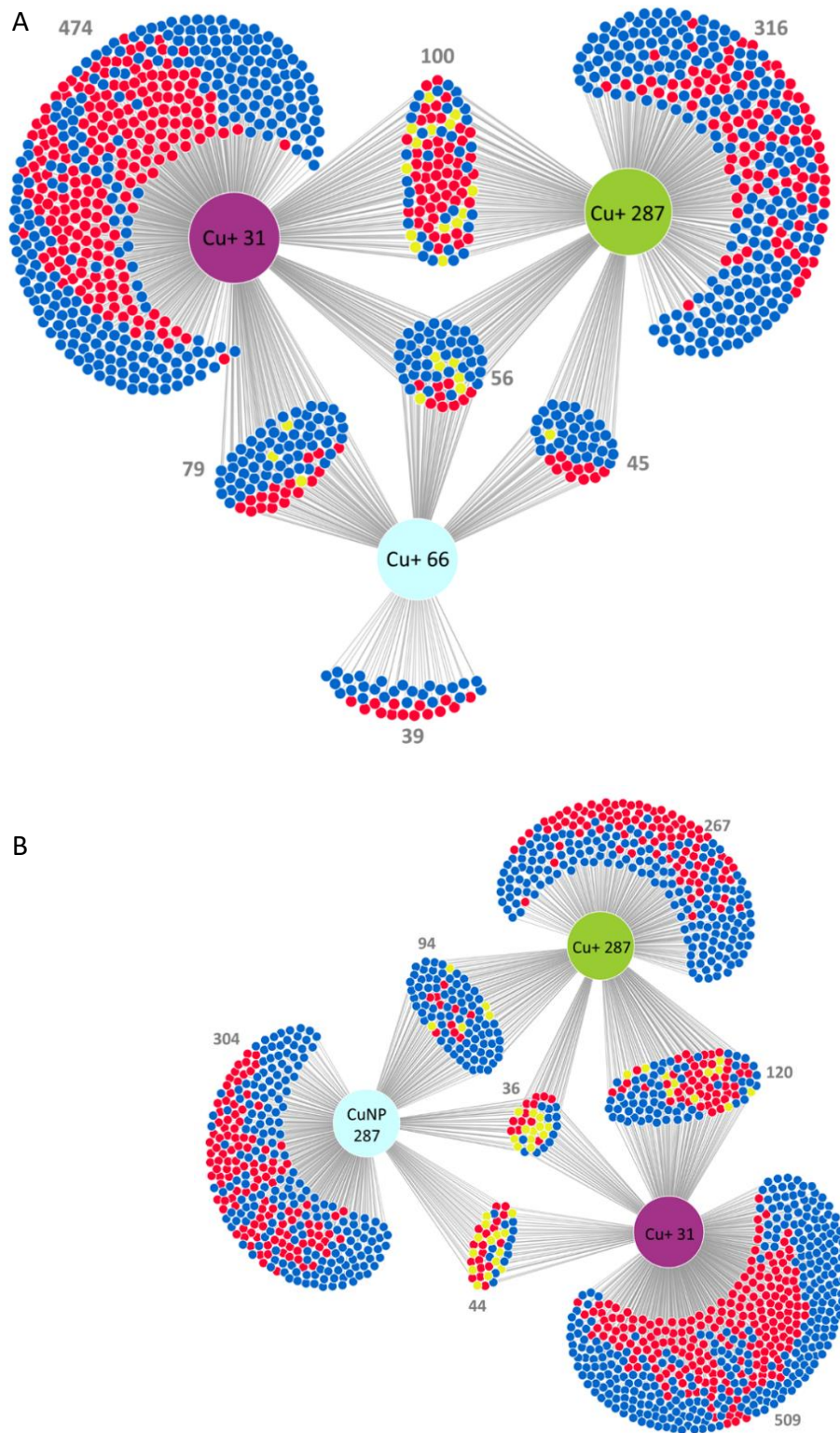


Figure 71: Overlap of differentially expressed genes in *E. fetida* exposed to different concentrations of copper-ion (A) and the between the low and high copper-ion and high dose copper NPs exposure (B). Red and blue colours represent consistent up, down-regulation respectively whilst yellow indicates differential regulation between the conditions being compared.

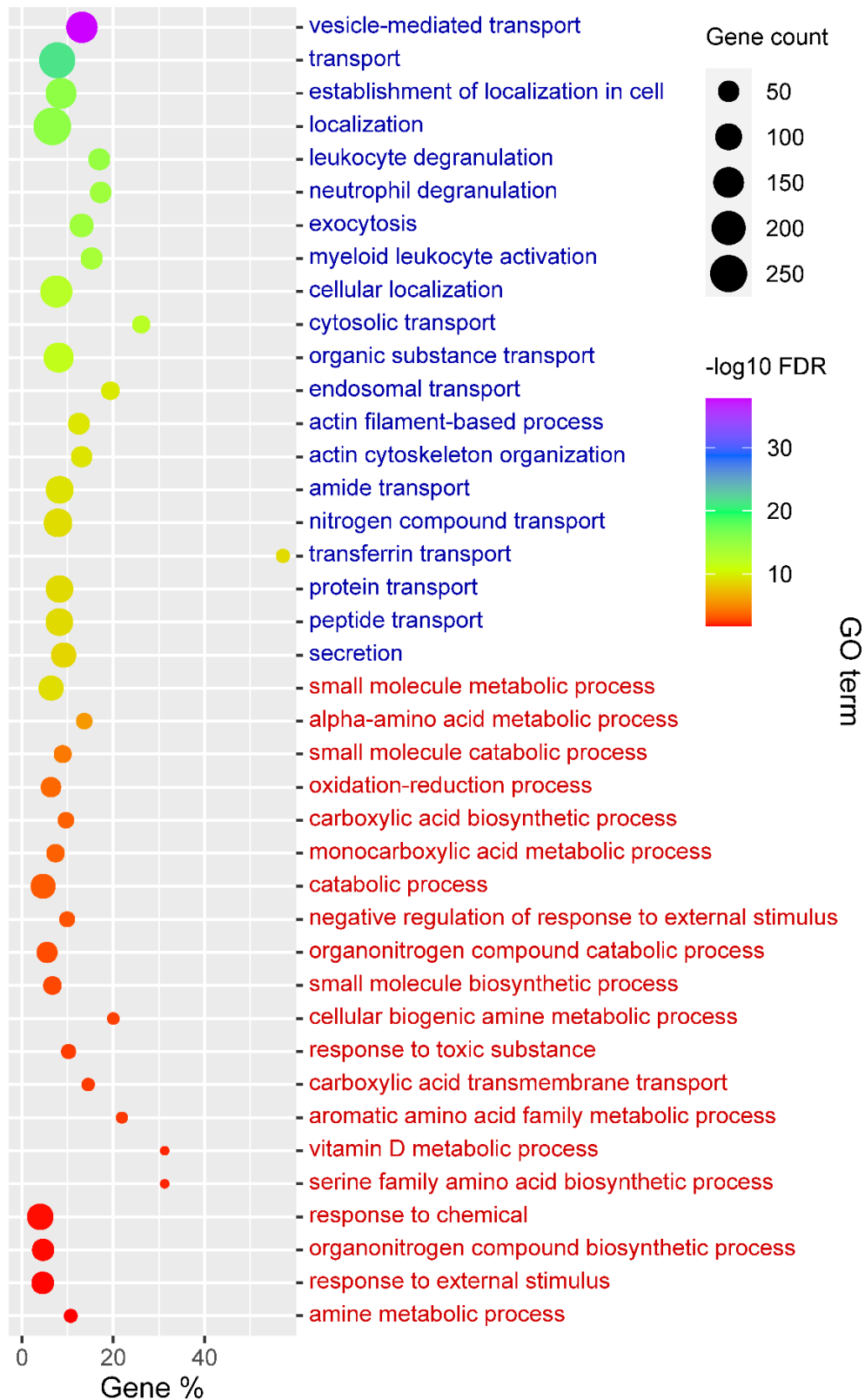


Figure 72: Dot plot showing the top 20 significantly enriched Gene Ontology (GO) terms, in the case of the Cu+31 (low-dose) treatment. GO terms in blue and red represents the analysis of down and up regulated genes respectively. The colour of the dots shows the level of significance based on Benjamini-Hochberg corrected p-values (FDR), while sizes represent the number of genes associated with the specific terms.

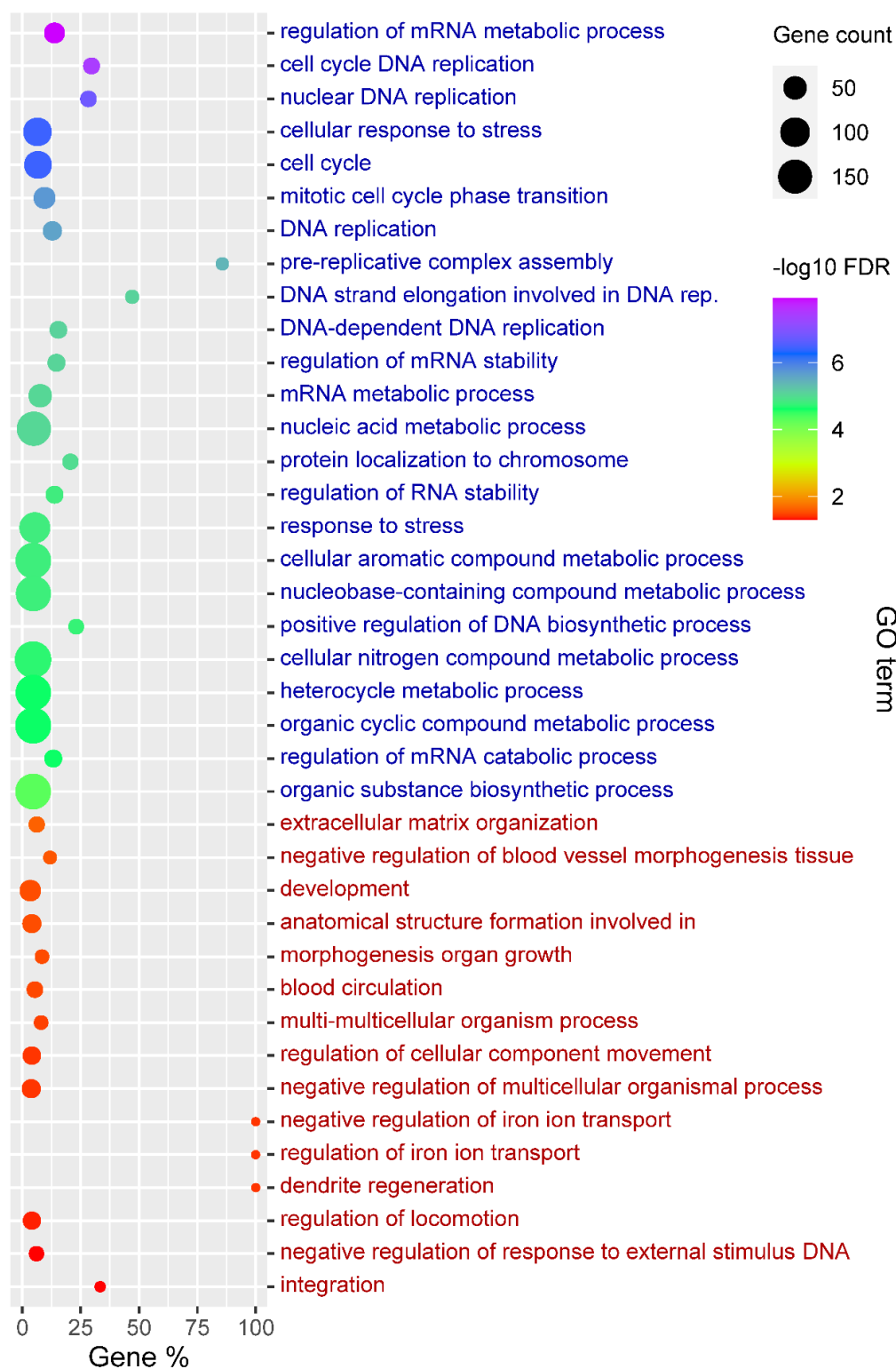


Figure 73: Dot plot showing the top 20 significantly enriched Gene Ontology (GO) terms, in the case of the Cu+287 (high-dose) treatment. GO terms in blue and red represents the analysis of down and up regulated genes respectively. The colour of the dots shows the level of significance based on Benjamini-Hochberg corrected p-values (FDR), while sizes represent the number of genes associated with the specific terms.

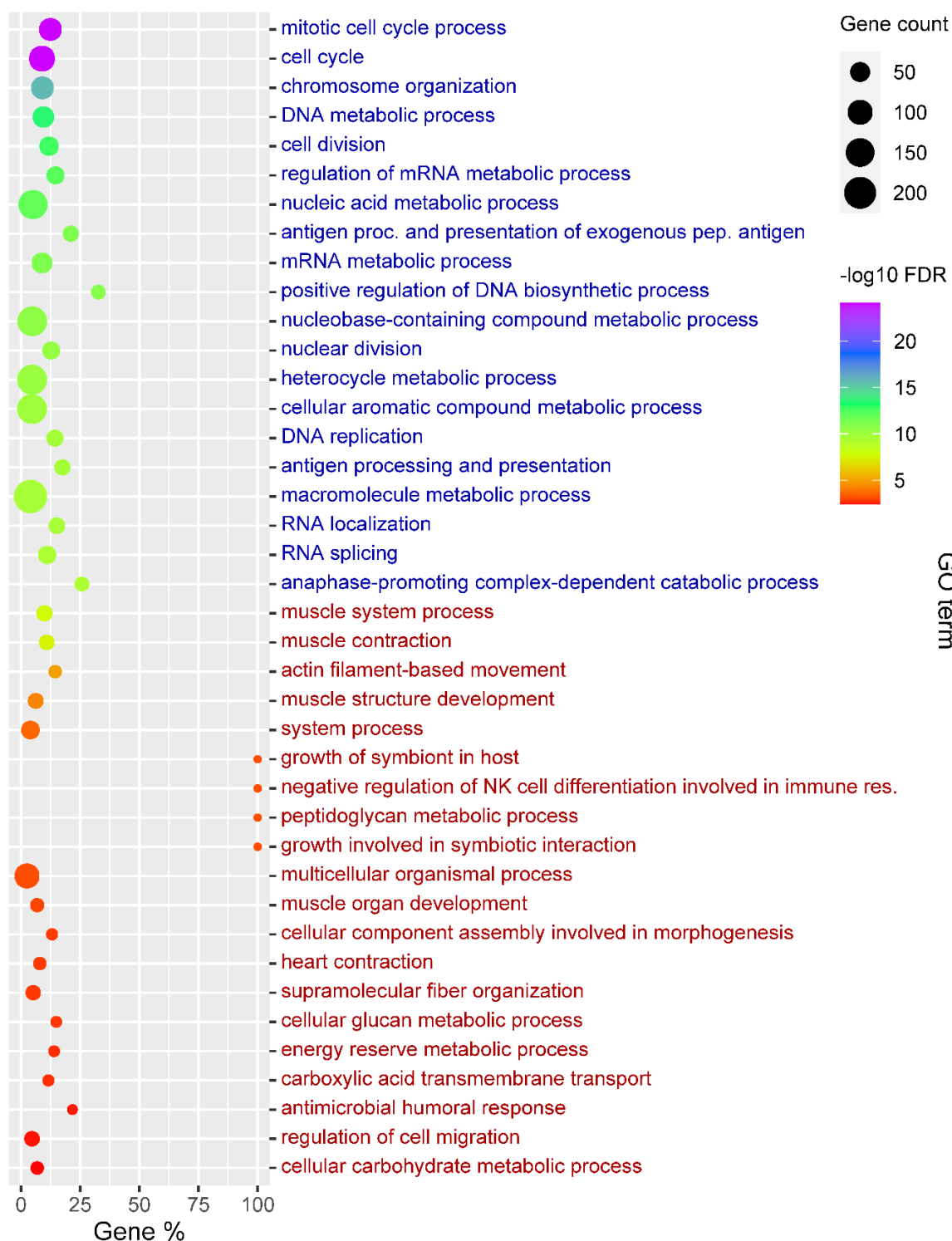


Figure 74: Dot plot showing the top 20 significantly enriched Gene Ontology (GO) terms, in the case of the Cu-NP287 (high-dose Cu-NP) treatment. GO terms in blue and red represents the analysis of down and up regulated genes respectively. The colour of the dots shows the level of significance based on Benjamini-Hochberg corrected p-values (FDR), while sizes represent the number of genes associated with the specific terms.

6.6 Pathway enrichment analysis

To gain a more detailed picture of the biological impact of the different experimental conditions, a pathway enrichment analysis was conducted based on the KEGG and Reactome databases using the ClueGO software (Bindea et al. 2009). In general, pathway enrichment results showed a large overlap with the GO based over-representation analysis.

In the case of the low copper ion (Cu+31) treatment, the most significant pathways were organised around vesicular transport and membrane trafficking, where “Clathrin-mediated endocytosis” appeared to be highly affected. Although in the case of Cu+31 treatment, several pathways associated with oxidative stress could be identified, such as “ROS, RNS production in phagocytes” and “Glutathione conjugation”, DNA damage-related pathways were not observed in either Cu+31 or Cu+66 samples. Innate immune system, or the more generic term ‘Immune system’ represented the immune-related GO terms and these were observed with high significance following analysis of both Cu+31 and Cu+66 DEGs (Figure 75).

Similar to the GO enrichment, high dose copper ion and CuNPs displayed a large number of overlapping terms in their enriched pathways. Whilst contrasting results were observed for the lower concentration copper exposures. In the case of both Cu+287 and CuNP-287, pathways associated with cell cycle, ‘DNA damage’ and DNA repair appeared to be highly abundant. Despite these similarities, a few CuNP exclusive pathways associated with muscle tissue, including: “Muscle contraction” and “Strained muscle contraction” together with Immune system processes (“MHC class II antigen presentation”) could be identified (Figure 76).

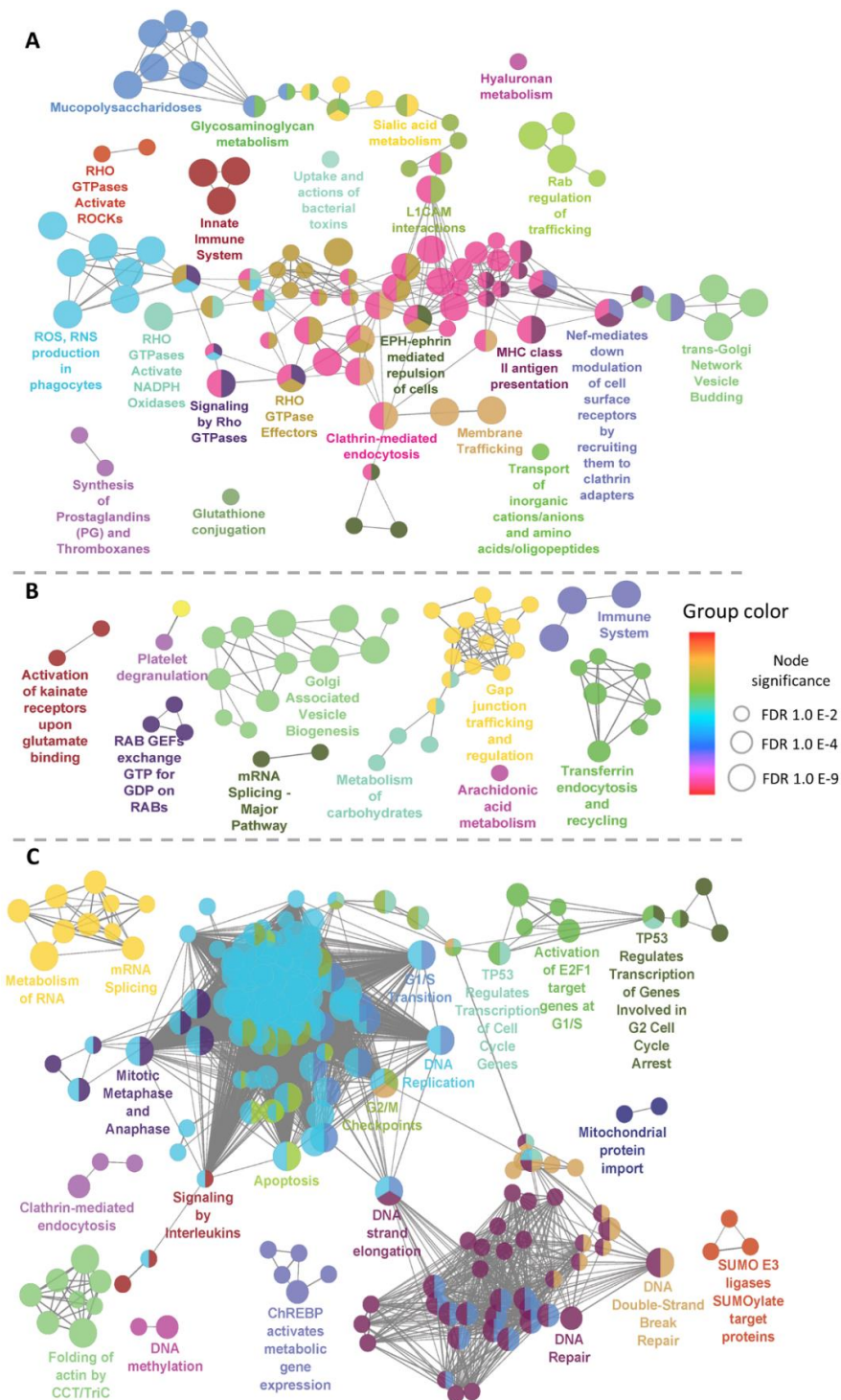


Figure 75: Pathway enrichment network generated by ClueGO analysis using DEGs from Cu+31 (A), Cu+66 (B) and Cu+287 (C) experimental conditions. The size of the circles represents the level of significance based on Benjamini-Hochberg adjusted p -values. Nodes were coloured according to their group identity, based on related functions. From terms associated with the same group, only the pathway with the highest significance is labelled. To reduce redundancy, pathways were functionally grouped (colours) based on gene membership scores (κ score) where only terms with the highest significance are shown. The edges represent the statistical association (overlaps) between the enriched terms (Bindea et al. 2009).

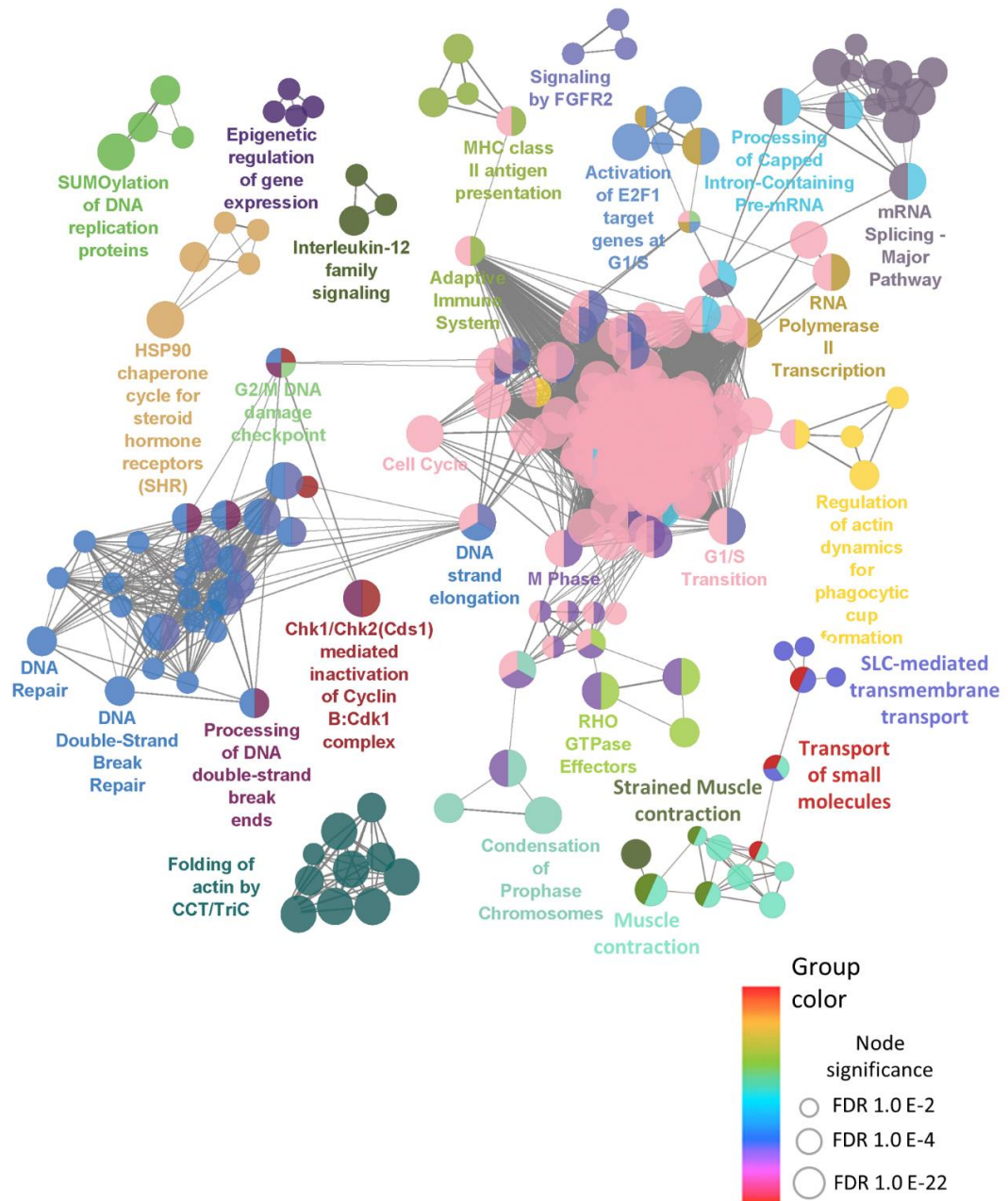


Figure 76: Pathway enrichment network generated by ClueGO analysis using DEGs from the non-coated Cu-NP287 experimental condition. The size of the circles represents the level of significance based on Benjamini-Hochberg adjusted p -values. Nodes were coloured according their group identity, based on related function. Pathways were functionally grouped (colours) based on gene membership scores (kappa score) where only terms with the highest significance are shown. The edges represent the statistical association (overlaps) between the enriched terms (Bindea et al. 2009).

6.7 Discussion

6.7.1 Differential gene expression analysis

Principle component analysis (PCA) provides insight into the overall variability between control and treated samples. For many of the PCA1 and PCA2, which represented the dimensions with the highest variability, separation within these dimensions was not correlated with the experimental treatment. Although on many occasions biological conditions show separation on PCA1 and PCA2, it has been pointed out by Nguyen and Holmes (Nguyen and Holmes 2019) that, in some cases, the variation of interest is only captured in higher-order PCs (i.e. PCA3, PCA4, PCA5). Manual refinement of PCs allows the identification of dimension that best illustrated separation in 2-dimensional space. The number of differentially expressed genes that displayed a clear correlation with the amount of variance, accounting for the variation between replicates represented by the intersection of the 95% confidence intervals in the various PC groups. In the case of the analysis of the Cu⁺ and Cu-NP287 samples, control and treatment conditions showed a clear separation, resulting in the highest number of DEGs. However, in the case of all CuSi-NP and CuNP-66 samples, only a relatively low portion of variation was detected between the control and treatment conditions and this manifested in the low number of identified DEGs between these conditions.

6.7.2 Bioavailability of copper ions, released from NPs

During the last few years, ecotoxicological studies have shown that due to their highly reactive nature, many metal nanoparticles tend to undergo transformations (aggregation, dissolution, oxidation, reduction) as soon as they come in contact with soil (Keller et al. 2017, Sekine et al. 2017). Based on the results of studies investigating the environmental fate of these particles, it seems that the observed toxic effect reflects uptake and accumulation of the released metal ions that results from the dissolution of the NPs (van den Brink et al. 2019, Lee et al. 2018).

The numerical differences in DEGs observed when comparing the different treatments used in the present study suggests the silica encapsulated nanoparticles produce only a mild impact on the earthworms relative to the ionic copper or even non-coated CuNPs. A possible explanation for the decreased impact of the silica encapsulated nanoparticles

is that encapsulation may greatly reduce the rate of copper core dissolution, thus limiting the bioavailability of the nanoparticle's copper component. This is in agreement with previous studies that have described a reduced toxic effect for silica coated NPs relative to their uncoated forms (Liu and Han 2010, Shiomi et al. 2015) .

6.7.3 Gene ontology and pathway enrichment analysis

Although, copper is an essential trace element for most terrestrial invertebrate organism and specific biological mechanisms can often benefit from excess bioavailable copper, in high soil concentrations copper represents one of the most toxic heavy metals for small invertebrate species (Gerhardt, 1993 (Johnson et al. 2017)). This duality of the copper dose dependent biological effect on earthworms was easily recognisable when comparing the results of the enrichment analysis of earthworms exposed to the low copper ion concentration with those exposed to higher copper soil content. The results of the gene enrichment studies suggested a biphasic dose response between the utilised experimental soil concentrations. In the case of low concentrations of copper ion, only a moderate evidence for oxidative stress was observed, presumably the consequence of a hormetic effect of the excess copper content of the soil (Tyne et al. 2015). The increased copper challenge of the cells was underpinned by the observed large impact on the vesicle-mediated endocytic pathways, which presumably was a consequence of the copper transported (CTR1) internalisation and recycling processes (Clifford et al. 2016), which likely represents an attempt to decrease the extracellular copper intake by lowering the rate of copper entry to cells (Petris et al. 2003). Although increased oxidative stress was evident in this exposure (Cu+31), signs of cytotoxicity or DNA damage could not be identified. This, alongside the observed upregulation in metabolic processes, seems to also support the hypothesis of the beneficial, hormetic, biological response caused by the slightly toxic effect of low dose copper (Calabrese and Baldwin 2000, Stebbing 2002).

In contrast to the hormetic effect described above, both the high concentration of ionic copper and non-coated copper NPs triggered a cytotoxic response with increased DNA damage and DNA repair, identified in the results of both the GO (Biological Processes) and the pathway enrichment. In this case, likely as a consequence of the copper induced Fenton and redox reactions (Sutton and Winterbourn 1989), the level of oxidative stress

appears to have reached the point where evidence of apoptotic processes are present (Li et al. 2019, Lloyd and Phillips 1999). The high abundance of the cell-cycle related pathways and biological processes at these exposures suggests the accumulation of copper-related genotoxic stress has resulted in sufficient DNA damage to induce cell cycle arrest (Mitra et al. 2012). In addition, the appearance of DNA damage checkpoint related terms in both the high concentration copper ion and non-coated nanoparticle exposed worms suggests the extent of DNA damage is overwhelming repair mechanisms, leading to cell cycle checkpoint mechanisms initiating apoptotic pathways to eliminate damaged cells (Ishikawa et al. 2006). The key characteristics of the hormetic effect of the copper ion exposure are shown of Figure 77.

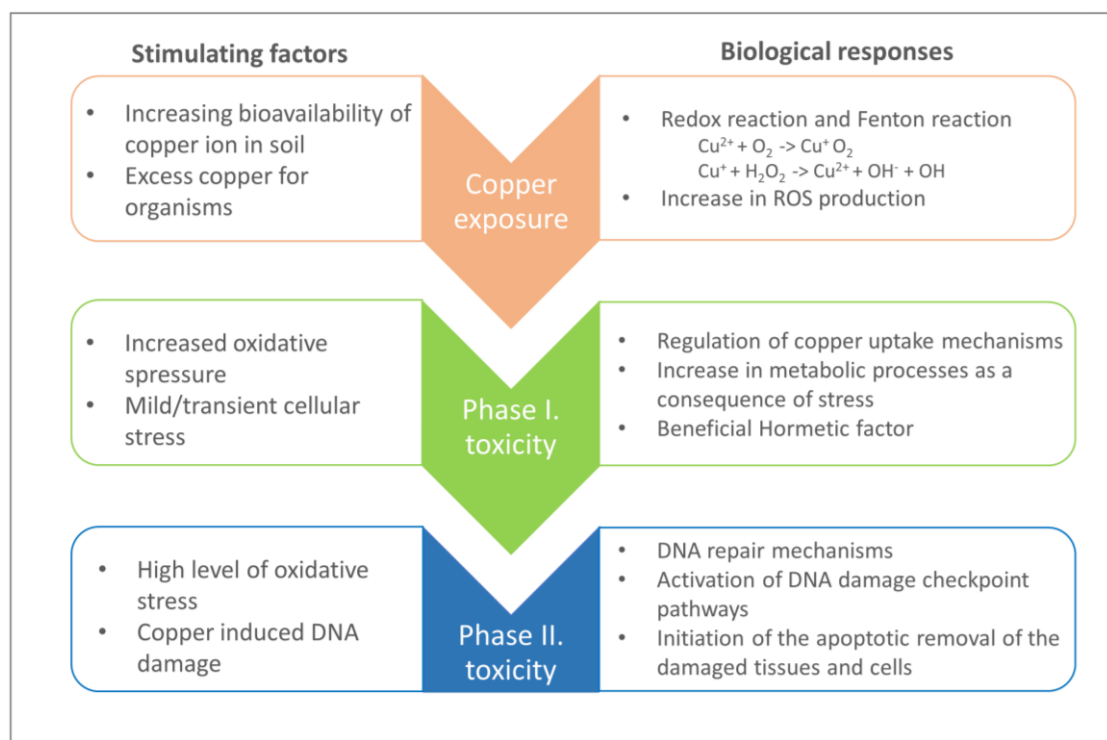


Figure 77: Flowchart summarising the main characteristics of observed hormetic effect of the copper ion exposure.

6.7.4 Shared effects in the high concentration of Cu+ and Cu-NP exposures

In total more than 800 DEGs were identified when analysing the high concentration Cu+ treatment, while this number was approximately the same (727) when analysing the high concentration Cu-NPs condition. The identified overlap between the lists of DEGs accounted for only approximately 18% of the number of DEGs which appeared within

the two conditions. A much higher overlap was identified when analysing the overlap in significantly enriched biological processes and pathways rather than the lists of differentially expressed genes (Figure 78). Relative to controls, the enriched terms present in earthworms exposed to high ionic copper and CuNPs appear to be related to the same important copper homeostasis, transport and toxic effect associated biological pathways. Despite the high degree of similarities, a few differences could be identified between the copper ion and CuNP exposures. Specifically, in the case of high Cu⁺ exposure, an angiogenesis suppressor effect was observed in the list of up-regulated genes when compared to Gene Ontology database. The excess copper induced anti-angiogenic effect has been described in higher order vertebrates and it is clear that copper acts as an obligatory co-factor for angiogenesis, although the exact mechanisms are still unknown in the case of earthworms (Urso and Maffia 2015, Finney et al. 2009, Lee et al. 2020). In the case of the Cu-NP287 treatment, a unique NP specific effect was identified, namely the upregulation of muscle tissue development and related biological processes. This finding seems to be supported by a recent publication which observed muscle toxicity in adult zebrafish as a consequence of chronic exposure to CuNPs (Mani et al. 2020).

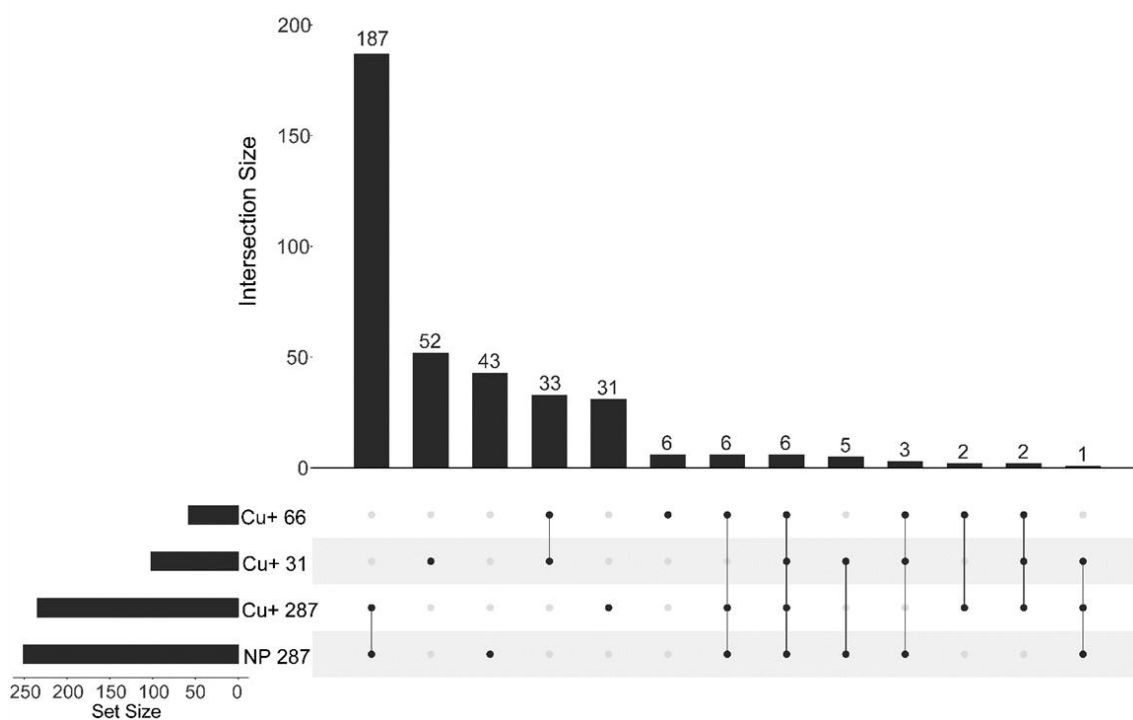


Figure 78: Shared KEGG and Reactome pathways identified by enrichment analysis of treatments with Copper ions and high dose CuNPs. A high overlap (~40%) can be observed between exposure to 287 mg Cu/kg CuNP and the equivalent dose of Cu ions.

6.7.5 Conclusion

In this chapter we compared the effect of copper ion with copper nanoparticles both delivered in a pure metallic form or encapsulated in silica. In the case of CuSiNPs both the identified total variance between the control and treatment condition and the low number of identified DEGs suggested that the biological impact of the particles was minimal, presumably due to the slow dissociation caused by the silica coating. In the case of applied different Cu⁺ concentrations, a two-phase toxic effect could be observed where the low concentration provided a low-level oxidative stress thus resulting in a hormetic effect. The most affected biological pathways were vesicular transport and increased metabolic processes. Opposite to Cu⁺31 the high concentration of Cu⁺ exposure produced a real toxic effect, where increased DNA damage and programmed cell death could be observed as the consequence of the high oxidative stress. To be able to observe the possible differences between the ionic and nanoparticle copper DEG results of the two conditions were compared on the level of both DEGs and biological processes, which showed high similarity between the major impacted pathways with only minor differences.

7 Final discussion

7.1 Background of the study

Recent innovations in the field of nanotechnology has exponentially grown the portfolio of available nanomaterials (Ogunsona et al. 2020, D'Mello et al. 2017), resulting in a substantial increase in their use and disposal into the environment. Despite the increasing number of studies targeting the possible biological risks associated with nanomaterials, their effects on the ecosystem are only modestly understood. A growing body of evidence has shown nanoparticles interact with different components of the innate immune system (Boraschi et al. 2017, Liu et al. 2017) and the relationship between nanomaterials and immune response has become a key factor when assessing the safety of nanomaterials (Pallardy et al. 2017). The aim of this study was to gain a better understanding of the impacts of engineered nanoparticles on the innate immunity of earthworms (*E. fetida*). However, to achieve this goal, the project had to first address the dearth of information about the cellular and molecular components of the earthworm immune system.

7.2 Challenges

Although earthworms are widely used as ecotoxicological sentinel species, their immune system is only modestly characterised. The available molecular biology resources associated with earthworm immunity is limited relative to other ecotoxicologically relevant species, with only a handful of immune-related genes having been characterised. Understanding how different nanomaterials interact and modulate immune pathways can be highly challenging, even when studying species with a well-defined immune system that have characterised cellular and humoral mechanisms. Therefore, the lack of detailed molecular information on earthworm immunity severely limits our capacity to investigate the innate immune responses to any challenge, including those posed by engineered nanomaterials. To overcome the issues associated with the limited resources available for investigating earthworm immunity, we focused the project around three major aims: 1) Identification of the key components (both well conserved and taxa specific) of the *E. fetida* innate immune system. For this we use high-throughput techniques, including transcriptomics and genomics, to build putative pathways that have the potential to be impacted by nanoparticles. 2) Profiling the

spatial (cell-specific) and temporal system immunity response under bacterial challenge both in the presence and absence of copper oxide nanoparticles (CuNPs). This would deliver a spatially and temporally resolved framework for the interaction between CuNPs and the earthworm innate immune system. 3) Finally, the project investigated the effect of CuNPs on the earthworm immune system via a soil-based exposure. This was done to observe the possible differences between the direct NP effect and a more ecotoxicologically realistic scenario. The CuNP injection results were compared to those of the soil-based exposure to identify immune genes and pathways modulated by both exposures.

7.3 Summary and brief interpretation of key findings

7.3.1 Genomic and transcriptomic framework of earthworm immunity

First, a high-quality tissue-specific transcriptome atlas was successfully established for the two closely related *Eisenia species*, *E. fetida* (NCBI accession ID: PRJNA608692) and *E. andrei* (NCBI accession ID: PRJNA624103). In total, six different tissues were selected for analysis (pharynx, crop, gizzard, gut/chloragogen, ventral nerve cord, coelomic fluid), each of which represented by a similar sequencing depth. Using only a low number of individuals from a relatively inbred population and sequencing the selected tissues with similar sequencing coverage has resulted in a highly complete and contiguous *de novo* *E. fetida* and *E. andrei* reference transcriptome. This data will have wide spread applications beyond the study of CuNPs and immune system biology. Indeed, it has already been used in an unrelated study investigating the basis of earthworm sensitivity to pesticides (Short 2021). Utilising the generated transcriptomes allowed the innate immune system of two earthworm species to be characterised at a high level of detail. In total more than 3,900 putative immune system related transcripts were annotated in both earthworm species, from which more than 1,860 annotation were assigned to at least one of the major Innate Immune pathways. The tissue-specific dataset also provided the first overview of the tissue expression profile of these immune system genes, giving insight into the immunological importance of the individual tissues. Despite the tissue-specific reference transcriptomes providing an excellent resource, the limitations of the *de novo* assembly pipeline became clear when analysing the numbers and the continuity of retrieved transcripts annotating as earthworm Toll-like

receptors (TLRs). While the resultant tissue-specific transcriptomes provided a great baseline resource for both earthworm immunology or any other molecular biology related study targeting *E. fetida*, its limitations indicated the necessity for a high-quality genomic template.

Therefore, to support the tissue-specific reference transcriptome data, a new highly contiguous and complete reference genome was established for *E. fetida*, this contained only 7,208 scaffolds with an overall N50 of 1.1 Mb and a BUSCO completeness score of 95%. Following the gene prediction, approximately 130,000 gene objects were identified across the reference genome, from which more than 38,000 were annotated by homology to provide putative functional assignment. The generation of an annotated genome allowed us to refine the *E. fetida* reference transcriptome and increase its contiguity. The new genome-based reference transcriptome resulted in a significantly higher number of complete transcripts than was observed with the tissue-specific *de novo* assemblies. Based on the annotated reference genome, a total of 39 full-length Toll-like receptor were successfully identified, allowing functional domain prediction and phylogenetic categorisation. Combining the tissue-specific resources with the newly identified TLR genes, we were able to determine their tissue-specific expression profiles to provide the foundation for both this work and future functional studies.

The mixed populations of coelomocytes cells are thought to play critical roles in earthworm immunity (Stein et al. 1977). However, our understanding of cell type specific immune functions had been mostly derived using histochemical observations and the genomic and transcriptomic knowledge was markedly lacking. The coelomocyte-specific transcriptome, generated in Chapter 4, delivered high detailed transcriptomic fingerprints for the three major populations of free-floating coelomocytes, a first for earthworm immunology. This data gave us better understanding of the functions associated with the different coelomocyte populations allowing us to trace the immune pathways affected by NPs. Analysis of this data permitted the characterization of the cell-specific expression profiles of several important immune and defence-related molecules. The benefits of determining cell specificity were notably illustrated when considering the TLR family, which displayed a strong cell-type dependent expression pattern. The identification of the 'cell-type

specific' TLR expression profile opens the possibility for future studies to characterise the possible pathogen specific recognition processes in the different coelomocyte populations.

7.3.2 Spatio-Temporal characterisation of the direct NP effect

RNAseq was used to study the spatio-temporal response of the earthworm immune system when challenged with *B. subtilis*, both in the presence and absence of previous copper-oxide nanoparticle (CuNPs) exposure. To resolve the spatial aspects of the cellular responses, the major coelomocyte populations were separated by cell sorting, enabling cell-type specific profiling, while temporal information was provided by multi-sampling throughout an extended time course encompassing exposure through to steady-state recovery. The experimental design required low mortality throughout the extended exposure period but, unfortunately, the sublethal dermal dose of *B. subtilis* required to maintain viability did not illicit significant differential gene expression. However, the pre-injection of CuNPs resulted in significant gene expression changes in the case of all the three coelomocyte populations. Differentially expressed genes from each cell-type were clustered based on their temporal profiles, which allowed the characterisation of the early, middle, and late cellular responses in a cell type-specific manner. It was intriguing to see that the early-response in granular amoebocytes appeared to be organised around interleukin-1-signalling, while TLR signalling was acutely activated in eleocytes. In contrast, only an oxidative stress associated glutathione response was observed in the hyaline amoebocytes. This was followed by a mid-response organised around RNA processing and metabolism-related processes in all the three cell populations. Although during the late response, the enrichment analysis produced more generic terms, eleocytes showed over-representation for "Glutathione biosynthetic processes", which suggests that the oxidative stress caused by the CuNPs was still present after 120 h post-injection. To understand how CuNPs effect earthworm copper homeostasis, targeted analysis was performed on genes involved in copper trafficking pathways. By far the largest impact of CuNPs on components of the copper pathway was identified in eleocytes, with the elevation of generalist copper chaperones (CutC) and metal transcription factor (MTF1), this mirrored the reduction in expression of copper transporters (CTR1 and ATP7A) and functionally specific chaperones (CCS and

COX11). Although the strongest copper pathway response was observed in leucocytes, granular amoebocytes did show a modest induction of genes encoding phytochelatin synthase and Metallothionein, both of which act as generalised metal chaperones distinct from CutC.

7.3.3 Comparing direct NP impact with ecotoxicological relevance

To compare the direct CuNP effect described in Chapter 5 with an ecotoxicologically more realistic scenario, an additional experiment was conducted where earthworms were exposed to copper ions together with different copper containing nanoparticles using soil as an exposure medium. Differentially expressed genes could be identified under all treatment conditions (Cu Ions, CuNP, CuSi), CuSi and low dose CuNP treated samples affected only a small number of genes, for which reason functional analysis could not be conducted. In terms of the effect of copper ions, low and high concentrations induced distinct biological responses. While the low dose of copper induced a moderate environmental challenge associated with a possible hormetic effect, higher soil copper content resulted in a cytotoxic effect which was represented by increased stress response-related pathways and DNA damage. Although the specific suite of genes induced were different, Gene Ontology and pathway enrichment revealed that high dose CuNPs induce a similar biological effect as the equivalent ionic copper dose. However, CuNP exposure appeared to distinctly induce some immunity-related biological processes, such as “antigen processing and presentation”.

7.4 Limitations

7.4.1 The genomic framework of earthworm immunity

7.4.1.1 Tissue-specific transcriptomic atlas

Resources provided by the tissue-specific transcriptomic atlas delivered an excellent overall insight into the earthworm innate immune system, however, there are still limitations that could only be partially resolved during this study. It was only possible to examine a restricted number of tissues and, while the experiments were designed to be as comprehensive as possible, tissues with different physiological functions could be incorporated to give further insights. Since obvious limitations in time and resources only enabled analysis of a limited number of samples with tissues such as the skin, which

provide the first-line defence against pathogens, not included in this study. Another limiting factor was the lack of separation between the gut and chloragogen tissues. Despite the endeavour of extracting RNA from only tissue samples with high purity, in the case of gut/chloragogen tissues the microscopic dissection based separation appeared to be technically challenging due to the delicate nature of the chloragogen. For this reason, the distinct functional characteristics of the gut and chloragogen tissues could not be identified, even though it is reasonable to assume they possess distinct immunological roles.

Important developments in sequencing, such as 10x spatial transcriptomics, provide the opportunity to overlay transcriptional expression on top of a tissue section to directly associate gene expression to morphologically distinct cells (Ståhl et al. 2016). This would allow the full heterogeneity linked to tissue specific gene expression to be explored. Tissues with complex functions, such as the earthworm chloragogenous tissue or cerebral ganglia, may be explored in exquisite detail, assigning specific functions to histologically distinct cell types.

7.4.1.2 Reference genome

The conducted *ab initio* gene prediction resulted in a high number of structural transcript completeness compared to the *de novo* transcriptomes. These prediction algorithms require long and tedious species-specific training to achieve the most accurate gene prediction model for each organism. However, to avoid this highly time-consuming method, we used the well-established *D. melanogaster* training set (BioBam 2019). In the absence of a purely earthworm-specific gene prediction model, the annotation success and accuracy mostly depends on similarities in gene structure between the model organism and the applied training set, with discrepancies potentially producing a less accurate annotation. The application of the *D. melanogaster* based training set was not perfectly optimised for the earthworm genome, however, with the aid of the RNA-Seq based exon and intron prediction, it did provide a high number of accurate gene prediction, that was significantly better than those derived using training sets from other well-known model organisms. One future avenue to increase gene prediction accuracy could be to create an earthworm specific training set using data from a closely related earthworm species, or via the application of long-read cDNA

sequencing that could highly simplify the challenge of accurate gene prediction by providing full, gene-length, cDNA sequences.

7.4.2 Spatio-temporal immune effect of CuNP

Treating the major coelomocyte populations as separate samples, combined with the three treatment types and nine different sampling points, resulted in the spatio-temporal experiment becoming inflated. To avoid further increase in cost and the required biological material, the experiment was designed without using sample replication. This prohibited the application of the most routinely used pairwise DE analysis pipelines and greatly reduced the number of software packages that could be used for DEG identification. It also allowed the observation of temporal gene expression patterns. However, by using replication at each time-point the statistical power of DEG identification could possibly be increased.

Another major limitation of the experimental design was the lack of observable transcriptomic changes in the case *B. subtilis* only treatment. Although a non-lethal concentration was chosen to avoid mortality during the length of the experiment, the total lack of measurable response meant the evaluation of the combined treatment became extremely challenging. Without the possibility of comparing the results of the combined exposure to an immune response “truth set”, we could not identify if the observed biological response was the result of an immune sensitizer reaction caused by the combination of NPs injection and bacterial challenge or it was a response only to the CuNP treatment. Although the clarification of this question requires further investigation, the lack of DEGs further verified the stringency of the used DE analysis pipeline.

7.4.3 Indirect Copper exposure

The key limitation factor during the indirect copper exposure experiment was the COVID caused delay in the body burden results. Outlier organisms could not be assigned during the differential expression without information about individual Cu body burdens, to provide ‘real’ exposure dose affecting the given individuals. For this reason, the possible differences due to the feeding habits of each earthworm could not be incorporated. Other limitations were linked to the lack of tissue specificity associated with the

experimental design. A gut/chloragogen, coelomic fluid, nephridia and body wall specific RNA-Seq design would be useful to link cell-specific burdens to Cu derived from nanomaterial or metal ions with transcriptional impact on a cell by cell basis. It may be possible to compare metal load mapping, determined using synchrotron or micro-laser ablation-ICPMS approaches, with consecutive cryo-sections analysed using 10x spatial transcriptomics.

7.5 Future developments

7.5.1 Metal metabolism in earthworm – a cell specific affair

One of the important challenges for earthworm toxicology is that the dose of the toxic agent often depends on the feeding habits (soil consumption) of the individual animal. The differences in the absorbed agent can result in an increased deviation in the biological response for each measured individual. The injection method introduced in this study provides a highly reproducible method for precise metal dosing and studying the dynamics of metal trafficking responses independently of the soil consumption habits of the earthworms. The injection method could also enable us to address fundamental questions relating to the metal homeostasis of the coelomic cavity. Using the earthworm coelomic cavity as '*in vivo*' cell culture vessel (96 h experiment length) allows pharmaceutical or genetic manipulation. This may allow chemical blocking of particular channels or suppression/editing of specific genes using RNAi or CRISPR technologies (Barrangou and Doudna 2016, Hannon 2002) followed by the analysis of cell type-specific responses. This approach may also allow us to address specific key questions, such as the processes controlling coelomocyte differentiation.

7.5.2 Differential transcripts usage and non-polyadenylated RNAs

Due to the last few years of innovation in the field of third-generation sequencing technologies, long-read cDNA sequencing has become more affordable. The use of long-read cDNA and direct mRNA sequencing could not only greatly reduce the difficulties of reference genome annotation but also provide more accurate gene-objects definition for functional studies (Cook et al. 2019). These innovations will offer new insights into alternative splicing events, which are known to play critical roles in the regulation of innate immune responses (Rhoads and Au 2015, Carpenter et al. 2014).

According to the presence or absence of a poly(A) tail, RNA molecules can be classified into polyadenylated and non-polyadenylated transcripts. Several poly(A) negative RNA species, such as miRNAs and lncRNAs have been identified as important regulatory components of the innate immune response (Zhang and Cao 2016, Li et al. 2012, Nejad et al. 2018). Most of the traditional RNA-Seq library preparation protocols are based on A-tailing based amplification methods that exclude these RNAs from the sequencing procedure. By using miRNA compatible amplification procedures or performing direct RNA sequencing, these RNA species could be characterized and their importance in invertebrate immune regulation processes explored.

7.5.3 Origin and maturation of coelomocytes

Even though a relatively high number of experiments have targeted the immunological functions of both eleocytes and amebocytes (Plytycz and Morgan 2011), the origin of these cell populations is still yet to be fully described. Current studies suggest eleocytes are derived from the chloragogenous tissue, following detachment of chloragocytes from the blood sinuses, while amebocytes originate from the mesenchymal linings of the coelomic cavity (Engelmann et al. 2016a). However, the details about the location of the progenitor cells and the mechanisms behind these cellular maturation processes are rather limited.

Biologically crucial questions may be addressed by supplementing the tissue and cell-specific resources described here with the recent advantages of novel sequencing-based technologies, such as the spatial transcriptomics. Using the benefits of spatial transcriptomics on cross-sections of intestine tissue could resolve the challenge of mechanical separation of chloragogen and gut tissue, to provide the first comprehensive transcriptomic profile for the specific functions of these tissues. Furthermore, it could create the possibility to follow the chloragocytes to eleocytes differentiation processes at the transcriptomic level. The spatial analysis of the longitudinal intestine sections could help to identify the location of the progenitor cells involved in eleocyte maturation.

7.5.4 Coelomocyte classification based on transcriptomic fingerprints

The proportions of the three different coelomocyte populations has previously been identified as critical to earthworm health (Cooper 1996). The establishment of cell-type specific markers allows the development of a profiling method to determine coelomocyte population ratios from bulk RNA-Seq experiments produced using a mixed coelomocyte population. This could allow rapid insights into the immune response of earthworms to different cytotoxic challenges.

Future utilisation of the recently available single-cell transcriptomics (Kulkarni et al. 2019) means the cell morphology-based coelomocyte separation (using FACS) could be bypassed and coelomocytes populations could be separated based on their transcriptomic fingerprints. Performing coelomocyte separation based purely on the transcriptomic fingerprints of the different cell-types could identify functional cell populations that cannot be separated using morphological differences, as well as provide an insight into the coelomocyte maturation processes by separating coelomocytes at different stages of cell differentiation.

7.5.5 Conclusion

The last few years of research from the field of comparative immunology highlighted how little is known about the highly diverse invertebrate immune system. This is especially true in the case of less well-established model organisms, such as the earthworm. The major aim of this thesis was to provide an omics level template to study the components of the earthworm and other invertebrates innate immune systems. Although the results described are far from complete, the generated genomic and transcriptomic resources provided the foundation for studying the earthworm innate immune system at the transcriptomic level. The high number of newly identified TLR genes gave a great indication of how much more there is to learn, even about some of the most conserved components of the earthworm innate immune system. To gain an even higher-level understanding of earthworm immune system mechanisms, the resources developed in this study need to be supported with more functional studies.

The analysis of interactions between engineered MNPs and the innate immune system has shown that, if administered directly, CuNPs will impact the innate immune system

through 'sensitizing' it, arguably justifying the immune safety concerns relating to the applications of these particles (Zolnik et al.). However, when NPs go through biotransformation in soil, the exposure of the animals to 'native' NPs is very limited (Novo et al. 2020). Although, in this case, a CuNP specific response can be identified when exposed via soil, the responsive pathways overlap significantly with those of copper ions. Therefore, in comparison to the direct exposure, the ecological risk posed by environmentally available NPs is less well-supported when considered from the perspective of inducing an immune response.

8 References

- ABU-JAMOUS, B. & KELLY, S. 2018. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biology*, 19, 172.
- ADEMA, C. M. 2015. Fibrinogen-Related Proteins (FREPs) in Mollusks. *Results Probl Cell Differ*, 57, 111-29.
- AFFAR, E. B., DUFOUR, M., POIRIER, G. G. & NADEAU, D. 1998. Isolation, purification and partial characterization of chloragocytes from the earthworm species *Lumbricus terrestris*. *Molecular and Cellular Biochemistry*, 185, 123-133.
- AKIRA, S., TAKEDA, K. & KAISHO, T. 2001. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nature Immunology*, 2, 675-680.
- AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21, 30.
- ANDERSON, K. V. 2000. Toll signaling pathways in the innate immune response. *Current Opinion in Immunology*, 12, 13-19.
- ANDREWS, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data* [Online]. Available: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [Accessed].
- AREAL, H., ABRANTES, J. & ESTEVES, P. J. 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. *BMC Evolutionary Biology*, 11, 368.
- AUFFAN, M., ROSE, J., BOTTERO, J.-Y., LOWRY, G. V., JOLIVET, J.-P. & WIESNER, M. R. 2009. Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective. *Nature Nanotechnology*, 4, 634-641.
- BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19, 455-477.
- BARRANGOU, R. & DOUDNA, J. A. 2016. Applications of CRISPR technologies in research and beyond. *Nature biotechnology*, 34, 933-941.
- BAUN, A., HARTMANN, N. B., GRIEGER, K. & KUSK, K. O. 2008. Ecotoxicity of engineered nanoparticles to aquatic invertebrates: a brief review and recommendations for future toxicity testing. *Ecotoxicology*, 17, 387-395.
- BELVIN, M. P. & ANDERSON, K. V. 1996. A conserved signaling pathway: the *Drosophila* toll-dorsal pathway.
- BHAMBRI, A., DHAUNTA, N., PATEL, S. S., HARDIKAR, M., BHATT, A., SRIKAKULAM, N., SHRIDHAR, S., VELLARIKKAL, S., PANDEY, R., JAYARAJAN, R., VERMA, A., KUMAR, V., GAUTAM, P.,

- KHANNA, Y., KHAN, J. A., FROMM, B., PETERSON, K. J., SCARIA, V., SIVASUBBU, S. & PILLAI, B. 2018. Large scale changes in the transcriptome of *Eisenia fetida* during regeneration. *PLoS One*, 13, e0204234.
- BHATTACHARJEE, A., CHAKRABORTY, K. & SHUKLA, A. 2017. Cellular copper homeostasis: current concepts on its interplay with glutathione homeostasis and its implication in physiology and human diseases. *Metallomics*, 9, 1376-1388.
- BILEJ, M., PROCHAZKOVA P FAU - SILEROVA, M., SILEROVA M FAU - JOSKOVA, R. & JOSKOVA, R. 2013. Earthworm Immunity. In: Madame Curie Bioscience Database Earthworm immunity.
- BILEJ, M., ROSSMANN, P., ŠINKORA, M., HANUŠOVÁ, R., BESCHIN, A., RAES, G. & DE BAETSELIER, P. 1998. Cellular expression of the cytolytic factor in earthworms *Eisenia foetida*. *Immunology Letters*, 60, 23-29.
- BINDEA, G., MLECNIK, B., HACKL, H., CHAROENTONG, P., TOSOLINI, M., KIRILOVSKY, A., FRIDMAN, W.-H., PAGÈS, F., TRAJANOSKI, Z. & GALON, J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25, 1091-1093.
- BIOBAM. 2019. *Bioinformatics Made Easy, BioBam Bioinformatics* [Online]. OmicsBox. Available: <https://www.biobam.com/omicsbox> [Accessed March 3].
- BJÖRN, L. O. & GOVINDJEE 2008. The Evolution of Photosynthesis and Its Environmental Impact. In: BJÖRN, L. O. (ed.) *Photobiology: The Science of Life and Light*. New York, NY: Springer New York.
- BODÓ, K., ERNSZT, D., NÉMETH, P. & ENGELMANN, P. 2018. Distinct immune- and defense-related molecular fingerprints in separated coelomocyte subsets of *Eisenia andrei* earthworms. *Invertebrate Survival Journal*, 338-345%V 15.
- BODÓ, K., HAYASHI, Y., GERENCSÉR, G., LÁSZLÓ, Z., KÉRI, A., GALBÁCS, G., TELEK, E., MÉSZÁROS, M., DELI, M. A., KOKHANYUK, B., NÉMETH, P. & ENGELMANN, P. 2020. Species-specific sensitivity of *Eisenia* earthworms towards noble metal nanoparticles: a multiparametric in vitro study. *Environmental Science: Nano*, 7, 3509-3525.
- BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114-2120.
- BORASCHI, D., ITALIANI, P., PALOMBA, R., DECUZZI, P., DUSCHL, A., FADEEL, B. & MOGHIMI, S. M. 2017. Nanoparticles and innate immunity: new perspectives on host defence. *Seminars in Immunology*, 34, 33-51.
- BRENNAN, J. J. & GILMORE, T. D. 2018. Evolutionary Origins of Toll-like Receptor Signaling. *Molecular Biology and Evolution*, 35, 1576-1587.

- BROAD INSTITUTE. 2018. *Picard Toolkit* [Online]. Broad Institute. Available: <http://broadinstitute.github.io/picard/> [Accessed].
- BROWN, C. G. & CLARKE, J. 2016. Nanopore development at Oxford Nanopore. *Nature Biotechnology*, 34, 810-811.
- BRŮNA, T., LOMSADZE, A. & BORODOVSKY, M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, 2.
- BRYANT, D. M., JOHNSON, K., DITOMMASO, T., TICKLE, T., COUGER, M. B., PAYZIN-DOGRU, D., LEE, T. J., LEIGH, N. D., KUO, T.-H., DAVIS, F. G., BATEMAN, J., BRYANT, S., GUZIKOWSKI, A. R., TSAI, S. L., COYNE, S., YE, W. W., FREEMAN, R. M., JR., PESHKIN, L., TABIN, C. J., REGEV, A., HAAS, B. J. & WHITED, J. L. 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, 18, 762-776.
- BUCKLEY, K. & RAST, J. 2012a. Dynamic Evolution of Toll-Like Receptor Multigene Families in Echinoderms. *Frontiers in Immunology*, 3.
- BUCKLEY, K. M. & RAST, J. P. 2012b. Dynamic evolution of toll-like receptor multigene families in echinoderms. *Frontiers in immunology*, 3, 136-136.
- BUNDY, J. G., SIDHU, J. K., RANA, F., SPURGEON, D. J., SVENDSEN, C., WREN, J. F., STÜRZENBAUM, S. R., MORGAN, A. J. & KILLE, P. 2008. 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology*, 6, 25.
- CALABRESE, E. J. & BALDWIN, L. A. 2000. Chemical hormesis: its historical foundations as a biological hypothesis. *Human & Experimental Toxicology*, 19, 2-31.
- CANCIO, I., GWYNN, I., IRELAND, M. P. & CAJARAVILLE, M. P. 1995. The effect of sublethal lead exposure on the ultrastructure and on the distribution of acid phosphatase activity in chloragocytes of earthworms (Annelida, Oligochaeta). *Histochem J*, 27, 965-73.
- CANESI, L., CIACCI, C. & BALBI, T. 2016. Invertebrate Models for Investigating the Impact of Nanomaterials on Innate Immunity: The Example of the Marine Mussel *Mytilus* spp. *Current Bionanotechnology*, 2, 77-83.
- CARLHOFF, D. & D'HAESE, J. 1987. Slow type muscle cells in the earthworm gizzard with a distinct, Ca²⁺-regulated myosin isoform. *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology*, 157, 589-597.
- CARPENTER, S., RICCI, E. P., MERCIER, B. C., MOORE, M. J. & FITZGERALD, K. A. 2014. Post-transcriptional regulation of gene expression in innate immunity. *Nature Reviews Immunology*, 14, 361-376.

- CASALS, E. & PUNTES, V. F. Inorganic nanoparticle biomolecular corona: formation, evolution and biological impact.
- CERENIUS, L., LEE, B. L. & SÖDERHÄLL, K. 2008. The proPO-system: pros and cons for its role in invertebrate immunity. *Trends in Immunology*, 29, 263-271.
- CERENIUS, L. & SÖDERHÄLL, K. 2013. Variable immune molecules in invertebrates. *The Journal of Experimental Biology*, 216, 4313.
- CHALLIS, R., RICHARDS, E., RAJAN, J., COCHRANE, G. & BLAXTER, M. 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3: Genes/Genomes/Genetics*, 10, 1361-1374.
- CHEN, P., KANEHIRA, K. & TANIGUCHI, A. 2013. Role of toll-like receptors 3, 4 and 7 in cellular uptake and response to titanium dioxide nanoparticles. *Science and Technology of Advanced Materials*, 14, 015008.
- CLIFFORD, R. J., MARYON, E. B. & KAPLAN, J. H. 2016. Dynamic internalization and recycling of a metal ion transporter: Cu homeostasis and CTR1, the human Cu⁺ uptake system. *Journal of Cell Science*, 129, 1711-1721.
- COATES, C. J. & NAIRN, J. 2014. Diverse immune functions of hemocyanins. *Developmental & Comparative Immunology*, 45, 43-55.
- CONESA, A., GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., TALÓN, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674-3676.
- COOK, D. E., VALLE-INCLAN, J. E., PAJORO, A., ROVENICH, H., THOMMA, B. P. & FAINO, L. 2019. Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant physiology*, 179, 38-54.
- COOMBE, L., ZHANG, J., VANDERVALK, B. P., CHU, J., JACKMAN, S. D., BIROL, I. & WARREN, R. L. 2018. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*, 19, 234.
- COOPER, E. 1996. Earthworm immunity. *Invertebrate immunology*. Springer.
- COOPER EDWIN, L., RINKEVICH, B., UHLENBRUCK, G. & VALEMBOIS, P. 1992. Invertebrate Immunity: Another Viewpoint. *Scandinavian Journal of Immunology*, 35, 247-266.
- COOPER, E. L. 2003. Comparative Immunology. *Current Pharmaceutical Design*, 9, 119-131.
- CZURYŁO, E., KULIKOVA, N. & SOBOTA, A. 2008. Disturbance of smooth muscle regulatory function by Eisenia foetida toxin lysenin: Insight into the mechanism of smooth muscle contraction. *Toxicon : official journal of the International Society on Toxinology*, 51, 1090-102.

- D'MELLO, S. R., CRUZ, C. N., CHEN, M.-L., KAPOOR, M., LEE, S. L. & TYNER, K. M. 2017. The evolving landscape of drug products containing nanomaterials in the United States. *Nature Nanotechnology*, 12, 523-529.
- DAVIDSON, C. R., BEST, N. M., FRANCIS, J. W., COOPER, E. L. & WOOD, T. C. 2008. Toll-like receptor genes (TLRs) from *Capitella capitata* and *Helobdella robusta* (Annelida). *Developmental & Comparative Immunology*, 32, 608-612.
- DEV, A., IYER, S., RAZANI, B. & CHENG, G. 2011. NF- κ B and Innate Immunity. In: KARIN, M. (ed.) *NF- κ B in Health and Disease*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- DHEILLY, N. M., ADEMA, C., RAFTOS, D. A., GOURBAL, B., GRUNAU, C. & DU PASQUIER, L. 2014. No more non-model species: The promise of next generation sequencing for comparative immunology. *Developmental & Comparative Immunology*, 45, 56-66.
- DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.
- DRAKE, H. L. & HORN, M. A. 2007. As the Worm Turns: The Earthworm Gut as a Transient Habitat for Soil Microbial Biomes. *Annual Review of Microbiology*, 61, 169-189.
- DU, J., ZHANG, Y. S., HOBSON, D. & HYDBRING, P. 2017. Nanoparticles for immune system targeting. *Drug Discovery Today*, 22, 1295-1301.
- DUBEY, A. & MAILAPALLI, D. R. 2016. Nanofertilisers, Nanopesticides, Nanosensors of Pest and Nanotoxicity in Agriculture. In: LICHTFOUSE, E. (ed.) *Sustainable Agriculture Reviews: Volume 19*. Cham: Springer International Publishing.
- DUDZIC, J. P., KONDO, S., UEDA, R., BERGMAN, C. M. & LEMAITRE, B. 2015. Drosophila innate immunity: regional and functional specialization of prophenoloxidases. *BMC Biology*, 13, 81.
- DVOŘÁK, J., ROUBALOVÁ, R., PROCHÁZKOVÁ, P., ROSSMANN, P., ŠKANTA, F. & BILEJ, M. 2016. Sensing microorganisms in the gut triggers the immune response in *Eisenia andrei* earthworms. *Developmental & Comparative Immunology*, 57, 67-74.
- EALIA, S. & SARAVANAKUMAR, M. P. 2017. A review on the classification, characterisation, synthesis of nanoparticles and their application. *IOP Conference Series: Materials Science and Engineering*, 263, 032019.
- EL HADRI, H., LOUIE, S. M. & HACKLEY, V. A. 2018. Assessing the interactions of metal nanoparticles in soil and sediment matrices – a quantitative analytical multi-technique approach. *Environmental Science: Nano*, 5, 203-214.
- ENGELMANN, P., HAYASHI, Y., BODÓ, K., ERNSZT, D., SOMOGYI, I., STEIB, A., ORBÁN, J., POLLÁK, E., NYITRAI, M., NÉMETH, P. & MOLNÁR, L. 2016a. Phenotypic and functional characterization of earthworm coelomocyte subsets: Linking light scatter-based cell

typing and imaging of the sorted populations. *Developmental & Comparative Immunology*, 65, 41-52.

- ENGELMANN, P., HAYASHI, Y., BODÓ, K. & MOLNÁR, L. 2016b. Chapter 4 - New Aspects of Earthworm Innate Immunity: Novel Molecules and Old Proteins With Unexpected Functions. *In: BALLARIN, L. & CAMMARATA, M. (eds.) Lessons in Immunity*. Academic Press.
- EWELS, P., MAGNUSSON, M., LUNDIN, S. & KÄLLER, M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32, 3047-3048.
- FINNEY, L., VOGT, S., FUKAI, T. & GLESNE, D. 2009. Copper and angiogenesis: unravelling a relationship key to cancer progression. *Clinical and experimental pharmacology & physiology*, 36, 88-94.
- FISCHER, E. 1993. The myelo-erythroid nature of the chloragogenous-like tissues of the annelids. *Comparative Biochemistry and Physiology Part A: Physiology*, 106, 449-453.
- FOLEY, E. & O'FARRELL, P. H. 2004. Functional Dissection of an Innate Immune Response by a Genome-Wide RNAi Screen. *PLOS Biology*, 2, e203.
- FOLMER, O., BLACK, M., HOEH, W., LUTZ, R. & VRIJENHOEK, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology*, 3, 294-299.
- FRANCHI, L., WARNER, N., VIANI, K. & NUÑEZ, G. 2009. Function of Nod-like receptors in microbial recognition and host defense. *Immunol Rev*, 227, 106-28.
- FRANCIS, J., WREESMAN, S., YONG, S., REIGSTAD, K., KRUTZIK, S. & COOPER, E. L. 2007. Analysis of the earthworm coelomocyte cell surface for the presence of Toll-like immune receptors. *European Journal of Soil Biology*, 43, S92-S96.
- GHOSH, J., LUN, C. M., MAJESKE, A. J., SACCHI, S., SCHRANKEL, C. S. & SMITH, L. C. 2011. Invertebrate immune diversity. *Developmental & Comparative Immunology*, 35, 959-974.
- GIBLIN, S. P., SCHWENZER, A. & MIDWOOD, K. S. 2020. Alternative splicing controls cell lineage-specific responses to endogenous innate immune triggers within the extracellular matrix. *Matrix Biology*, 93, 95-114.
- GILBERT, D. 2013. Gene-omes built from mRNA seq not genome DNA.
- GÖTZ, P. Encapsulation in Arthropods. 1986 Berlin, Heidelberg. Springer Berlin Heidelberg, 153-170.
- GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., WILLIAMS, T. D., NAGARAJ, S. H., NUEDA, M. J., ROBLES, M., TALÓN, M., DOPAZO, J. & CONESA, A. 2008. High-throughput functional

annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36, 3420-3435.

GRABHERR, M. G., HAAS, B. J., YASSOUR, M., LEVIN, J. Z., THOMPSON, D. A., AMIT, I., ADICONIS, X., FAN, L., RAYCHOWDHURY, R., ZENG, Q., CHEN, Z., MAUCELI, E., HACOEN, N., GNIRKE, A., RHIND, N., DI PALMA, F., BIRREN, B. W., NUSBAUM, C., LINDBLAD-TOH, K., FRIEDMAN, N. & REGEV, A. 2011a. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29, 644-652.

GRABHERR, M. G., HAAS, B. J., YASSOUR, M., LEVIN, J. Z., THOMPSON, D. A., AMIT, I., ADICONIS, X., FAN, L., RAYCHOWDHURY, R., ZENG, Q., CHEN, Z., MAUCELI, E., HACOEN, N., GNIRKE, A., RHIND, N., DI PALMA, F., BIRREN, B. W., NUSBAUM, C., LINDBLAD-TOH, K., FRIEDMAN, N. & REGEV, A. 2011b. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, 29, 644-652.

GRIFFIN, F. M. 1977. Opsonization. In: DAY, N. K. & GOOD, R. A. (eds.) *Biological Amplification Systems in Immunology*. Boston, MA: Springer US.

HAAS, B. J., DELCHER, A. L., MOUNT, S. M., WORTMAN, J. R., SMITH, R. K., JR., HANNICK, L. I., MAITI, R., RONNING, C. M., RUSCH, D. B., TOWN, C. D., SALZBERG, S. L. & WHITE, O. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 31, 5654-66.

HADRUP, N., AIMONEN, K., ILVES, M., LINDBERG, H., ATLURI, R., SAHLGREN, N. M., JACOBSEN, N. R., BARFOD, K. K., BERTHING, T., LAWLOR, A., NORPPA, H., WOLFF, H., JENSEN, K. A., HOUGAARD, K. S., ALENIUS, H., CATALAN, J. & VOGEL, U. 2020. Pulmonary toxicity of synthetic amorphous silica – effects of porosity and copper oxide doping. *Nanotoxicology*, 1-18.

HANNON, G. J. 2002. RNA interference. *Nature*, 418, 244-251.

HAYASHI, Y. & ENGELMANN, P. 2013. Earthworm's immunity in the nanomaterial world: New room, future challenges. *Invertebrate Survival Journal*, 10, 69-76.

HAYASHI, Y., MICLAUS, T., SCAVENIUS, C., KWIATKOWSKA, K., SOBOTA, A., ENGELMANN, P., SCOTT-FORDSMAND, J. J., ENGHILD, J. J. & SUTHERLAND, D. S. 2013. Species Differences Take Shape at Nanoparticles: Protein Corona Made of the Native Repertoire Assists Cellular Interaction. *Environmental Science & Technology*, 47, 14367-14375.

HAYNES, W. 2013. Benjamini–Hochberg Method. In: DUBITZKY, W., WOLKENHAUER, O., CHO, K.-H. & YOKOTA, H. (eds.) *Encyclopedia of Systems Biology*. New York, NY: Springer New York.

HE, S., JOHNSTON, P. R. & MCMAHON, D. P. 2020. Analyzing Immunity in Non-model Insects Using De Novo Transcriptomics. In: SANDRELLI, F. & TETTAMANTI, G. (eds.) *Immunity in Insects*. New York, NY: Springer US.

HEILIGTAG, F. J. & NIEDERBERGER, M. 2013. The fascinating world of nanoparticle research. *Materials Today*, 16, 262-271.

- HESS, W. N. 1925. Nervous system of the earthworm, *lumbricus terrestris* L. *Journal of Morphology*, 40, 235-259.
- HEWARD, J. A. & LINDSAY, M. A. 2014. Long non-coding RNAs in the regulation of the immune response. *Trends in Immunology*, 35, 408-419.
- HOMA, J. 2018. Earthworm coelomocyte extracellular traps: structural and functional similarities with neutrophil NETs. *Cell and Tissue Research*, 371, 407-414.
- HOMA, J., ORTMANN, W. & KOLACZKOWSKA, E. 2016. Conservative Mechanisms of Extracellular Trap Formation by Annelida *Eisenia andrei*: Serine Protease Activity Requirement. *PLOS ONE*, 11, e0159031.
- HOSTETTER, R. K. & COOPER, E. L. 1974. Earthworm Coelomocyte Immunity. In: HANNA, M. G. & COOPER, E. L. (eds.) *Contemporary Topics in Immunobiology: Volume 4 Invertebrate Immunology*. Boston, MA: Springer US.
- HUBLEY, R., FINN, R. D., CLEMENTS, J., EDDY, S. R., JONES, T. A., BAO, W., SMIT, A. F. A. & WHEELER, T. J. 2015. The Dfam database of repetitive DNA families. *Nucleic Acids Research*, 44, D81-D89.
- HUMPHREYS, T. & REINHERZ, E. L. 1994. Invertebrate immune recognition, natural immunity and the evolution of positive selection. *Immunology Today*, 15, 316-320.
- HUNTER, S., APWEILER, R., ATTWOOD, T. K., BAIRUCH, A., BATEMAN, A., BINNS, D., BORK, P., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LAUGRAUD, A., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., MULDER, N., NATALE, D., ORENGO, C., QUINN, A. F., SELENGUT, J. D., SIGRIST, C. J. A., THIMMA, M., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C. H. & YEATS, C. 2009. InterPro: the integrative protein signature database. *Nucleic acids research*, 37, D211-D215.
- ISHIKAWA, K., ISHII, H. & SAITO, T. 2006. DNA Damage-Dependent Cell Cycle Checkpoints and Genomic Stability. *DNA and Cell Biology*, 25, 406-411.
- IWANAGA, S. & LEE, B. L. 2005. Recent advances in the innate immunity of invertebrate animals. *Journal of biochemistry and molecular biology*, 38, 128-150.
- J DIOGÈNE, M. D., G G POIRIER, D NADEAU 1997. Extrusion of earthworm coelomocytes: comparison of the cell populations recovered from the species *Lumbricus terrestris*, *Eisenia fetida* and *Octolasion tyrtaeum*. *Laboratory Animals*, 31, 326-336.
- JAMIESON, B. G. M. 1981. *Ultrastructure of the oligochaeta*, Academic Press, London.
- JASSAL, B., MATTHEWS, L., VITERI, G., GONG, C., LORENTE, P., FABREGAT, A., SIDIROPOULOS, K., COOK, J., GILLESPIE, M., HAW, R., LONEY, F., MAY, B., MILACIC, M., ROTHFELS, K., SEVILLA, C., SHAMOVSKY, V., SHORSER, S., VARUSAI, T., WEISER, J., WU, G., STEIN, L., HERMJAKOB, H. & D'EUSTACHIO, P. 2020. The reactome pathway knowledgebase.

- JIN, M. S. & LEE, J.-O. 2008. Structures of the Toll-like Receptor Family and Its Ligand Complexes. *Immunity*, 29, 182-191.
- JOHNSON, A. C., DONNACHIE, R. L., SUMPTER, J. P., JÜRGENS, M. D., MOECKEL, C. & PEREIRA, M. G. 2017. An alternative approach to risk rank chemicals on the threat they pose to the aquatic environment. *Sci Total Environ*, 599-600, 1372-1381.
- JOHNSON, G. B., BRUNN, G. J., TANG, A. H. & PLATT, J. L. 2003. Evolutionary clues to the functions of the Toll-like family as surveillance receptors. *Trends in Immunology*, 24, 19-24.
- JOSKOVÁ, R., ŠILEROVÁ, M., PROCHÁZKOVÁ, P. & BILEJ, M. 2009. Identification and cloning of an invertebrate-type lysozyme from *Eisenia andrei*. *Developmental & Comparative Immunology*, 33, 932-938.
- JOUQUET, P., DAUBER, J., LAGERLÖF, J., LAVELLE, P. & LEPAGE, M. 2006. Soil invertebrates as ecosystem engineers: Intended and accidental effects on soil and feedback loops. *Applied Soil Ecology*, 32, 153-164.
- JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110, 462-467.
- KANEHISA, M. & SATO, Y. 2020. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci*, 29, 28-35.
- KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2015. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44, D457-D462.
- KAWAI, T. & AKIRA, S. 2008. Toll-like receptor and RIG-I-like receptor signaling. *Ann N Y Acad Sci*, 1143, 1-20.
- KELLER, A. A., ADELEYE, A. S., CONWAY, J. R., GARNER, K. L., ZHAO, L., CHERR, G. N., HONG, J., GARDEA-TORRESDEY, J. L., GODWIN, H. A., HANNA, S., JI, Z., KAWETEERAWAT, C., LIN, S., LENIHAN, H. S., MILLER, R. J., NEL, A. E., PERALTA-VIDEA, J. R., WALKER, S. L., TAYLOR, A. A., TORRES-DUARTE, C., ZINK, J. I. & ZUVERZA-MENA, N. 2017. Comparative environmental fate and toxicity of copper nanomaterials. *NanoImpact*, 7, 28-40.
- KHAN, I., SAEED, K. & KHAN, I. 2019. Nanoparticles: Properties, applications and toxicities. *Arabian Journal of Chemistry*, 12, 908-931.
- KIM, B. Y. S., RUTKA, J. T. & CHAN, W. C. W. 2010. Nanomedicine. *New England Journal of Medicine*, 363, 2434-2443.
- KOLDE, R. 2012. Pheatmap: pretty heatmaps. *R package version*, 1.

- KOLMOGOROV, M., YUAN, J., LIN, Y. & PEVZNER, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37, 540-546.
- KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H. & PHILLIPPY, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 27, 722-736.
- KORF, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics*, 5, 59.
- KUKURBA, K. R. & MONTGOMERY, S. B. 2015. RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015, 951-969.
- KULKARNI, A., ANDERSON, A. G., MERULLO, D. P. & KONOPKA, G. 2019. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology*, 58, 129-136.
- KUMAR, L. & FUTSCHIK, M. 2007. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, 2, 5-7.
- KUMAR, S., STECHER, G., LI, M., KNYAZ, C. & TAMURA, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*, 35, 1547-1549.
- KUNO, K. & MATSUSHIMA, K. 1994. The IL-1 receptor signaling pathway. *Journal of Leukocyte Biology*, 56, 542-547.
- LAULAN, A., LESTAGE, J., BOUC, A. M. & CHATEAUREYNAUD-DUPRAT, P. 1988. The phagocytic activity of *Lumbricus terrestris* leukocytes is enhanced by the vertebrate opsonins: IgG and complement C3b fragment. *Dev Comp Immunol*, 12, 269-77.
- LEE, J.-H., PARTHIBAN, P., JIN, G.-Z., KNOWLES, J. C. & KIM, H.-W. 2020. Materials roles for promoting angiogenesis in tissue regeneration. *Progress in Materials Science*, 100732.
- LEE, W. S., KIM, E., CHO, H.-J., KANG, T., KIM, B., KIM, M. Y., KIM, Y. S., SONG, N. W., LEE, J.-S. & JEONG, J. 2018. The Relationship between Dissolution Behavior and the Toxicity of Silver Nanoparticles on Zebrafish Embryos in Different Ionic Environments. *Nanomaterials (Basel, Switzerland)*, 8, 652.
- LETUNIC, I. & BORK, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, 47, W256-W259.
- LI, B. & DEWEY, C. N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
- LI, F., PIGNATTA, D., BENDIX, C., BRUNKARD, J. O., COHN, M. M., TUNG, J., SUN, H., KUMAR, P. & BAKER, B. 2012. MicroRNA regulation of plant innate immune receptors. *Proceedings of the National Academy of Sciences*, 109, 1790-1795.

- LI, H. 2016. Minimap and miniiasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32, 2103-2110.
- LI, X., HAO, S., HAN, A., YANG, Y., FANG, G., LIU, J. & WANG, S. 2019. Intracellular Fenton reaction based on mitochondria-targeted copper(ii)-peptide complex for induced apoptosis. *Journal of Materials Chemistry B*, 7, 4008-4016.
- LIEBMANN, E. 1942. The coelomocytes of Lumbricidae. *Journal of Morphology*, 71, 221-249.
- LIU, J. & CAO, X. 2016. Cellular and molecular regulation of innate inflammatory responses. *Cellular & Molecular Immunology*, 13, 711-721.
- LIU, S. & HAN, M.-Y. 2010. Silica-Coated Metal Nanoparticles. *Chemistry – An Asian Journal*, 5, 36-45.
- LIU, Y., HARDIE, J., ZHANG, X. & ROTELLO, V. M. 2017. Effects of engineered nanoparticles on the innate immune system. *Seminars in Immunology*, 34, 25-32.
- LLOYD, D. R. & PHILLIPS, D. H. 1999. Oxidative DNA damage mediated by copper(II), iron(II) and nickel(II) Fenton reactions: evidence for site-specific mechanisms in the formation of double-strand breaks, 8-hydroxydeoxyguanosine and putative intrastrand cross-links. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 424, 23-36.
- LOKER, E. S., ADEMA, C. M., ZHANG, S.-M. & KEPLER, T. B. 2004. Invertebrate immune systems – not homogeneous, not simple, not well understood. *Immunological reviews*, 198, 10-24.
- LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.
- LÓW, P., MOLNÁR, K. & KRISKA, G. 2016. Dissection of the Earthworm (*Lumbricus terrestris*). In: LÓW, P., MOLNÁR, K. & KRISKA, G. (eds.) *Atlas of Animal Anatomy and Histology*. Cham: Springer International Publishing.
- LUNDQVIST, M. 2013. Tracking protein corona over time. *Nature Nanotechnology*, 8, 701-702.
- LUNDQVIST, M., STIGLER, J., ELIA, G., LYNCH, I., CEDERVALL, T. & DAWSON, K. A. 2008. Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. *Proceedings of the National Academy of Sciences*, 105, 14265.
- MÁCSIK, L., SOMOGYI, I., OPPER, B., BOVÁRI-BIRI, J., POLLÁK, E., MOLNAR, L., NEMETH, P. & ENGELMANN, P. 2015. Induction of apoptosis-like cell death by coelomocyte extracts from *Eisenia andrei* earthworms. *Molecular immunology*, 67.

- MANI, R., BALASUBRAMANIAN, S., RAGHUNATH, A. & PERUMAL, E. 2020. Chronic exposure to copper oxide nanoparticles causes muscle toxicity in adult zebrafish. *Environ Sci Pollut Res Int*, 27, 27358-27369.
- MARCHEGGIANO, A., IANNONI, C. & DAVOLI, C. 1985. Thyroglobulin-like immunoreactivity in the nervous system of *Eisenia foetida* (Annelida, Oligochaeta). *Cell and Tissue Research*, 241, 429-433.
- MARINI, F. & BINDER, H. 2019. pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*, 20, 331.
- MAURICIO, M. D., GUERRA-OJEDA, S., MARCHIO, P., VALLES, S. L., ALDASORO, M., ESCRIBANO-LOPEZ, I., HERANCE, J. R., ROCHA, M., VILA, J. M. & VICTOR, V. M. 2018. Nanoparticles in Medicine: A Focus on Vascular Oxidative Stress. *Oxidative Medicine and Cellular Longevity*, 2018, 6231482.
- MAZUR, A. I., KLIMEK, M., MORGAN, A. J. & PLYTYCZ, B. 2011. Riboflavin storage in earthworm chloragocytes and chloragocyte-derived leucocytes and its putative role as chemoattractant for immunocompetent cells. *Pedobiologia*, 54, S37-S42.
- MCGINNIS, S. & MADDEN, T. L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32, W20-W25.
- MEDZHITOV, R. 2001. Toll-like receptors and innate immunity. *Nature Reviews Immunology*, 1, 135-145.
- METCHNIKOFF, E. 1968. *Lectures on the comparative pathology of inflammation; delivered at the Pasteur Institute in 1891*, New York, Dover Publications.
- MICHNA, A., BRASELMANN, H., SELMANSBERGER, M., DIETZ, A., HESS, J., GOMOLKA, M., HORNHARDT, S., BLÜTHGEN, N., ZITZELSBERGER, H. & UNGER, K. 2016. Natural Cubic Spline Regression Modeling Followed by Dynamic Network Reconstruction for the Identification of Radiation-Sensitivity Gene Association Networks from Time-Course Transcriptome Data. *PLOS ONE*, 11, e0160791.
- MITRA, S., KESWANI, T., DEY, M., BHATTACHARYA, S., SARKAR, S., GOSWAMI, S., GHOSH, N., DUTTA, A. & BHATTACHARYYA, A. 2012. Copper-induced immunotoxicity involves cell cycle arrest and cell death in the spleen and thymus. *Toxicology*, 293, 78-88.
- MITTAG, J., BEHRENDTS, T., NORDSTRÖM, K., ANSELMO, J., VENNSTRÖM, B. & SCHOMBURG, L. 2012. Serum copper as a novel biomarker for resistance to thyroid hormone. *Biochemical Journal*, 443, 103-109.
- MOUROUZIS, I., LAVECCHIA, A. M. & XINARIS, C. 2020. Thyroid Hormone Signalling: From the Dawn of Life to the Bedside. *Journal of Molecular Evolution*, 88, 88-103.

- MWAANGA, P., CARRAWAY, E. R. & VAN DEN HURK, P. 2014. The induction of biochemical changes in *Daphnia magna* by CuO and ZnO nanoparticles. *Aquatic Toxicology*, 150, 201-209.
- NAGARAJAN, R. 2008. Nanoparticles: Building Blocks for Nanotechnology. *Nanoparticles: Synthesis, Stabilization, Passivation, and Functionalization*. American Chemical Society.
- NAKASUGI, K., CROWHURST RN FAU - BALLY, J., BALLY J FAU - WOOD, C. C., WOOD CC FAU - HELLENS, R. P., HELLENS RP FAU - WATERHOUSE, P. M. & WATERHOUSE, P. M. 2013. De novo transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*.
- NAPPI, A. J. 1973. Hemocytic changes associated with the encapsulation and melanization of some insect parasites. *Experimental Parasitology*, 33, 285-302.
- NEJAD, C., STUNDEN, H. J. & GANTIER, M. P. 2018. A guide to miRNAs in inflammation and innate immune responses. *The FEBS Journal*, 285, 3695-3716.
- NG, T. H. & KURTZ, J. 2020. Dscam in immunity: A question of diversity in insects and crustaceans. *Dev Comp Immunol*, 105, 103539.
- NGUYEN, L. H. & HOLMES, S. 2019. Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, 15, e1006907.
- NONAKA 2011. The complement C3 protein family in invertebrates. *Invertebrate Survival Journal*, 8.
- NOVO, M., LAHIVE, E., DÍEZ ORTIZ, M., SPURGEON, D. J. & KILLE, P. 2020. Toxicogenomics in a soil sentinel exposure to Zn nanoparticles and ions reveals the comparative role of toxicokinetic and toxicodynamic mechanisms. *Environmental Science: Nano*, 7, 1464-1480.
- OFFORD, V., COFFEY, T. J. & WERLING, D. 2010. LRRfinder: a web application for the identification of leucine-rich repeats and an integrative Toll-like receptor database. *Dev Comp Immunol*, 34, 1035-41.
- OGUNSONA, E. O., MUTHURAJ, R., OJOGBO, E., VALERIO, O. & MEKONNEN, T. H. 2020. Engineered nanomaterials for antimicrobial applications: a review. *Applied Materials Today*, 18, 100473.
- OTTAVIANI, E. 2011. Is the distinction between innate and adaptive immunity in invertebrates still as clear-cut as thought? *Italian Journal of Zoology*, 78, 274-278.
- PALLARDY, M. J., TURBICA, I. & BIOLA-VIDAMMENT, A. 2017. Why the immune system should be concerned by nanomaterials? *Frontiers in Immunology*, 8, 544.

- PAREKH, S., ZIEGENHAIN, C., VIETH, B., ENARD, W. & HELLMANN, I. 2016. The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, 6, 25533.
- PETERS-GOLDEN, M., CANETTI, C., MANCUSO, P. & COFFEY, M. J. 2005. Leukotrienes: Underappreciated Mediators of Innate Immune Responses. *The Journal of Immunology*, 174, 589-594.
- PETERS, W. & WALLDORF, V. 1986. Endodermal secretion of chitin in the 'cuticle' of the earthworm gizzard. *Tissue and Cell*, 18, 361-374.
- PETRIS, M. J., SMITH, K., LEE, J. & THIELE, D. J. 2003. Copper-stimulated endocytosis and degradation of the human copper transporter, hCtr1. *J Biol Chem*, 278, 9639-46.
- PICCIRILLO, C. A., BJUR, E., TOPISIROVIC, I., SONENBERG, N. & LARSSON, O. 2014. Translational control of immune responses: from transcripts to translatoemes. *Nature Immunology*, 15, 503-511.
- PLOEG, M. J. C., VAN DEN BERG, J. H. J., BHATTACHARJEE, S., DE HAAN, L. H. J., ERSHOV, D. S., FOKKINK, R. G., ZUILHOF, H., RIETJENS, I. M. C. M. & VAN DEN BRINK, N. W. 2014. In vitro nanoparticle toxicity to rat alveolar cells and coelomocytes from the earthworm *Lumbricus rubellus*. *Nanotoxicology*, 8, 28-37.
- PLTYCZ, B., HOMA, J., KOZIOŁ, B., RÓZANOWSKA, M. & MORGAN, A. J. 2006. Riboflavin content in autofluorescent earthworm coelomocytes is species-specific. *Folia histochemica et cytobiologica*, 44, 275-280.
- PLYTYCZ, B. & MORGAN, A. 2011. Riboflavin storage in earthworm chloragocytes/eleocytes in an eco-immunology perspective. *ISJ-Invertebrate Survival Journal*, 8, 199-209.
- PROCHAZKOVA, P., ROUBALOVA, R., SKANTA, F., DVORAK, J., PACHECO, N. I. N., KOLARIK, M. & BILEJ, M. 2019a. Developmental and Immune Role of a Novel Multiple Cysteine Cluster TLR From *Eisenia andrei* Earthworms. *Frontiers in Immunology*, 10, 1277.
- PROCHAZKOVA, P., ROUBALOVA, R., SKANTA, F., DVORAK, J., PACHECO, N. I. N., KOLARIK, M. & BILEJ, M. 2019b. Developmental and Immune Role of a Novel Multiple Cysteine Cluster TLR From *Eisenia andrei* Earthworms. *Frontiers in immunology*, 10, 1277-1277.
- PROCHÁZKOVÁ, P., ŠILEROVÁ, M., STIJLEMANS, B., DIEU, M., HALADA, P., JOSKOVÁ, R., BESCHIN, A., DE BAETSELIER, P. & BILEJ, M. 2006. Evidence for proteins involved in prophenoloxidase cascade *Eisenia fetida* earthworms. *Journal of Comparative Physiology B*, 176, 581-587.
- PUTNAM, N. H., SRIVASTAVA, M., HELLSTEN, U., DIRKS, B., CHAPMAN, J., SALAMOV, A., TERRY, A., SHAPIRO, H., LINDQUIST, E., KAPITONOV, V. V., JURKA, J., GENIKHOVICH, G., GRIGORIEV, I. V., LUCAS, S. M., STEELE, R. E., FINNERTY, J. R., TECHNAU, U., MARTINDALE, M. Q. & ROKHSAR, D. S. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317, 86-94.

- RAHEMTULLA, F. & LØVTRUP, S. 1974. The comparative biochemistry of invertebrate mucopolysaccharides—II. nematoda; annelida. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 49, 639-646.
- RÄMET, M., PEARSON, A., BAKSA, K. & HARIKRISHNAN, A. 2003. Pattern Recognition Receptors in *Drosophila*. In: EZEKOWITZ, R. A. B. & HOFFMANN, J. A. (eds.) *Innate Immunity*. Totowa, NJ: Humana Press.
- RÄMET, M., PEARSON, A., MANFRUELLI, P., LI, X., KOZIEL, H., GÖBEL, V., CHUNG, E., KRIEGER, M. & EZEKOWITZ, R. A. B. 2001. *Drosophila* Scavenger Receptor C1 Is a Pattern Recognition Receptor for Bacteria. *Immunity*, 15, 1027-1038.
- RAST, J. P. & MESSIER-SOLEK, C. 2008. Marine Invertebrate Genome Sequences and Our Evolving Understanding of Animal Immunity. *The Biological Bulletin*, 214, 274-283.
- RAST, J. P., SMITH, L. C., LOZA-COLL, M., HIBINO, T. & LITMAN, G. W. 2006. Genomic Insights into the Immune System of the Sea Urchin. *Science*, 314, 952.
- RATCLIFFE, N. A., WHITE, K. N., ROWLEY, A. F. & WALTERS, J. B. 1982. Cellular Defense Systems of the Arthropoda. In: COHEN, N. & SIGEL, M. M. (eds.) *Phylogeny and Ontogeny*. Boston, MA: Springer US.
- RAUDVERE, U., KOLBERG, L., KUZMIN, I., ARAK, T., ADLER, P., PETERSON, H. & VILO, J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47, W191-W198.
- REIMAND, J., ISSERLIN, R., VOISIN, V., KUCERA, M., TANNUS-LOPES, C., ROSTAMIANFAR, A., WADI, L., MEYER, M., WONG, J., XU, C., MERICO, D. & BADER, G. D. 2019. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, 14, 482-517.
- RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13, 278-289.
- RIMMINGTON, O. J. 2018. *Modelling latent information as a survival mechanism in earthworm genomes*. Doctor of Philosophy, Cardiff University.
- ROACH, J. C., GLUSMAN, G., ROWEN, L., KAUR, A., PURCELL, M. K., SMITH, K. D., HOOD, L. E. & ADEREM, A. 2005. The evolution of vertebrate Toll-like receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 9577.
- ROEHR, J. T., DIETERICH, C. & REINERT, K. 2017. Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics*, 33, 2941-2942.
- ROWLEY, A. F. & POWELL, A. 2007. Invertebrate Immune Systems—Specific, Quasi-Specific, or Nonspecific? *The Journal of Immunology*, 179, 7209.

- ROY, R., KUMAR, D., SHARMA, A., GUPTA, P., CHAUDHARI, B. P., TRIPATHI, A., DAS, M. & DWIVEDI, P. D. 2014. ZnO nanoparticles induced adjuvant effect via toll-like receptors and Src signaling in Balb/c mice. *Toxicology Letters*, 230, 421-433.
- RUAN, J. & LI, H. 2019. Fast and accurate long-read assembly with wtdbg2. *bioRxiv*, 530972.
- SALATA, O. V. 2004. Applications of nanoparticles in biology and medicine. *Journal of Nanobiotechnology*, 2, 3.
- SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74, 5463.
- SAYOLS, S., SCHERZINGER, D. & KLEIN, H. 2016. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*, 17, 428.
- SBONER, A., MU, X. J., GREENBAUM, D., AUERBACH, R. K. & GERSTEIN, M. B. 2011. The real cost of sequencing: higher than you think! *Genome biology*, 12, 125-125.
- SCHMIDT, B. & HILDEBRANDT, A. 2017. Next-generation sequencing: big data meets high performance computing. *Drug Discov Today*, 22, 712-717.
- SCHULENBURG, H., KURTZ, J., MORET, Y. & SIVA-JOTHY, M. T. 2009. Introduction. Ecological immunology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 3-14.
- SCHULTZ, J. H. & ADEMA, C. M. 2017. Comparative immunogenomics of molluscs. *Developmental & Comparative Immunology*, 75, 3-15.
- SCHUSTER, S. C. 2008. Next-generation sequencing transforms today's biology. *Nature Methods*, 5, 16-18.
- SEKINE, R., MARZOUK, E. R., KHAKSAR, M., SCHECKEL, K. G., STEGEMEIER, J. P., LOWRY, G. V., DONNER, E. & LOMBI, E. 2017. Aging of Dissolved Copper and Copper-based Nanoparticles in Five Different Soils: Short-term Kinetics vs. Long-term Fate. *Journal of environmental quality*, 46, 1198-1205.
- SEPPEY, M., MANNI, M. & ZDOBNOV, E. M. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol*, 1962, 227-245.
- SHAKOR, A.-B. A., CZURYŁO, E. A. & SOBOTA, A. 2003. Lysenin, a unique sphingomyelin-binding protein. *FEBS Letters*, 542, 1-6.
- SHANG, L., NIENHAUS, K., JIANG, X., YANG, L., LANDFESTER, K., MAILÄNDER, V., SIMMET, T. & NIENHAUS, G. U. 2014a. Nanoparticle interactions with live cells: Quantitative fluorescence microscopy of nanoparticle size effects. *Beilstein Journal of Nanotechnology*, 5, 2388-2397.

- SHANG, L., NIENHAUS, K. & NIENHAUS, G. U. 2014b. Engineered nanoparticles interacting with cells: size matters. *Journal of Nanobiotechnology*, 12, 5.
- SHIOMI, S., KAWAMORI, M., YAGI, S. & MATSUBARA, E. 2015. One-pot synthesis of silica-coated copper nanoparticles with high chemical and thermal stability. *Journal of Colloid and Interface Science*, 460, 47-54.
- SHORT, S., ROBINSON, A., LAHIVE, E., GREEN ETXABE, A., HERNÁDI, S., GLÓRIA PEREIRA, M., KILLE, P., AND SPURGEON, D.J. 2021. Off-target stoichiometric binding identified from toxicogenomics explains why some species are more sensitive than others to a widely-used neonicotinoid. *Environmental Science & Technology* *Environmental Science & Technology (In press)*.
- ŠILEROVÁ, M., PROCHÁZKOVÁ, P., JOSKOVÁ, R., JOSENS, G., BESCHIN, A., DE BAETSELIER, P. & BILEJ, M. 2006. Comparative study of the CCF-like pattern recognition protein in different Lumbricid species. *Developmental & Comparative Immunology*, 30, 765-771.
- ŠÍMA, P. 1994. Annelid coelomocytes and haemocytes: roles in cellular immune reactions. *Immunology of Annelids*. CRC Press Boca Raton.
- SIMAO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.
- SIMONET, B. M. & VALCÁRCEL, M. 2008. Monitoring nanoparticles in the environment. *Analytical and Bioanalytical Chemistry*, 393, 17.
- ŠKANTA, F., PROCHÁZKOVÁ, P., ROUBALOVÁ, R., DVOŘÁK, J. & BILEJ, M. 2016. LBP/BPI homologue in *Eisenia andrei* earthworms. *Developmental & Comparative Immunology*, 54, 1-6.
- ŠKANTA, F., ROUBALOVÁ, R., DVOŘÁK, J., PROCHÁZKOVÁ, P. & BILEJ, M. 2013. Molecular cloning and expression of TLR in the *Eisenia andrei* earthworm. *Developmental & Comparative Immunology*, 41, 694-702.
- SLENTER, D. N., KUTMON, M., HANSPERS, K., RIUTTA, A., WINDSOR, J., NUNES, N., MÉLIUS, J., CIRILLO, E., COORT, S. L., DIGLES, D., EHRHART, F., GIESBERTZ, P., KALAFATI, M., MARTENS, M., MILLER, R., NISHIDA, K., RIESWIJK, L., WAAGMEESTER, A., EIJSSEN, L. M. T., EVELO, C. T., PICO, A. R. & WILLIGHAGEN, E. L. 2017. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46, D661-D667.
- SMIT AFA, H. R. 2008. *RepeatModeler* [Online]. Institute for Systems Biology. Available: <http://www.repeatmasker.org> [Accessed Repeatmodeler].
- SMITH, D. M., SIMON, J. K. & BAKER JR, J. R. 2013. Applications of nanotechnology for immunology. *Nature Reviews Immunology*, 13, 592-605.

- SRIVASTAVA, M., SIMAKOV, O., CHAPMAN, J., FAHEY, B., GAUTHIER, M. E. A., MITROS, T., RICHARDS, G. S., CONACO, C., DACRE, M., HELLSTEN, U., LARROUX, C., PUTNAM, N. H., STANKE, M., ADAMSKA, M., DARLING, A., DEGNAN, S. M., OAKLEY, T. H., PLACHETZKI, D. C., ZHAI, Y., ADAMSKI, M., CALCINO, A., CUMMINS, S. F., GOODSTEIN, D. M., HARRIS, C., JACKSON, D. J., LEYS, S. P., SHU, S., WOODCROFT, B. J., VERVOORT, M., KOSIK, K. S., MANNING, G., DEGNAN, B. M. & ROKHSAR, D. S. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466, 720.
- STÅHL, P. L., SALMÉN, F., VICKOVIC, S., LUNDMARK, A., NAVARRO, J. F., MAGNUSSON, J., GIACOMELLO, S., ASP, M., WESTHOLM, J. O., HUSS, M., MOLLBRINK, A., LINNARSSON, S., CODELUPPI, S., BORG, Å., PONTÉN, F., COSTEA, P. I., SAHLÉN, P., MULDER, J., BERGMANN, O., LUNDEBERG, J. & FRISÉN, J. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353, 78-82.
- STANKE, M., SCHÖFFMANN, O., MORGENSTERN, B. & WAACK, S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7, 62.
- STEBBING, A. 2002. Tolerance and hormesis - Increased resistance to copper in hydroids linked to hormesis. *Marine environmental research*, 54, 805-9.
- STEIN, E., AVTALION, R. R. & COOPER, E. L. 1977. The coelomocytes of the earthworm *Lumbricus terrestris*: morphology and phagocytic properties. *Journal of Morphology*, 153, 467-477.
- STRAMBEANU, N., DEMETROVICI, L. & DRAGOS, D. 2015a. Anthropogenic Sources of Nanoparticles. In: LUNGU, M., NECULAE, A., BUNOIU, M. & BIRIS, C. (eds.) *Nanoparticles' Promises and Risks: Characterization, Manipulation, and Potential Hazards to Humanity and the Environment*. Cham: Springer International Publishing.
- STRAMBEANU, N., DEMETROVICI, L., DRAGOS, D. & LUNGU, M. 2015b. Nanoparticles: Definition, Classification and General Physical Properties. 3-8.
- SUN, L., DONG, S., GE, Y., FONSECA, J. P., ROBINSON, Z. T., MYSORE, K. S. & MEHTA, P. 2019. DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists. *Frontiers in Genetics*, 10.
- SUPEK, F., BOŠNJAK, M., ŠKUNCA, N. & ŠMUC, T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, 6, e21800.
- SUTTON, H. C. & WINTERBOURN, C. C. 1989. On the participation of higher oxidation states of iron and copper in fenton reactions. *Free Radical Biology and Medicine*, 6, 53-60.
- ŚWIĄTEK, Z. M., WOŹNICKA, O. & BEDNARSKA, A. J. 2020. Unravelling the ZnO-NPs mechanistic pathway: Cellular changes and altered morphology in the gastrointestinal tract of the earthworm *Eisenia andrei*. *Ecotoxicology and Environmental Safety*, 196, 110532.
- SWIDERSKA, B., KEDRACKA-KROK, S., PANZ, T., MORGAN, A. J., FALNIOWSKI, A., GRZMIL, P. & PLYTYCZ, B. 2016. Lysenin family proteins in earthworm coelomocytes - Comparative approach.

- TAKEDA, K. & AKIRA, S. 2015. Toll-Like Receptors. *Current Protocols in Immunology*, 109, 14.12.1-14.12.10.
- TAKEDA, K., KAISHO, T. & AKIRA, S. 2003. Toll-Like Receptors. *Annual Review of Immunology*, 21, 335-376.
- TARAILO-GRAOVAC, M. & CHEN, N. 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, 25, 4.10.1-4.10.14.
- TARAZONA, S., FURIÓ-TARÍ, P., TURRÀ, D., PIETRO, A. D., NUEDA, M. J., FERRER, A. & CONESA, A. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic acids research*, 43, e140-e140.
- TAYLOR, E. & HEYLAND, A. 2017. Evolution of thyroid hormone signaling in animals: Non-genomic and genomic modes of action. *Molecular and Cellular Endocrinology*, 459, 14-20.
- TCHOUNWOU, P. B., NEWSOME, C., WILLIAMS, J. & GLASS, K. 2008. Copper-Induced Cytotoxicity and Transcriptional Activation of Stress Genes in Human Liver Carcinoma (HepG(2)) Cells. *Metal ions in biology and medicine : proceedings of the ... International Symposium on Metal Ions in Biology and Medicine held ... = Les ions metalliques en biologie et en medecine : ... Symposium international sur les ions metalliques ...* 10, 285-290.
- TENZER, S., DOCTER, D., KUHAREV, J., MUSYANOVYCH, A., FETZ, V., HECHT, R., SCHLENK, F., FISCHER, D., KIOUPTSI, K., REINHARDT, C., LANDFESTER, K., SCHILD, H., MASKOS, M., KNAUER, S. K. & STAUBER, R. H. 2013. Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nature Nanotechnology*, 8, 772-781.
- THE GENE ONTOLOGY CONSORTIUM 2018. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47, D330-D338.
- THOMAS, R. H. 2001. Molecular Evolution and Phylogenetics. Masatoshi Nei and Sudhir Kumar. Oxford University Press, Oxford. 2000. pp. 333. Price £65.00, hardback. ISBN 0 19 513584 9. *Heredity*, 86, 385-385.
- TOURINHO, P. S., VAN GESTEL, C. A. M., LOFTS, S., SVENDSEN, C., SOARES, A. M. V. M. & LOUREIRO, S. 2012. Metal-based nanoparticles in soil: Fate, behavior, and effects on soil invertebrates. *Environmental Toxicology and Chemistry*, 31, 1679-1692.
- TYNE, W., LITTLE, S., SPURGEON, D. J. & SVENDSEN, C. 2015. Hormesis depends upon the life-stage and duration of exposure: Examples for a pesticide and a nanomaterial. *Ecotoxicol Environ Saf*, 120, 117-23.
- UNGARO, A., PECH, N., MARTIN, J.-F., MCCAIRNS, R. J. S., MÉVY, J.-P., CHAPPAZ, R. & GILLES, A. 2017. Challenges and advances for transcriptome assembly in non-model species. *PLOS ONE*, 12, e0185020.
- UNIPROT 2019. UniProt: a worldwide hub of protein knowledge. *In*: CONSORTIUM, U. (ed.).

- URSO, E. & MAFFIA, M. 2015. Behind the Link between Copper and Angiogenesis: Established Mechanisms and an Overview on the Role of Vascular Copper Transport Systems. *Journal of Vascular Research*, 52, 172-196.
- VALEMBOIS, P., LASSÈGUES, M., ROCH, P. & VAILLIER, J. 1985. Scanning electron-microscopic study of the involvement of coelomic cells in earthworm antibacterial defense. *Cell and Tissue Research*, 240, 479-484.
- VALEMBOIS, P., SEYMOUR, J. & LASSÈGUES, M. 1994. Evidence of lipofuscin and melanin in the brown body of the earthworm *Eisenia fetida andrei*. *Cell and Tissue Research*, 277, 183-188.
- VAN DEN BRINK, N. W., JEMEC KOKALJ, A., SILVA, P. V., LAHIVE, E., NORRFORS, K., BACCARO, M., KHODAPARAST, Z., LOUREIRO, S., DROBNE, D., CORNELIS, G., LOFTS, S., HANDY, R. D., SVENDSEN, C., SPURGEON, D. & VAN GESTEL, C. A. M. 2019. Tools and rules for modelling uptake and bioaccumulation of nanomaterials in invertebrate organisms. *Environmental Science: Nano*, 6, 1985-2001.
- VAN DIJK, E. L., JASZCZYSZYN, Y., NAQUIN, D. & THERMES, C. 2018. The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34, 666-681.
- VAN SINAY, E., MIRABEAU, O., DEPUYDT, G., VAN HIEL, M. B., PEYMEN, K., WATTEYNE, J., ZELS, S., SCHOofs, L. & BEETS, I. 2017. Evolutionarily conserved TRH neuropeptide pathway regulates growth in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 114, E4065.
- VANHOOK, A. M. 2012. Copper as a Kinase Cofactor. *Science Signaling*, 5, ec84-ec84.
- VARET, H., BRILLET-GUÉGUEN, L., COPPÉE, J.-Y. & DILLIES, M.-A. 2016. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLOS ONE*, 11, e0157022.
- VASER, R., SOVIC, I., NAGARAJAN, N. & SIKIC, M. A.-O. 2017. Fast and accurate de novo genome assembly from long uncorrected reads.
- VITTURI, R., COLOMBA, M. S., PIRRONE, A. & LIBERTINI, A. 2000. Physical mapping of rDNA genes, (TTAGGG)_n telomeric sequence and other karyological features in two earthworms of the family Lumbricidae (Annelida: Oligochaeta). *Heredity (Edinb)*, 85 Pt 3, 203-7.
- WAALEWIJN-KOOL, P., DÍEZ-ORTIZ, M. & GESTEL, C. 2012. Effect of different spiking procedures on the distribution and toxicity of ZnO nanoparticles in soil. *Ecotoxicology (London, England)*, 21, 1797-804.
- WALKER, B. J., ABEEL, T., SHEA, T., PRIEST, M., ABOUELLIEL, A., SAKTHIKUMAR, S., CUOMO, C. A., ZENG, Q., WORTMAN, J., YOUNG, S. K. & EARL, A. M. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9, e112963.

- WANG, J. R., HOLT, J., MCMILLAN, L. & JONES, C. D. 2018. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, 19, 50.
- WANG, X. & LIU, Y. 2007. Regulation of innate immune response by MAP kinase phosphatase-1. *Cellular Signalling*, 19, 1372-1382.
- WATERHOUSE, R. M., SEPPEY, M., SIMÃO, F. A., MANNI, M., IOANNIDIS, P., KLIOUTCHNIKOV, G., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, 35, 543-548.
- WEBER, A., WASILIEW, P. & KRACHT, M. 2010. Interleukin-1 (IL-1) Pathway. *Science Signaling*, 3, cm1-cm1.
- WEISENFELD, N. I., KUMAR, V., SHAH, P., CHURCH, D. M. & JAFFE, D. B. 2017. Direct determination of diploid genome sequences. *Genome Res*, 27, 757-767.
- WICKHAM, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer International Publishing.
- WIENS, M., KORZHEV, M., PEROVIĆ-OTTSTADT, S., LUTHRINGER, B., BRANDT, D., KLEIN, S. & MÜLLER, W. E. G. 2007. Toll-Like Receptors Are Part of the Innate Immune Defense System of Sponges (Demospongiae: Porifera). *Molecular Biology and Evolution*, 24, 792-804.
- WOOD, E. J. 1983. Molecular cloning. A laboratory manual by T Maniatis, E F Fritsch and J Sambrook. pp 545. Cold Spring Harbor Laboratory, New York. 1982. \$48 ISBN 0-87969-136-0. *Biochemical Education*, 11, 82-82.
- XU, G. C., XU, T. J., ZHU, R., ZHANG, Y., LI, S. Q., WANG, H. W. & LI, J. T. 2019. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience*, 8.
- XUE, W., LI, J.-T., ZHU, Y.-P., HOU, G.-Y., KONG, X.-F., KUANG, Y.-Y. & SUN, X.-W. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics*, 14, 604.
- ZHANG, Q., ZMASEK, C. M. & GODZIK, A. 2010. Domain architecture evolution of pattern-recognition receptors. *Immunogenetics*, 62, 263-272.
- ZHANG, Y. & CAO, X. 2016. Long noncoding RNAs in innate immunity. *Cellular & Molecular Immunology*, 13, 138-147.
- ZHAO, H., SHAO, D., JIANG, C., SHI, J., LI, Q., HUANG, Q., RAJOKA, M. S. R., YANG, H. & JIN, M. 2017. Biological activity of lipopeptides from Bacillus. *Applied Microbiology and Biotechnology*, 101, 5951-5960.
- ZIMIN, A. V., MARÇAIS, G., PUIU, D., ROBERTS, M., SALZBERG, S. L. & YORKE, J. A. 2013. The MaSuRCA genome assembler. *Bioinformatics*, 29, 2669-2677.

ZIPFEL, C. 2014. Plant pattern-recognition receptors. *Trends in Immunology*, 35, 345-351.

ZOLNIK, B. S., GONZALEZ-FERNANDEZ A FAU - SADRIEH, N., SADRIEH N FAU - DOBROVOLSKAIA, M. A. & DOBROVOLSKAIA, M. A. Nanoparticles and the immune system.

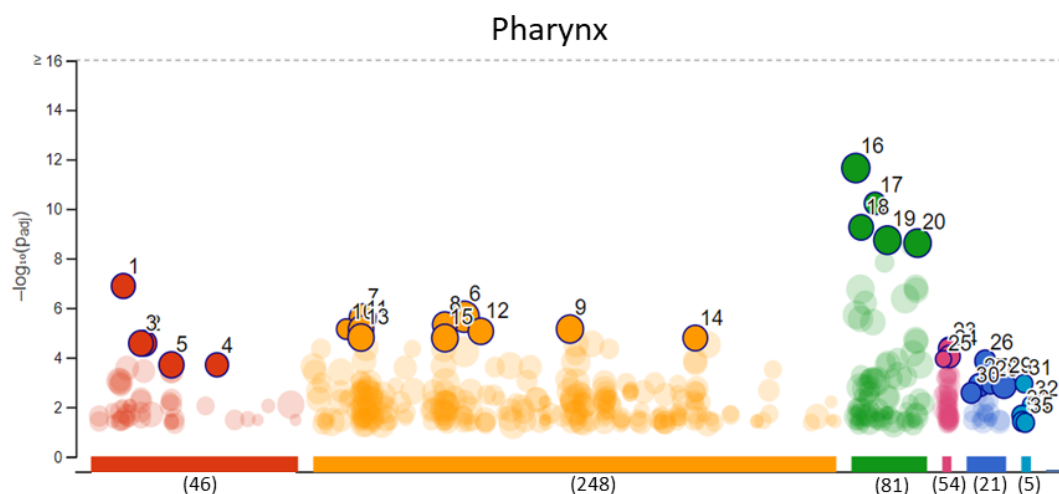
ZOLNIK, B. S., GONZÁLEZ-FERNÁNDEZ, A. F., SADRIEH, N. & DOBROVOLSKAIA, M. A. 2010. Minireview: Nanoparticles and the Immune System. *Endocrinology*, 151, 458-465.

ZWARYCZ, A. S., NOSSA, C. W., PUTNAM, N. H. & RYAN, J. F. 2015. Timing and Scope of Genomic Expansion within Annelida: Evidence from Homeoboxes in the Genome of the Earthworm *Eisenia fetida*. *Genome Biology and Evolution*, 8, 271-281.

9 Appendices

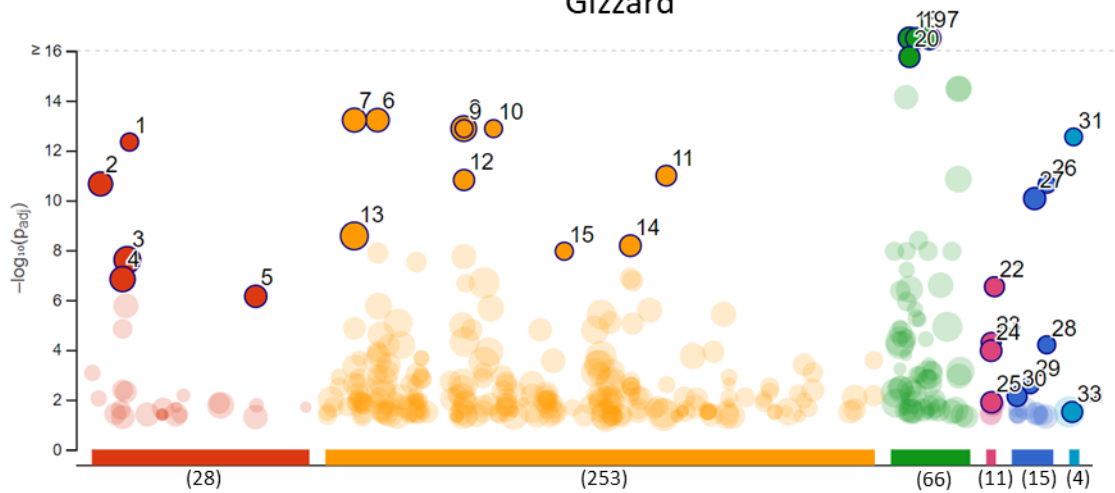
Appendix - Chapter 2

2.1.1 Over-representation analysis of the tissue-specific genes



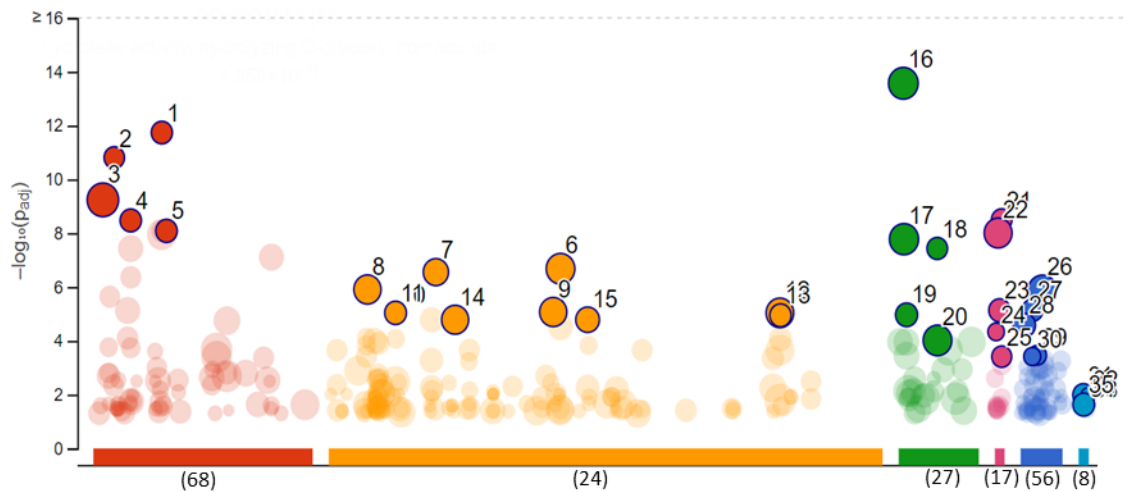
ID	Term name	Source	Negative log10 FDR
1	calcium ion binding	GO:MF	6.9
2	inorganic molecular entity transmembrane transporter activity	GO:MF	4.6
3	ion transmembrane transporter activity	GO:MF	4.6
4	metal ion transmembrane transporter activity	GO:MF	3.7
5	transmembrane transporter activity	GO:MF	3.7
6	multicellular organismal process	GO:BP	5.6
7	movement of cell or subcellular component	GO:BP	5.6
8	metal ion transport	GO:BP	5.3
9	cell development	GO:BP	5.1
10	cilium movement	GO:BP	5.1
11	cation transport	GO:BP	5.1
12	ion transmembrane transport	GO:BP	5.1
13	ion transport	GO:BP	4.8
14	inorganic ion transmembrane transport	GO:BP	4.8
15	cell projection organization	GO:BP	4.8
16	extracellular region	GO:CC	11.6
17	motile cilium	GO:CC	10.2
18	cilium	GO:CC	9.3
19	cell projection	GO:CC	8.7
20	plasma membrane bounded cell projection	GO:CC	8.6
21	Gastric acid secretion	KEGG	4.4
22	Salivary secretion	KEGG	4.4
23	Glucagon signaling pathway	KEGG	4.4
24	Proteoglycans in cancer	KEGG	4.0
25	Mucin type O-glycan biosynthesis	KEGG	3.9
26	Ion channel transport	REAC	3.9
27	Extracellular matrix organization	REAC	2.9
28	O-linked glycosylation of mucins	REAC	2.9
29	Transport of small molecules	REAC	2.9
30	Cardiac conduction	REAC	2.6
31	Glycolysis and Gluconeogenesis	WP	2.9
32	Computational Model of Aerobic Glycolysis	WP	2.1
33	Spinal Cord Injury	WP	1.7
34	Alzheimers Disease	WP	1.4
35	Notch Signaling Pathway Netpath	WP	1.4

Gizzard



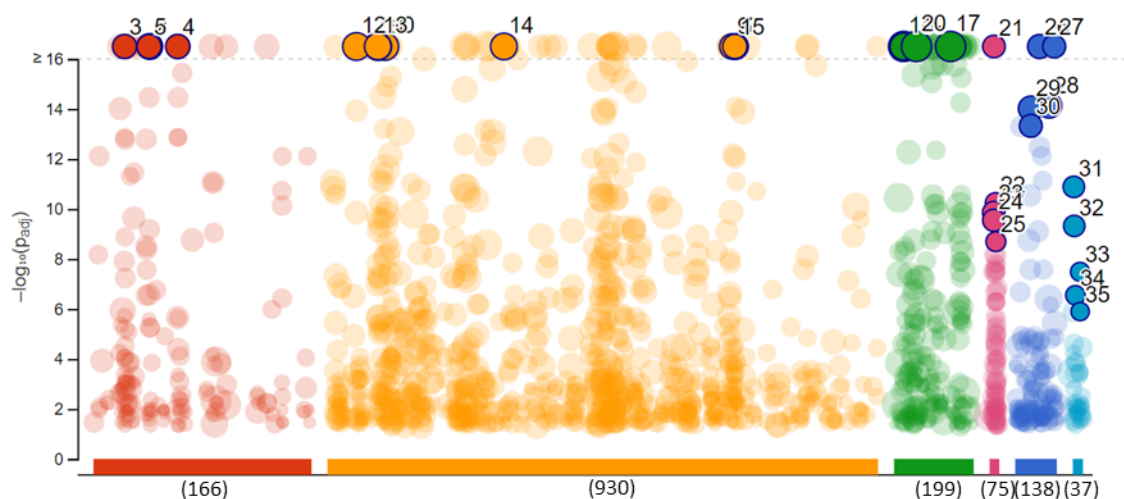
ID	Term name	Source	Negative log10 FDR
1	structural constituent of muscle	GO:MF	12.3
2	actin binding	GO:MF	10.7
3	cytoskeletal protein binding	GO:MF	7.6
4	structural molecule activity	GO:MF	6.8
5	actin filament binding	GO:MF	6.1
6	muscle contraction	GO:BP	13.2
7	muscle system process	GO:BP	13.2
8	actin filament-based process	GO:BP	12.9
9	muscle filament sliding	GO:BP	12.9
10	actin-myosin filament sliding	GO:BP	12.9
11	actin-mediated cell contraction	GO:BP	11.0
12	actin filament-based movement	GO:BP	10.8
13	system process	GO:BP	8.6
14	muscle cell development	GO:BP	8.2
15	sarcomere organization	GO:BP	8.0
16	myofibril	GO:CC	24.2
17	contractile fiber	GO:CC	24.0
18	sarcomere	GO:CC	23.4
19	I band	GO:CC	16.1
20	Z disc	GO:CC	15.7
21	Hypertrophic cardiomyopathy (HCM)	KEGG	6.5
22	Dilated cardiomyopathy (DCM)	KEGG	6.5
23	Cardiac muscle contraction	KEGG	4.3
24	Adrenergic signaling in cardiomyocytes	KEGG	4.0
25	Tight junction	KEGG	1.9
26	Muscle contraction	REAC	10.1
27	Smooth Muscle Contraction	REAC	4.2
28	Interaction between L1 and Ankyrins	REAC	2.6
29	Cell-Cell communication	REAC	2.1
30	Cell junction organization	REAC	1.6
31	Striated Muscle Contraction Pathway	WP	12.5
32	Arrhythmogenic Right Ventricular Cardiomyopathy	WP	1.5
33	Myometrial Relaxation and Contraction Pathways	WP	1.5

Gut - Chloragogen

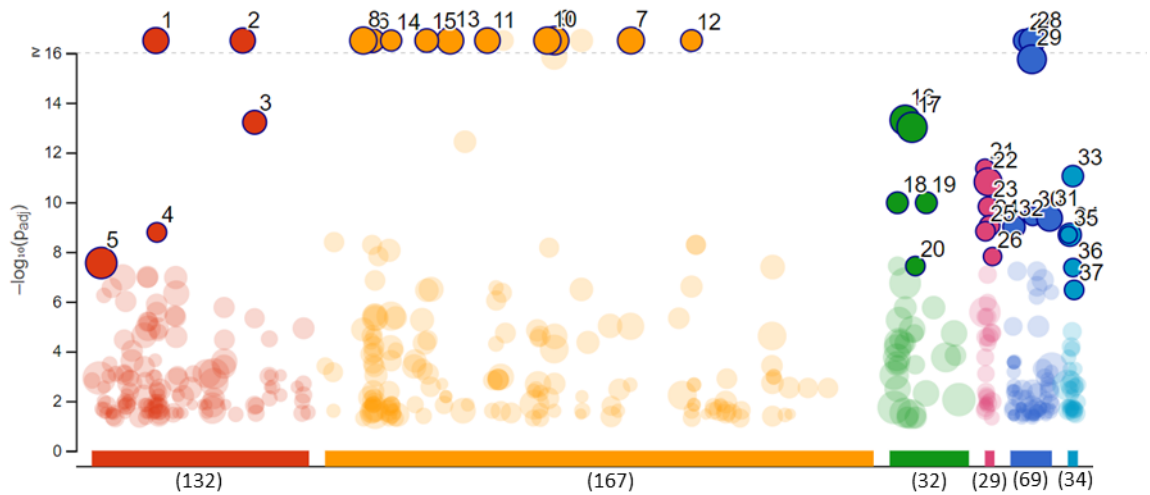


ID	Term name	Source	Negative log10 FDR
1	hydrolase activity, acting on glycosyl bonds	GO:MF	11.7
2	hydrolase activity, hydrolyzing O-glycosyl compounds	GO:MF	10.8
3	catalytic activity	GO:MF	9.2
4	serine-type peptidase activity	GO:MF	8.5
5	serine hydrolase activity	GO:MF	8.1
6	small molecule metabolic process	GO:BP	6.7
7	drug metabolic process	GO:BP	6.6
8	organic acid metabolic process	GO:BP	5.9
9	oxoacid metabolic process	GO:BP	5.1
10	aromatic amino acid family catabolic process	GO:BP	5.0
11	cellular amino acid catabolic process	GO:BP	5.0
12	organonitrogen compound catabolic process	GO:BP	5.0
13	alpha-amino acid catabolic process	GO:BP	4.9
14	carboxylic acid metabolic process	GO:BP	4.8
15	carboxylic acid catabolic process	GO:BP	4.8
16	extracellular region	GO:CC	13.6
17	extracellular space	GO:CC	7.8
18	lysosomal lumen	GO:CC	7.4
19	vacuolar lumen	GO:CC	5.0
20	extracellular organelle	GO:CC	4.0
21	Salivary secretion	KEGG	3.1
22	Galactose metabolism	KEGG	2.6
23	Carbohydrate digestion and absorption	KEGG	1.9
24	Renin-angiotensin system	KEGG	1.7
25	Phenylalanine metabolism	KEGG	1.7
26	Metabolism	REAC	5.9
27	Diseases of metabolism	REAC	5.1
28	Biological oxidations	REAC	4.6
29	Glycosphingolipid metabolism	REAC	3.5
30	Digestion	REAC	3.4
31	Amino Acid metabolism	WP	2.0
32	Amino Acid Metabolism Pathway Excerpt	WP	2.0
33	Tyrosine Metabolism	WP	1.6
34	NAD Biosynthesis II (from tryptophan)	WP	1.6
35	Metapathway biotransformation Phase I and II	WP	1.6

Nerve cord

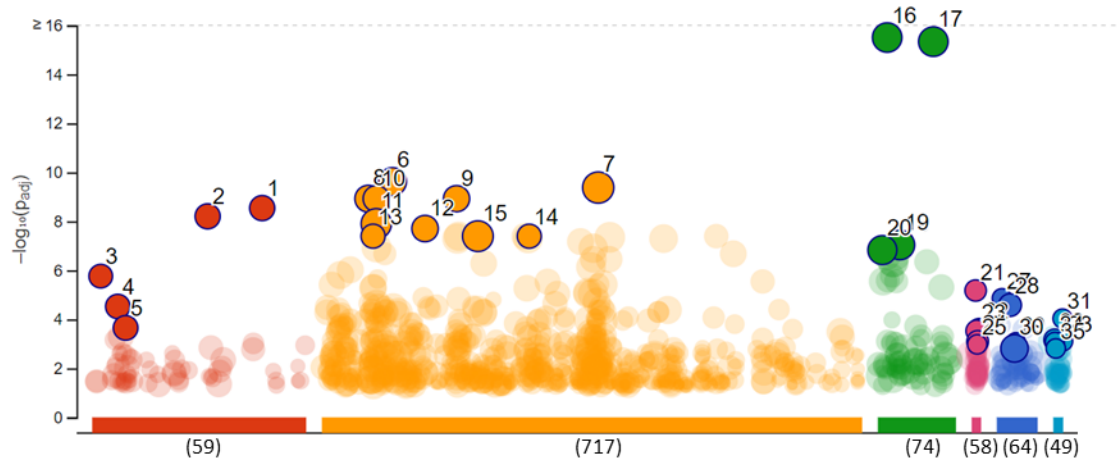


ID	Term name	Source	Negative log10 FDR
1	gated channel activity	GO:MF	32.25
2	inorganic molecular entity transmembrane transporter activity	GO:MF	32.12
3	ion channel activity	GO:MF	32.04
4	passive transmembrane transporter activity	GO:MF	31.89
5	channel activity	GO:MF	31.36
6	synaptic signaling	GO:BP	43.65
7	trans-synaptic signaling	GO:BP	43.65
8	chemical synaptic transmission	GO:BP	42.41
9	anterograde trans-synaptic signaling	GO:BP	42.41
10	cell-cell signaling	GO:BP	36.61
11	ion transport	GO:BP	36.61
12	system process	GO:BP	33.32
13	cation transport	GO:BP	32.63
14	ion transmembrane transport	GO:BP	32.48
15	regulation of trans-synaptic signaling	GO:BP	28.80
16	plasma membrane	GO:CC	77.17
17	cell periphery	GO:CC	77.14
18	integral component of plasma membrane	GO:CC	65.36
19	intrinsic component of plasma membrane	GO:CC	64.09
20	intrinsic component of membrane	GO:CC	53.53
21	Neuroactive ligand-receptor interaction	KEGG	25.52
22	Serotonergic synapse	KEGG	10.24
23	cAMP signaling pathway	KEGG	9.85
24	Calcium signaling pathway	KEGG	9.55
25	Insulin secretion	KEGG	8.69
26	Neuronal System	REAC	33.55
27	Transmission across Chemical Synapses	REAC	18.45
28	Signaling by GPCR	REAC	14.14
29	GPCR downstream signalling	REAC	14.02
30	GPCR ligand binding	REAC	13.32
31	Sudden Infant Death Syndrome (SIDS) Susceptibility Pathways	WP	10.88
32	GPCRs, Class A Rhodopsin-like	WP	9.32
33	Peptide GPCRs	WP	7.48
34	Synaptic Vesicle Pathway	WP	6.55
35	Splicing factor NOVA regulated synaptic proteins	WP	5.89



ID	Term name	Source	Negative log10 FDR
1	oxidoreductase activity	GO:MF	21.3
2	cofactor binding	GO:MF	19.6
3	coenzyme binding	GO:MF	13.2
4	oxidoreductase activity, acting on the CH-CH group of donors	GO:MF	8.8
5	catalytic activity	GO:MF	7.6
6	fatty acid metabolic process	GO:BP	25.7
7	oxidation-reduction process	GO:BP	23.7
8	organic acid metabolic process	GO:BP	23.1
9	small molecule metabolic process	GO:BP	22.9
10	oxoacid metabolic process	GO:BP	22.3
11	monocarboxylic acid metabolic process	GO:BP	22.3
12	monocarboxylic acid catabolic process	GO:BP	21.3
13	carboxylic acid metabolic process	GO:BP	21.2
14	fatty acid catabolic process	GO:BP	21.1
15	lipid catabolic process	GO:BP	20.5
16	integral component of membrane	GO:CC	13.3
17	intrinsic component of membrane	GO:CC	13.0
18	peroxisome	GO:CC	10.0
19	microbody	GO:CC	10.0
20	microbody lumen	GO:CC	7.4
21	Fatty acid degradation	KEGG	11.4
22	Metabolic pathways	KEGG	10.8
23	Fatty acid metabolism	KEGG	9.8
24	Peroxisome	KEGG	9.0
25	Valine, leucine and isoleucine degradation	KEGG	8.8
26	Fatty acid metabolism	REAC	23.9
27	Metabolism of lipids	REAC	16.1
28	Metabolism	REAC	15.7
29	Mitochondrial Fatty Acid Beta-Oxidation	REAC	9.4
30	Transport of small molecules	REAC	9.3
31	Metapathway biotransformation Phase I and II	WP	11.0
32	Nuclear Receptors Meta-Pathway	WP	8.7
33	Mitochondrial LC-Fatty Acid Beta-Oxidation	WP	8.7
34	Fatty Acid Beta Oxidation	WP	7.4
35	Oxidation by Cytochrome P450	WP	6.5

Coelomic fluid



ID	Term name	Source	Negative log10 FDR
1	molecular transducer activity	GO:MF	8.5
2	signaling receptor activity	GO:MF	8.2
3	actin binding	GO:MF	5.8
4	transmembrane signaling receptor activity	GO:MF	4.5
5	calcium ion binding	GO:MF	3.7
6	response to external stimulus	GO:BP	9.6
7	response to stimulus	GO:BP	9.4
8	lipid metabolic process	GO:BP	8.9
9	biological adhesion	GO:BP	8.9
10	cell adhesion	GO:BP	8.9
11	cell communication	GO:BP	7.9
12	cell migration	GO:BP	7.7
13	chemotaxis	GO:BP	7.4
14	taxis	GO:BP	7.4
15	multicellular organismal process	GO:BP	7.4
16	plasma membrane	GO:CC	15.5
17	cell periphery	GO:CC	15.3
18	intrinsic component of plasma membrane	GO:CC	7.0
19	intrinsic component of membrane	GO:CC	7.0
20	extracellular region	GO:CC	6.8
21	Rap1 signaling pathway	KEGG	5.2
22	Proteoglycans in cancer	KEGG	3.6
23	Focal adhesion	KEGG	3.5
24	Regulation of actin cytoskeleton	KEGG	3.1
25	B cell receptor signaling pathway	KEGG	3.0
26	Cell-extracellular matrix interactions	REAC	4.9
27	Cell-Cell communication	REAC	4.9
28	Extracellular matrix organization	REAC	4.6
29	Interleukin-4 and Interleukin-13 signaling	REAC	3.1
30	Immune System	REAC	2.8
31	Melanoma	WP	4.0
32	Focal Adhesion	WP	3.2
33	Cholesterol metabolism	WP	3.1
34	Primary Focal Segmental Glomerulosclerosis FSGS	WP	3.1
35	RAC1/PAK1/p38/MMP2 Pathway	WP	2.8

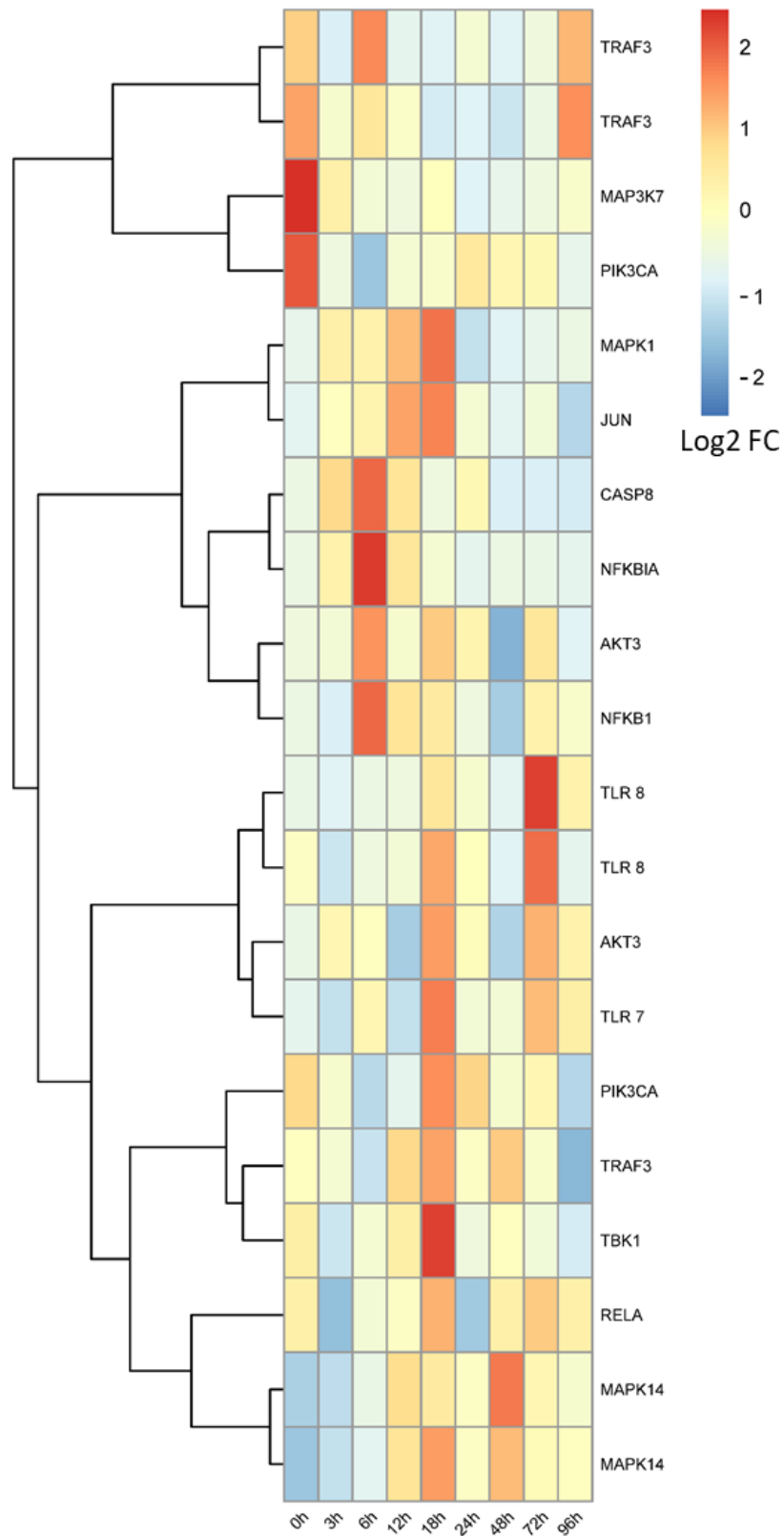
Appendix chapter 5

5.1 KEGG mapping table

UniProt ID	KEGG ID	Gene symbol
Q13158	hsa:8772	FADD; Fas associated via death domain
P31749	hsa:207	AKT1; AKT serine/threonine kinase 1
Q9Y243	hsa:10000	AKT3; AKT serine/threonine kinase 3
Q14790	hsa:841	CASP8; caspase 8
O15111	hsa:1147	CHUK; component of inhibitor of nuclear factor kappa B kinase complex
P43235	hsa:1513	CTSK; cathepsin K
P01100	hsa:2353	FOS; Fos proto-oncogene, AP-1 transcription factor subunit
Q14164	hsa:9641	IKBKE; inhibitor of nuclear factor kappa B kinase subunit epsilon
P51617	hsa:3654	IRAK1; interleukin 1 receptor associated kinase 1
Q9NWZ3	hsa:51135	IRAK4; interleukin 1 receptor associated kinase 4
Q92985	hsa:3665	IRF7; interferon regulatory factor 7
P05412	hsa:3725	JUN; Jun proto-oncogene, AP-1 transcription factor subunit
P18428	hsa:3929	LBP; lipopolysaccharide binding protein
Q02750	hsa:5604	MAP2K1; mitogen-activated protein kinase kinase 1
P45985	hsa:6416	MAP2K4; mitogen-activated protein kinase kinase 4
P52564	hsa:5608	MAP2K6; mitogen-activated protein kinase kinase 6
O14733	hsa:5609	MAP2K7; mitogen-activated protein kinase kinase 7
O43318	hsa:6885	MAP3K7; mitogen-activated protein kinase kinase kinase 7
P41279	hsa:1326	MAP3K8; mitogen-activated protein kinase kinase kinase 8
P28482	hsa:5594	MAPK1; mitogen-activated protein kinase 1
P53779	hsa:5602	MAPK10; mitogen-activated protein kinase 10

Q16539	hsa:1432	MAPK14; mitogen-activated protein kinase 14
P45983	hsa:5599	MAPK8; mitogen-activated protein kinase 8
Q99836	hsa:4615	MYD88; MYD88 innate immune signal transduction adaptor
P19838	hsa:4790	NFKB1; nuclear factor kappa B subunit 1
P25963	hsa:4792	NFKBIA; NFKB inhibitor alpha
P42336	hsa:5290	PIK3CA; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
P42338	hsa:5291	PIK3CB; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta
O00329	hsa:5293	PIK3CD; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta
P27986	hsa:5295	PIK3R1; phosphoinositide-3-kinase regulatory subunit 1
P63000	hsa:5879	RAC1; Rac family small GTPase 1
Q04206	hsa:5970	RELA; RELA proto-oncogene, NF-kB subunit
Q15750	hsa:10454	TAB1; TGF-beta activated kinase 1 (MAP3K7) binding protein 1
Q9NYJ8	hsa:23118	TAB2; TGF-beta activated kinase 1 (MAP3K7) binding protein 2
Q9UHD2	hsa:29110	TBK1; TANK binding kinase 1
Q86XR7	hsa:353376	TICAM2; toll like receptor adaptor molecule 2
Q15399	hsa:7096	TLR1; toll like receptor 1
O60603	hsa:7097	TLR2; toll like receptor 2
O00206	hsa:7099	TLR4; toll like receptor 4
Q9Y2C9	hsa:10333	TLR6; toll like receptor 6
Q9NYK1	hsa:51284	TLR7; toll like receptor 7
Q9NR97	hsa:51311	TLR8; toll like receptor 8
Q9NR96	hsa:54106	TLR9; toll like receptor 9
Q9H0E2	hsa:54472	TOLLIP; toll interacting protein
Q13114	hsa:7187	TRAF3; TNF receptor associated factor 3
Q9Y4K3	hsa:7189	TRAF6; TNF receptor associated factor 6

5.2 Expression profile of the components of 'Toll-like receptor signalling' pathway in granular amoebocytes.



Temporal analysis expression changes in granular amoebocytes under the combined bacterial (B. subtilis) and CuNP challenge. Differentially expressed genes associated with the “Toll-like receptor signalling” KEGG pathway were evaluated across the full-time course of analysis. Changes in expression were calculated against time-matched controls (no CuNPs or bacterial challenge). Hierarchical clustering was performed in R using pheatmap (Kolde 2012) with the Pearson correlation based distance measure option.