

**Maintenance of genetic and
physiological stability in
*Caenorhabditis elegans***

Yannic Chen

School of Molecular and Cellular Biology
The University of Leeds

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

January 2020

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Part of the experimental work described in Chapter 4 of the thesis is included in the publication below and is directly attributable to the candidate.

Pokhrel, B., Chen, Y. & Biro, J. J. (2019) CFP-1 interacts with HDAC1/2 complexes in *C. elegans* development. FEBS J, 286: 2490-2504.
doi:10.1111/febs.14833.

The candidate performed the RNAi sensitivity assay of the COMPASS mutants *set-2(bn129)* and *cfp-1(tm6369)*. Bharat Pokhrel performed most of the work, devised the layout and wrote the manuscript. Jonathan Biro performed heat shock reporter assay.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

The right of Yannic Chen to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

I am grateful to the University of Leeds for funding my PhD without which this project would have been impossible. I would like to express my deep gratitude to my supervisor Professor David Westhead and co-supervisor Dr Patricija van Oosten-Hawle for their professional supervision and consistent guidance. My special thanks goes to Dr Ron Chen for initiating and giving me the opportunity to research this PhD project, as well as supervising me for the first half of the PhD program.

I am very thankful to my friends Bharat Pokhrel and Dovile Milonaityte for their academic support and their friendship that made my time at Leeds much more enjoyable. My sincere thanks also goes to Dr Laura Jones for her friendly support throughout the second part of the project. I would also like to thank my fellow labmates Rosamund Clifford, Laura Warwick and Hann Ng for their technical support and uplifting the working environment. My sincere appreciation goes to Jeanne Rivera and Dr Edwin Chen for their advice and help. I also want to mention my appreciation to LIDA for their help regarding the Bioinformatics aspect of the project.

Last but not least, I am particularly grateful to my father and my mother for their continuous support throughout my life, that allowed me to be where I am today. I give my deep thanks to Wei Li for all her love and support.

我特别感谢我父母不断地支持才能让我有读博的机会

Abstract

The genetic and physiological stability of an organism is essential to ensure its well-being and survival. This project investigates the effect of epigenetic changes on genetic stability, and the association between different stressors that threaten the physiological stability of *C. elegans*.

Part 1:

Gene expression is controlled by epigenetic effects such as DNA methylation and histone modification. The histone modification H3K4me3 is associated with actively transcribed genes and co-localizes with a DNA:RNA hybrid structure known as R-loops which are associated with DNA instability. To investigate the link between H3K4me3 and R-loops, I use *Caenorhabditis elegans* COMPASS mutants *set-2(bn129)* and *cfp-1(tm6369)*, that have drastically reduced global H3K4me3 marks. I found that *set-2(bn129)* has a consistent reduction of R-loop levels compared to wild-type worms, suggesting that SET-2 (or H3K4me3) is vital in sustaining the R-loop levels observed in wild-type worms. Furthermore, seven helicases have been identified to rescue the R-loop levels in *set-2(bn129)* mutants, four of which are chromatin remodelers, suggesting a link between chromatin remodelling and R-loop aggregation.

Part 2:

Environmental stress is a common influence that threatens the health of an organism. While different stressors elicit different responses, how these different responses are interconnected is not well understood. To investigate this, I use a bioinformatic approach to compare the response of *C. elegans* under heat stress and biotic stress inflicted by pathogen infection. Comparison of transcriptomic data from *C. elegans* infected by different pathogens indicates an overall dissimilar gene expression response. However, a small set of “general pathogen responsive genes” are consistently differentially expressed at a low level under most pathogen infections. Comparing these general pathogen responsive genes with heat shock responsive genes identified a significant overlap of 50 genes. This suggests that the heat shock response and innate immune response partially overlap.

Table of Contents

<i>Acknowledgements</i>	<i>ii</i>
<i>Abstract</i>	<i>iii</i>
<i>List of Figures</i>	<i>ix</i>
<i>List of Tables</i>	<i>xii</i>
<i>List of Abbreviations</i>	<i>xiii</i>
Part 1: Investigating the functional relationship between R-loops and the evolutionarily conserved COMPASS complex	
Chapter 1 Introduction of epigenetics	2
1.1. Mechanisms of epigenetics	2
1.1.1. DNA methylation.....	3
1.1.2. Chromatin.....	4
1.1.3. Histone modification	6
1.2. COMPASS complex and H3K4me3	10
1.2.1. The COMPASS complex.....	10
1.2.2. The H3K4me3 epigenetic mark and its function	12
1.3. R-loops	14
1.3.1. Formation of R-loops.....	15
1.3.2. Preventing R-loop formation.....	17
1.3.3. Structure and stability	18
1.3.4. Resolution of R-loops.....	18
1.4. Function and effect of R-loops	20
1.4.1. Transcription regulation of R-loops	20
1.4.2. R-loop dependent DNA methylation state.....	21
1.4.3. Histone modifications and chromatin compaction.....	21
1.4.4. DNA instability.....	24
1.4.5. DNA repair	25
1.4.6. Health implications	25
1.5. Helicases	26
1.5.1. Types of helicases.....	27
1.5.2. Known helicases that resolve R-loops.....	31
1.5.3. Other functions of helicases that can affect R-loops.....	32
1.6. Aim and objectives	35
Chapter 2: Methods for the epigenetic analysis	36
2.1. Basic maintenance	36
2.1.1. List of strains.....	36
2.1.2. Nematode Growth Medium (NGM) plate preparation	36
2.1.3. OP50 maintenance.....	37

2.1.4. Worm maintenance.....	37
2.1.5. M9 buffer preparation.....	37
2.1.6. Liquid culture bleach preparation.....	37
2.1.7. Worm bleaching and synchronization.....	37
2.1.8. Tris-buffered Saline (TBS) and TBS + Tween (TBST) preparation.....	38
2.2. Genotyping and outcrossing.....	38
2.2.1. List of primers.....	38
2.2.2. Genotyping and single worm PCR.....	39
2.2.3. Outcrossing.....	40
2.3. RNA interference.....	40
2.3.1. RNAi LB culture plate preparation.....	40
2.3.2. Liquid RNAi LB preparation.....	41
2.3.3. RNAi bacteria feeding plates preparation.....	41
2.4. Developmental assay.....	41
2.5. RNAi sensitivity assay.....	41
2.5.1. Dumpy phenotype.....	42
2.5.2. Uncoordinated phenotype.....	42
2.5.3. <i>lin-1</i> phenotype.....	42
2.5.4. <i>bmr-1</i> phenotype.....	42
2.6. Sample collection for staging and genomic DNA (gDNA) extraction.....	43
2.6.1. Young adult collection.....	43
2.6.2. Late embryo collection.....	43
2.6.3. Staging.....	44
2.6.4. L1 Worm Collection.....	44
2.7. DNA extraction.....	45
2.8. R-loop slot blot.....	46
2.8.1. Membrane blotting.....	46
2.8.2. Membrane development.....	48
2.8.3. DNA loading control.....	48
2.9. Computational analysis of H3K4me3 ChIP-seq.....	49
2.9.1. Identifying differential H3K4me3 signal in COMPASS mutants.....	49
2.9.2. Hypergeometric testing.....	49
2.9.3. Average H3K4me3 signal change around TSS (SeqPlot).....	50
2.9.4. Gene list analysis.....	50
2.9.5. Motif discovery.....	50
Chapter 3: Bioinformatic analysis of H3K4me3 levels in COMPASS mutants.....	51
3.1. Introduction.....	51
3.2. <i>set-2(bn129)</i> and <i>cfp-1(tm6369)</i> mutants have reduced H3K4me3 levels.....	51
3.3. Loss of H3K4me3 signal is associated with housekeeping genes.....	54
3.4. Gain of H3K4me3 signal is associated with developmental and chromatin genes.....	56
3.5. Motif discovery and nucleotide frequency analysis.....	58

3.6.	Discussion.....	63
Chapter 4: R-loop levels in COMPASS mutants.....		66
4.1.	Introduction	66
4.2.	Characterisation of R-loops in the COMPASS mutants	66
4.2.1.	Dot/Slot blot method optimisation.....	67
4.2.2.	The R-loop signal from adult worms show large variation	72
4.2.3.	Late embryo and L1 show reduced R-loop signal in <i>set-2(bn129)</i> mutants compared to wild-type worms.....	73
4.2.4.	OP50 and EV diet results in the same R-loop pattern between Wild-type and <i>set-2(bn129)</i> mutants	75
4.2.5.	Discussion.....	76
4.3.	Development of antibody-independent R-loop purification	79
4.3.1.	R-loop purification by nuclease digestion and mass separation.....	79
4.3.2.	dsDNAse endonuclease digestion.....	81
4.3.3.	Restriction enzyme cutting	84
4.3.4.	Buffer exchange	85
4.3.5.	Discussion and future work	87
4.4.	Effect of various helicases on R-loops formation	88
4.4.1.	<i>C. elegans</i> helicase mutants <i>rba-1(tm329)</i> and <i>rcq-5(ok660)</i> both have increased R-loop accumulation	89
4.4.2.	<i>set-2(bn129)</i> and <i>cfp-1(tm6369)</i> are mildly resistant to specific RNAi bacteria	91
4.4.3.	<i>set-2(bn129)</i> R-loop suppressor screen with helicase genes	94
4.4.4.	Discussion.....	95
4.5.	Conclusion and future work of the epigenetic study	102
Part 2: Bioinformatic analysis to investigate the association between the innate immune response and heat shock response in <i>C. elegans</i>		
Chapter 5: Introduction of stress response and innate immunity in <i>C. elegans</i>		105
5.1.	Oxidative stress response	106
5.2.	Unfolded protein response.....	107
5.3.	Heat shock response.....	109
5.3.1.	Crosstalk between the HSR and other stress response pathways.....	110
5.3.2.	Transcellular chaperone signalling	112
5.4.	Innate immunity in <i>C. elegans</i>	113
5.4.1.	Physical defence against pathogens.....	114
5.4.2.	Innate immune response.....	115
5.5.	<i>C. elegans</i> pathogens investigated in this study.....	122
5.5.1.	Gram-positive bacteria	123
5.5.2.	Gram-negative bacteria	128
5.5.3.	Fungi.....	135
5.5.4.	Orsay Virus.....	139
5.6.	Aim and objective	140

Chapter 6: Methods for the bioinformatic analysis of high-throughput datasets	141
6.1. Data selection	141
6.2. Data processing	141
6.2.1. Affymetrix Microarray	141
6.2.2. Agilent Microarray	142
6.2.3. Nimblegen Microarray	142
6.2.4. RNA-seq Data	142
6.2.5. ChIP-seq Data	143
6.3. Meta-analysis/systematic review	144
6.3.1. Identifying common differentially expressed genes	144
6.3.2. K-means clustering	145
6.3.3. Determining p-value of filtering Criteria	146
6.3.4. Gene Enrichment Analysis	146
6.3.5. TF binding around the transcript Start Site (SeqPlot)	147
6.3.6. Hypergeometric test	147
6.4. Motif enrichment	148
Chapter 7: The association of the heat shock and innate immune response	149
7.1. Data used and quality control	149
7.1.1. Microarray data	150
7.1.2. RNA-seq data	152
7.2. Change of the gene expression landscape as a result of pathogen infection	154
7.2.1. Pathogen infection datasets show low overlapping up- and down-regulated genes	155
7.2.2. A small set of genes is consistently differentially expressed under various pathogens	159
7.2.3. Expression of immune effector protein after pathogen infection	177
7.2.4. Discussion	179
7.3. Transcription factors related to general pathogen response	181
7.4. Change of the gene expression landscape as a result of heat shock	190
7.4.1. Differentially expressed genes are consistent across the heat shock datasets	190
7.4.2. Heat shock responsive genes are related to the immune system	195
7.5. Comparison of pathogen response genes and heat shock response genes	197
7.6. Conclusion and future work of the stress resistance study	202
References	204
Appendix 1	231
Appendix 2	232
Appendix 3	233
Appendix 4	234
Appendix 5	235
Appendix 6	237
Appendix 7	238
Appendix 8	239

<i>Appendix 9</i>	240
<i>Appendix 10</i>	241
<i>Appendix 11</i>	242
<i>Appendix 12</i>	243
<i>Appendix 13</i>	244
<i>Appendix 14</i>	245
<i>Appendix 15</i>	246
<i>Appendix 16</i>	248
<i>Appendix 17</i>	249
<i>Appendix 18</i>	250
<i>Appendix 19</i>	251
<i>Appendix 20</i>	252
<i>Appendix 21</i>	257

List of Figures

Figure 1.1 Process of DNA methylation pattern inheritance.....	4
Figure 1.2 Inheritance of histone modification across DNA replication.....	7
Figure 1.3 Pattern of histone modifications found around the gene.....	9
Figure 1.4 Evolutionary conserved COMPASS and COMPASS-like complex.....	11
Figure 1.5 Formation of R-loop during transcription.....	14
Figure 1.6 Theoretical structure of R-loop <i>in cis</i> and <i>in trans</i>	16
Figure 1.7 Management of R-loop formation.....	19
Figure 1.8 Families of the SF1 and SF2 superfamily.....	28
Figure 2.1 The developmental stages of the embryo from fertilization to hatching.....	44
Figure 2.2 Set-up of the slot blot machine.....	47
Figure 3.1 Section of COMPASS mutants H3K4me3 ChIP-seq.....	52
Figure 3.2 ChIP-seq results comparing the H3K4me3 enriched and depleted genes in the COMPASS mutants.....	53
Figure 3.3 Average change in H3K4me3 around the TSS in the COMPASS mutants.....	54
Figure 3.4 Gene set enrichment analysis H3K4me3 enriched genes.....	55
Figure 3.5 Gene Ontology analysis and protein-protein network.....	57
Figure 3.6 <i>de novo</i> motif discovery of the 4499 genes depleted with H3K4me3 in both COMPASS mutants.....	58
Figure 3.7 Plots of sequence motif occurrence and nucleotide frequency upstream and downstream of TSS.....	60
Figure 3.8 Sequence analysis around the TSS of the 179 H3K4me3 enriched genes in <i>set-2(bn129)</i> and <i>cfp-1(tm6369)</i>	62
Figure 3.9 Average thymine frequency for each amino acid codon and the degradation pathway the amino acid is part of.....	65
Figure 4.1 Dot blot of the R-loop pilot experiment.....	68
Figure 4.2 R-loop pilot experiment using the slot blot apparatus and anti-dsDNA loading control procedure.....	70
Figure 4.3 Representative image of the optimized slot blot method and methylene blue loading control.....	71
Figure 4.4 Representative images showing the output of the G:BOX.....	71
Figure 4.5 Scoring of adult <i>C. elegans</i> developmental stage.....	72
Figure 4.6 Embryo developmental stage scoring.....	74
Figure 4.7 G:BOX image comparing the effect of different diets on the R-loop signal using hatched L1 worms.....	76
Figure 4.8 The theoretical mechanism of the antibody-independent R-loop purification protocol.....	80
Figure 4.9 Comparison of the efficiency of dsDNase in dsDNase buffer and NEB CutSmart® buffer.....	82
Figure 4.10 Nucleic acid mass composition after each nuclease digestion reaction using dsDNase.....	83
Figure 4.11 R-loop slot blot of nucleic acid samples after various digestion steps.....	83

Figure 4.12 Nucleic acid mass composition after each nuclease digestion reaction using restriction enzymes	85
Figure 4.13 The effect of Buffer exchange after restriction enzyme digestion.	86
Figure 4.14 The effect of Buffer exchange after dsDNAse digestion.	87
Figure 4.15 R-loop levels of the helicase mutants <i>rba-1(tm329)</i> and <i>rcq-5(ok660)</i>	90
Figure 4.16 Common body morphology phenotypes in <i>C. elegans</i> research.....	91
Figure 4.17 <i>bmr-1</i> RNAi effect on dead embryos in N2, <i>set-2(bn129)</i> and <i>cfp-1(tm6369)</i>	92
Figure 4.18 RNAi sensitivity of wild-type worms and COMPASS mutants on <i>dpy-10</i> and <i>unc-15</i> RNAi.....	93
Figure 4.19 Hatching assay of the seven RNAi genes resulting in suppression of the R-loop phenotype.....	95
Figure 5.1 Signalling pathway of all three branches of the UPR ^{ER}	108
Figure 5.2 Heat shock response pathway.....	110
Figure 5.3 Diagram depicting the three MAPK pathways associated with the innate immune response.....	119
Figure 5.4 <i>C. elegans</i> infected with <i>Bacillus thuringiensis</i>	124
Figure 5.5 Cross-section of <i>C. elegans</i> fed on Enterococcus bacteria after 8 hours.	125
Figure 5.6 <i>C. elegans</i> infected by <i>M. nematophilum</i>	127
Figure 5.7 <i>S. aureus</i> accumulation inside <i>C. elegans</i> gut.....	128
Figure 5.8 <i>S. enterica</i> (<i>Typhimurium</i>) accumulation inside <i>C. elegans</i> gut.....	129
Figure 5.9 <i>S. maltophilia</i> accumulation inside <i>C. elegans</i> gut.	130
Figure 5.10 Crystal structure formation inside <i>C. elegans</i> intestine following <i>P. luminescence</i> infection	131
Figure 5.11 <i>Y. tuberculosis</i> accumulation on <i>C. elegans</i> head.	132
Figure 5.12 <i>P. aeruginosa</i> accumulation inside <i>C. elegans</i>	133
Figure 5.13 <i>V. cholerae</i> accumulation inside <i>C. elegans</i> pharynx and intestine.	134
Figure 5.14 <i>D. coniospora</i> infection cycle in <i>C. elegans</i>	136
Figure 5.15 <i>C. elegans</i> infected by <i>N. parisii</i>	137
Figure 5.16 <i>C. elegans</i> infected with <i>Harposporium sp</i> (Juf27).....	138
Figure 5.17 <i>C. albicans</i> emerging from dead <i>C. elegans</i> hosts.....	139
Figure 7.1 Representative quality control using the dataset from Estes et al. 2010.	151
Figure 7.2 Histogram plot of the log2 Fold Change distribution for all genes in each dataset.....	153
Figure 7.3 Comparison of significantly differentially expressed genes from microarray datasets using the same pathogens.....	157
Figure 7.4 Venn diagram of significantly differentially expressed genes of closely related pathogens from the microarray datasets.	158
Figure 7.5 Venn diagram of significantly differentially expressed genes of pathogens in the same domain from the RNA-seq datasets.....	159
Figure 7.6 Heatmap of the 331 filtered genes from all 29 datasets.....	162
Figure 7.7 Within-groups sum of squares of the datasets by k-means clustering.....	163
Figure 7.8 Heatmap of the 585 filtered genes from the 25 datasets.....	166

Figure 7.9 Heatmap of the 383 genes from the 25 datasets using the more stringent filtering criteria.	169
Figure 7.10 Comparison of the three filtered gene lists.	170
Figure 7.11 Wormbase Enrichment Analysis on list585 and list383.	171
Figure 7.12 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of Figure 7.8.	172
Figure 7.13 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of Figure 7.9.	173
Figure 7.14 ClueGO network analysis for ‘GO:biological processes’ on all genes from list383.	175
Figure 7.15 ClueGO analysis for GO:biological processes for the three largest groups of list383.	176
Figure 7.16 Heatmaps of gene expression of protein families considered to be immune effectors in each pathogen dataset.	178
Figure 7.17 Heatmap of transcription factor and co-factor expression changes in each pathogen dataset.	182
Figure 7.18 Seqplot of the ChIP-seq signal around the transcript start site.	184
Figure 7.19 de novo Motif discovery result using HOMER on list383.	185
Figure 7.20 Comparison of the target genes of different TFs associated with immune response.	186
Figure 7.21 Comparison of the differentially expressed genes between <i>daf-16(mu86)</i> and <i>pqm-1(ok485)</i> at different temperatures.	188
Figure 7.22 Overlap between TF ChIP-seq binding data and the RNA-seq differential expression data of DAF-16 and PQM-1.	188
Figure 7.23 Comparison of up and down-regulated genes in the heat shock datasets.	191
Figure 7.24 Heatmap of the 255 genes from the 4 heat shock datasets after filtering.	192
Figure 7.25 Wormbase Enrichment Analysis of the 255 HSR genes.	193
Figure 7.26 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of the heat shock datasets (Figure 7.24).	194
Figure 7.27 Heatmap of the heat shock datasets for caenacin (CNC).	196
Figure 7.28 Comparison of pathogen responsive gene lists with the heat shock gene list.	197
Figure 7.29 Comparison of the genes in each pathogen responsive gene list that are also in the heat shock responsive gene list.	198
Figure 7.30 Wormbase Enrichment Analysis and STRING analysis for the 50 genes that are responsive to both pathogen infection and heat shock.	199

List of Tables

Table 1.1 Summary of methylation modification found to affect transcription.....	8
Table 1.2 Function of the families of the SF2 helicase superfamily.....	29
Table 2.1 List of strains used in the epigenetic study.	36
Table 2.2 Table displaying the sequence of primers used for genotyping the mutant <i>C. elegans</i> strains.....	38
Table 2.3 Composition of the PCR master mix for either one reaction or eight (8.5) reactions.	39
Table 2.4 PCR program of the thermocycler for different gene/strain.	39
Table 4.1 R-loop RNAi helicase suppressor screen on <i>set-2(bn129)</i>	94
Table 7.1 Summary of the number of significantly differentially expressed genes of all 29 pathogen infection datasets.	154
Table 7.2 Grouping of the 29 datasets into 7 groups via k-means clustering.	164
Table 7.3 Grouping of the 585 differentially expressed genes from the 25 datasets into 6 groups via k-means clustering.	165
Table 7.4 Grouping of the 1671 genes from the 25 datasets into seven groups via k-means clustering.	168
Table 7.5 Number of up and down-regulated genes in the <i>daf-16(mu86)</i> and <i>pqm-1(ok485)</i> mutants at different temperatures.	187
Table 7.6 Number of significantly differentially expressed genes for each of the 4 heat shock datasets.	190

List of Abbreviations

ABF	Antibacterial factor related
AID	Activation-Induced Cytidine Deaminase
ATF	Activating Transcription Factor
ATFS-1	Activating Transcription Factor associated with Stress
<i>bar-1</i>	Beta-catenin/Armadillo Related
BLM	Bloom Syndrome Protein
BRCA1/2	Breast Cancer Type 1/2
bus	Bacterial Un-Swollen
bZIP	Basic Leucine Zipper
CFP1	CxxC Finger Protein 1
CGI	CpG Island
CHD	Chromodomain
CLEC	C-type Lectin
CNC	Caenacin
COMPASS	Complex of Proteins Associated with Set1
Cps	COMPASS
DAF	Abnormal Dauer Formation
DAPI	4',6-diamidino-2-phenylindole
DDR	DNA Damage Response
DDX	DEAD-Box helicase
DHX9	DExH-Box Helicase 9
DNMT	DNA methyltransferase
<i>dod</i>	Downstream of DAF
<i>dpy</i>	Dumpy
DRB	5,6-Dichloro-1- β -D-Ribofuranosyl-Benzimidazole
DSB	Double-Stranded Break
ECL	Electrochemical Luminescence
EDTA	Ethylenediaminetetraacetic acid
eIF	Eukaryotic initiation factor
ELT	Erythroid-Like Transcription Factor Family
ERK	Extracellular Signal-Regulated Kinase
EV	Empty Vector
FANCM	Fanconi Anemia, complementation Group M
FOXO	Forkhead Box Protein O
GADD45A	Growth Arrest and DNA Damage inducible protein 45 Alpha
GO	Gene Ontology
H3K27me3	Histone 3 Lysine 27 trimethylation
H3K36me3	Histone 3 Lysine 36 trimethylation
H3K4me3	Histone 3 Lysine 4 trimethylation
H3K9me3	Histone 3 Lysine 9 trimethylation

H3S10P	Histone 3 Serine 10 Phosphorylation
HLTF	Helicase Like Transcription Factor
<i>hmr-1</i>	Hammerhead embryonic lethal 1
HOT	High-occupancy Targets
HR	Homologous Recombination
HSE	Heat Shock Element
<i>hsf-1</i>	Heat Shock Factor 1
HSP	Heat Shock Protein
HSR	Heat Shock Response
ILS	Insulin-Like Signalling
ILYS	Invertebrate lysozyme
ING	Inhibitor of Growth
IPTG	Isopropyl β -d-1-thiogalactopyranoside
IRE1	Inositol-requiring Protein 1 α
ISWI	Imitation SWI
JMJD2C	Jumanji Domain 2C
JNK	C-Jun N-Terminal Kinase
JUN-1	JUN Transcription Factor Homolog
LB	Luria Broth
LET	Lethal
<i>lin-1</i>	Abnormal cell lineage 1
LYS	Lysozyme
m⁵C	5-methylcytosine
m⁶A	6-methyladenosine
MAPK	Mitogen Activated Protein Kinase
MDT-15	Mediator 15
MEK	MAP Kinase Kinase
MLL	Mixed Lineage Leukemia
<i>mog-5</i>	Masculinisation of Germline 5
MPK-1	MAP Kinase 1
NEB	New England Biolabs
NGM	Nematode Growth Media
NLP	Neuropeptide-Like Protein
Nrf2	Nuclear Factor Erythroid 2 – related factor 2
NS3	Nonstructural Protein 3
NSY-1	Neural Symmetry
NURF	Nucleosome Remodelling Factor
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PERK	Protein Kinase RNA-Like Endoplasmic Reticulum Kinase
PHD	Plant Homeodomain
PK	Proteinase K

PMK-1	P38 MAP Kinase 1
PolyQ	Polyglutamine
PQM-1	Paraquat (Methylviologen) Responsive 1
PTM	Post-Translational Modification
<i>rad-54</i>	Radiation sensitivity abnormal 54
<i>rcq-5</i>	RecQ DNA Helicase family 5
<i>rha-1</i>	RNA helicase A
RNP	Ribonucleoprotein
ROS	Reactive Oxygen Species
RPA	Replication Protein A
SEK-1	SAP/ERK Kinase
SET	<i>Su(var)3-9</i> , <i>Enhancer of zeste (E(z))</i> , and <i>trx</i>
SETX	Senataxin
SF	(Helicase) Superfamily
SKN-1	Skinhead 1
SL1	Spliced Leader 1
SPP	Saposin-like protein
SSB	Single-Stranded Break
SWI/SNF	Switch/Sucrose Non-Fermentable
TAF3	TATA-box binding protein associated Factor 3
TBE	Tris-Borate EDTA
TBS (TBST)	Tris-Buffered Saline (+ Tween)
TCS	Transcellular Chaperone Signalling
TDRD3	Tudor Domain Containing 3
TET1	Ten-Eleven Translocase 1
TF	Transcription Factor
TFIID	Transcription Factor II D
TGF-β	Transforming Growth Factor β -like pathway
TIR-1	Toll-Interleukin-1 Receptor
TLR	Toll-Like Receptor
TOP3B	Topoisomerase III β
topA	Topoisomerase 1
TREX	Transcription Export
TSS	Transcription Start Site
TTF2	Transcription Termination factor 2
TTS	Transcription Termination Site
<i>unc</i>	uncoordinated
UPR (UPR^{ER})	Unfolded Protein Response (Endoplasmic Reticulum)
UTR	Untranslated Region
<i>vbh-1</i>	Vasa- and Belle-like helicase 1
WRN-1	Werner Syndrome ATP-dependent Helicase 1
XBP-1	X-box Protein 1

Part 1

Investigating the functional
relationship between R-loops and the
evolutionarily conserved COMPASS
complex

Chapter 1 Introduction of epigenetics

Efforts in genetic and genomic research have revealed the genome of many organisms. However, knowing the genome does not automatically translate to knowing the phenotype of a cell. A prime example is that cells in a multicellular organism have the same genome, but can be different types of cells. Conrad Waddington put forward this thought that “between genotype and phenotype, and connecting them to each other, there lies a whole complex of developmental processes” and termed the study of these processes *epigenetics* in 1942 (Waddington, 1942). As science has progressed, molecular pathways have been found that link specific genotypes with phenotypes and as such, the definition of epigenetics has been gradually narrowed (Dupont, et al., 2009). Currently, the definition of epigenetics is widely accepted as “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in the DNA sequence” (Wu & Morris, 2001; Dupont, et al., 2009). Epigenetic effects are controlled through various mechanisms such as the methylation of DNA residues, thereby preventing transcription, and modification of the amino acids near the amino-terminal of histones (histone tail), which is associated with altering the chromatin landscape and DNA accessibility (Handy, et al., 2011). Non-coding RNAs have been considered epigenetic regulators for their ability to control gene expression and their effect on DNA methylation and histone modification (Wei, et al., 2016). A new candidate has recently gained interest in the epigenetics community as an epigenetic regulator. This candidate is the DNA:RNA hybrid called the “R-loop” and is implicated in transcriptional regulation of active genes (Al-Hadid & Yang, 2016). Each type of epigenetic regulator uses different mechanisms to regulate gene expression, and their interaction adds another layer of complexity in gene expression regulation.

This chapter provides a summary of epigenetics with regards to chromatin and histone modification, focusing specifically on the trimethylation of the Histone 3 Lysine 4 (H3K4) residue. Furthermore, the new epigenetic regulator candidate “R-loop” is introduced and its crosstalk with other epigenetic modifiers is described. Finally, helicases are discussed for their ability to resolve R-loops.

1.1. Mechanisms of epigenetics

Epigenetic regulation by DNA methylation and histone modification controls gene expression at the transcriptional level by controlling the binding of proteins such as transcription factors and RNA Polymerase II to DNA. For example, dosage compensation

(equalizing gene expression between members of different sexes) by X-inactivation in female mammals is attributed to the tight packaging of the inactive X chromosome (heterochromatin) through histone modifications, thereby preventing proteins from accessing the relevant DNA region (Dupont, et al., 2009). Another example is genomic imprinting in diploid organisms, where only one parental copy of a gene is expressed, while the other copy is inactive. This parent-of-origin specific expression of certain genes is mainly attributed to DNA methylation, but also non-canonically controlled by histone modification (Ferguson-Smith & Bourc'his, 2018).

1.1.1. DNA methylation

Mammalian DNA methylation, in the context of epigenetics, is the addition of methyl groups to the carbon at the fifth (m^5C) position of nucleotides, which acts as a transcriptional repressor. Methylation marks at other positions are associated with different functions. The methylation on the first (m^1A) and third (m^3C) position are considered markers for DNA damage. In prokaryotes, methylation at the fourth (m^4C) and sixth position (m^6A) are used to differentiate between own-DNA and foreign-DNA (Greer, et al., 2015). DNA methylation (m^5C) mainly occurs at cytosine bases, especially at CpG dinucleotides (cytosine followed by guanine base) (Jin, et al., 2011). In mammals, 60-90% of CpG dinucleotides in CpG poor regions are methylated. CpG rich regions, known as CpG Islands (CGI) on the other hand often acts as regulatory regions such as promoters, and are often hypomethylated (Siegfried & Cedar, 1997; Cross & Birds, 1995). Roughly 70% of actively transcribed gene promoters are found at CGIs (Saxonov, et al., 2006). However, methylated CpG rich areas do exist and are found at silenced genes and inactivated X chromosome (Riggs & Pfeifer, 1992; Neumann & Barlow, 1996). The methylation-dependent silencing can occur by either interfering with the binding of transcription factors, the recruitment of repressor complexes that specifically bind to methylated DNA and/or the modification of the chromatin landscape and other epigenetic modifications (Curradi, et al., 2002).

Inheritance of methylated DNA occurs alongside replication. During DNA replication, the parent strand with the methylated nucleotides is split to become the template for the newly synthesized daughter strand. The newly synthesized daughter strand is unmethylated and with the methylated parent strand form hemimethylated DNA. This dilution of methylation is reversed when the daughter strand becomes methylated according to the methylation pattern at the parent strand (**Figure 1.1**) (Sharif & Koseki, 2018). The main protein responsible for this is DNA-methyltransferase 1 (DNMT1) (Yu, et al., 2011). However, not

all organisms have m^5C . The model organism *Caenorhabditis elegans*, for example, has neither detectable levels of m^5C nor has homologs for DNMT1. Nevertheless, methylation has been found on the nematodes DNA, but at the sixth position (m^6A) (Greer, et al., 2015). In *C. elegans*, m^6A is distributed broadly across the genome and is inherited across generations, making this modification a potential epigenetic information carrier. m^6A has been shown to crosstalk with methylated H3K4, but whether it has transcription regulatory functions in *C. elegans* remains to be determined (Greer, et al., 2015).

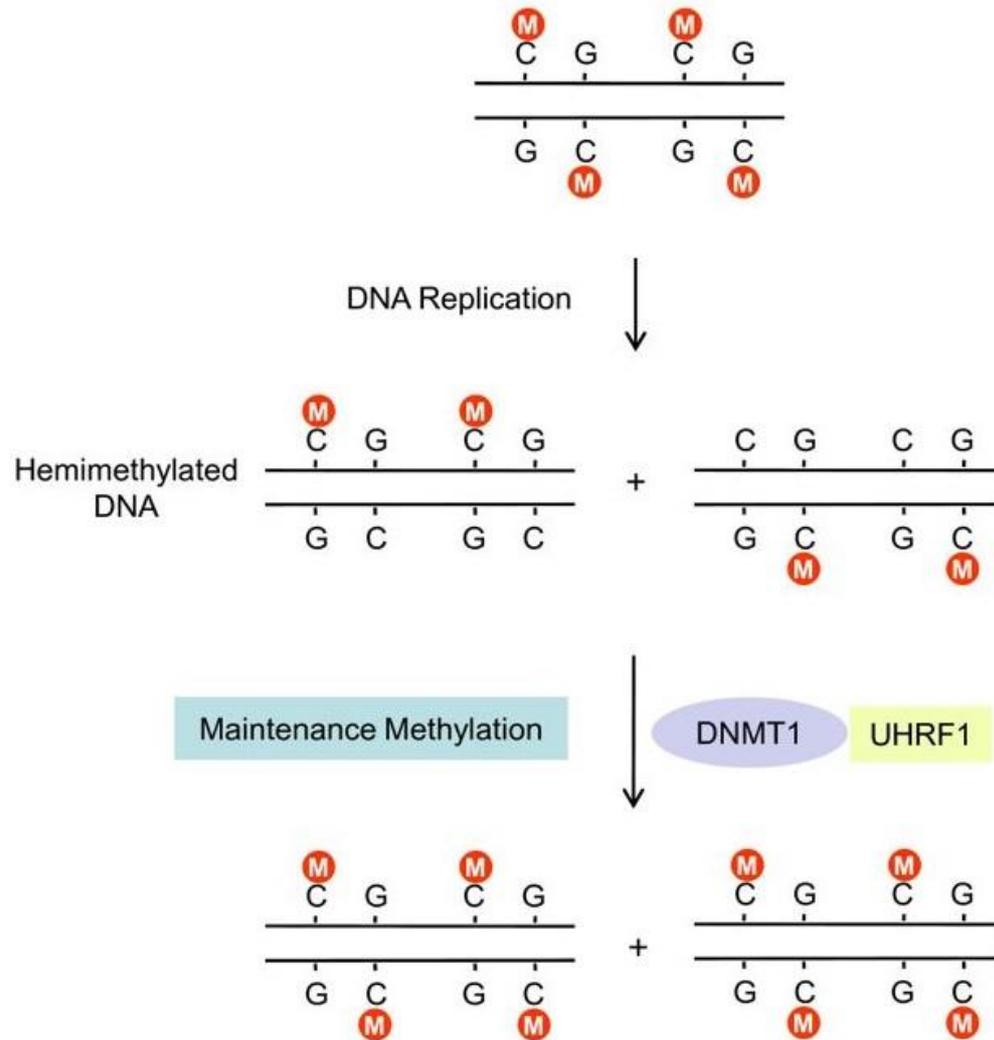


Figure 1.1 Process of DNA methylation pattern inheritance. DNA replication results in two hemimethylated double-stranded DNAs consisting of one parent strand and one daughter strand each. Methylation is subsequently performed on the daughter strands by a specific methyltransferase (DNMT1). The human UHRF1 mediates the binding of DNMT1 to hemimethylated DNA. Image taken from Yu, et al. (2011).

1.1.2. Chromatin

In eukaryotes, the genome is packaged in a highly ordered structure comprising of DNA, RNA and proteins called chromatin (Black & Whetstone, 2011). The varying extent of this

packaging across the genome is called the chromatin landscape and changes depending on many factors such as developmental stage, type of cell and environmental effects. Apart from DNA compaction, chromatin is implicated in transcription regulation through many aspects, from being a steric hindrance for DNA accessibility (Orphanides & Reinberg, 2000) to the regulation of transcription factors (Li, et al., 2007).

Chromatin is made of smaller units called nucleosomes, which comprise less than two turns (146 base pairs) of DNA wrapped around a histone protein complex. The histone protein complex itself is an octamer made of two of each H2A, H2B, H3 and H4 protein, also known as core histone proteins (Kornberg, 1974; Kornberg & Lorch, 1999). The histone H1 protein is known as a linker protein that binds to the core histone “bead” at the DNA entry and exit sites and stabilizes more condensed chromatin architecture (Hergeth & Schneider, 2015).

Chromatin exists in two forms: heterochromatin and euchromatin. These forms are distinguished by distinct chromosomal proteins and histone posttranslational modifications, which separates them into densely packed gene-poor (heterochromatin), and loosely packed gene-rich (euchromatin) regions (Huisinga, et al., 2006; Handy, et al., 2011).

1.1.2.1. Euchromatin

Euchromatin is the loose chromatin structure identified as 11 nm fibre (11 nm diameter of the chromatin strand). This loose structure allows for different proteins such as RNA Polymerase II to bind and transcribe genes. Owing to the presence of active transcription of genes, this type of chromatin is also referred to as active chromatin (Trojer & Reinberg, 2007).

Histone modifications in euchromatin are variable with modification-rich “islands” at or near genes and transcription regulation sites. Modifications include H3K4 mono- and trimethylation at the transcription start sites (TSS) of a gene and H3K36 trimethylation in the gene body. Hyperacetylated histones are also frequently found in euchromatin (Trojer & Reinberg, 2007; Bannister & Kouzarides, 2011).

1.1.2.2. Heterochromatin

Heterochromatin is the densely packaged form of chromatin that hinders transcription by reducing DNA accessibility. As such it is also called inactive chromatin. Heterochromatin is categorized into constitutive and facultative heterochromatin. Both are densely packed, but the latter retains the ability to convert between heterochromatin and euchromatin. This ability to change between transcriptionally active and inactive chromatin states is important

to control cell type or developmental stage-specific genes as well as dosage compensation. Constitutive heterochromatin, on the other hand, is mainly found at regions important for gene stability (e.g. centromeres and telomeres) as well as repetitive and non-coding regions (Trojer & Reinberg, 2007; Huisinga, et al., 2006).

Constitutive heterochromatin is organized into 30 nm fibres (30 nm diameter) while facultative heterochromatin is compacted into a mixture of both 11 nm fibres and 30 nm fibres, consistent with the idea that it can convert between heterochromatin and euchromatin (Trojer & Reinberg, 2007). These two types of heterochromatin need to be differentiable by cellular mechanisms to determine which regions of heterochromatin to expand. It is currently unknown how these two are distinguished, but it could be likely that the two heterochromatins adopt different chromatin compaction architectures or various histone modifications to differentiate between them.

Histone modifications found on constitutive heterochromatin are mainly hypoacetylation with H3K9 and H4K20 trimethylations (Trojer & Reinberg, 2007). Facultative heterochromatin shares many of the same modifications with the addition of some unique modifications such as H3K27 trimethylation found on the inactivated X-chromosome in females (Bannister & Kouzarides, 2011).

1.1.3. Histone modification

Histone modification is the addition of functional groups (e.g. methyl, acetyl and phosphate groups) to the histone tail towards the amino-terminal. Proteins with the ability to add modifications are termed “writers”. Modifications can be subsequently removed by proteins called “erasers”. Since the histone tail has many residues where various functional groups can be added, and the same functional groups can be added more than once to the residue effectively stacking on top of each other. This allows for numerous modification patterns that are identified and interpreted by proteins known as “readers” (Biswas & Rao, 2018). Each modification can affect the chromatin state and transcription. The large number of possible combinations of histone modifications is termed the histone code (Jenuwein & Allis, 2001). The effect of each modification depends on which residue this modification is found and how often this modification is found. For instance, acetylation of histone residues is often associated with active chromatin regions, but can also indicate DNA damage (H3K56ac) (Masumoto, et al., 2005) similarly to phosphorylation of H2A histone (γ H2AX) (Rogakou, et al., 1998). Mono-methylation on H3K9 is mainly associated with gene

activation, but di- and tri-methylation is associated with transcriptional repression (Barski, et al., 2007).

Similar to DNA methylation, histone modifications can also be inherited. During replication, modified histones from the parent DNA double helix are randomly distributed to the two daughter helices through an unknown mechanism. New and unmodified histones are then added to fill in unoccupied regions, effectively diluting the concentration of the histone modification. Afterwards, another mechanism then copies the histone modification from the inherited parent histone to neighbouring newly incorporated nascent histones (**Figure 1.2**) (Moazed, 2011; Whitehouse & Smith, 2013).

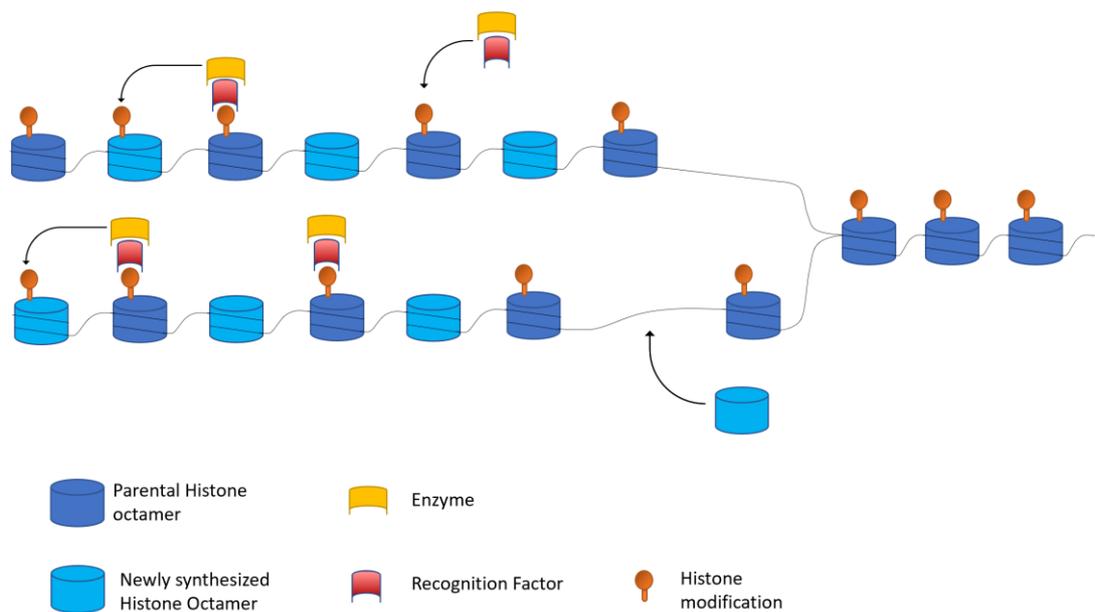


Figure 1.2 Inheritance of histone modification across DNA replication. During DNA replication, histones from the parent helix are split between the daughter helices. New unmodified histones are then added to fill the empty gaps. Histone modification is then re-established based on the neighbouring parental histones. Image based on Moazed (2011).

There are many post-translational modifications (PTMs) of histone proteins. Methylation and acetylation are the two main modifications with regards to transcriptional regulation (An, 2007). Methylation occurs mainly on lysine and arginine residues and has various effects on transcription, which are summarized in **Table 1.1**. Acetylation is only found on lysine residues and is associated with a less packed chromatin structure as it neutralizes the positive charge from lysine and reduces the electrostatic attraction to the negatively charged DNA that condenses chromatin (Bannister & Kouzarides, 2011). As such, they are mainly found associated with transcriptional activation. Ubiquitination, SUMOylation (Small Ubiquitin-like Modifier) and phosphorylation are PTMs suggested to affect gene expression through the change in chromatin condensation state or recruitment of DNA-binding proteins (Bannister & Kouzarides, 2011).

	mono-methylation	di-methylation	tri-methylation
H2BK5	Activation		
H3R2		Activation	
H3K4	Activation & Repression	Activation	Activation
H3R8	Repression	Repression	
H3K9	Activation & Repression	Repression	Repression
H3R17	Activation	Activation	
H3R26	Activation		
H3K27	Activation & Repression	Repression	Repression
H3K36	Activation	Depends on species	Activation
H3R43	Activation		
H3K79	Activation	Activation	Activation & Repression
H4R3	Activation		
H4K20	Activation & Repression	Repression	Repression

Table 1.1 Summary of methylation modification found to affect transcription. The figure is based on published information (Zhao & Garcia, 2015; Barski, et al., 2007; Li, et al., 2007; Rosenfeld, et al., 2009).

As mentioned above, histone modifications effect on transcription is either through changing the chromatin landscape (electrostatic attraction) or by promoting the recruitment of non-histone proteins. One type of non-histone protein recruited is chromatin remodelers that reshape the chromatin landscape by moving nucleosomes. This can create areas of dense nucleosomes for expressional inactivation or nucleosome-sparse regions, thereby making the DNA accessible for gene expression. For example, the human chromodomain (CHD1), a chromatin remodeler likely functioning in disassembling nucleosomes, specifically recognizes H3K4me3 (Petty & Pillus, 2013). Other recognition domains include the bromodomain that recognizes acetylated lysine residues and plant homeodomain (PHD) domain that recognizes H3K4me3 (Petty & Pillus, 2013). Other non-histone proteins that are recruited by PTMs are transcription factors. The basal transcription factor TFIID for example directly binds to H3K4me3 through the PHD finger domain of its subunit TAF3 and directs the formation of the preinitiation complex (Lauberth, et al., 2014; van Ingen, et al., 2008).

The location where the histone modification is found relative to the gene position is dependent on the type of modification (**Figure 1.3**). H3K4me3 is mainly seen as a sharp peak at the TSS, whereas its mono- and di-methylated counterparts are broadly distributed peaking towards the end of the gene and the middle of the gene body, respectively. The specific location of H3K9me on the gene determines if it acts as a transcriptional suppressor (before the TSS) or transcriptional activator (gene body) (Li, et al., 2007). Proteins associated/recruited by specific histone PTMs will show a similar trend in occupancy as them. For example, TFIID/TAF3, which is recruited by H3K4me3, also has an occupancy pattern with a peak at the TSS (Lauberth, et al., 2014).

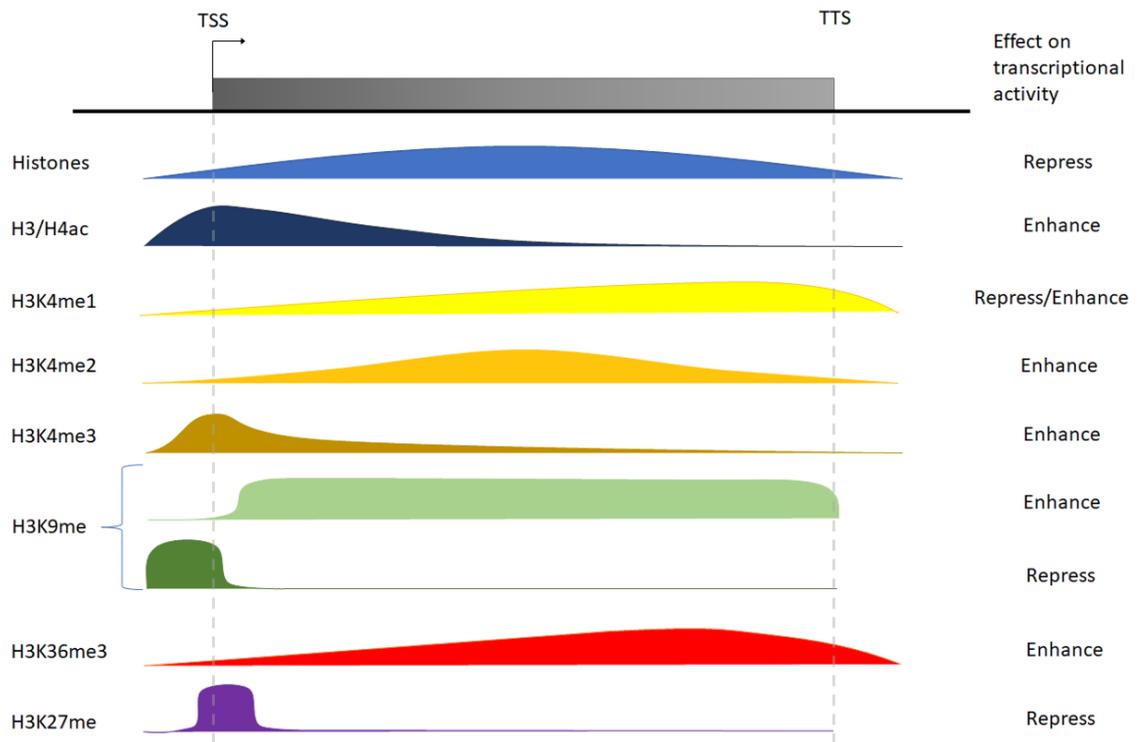


Figure 1.3 Pattern of histone modifications found around the gene. The dashed line encompasses the gene body. On the right denotes the effect of the histone modification on transcriptional activity. Image based on Li, et al. (2007).

1.2. COMPASS complex and H3K4me3

Methylation of histone residues is achieved by the writer enzymes classified as histone methyltransferase (HMT) (Torres & Fujimori, 2015; Falkenberg & Johnstone, 2014). H3K4me3 is one of the more well-known histone PTMs which has been extensively researched (EpiGenie, n.d.; Kusch, 2012). It is an epigenetic marker associated with active transcription, more specifically of active promoters, because they peak at the TSS (**Figure 1.3**), which is located within the core promoter region (Heintzman, et al., 2007). This marker is conserved in eukaryotes, from yeast to mammals (Zhang, et al., 2015) and is deposited mainly by the evolutionary conserved COMPASS (Complex Protein Associated with Set1) (Miller, et al., 2001).

1.2.1. The COMPASS complex

The SET1 methyltransferase was first identified in *Saccharomyces cerevisiae* (Stasser, et al., 1995; Shilatifard, 2012), with all six other subunits that form the Set1/COMPASS complex being found six years later (Miller, et al., 2001). SET1 is the main catalytic subunit that facilitates the trimethylation of lysine residues (Ardehali, et al., 2011). This methylation function is conserved in the SET domain, which is found in most histone lysine methyltransferases (Dillon, et al., 2005). The other subunits are also essential for the proper function of the COMPASS complex. Cps50 and Cps30 are required for the assembly and stability of the complex. Cps25, Cps35 and Cps60 are essential for di- and trimethylation, while Cps40 is only needed for trimethylation of H3K4 (Shilatifard, 2012; Dehe, et al., 2006).

While *S. cerevisiae* only has one COMPASS complex that facilitates all trimethylation, *C. elegans*, *Drosophila melanogaster* and humans have multiple COMPASS or distantly related COMPASS-like complexes (**Figure 1.4**). In humans, the COMPASS-like complexes Mixed lineage leukemia protein (MLL) trimethylate H3K4 at a subset of genes (MLL1 and MLL2) or are responsible for monomethylating H3K4 (MLL3 and MLL4) whereas the “main” COMPASS complexes (SET1A and SET1B) are responsible for the majority of H3K4me3 in the genome (Shilatifard, 2012).

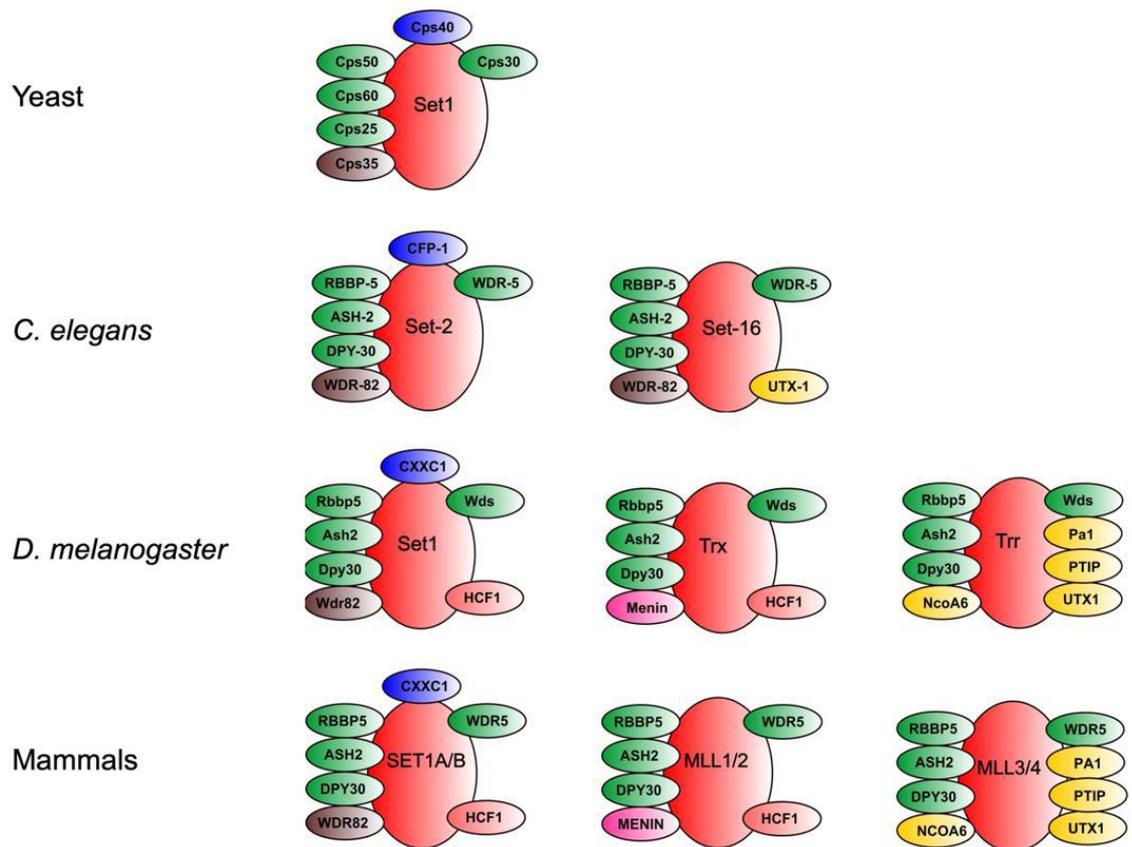


Figure 1.4 Evolutionary conserved COMPASS and COMPASS-like complex. The COMPASS complex for each organism is found in the first column. The other columns show COMPASS-like complexes. The catalytic subunit is shown in red, and the same colours indicate conserved subunits. Image taken from Pokhrel (2019).

1.2.1.1. *set-2*, the *C. elegans* ortholog of *Set1*

In *C. elegans*, *set-2* is the ortholog of the yeast Set1 COMPASS complex that facilitates the majority of trimethylation of H3K4, while the COMPASS-like, *set-16*, is the ortholog of mammalian MLL3/4 which are responsible for monomethylation of H3K4 (Pokhrel, et al., 2019). Mutation of *set-2* in the C-terminal domain, where the SET domain is located, reduces the global H3K4me3 levels in *C. elegans* (Xiao, et al., 2011; Pokhrel, et al., 2019). Other phenotypes associated with this include slow growth with increased lifespan (Han, et al., 2017) and transgenerational progressive sterility (Xiao, et al., 2011). In yeast, null mutants of Set1 or any other subunit result in a slow-growth phenotype (Miller, et al., 2001). A similar phenotype has been shown in human cancer cell lines depleted with SETD1A (Tajima, et al., 2015).

1.2.1.2. *cfp-1*, *C. elegans* ortholog of *Cps40*

cfp-1 (CXXC finger protein 1), one of the core subunits of SET-2/COMPASS and ortholog of yeast Cps40, contains a DNA binding domain that has the ability to bind to unmethylated DNA at CGIs (Lee & Skalnik, 2005; Long, et al., 2013). Similar to *set-2*, mutation in *cfp-1* causes reduction in global H3K4me3 levels (Pokhrel, et al., 2019; Beurton, et al., 2019). It has been implicated to be essential for targeting of the COMPASS complex to the correct sites (Brown, et al., 2017). In mouse embryonic stem cells, Cfp1 (ortholog of *C. elegans* Cfp-1) null mutants cause aberrant accumulation of H3K4me3 at many non-promoter regions. These aberrant accumulations of H3K4me3 peaked at sites correlating with transcriptional enhancers (Clouaire, et al., 2012; Clouaire, et al., 2014).

1.2.2. The H3K4me3 epigenetic mark and its function

1.2.2.1. H3K4me3 and transcription activation

The presence of H3K4me3 at active gene promoters is observed universally from yeast to human, and the amount of this modification reflects the transcriptional levels of the gene (Howe, et al., 2017). This correlation between H3K4me3 levels and transcription led to the hypothesis that H3K4me3 could be instructive for active transcription. For example, H3K4me3 could act as a binding site for chromatin remodelers and chromatin modifiers such as the NURF (Nucleosome Remodelling Factor) complex, which has a PHD finger that directly associates with H3K4me3 (Wysocka, et al., 2006). NURF has been shown to be required for active transcription by remodelling the chromatin landscape in a way that promotes the recruitment of transcriptional machinery and the loss of H3K4me3 results in the partial release of NURF subunit from chromatin (Badenhorst, et al., 2002; Wysocka, et al., 2006). Genome-wide studies, however, did not find significant transcriptional changes upon removal of most H3K4me3 marks, thereby implying that transcriptional activity is instructive for H3K4me3 deposition, rather than the other way round, perhaps to act as a marker of actively transcribed genes (Clouaire, et al., 2012).

Other chromatin modifiers that contain a PHD finger are the human Inhibitor of Growth ING4 and ING5, which are subunits of the histone acetyltransferase (HAT) complex HBO1 (Hung, et al., 2010; Lee, et al., 2018) and yeast Yng1 and Yng2 which are part of NuA3 and NuA4 HAT complexes respectively (Taverna, et al., 2015; Shi, et al., 2009; Lee, et al., 2018). HATs acetylate histones, and as mentioned in a section 1.1.3, acetylation of histones opens up chromatin to allow the binding of transcription machinery to facilitate active

transcription. The histone demethylase Jumonji Domain 2C (JMJD2C) has also been shown to bind to H3K4me₃, but through the double Tudor domain, and reduces the transcriptionally repressive H3K9me₃ and H3K36me₃ modifications (Huang, et al., 2006; Pedersen, et al., 2014). A non-chromatin modifier protein that can recognize H3K4me₃ is the TFIID basal transcription factor complex through its PHD finger on the TAF3 subunit, that regulates the transcription of specific genes in response to DNA damage (Lauberth, et al., 2014; van Ingen, et al., 2008). CHD1 has been demonstrated to interact with H3K4me₃ to function in mRNA maturation via transcription elongation and pre-mRNA processing (Sims III, et al., 2007).

1.2.2.2. H3K4me₃ in transcriptional repression

Emerging evidence indicates that H3K4me₃ could also act to repress transcription. Proteins related to transcriptional repression can also have H3K4me₃ binding domains (e.g. PHD finger) that are recruited by H3K4me₃. One example is the yeast histone deacetylase (HDAC) Rpd3L whose subunits, Pho23 and Cti6, have a PHD finger (Lee, et al., 2018; Shi, et al., 2009). Similarly, the human ortholog of Pho23, ING2, which is a subunit of mSin3a-HDAC1 can also directly bind to H3K4me₃ and repress transcription of proliferation genes in response to DNA damage (Shi, et al., 2011).

In yeast, H3K4ac was shown to be limited by COMPASS complex, implying that H3K4me₃ and H3K4ac levels are oppositely controlled. This suggests that the transcription promoting effect of H3K4ac are hindered by H3K4me₃ deposition (Guillemette, et al., 2011). Furthermore, comparison of mRNA levels identified H3K4me₂ as a repressive marker of transcription that relies on H3K4me₃-dependent antisense transcription (Margaritis, et al., 2012). Set1-dependent H3K4 methylation acts as a transcriptional repressor during stress, specifically of genes required for ribosome biosynthesis (Weiner, et al., 2012).

1.2.2.3. H3K4me₃ in DNA damage

H3K4me₃ could play a role in the DNA damage response. It has been observed that H3K4me₃ accumulates at the DNA damage-induced genes Growth Arrest and DNA Damage Protein Inducible 45 Alpha (GADD45A) and p21 during DNA damage (Lauberth, et al., 2014; Kim, et al., 2010). Since H3K4me₃ can act as a binding site for TAF3/TFIID, that also binds with the tumour suppressor protein p53 (Coleman, et al., 2017), this could potentially mean that H3K4me₃ responds to DNA damage through a p53 mediated response.

1.3. R-loops

The traditional paradigm dictates that DNA exists as double-stranded helix while RNA is single-stranded. However, “non-traditional” moieties like the DNA:RNA hybrid structures exist and occur naturally, such as Okazaki fragments that are formed during replication. Another such structure exists naturally, called the R-loop, which is a longer-lasting hybrid that is mainly produced as a product of transcription (Aguilera & Garcia-Muse, 2012). The structure is formed by the incorporation of single-stranded RNA (ssRNA) into unwound double-stranded DNA (dsDNA), producing a DNA:RNA hybrid and a displaced ssDNA (Aguilera & Garcia-Muse, 2012; Santos-Pereira & Aguilera, 2015). It was first described in 1976 (Thomas, et al., 1976), but only started to gain more attention in the past few decades. While they are naturally occurring in many different organisms, from yeast to humans, occupying as much as 5% of the mammalian genome (Zeller, et al., 2016; Sanz, et al., 2016), high aggregation of them has often been associated with reduced DNA stability through single and double-stranded breaks (Skourti-Stathaki & Proudfoot, 2014). However, they have also been linked to crucial biological roles, such as transcriptional regulation (Aguilera & Garcia-Muse, 2012). It is debatable if R-loops can be classified as epigenetic since their heritability is unknown, but they do have an effect on transcriptional regulation without changing the DNA sequence (see **Section 1.4**), thereby fitting at least one of the two criteria defining epigenetics.

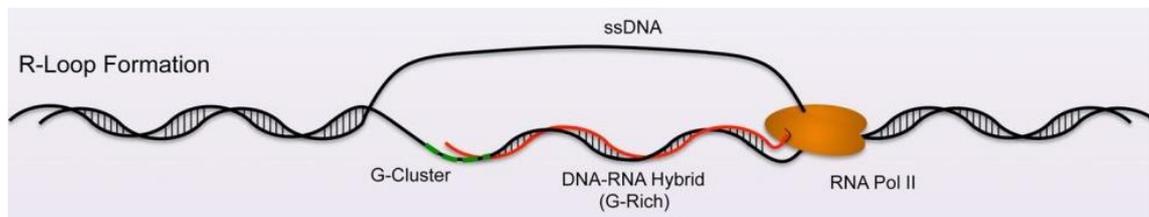


Figure 1.5 Formation of R-loop during transcription. DNA strands are coloured black, and the RNA strand is coloured red. GC rich clusters are coloured green. DNA-RNA formation is enriched at regions with high GC content, where the RNA strand has a higher affinity to bind to its complementary C-rich DNA strand. Image taken from Allison & Wang (2019).

Methods of mapping R-loops throughout the genome have identified properties that correlate with R-loop occupancy. R-loops are found disproportionately at the promoter and terminal regions with a 2-3 fold overrepresentation relative to the respective region size (Sanz, et al., 2016). In humans, terminal R-loops show a broader signal covering the transcription termination site (TTS) and peaking just before the polyadenylation site, while promoter-proximal R-loops are mostly observed as sharp peaks rising immediately after the TSS and peaking at around 1.5kb downstream of the TSS (Sanz, et al., 2016). Not all gene

promoters have R-loops. The presence of R-loops has been observed to increase at unmethylated CGIs, as well as high GC skew (asymmetric distribution of guanine and cytosine between strands). Furthermore, H3K4me2/me3, as well as H3K9ac and H3K27ac, are also found enriched at these sites, suggesting that R-loop formation is associated with highly active genes and open chromatin regions (Ginno, et al., 2012; Sanz, et al., 2016).

1.3.1. Formation of R-loops

R-loop formation has been found to be enhanced by high GC skew, negative supercoiling, and DNA nicks (Roy, et al., 2010). Although the mechanism behind R-loop formation is still being investigated, current research indicates that the majority of R-loops are produced during transcription (Frederic, 2016). R-loops are found at CGIs, where the majority of active genes are also located at (Ginno, et al., 2012). CGIs with high GC skews are especially enriched in R-loops (Ginno, et al., 2013) as these promoters are highly active (Illingworth & Bird, 2009). Around 97% of GC-skewed promoters are located at CGIs, and 67% of R-loop enriched sites are found in these GC-skewed CGI promoters (Ginno, et al., 2012). Likewise, GC-skew at the terminal region of the gene was also found to associate with increased R-loops (Ginno, et al., 2013). Since the most active genes are constitutively active “housekeeping” genes, R-loops are also associated with these “housekeeping” genes such as genes with functions in cellular metabolic processes and translation elongation (Ginno, et al., 2013).

In mammals, the CGIs are scarce in DNA methylation (m^5C), which acts as a heritable transcriptional silencer, and are predominantly found at the 5' ends of genes where they act as promoters. The more skewed the GC skew is, the less DNA methylation is observed at the transcriptional start site (TSS) (Ginno, et al., 2012). Furthermore, the transcriptional activity of the CGI promoters itself protects DNA from methylation (Bird, 2002), as DNA methylation follows the transcriptional inactivity of a gene (Bachman, et al., 2003). These findings indicate that R-loops and DNA methylation are anti-correlated.

The “thread back” model (Aguilera & Garcia-Muse, 2012) provides the most accepted explanation for how R-loops are formed. It suggests that G-rich nascent RNA produced during transcription anneals to the single-stranded C-rich template DNA. Since the nascent RNA leaves the RNA polymerase II at a site far away from the DNA strands, it would typically be outcompeted by the non-template DNA strand (coding strand) for binding to the template DNA strand due to its closer proximity. However, G-rich clusters of the nascent RNA have a higher affinity for binding to the template DNA strand and can outcompete the

much closer coding strand (Roy & Lieber, 2009). While both the nascent RNA and the non-template strand will have G-rich clusters, Hung, et al. (1994) showed that ribose-purine:deoxyribose-pyrimidine paired molecules display higher levels of thermal stability than deoxyribose-purine:deoxyribose-pyrimidine base pair, thus explaining why G-rich RNA can outcompete G-rich DNA. R-loops are unlikely to form when the template strand is G-rich, as ribose-pyrimidine:deoxyribose-purine base pairing is thermally least stable out of all possible combinations (Hung, et al., 1994), highlighting the importance of GC skews in the formation of R-loops.

The formation of R-loops during transcription results in *cis* R-loops (**Figure 1.6**), where the guanine rich RNA strand anneals to the cytosine rich template DNA strand, and are the most commonly observed type of R-loops (Frederic, 2016). Little is known about *trans* R-loops, where the RNA binds to the non-template G-rich strand (**Figure 1.6**). They were first artificially generated and confirmed by Wahba et al. (2013) in yeast. *trans* R-loops are believed to form non-co-transcriptionally, and recent studies show that they do occur naturally (Nadel, et al., 2015).

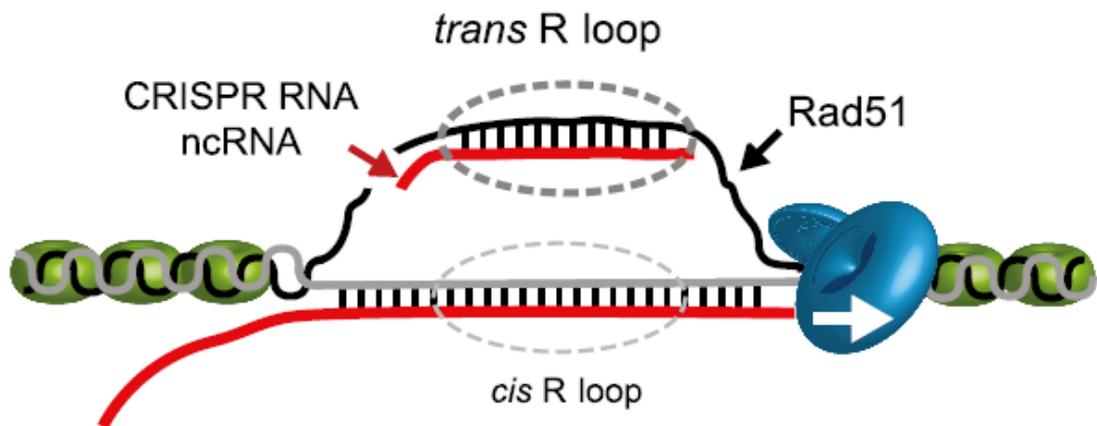


Figure 1.6 Theoretical structure of R-loop *in cis* and *in trans*. Rad51-dependent *trans* R-loops were observed by Wahba, et al. (2005). CRISPR-cas9 functions by hybridizing its guide RNA onto matching DNA to cause DNA strand breaks. Image taken from Skourti-Stathaki & Proudfoot (2014).

Specific instances of DNA:RNA hybrids have been identified that are formed non-co-transcriptionally. DNA negative supercoiled stress can promote R-loop formation. Similar to the unwinding of DNA during transcription, negative supercoiling can unwind the DNA to produce ssDNA that allows RNA to bind and form R-loops. This is supported by the fact that *topA* (topoisomerase enzyme that helps to remove negative supercoiling of the DNA) *E. coli* mutants have increased R-loops, which are suppressed by RNase H (DNA:RNA specific nuclease) overexpression (Drolet, et al., 1995).

There are a few cases where R-loops could be formed in the absence of active transcription, for example, as part of the DNA damage response. Kasahara, et al. (2000) artificially formed R-loops *in vitro* in the absence of transcription by utilizing the nucleic acid binding function of RecA, a protein related to DNA repair and maintenance (Kasahara, et al., 2000). In addition, Ohle, et al. (2016) found that R-loops form as part of the homologous-recombination mediated double-stranded repair and are essential for DNA repair.

1.3.2. Preventing R-loop formation

R-loop formation needs to be carefully controlled in order to avoid R-loop accumulation. Various mechanisms have been proposed that play a role in preventing R-loop formation. The binding of proteins (eukaryotes) and ribosomes (prokaryotes) to the nascent RNA can reduce the formation of R-loops. They act as a physical hindrance preventing the nascent RNA and ssDNA strand from hybridizing (**Figure 1.7**) (Hamperl & Cimprich, 2014; Garcia-Benitez, et al., 2017). The transport complex THO/TREX (Transcription export complex) for example immediately binds to nascent RNA. Since its function is to move the RNA outside the nucleus, it could have an additional effect in preventing nuclear RNA accumulation, which can further reduce R-loop formation (Dominguez-Sanchez, et al., 2011). It is unclear whether the prevention of RNA accumulation through transport or the direct binding of THO/TREX to the nascent RNA is the main factor preventing R-loop formation. Even if ssRNA accumulates in the nucleus, many RNA binding molecules can bind to the RNA, forming ribonucleoproteins (RNPs), and modify the RNA (Glisovic, et al., 2008), which disfavors reannealing of the RNA to the DNA (Santos-Pereira & Aguilera, 2015; Garcia-Benitez, et al., 2017). Similarly, the RNA (or RNP) will fold into an energetically favourable structure, making reannealing to DNA more difficult. Likewise, accessibility to open DNA is also vital for R-loop formation. Negative supercoiled DNA is susceptible to unwinding, which provides the accessibility for RNA binding.

Overexpression of DNA gyrase, which functions to alleviate positive supercoiling, can induce negative supercoiling and enhance R-loop formation (Drolet, et al., 1995; Drolet, et al., 2003). Enzymes like DNA topoisomerase 1 (TOP1) exist that relax negative DNA supercoiling, thus reducing open DNA double helix structures and diminishes accessibility of open annealing sites for ssRNA (Drolet, et al., 1994). Absence of DNA topoisomerase I in human cells and *E. coli* results in DNA instability and reduced growth, respectively, which can be rescued by overexpression of RNase H (Masse & Drolet, 1999; Tuduri, et al., 2009), or compensated by mutating DNA gyrase (Drolet, et al., 1995).

1.3.3. Structure and stability

Once formed, R-loops are thermodynamically more stable than their DNA:DNA double helix counterpart (Ratmeyer, et al., 1994). They are proposed to form a more stable conformation that is intermediate between the A and B form of the double helix (Shaw & Arya, 2008). The formation of a stable G-quadruplex structure by the non-template ssDNA strand, called a G-loop (Duquette, et al., 2004), could also enhance the stability of R-loops. As previously noted, a large GC skew would both enhance the thermodynamic stability of DNA:RNA hybrids (Ginno, et al., 2012) and also help G-loop formation on the displaced ssDNA.

Despite the fact that R-loops are more stable than their dsDNA counterpart, they are resolved relatively quickly. They exist as a transient structure whose quantity is kept at equilibrium through constant R-loop formation and removal events. It has been shown that blocking transcription by 5,6-dichloro-1- β -D-ribofuranosyl-Benzimidazole (DRB) resolves most R-loops within 30 minutes with an average half-life of 10 minutes. DRB inhibits cyclin-dependent kinase 9 (CDK9), which phosphorylates the C-terminal domain of RNA polymerase II required for elongation initiation. Upon removal of DRB, the reappearance of R-loops was observed within 10 minutes, showing the short turnover rate of this hybrid structure (Sanz, et al., 2016).

1.3.4. Resolution of R-loops

Currently, two types of proteins have been identified that can resolve R-loops once they have formed: nucleases and helicases. Nucleases are enzymes that degrade the phosphodiester bonds binding nucleotides together. The endonuclease RNase H enzyme, which is found in nearly all organisms, specifically targets and resolves DNA:RNA hybrids. It only degrades the RNA part of the DNA:RNA hybrid leaving the DNA portion intact (Cerritelli & Crouch, 2009; Ohle, et al., 2016). Another enzyme is the Mung Bean Nuclease. While it strongly prefers single-stranded nucleic acids, it is able to cleave double-stranded nucleic acid, including the DNA:RNA hybrid (Takara, n.d.). Similarly, the Exonuclease III from *E. coli* can degrade both strands in the hybrid (Keller & Crouch, 1972; Valsala & Sugathan, 2017). The nuclease from the bacteria *Serratia marcescens* (also known as Benzonase) is a non-specific endonuclease that can degrade both DNA and RNA in double and single-stranded form. Its potency suggests its use as a scavenging enzyme released outside the bacteria rather than being used inside the nucleus (Benedik & Styrch, 1998). The unwinding of the DNA:RNA

helix by a group of DNA:RNA helicases is another way to resolve R-loops. A number of such helicases have been found across many organisms. These including RecG (Hong, et al., 1995), DHX9 (Chakraborty & Grosse, 2011) and Senataxin (SETX) (see **Figure 1.7a**) (Kim, et al., 1999; Becherel, et al., 2013). Senataxin has been identified to act on terminal R-loops in human cell lines (Skourti-Stathaki, et al., 2011) (for more detail on helicases refer to **Section 1.5**). In *C. elegans*, WRN-1, a RecQ helicase, has been shown to have the ability to resolve R-loops (Hyun, et al., 2008) (see **Section 1.5.2** for more details).

Although a number of proteins have been identified to reduce R-loop levels, the exact mechanisms and the choice of the protein remains to be determined. Resolving by RNase H and helicase result in different outcomes. With RNase H, the RNA part of the helix is cleaved into ribonucleotides, leaving the DNA strand intact (New England Biolabs, no date). Helicase unwinding could rescue the “stuck” ssRNA and allow it to continue with post-transcriptional processing.

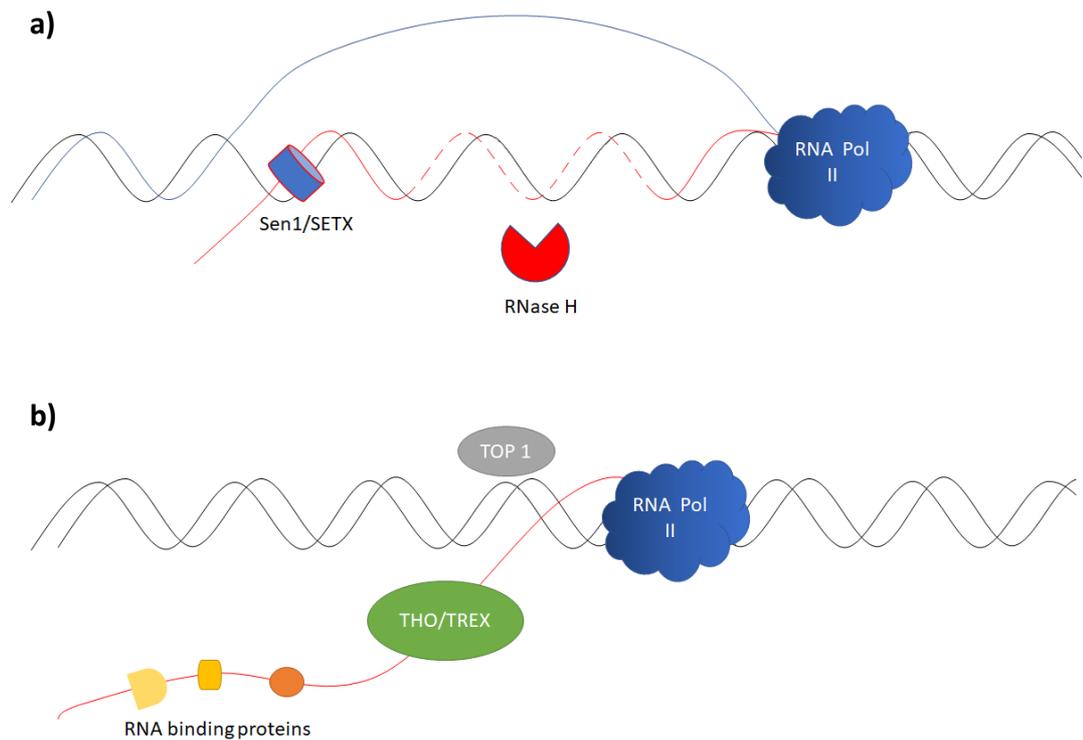


Figure 1.7 Management of R-loop formation. R-loop levels are maintained by (a) the removal of existing R-loops through RNase H digestion or helicase unwinding, and (b) the prevention of formation of R-loops by RNA binding proteins and topoisomerase rewinding. Image based on Santos-Pereira & Aguilera (2015).

1.4. Function and effect of R-loops

R-loops are naturally found in a variety of organisms, making up as much as 5% of the mammalian genome (Sanz, et al., 2016). DNA:RNA hybrids are generated as intermediates during replication, genome rearrangement and gene expression (Aguilera & Garcia-Muse, 2012; Santos-Pereira & Aguilera, 2015). Whether R-loops have an intended function or are unplanned consequences of chemical and physical laws is unknown. Regardless of either possibility, they have a considerable effect on the organism at the genetic level, with far-reaching health implications.

1.4.1. Transcription regulation of R-loops

The most well-documented aspect of R-loops is their association with transcription regulation. Multiple studies identified that R-loops impair transcription. The defects in transcription of *E. coli topA* mutant and yeast *hpr1* (subunit of the THO complex) mutant are attributed to the accumulation of R-loops, where overexpression of RNase H was able to overcome the defect (Aguilera & Huertas, 2003; Baaklini, et al., 2004). The mechanism by which R-loops impair transcription is that they stop or slow the transcriptional elongation by acting as a roadblock for incoming RNA polymerases II. Since multiple RNA polymerase II can transcribe a gene concurrently, R-loops formed as a result of the nascent RNA produced by one RNA polymerase II would inhibit the elongating RNA polymerase II behind it (Aguilera & Huertas, 2003). Another non-exclusive hypothesis proposed by Tous & Aguilera (2007), suggests that RNA overhangs of the R-loop could establish undesired contact with the C-terminal domain of the RNA polymerase II, thereby triggering a checkpoint mechanism to inhibit elongation (Tous & Aguilera, 2007).

R-loops have also been proposed to have transcription supporting features. R-loops formed at transcription termination pause sites can recruit the helicase Senataxin, which helps the separation of the nascent RNA from the RNA polymerase II and promote efficient Xrn2-dependent transcription termination (Skourti-Stathaki, et al., 2011). This transcription termination may be highly crucial since terminal R-loops are found at gene-rich sites, and immediate termination prevents transcriptional read-through that could unintentionally express downstream genes (Santos-Pereira & Aguilera, 2015). Another line of evidence supporting the transcription termination role of terminal R-loops is that they induce antisense-transcription leading to dsRNA formation, which reinforces RNA Polymerase II pausing and transcription termination (Skourti-Stathaki, et al., 2014). This is further

supported by the observation that G-quadruplexes are found enriched at 3'-UTR of genes in gene-rich regions (Huppert, et al., 2008), which could be formed from the displaced ssDNA as a result of R-loop formation.

1.4.2. R-loop dependent DNA methylation state

CGIs function as promoters for the majority of constitutively active genes and are devoid of DNA methylation (Bachman, et al., 2003; Illingworth & Bird, 2009), but enriched with R-loops (Ginno, et al., 2012). This correlation incited the search for a link between DNA hypomethylation and R-loop formation. Ginno et al. (2012) found that the GC skew of CGIs plays a role in predicting DNA methylation and subsequently observed that R-loop formation protected DNA from methylation by the primary *de novo* DNA methyltransferases 3B1 (DNMT3B1) and the DNMT3A stimulating factor DNMT3L. The mechanism behind this is hypothesized to be due to R-loop induced transcriptional pausing (Ginno, et al., 2012). RNA polymerase II occupancy on the DNA (regardless of being active or stalled) has been shown to be sufficient to prevent DNA methylation, likely by hindering DNA methylation machinery binding (Takeshima, et al., 2009).

R-loops have not only been associated with preventing DNA methylation but also with DNA demethylation. GADD45A is a stress response protein that is linked with DNA demethylation, as overexpression of this protein promotes global DNA demethylation (Barreto, et al., 2007). GADD45A recruits the methylcytosine dioxygenase (an enzyme that oxidates the methyl groups on cytosine) Ten-Eleven Translocation 1 (TET1) that demethylates the DNA at specific CGI promoters. GADD45A is able to bind R-loops and removal of R-loops by RNase H reduces GADD45A binding and DNA methylation at a subset of TET1-targeted gene promoters (Arab, et al., 2019).

1.4.3. Histone modifications and chromatin compaction

R-loops are linked with various histone modifications and are implicated in affecting chromatin compaction and transcriptional activity. Below is a summary of histone modifications that are associated with R-loops.

1.4.3.1. H3K4me3

H3K4me3 is found at the TSS (Heintzman, et al., 2007) while R-loops are found to peak downstream of the TSS, as much as 1.5kb downstream in humans (Sanz, et al., 2016). Both

of them also have a direct correlation with transcriptional activity, with highly transcribed genes showing higher enrichment of H3K4me3 and R-loops, while weakly transcribed or silenced genes have low occupancy of both (Barski, et al., 2007; Kuznetsov, et al., 2018; L. Chen, et al., 2017; Ginno, et al., 2012). Several members of methyltransferases (trithorax and polycomb family) contain SET domains which function in methylating specifically the lysine residue of histones associated with active transcription (Dillon, et al., 2005). These SET domains can bind to ssDNA and ssRNA (Krajewski, et al., 2005). The displaced ssDNA and the ssRNA overhang of the R-loop can be potential targets for these methyltransferases, thereby increasing histone lysine methylation.

Ginno et al. (2012) proposed a model in which the displaced ssDNA portion of the R-loops recruit H3K4 methyltransferases, based on the findings that the SET-domain of H3K4 methyltransferases can bind ssDNA, which then deposits the H3K4me3 mark. Since R-loop exists transiently, with a half-life of 10 minutes (Sanz, et al., 2016), the deposition of H3K4me3 induced by R-loops could act as a long-lasting marker of active genes, that does not directly hinder transcription by RNA polymerase unlike R-loop does. In addition, the H3K4me3 mark has the added advantage of being heritable, (Moazed, 2011; Whitehouse & Smith, 2013), indirectly preserving the information of R-loop prone sites.

1.4.3.2. H3K9 methylation

The *C. elegans* double mutants lacking the H3K9 methyltransferases *met-2* and *set-25* show no detectable levels of H3K9 methylation marks (H3K9me1/2/3). These worms also have enhanced R-loop accumulation, comparable to mutants defective for the THO/TREX complex (*thoc-2*). Specifically, tandem repeats that are derepressed in the absence of H3K9 methylation show particularly high R-loop enrichments (Zeller, et al., 2016).

The repressive H3K9me2 modification is deposited by the G9a histone methyltransferase in humans. H3K9me2 is found to be recruited by dsRNA generated as a result of terminal R-loop formation to help transcription termination (Skourti-Stathaki, et al., 2014). Consequently, H3K9me2 recruits heterochromatin protein 1 γ (HP1 γ) that plays a role in heterochromatin formation (Skourti-Stathaki, et al., 2014). H3K9me2, however, has no influence on R-loop formation (Groh, et al., 2014).

1.4.3.3. H4R3me2 and H3R17me2

Dimethylation marks on H4R3 and H3R17 are found to inhibit R-loops at the *c-MYC* locus in human cell lines. These marks, especially H4R3me2, are recognized by the Tudor domain of Tudor Domain Containing 3 (TDRD3) which in turn recruits topoisomerase III β

(TOP3B). TOP3B relaxes negative supercoiling during transcription of the *c-MYC* gene, thereby discouraging the formation of R-loops (Yang, et al., 2014).

1.4.3.4. H3K79me2, H4K20me1 and H3K27me3

Both H3K79me2 and H4K20me1 are linked with R-loops due to the occupancy of them at the R-loop prone CGIs (Ginno, et al., 2013). Similar to H3K4me3, these two histone modifications are also linked to active transcription (Santos-Pereira & Aguilera, 2015). Apart from their co-occupancy with R-loops, there is limited research regarding their relationship.

H3K27me3, on the other hand, is enriched at CGIs with a reverse GC skew, suggesting that H3K27me3 is found at CGI sites unlikely to form R-loops, thus implying an inverse relationship. Ginno et al. (2013) hypothesise that since H3K27me3 and DNA methylation are mutually exclusive, that H3K27me3 acts as a substitute for R-loops to protect these CGIs with reverse GC skew from DNA methylation by recruiting polycomb complexes (Ginno, et al., 2013).

1.4.3.5. H3 acetylation and open chromatin

The characteristics associated with active transcription, such as CGI, H3K4me3 and DNA hypomethylation, are found in open chromatin regions. Therefore it suggests that R-loops are also linked with open chromatin. Indeed, acetylated H3K9 and H3K27 are enriched at promoter R-loops, which promotes an open chromatin state (Sanz, et al., 2016).

R-loops may prevent chromatin compaction, as unresolved R-loops might hinder the DNA wrapping around the histones (Dunn & Griffith, 1980), especially if the displaced ssDNA forms G-quadruplex structures. Indeed, an increase in overall chromatin accessibility has been observed through increased DNase accessibility, reduced MNase accessibility and FAIR-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) signal overlapping with R-loop signals (Tsompana & Buck, 2014; Sanz, et al., 2016). Reduced nucleosome density is observed with increased R-loop formation by ncRNA invasion, while reduced ncRNA invasion enhances chromatin compaction (Boque-Sastre, et al., 2015).

1.4.3.6. H3S10P

Phosphorylation of H3S10 (H3S10P), which is associated with chromatin condensation and compaction, has also been linked with R-loops as both are found elevated at the centromere, pericentromeres and a large number of open reading frames. RNase H overexpression found

decreased levels of H3S10P, suggesting that R-loops trigger this histone phosphorylation modification (Castellano-Pozo, et al., 2013).

The observation that R-loop triggers H3S10P and chromatin condensation oppose the idea that R-loops are associated with active transcription and open chromatin state. It could be possible that R-loop can promote both active and inactive chromatin states in a context-dependent manner.

1.4.4. DNA instability

Aberrant R-loops per se do not cause damage directly to the DNA but instead by subsequent mechanisms that result in various genetic implications such as mutagenesis, hyperrecombination, rearrangements and transcription/replication collisions (Hamperl & Cimprich, 2014; Sanz, et al., 2016; Garcia-Picardo, et al., 2017). For example, the R-loop dependent DNA breaks observed in THO-depleted human cells are linked to replication failure due to R-loop formation, rather than R-loops themselves (Dominguez-Sanchez, et al., 2011).

The displaced ssDNA in the R-loop is chemically unstable and more susceptible to DNA damage (Lindahl, 1993; Beletskii & Bhagwat, 1996), but can be more stable if it forms G-quadruplexes (Lane, et al., 2008) as the non-template strand is already G-rich. Indeed, Huppert, et al. (2008) observed overrepresentation of G-quadruplexes near the 5'-UTR and 3'-UTR (Huppert, et al., 2008), corresponding to regions of elevated R-loops. Single-stranded breaks (SSB) are the most prominent problems associated with the displaced ssDNA. One proposed mechanism is the recruitment of Activation-Induced Cytidine Deaminase (AID), an enzyme that can deaminate cytosine to uracil on ssDNA (Muramatsu, et al., 2000), which can then be processed by base excision repair and abasic endonuclease to create a SSB or abasic site (a site where the nucleotide lost its base) (Di Noia & Neuberger, 2002). This mechanism, however, is unlikely to happen at R-loops formed at GC skewed sites because the displaced ssDNA is G-rich and C-poor. Furthermore, Garcia-Picardo et al., (2017) were unable to measure increased AID dependent ssDNA damage in R-loop-accumulating mutants (Garcia-Picardo, et al., 2017). The argument that AID is recruited by R-loops has been discussed by Pacri (2017), who concluded that current scientific evidence does not support that R-loops per se recruit AID (Pacri, 2017). A G-quadruplex can also lead to SSB despite its stability, as there are human nucleases that specifically cleaves DNA G-quadruplexes regardless of its sequence (Sun, et al., 2001).

Double-stranded breaks (DSB) have seen an increase in R-loop accumulating mutants. The mechanism by which R-loops contribute to DSB is proposed by the transcription or replication machinery colliding with the R-loop structure during elongation, as these DNA:RNA hybrids present a physical hindrance to the elongation mechanisms (Gan, et al., 2011; Hamperl & Cimprich, 2014; Sollier & Cimprich, 2015).

1.4.5. DNA repair

DSB repair by homologous recombination (HR) includes a step of generating single-stranded DNA overhang that is later used for strand invasion into the sister chromatid. This unstable ssDNA overhang is stabilized through the binding of replication protein A (RPA) until it is replaced by Rad51, which is essential for homology search (Ohle, et al., 2016). Ohle et al. (2016) suggest that RNA polymerase II transcribes the ssDNA overhang before being occupied by RPA, generating nascent RNA that then competes with RPA for binding onto the ssDNA overhang. Both overexpression and deletion of RNase H1 in *S. pombe* inhibit HR, while the wild-type showed complete DSB recovery. The absence of RNase H stabilizes R-loops, thus impairing RPA binding and preventing Rad51 recruitment. This results in stalling of the HR repair process. On the other hand, when R-loop accumulation is reduced by the overexpression of RNase H, repeat regions around the DSB become destabilized. In addition, there is excessive recruitment of RPA that result in long ssDNA that will be prone to excessive strand resection. As such, it suggests that R-loops are required transiently to manage HR repair (Ohle, et al., 2016).

1.4.6. Health implications

The formation of R-loops at the 5'-UTR of the Fragile X mental Retardation 1 (FMR1) gene causes reduced expression, ultimately leading to the two neurodegenerative diseases fragile X syndrome and fragile X-associated tremor/ataxia syndrome (Loomis, et al., 2014; Groh, et al., 2014). Many genetic disorders have GC-rich trinucleotide repeat expansions that have the potential to form R-loops. For example, the trinucleotide repeats CAG (associated with spinocerebellar ataxia) and CGG (associated with fragile X syndrome type A) are observed to form R-loops during *in vitro* transcription (Reddy, et al., 2011). Aicardi-Goutières Syndrome has been linked with R-loops, as primary cells isolated from patients with this syndrome contain elevated levels of R-loops (Lim, et al., 2015).

Cancer has been suggested as a possible result of R-loop accumulation, due to the effect of R-loops on genome stability and replication stress. Cells infected with cancer-causing viruses such as the Kaposi's-sarcoma associated herpesvirus have increased R-loop levels (Santos-Pereira & Aguilera, 2015). Genes mutated in cancer have also been associated with R-loops. For example, the tumour-suppressor genes *BRC A1* and *BRC A2* in humans also prevent R-loop accumulation (Bhati, et al., 2014; Sollier & Cimprich, 2015). *BRC A1* and *BRC A2* are part of the Fanconi Anemia mediated DSB repair pathway and many of the DNA breaks found in Fanconi Anemia disrupted cells are R-loop dependent, providing a link to the disease of the same name (Garcia-Rubio, et al., 2015). In mice, R-loop accumulation at the proto-oncogene *c-MYC*, as a result of TDRD3 mutation, results in DNA damage that leads to *c-Myc/Igh* translocation. TDRD3 forms a complex with topoisomerase IIIB (TOP3B) to relax negative supercoiling and reduce R-loop formation. The *c-Myc/Igh* translocation is associated with *c-MYC* misregulation and is commonly observed in lymphomas (Yang, et al., 2014).

1.5. Helicases

Helicases are found and classified by sequence homology, specifically by the Walker A and B motif. These two motifs allow for ATP binding and ATP hydrolysis and are required to move the enzyme forward on the nucleic acid strand (Caruthers & McKay, 2002). Helicases can form oligomeric structures that enhance their activity (Patel & Donmez, 2006). A relatively large part of the eukaryotic genome (approximately 1% of the protein-coding genes) encodes helicases (Wu, 2012) and are thus among the largest class of proteins (Jankowsky & Fairman-Williams, 2010). The human genome, for example, encodes 95 non-redundant (based on sequence similarity) helicases with 64 classified as RNA helicases and 31 classified as DNA helicases (Umate, et al., 2011). Although their classification is based on their substrate, some of them are not limited to only one type of nucleic acid and can function on both DNA and RNA, as well as DNA:RNA hybrids (Singleton, et al., 2007; Wu, 2012).

The mechanism by which helicases unwind is still unknown. Two models hypothesise a possible mechanism. The passive model takes the context of the enzyme as a catalyst that reduces the activation energy, the energy required to separate the two strands of nucleic acid. The reduced activation energy threshold allows smaller fluctuation of thermal/kinetic energy to be sufficient to overcome the bond energy and breaking the hydrogen bond holding the two strands together. After the helix separates, the helicase traps the nucleic acid in their single-stranded state, thereby preventing them from reannealing. The other model, known

as the active model, hypothesizes that the energy from the enzyme moving along the DNA is sufficient enough to destabilise and force apart the double-stranded nucleic acid, i.e. the helicase has a “blade” that cuts the hydrogen bonds holding the strands together as it moves forward (Manosas, et al., 2010). This mechanism fits well to the observations and measurements of several helicases (Singleton & Wigley, 2002; Raney, et al., 2013).

The below section provides more detail about the types of helicases and gives an overview of known helicases that are able to resolve R-loops in a variety of mechanisms, not limited to unwinding.

1.5.1. Types of helicases

Currently, all helicases are assigned to one of six superfamilies (SF1-6) based on their sequence and structure similarity (Gorbalenya & Koonin, 1993; Singleton, et al., 2007). SF1 and SF2 are characterized by a helicase core that is formed from two nearly identical recA folds and are active as either monomer or dimer (Tanner & Linder, 2001). They also share a large number of motifs at the same location within the gene (Singleton & Wigley, 2002; Singleton, et al., 2007; Fairman-Williams, et al., 2010). These two superfamilies contain most of the known helicases, with SF2 being the largest superfamily by far. Helicases from SF3 to SF6, on the other hand, have only one recA fold and form hexameric rings with five others (Singleton, et al., 2007). The superfamilies are further divided into families, that are classified by the enzyme’s functional characteristics and motifs (Jankowsky & Fairman-Williams, 2010). This classification, however, does not differentiate between RNA and DNA helicases, as both these classes exist in all superfamilies except for SF6, which only contains DNA helicases.

1.5.1.1. SF1 helicases

SF1 and SF2 share many motifs, only some of which are used to distinguish between these two superfamilies such as the motif III, which is diagnostic of SF1 helicase. Three families are associated with SF1: Rep/UvrD, Pif1/RecD, and Upf1-like (**Figure 1.8**) (Fairman-Williams, et al., 2010; Raney, et al., 2013). The Rep/UvrD family is mostly associated with DNA repair and maintenance (Gilhooly, et al., 2013). Pif1/RecD family helicases have roles maintaining nuclear and mitochondrial DNA by resolving G-quadruplexes and help in directing and maintaining Okazaki fragments during DNA replication (Bochman, et al., 2010). Upf1-like helicases are associated with various roles within transcription-associated

events, such as mRNA quality control (nonsense-mediated decay) (Chang, et al., 2007) and R-loop removal (Mischo, et al., 2011).

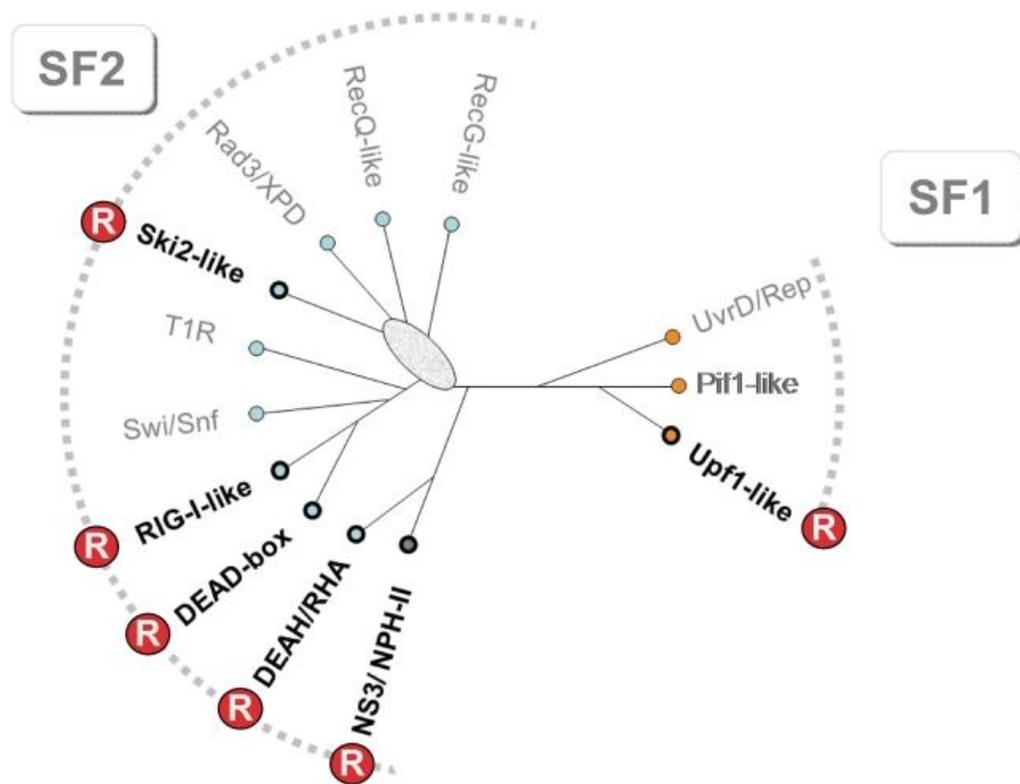


Figure 1.8 Families of the SF1 and SF2 superfamily. Their evolutionary relationship is represented by an unrooted cladogram based on sequence homology from human and *S. cerevisiae*. Highlighted families marked with “R” contain RNA-helicases. Other families contain DNA helicases. Image taken from Jankowsky, et al. (2011).

1.5.1.2. SF2 helicases

SF2 helicases are subdivided into many families, five of which are RNA specific helicase (DEAD-box, RIG-I like, DEAH/RHA, Ski2-like and NS3/NPH-II), collectively called DExD/H helicases. DExD/H helicases share the conserved motif DEAD/DEAH and other closely related DExD/H variants (where “x” can be any amino acid) (Jankowsky & Fairman-Williams, 2010; Bryd & Raney, 2012). They are found to be active in all RNA related interactions, including DNA:RNA hybrids (Tanner & Linder, 2001). The other families of SF2 encompass single and double-stranded DNA translocases and helicases (**Figure 1.8** and **Table 1.2**) (Beyer, et al., 2013).

Family	Function	Examples
DEAD-box	dsRNA unwinding	eIF4a Mss116p Ded1p DbpA
DEAH/RHA	ssRNA translocase dsRNA unwinding	Prp16p Prp22p Prp43p
RecQ-like	ssDNA translocase dsDNA unwinding	BLM WRN RecQ1 RecQ4
Rad3/XPD	ssDNA translocase dsDNA unwinding	XPD
Swi/Snf	dsDNA translocase	CSB ATRX INO80 ISWI Rad54 SWI2/SNF2
RIG-I like	dsRNA translocase	DICER RIG-I
Ski-2 like	ssRNA translocase dsRNA unwinding	Ski2p Mtr4
RecG-like	dsDNA translocase branched DNA unwinding	RecG PriA
NS3/NPH-II	ssRNA translocase dsRNA unwinding	NS3 NPH-II

Table 1.2 Function of the families of the SF2 helicase superfamily. The right-most column includes examples for each family. Table based on Bryd & Raney (2012).

SF2 helicases also have a diverse set of functions other than unwinding nucleic acid, mainly owing to the wide range of accessory domains. These functions include chromatin remodelling and peptide export (Beyer, et al., 2013). The RecG and RIG-I family functions in resolving a variety of branched nucleic acid structures which include Holliday junctions, D-loops and R-loops (Rudolph, et al., 2010; Beyer, et al., 2013). The RecQ family can also resolve triple helices and G-quadruplexes. This family is of particular interest for their association to a variety of diseases such as cancer and Bloom syndrome (Beyer, et al., 2013).

The Swi/Snf family are also classified as chromatin remodelers owing to their ability to alter the chromatin landscape when they form an 8-14 subunit multiprotein complex (Zhang, et al., 2006). Family members include ISWI, CHD and INO80 (**Table 1.2**).

1.5.1.3. Hexameric helicases (superfamily 3-6)

Hexameric helicases have been mainly found in viruses and bacteria (Jankowsky & Fairman-Williams, 2010). Little is known about these helicases. SF3 helicases are found in small DNA and RNA viruses (e.g. Similan virus, Human Papillomavirus and Adeno-associated virus). They have four conserved motifs A, B, B', and C, of which the former two are the Walker A and B equivalent and the last motif is SF3 specific, which is suggested to act as a sensor of ATP binding and hydrolysis. The crystal structure of only a few SF3 helicases have been solved, showing that they form hexamers and double hexamers. Their function has so far only been associated with viral replication (Hickman & Dyad, 2005; Singleton, et al., 2007; Jankowsky & Fairman-Williams, 2010; Tuma, 2010).

SF4 helicases are structurally very similar to SF3 and are found in bacteria and bacteriophages (Jankowsky & Fairman-Williams, 2010). They have five motifs (H1, H1a, H2, H3, and H4) of which two are the Walker A (H1) and B (H2) equivalent, while the rest are specific to SF4 (Singleton, et al., 2007). The H3 motif is thought to be the functional equivalent of motif C from helicases of SF3. All the currently identified SF4 helicases have primary roles in DNA replication (Tuma, 2010). The primase from bacteriophage T7 is one of the most extensively studied SF4 helicase. It unwinds the DNA for replication fork progression and is vital for inducing Okazaki fragment synthesis (Frick & Richardson, 2001). The closely related P4 helicase from dsRNA bacteriophages plays a role in RNA packaging into viral capsids and dsRNA unwinding. RNA helicases belonging to SF3 and SF4 have only been found in viruses (Jankowsky & Fairman-Williams, 2010).

SF5 is closely related to SF4 and contains the Rho proteins that were initially identified in *E. coli*. This superfamily of helicases is responsible for transcription termination by binding to the nascent RNA and dissociate the RNA polymerase II. It prevents the aggregation of R-loops by either blocking the formation or unwinding already formed hybrids (Gogol, et al., 1991; Singleton, et al., 2007; Jankowsky & Fairman-Williams, 2010).

SF6 helicases are all DNA helicases (Jankowsky & Fairman-Williams, 2010). They have a core AAA+ (ATPases associated with diverse cellular activities) fold, which is used for energy conversion from chemical (ATP) to forward motion along the nucleic acid strand. The eukaryotic MCM helicase is an example of an SF6 helicase and is essential for DNA replication initiation and elongation. The three prokaryotic RuvA/B/C helicases are responsible for resolving Holliday junctions during homologous recombination (Erzberger & Berger, 2006; Singleton, et al., 2007).

1.5.2. Known helicases that resolve R-loops

A number of helicases have been found to resolve DNA:RNA structures either *in vivo* or *in vitro*. Within the SF1, human and mouse Senataxin (SETX) and its ortholog SEN1 (a member of the Upf-1 like family) in *S. cerevisiae* and *Schizosaccharomyces pombe* have been shown to unwind DNA:RNA hybrids *in vivo* (Kim, et al., 1999; Mischo, et al., 2011; Skourti-Stathaki, et al., 2011; Becherel, et al., 2013). *S. cerevisiae* Pif1 has shown the ability to unwind DNA:RNA hybrids *in vitro* and has a stronger preference and higher unwinding rate for the hybrids compared to dsDNA (Boule & Zakian, 2007). The human Aquarius (AQR) helicase, belonging to the same subfamily as SETX, is essential for preventing R-loop mediated DSB (Sollier, et al., 2014).

SF2 RNA helicases (DExD/H helicases) are hypothesized to have extensive DNA:RNA binding capability. Human DEAH-box DHX9, also known as RNA helicase A (RHA), has the ability to unwind R-loops *in vitro* (Chakraborty & Grosse, 2011). However, the same lab also found that DHX9 can help the formation of some R-loops in the absence of splicing factors (Chakraborty, et al., 2018). The DEAD-box helicase DDX21 can reduce R-loops both *in vitro* and *in vivo* (Song, et al., 2017). DDX19, normally functioning in the export of mRNA, can unwind R-loops *in vitro* and plays a role in R-loop resolution *in vivo*, potentially as part of the DNA damage response during DNA replication (Hodroj, et al., 2017). The DEAH-box helicase FANCM, associated with genome stability, can unwind R-loops *in vitro* (Schwab, et al., 2015). Within the SF2 DNA helicases, RecG from *E. coli* has been shown to be required for maintaining low R-loop levels (Hong, et al., 1995). The Human RecQ helicase WRN-1 and the *C. elegans* orthologue WRN-1 have the ability to unwind R-loops *in vitro* (Hyun, et al., 2008; Chakraborty & Grosse, 2011). Another RecQ helicase, BLM, has DNA:RNA helicase activity *in vitro* (Grierson, et al., 2013), while the *S. cerevisiae* orthologue SGS1 is required to avoid R-loop accumulation *in vivo* (Chang, et al., 2017). Cas3, an ssDNA nuclease, has been shown to disassemble R-loops in the presence of ATP but promotes their formation in the absence of ATP *in vitro* (Howard, et al., 2011).

Viral SF2 helicases with the ability to unwind DNA:RNA hybrids have also been identified. The nonstructural protein 3 (NS3) from the Hepatitis C virus has been shown to have the capability to unwind the hybrids *in vitro* (Gwack, et al., 1997; Pang, et al., 2002). UvsW helicase, one of the three helicases found in Bacteriophage T4, unwinds R-loops both *in vitro* and *in vivo* (Mosig, et al., 1995; Dudas & Kreuzer, 2001). UvsW has been suggested to be functionally analogous to *E. coli* recG as ectopic expression of UvsW in recG *E. coli* mutants can rescue the bacteria (Carles-Kinch, et al., 1997).

Within the superfamily 5, the well studied Rho helicase has the ability to unwind R-loops. Its role in transcription termination by dissociating the nascent RNA is crucial for the survival of *E. coli* (Gogol, et al., 1991; Singleton, et al., 2007; Jankowsky & Fairman-Williams, 2010). Interestingly, an otherwise lethal mutant lacking the Rho helicase can be rescued through the ectopic expression of UvsW from Bacteriophage T4 (Leela, et al., 2013).

1.5.3. Other functions of helicases that can affect R-loops

Although helicases are known for their nucleic acid unwinding characteristics, not all proteins classified as helicases are able to unwind nucleic acids and are better understood as translocases, that mainly function as motors to drive a unidirectional movement (Singleton, et al., 2007). This discrepancy stems from the method of classifying helicases based on their conserved amino acid sequence, especially the Walker A and Walker B motif (Jankowsky & Fairman-Williams, 2010). For example, some of these “helicase-classified translocases” lack a wedge domain (“blade”) used to break the hydrogen bonds and force open the double helix (Saha, et al., 2006). Furthermore, as previously noted, helicases (and translocases) have a wide variety of accessory domains that give specific helicases unique functions (Beyer, et al., 2013). Here, other possible functions of helicases are discussed that can affect R-loop accumulation independent of the unwinding mechanism summarized in the previous section.

1.5.3.1. Chromatin remodelers

There are at least four families of ATP dependent chromatin remodelers: SWI/SNF, ISWI, NURD/Mi-2/CHD and INO80. They all belong to the helicase SF2 and contain the Swi2/Snf2 motor core to drive the translocation movement (Saha, et al., 2006; Liu, et al., 2011; Beyer, et al., 2013). Chromatin remodelers have the ability to move histones along the DNA, thereby changing the chromatin landscape and dictating the accessibility of parts of the DNA for protein binding such as transcription factors (Längst & Manelyte, 2015). While some chromatin remodelers move histones away from the gene to be transcribed, others “reorganize” the spacing of the histones and promote transcriptional repression (Längst & Manelyte, 2015). Since chromatin remodelling has a direct impact on transcriptional activity, it should also affect R-loop formation.

R-loops themselves could affect the recruitment of chromatin remodelers. Some of the chromatin remodelers binding targets are formed as a consequence of R-loop formations. For example, acetylated H3 and G-quadruplexes are recognized by the chromatin remodeler of the ISWI family, which are histone “reorganizers” repressing transcription (Längst &

Manelyte, 2015). This could potentially create a self-regulatory system in which transcriptional activity is moderated.

1.5.3.2. Helicase rewinding ability

Helicases are extensively researched and used for their unwinding ability. However, emerging evidence suggests that helicases can have rewinding capabilities (Wu, 2012). Examples of RNA helicases with rewinding ability include human DDX42p, *S. cerevisiae* DED1 and Mss16p, and Dengue virus NS3. Reported DNA helicases are human RECQ5 β , BLM, WRN, RECQ1, DNA2, PIF1, HARP, AH2 (Wu, 2012) and UvsW (Nelson & Benkovic, 2006). DNA:RNA annealing helicases that have been reported are DHX9 and Cas3 (Howard, et al., 2011; Chakraborty, et al., 2018). The purpose of this rewinding activity is currently unknown, but Wu (2012) suggests multiple functions. The helicases could help in stabilizing stalled replication fork by reannealing long stretches of parental ssDNA until replication restarts. In homologous recombination, the helicases (particularly BLM and WRN) could coordinate both the unwinding of the intact dsDNA for the invasion of the ssDNA and the annealing of the invading ssDNA to the complementary template (Wu, 2012). Along with the same reasoning, some of the functions of R-loops described in section 1.4 could profit from DNA:RNA rewinding capabilities, such as DNA repair, control of transcription and histone modification/chromatin compaction. Information regarding helicase rewinding is scarce, and more research is required to begin understanding the impact of rewinding nucleic acids.

1.5.3.3. Steric hindrance and export

DDX19 is able to unwind R-loops *in vitro*, but its effects on reducing R-loop accumulation could also be related to its function as an mRNA export protein that binds to the nascent RNA. Although export defective DDX19 protein has been shown to resolve R-loop accumulation as effective as wild-type (Hodroj, Serhal & Maiorano, 2017), it cannot be excluded that the binding of DDX19 (or any other protein-related to RNP biogenesis) to the nascent RNA itself is enough of a hindrance to prevent reannealing of the nascent RNA to the ssDNA. It has indeed been shown that pre-mRNA processing proteins that are co-transcriptionally loaded to the nascent RNA can reduce R-loop formation (Skourti-Stathaki & Proudfoot, 2014). The THO/TREX export contains the DEAD-box RNA helicase Sub2, which is required for mRNA splicing and export. Depletion of the THO subunit results in transcription elongation impairment linked with increased R-loop formation (Dominguez-Sanchez, et al., 2011). Overexpression of other RNA binding factors, including the Sub2

subunit, suppresses the hyperrecombination of THO depleted mutants, supporting the idea that covering the nascent RNA with RNPs prevents R-loop formation (Garcia-Benitez, et al., 2017).

1.5.3.4. Recruiting other proteins

Helicases could act as binding targets for other proteins after binding to R-loops and recruit proteins that then function in removing R-loops. The eukaryotic initiation factor 4A-III (eIF4A-III) DEAD-box helicase, for example, acts as a stationary ATP-dependent clamp around RNA on which other proteins assemble (Jankowsky & Fairman-Williams, 2010). This could affect R-loop levels in a manner similar to the non-helicase example, FANCI-FANCD2 (ID2). This DNA binding protein binds R-loops via the recognition of ssRNA or G-rich ssDNA, but not the DNA:RNA hybrid itself, which activates the Fanconi Anemia pathway (including the FANCM helicase) that then resolves R-loops (Liang, et al., 2019).

1.6. Aim and objectives

Transcriptional activity depends on epigenetic regulation and can be correlated with the formation of R-loops and the deposition of the histone mark H3K4me3. What makes these two in particular interesting is that they show high aggregation at the promoter location of the gene. Whether this correlation has a causal effect is unknown and is the central question that initiated this study. I hypothesize that:

Reducing the global level of H3K4me3 has an impact on R-loop aggregation.

To test this hypothesis, I use the model organism *C. elegans*, which only has one COMPASS complex responsible for the majority of H3K4me3 marks. The *C. elegans* loss-of-function mutants *cfp-1(tm6369)* and *set-2(bn129)* produce viable progenies and have a drastic reduction in global H3K4me3 levels (Pokhrel, et al., 2019). First, I analyse the changes in H3K4me3 levels in both mutants. Then, by comparing the R-loop levels in the COMPASS mutants with wild-type controls, changes in R-loop aggregation as a result of H3K4me3 depletion can be identified. The results of this investigation could enhance our understanding of epigenetic regulation and potentially uncover the purpose of the H3K4me3 active transcription marker.

In addition, this project aims to identify helicases that can resolve R-loops, independent or dependent on the H3K4me3 marker by performing RNA interference (RNAi) screening on the *C. elegans* COMPASS mutants. Results from this screen provide a comprehensive list of DNA:RNA helicase candidates that play a role in regulating R-loop aggregation.

Chapter 2: Methods for the epigenetic analysis

2.1. Basic maintenance

2.1.1. List of strains

The following strains were used in this thesis:

Strain Name	Genotype	Nature of Mutation
N2 Bristol	wild-type	wild-type
	<i>set-2(bn129)</i>	748 bp deletion; frameshift and nonsense mutation (Xiao, et al., 2011)
	<i>cfp-1(tm6369)</i>	254 bp deletion (Pokhrel, et al., 2019); frameshift mutation
RB835	<i>rcq-5(ok660)</i>	1299 bp deletion (The <i>C. elegans</i> Deletion Mutant Consortium, 2012); frameshift and nonsense mutation
KMW1	<i>rha-1(tm329)</i>	1059 bp deletion; nonsense mutation and genetic null mutation (Walstrom, et al., 2005)

Table 2.1 List of strains used in the epigenetic study. All genotypes are listed with their mutant name and strain name when available. As of writing, *set-2(bn129)* and *cfp-1(tm6369)* did not have a designated strain name.

2.1.2. Nematode Growth Medium (NGM) plate preparation

For 1 L of NGM, 17 g of Agar (ash 2.0-4.5%) (Sigma-Aldrich®), 2.5 g of peptone (from meat) (Sigma-Aldrich®) and 3 g NaCl (Acros Organics®) were mixed in a 1 L bottle and made up to 1 L with deionized water and left for autoclaving.

Before usage, if the mixture has solidified, it was heated up in a microwave until completely liquified and let to cool to ~60°C. Afterwards the following were added: 1 mL 1 M CaCl₂ (Sigma-Aldrich®), 25 mL 1 M KPO₄ (Honeywell Fluka™), 1 mL 1 M MgSO₄ (Sigma-Aldrich®) and 1 mL of 5 mg/mL cholesterol in ethanol (Sigma-Aldrich®). KPO₄ was prepared by mixing 108.3 g of KH₂PO₄ with 35.6 g of K₂HPO₄ in 1 L water. The mixture was then poured into the Petri dishes using sterile serological pipettes under a laminar flow hood. For small plates (diameter: 55 mm), 10 mL of NGM was used. For large plates (diameter: 135 mm) 75 mL of NGM was used.

After the NGM has solidified, the plates were turned upside down and dried for one (small plates) or two weeks (large plates) at room temperature. OP50 bacteria liquid culture was then spotted on the plates (150 µL for small plates and 2 mL for large plates) and air-dried at room temperature. OP50 was allowed to grow for three to seven days to produce a large lawn of bacteria.

2.1.3. OP50 maintenance

OP50 was maintained in Luria-Broth (LB) and kept at 4°C. 1 L of LB was made by mixing 10 g of Tryptone (vegetable) (Sigma-Aldrich®), 5 g yeast extract (Fluka Analytica®) and 5 g of NaCl (Acros Organics®) and made up to 1 L with deionized water. The pH was adjusted to 7 with 1M NaOH (AnalaR NORMAPUR®) and autoclaved.

2.1.4. Worm maintenance

Worm stocks were maintained on NGM (Nematode Growth Medium) Petri plates with seeded OP50 (*Escherichia coli*) and kept at 15°C. For experiments, worms were shifted to 20°C (standard growth condition) (Stiernagle, 2006) and maintained for four generations before use (except *rha-1*, which has reduced brood size at 20°C and thus experiments were done at 15°C). 5-10 Worms were transferred to new NGM plates before the old NGM plate runs out of OP50 food.

Male stock worms were maintained at 15°C. Hermaphrodites were paired with males in a 1:3 ratio to improve the chance of sexual reproduction and male progenies.

2.1.5. M9 buffer preparation

1 L of M9 buffer was prepared by mixing 3 g of KH_2PO_4 (Honeywell Fluka™) with 6 g of Na_2HPO_4 (Acros Organics®) and 5 g of NaCl (Acros Organics®) and made up to 1 L with deionised water. The mixture was then autoclaved. Before usage, 1 mL of 1 M MgSO_4 (Sigma-Aldrich®) was added to the mixture.

2.1.6. Liquid culture bleach preparation

50 mL of liquid culture bleach was prepared by mixing 5 mL of 10 M NaOH (AnalaR NORMAPUR®) with 15 mL 4% NaClO and 30 mL water.

2.1.7. Worm bleaching and synchronization

Bleaching worms from NGM plates were done by washing off all the worms by adding 1 mL of water to the 55 mm plates (5 mL of water for 135 mm plates). The water with suspended worms was transferred to a centrifugal tube. An equal amount of liquid culture

bleach was added to the suspension and placed on a shaker or vortex at 600 rpm for 4-6 min. The mixture was then centrifuged at 2000 rpm for 2 min, and the liquid was decanted to 0.1-0.5 mL (depending on the number of embryos). Water was added to fill up the centrifugal tube to the limit, shaken for 10 seconds, centrifuged at 2000 rpm for 2 min and decanted. This cycle was repeated once more and then again with M9 buffer instead of water. After the last decanting step, the embryos were transferred to a plate directly or left to hatch inside the centrifugal tube to become synchronized larvae stage 1 (L1) animal. For the latter option, M9 buffer was added to fill up 2/3 of the centrifugal tube and left on the shaker overnight at 20°C.

2.1.8. Tris-buffered Saline (TBS) and TBS + Tween (TBST) preparation

For 1 L of 10x TBS, 24 g of Tris base (AppliChem Panreac) was mixed with 88 g of NaCl (Acros Organics®) and dissolved in 900 mL deionised water. The pH was adjusted to 7.6 with 12 M HCl, and deionized water was added to make up 1 L.

For 1 L of 1x TBST, 100 mL 10x TBS was mixed with 900 mL distilled water and 1 mL Tween 20.

2.2. Genotyping and outcrossing

2.2.1. List of primers

Genotype	Primer	Wild-type	Mutant
<i>cfp-1(tm6369)</i>	F: 5'-ACA CGG GGC AGT TTG TGC GA-3' R: 5'-AGG AGT GCA CGA GCC ACG TA-3'	1.1 kb	846 bp
<i>set-2(bn129)</i>	F: 5'-TGGAAGAGTTAGTGGAGAATTTGG-3' R: 5'-TGTGCGAAAAATTGCAGTGC-3'	1.3 Kb	572 bp
<i>rcq-5(ok660)</i>	F: 5'-TTTCAGCTTTCTCCCCCTCT-3' R: 5'-TGAAAACCCTAATTGCCAGA-3'	1782 bp	483 bp
<i>rha-1(tm329)</i>	F: 5'-TAATCCGTTCTCCATCATTCG-3' R: 5'-GATTTGGCTACTGCTTTTCG-3'	1565 bp	506 bp

Table 2.2 Table displaying the sequence of primers used for genotyping the mutant *C. elegans* strains. The expected size of PCR products in wild-type and mutants using the respective primers are shown on the right side of the table.

2.2.2. Genotyping and single worm PCR

The genotype of samples was confirmed using PCR. A single hermaphrodite (mother) or multiple offsprings were transferred to the PCR tubes containing the Proteinase K (PK) solution with as little bacteria as possible. The PK solution was made up of 7.5 μL nuclease-free water, 0.5 μL of 20 mg/mL Proteinase K (ThermoFisher Scientific™) and 2 μL 5x HF Phusion PCR buffer (ThermoFisher Scientific™). The PCR tube was kept in a thermocycler (PCR machine) for 75 minutes at 55°C, then 20 minutes at 98°C and were kept at 4°C.

After the worm was completely digested in the PK solution, the PCR master mix is prepared by mixing various chemicals, as shown below:

The number of reaction:	1x	8.5x
5x HF Phusion PCR buffer	2.5 μL	21.3 μL
10 mM dNTP	0.25 μL	2.1 μL
10 μM Forward primer	0.625 μL	5.3 μL
10 μM Reverse primer	0.625 μL	5.3 μL
Phusion Polymerase (2 U/μL)	0.125 μL	1.06 μL
Water	8.38 μL	71.23 μL
Total	12.5 μL	106.25 μL

Table 2.3 Composition of the PCR master mix for either one reaction or eight (8.5) reactions.

For single worm PCR, 2 μL of the PK digest was mixed with 12.5 μL of the PCR master mix (For genotyping from purified DNA, 0.1 μL of the template was sufficient). The mixture was then kept in the thermocycler (PCR machine), with a PCR program depending on the primers used.

	Temperature	<i>cfp-1/set-2</i>	<i>rcq-5/rha-1</i>
Initial denaturation	98.0°C	5 min	5 min
Denaturation	98.0°C	10 sec	10 sec
Annealing	58.0°C	20 sec	15 sec
Extension	72.0°C	1:30 min	2:15 min
Final extension	72.0°C	20 min	5 min
Storing	4.0°C	∞	∞

} 35 cycle

Table 2.4 PCR program of the thermocycler for different gene/strain.

Gel electrophoresis on a 1.2% agarose gel was done for PCR samples. The gel was prepared by mixing 1.2 g of agarose (SERVA) with 100 mL of 0.5x TBE (Tris Borate EDTA) and heated up until homogenous. Next, 5 μL of 10 mg/mL Ethidium Bromide was added to the solution and mixed. The gel was then poured into a casting tray to settle. The PCR samples were prepared by adding 2 μL of 6x orange loading dye to them. The agarose gel was placed in the electrophoretic tank with 0.5x TBE. 10-15 μL of the sample was loaded onto the wells of the gel. 2 μL of DNA hyperladder was used (100 bp or 1 kbp) as a molecular marker.

Electrophoresis was run at 100 V for 30 min or 50 V for 1 hour. The Gel was then visualized under a gel documentation system (FAS-Digi PRO).

2.2.3. Outcrossing

C. elegans strains were outcrossed at least four times with the N2 males before being used in experiments to ensure that enough recombination event has occurred to limit the mutation sites to the area around the gene of interest (Zuryn & Jarriault, 2013). Worms were kept under standard experimental condition during the outcrossing process. Two to three L3-L4 hermaphrodites of the strain to be outcrossed were placed on a seeded NGM plate with male N2 worms in a 1:3 ratio. After two to three days, 6 hermaphrodite offsprings were transferred to new seeded NGM plate each. After the 6 worms have reached the adult stage and have laid a few eggs, the adult worm was genotyped using the single worm PCR protocol. One plate where the adult worm was confirmed to be heterozygous for the gene of interest is chosen for the next process. 8 offspring from the chosen plate were separated onto a new seeded NGM plate each. After those 8 offspring have started laying eggs, each of them was genotyped to find a homozygous mutant for the gene of interest. The plate from which a homozygous mutant was identified was kept and grown. For re-confirmation of the homozygosity, 4 worms from the homozygous plate were genotyped. This concluded one round of outcrossing and the process was repeated at least 3 more times.

2.3. RNA interference

2.3.1. RNAi LB culture plate preparation

LB agar plates were used to grow the RNAi bacteria from frozen stock. 1 L of LB agar was made by mixing 10 g of Tryptone (vegetable) (Sigma-Aldrich®), 5 g yeast extract (Fluka Analytica®), 5 g of NaCl (Acros Organics®) and 15 g of agar (ash 2.0-4.5%) (Sigma-Aldrich®) together and made up with deionized water to 1 L. The pH was adjusted to 7.5 with 10 M NaOH (AnalaR NORMAPUR®), followed by autoclaving. The solidified LB agar was heated up in a microwave until it was completely liquified and left to cool to ~60°C. 1 mL of 10 mg/mL Tetramycin and 1 mL of 50 mg/mL Ampicillin was added to the 1 L liquified LB Agar. Plates were poured the same way as NGM plates are made (see **Section 2.1.2**).

2.3.2. Liquid RNAi LB preparation

RNAi bacteria liquid culture was prepared by mixing 1 L of LB with 1 mL of 50 mg/mL Ampicillin. For overnight inoculation, another 1 mL of 1 mg/mL Tetramycin was added.

2.3.3. RNAi bacteria feeding plates preparation

RNAi bacteria of interest were streaked from the glycerol stock (from Julie Ahringer library) to RNAi LB culture plates under sterile condition. The streaked plates were left to incubate overnight at 37°C. A small sample from the grown bacteria was then transferred to liquid RNAi LB and incubated for 6-8 hours (or overnight) at 37°C in a shaking incubator. The liquid bacteria culture was then seeded onto dried NGM plates containing 1 mM Isopropyl β -d-1-thiogalactopyranoside (IPTG) and 50 μ g/mL ampicillin (the same method as OP50, see **Section 2.1.2**). The plates were then left to dry and grow for 3-7 days.

2.4. Developmental assay

Determining the time at which worms reach L4, young adult or adult stage was done using the developmental assay. Worms were bleached and synchronized to L1. 100-200 L1 worms were transferred to an NGM plate seeded with OP50. The worms were left to grow at 20°C or 25°C. Worms were observed every day starting on the 3rd day until they reach the adult stage and the development stage at each time point of observation was noted down.

2.5. RNAi sensitivity assay

RNAi plates were set-up as described above. RNAi bacteria used were: *dpy-10*, *unc-15*, *bmr-1*, *dpy-8*, *lin-1*, *dpy-13* and *unc-73*.

For each strain (wild-type(N2), *set-2(bn129)* and *cfp-1(tm6369)*), three L3-L4 worms were spotted on the NGM plate containing the RNAi bacteria of interest. After 24-48 hours (depending on the growth rate), the three worms were transferred to a new plate to feed for another 24 hours before being transferred once more to a new plate. The number of eggs/progeny worms were counted and added together to find the total brood size of the three worms. The plates were kept until the progenies of the three worms have grown to

L4/adult stage in order to score for the relevant RNAi phenotype (see the relevant section for each phenotype). Biological replicates were done in parallel.

2.5.1. Dumpy phenotype

For scoring “dumpy” phenotype (*dpy-10*, *dpy-8* and *dpy-13*), the shape of the worm was compared to control worms (worms treated with Empty vector (HT115) *E. coli* bacteria). The number of worms with the dumpy phenotype (shorter and fatter than the control worms) were counted in each plate and added together. The total number of dumpy worms were then compared to the total brood size to find the percentage of dumpy worms. A few dumpy worms were picked, and their body length was compared to the control worm. Both criteria were used to determine RNAi strength.

2.5.2. Uncoordinated phenotype

The “uncoordinated” phenotype (*unc-15* and *unc-73*) was scored based on the worms body paralysis. In this assay, a worm is considered to be paralyzed when it cannot move its body (but may still have some limited head movement) even when gently touching them or tapping the NGM plate on the table. The pharynx activity was not considered.

The number of worms that were paralyzed were counted and added together and compared to the total brood size to determine the strength of the RNAi phenotype.

2.5.3. *lin-1* phenotype

Worms treated with *lin-1* RNAi display multi-vulva phenotype (worms with one or more than one vulva-like protrusion). The number of worms with multiple vulvae were counted and compared to the total brood size to determine the strength of RNAi.

2.5.4. *hmr-1* phenotype

hmr-1 RNAi phenotype was scored based on the number of dead embryos. The number of dead embryos was counted and compared to the total brood size.

2.6. Sample collection for staging and genomic DNA (gDNA) extraction

2.6.1. Young adult collection

Synchronized worms were left to grow into very early young adults (around 60 hours for wild-type and 65 hours for COMPASS mutants). Worm collection was done by washing the worms off the plates using water and transferring to a centrifuge tube. The collection was centrifuged at 2000 rpm for 2 min, followed by decanting the liquid and filling the tube with M9 buffer. This was repeated at least one more time until the mixture was clear (without the cloudy bacteria) and left for 5 min at room temperature. The worms were centrifuged one more time and M9 buffer was removed as much as possible. A small sample of the collection was preserved in 100% methanol at -20°C for scoring the developmental stage while the rest was frozen in dry ice and stored in -80°C.

2.6.2. Late embryo collection

The worms were left to grow into adults that just started laying embryos and carry large amounts of embryos. The time between transferring the synchronized L1 worms to NGM plates and collecting them was: wild-type(N2) = 60h, *cfp-1(tm6369)* = 68h and *set-2(bn129)* = 68h. After collection, the worms were bleached, and the embryos were left to develop in M9 buffer at 20°C for 5.5 h and 6 h for wild-type(N2) and mutants, respectively. Before proceeding to the sucrose floating step, a small sample is scored to determine whether the embryos were at the desired developmental stage (2-fold and 3-fold stage). The embryos were centrifuged, and M9 buffer was decanted to 3 mL. The mixture was transferred to 15mL centrifugal tube if applicable. An equal amount of 60% sucrose was added, mixed and centrifuged at 1000 rpm for 5 minutes. 1 mL of M9 buffer + 0.1% triton was carefully added to the surface creating a partition of immiscible liquid. Live embryos will float at the boundary of the 2 liquids. The embryos were collected by carefully draining the M9 buffer + 0.1% triton at the partition boundary. The embryos were then washed twice in M9 buffer +0.1% triton and once more in M9 buffer only before being frozen in dry ice and stored in -80°C.

2.6.3. Staging

Young adult samples preserved in methanol were staged by 4',6-diamidino-2-phenylindole (DAPI) staining. Methanol was removed from the sample as much as possible by centrifuging and decanting. The samples were then washed twice in M9 buffer +1% triton. 50 μ L of 1 μ g/mL DAPI was added to the sample and left on the shaker for 5 minutes and then visualized under a Nomarski Microscope. The worms are categorized in either as young adults by the presence of embryos and vulva or as L4 by the absence of these features.

Late embryos were staged without any staining under the Nomarski Microscope. They were classified as either younger than comma stage, comma stage, 2-fold, 3-fold or hatched (see **Figure 2.1**).

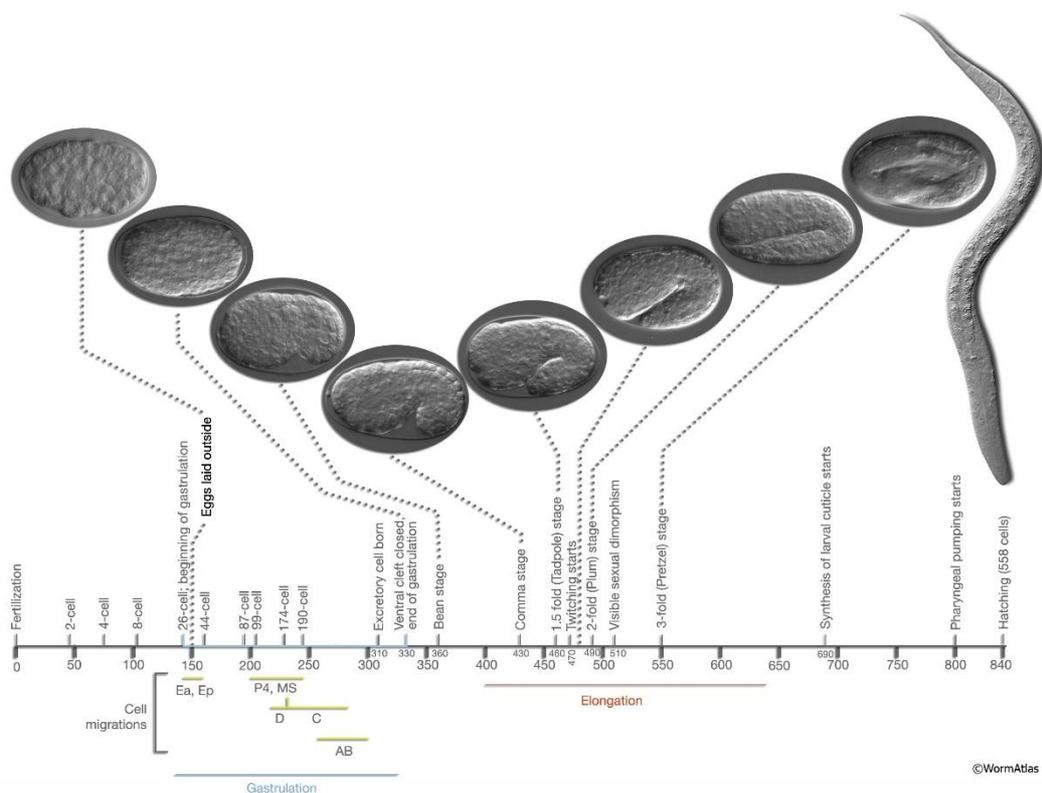


Figure 2.1 The developmental stages of the embryo from fertilization to hatching. Image taken from Altun & Hall (2009).

2.6.4. L1 Worm Collection

The worms were left to grow into adults that just started laying embryos and carry large amounts of embryos. The time between transferring the synchronized L1 worms to NGM plates and collecting them was: wild-type(N2) = 60h, *cfp-1(tm6369)* = 68h and *set-2(bn129)* =

68h. After collection, the worms were bleached. A small sample was transferred to a small NGM plate for the hatching assay. The rest were left to grow and hatch in M9 buffer overnight at 20°C on a shaker. On the next day, the worm sample was estimated by spotting 0.1% of the mixture on an empty petri dish and counting the number of worms. The rest of the mixture was centrifuged, decanted and frozen in dry ice and stored at -80°C.

For the hatching assay, the number of eggs was counted immediately after spotting. After leaving for 24 hours, the number of hatched worms were counted to determine the fraction of hatched eggs.

2.7. DNA extraction

The DNA extraction process was done using the Invitrogen™ PureLink™ Genomic DNA mini kit. Collected worm and embryo samples were thawed on ice and centrifuged to remove as much M9 buffer as possible. Removal was repeated if necessary since M9 buffer contamination reduces gDNA extraction efficiency. 20 µL of 20 mg/mL Proteinase K and 180 µL of Digestion Buffer was added to the centrifuge tube. After short vortexing, the tubes were left in a 55°C water bath for 1 h (L1 worms), 4 h (embryos) or 5 h (adult worms) with occasional shaking. The lysate was centrifuged at 17,000 g, and the supernatant was transferred to an Eppendorf tube. The Eppendorf tube was then centrifuged at 17,000 g for 3 min to remove any more residues, and the supernatant containing the DNA was transferred to a new Eppendorf tube. Centrifugation was repeated if necessary to remove further residues (only applicable to adult worm gDNA extraction). 20 µL of 20 mg/mL RNase A was added to the supernatant and left at room temperature for 2-3 minutes. An equal volume to the supernatant (~200-300 µL) of Lysis/Binding Buffer was added and vortexed until the mixture became homogenous. The same volume of 100% ethanol was added and vortexed until homogenous.

The mixture was then added to the spin column in a collection tube (max capacity is 640 µL) and centrifuged at 10,000 g for 1 minute at room temperature. The eluate inside the collection tube was discarded. The collection tube was reused if there was more DNA mixture to be eluted; otherwise, a new collection tube was used. 500 µL of Wash Buffer 1 was added to the spin column and centrifuged at 10,000 g for 1 minute, and the eluate and collection tube was discarded. This was repeated for Wash Buffer 2 but centrifuged at 17,000g for 3 minutes.

The spin column was then placed into an Eppendorf tube to collect the DNA. 50 μL of nuclease-free water was added to the spin column and left at room temperature for 5 minutes. The tube was then centrifuged at 17,000 g for 2 minutes.

The DNA concentration was measured on a Nanodrop by blanking with nuclease-free water and adding 1 μL of the elution onto the sensor or by using Qubit 4 (Invitrogen™) with the dsDNA BR assay kit.

2.8. R-loop slot blot

2.8.1. Membrane blotting

gDNA samples were prepared by diluting the stock gDNA to a standard concentration (typically 100 ng/ μL). Three samples of varying amount of DNA were prepared (see individual blots) by adding the respective amount of DNA to 1x RNase H buffer for a total volume of 100 μL . An additional RNase H negative control was prepared by adding 2 μL of 5 U/ μL RNase H to an identical sample with the highest amount of gDNA. All samples were left at 37°C for 20 minutes. This allows RNase H to degrade R-loops while exposing all other samples to the same condition.

Amersham Hybond-N+ Nylon membrane and filter paper were cut to the required size. The filter paper must cover all wells of the machine, while the nylon membrane only needs to accommodate all samples. Both the membrane and the filter paper were pre-wetted in autoclaved water. The slot blot machine was then set up as shown in **Figure 2.2**:

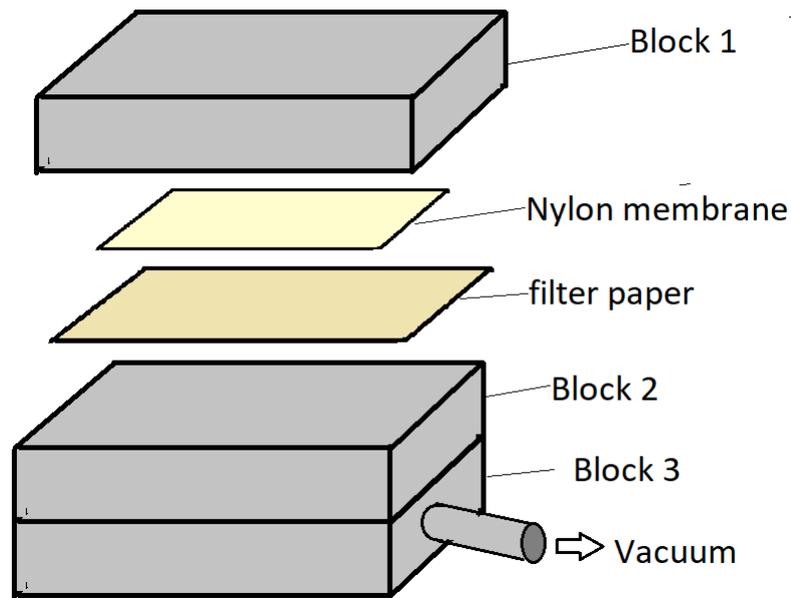


Figure 2.2 Set-up of the slot blot machine. The filter paper needs to be cut to cover all the wells in the slot blot machine. The nylon membrane only needs to be cut to accommodate all samples. Both the membrane and filter paper were pre-wetted in autoclaved water.

The vacuum is turned on and left running for 1 minute to produce a negative pressure before continuing.

For each well that was being used, 100 μL of autoclaved water was added to rehydrate the membrane. Then the DNA samples were added, and afterwards another 100 μL of autoclaved water was added to wash down any DNA residues that might be stuck at the walls. Before each addition, the previous liquid must be completely pulled through. The apparatus was dismantled, and the nylon membrane was air-dried for 5 minutes and then placed between dry filter paper and wrapped with aluminium foil before baking it at 80°C for 2 hours.

The nylon membrane was then blocked in blocking solution (2 g milk powder in 40 mL 1x TBST) at room temperature for 1 hour. The membrane was then transferred to a 50 mL centrifugal tube. 1 μL of S9.6 antibody (1 mg/mL) (EMD Millipore) in 5 mL of blocking solution was transferred to the centrifugal tube and left on a roller at room temperature for 2-4 hours. The antibody/blocking solution mixture was then replaced by 1x TBST and left on the roller for 5 minutes to wash. Washing was repeated 2 times for a total of 3 washes. 1 μL of anti-mouse IgG antibody was added to 5 mL of blocking solution and replaced the 1x TBST. The centrifuge tube was left on the roller for 2 hours. The membrane was then washed once more in 1x TBST for 10 minutes followed by developing.

2.8.2. Membrane development

The nylon membrane signal was visualized either using the traditional X-ray film development or the G:BOX (Syngene) imaging system. The nylon membrane was placed onto a clear plastic sheet. 1.5 mL of ECL solution A and B (SuperSignal™ West Pico PLUS) were mixed and poured onto the membrane and covered by another clear plastic sheet before being placed in an X-ray film cassette. In a dark room, an X-ray film was placed on top of the clear plastic sheet with the membrane (and spotted DNA) facing upward. The length at which the film needed to be kept on top depends on the strength of the signal and can require multiple trials. Finally, the x-ray film was sent through the X-ray Film Processing machine to fix the film. Visualization using the G:BOX imaging system was performed by placing the membrane onto the target area and pouring the ECL mixture onto the membrane. The type of blotting was chosen as “chemiblot”, and the type of ECL was set as “superluminescent EZ-ECL”.

2.8.3. DNA loading control

2.8.3.1. Methylene blue Staining

The membrane was soaked in 5% acetic acid at room temperature on a shaker for 15 minutes. The acetic acid was then replaced by 0.5M sodium acetate (pH 5.2) and 0.04% methylene blue and left for 5-10 minutes. The membrane was then washed with water for 5-10 minutes or until the desired colour saturation was reached.

2.8.3.2. Anti-dsDNA reprobing

After the membrane has been developed and results captured on the X-ray film, the membrane was stripped by submerging it in mild stripping buffer (5 g glycerin, 10 g SDS (1%), 10 mL Tween 20 in 1 L of deionized water and adjusted to pH 2.2) at room temperature for 10 minutes and again for 30 minutes, followed by two 10 minutes TBS wash and two 5 minutes TBST wash. The membrane was blocked for 30 minutes with blocking solution and incubated directly with secondary antibody for 2 hours. A 10 minutes wash with TBST followed by developing the film with ECL can then indicate whether the stripping was successful or not. After a successful stripping, the membrane was directly incubated with anti-dsDNA primary antibody (1:180 dilution in blocking solution), which only targets double-stranded DNA, for 4 hours. Afterwards, the membrane was washed three times with

TBST for 5 minutes each and then incubated with the secondary antibody (anti-mouse IgG antibody in blocking solution) for 1 hour. Washing and developing was done the same as previously.

2.9. Computational analysis of H3K4me3 ChIP-seq

2.9.1. Identifying differential H3K4me3 signal in COMPASS mutants

The H3K4me3 ChIP-seq data (provided by Dr Ron Chen) for *C. elegans* mixed embryo, from wild-type(N2), *set-2(bn129)* and *cfp-1(tm6369)* mutants, were given in a bigwig file format (linear normalized to input and aligned to the ce10 reference genome with 1bp bin size). *set-2(bn129)* and *cfp-1(tm6369)* bigwig tracks were subtracted from wild-type(N2) track, to obtain tracks showing the differential H3K4me3 level. Each of the resulting track was then separated into two tracks according to the sign of the numbers. The track consisting of negative numbers denote an increase in H3K4me3 levels, while the track consisting of positive numbers denote a decrease in H3K4me3 levels. For further analysis, the negative values of the track denoting increased H3K4me3 were converted to positive values. Next, MACS2 (Zhang, et al., 2008) peak calling with the p-value cutoff of 0.0001 was done on each of the tracks to identify regions of significant H3K4me3 changes. In order to identify which genes show significant H3K4me3 changes, the region of significant H3K4me3 changes found through MACS2 peak calling were intersected with the promoter region of all protein-coding genes. The promoter region was defined as the transcript start site (taken from Ensembl BioMart (Ensembl, 2019) for WBcel215) \pm 500 bp. Any number of bp overlap was deemed sufficient to assign a peak-call region with a gene.

2.9.2. Hypergeometric testing

For comparison between two gene lists, the hypergeometric distribution test was done in R using the `phyper()`¹ function from the *stats*² package (R Core Team, 2014), with `lower.tail` set to `FALSE`. For comparison and visualization of multiple gene lists and

¹ Code text are written in the monospaced font Courier New

² Software packages for R are italicized

their overlaps, the `supertest()` function from the *SuperExactTest* package (version 1.0.7) (Wang, et al., 2015) was used.

2.9.3. Average H3K4me3 signal change around TSS (SeqPlot)

The visualization of the average H3K4me3 level change was visualized using SeqPlot (Stempor & Ahringer, 2016). The tracks used were the H3K4me3 enriched and depleted for *set-2(bn129)* and *cpf-1(tm6369)*. The features dataset was set to .bed file of all *C. elegans* (WBcel215) protein-coding TSS extracted from Ensembl BioMart (Ensembl, 2019). The type of plot was set to “point feature”, and the plotting distance was set to 1000 bp up- and downstream.

2.9.4. Gene list analysis

List of genes that see an enrichment or depletion in H3K4me3 in the COMPASS mutants were analysed using various web-based software. Protein-protein interaction network was generated using StringDB (Szklarczyk, et al., 2019) and gene set enrichment analysis was done using ShinyGO (Ge & Jung, 2018) and g:Profiler (Raimand, et al., 2007).

2.9.5. Motif discovery

The *de novo* motif discovery software DREME from the MEME suite (Bailey, 2011) and BMM motif (Siebert & Söding, 2016) were used to identify potential enriched motifs. In both cases, the input sequences were the promoter regions (500 bp upstream and downstream of the TSS) of the genes with changes (enriched or depleted) in H3K4me3 levels in the COMPASS mutants. The sequences were extracted from the ce10 reference genome using the *BEDTools* `getfasta` command (Quinlan & Hall, 2010) and providing the .bed file containing the target promoter regions. Default options were used, except for BMMmotif, where the JASPAR2018 Motif Database was chosen.

Chapter 3: Bioinformatic analysis of H3K4me3 levels in COMPASS mutants

3.1. Introduction

Histone methylation is a highly conserved PTM of histone proteins that is implicated in chromatin packaging and transcription regulation. H3K4me3 is associated with both transcriptional activation and repression in a context-dependent manner (Howe, et al., 2017; Pokhrel, et al., 2019). Research on H3K4me3 has proven to be difficult because mutations in the core subunits of the COMPASS complex are often lethal in mammals (Bledau, et al., 2014). On the other hand, *C. elegans* COMPASS loss-of-function mutants are still viable in the absence of some of its core subunits (Xiao, et al., 2011; Pokhrel, 2019), which makes it one of the few organisms suitable for studying H3K4me3 and COMPASS. *C. elegans* has two histone methyltransferases that methylate H3K4: SET-2 and SET-16. SET-2 is the main methyltransferase of the COMPASS complex, while SET-16 is the enzyme of the COMPASS-like complex that only affects a small subset of genes (Xiao, et al., 2011). In *C. elegans*, mutation in *set-2* and *cfp-1* (an important subunit) causes global reduction in H3K4me3 levels (Pokhrel, et al., 2019).

Here I used a bioinformatics approach to investigate the pattern of H3K4me3 signal change in *set-2(bn129)* and *cfp-1(tm6369)* mutants around gene promoters, to identify potential characteristics that determine if a gene is a target for COMPASS-dependent H3K4 trimethylation. The data used in this chapter was contributed by Dr Ron Chen.

3.2. *set-2(bn129)* and *cfp-1(tm6369)* mutants have reduced H3K4me3 levels

The H3K4me3 ChIP-seq data analysed here were provided by Dr Ron Chen. In total, three ChIP-seq datasets for *C. elegans* mixed embryo, one each from wild-type(N2), *set-2(bn129)* and *cfp-1(tm6369)* mutants (using ab8580 anti.histone H3K4me3 antibody (abcam)), were given in a bigwig file format (linear normalized to input and aligned to the ce10 reference genome with 1bp bin size). The files were processed by subtracting the control track (wild-type) from the treatment tracks (*set-2(bn129)* and *cfp-1(tm6369)* mutants), to generate a track of H3K4me3

increase and decrease. The file was then separated by positive and negative values to generate H3K4me3 enriched and H3K4me3 depleted files, respectively, for each mutant (**Figure 3.1**). MACS2 peak calling (Zhang, et al., 2008) was done on the depleted and enriched H3K4me3 files with a p-value cutoff of 1E-04 to identify regions of significant H3K4me3 changes. By intersecting the peaks identified with MACS2 and the gene TSS/promoter region (which was defined as 500bp upstream and downstream of the transcript start site), the genes affected significantly by dysfunctional COMPASS complex were found. The number of genes with increased and decreased H3K4me3 levels is shown below (**Figure 3.2a**).

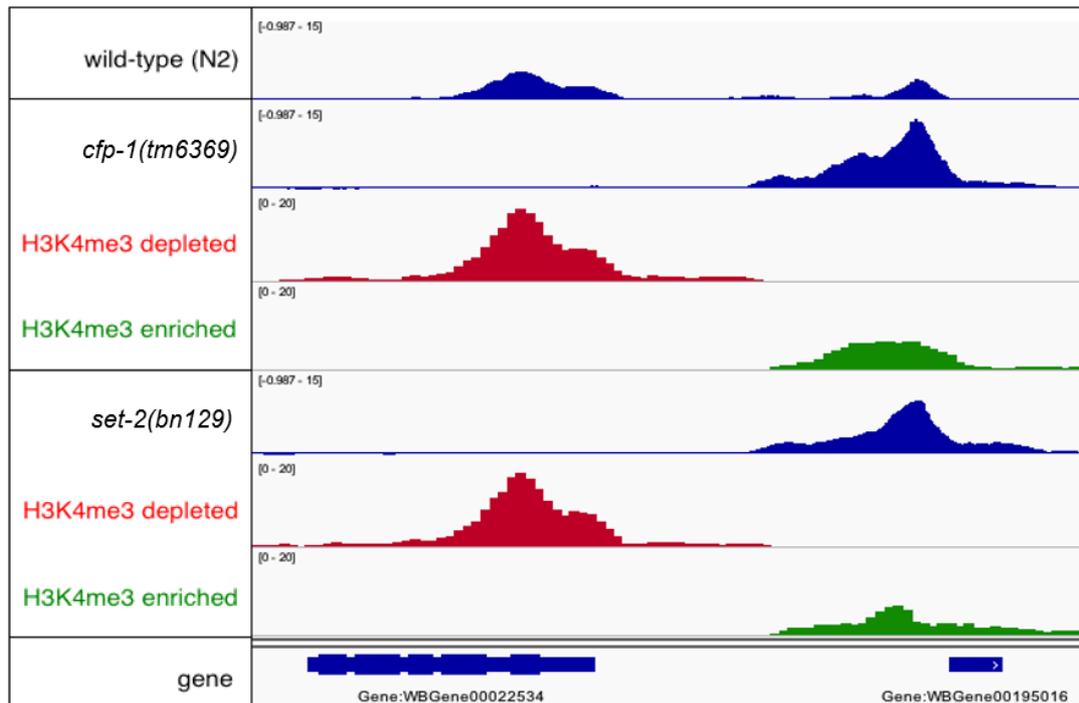


Figure 3.1 Section of COMPASS mutants H3K4me3 ChIP-seq. IGV viewer image displaying a 5kb section in chromosome III, showing the ChIP-seq track of wild-type(N2), *cfp-1(tm6369)* and *set-2(bn129)* in blue. Red tracks show the H3K4me3 depletion and green tracks represent H3K4me3 enrichment for the mutants relative to wild-type. Both red and green tracks have a bin size of 50. The y-axis scale is shown on the top left corner of each track.

Both *set-2(bn129)* and *cfp-1(tm6369)* mutants show a significant level of H3K4me3 depletion at the promoter region of nearly 1/4 of all protein-coding genes relative to wild-type. The number is nearly identical in both COMPASS mutants (4614 in *cfp-1(tm6369)* and 4612 *set-2(bn129)*) with 95.75% of the genes overlapping between the two sets (4499 genes out of 4614 and 4612 genes) (**Figure 3.2b & c** and **Figure 3.4b**). This result indicates that SET-2 and CFP-1 target the same subset of genes.

Enrichment of H3K4me3 in a small set of genes was also identified in both *set-2(bn129)* and *cfp-1(tm6369)* mutants. For *set-2(bn129)*, 374 genes show an enrichment, while in *cfp-1(tm6369)* the number is 233. Within these two subsets, a very significant number of genes (179) is shared among them (**Figure 3.2c**). To investigate how the signal of the enriched and

depleted genes changes, a plot was generated using SeqPlot (Stempor & Ahringer, 2016), that shows the average H3K4me3 signal change for each set at each position relative to the TSS (Figure 3.3).

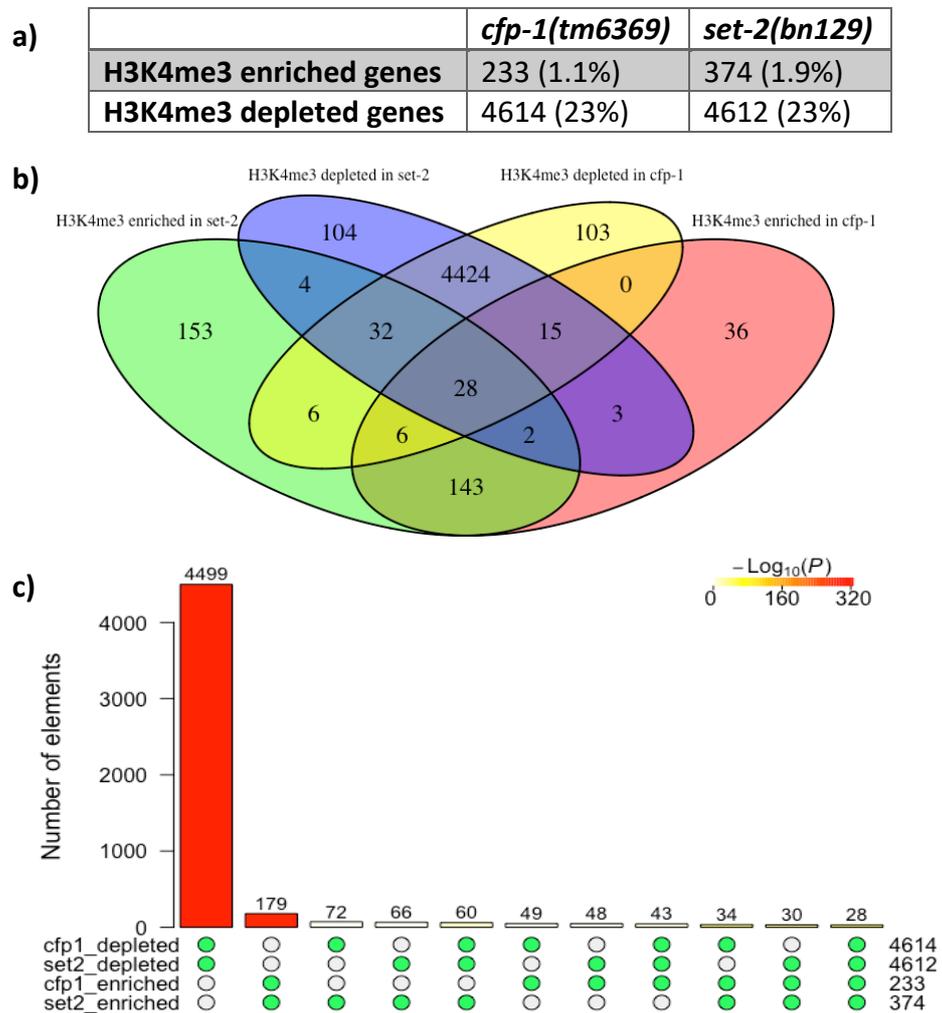


Figure 3.2 ChIP-seq results comparing the H3K4me3 enriched and depleted genes in the COMPASS mutants. a) Table showing the number of protein-coding genes with either enrichment or depletion of H3K4me3 in *set-2(bn129)* or *cfp-1(tm6369)* mutants. Numbers in brackets show the percentage of the total number of protein-coding genes. b) Venn Diagram visualizing the shared genes among each of the groups from a). c) Bar Graph showing the hypergeometric test for each combination of groups (indicated by the green circles). The numbers on the bottom right side show the number of genes in each group. The number above each column shows the number of shared genes between the compared groups. The bar fill colours indicate the p-value according to the legend on the top right. The higher the number (more red) the lower the p-value (limited to 1×10^{-320}).

The extent to which the H3K4 trimethylation modification is lost is much greater compared to how much the H3K4me3 enriched genes gain this modification (Figure 3.3). The “shape” of the lost modification correlates with the “double-peak shape” previously observed by Chen et al. (2014). This double-peak shape is not exclusive to *C. elegans* ChIP-seq and is also found in other organisms such as humans and *S. cerevisiae*. The smaller upstream peak is thought to be due to closely spaced divergent transcription of both coding and non-coding DNA. A nucleosome-depleted region separates the two peaks (Howe, et al., 2017). Similarly,

other histone modifications of active transcription such as H3K9ac (Mosesson, et al., 2014) and H3K27ac (Han, et al., 2019) also show this shape. The H3K4me3 enrichment is relatively low and has a relatively large confidence interval. Due to the large number of genes enriched with H3K4me3 overlapping between *set-2(bn129)* and *cfp-1(tm6369)* mutants (**Figure 3.2c**), it is unlikely that the hits are random (but could be due to systematic effects).

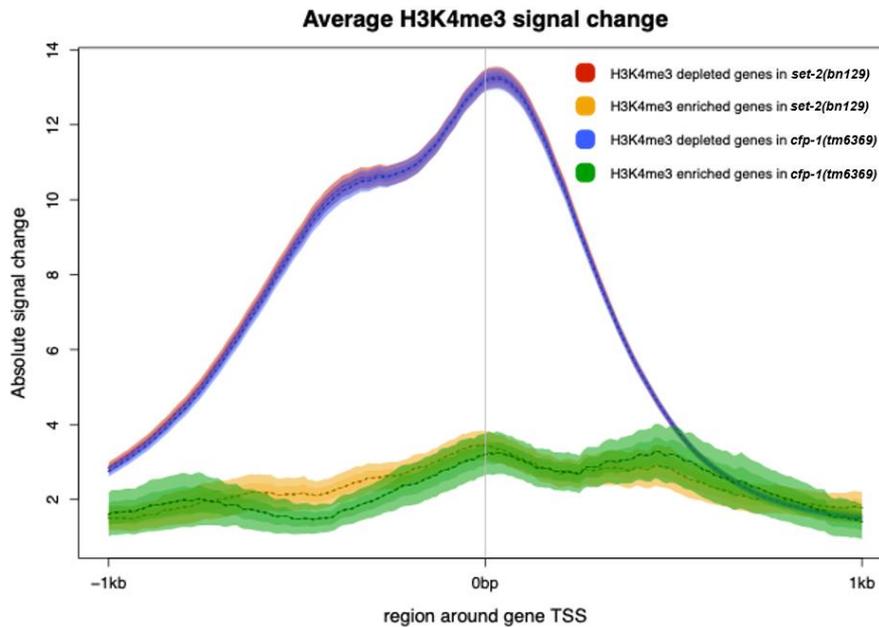


Figure 3.3 Average change in H3K4me3 around the TSS in the COMPASS mutants. SeqPlot diagram showing a 2kb window around the TSS of the respective gene sets. The red and blue lines show how much H3K4me3 signal is lost around the TSS region, while the yellow and green lines show how much H3K4me3 signal is gained in the *cfp-1(tm6369)* and *set-2(bn129)* mutants respectively compared to wild-type(N2). The dashed line denotes the mean, the dark area is the standard error, and the light area indicates a 95% confidence interval.

3.3. Loss of H3K4me3 signal is associated with housekeeping genes

With the Loss-of-function of SET-2 and CFP-1, a global loss of H3K4me3 is expected as these two genes are key subunits of the COMPASS complex. This is reflected in the ChIP-seq data presented above as well as the published western blot (Pokhrel, et al., 2019). 4499 genes were found to lose the H3K4me3 modification around their TSS region in both *set-2(bn129)* and *cfp-1(tm6369)* mutants (**Figure 3.2b & c** and **Figure 3.4b**). A gene ontology (GO) term analysis on all the shared genes reveals that these are predominantly genes required for maintenance of basic cellular functions (housekeeping genes) (**Figure 3.4a**), such as the actin genes (*act-1, -2, -3, -4*), the RHO GTPase *cdc-42*, the RNA polymerase II subunit *ama-1* and the Acyl-CoA transporter *pmp-3*.

GO term analysis on the 113 and 115 protein-coding genes with significant H3K4me3 depletion found exclusively in *set-2(bn129)* and *cfp-1(tm6369)* mutants respectively (**Figure 3.4b**), found limited number enriched GO terms. For the 113 genes exclusive to *set-2(bn129)*, ShinyGO identified enrichment in the terms 'lipopolysaccharide metabolic process' and 'lipopolysaccharide biosynthetic process'. The 115 *cfp-1(tm6369)* exclusive genes only showed enrichment in the 'Sex determination' GO term identified by g:Profiler. Protein-Protein interaction analysis revealed limited interaction in both of these sets (**Figure 3.4c & d**).

3.4. Gain of H3K4me3 signal is associated with developmental and chromatin genes

A small number of genes had an increased level of H3K4me3 in the *set-2(bn129)* or *cfp-1(tm6369)* mutants compared to wild-type (**Figure 3.2**). *set-2(bn129)* had 374 genes with enriched H3K4me3 while *cfp-1(tm6369)* had 233 genes. Out of these two sets, 179 were shared between the two mutants (**Figure 3.2** and **Figure 3.5a**). GO term analysis identified that these 179 genes were enriched in chromatin activity and DNA binding processes (**Figure 3.5b**). Protein-protein interaction analysis further identified a large cluster of histone proteins that interact with each other (**Figure 3.5c**).

The 195 H3K4me3 enriched genes unique to *set-2(bn129)* mutants show mainly developmental related GO terms (**Appendix 1a**). The 54 H3K4me3 enriched genes specific to *cfp-1(tm6369)* on the other hand, only identified three significant GO terms using Wormbase Gene Ontology Enrichment Analysis. These terms are: 'development of primary sexual characteristics', 'reproductive development' and 'protein heterodimerization' (**Appendix 1b**).



b)

GO:BP		stats						
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	≤ 16	T	Q	TnQ	U
nucleosome assembly	GO:0006334	5.727×10^{-8}			80	90	11	10175
chromatin assembly	GO:0031497	6.580×10^{-8}			81	90	11	10175
chromatin assembly or disassembly	GO:0006333	9.866×10^{-8}			84	90	11	10175
nucleosome organization	GO:0034728	1.652×10^{-7}			88	90	11	10175
DNA packaging	GO:0006323	1.026×10^{-6}			104	90	11	10175
chromatin organization	GO:0006325	1.577×10^{-6}			280	90	16	10175
protein-DNA complex assembly	GO:0065004	1.781×10^{-5}			136	90	11	10175
protein-DNA complex subunit organization	GO:0071824	3.724×10^{-5}			146	90	11	10175
DNA conformation change	GO:0071103	7.865×10^{-5}			157	90	11	10175
chromosome organization	GO:0051276	2.315×10^{-3}			476	90	16	10175

c)

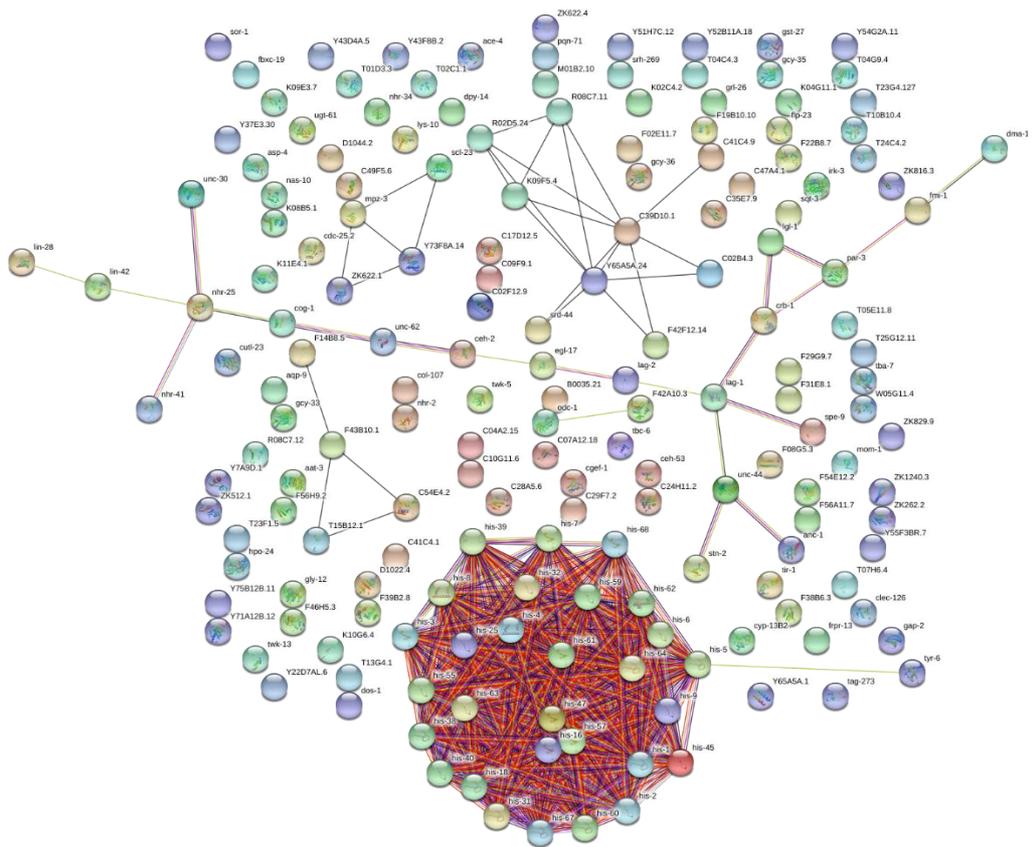


Figure 3.5 Gene Ontology analysis and protein-protein network. a) Venn diagram of genes with H3K4me3 enrichment in *set-2(bn129)* and *cfp-1(tm6369)* mutants. b) Results from the g:Profiler GO term enrichment analysis on the 179 shared genes between *set-2(bn129)* and *cfp-1(tm6369)* genes that have H3K4me3 enrichment. Only the GO:Biological Processes Terms are shown. ShinyGO did not found significant term enrichment. c) STRING protein-protein interaction network of the 179 shared genes.

3.5. Motif discovery and nucleotide frequency analysis

Following the identification that changes in H3K4me3 levels were associated with certain groups of genes, I wondered if there were specific DNA sequence motifs on which the COMPASS complex orientates itself to determine which H3K4 to methylate. The *de novo* motif discovery software DREME (Bailey, 2011) and BMM motif (Siebert & Söding, 2016) were used to identify potential enriched motifs. The results for the 4499 genes depleted with H3K4me3 in both *set-2(bn129)* and *cfp-1(tm6369)* are shown in **Figure 3.6**.

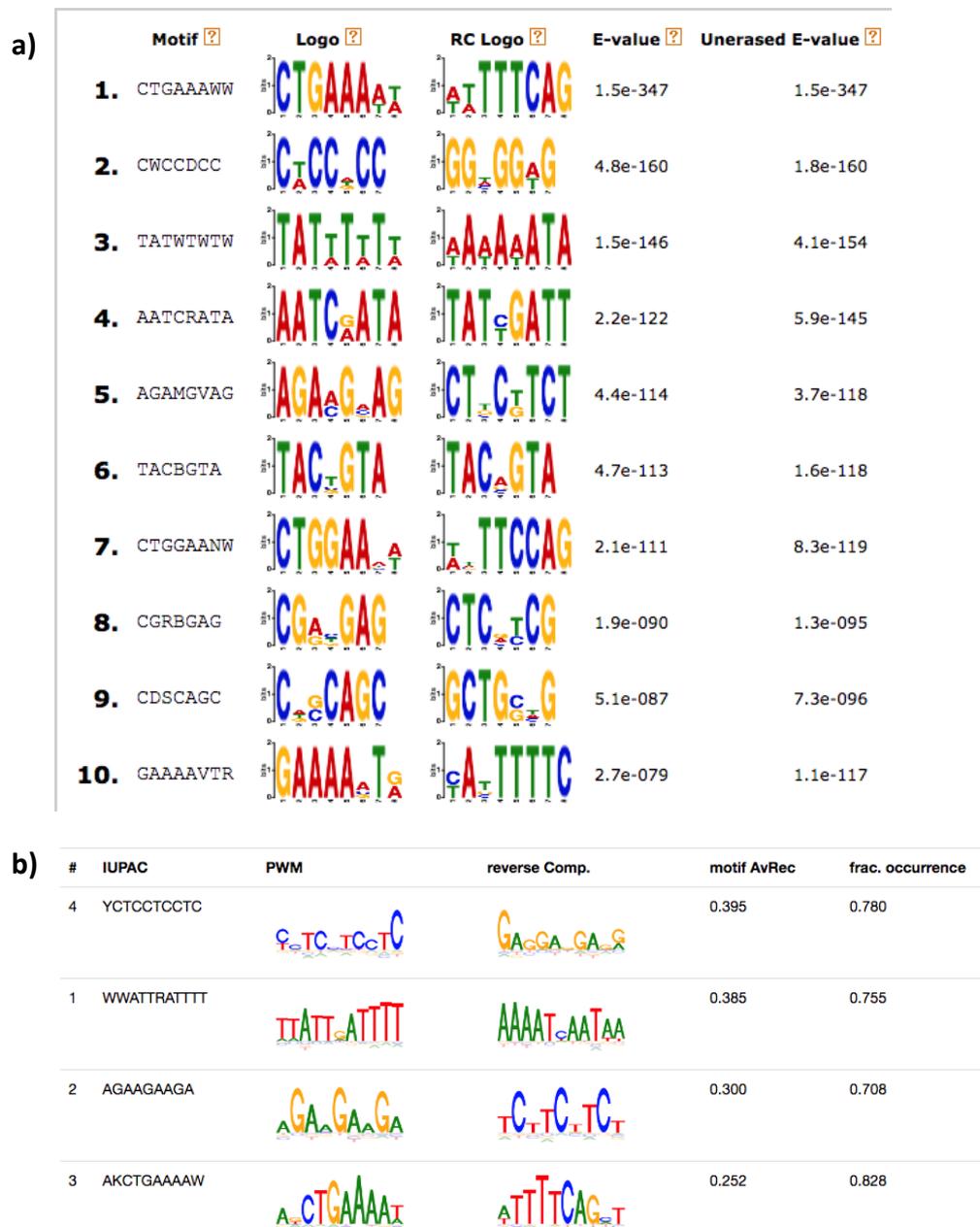


Figure 3.6 *de novo* motif discovery of the 4499 genes depleted with H3K4me3 in both COMPASS mutants. a) Results from DREME showing the top 10 motifs with the lowest E-value. b) Results from BMM motif.

The results from both software identified similar motifs. The motif with the lowest E-value and highest occurrence is CTGAAA and its reverse complement TTTCAG. The TTTCAG motif is the splicing motif for Spliced Leader 1 (SL1) of *C. elegans* that is responsible for trans-splicing of more than half of the pre-mRNAs (Blumenthal, 2012; Saito, et al., 2013). Next, I asked how much this splicing motif was enriched in those 4499 genes compared to all *C. elegans* genes and the set of genes containing all but the 4499 genes (complement set). The SeqPlot plot showed that the motif occurrence in the H3K4me3 depleted genes peaked at around 0.53 times per gene per 200bp window, while for the complement set this number was at around 0.39. For all protein-coding genes, the motif occurred 0.42 times at the TSS (**Figure 3.7a**). This means that the splicing motif is 1.36-fold enriched in the 4499 H3K4me3 depleted genes compared to its complement set.

An interesting core promoter motif called 'T-block', discovered by the Yanai lab, is characterized by three to five consecutive thymine nucleotides. The frequency of this motif is correlated with expression levels. Genes with six or more T-blocks are up to five times higher expressed than genes with three or fewer T-blocks (Grishkevich, et al., 2011). Such T-blocks were identified by both motif discovery software (last entry in both results table in **Figure 3.6**). Grishkevich, et al. (2011) further observed that T-blocks frequently occur with SL1 as a supra-motif, but are strongly depleted of the TATA-box motif (Grishkevich, et al., 2011). My data supports their observation as both SL1 and T-blocks were identified by both software, but not the TATA-box. Additionally, the link between T-blocks and active gene expression also correlates with this set of 4499 H3K4me3 depleted genes being constitutively active genes.

Apart from *de novo* motif discovery, there could be a general over or underrepresentation of specific nucleotides at particular regions of the promoter, which makes the H3K4me3 depleted genes stand out. The average nucleotide percentage at each base upstream and downstream of the TSS were plotted and compared (**Figure 3.7b-j**). The difference in the average nucleotide composition between the H3K4me3 depleted genes and its complement set is subtle. The notably different characteristic is that the 4499 H3K4me3 depleted genes have a larger difference between adenine and thymine density (stronger AT skew) just upstream of the TSS and there was a strong drop in thymine density just after the TSS (even lower than adenine). At the same time, the guanine and cytosine percentage are a little elevated compared to the complement set and all protein-coding genes (**Figure 3.7b** compared to **c & d**). The TTTCAG motif is more pronounced in the 4499 H3K4me3 depleted set (**Figure 3.7e & h**) compared to the other two sets (**Figure 3.7f, g, i & j**).

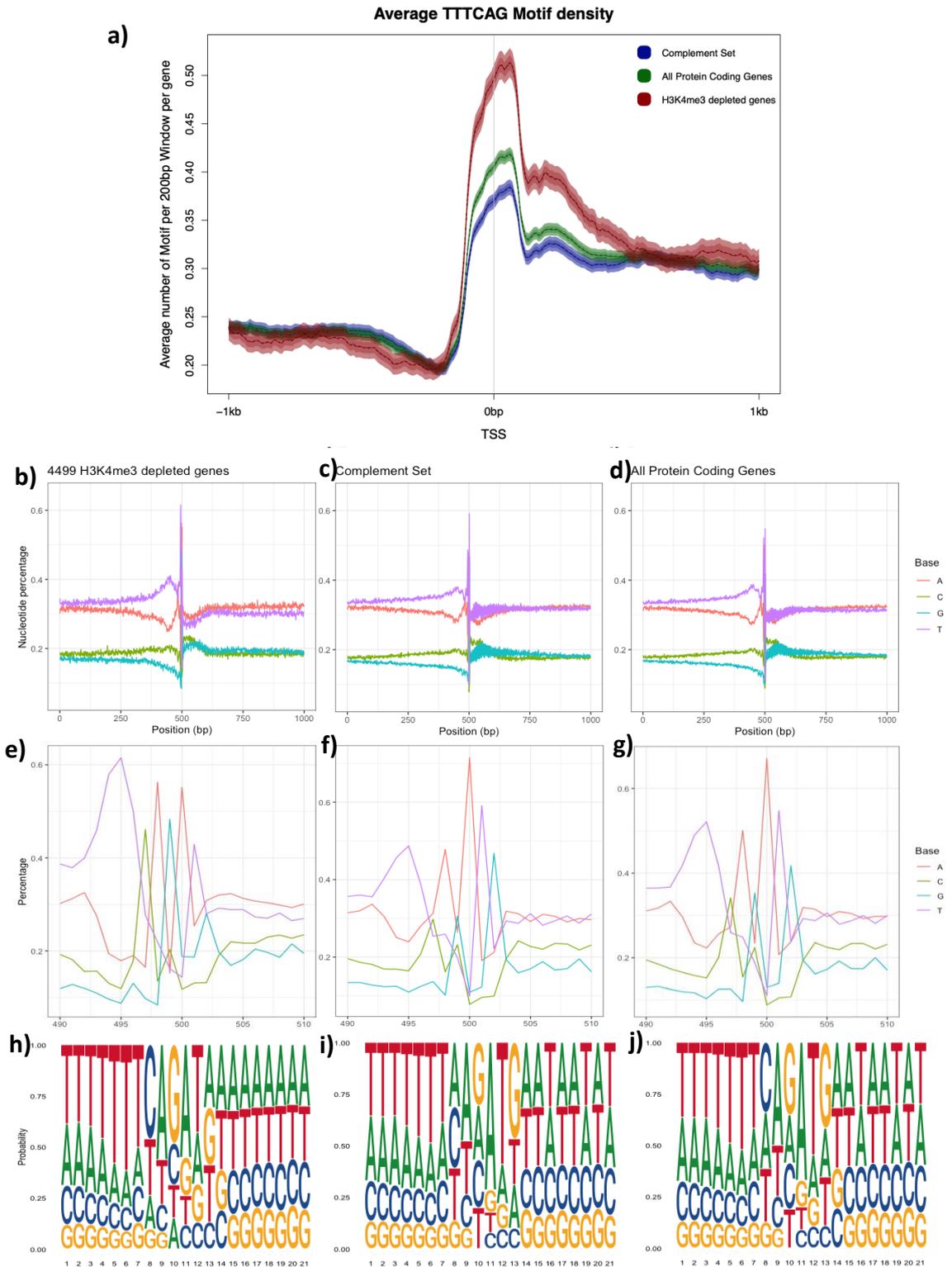


Figure 3.7 Plots of sequence motif occurrence and nucleotide frequency upstream and downstream of TSS. a) SeqPlot graph showing the average number of occurrences of the splicing motif TTTCAG in a 200bp sliding window 1.5kb upstream and downstream of the TSS. The red line represents the 4499 H3K4me3 depleted genes. The green line represents all *C. elegans* protein-coding genes and the blue line is the complement set of genes that includes all protein-coding genes, except the 4499 H3K4me3 depleted genes. b) - g) Graphs showing the nucleotide frequency at each position around the TSS. The TSS is located at the centre (position 500). b) & e) 4499 H3K4me3 depleted genes. c) & f) complement set. d) & g) All *C. elegans* protein-coding genes. h) - j) Sequence logo of 10bp upstream and downstream of TSS. h) 4499 H3K4me3 depleted genes. i) complement set. j) All *C. elegans* protein-coding genes.

The same analysis was conducted on the 179 H3K4me3 enriched genes in *set-2(bn129)* and *cfp-1(tm6369)*. The motifs found by DREME and BMM motif have a much lower E-value and occurrence compared to the motifs found with the H3K4me3 depleted genes. The main finding is the inverse repeat TACNGTA (**Figure 3.8a & b** 2nd from the top). This motif is also enriched in the H3K4me3 depleted set (**Figure 3.6a** 6th place). A motif search using Tomtom (Gupta, et al., 2007) was not able to find a known motif. Since SeqPlot does not accept ambiguous nucleotide code, I used FIMO (Grant, et al., 2011) and CentriMO (Bailey & Machanick, 2012) to identify the location at which the motif could be enriched relative to the TSS. However, none of the software were able to identify a location with enrichment. A nucleotide frequency plot was also generated for these sets of genes. However, no obvious pattern or motif was identified (**Figure 3.8c-e**).

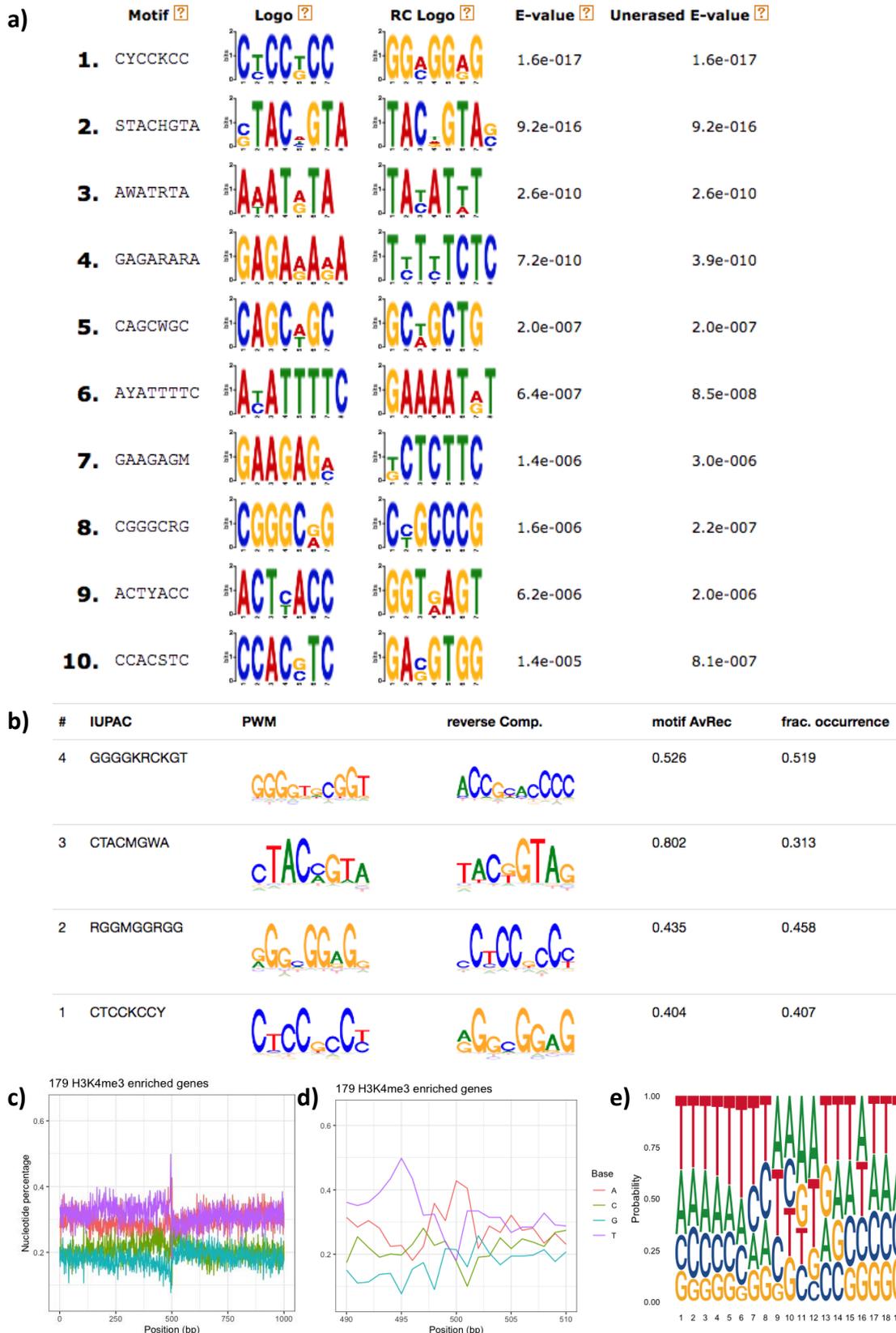


Figure 3.8 Sequence analysis around the TSS of the 179 H3K4me3 enriched genes in *set-2(bn129)* and *cfp-1(tm6369)*.

a) DREME output of the 179 sequences. b) BMM motif output. c) nucleotide frequency plot showing the average nucleotide at each position 500bp upstream and downstream of the TSS. d) A zoomed-in view of the nucleotide frequency plot of a 20bp window (10bp upstream and downstream of TSS). e) Sequence motif 10bp up and downstream of the TSS (TSS is at position 11).

3.6. Discussion

This chapter analysed the change in H3K4me3 levels around the TSS in *set-2(bn129)* and *cpf-1(tm6369)* mutants. The mutants showed a reduction of H3K4me3 in nearly 1/4 of all the protein-coding genes, while only a small set of genes have increased H3K4me3 levels (379 genes and 233 genes respectively). The extent of the H3K4me3 loss is much stronger than the H3K4me3 gain (**Figure 3.3**), supporting previous studies indicating a net loss of global H3K4me3 (Pokhrel, et al., 2019). The genes corresponding to loss of H3K4me3 are housekeeping genes, while the genes that saw a gain in H3K4me3 were associated with chromatin activities and development.

It is not surprising that genes with loss of H3K4me3 are housekeeping genes since housekeeping genes are constitutively active and are thus constantly transcribing, supporting the idea that H3K4me3 are markers of constitutively active genes (Barski, et al., 2007). Surprisingly, some genes have an increase in H3K4me3 levels in the COMPASS mutants. These genes are associated with developmental functions which are facultative genes (only transcribed when needed) and represent the “opposite” type to constitutive genes. On the other hand, the enrichment of many histone genes is very interesting. An explanation for this observation might be that due to the lack of H3K4 trimethylation post-translational modification, histone crosstalk is disrupted, which signals the cell that there is a problem with histone homeostasis, thus increasing the production of new histones. This increased histone transcription is then marked by H3K4me3 marks. The question is then, which protein is doing the methylation after the COMPASS-complex is non-functional. It could be SET-16/COMPASS, which has been observed to methylate a small subgroup of genes (Xiao, et al., 2011) or there could be another yet undiscovered methyltransferase in *C. elegans*.

The *de novo* motif discovery and sequence analysis aimed to identify characteristics in the DNA sequence that allowed them to be specifically targeted by the COMPASS complex. The most enriched motif for the 4499 H3K4me3 depleted genes is the SL1 splicing motif TTTCAG (**Figure 3.6**). Considering that more than half (~55%) of the genes in *C. elegans* are trans-spliced by SL1 (Saito, et al., 2013), this motif is expected to be found. However, when comparing the occurrence of this motif with the complement set (negative control) and all *C. elegans* genes (background control), it becomes apparent that the H3K4me3 depleted genes are enriched in this motif (**Figure 3.6**). If we consider that the H3K4me3 genes are constitutively active genes, this result indicates that constitutively active genes are enriched in SL1 trans-splicing. Currently, the function of trans-splicing is unknown, but it is hypothesized that it may play a role in translation initiation, as the motif is very close or

immediately adjacent to the start codon (Blumenthal, 2012) (**Figure 3.6j**). Proteins other than SL1 could use this motif as a binding target. Since the H3K4me3 levels of the 4499 genes are COMPASS dependent genes, the COMPASS complex could be recruited by this motif directly or indirectly. Further support for this observation and hypothesis comes from the reduced occurrence of this motif at the complement set (**Figure 3.7f & i**) and the absence of it in the 179 H3K4me3 enriched genes, which are suggested to be facultative genes.

Another characteristic identified in the H3K4me3 depleted gene set compared to the complement set (and the set of all protein-coding genes) is a larger AT skew just before the TSS and a drop in thymine frequency immediately after it (**Figure 3.7b-d**). While GC skew is believed to enhance R-loop formation (**Section 1.3.1**), not much is known about AT skew. Could this AT skew be an important feature for the H3K4me3 depleted gene set? More research would be required to answer this question. The reduction in thymine after the TSS is very interesting since the region is translated by the ribosome. How much this reduction in thymine frequency contribute to an amino acid preference/bias at the N-terminal and how this affects the protein is also a question that remains to be answered. I would hypothesize that in the case an amino acid preference/bias exists, it would function as an N-terminal signal peptide sequence or determining the half-life of the protein (N-end rule) (Varshavsky, 1997). A decreasing thymine frequency could see an increase in amino acids with codons low in uracils, such as lysine and arginine, while amino acids with uracil rich codons such as phenylalanine would decrease (**Figure 3.9**). Research on N-degrons (N-terminal residues that signal protein for degradation) identified three N-degron pathways: Pro/N-degron pathway, Arg/N-degron pathway and Ac/N-degron pathway. The Pro/N-pathways require proline to be the 2nd or 3rd amino acid to be recognized for degradation. The other two pathways recognize a large number of amino acids. However, Ac/N-degron pathway mainly acts on alanine, threonine, serine and methionine (Varshavsky, 2019). All pathways except the primary Arg/N-degron pathway, tend to require amino acids with low thymine frequency in their codons (except methionine and serine). This suggests that the 4499 H3K4me3 depleted genes, which are constitutively active genes, tend to be targeted by all but the Arg/N-degron pathway, resulting in shorter protein half-life. This may perhaps counteract the high transcription/translation rate of the constitutively active gene, to correctly balance the concentration of the protein quickly according to the needs of the cells.

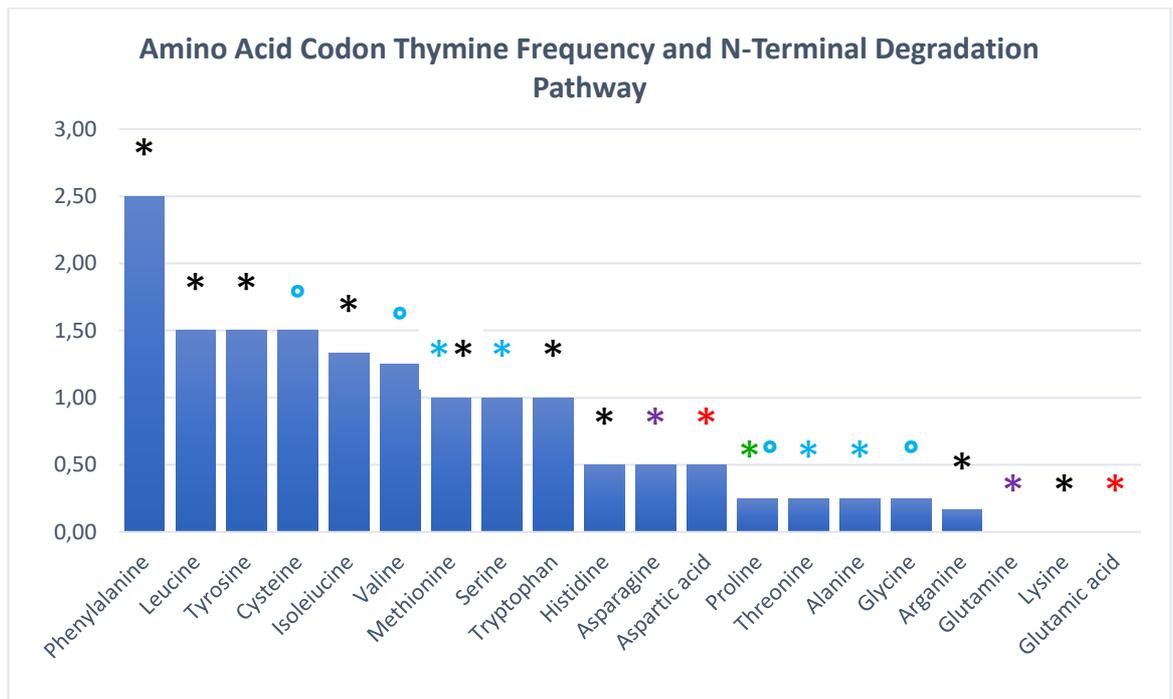


Figure 3.9 Average thymine frequency for each amino acid codon and the degradation pathway the amino acid is part of. The number of uracil in all codons of the same amino acid are added together and divided by the number of codons to obtain the average. The symbol colour above each bar indicates the N-degron pathway the amino acid is involved in. Black: primary Arg/N-degron pathway. Red: secondary Arg/N-degron pathway. Purple: tertiary Arg/N-degron pathway. Green: Pro/N-degron pathway. Blue: Ac/N-degron pathway. Circle shapes indicate that the amino acid is rarely used in the pathway. Certain amino acids are used by multiple pathways.

Chapter 4: R-loop levels in COMPASS mutants

4.1. Introduction

This section investigates the relationship between the epigenetic marker H3K4me₃, which was the focus in the previous chapter, and R-loops. The reason why R-loop is a suitable subject to associate with H3K4me₃ is that at the genomic level, the location where H3K4me₃ is found (at the TSS of active genes) coincide with the location of R-loop signal enrichment (Ginno, et al., 2012). This correlation prompted the investigation to identify the relationship between R-loops and H3K4me₃.

Here I am optimizing a method to quantify the level of R-loops in *C. elegans*. Using this optimised method, I then compare the R-loop levels between wild-type *C. elegans* and COMPASS mutants to identify how the deficiency in the epigenetic marker H3K4me₃ (or the absence of the COMPASS subunit) affect R-loop levels. Finally, I screen for helicases that affect R-loop levels in a COMPASS dependent or independent manner.

4.2. Characterisation of R-loops in the COMPASS mutants

Research on R-loop has mainly focused on yeast and human cells, but not so much on the model organism *C. elegans*. Hence, to use *C. elegans*, methodologies need to be adjusted and optimised to reliably measure R-loop levels that correctly reflect the underlying biology. Since this project aims to identify factors that affect the accumulation of R-loops, a quantitative method was required to quantify R-loop levels and compare between different samples. A straightforward method for this is the use of dot or slot blots (Vanoosthuysse, 2018). While this technique is relatively basic, its use with R-loops and especially with *C. elegans* nucleic acid has not been described in much detail. The standard dot blot method used for single-stranded DNA or RNA is not optimal for use in detecting R-loops since the main feature of R-loop is the association of DNA with RNA and thus cannot be denatured. In this section, the optimisation of the R-loop dot blot is described, and the reasoning behind various

changes is explained. Furthermore, the R-loop levels of *C. elegans* wild-type and COMPASS mutants are compared.

4.2.1. Dot/Slot blot method optimisation

The dot blot can be considered as a reduced version of other widely used blotting methods such as the western or southern blot. The sample of interest (protein or nucleic acid) is directly spotted on a membrane (e.g. nitrocellulose or nylon), instead of separating the sample first based on size and weight via electrophoresis and then transferred to a membrane. The subsequent steps of the dot blot are the same as other blotting methods. The samples are visualised via various stains (e.g. silver stain) or labels (e.g. antibodies) to identify the presence or the quantity of the target of interest. The dot blot does not require any specialised equipment, and the samples can be applied directly to the membrane by hand (creating circular blots). The slot blot uses an apparatus to make the sample loading more consistent and shaping the loading area into a rectangular slot. For R-loop dot blot, the conventional procedure starts with the extraction and purification of the gDNA sample of interest, followed by quantification of the sample. Then a known amount of the gDNA is directly spotted onto a membrane (and fixed). Afterwards, the membrane is blocked with a blocking solution and washed with a solution containing the primary antibody. Here the primary antibody is S9.6 as it is the only currently available antibody that specifically targets and binds to DNA:RNA hybrids. The antibodies that are unable to attach to the sample on the membrane are washed away, and the secondary antibody is applied that binds to the primary antibody to amplify the signal during visualisation.

Initially, the dot blot setup was done by spotting a known amount of genetic material onto a nitrocellulose membrane, leaving it to air dry, followed by blocking with 5% milk in tris-buffered saline (TBS) (blocking solution) overnight at 4°C. The membrane was then incubated with the S9.6 antibody for 4 hours at room temperature and washed three times with tris-buffered saline + Tween 20 (TBST) for 5 minutes each. Incubation with secondary antibody (2 µL of anti-mouse IgG antibody in 10 mL of blocking solution) was done for 2 hours at room temperature followed by a final 10 minute TBST wash. The membrane was then developed on an X-ray film using electrochemical luminescence (ECL). The resulting film of the pilot experiment is shown below (**Figure 4.1**).

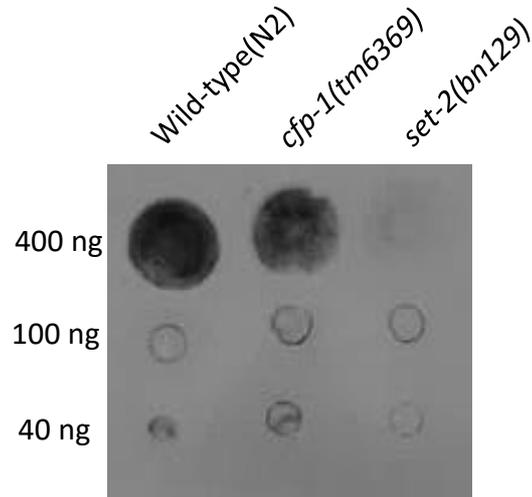


Figure 4.1 Dot blot of the R-loop pilot experiment. Genetic samples used in this blot were collected from adult wild-type(N2), *cfp-1(tm6369)* and *set-2(bn129)* worms fed on the standard laboratory *E. coli* strain OP50. The quantity of gDNA spotted is shown on the left. The R-loop signal is not evenly distributed within the “dot” but concentrated as rings for the samples with 100 ng and 40 ng of gDNA. This effect is due to the coffee ring effect where the surface flow concentrates most of the nucleic acid at the edge of the sample droplet.

The pilot dot blot shows that wild-type (N2) has a strong R-loop signal and *set-2(bn129)* shows a very weak R-loop signal, while *cfp-1(tm6369)* has a medium signal strength in between wild-type and *set-2(bn129)* at the 400 ng level (**Figure 4.1**). This indicates that R-loop levels are reduced in worms with a reduced H3K4me3 level. A coffee ring effect can be seen in this blot, where most of the signal is concentrated on a small area in the shape of a ring, instead of being evenly distributed. Such an effect happens when the suspended particles (e.g. nucleic acids) at the surface flows from an area of high surface tension (top of the drop) to an area of low surface tension (edge of the drop), known as the Marangoni effect. This effect happens in any non-equilibrium system with a surface tension gradient (Yunker, et al., 2011; Seo, et al., 2017).

In order to avoid the coffee ring effect, a slot blot machine was used. This machine pulls the liquid samples in the wells through the membrane by force using a vacuum pump, reducing the surface tension gradient and thus minimising the Marangoni effect. The setup and use of the machine are described in section 2.8. The resulting dot blot of the pilot experiment using a slot blot apparatus is shown in **Figure 4.2a**.

Although the concentration (and hence the quantity) of the sample gDNA was carefully measured and pipetted, it was not a guarantee that the intended quantity of gDNA pipetted reflects the actual quantity of gDNA loaded onto the membrane. To account for loading uncertainties, an additional loading control was included, which visualises the actual quantity of gDNA on the membrane. After the membrane has been developed and results captured on the X-ray film, the membrane was stripped by submerging it in mild stripping buffer (5 g

glycerin, 10 g SDS (1%), 10 mL Tween 20 in 1 L of deionized water and adjusted to pH 2.2) at room temperature for 10 minutes and again for 30 minutes, followed by two 10 minutes TBS wash and two 5 minutes TBST wash. The membrane was then blocked for 30 minutes with blocking solution and incubated directly with secondary antibody for 2 hours. A 10 minute wash with TBST followed by developing the film with ECL was then able to indicate whether the stripping was successful or not. Successfully stripped membranes have very weak or no signal when developed with no or only secondary antibody. After a successful stripping, the membrane was incubated with anti-dsDNA primary antibody (1:180 dilution in blocking solution) for 4 hours, which only targets double-stranded DNA. Afterwards, the membrane was washed three times with TBST for 5 minutes each and then incubated with the secondary antibody (anti-mouse antibody) for 1 hour. Washing and developing was done the same as previously. The resulting loading control (**Figure 4.2b**) revealed further problems with the current method design. Comparison of the two images in **Figure 4.2** shows that some of the nucleic acids on the membrane were lost (probably as the result of the numerous washing steps). This was because a uniform rectangular slot-shaped pattern was expected as the gDNA is evenly distributed along with the slot. However, **Figure 4.2b** shows that “chunks” of signals were lost especially in the 200 ng and 400 ng samples only leaving parts of the original rectangular slot-shaped signal (**Figure 4.2a**).

This loss of signal might be due to the nucleic acid not being adequately fixed on the membrane. The DNA in such a small spot might be too concentrated and could stack on top of each other, making them susceptible to be washed away. According to the manufacturer of the nitrocellulose membrane (Amersham™ Protran®), fixation of nucleic acids can be done by UV-crosslinking or baking the membrane in a vacuum oven. However, UV-crosslinking could damage double-stranded nucleic acids (Pall, et al., 2007), and in the Leeds laboratory, there is no access to a vacuum oven to bake the membrane (a regular oven cannot be used as nitrocellulose membrane is highly flammable).

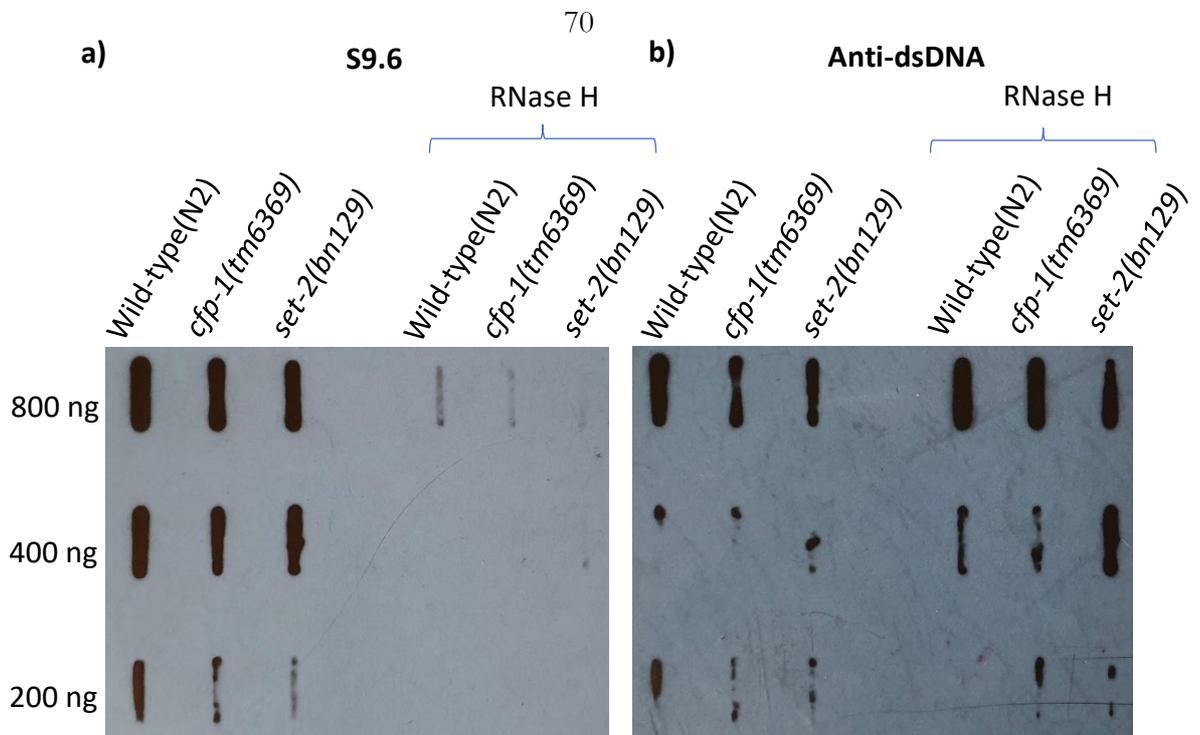


Figure 4.2 R-loop pilot experiment using the slot blot apparatus and anti-dsDNA loading control procedure. a) slot blot using the S9.6 primary antibody. b) loading control using the anti-dsDNA primary antibody, after stripping the S9.6 antibody from the membrane a). The sample nucleic acid was extracted from young adult wild-type(N2), *cfp-1(tm6369)* and *set-2(bn129)* worms fed on OP50. RNase H treated samples were used as a control to show that the signals indeed represent R-loops, as RNase H digests the RNA portion of the DNA:RNA hybrid. Mouse secondary antibody was used for both blots.

Consequently, the nitrocellulose membrane was replaced with a nylon membrane, which can be baked in a regular oven. Although the nylon membrane manufacturer suggests the use of 0.4M NaOH for nucleic acid dot blot, this was replaced by deionized water, as NaOH denatures double-stranded nucleic acids (Wang, et al., 2014), which is standard procedure for normal DNA blots, but not viable for R-loop blots. A slot blot using nylon membrane can be seen in **Figure 4.3a**. The result shows a similar trend of a weak R-loop in *set-2(bn129)* mutants. *cfp-1(tm6369)* mutants, on the other hand, has a signal strength comparable to wild-type.

The nucleic acid loading control method was also changed due to the time-consuming methodology of stripping the S9.6 primary antibody and reprobing with the anti-dsDNA antibody. The DNA loading control was visualized using methylene blue staining (as described in **Section 2.8.3.1**). Methylene blue is positively charged and binds to the negatively charged phosphate backbone of the nucleic acid (Vardevanyan, et al., 2013). It is a blue dye that attaches to DNA, and a higher concentration of DNA is visualized as darker blue hue under visible light (**Figure 4.3b**). However, once stained with methylene blue, the membrane cannot be reused.

Finally, owing to the difficulty of obtaining a suitable X-ray image, the imaging system G:BOX (Syngene) was used. This system has the advantage to automatically determine the

time required to capture the signal and displays an optimum image based on the strongest signal, reducing the amount of time and film needed (**Figure 4.4**).

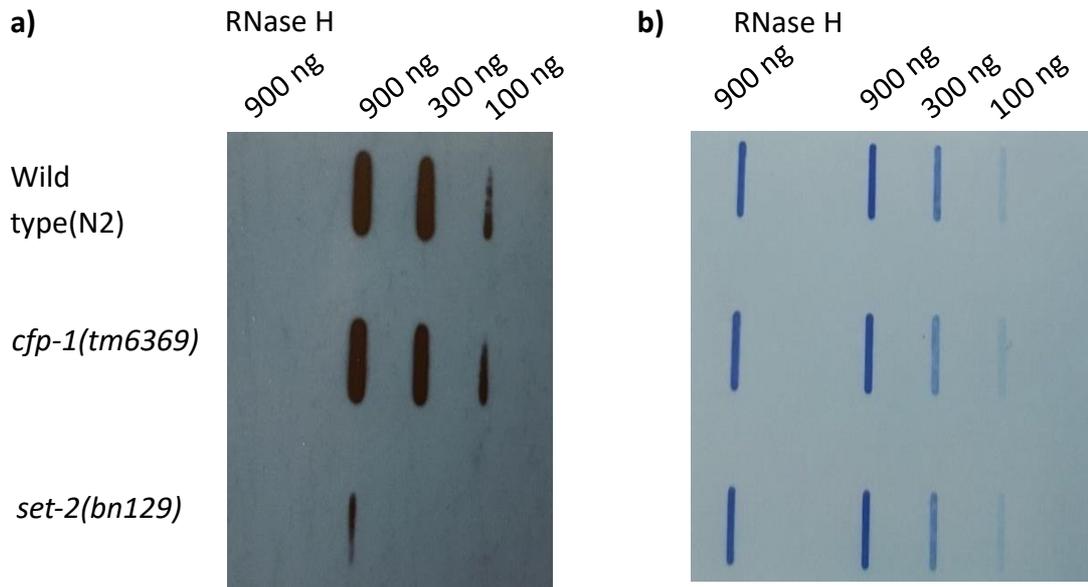


Figure 4.3 Representative image of the optimized slot blot method and methylene blue loading control. a) slot blot using the S9.6 primary antibody. b) loading control of a) using methylene blue staining. The amount of gDNA used is shown at the top. RNase H treated samples serve as a negative control. The sample nucleic acid was extracted from wild-type(N2), *cfp-1(tm6369)* and *set-2(bn129)* late embryo. The mothers were fed on OP50. The details on late embryo sample collection can be found in Section 4.2.3, and the staging of the late embryo can be found in Figure 4.6 replicate 1. The methylene blue staining has a larger linear range of detection, making a comparison between larger differences easier. The ECL used in the slot blot has a smaller range, allowing for better differentiation of target abundance within a narrow range.

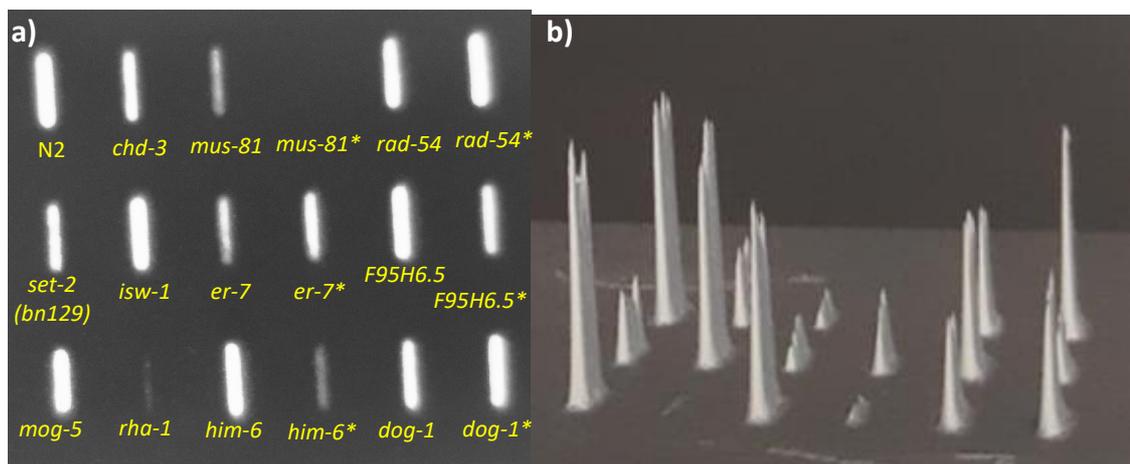


Figure 4.4 Representative images showing the output of the G:BOX. The slot blot results shown here are part of the helicase screen experiment covered in Section 4.4.3. a) image shows a classical view of the R-loop slot blot generated by G:BOX. b) image shows the “quantification” of the slot blot by G:BOX, with the height of peak correlating with the signal intensity. The gDNA samples were collected from L1 worms hatched in M9 buffer. The mothers were fed with various RNAi bacteria. N2 and *set-2(bn129)* worms were fed RNAi control bacteria (EV). 400ng of DNA was loaded in each well. Stars indicate biological replicates. A more detailed version can be found in Appendix 21.

4.2.2. The R-loop signal from adult worms show large variation

Using the optimized R-loop slot blot method (from the previous section), I wanted to measure the R-loop levels in young adult *C. elegans* worms. Specifically, I wanted to target co-transcriptional R-loops and avoid non-co-transcriptional R-loops. For this, young adult *C. elegans* were chosen, since they have a fixed number of somatic cells, which do not undergo replication. Due to the natural variation in developmental speed of individual worms, the developmental stage of a population consists of worms between the L4 and adult stage. In order to assess the stage of the collected worms, a small sample was stained with DAPI to identify the presence or absence of embryos. The target stage was between L4 and adult, where divisions are completed, and self-fertilization has just started. Optimally, animals would only carry a couple of embryos at most inside them. It is important to collect *C. elegans* in their early adulthood, as older adult animals can carry up to 15 embryos (Schafer, 2005), which would contaminate the sample with cells undergoing replication and thus potentially include R-loops formed as a result of DNA replication (transcription-independent R-loops). The timing for collecting the worms after spotting L1 onto NGM plate was around 60 hours for wild-type and 65 hours for COMPASS mutant worms. DAPI staining showed that the majority of worms started carrying embryos (**Figure 4.5**), but not nearly as much as the maximum of 15, indicating that the samples did not contain significant amount of replicating cells.

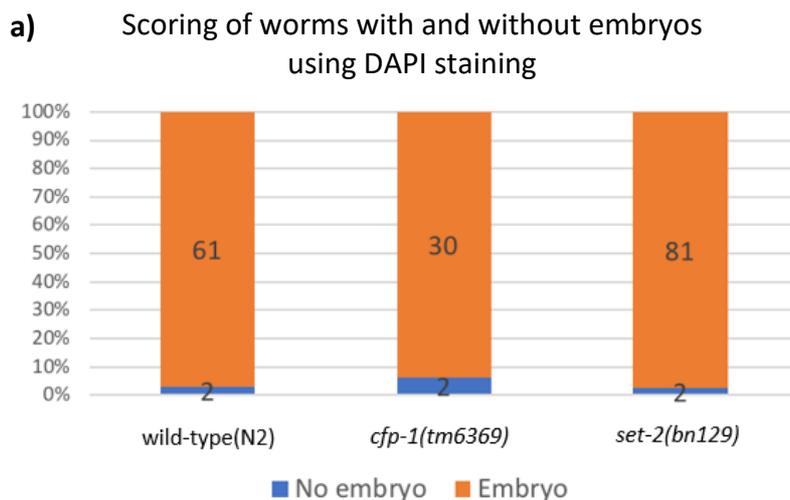


Figure 4.5 Scoring of adult *C. elegans* developmental stage. DAPI staining of worms shows the proportion of nematodes carrying embryos, indicating that reproduction has started. a) Proportion and the number of worms with and without embryos in N2, and *cfp-1(tm6369)* and *set-2(bn129)* mutants. b) representative image showing the DAPI staining of a young adult worm with embryos.

The R-loop dot blot with young adults gave varying results, making it difficult to reproduce robust data (see **Appendix 2** for an example of inconsistent wild-type and *set-2(bn129)* result). One reason for this discrepancy could be attributed to bacterial accumulation in the gut and pharynx (Portal-Celhay, et al., 2012), that could contaminate the *C. elegans* nucleic acid sample. It has been shown that R-loops are also present in *E. coli* (Kogoma, 1997). Furthermore, it is well observed that bacterial gut accumulation can have a harmful effect on *C. elegans*, including OP50 bacteria accumulation (Garigan, et al., 2002), which in turn could result in excessive stress or increased transcription and thus elevated R-loop formation.

While one possibility to account for this is to UV- or heat-kill bacteria, to prevent colony formation in the gut, arrested and dead OP50 will affect worms differently. For example, it has been shown that heat-killed OP50 has reduced nutrients, which would explain the favouritism of the nematode towards live OP50 (Qi, et al., 2017). Rather than experimenting with different food sources, I chose to extract the genetic sample from a different developmental stage, where the effect of the bacterial food source is minimized. The only developmental stage that fit the criteria is the time point when the nematode finishes embryonic development and before it requires bacterial food. The approach to collecting this stage is described in the following section (**Section 4.2.3**).

4.2.3. Late embryo and L1 show reduced R-loop signal in *set-2(bn129)* mutants compared to wild-type worms

In order to collect samples free of genetic material contamination originating from bacteria, I shifted focus to using progenies from bleach-synchronised animals. The bleaching protocol includes many washing steps that should remove any remaining bacterial contaminants, leaving embryos to develop in a “sterile” environment. The embryos were allowed to develop in M9 buffer, to either reach a developmental stage close enough to hatching (late embryo) or left to hatch overnight (L1). Embryos are collected 5.5-6 hours after bleaching, depending on the nematode strain. The developmental stage of the late embryo collection is summarized in **Figure 4.6**, and a representative slot blot is shown in **Figure 4.3**. Worms freshly hatched in M9 buffer find themselves in an environment devoid of food, which could lead to the L1 arrested (diapause) state. Worms in this state show various adaptations such as stress resistance, that allows L1 to survive in the absence of food for around 10 days in M9 buffer (Baugh, 2013). Alternatively, worms can develop into the Dauer stage at the end of L1 (and L2d) molting. This developmental program is mainly dependent on the population density of the worm but requires the presence of some scarcely available food

(Hu, 2007; Baugh, 2013). In this experimental setup, dauer formation was unlikely to occur because food sources were completely absent owing to the multiple washing steps, and the relatively short overnight hatching window (~16 hours) is not enough time for the worms to enter their first molting stage (~21 hours from embryo to L2d).

The idea of using overnight hatched L1 worms instead of late embryos was considered as it is less labour-intensive and would be advantageous for large scale experiments (such as the helicase screening in **Section 4.4.3**). Furthermore, the stage of the worm will be more synchronized, as the worms are unable to develop past L1 stage overnight. However, the effect of overnight hatching on R-loop formation is unknown and need to be tested first. The R-loop slot blot using genetic samples from overnight hatched L1 worms (**Figure 4.7**) showed a similar result as observed with the late embryos (**Figure 4.3**): The wild-type(N2) samples have strong R-loop signal, while *set-2(bn129)* samples show a relatively weak R-loop signal. The R-loop signal from *cfp-1(tm6369)* is comparable in strength to that of wild-type. This indicates that overnight hatched L1 can be used as an alternative to late embryos for comparing the R-loop levels between wild-type(N2) worms and COMPASS mutants.

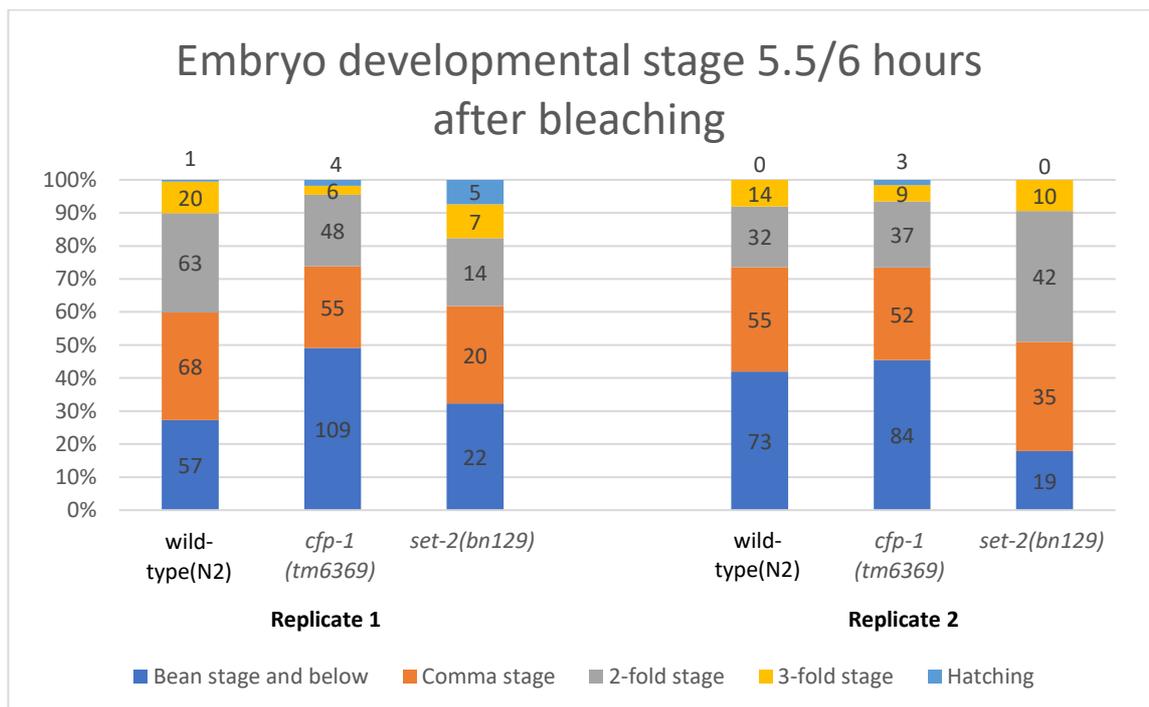


Figure 4.6 Embryo developmental stage scoring. The developmental stages were scored 5.5 hours (for wild-type) and 6 hours (COMPASS mutant) after bleach-synchronizing *C. elegans*. The numbers on the column are the number of animals scored. 2 biological replicates were done. The slot blot of Replicate 1 can be seen in Figure 4.3.

4.2.4. OP50 and EV diet results in the same R-loop pattern between wild-type and *set-2(bn129)* mutants

In order to use the powerful RNAi tool (which is used for the helicase suppressor screen in **Section 4.4.3**), the effect of the RNAi bacteria on R-loop levels needs to be investigated first. Specifically, the effect of empty vector bacteria (HT115) food needs to be compared to OP50. While OP50 is the standard laboratory food strain, the empty vector bacteria is a more suitable control food in RNAi experiments, as it is a closer resemblance to the RNAi bacteria. The empty vector bacteria is a modified *E. coli* strain that carries an “empty” L4440 plasmid. This plasmid is modified and used to express the dsRNA of target genes and is the main feature of the RNAi bacteria.

As R-loop levels naturally vary and dot/slot blot is not a highly sensitive method of quantification (for comparing small differences) (Vanoosthuyse, 2018), I focused on comparing the difference of R-loop signal strength between wild-type (control) and *set-2(bn129)* mutants, as this difference was sufficiently large to be reliable (**Figure 4.3**). The *cfp-1(tm6369)* mutants, on the other hand, were difficult to obtain consistent results. In addition, experimental difficulties (smaller brood size) and considerable variation in phenotypes (e.g. developmental speed) further argue against the use of the *cfp-1(tm6369)* mutant. Therefore, *cfp-1(tm6369)* mutants were not used in the helicase screen in sections 4.4.3.

Figure 4.7 shows the R-loop levels of L1 worms whose mothers were fed on the empty vector bacteria (HT115) and OP50. For both bacterial food diet, a strong signal was observed in wild-type(N2) and a weak signal was found in the *set-2(bn129)* mutants, showing that the difference in R-loop accumulation between wild-type and *set-2(bn129)* mutant is conserved under empty vector bacteria (HT115) diet. This indicates that the empty vector bacteria (HT115) can be used for RNAi work on R-loop.

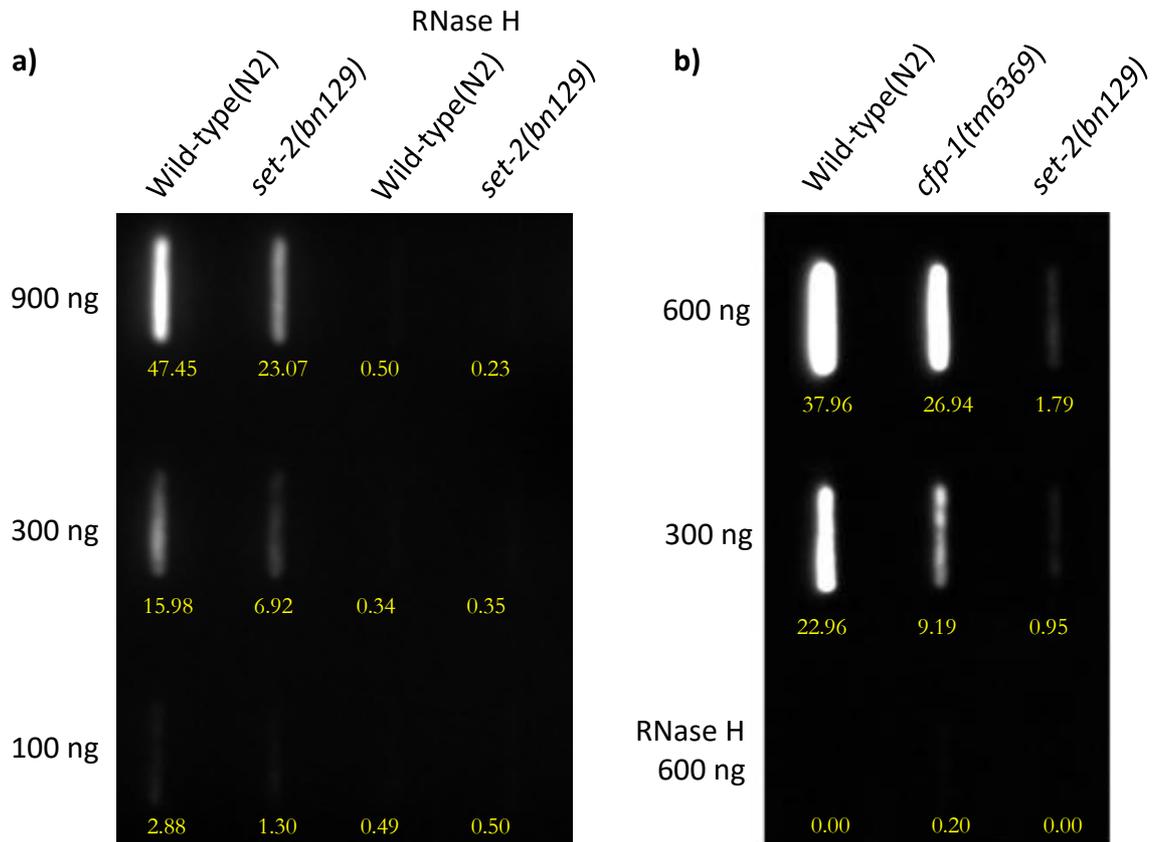


Figure 4.7 G:BOX image comparing the effect of different diets on the R-loop signal using hatched L1 worms. a) mothers were fed on OP50. b) mothers were fed on empty vector (HT115). In both cases, *set-2(bn129)* mutant has a much weaker R-loop signal compared to wild-type(N2). Yellow numbers show the signal quantification based on ImageJ.

4.2.5. Discussion

Initially, the plan is to use young adults, as these worms have finished development. The advantage of this is due to the unique feature of *C. elegans* to have a fixed number of somatic cells, which means that these cells do not undergo any more replication. R-loops found would be the result of transcription rather than replication (except the embryo and germ cells, which could be removed by using mutants that fail to develop germline such as *glp* mutants). However, the inconsistent result from using samples extracted from young adult worms, possibly due to bacterial accumulation inside the worm (Portal-Celhay, et al., 2012), prompted me to dismiss this developmental stage for further experiments. The DNA extraction from the adult worm would also extract DNA from residual bacteria inside the intestine, that could also contain R-loops. This would lead to the final DNA extract being a mix of *C. elegans* and *E. coli* DNA, which the S9.6 antibody cannot distinguish. The degree to which this affects the outcome depends on the extent of bacterial accumulation inside the worm during DNA extraction. Secondly, the formation of R-loop has been suggested to be

affected by stress (Lang, et al., 2017). Although OP50 is the standard laboratory food for the worms, it has been shown that OP50 might not be as optimal as initially thought. Accumulation of OP50 in *C. elegans* has a negative effect on nematode health (Garigan, et al., 2002). One way to reduce the bacterial impact is to use dead bacteria that are unable to form colonies inside the worm. However, a diet of dead bacteria affects the worm's health differently, such as vitamin B2 deficiency (Qi, et al., 2017). Furthermore, dead bacteria still contains nucleic acid and how long this nucleic acid persists inside the intestine/worm before it is broken down is also unknown. The effect of a dead bacterial diet on R-loop formation would have required extensive investigation and did not fit into the scope of this PhD project.

This limits the suitable sample collection from worms to developmental stages just before the worms come in contact with bacteria food. Embryos of worms that underwent the bleaching protocol (**Section 2.1.7**) can be kept in sterile M9 buffer while they undergo embryonic development until they hatch. Keeping these hatched worms for an extended period in the sterile and food-starved M9 buffer can arrest the worm's development (L1 diapause), which leads to specific physiological and transcriptional adaptation, such as lower activity and increased stress resistance (Baugh, 2013). Therefore, the genetic material needs to be extracted from freshly hatched worms. Practically, the bleached embryos are left to develop for 5.5-6 hours or left overnight (16 hours). While starvation under the overnight hatching method was a concern, as L1 might have entered diapause, the R-loop signal pattern was similar to the samples from the late embryo where feeding has not even started (**Figure 4.7** and **Figure 4.3**), indicating that the overnight hatching method is viable.

In this chapter, I have described an optimized method for the detection and comparison of R-loop levels in *C. elegans* using the slot blot method. This method is relatively straight forward and inexpensive; however, it does require a relatively large amount of DNA, requiring thousands of worms. Due to the natural variation in R-loop accumulation, this method is not suitable to distinguish small differences in R-loop levels.

Since both H3K4me3 and R-loop accumulate at actively transcribed genes, I wanted to identify if this correlation has a causal link. In order to do this, I measured the R-loop level in two COMPASS mutants *set-2(tm6369)* and *cfp-1(tm6369)*, which have drastically reduced H3K4me3 levels (Pokhrel, et al., 2019), using the optimized method described in this chapter. The results show that in the *set-2(tm6369)* mutants, there is a drastic reduction in R-loop level, supporting the idea that R-loop could be H3K4me3 dependent or at least dependent on SET-2. Results from *cfp-1(tm6369)* mutants are not yet conclusive (due to difficulty of reproducibility), which could be attributed to the large varying phenotype

observed in this mutant. They sometimes develop slower than *set-2(tm6369)* even though they were grown under the same condition. Other times they produced fewer progenies even though they are at a similar developmental stage to *set-2(tm6369)*. This variation of *cfp-1(tm6369)* mutants paired with the sensitivity of R-loop accumulation could be the reason that made the results unreliable.

It is difficult to ascertain whether SET-2 or H3K4me3 affect R-loop levels. Unfortunately, the results using *cfp-1(tm6369)* mutants is inconclusive; otherwise, it would be very useful in determining if H3K4me3 is the main player the phenotype is SET-2 specific. Apart from *cfp-1(tm6369)*, other COMPASS complex mutants could be used, where a strong H3K4me3 reduction has been observed in L1.

SET-2 has an RNA recognition motif domain (Wormbase, 2019), which is absent in CFP-1. Potentially, SET-2 could bind to the nascent RNA via the RNA recognition motif domain and recruit other proteins that help to form or stabilize R-loops. Helicases could be potential targets proteins that affect R-loop formation or resolution. Recent research suggests that apart from the standard unwinding function of helicases, they can also have an annealing/rewinding function thereby forming double-stranded nucleic acids from two single strands (Wu, 2012; Manosas, et al., 2013). SET-2 binding to R-loop could prevent helicases from unwinding R-loops or could recruit helicases that help re-winding/forming R-loops. If SET-2 does not directly bind to R-loops, it might still influence R-loop levels by targeting proteins that are responsible for R-loop formation/resolution. SET-2 could methylate proteins directly responsible for R-loop formation or removal. Proteins that negatively affect R-loop levels may be deactivated by methylation from SET-2, while proteins that positively affect R-loops could be activated. Although SET-2 has only been identified as part of the COMPASS complex, it does not exclude SET-2 from forming different complexes or have independent function altogether. CFP-1, for example, is inferred to have COMPASS independent functions (Pokhrel, et al., 2019). In *S. pombe*, the homolog Set2 has been shown to associate with RNA Polymerase II by directly binding to its C-terminal domain and plays an essential role in transcription elongation (Li, et al., 2002; Kizer, et al., 2005).

If we suppose that H3K4me3 itself is directly responsible for R-loop levels, then we can hypothesize that H3K4me3 could recruit relevant proteins such as helicases to support R-loop formation and accumulation. Another hypothesis could be that H3K4me3 promotes transcription, which then increases R-loop formation. However, current data suggest that H3K4me3 does not affect transcriptional activity much (Clouaire, et al., 2012).

4.3. Development of antibody-independent R-loop purification

The R-loop detection method optimized in the previous section (**Section 4.2**) relies on the specificity of the S9.6 primary antibody to target R-loops. This antibody specifically binds to DNA:RNA hybrid and not double-stranded DNA and RNA and ribosomal RNA (Boguslawski, et al., 1986). It is currently the standard tool for R-loop research. Recent concerns have been put forward regarding the reliability of the S9.6 antibody for binding to DNA:RNA hybrids. It has been suggested that S9.6 affinity towards the hybrid varies with sequence (König, et al., 2017). Furthermore, S9.6 can recognize RNase III-sensitive dsRNA and bind to similar structures (e.g. hairpin RNA); however, this is only impactful in organisms producing significant dsRNA loads (Hartono, et al., 2018). A relatively new affinity reagent has been developed to act as an alternative to S9.6 antibody. The mutated human RNASEH1 protein that lost its DNA:RNA specific endonuclease activity, but retains its binding competence. The advantage of RNase H compared to S9.6 is that it would reflect the biology more accurately, by binding to biological relevant DNA:RNA hybrids (Since RNase H is naturally produced in organisms). However, comparison of DRIP-seq (DNA:RNA immunoprecipitation) which uses the S9.6 antibody, with the technically equivalent DRIVE-seq (DNA:RNA *in vitro* Enrichment) that utilizes this mutated RNASEH1, shows that the latter produces a weaker signal and was only able to identify a fraction of the genes compared to the former method (Ginno, et al., 2012; L. Chen, et al., 2017).

Currently, R-loop research has a strong need for an alternative method of targeting (and purifying) R-loops. To this end, I propose a new method that does not rely on S9.6 antibody to ‘pull out’ R-loops. This chapter discusses the theoretical workings of the novel method and provides foundational work for practical implementation using commercially available reagents.

4.3.1. R-loop purification by nuclease digestion and mass separation

The theory behind an antibody-independent R-loop purification method follows the idea that, if it is not possible to pull out the R-loop, then removing everything that is not R-loops will yield the same result. The method to achieve this is by digesting the purified DNA using various endo- and exonucleases that only affect ssDNA, dsDNA and ssRNA. This would

hypothetically result in a solution of comparatively long strands of DNA:RNA hybrids and ideally mono- and dinucleotides of ssDNA, dsDNA and ssRNA (dsRNA is generally broken down in eukaryotes through RNAi pathways). By using mass or size separation techniques, such as gel electrophoresis and size exclusion chromatography, the long and heavy hybrid strands can then be separated from the short and light mono- and dinucleotides (**Figure 4.8**).

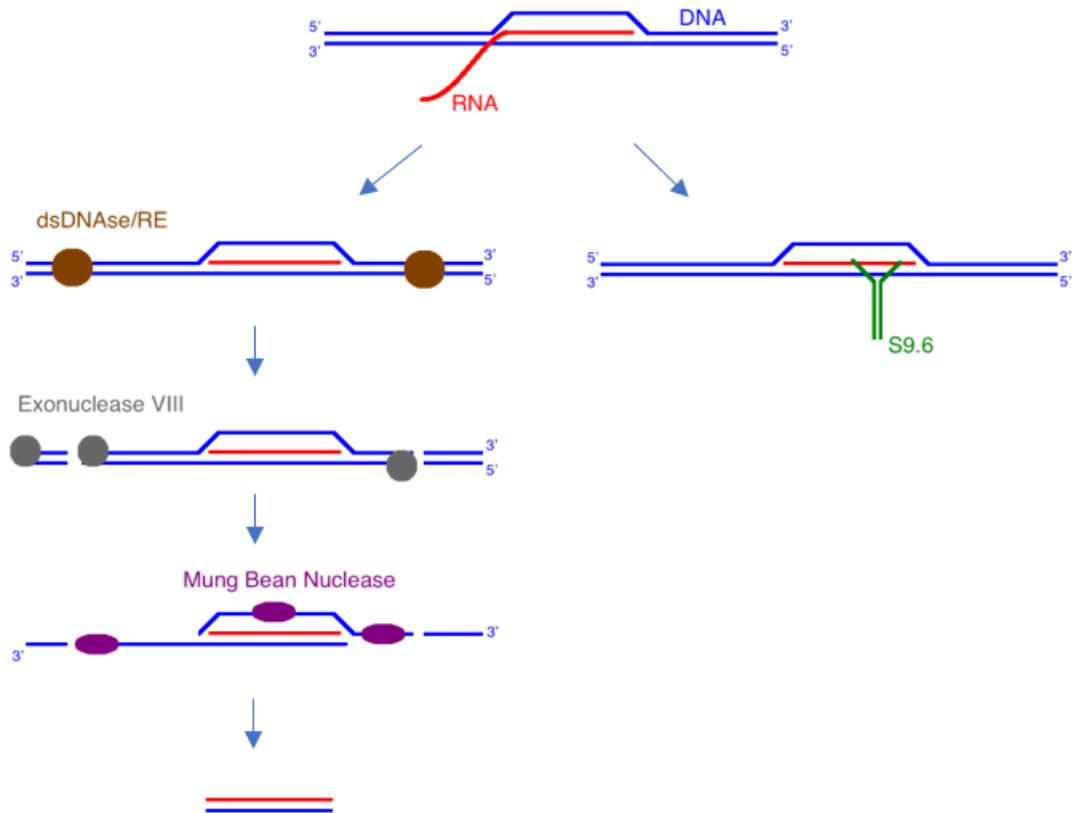


Figure 4.8 The theoretical mechanism of the antibody-independent R-loop purification protocol. The nucleic acid in its native state containing an R-loop will be purified from cell samples. During the purification step, RNase A removes the ssRNA overhang of the R-loop. Following the left path down, the addition of dsDNAse or RE (restriction enzyme) will cut dsDNA into smaller pieces, leaving R-loops intact. Next Exonucleases VIII will attach at the end of the dsDNA and cut the 5' strand into mononucleotides and leaving behind a 3'-ssDNA. Finally, the Mung Bean Nuclease will degrade any ssDNA and ssRNA overhangs into mononucleotides, leaving only the DNA:RNA hybrid portion. The right path shows the binding of S9.6 onto R-loops in gDNA.

Pilot Experimental Procedure

The experimental procedure is a series of nuclease digestion reactions that cleave the gDNA into smaller units with each digestion step, ultimately digesting all nucleic acids except DNA:RNA hybrids into mono- and dinucleotides.

For 1 μg of purified gDNA, 1 μl of dsDNAse (Thermos Fisher) and 1 μl of 10x NEB (New England Biolabs) CutSmart® buffer were mixed. The volume was then made up to 10 μl

with nuclease-free water and left in an incubator at 37°C for 15-30 minutes. For the alternative method utilizing restriction enzymes instead of dsDNase, the 1 µl dsDNase was replaced with 1 µl of the restriction enzyme “cocktail” (0.2 µl of each restriction enzyme: BsrGI-HF, XbaI, SspI-HF, HindIII-HF and EcoRI-HF) and 1-2mM of spermidine was added (spermidine increases the accuracy of restriction enzymes (Pingoud, 1985)). After the incubation, 2 µl of Exonuclease VIII truncated (NEB) was added to the mixture, followed by another 0.5 - 1 hour (dsDNase) or 2 hours (restriction enzyme) incubation at 37°C. Finally, 1-2 units of Mung Bean nuclease was added to the mixture. Mung Bean Nuclease requires zinc to function. The Mung Bean Nuclease (Promega) used here did not contain zinc donor in its storage buffer, thus 1 µl of 1mM ZnCl₂ was manually added to the reaction. The mixture was then incubated for 10-15 minutes at 37°C.

4.3.2. dsDNase endonuclease digestion

dsDNase from Thermo Scientific is an engineered DNase that has high specificity for dsDNA, leaving RNA and ssDNA intact. Thermo scientific stated that tests with fluorescently labelled oligonucleotides showed no activity on DNA:RNA hybrids under the recommended protocol and extremely low specificity for the hybrids even at 10x higher than the recommended concentration (personal communication). Since the supplied dsDNase buffer from Thermo scientific is proprietary, it is unknown whether the downstream enzymes would work with the buffer provided. Therefore, the activity of dsDNase was tested with the CutSmart® buffer from NEB, which is used for the enzymes downstream in the protocol and the composition of the buffer is publicly available (New England Biolabs, 2019). The testing for dsDNase activity in the NEB CutSmart® buffer showed that the activity of dsDNase is similar in both the NEB CutSmart® and dsDNase buffer (**Figure 4.9**).

Following the success of dsDNase digestion in NEB CutSmart® buffer, the DNA was restricted with Exonuclease VIII truncated. This exonuclease is functionally equivalent to Lambda Exonuclease, with the only difference being that Exonuclease VIII does not require phosphorylation at the 5'-end. Both degrade the 5'-strand of the dsDNA to mononucleotides (leaving the 3'-strand as an ssDNA) at a relatively slow pace of 19 nucleotides/second (Joseph & Kolodner, 1983; Lovett, 2011). Exonuclease VIII truncated (NEB) is RNase-free (personal communication).

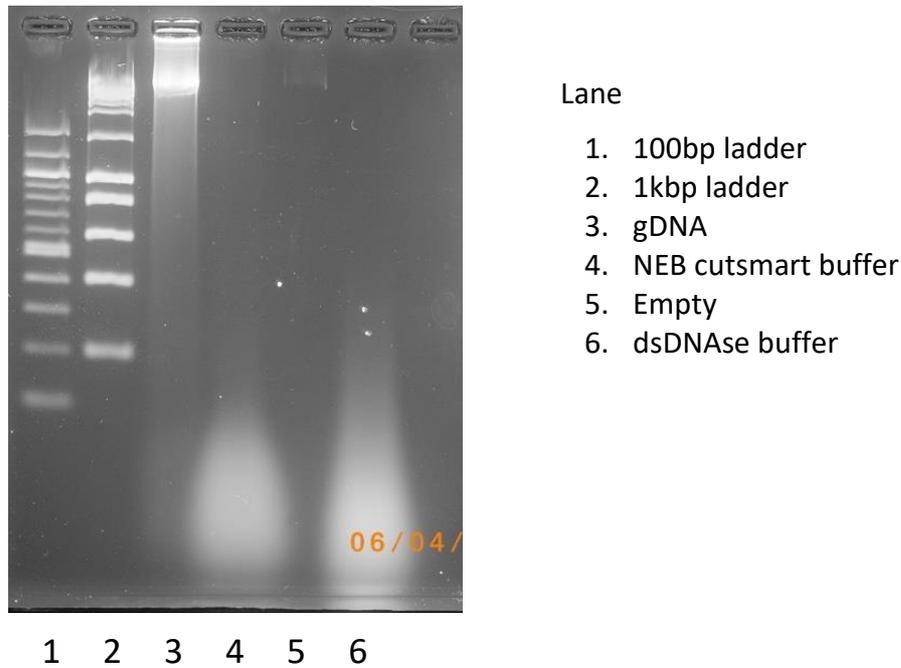


Figure 4.9 Comparison of the efficiency of dsDNase in dsDNase buffer and NEB CutSmart® buffer. The gDNA used here was extracted from HEK293T cells³, and the digestion reactions were done for 30 minutes at 37°C. The agarose gel did not contain any ethidium bromide and was stained with SYBR gold (Invitrogen) after electrophoresis for 40 minutes.

Finally, the nucleic acid was digested with Mung Bean Nuclease, which degrades both ssDNA and ssRNA in both directions, but does not digest double-stranded nucleic acid unless at a very high concentration (Promega, 2016; Valsala & Sugathan, 2017; Epicentre, n.d.). This digestion removes the single-stranded DNA left behind from the Exonuclease VIII digestion as well as any RNA overhangs from the R-loops and potentially the displaced single-stranded DNA. Furthermore, any remaining ssRNA that was not digested by RNase A during DNA purification would be degraded in this step. **Figure 4.10** shows the size composition of digested gDNA after each digestion step. The smear pattern on the gel electrophoresis images coincides with the expected shape. It can be seen that after dsDNase digestion, most of the gDNA were cut to below 100 bp (**Figure 4.10** left). The exonuclease VIII further reduced the mass of the nucleic acid marginally through the digestion of the 5'-strand (**Figure 4.10** middle). Finally, the Mung bean digestion removed all the remaining single-stranded nucleic acid (**Figure 4.10** right).

³gDNA samples are harvested from human embryonic kidney cells (HEK293T), due to the less labour intensive gDNA sample acquisition compared to *C. elegans*.

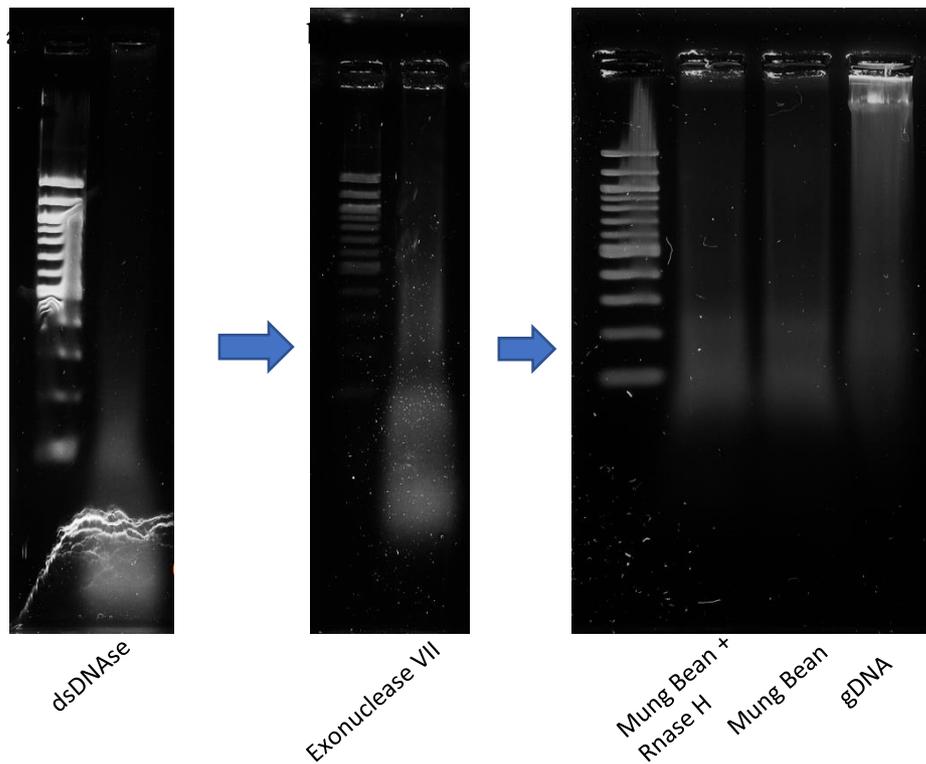


Figure 4.10 Nucleic acid mass composition after each nuclease digestion reaction using dsDNAse. The dsDNAse digestion was done at 37°C for 15 minutes (left), followed by an exonuclease VIII digestion at 37°C for 30 minutes (middle) and finally a Mung Bean nuclease digestion at 37°C for 10 minutes. The sample was then incubated for another 1 hour at 37°C with or without RNase H (right). The ladder is 100 bp ladder in all three gel images. All agarose gels were stained with SYBR gold for 40 minutes after electrophoresis. Each well should have around 900 ng DNA.

Theoretically, under complete digestion, the remaining large nucleic acid strands should be enriched in DNA:RNA hybrids. A slot blot of the final digest with the S9.6 antibody, however, did not show any R-loop signal and the methylene blue staining did not even measure any nucleic acid in the digest (**Figure 4.11**).

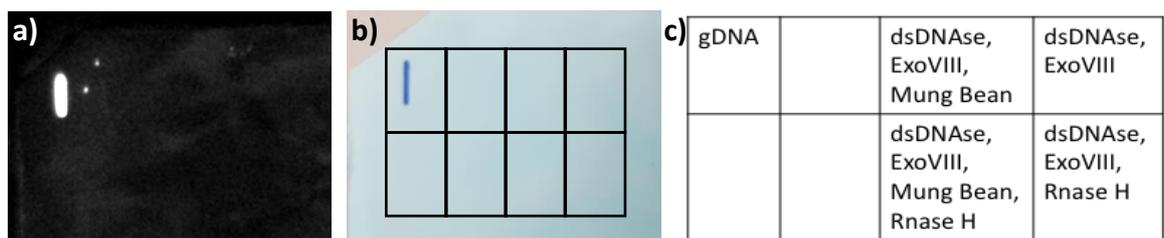


Figure 4.11 R-loop slot blot of nucleic acid samples after various digestion steps. a) Only the undigested gDNA shows R-loop signals. b) The DNA loading control shows the absence of any digested nucleic acid. c) The schema represents a section of the slot blot machine wells and shows the position where the nucleic acid was loaded onto the membrane. 1 μ g of DNA was loaded onto each well.

The nylon membrane (Amersham Hybond-N+) used in **Figure 4.11** has a pore size of 0.45 μ m and should retain nucleic acids larger than 50 bp (GE Healthcare Life Sciences, 2019). While the majority of nucleic acid was cut to below 100 bp and completely digested after Mung Bean nuclease digestion, there was still visible fluorescence of nucleic acids above 100

bp size (**Figure 4.10** right) and was expected to be retained on the membrane. Multiple factors could have played a role leading to the result seen in **Figure 4.11**. The digested mixture could contain too many enzymes that obstruct the DNA from attaching to the membrane and/or the remaining DNA could be too little to be detected by the methylene blue staining.

Another concern was that the dsDNAse from Thermo Fisher Scientific might digest the DNA:RNA hybrids, because this enzyme is an engineered shrimp DNase (Thermo Fisher Scientific, 2019) and DNases such as DNase I do have activity against DNA:RNA hybrids (Valsala & Sugathan, 2017). Since this experiment used the NEB CutSmart® buffer during the dsDNAse digestion instead of the provided dsDNAse buffer, Thermo Scientific was not able to guarantee that the DNA:RNA hybrids will remain intact (personal communication). Therefore, an alternative option using restriction enzymes was tested in the following section.

4.3.3. Restriction enzyme cutting

The ability of restriction enzymes to cut DNA:RNA hybrids is mostly unknown. Two studies have looked into specific type II restriction enzymes and found that only some restriction enzymes are able to cut DNA:RNA hybrids (Molly & Symons, 1980; Murray, et al., 2010). Out of the 223 enzymes in the Murray, et al. (2010) study, 5 restriction enzymes were chosen (that do not have any restrictive activity on DNA:RNA) that are available from NEB and work with the same buffer (CutSmart®) at the same temperature. These enzymes are BsrGI-HF, XbaI, SspI-HF, HindIII-HF and EcoRI-HF.

The restriction enzyme digestion follows the same protocol as dsDNAse digestion, but the dsDNAse was replaced with the restriction enzyme “cocktail” (0.2 µl of each restriction enzyme). The resulting digestion by the restriction enzyme “cocktail” can be seen in **Figure 4.12**.

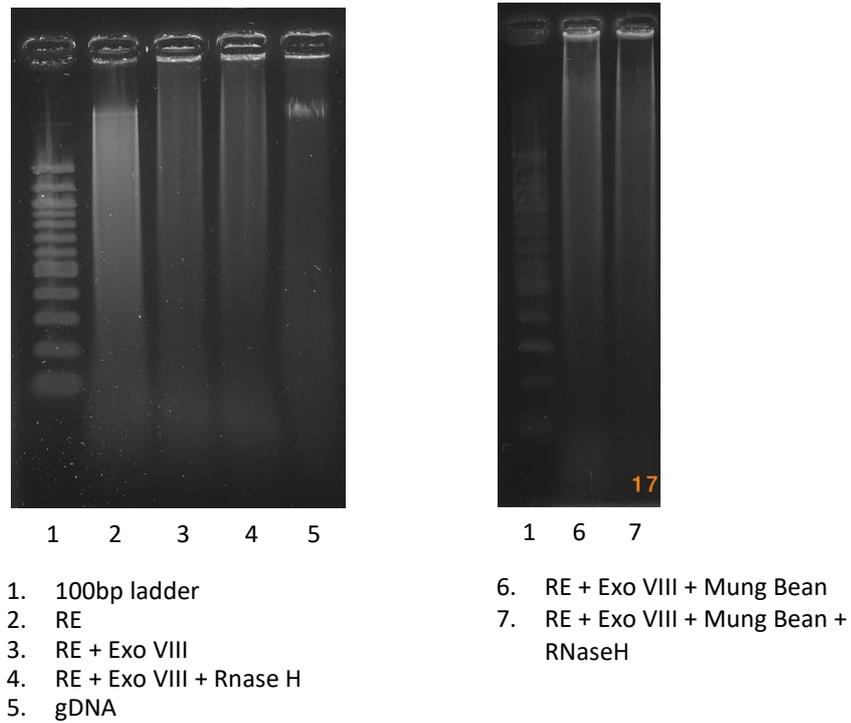


Figure 4.12 Nucleic acid mass composition after each nuclease digestion reaction using restriction enzymes. All digestion reactions were performed at 37°C for 2 hours, except Mung Bean digestion, which was incubated at 37°C for 15 minutes. RE = restriction enzyme cocktail. Exo VIII = Exonuclease VIII. Wells 2-5 contain 500 ng DNA, and wells 6 and 7 contain 1250 ng DNA.

The digestion with the restriction enzymes resulted in larger fragments compared to the digestion with dsDNAse, as the signal is stronger towards the top of the gel electrophoresis image (**Figure 4.12** lane 2 compared with **Figure 4.10** left). After the Exonuclease VIII and Mung Bean nuclease digestion steps, the final nucleic acids mixture was not digested as completely compared with the dsDNAse method. This can be seen due to the much stronger signal in lane 6 and 7 of **Figure 4.12** compared to **Figure 4.10c** (using the 100 bp ladder as reference).

4.3.4. Buffer exchange

A potential problem with using many nucleases is the aggregation of chemicals and proteins in the reaction mixture. For example, all the nucleases used here are stored in a 50% glycerol solution. The addition of each nuclease increases the concentration of glycerol in the digestion mixture. High glycerol concentration affects enzyme activity, such as reduced sequence specificity of restriction endonucleases (New England Biolabs, 2019) and reduced enzyme activity due to higher viscosity (Uribe & Sampedro, 2003). As such, digestion reaction mixtures are suggested to not exceed more than 5% glycerol concentration (New England Biolabs, 2019). In this experimental setup, however, the final glycerol concentration

nearly reaches 10%, which could negatively impact the digestion reaction. Therefore, in order to remove excess glycerol (and enzymes), a buffer exchange step (using the Monarch® PCR & DNA Cleanup Kit) that purifies the nucleic acid was incorporated after the Exonuclease VIII digestion step. The disadvantage of using an additional buffer exchange step is the loss of nucleic acid, with a typical recovery of between 50%-90% depending on the size of the nucleotide, ranging from 50 bp up to 25 kb (New England Biolabs, 2019). For more accurate quantification, gel electrophoresis was done on precast polyacrylamide gels (Novex™ TBE Gels, 4-20%) (**Figure 4.13**).

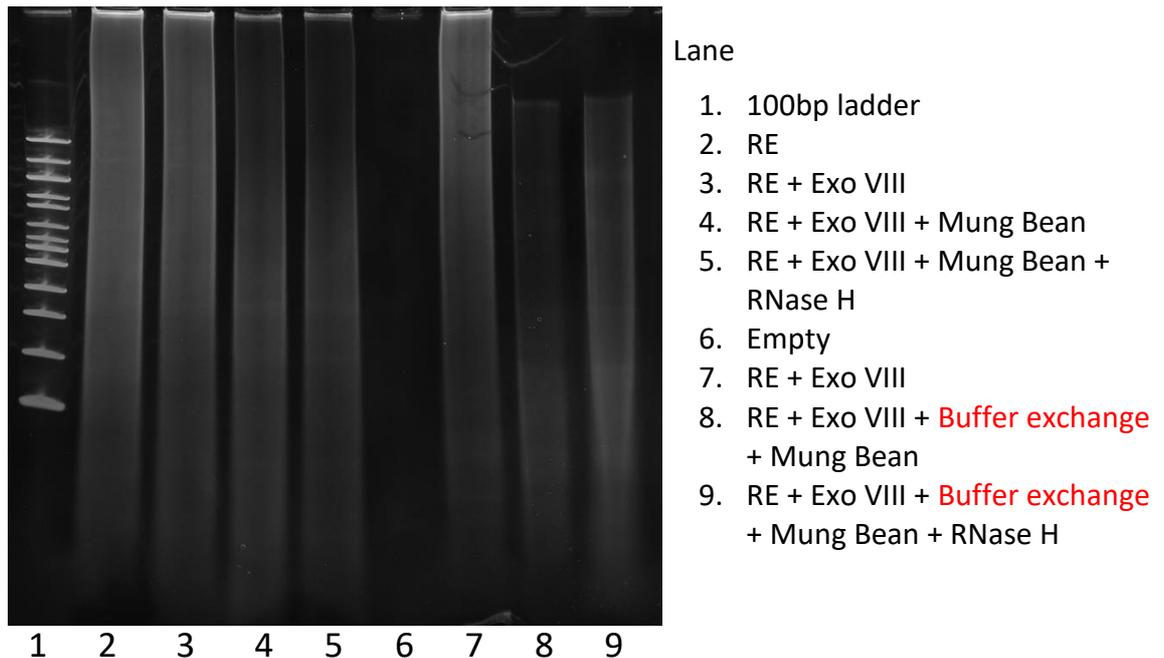
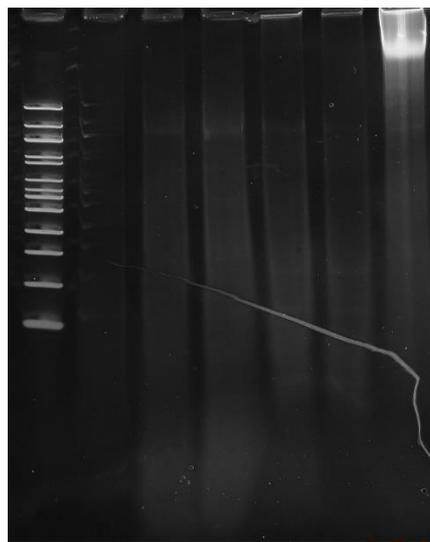


Figure 4.13 The effect of Buffer exchange after restriction enzyme digestion. The nuclease incubation conditions are the same as those in Figure 4.12. 500 ng of DNA was loaded into each lane.

Comparison of lane 8 with lane 4 in **Figure 4.13** shows a noticeable loss of nucleic acids, especially at the top end, where beyond a specific size, the nucleic acid was not able to be recovered. It is difficult to judge whether this step improved the downstream digestion activity of Mung Bean nuclease (and RNase H).

It was previously mentioned that dsDNAse might degrade R-loops in buffers not optimized for it, such as the NEBs CutSmart® buffer (**Section 4.3.2**). By implementing the buffer exchange step into the dsDNAse protocol, it allows the switching between the dsDNAse buffer and the NEB CutSmart® buffer between different steps of the protocol. Therefore, the dsDNAse digestion was performed with the dsDNAse buffer followed by a buffer exchange step and continued the downstream digestion with the NEB CutSmart® buffer (**Figure 4.14**). The Buffer exchange step did not affect the size composition of the sample (comparing lane 4 to lane 3 in **Figure 4.14**).



Lane

1. 100bp ladder
2. No DNA
3. dsDNAse
4. dsDNAse + Buffer exchange
5. dsDNAse + Buffer exchange + Exo VIII + Mung Bean
6. dsDNAse + Buffer exchange + Exo VIII + Mung Bean + RNase H
7. gDNA

1 2 3 4 5 6 7

Figure 4.14 The effect of Buffer exchange after dsDNAse digestion. The dsDNAse digestion was performed in the supplied dsDNAse buffer. The subsequent digestions after buffer exchange were performed in the NEB buffer. dsDNAse and Mung Bean Nuclease digestion were performed at 37°C for 15 minutes while Exonuclease VIII and RNase H digestion were incubated for 2 hours. 500 ng of DNA was loaded into each lane.

4.3.5. Discussion and future work

In this chapter, I have proposed a novel approach in purifying R-loops without the need of any antibodies or affinity reagents to “pull” these hybrids out. Currently, R-loop recognition is mainly dependent on S9.6. It is challenging to assess the accuracy of S9.6 without comparison to suitable alternative methodologies. Furthermore, it is unknown how S9.6 recognizes DNA:RNA hybrids. Multiple studies have emerged in recent years showing that S9.6 affinity is influenced by the DNA sequence and that it also recognizes dsRNAs, questioning the accuracy of DNA:RNA immunoprecipitation methods (Hartono, et al., 2018). As more studies on R-loops emerge, it becomes unavoidable to assess the accuracy of the current method. In this regard, a sample of purified R-loops rather than artificial DNA:RNA hybrids could help in the assessment of the accuracy of S9.6 and potentially help in the discovery of improved antibodies/affinity reagents. This also makes it possible to characterize the R-loop structure in more detail and identify the size of these hybrid structures.

The theoretical framework has been finalized; however, practical implementation requires more optimization. The Exonuclease VIII could be the main bottleneck due to its slow rate of digestion. In order to accelerate this, the DNA was cut into smaller pieces to provide more open ends for the Exonuclease VIII to work on. For this, dsDNAse seems to do a better job compared to the restriction enzyme “cocktail”, as the DNA is cut into smaller and more

uniform pieces (**Figure 4.12** lane 2 compared with **Figure 4.10** left). However, due to the proprietary composition of the optimized dsDNAse buffer, the NEB CutSmart® buffer was used, which might affect the specificity of Thermofisher’s dsDNAse. This problem was partially resolved using a buffer exchange step, in exchange for losing some of the digestion samples. Finally, this protocol requires a way to measure R-loops after digestion reaction to assess the extent the nucleases would resolve R-loops. For this, the S9.6 slot blot method could be used but requires some optimization of the sample to avoid a result seen in **Figure 4.11**. A “DNA immunoblotting” approach could be done after the nucleic acid has been separated by gel electrophoresis. The DNA on the agarose or polyacrylamide gel could be transferred to a membrane (dry-transfer) and then immunostained with S9.6 to identify the natural size (range) of R-loops.

An alternative method for quantifying R-loops could be to digest the DNA portion of the hybrid with, for example, DNase I and measure only the quantity of remaining RNA. One system that can distinguish and quantify specifically RNA is the Qubit system with the Qubit™ RNA HS Assay Kit.

4.4. Effect of various helicases on R-loops formation

Following the optimization of the R-loop slot blot method and the finding that *set-2(bn129)* mutants show a reduction in R-loop levels in section 4.2, I wondered how the absence of a functional SET-2 protein is able to reduce the level of R-loops. Two non-exclusive possibilities could explain the role of SET-2 in R-loop formation: SET-2 could activate or recruit proteins that increase the formation of R-loops or/and SET-2 could deactivate or prevent the binding of proteins that resolve R-loops. With the current knowledge regarding R-loop formation and resolution, investigating the latter possibility is a more viable approach, since the formation of R-loops is still not fully understood, much less the proteins involved in it. Proteins involved in the resolution of R-loops, however, have been identified to be nucleases (i.e. RNase H) and helicases. I focused the investigation on helicases, rather than nucleases as comparatively little is known about the role of helicases in the maintenance of R-loop levels in living organisms. Since helicases are among the largest class of proteins in eukaryotes (Jankowsky & Fairman-Williams, 2010), I used the RNAi by feeding method to screen a large set of helicase candidates for potential helicases that could affect R-loop aggregation dependent or independent of SET-2.

RNAi is a powerful tool for genetic analysis used in *C. elegans*. A comprehensive RNAi library was constructed by the Ahringer group that contains around 87% of all *C. elegans* genes

(Kamath & Ahringer, 2003). The “RNAi by feeding” method is inexpensive and relative fast. It allows for quick knockdown of genes at flexible developmental stages. However, as with many methods, this tool has its limitations. RNAi does not result in a complete knockout of the gene, only a knockdown. It does not deactivate the gene at the genome level but tries to prevent translation of the mRNA. The extent of the knockdown depends on many factors, including the *C. elegans* strain, preparation of the RNAi bacteria and the efficiency of the target sequence design. For example, hypersensitive mutant strains containing mutations in the RNA interference pathway can be more susceptible to RNAi. The preparation of the bacteria will affect the efficiency of the knockdown. For example, mixing two different RNAi bacteria or introducing IPTG to the bacterial culture solution diminishes the efficiency of RNAi (Kamath, et al., 2000). Seeding the bacteria on NGM plates that are too wet will also result in weaker phenotypes (Ahringer, 2005). Finally, the design of the cDNA template to be inserted into the L4440 vector also affects knockdown efficiency. For example, designing a sequence targeting the intronic region of the gene of interest will likely be inefficient (Conte, et al., 2015).

This chapter investigates the importance of various helicases in the maintenance of R-loop levels. First, orthologs of known R-loop resolving helicases were tested to verify their involvement in controlling R-loop levels are conserved in *C. elegans*. For this, the mutants *rha-1(tm329)* (Chakraborty & Grosse, 2011) and *rcq-5(ok660)* (Kanagaraj, et al., 2010) were tested. Next, the sensitivity of COMPASS mutants towards RNAi was measured. Finally, RNAi helicase suppressor screen in *set-2(bn129)* background mutants was carried out to identify helicases that when knocked down suppresses the reduced R-loop level phenotype (i.e. recovers the R-loop signal strength).

4.4.1. *C. elegans* helicase mutants *rha-1(tm329)* and *rcq-5(ok660)* both have increased R-loop accumulation

The two mutant strains *rha-1(tm329)* and *rcq-5(ok660)* are mutants of orthologous helicases known to resolve R-loops in humans. *rha-1* is the *C. elegans* ortholog of the human RHA gene (RNA helicase A), also known as DHX9, which has been shown to be able to unwind R-loops *in vitro* (Chakraborty & Grosse, 2011). *rcq-5* is the ortholog of the human RECQ5 gene which has been shown to reduce R-loop levels in human cell lines, but cannot by itself resolve R-loops *in vitro* (Kanagaraj, et al., 2010).

Both *C. elegans* strains *rha-1(tm329)* and *rcq-5(ok660)* have been generated via UV/TMP (Trimethylpsoralen) mutagenesis, which produces around 1-3kb deletions at a rate of 1

mutation every 1000 genes (Kutscher & Shaham, 2014). The mutants were outcrossed four times with our own wild-type strain to reduce the number of off-site mutations (Zuryn & Jarriault, 2013).

The R-loop level of both mutants was measured using the slot blot method optimized in section 4.2.1. The results show that both mutant strains have higher R-loop levels compared to wild-type (**Figure 4.15a**), suggesting that the R-loop resolving function of these two helicases is conserved in *C. elegans*.

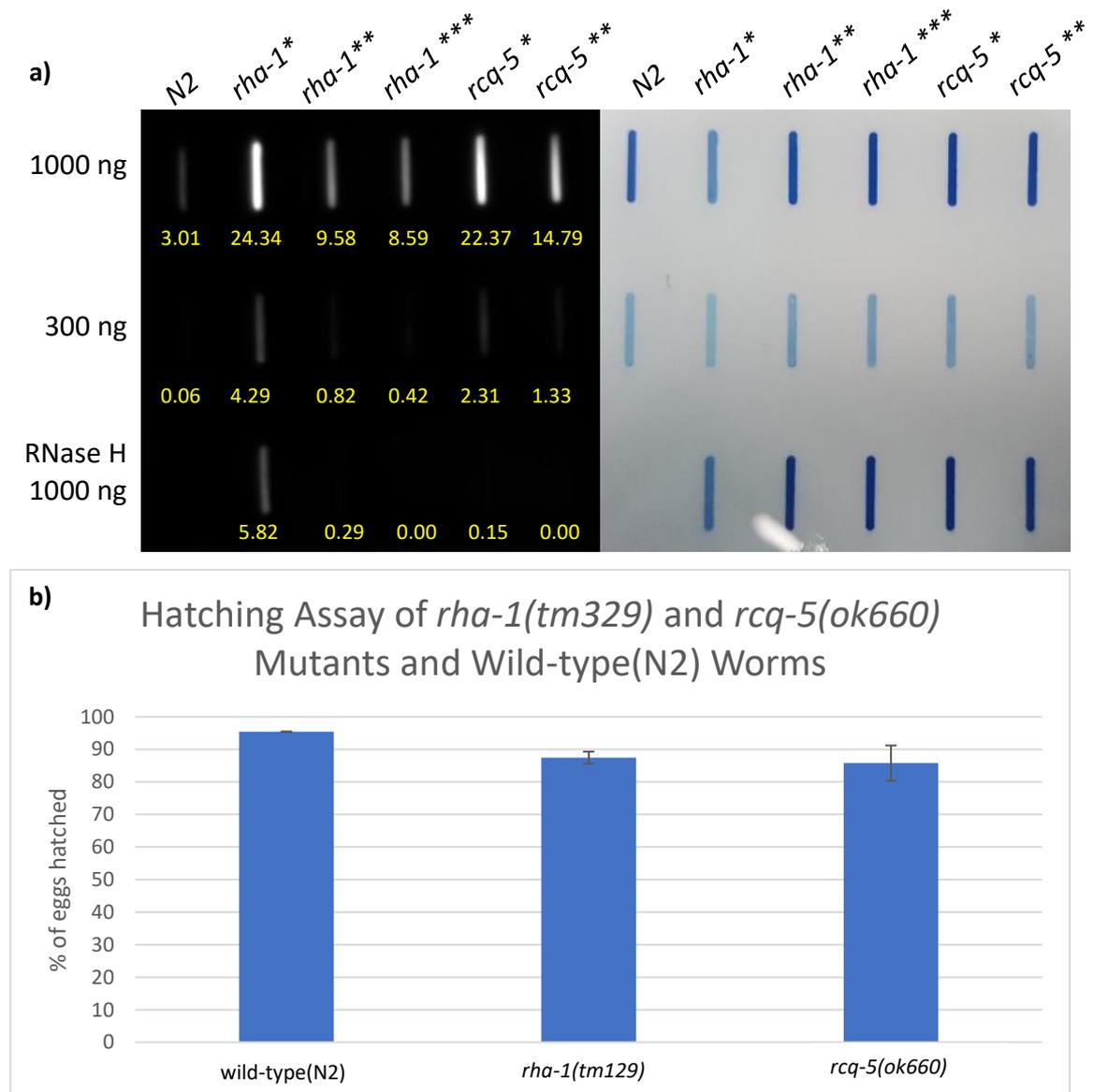


Figure 4.15 R-loop levels of the helicase mutants *rha-1(tm329)* and *rcq-5(ok660)*. a) Left: R-loop dot blot of L1 wild-type(N2) worms, *rha-1(tm329)* and *rcq-5(ok660)* mutants. Right: gDNA loading control using methylene blue. Yellow numbers show the signal quantification based on ImageJ. Stars (*) represents different samples. b) Hatching assay for the samples shown in a) (wild-type(N2), *rha-1(tm329)* and *rcq-5(ok660)* mutants). The error bar indicates Standard Error of the Mean.

4.4.2. *set-2(bn129)* and *cfp-1(tm6369)* are mildly resistant to specific RNAi bacteria

In the previous section, I confirmed that the two *C. elegans* helicases *rba-1* and *rcq-5* have a conserved role in R-loop maintenance. Following on, I aim to utilize RNAi in a suppressor screen to identify helicases that may similarly function in maintaining R-loop levels and revert back the low R-loop levels in *set-2(bn129)* mutant when knocked down. However, before proceeding, the COMPASS mutants need to be analysed whether they have a different sensitivity against RNAi bacteria relative to the wild-type control. This is important since when the mutants are more resistant to RNAi, it becomes difficult to distinguish whether a negative result is due to inefficient knockdown of the target gene (false negative) or actually does not contribute to the phenotype in question (true negative). Therefore I used various RNAi bacteria, that are used to test *C. elegans* strains sensitivity or resistance towards RNAi bacteria (Simonet, et al., 2007; Fischer, et al., 2013).

The RNAi bacteria to test for resistance target the genes *dpy-10*, *dpy-8*, *unc-15*. Both *dpy-10* and *dpy-8* show a strong ‘dumpy’ phenotype, where the worm is smaller and fatter than the wild-type phenotype (**Figure 4.16**). *unc-15* RNAi shows a strong paralysis phenotype (Simonet, et al., 2007). Both of these phenotypes become more apparent as *C. elegans* ages. Therefore, worms are scored after they develop into adults. Genes to be targeted by RNAi bacteria to test for sensitivity are: *dpy-13*, *lin-1*, *unc-73* and *bmr-1*. These show mild or no phenotypes in wild-type worms but are enhanced in RNAi sensitive worms such as *eri-1* (Enhanced RNA interference 1) mutants (Fischer, et al., 2013). *dpy-13* shows a weak dumpy phenotype in wild-type worms. *lin-1* does not show any phenotype in wild-type worms. However, in RNAi sensitive worms, this would manifest as a multi-vulva phenotype (**Figure 4.16**). Similarly, *unc-73* does not show any phenotype in wild-type worms but would result in limited-motility in RNAi sensitive mutants. *bmr-1* manifests itself as an increased number of dead eggs (embryonic lethal) as well as body morphology defects (Wormbase, 2019).

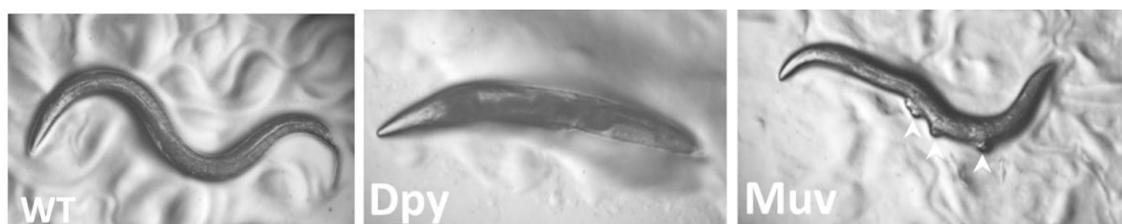


Figure 4.16 Common body morphology phenotypes in *C. elegans* research. Image depicting the phenotype of a wild-type worm (left), dumpy that can be observed in various dumpy RNAi (middle) and multi-vulva in *lin-1* RNAi (right). Image adjusted from Corsi, et al. (2015).

For the sensitivity assay using *dpy-13*, *lin-1* and *unc-73*, no phenotype was observed in both wild-type, and *set-2(bn129)* and *cfp-1(tm6369)* mutant worms. Published *dpy-13* RNAi result found mild dumpy phenotype in wild-type worms (Fischer, et al., 2013), however, here both mutant and wild-type fed on *dpy-13* RNAi did not show any difference compared to animals fed on empty vector (EV) control. The results for *lin-1* and *unc-73*, however, agree with Fischer et al. (2013), who also observed no phenotype when wild-type worms were fed on either bacteria. *hmr-1* RNAi also doesn't show significant phenotypic differences between wild-type and mutant worms (**Figure 4.17**). This suggests that the *set-2(bn129)* and *cfp-1(tm6369)* mutants have a comparable RNAi sensitivity to wild-type animals.

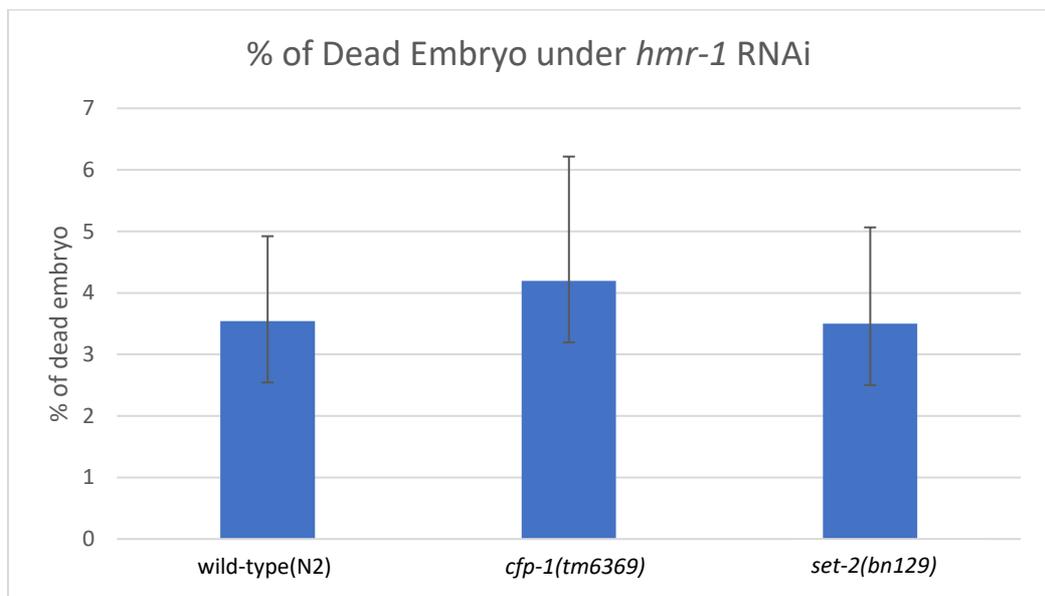


Figure 4.17 *hmr-1* RNAi effect on dead embryos in N2, *set-2(bn129)* and *cfp-1(tm6369)*. Each replicate consists of 3 worms, and the experiment is done with 2 replicates. The experiment has been done at least twice. Error bar represents standard deviation. Results are not significant.

Next, the COMPASS mutants are tested for resistance against RNAi. All three tested RNAi bacteria (*dpy-10*, *dpy-8*, *unc-15*) show phenotypes in both mutants and wild-type worms. However, the dumpy phenotype in *dpy-8* is very mild (i.e. the worms are only marginally shorter than control), making them difficult to distinguish. As such, counting the number of dumpy animals can be unreliable, and the results for *dpy-8* phenotype was not included here. This observation differs from the publicized result, where the *dpy-8* mutation is given a 3 out of 5 scoring for phenotypic strength (Simonet, et al., 2007). *dpy-10* shows a strong dumpy phenotype (30-50% shorter), making the adult animals easily distinguishable from normalized worms. Similarly, *unc-15* also shows strong phenotypes, where the whole body becomes paralyzed, as the worm ages. Even under external influences, such as touching the worm or hitting the Petri plates on the table which would typically stimulate them to move, no body movement would occur. Only the head can be observed to move occasionally, while the

pharynx is still relatively active. Mutant animals are found to be mildly resistant to *dpy-10* and *unc-15* RNAi compared to wild-type worms (**Figure 4.18**). *set-2(bn129)* mutant animals are weakly resistant compared to wild-type for *dpy-10* RNAi: on average the *set-2(bn129)* sample consisted of more normal-sized animals and the body length of the dumpy worms are less compromised compared to wild-type worms (**Figure 4.18a and b**). *cfp-1(tm6369)* mutants, on the other hand, are more resistant against *unc-15* RNAi, as there are fewer paralyzed worms (**Figure 4.18c**).

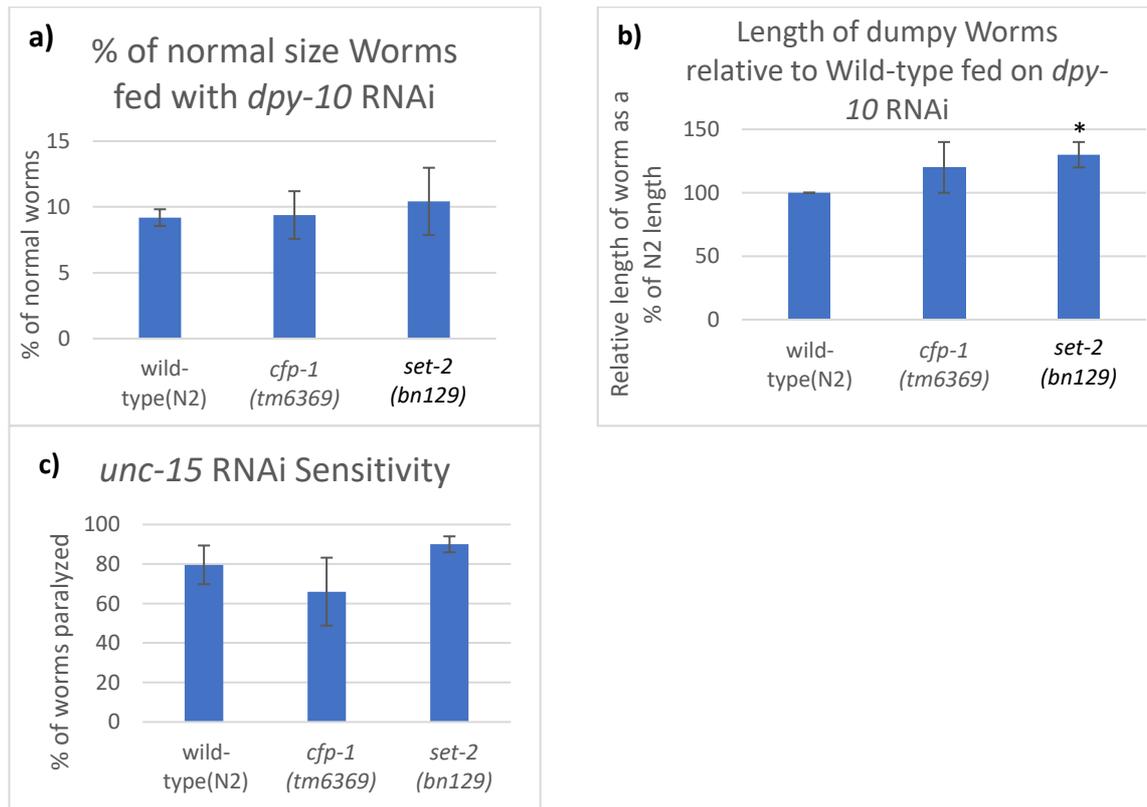


Figure 4.18 RNAi sensitivity of wild-type worms and COMPASS mutants on *dpy-10* and *unc-15* RNAi. a) percentage of worms that show dumpy body morphology when fed on *dpy-10* RNAi. b) severity of dumpy phenotype in affected worms based on body length relative to wild-type control. c) severity of paralysis of worms fed on *unc-15* RNAi. Each replicate consists of 3 mother worms, and the experiment is done with 2 replicates. Error bar represents standard deviation. Single asterisk (*) denotes p-value < 0.1.

In summary, the results show that the COMPASS mutants are not more sensitive to RNAi compared to wild-type worms. However, the individual mutants show mild resistance against specific RNAi bacteria. The results are not statistically significant, but more replicates would be needed, given the small effect size (difference in phenotype strength), to be sure. Since this observed RNAi resistance phenotypes between wild-type and mutant are minimal and only specific to certain RNAi bacteria, the resistance of the mutants towards RNAi is negligible. Therefore, the *set-2(bn219)* mutant is suitable for use in RNAi experiments, like the planned *set-2(bn219)* helicase suppressor screen described in the next section.

4.4.3. *set-2(bn129)* R-loop suppressor screen with helicase genes

All *C. elegans* helicase candidate genes were found using the Wormbase database (Wormbase, 2019), including genes that have a listed helicase protein domain but are not strictly classified as helicases in Wormbase. The 110 “helicases” were then filtered to only include observable non-lethal phenotypes upon RNAi. Furthermore, only the genes for which the Ahringer RNAi library contains an RNAi bacterial clone were considered. Finally, genes that have been observed to affect RNA interference, as well as redundant genes, were excluded. The final number of candidate helicases were 66 genes (see **Appendix 3** for the complete list), out of which 30 were screened. The suppressor screen was only performed in *set-2(bn129)* mutants (refer back to **Section 4.2.4**). The results of the *set-2(bn129)* suppressor screen are summarized in **Table 4.1**. 3 strong suppression (*mog-5*, *isw-1*, *vbb-1*) and 4 partial suppression hits (*rad-54*, *Y116A8C.13b*, *F54E12.2*, *F33H12.6*) were found. These helicases play a role in reducing R-loop accumulation in *C. elegans* in a *set-2* dependent or independent manner.

Strong Suppression	Partial Suppression	No Suppression	Inconclusive	
<i>mog-5</i>	<i>rad-54</i>	<i>mus-81</i>	<i>dog-1</i>	<i>F59H6.5</i>
<i>isw-1</i>	<i>Y116A8C.13b</i>	<i>eri-7</i>	<i>chd-3</i>	<i>him-6</i>
<i>vbb-1</i>	<i>F54E12.2</i>	<i>rcq-5</i>	<i>xpf-1</i>	<i>rha-1</i>
	<i>F33H12.6</i>	<i>Y54E2A4.c</i>	<i>polq-1</i>	<i>ZK250.9</i>
		<i>C46F11.4</i>	<i>ssl-1</i>	<i>ddx-15</i>
		<i>mtr-4</i>	<i>C24H12.4d</i>	<i>T05A12.4</i>
		<i>xpb-1</i>	<i>glh-2</i>	<i>glh-1</i>
			<i>F52B5.3</i>	<i>wrn-1</i>

Table 4.1 R-loop RNAi helicase suppressor screen on *set-2(bn129)*. RNAi suppression is assessed by the strength of R-loop signal in all replicates/duplicates relative to wild-type(N2) and *set-2(bn129)* fed on empty vector control bacteria. Strong suppression candidates are defined as samples whose R-loop signals recover to a level similar to or higher than wild-type(N2). Partial suppression candidates are defined as samples whose R-loop signal is stronger than *set-2(bn129)* (at least ~25% higher) but weaker than wild-type(N2). No suppression is defined as samples who have a comparable or weaker R-loop signal intensity than *set-2(bn129)*. Inconclusive candidates are those that could not be confidently classified. These include candidates whose R-loop signal intensity varies a lot between replicates or are at the borderline between two classifications and difficult to put into either. n ≥ 2

Five of the candidate genes RNAi-mediated knockdown resulted in phenotypes that made an R-loop slot blot not possible, which are for *nath-10*, *B0511.6*, *mcm-5*, *mog-4* and *F57B9.3*. RNAi-mediated knockdown of these five genes resulted in *C. elegans* to become sterile, and in three out of the five cases (*nath-10*, *B0511.6* and *mcm-5*) the adult hermaphrodites have a protruded vulva phenotype. Four of the candidate genes RNAi's (for *mog-5*, *mtr-4*, *isw-1* and *xpb-1*) resulted in adults having very few progenies. For two of the RNAi bacteria (*let-418* and *mcm-2*), I was unable to grow them on RNAi LB culture plates. The hatching assay

control of the seven suppressor hits is shown in **Figure 4.19**. The complete hatching assay for all tested genes can be found in **Appendix 4**.

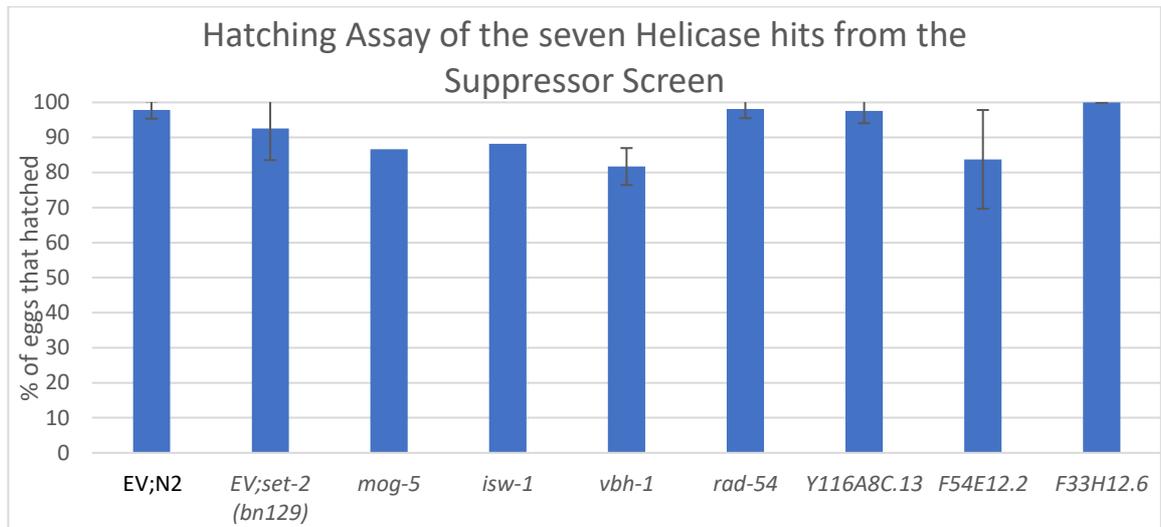


Figure 4.19 Hatching assay of the seven RNAi genes resulting in suppression of the R-loop phenotype. Both strong and partial suppression in *set-2(bn129)* mutants (Table 4.1) are considered. At least two samples have been counted for each RNAi except for *mog-5* and *isw-1* (as they have very few eggs and samples were pooled together). Error bars represent the standard deviation.

Four of the seven hits (strong and partial suppression in **Table 4.1**) are related to chromatin remodelling complexes SWI/SNF (*rad-54*, *Y116A8C.13b* and *F54E12.2*) and NURF (*isw-1*), which is part of the ISWI family (McAndrew & McManus, 2017; Jiang, et al., 2004; Andersen, et al., 2006). Chromatin remodelling is the movement of nucleosomes by sliding or disassembling in order to control the access of DNA for transcription, DNA repair and other activities that require a change in the packaging of the DNA (Lorch, et al., 2010). The ISWI complexes are involved in the equally spaced assembly of nucleosomes following replication, while SWI/SNF complexes alter the nucleosome positioning to promote transcription (Owen-Hughes, et al., 1996; Varga-Weisz, et al., 1997). Although these chromatin remodelling complexes have helicase domains, the helicase function is used as a motor to move the protein forward along the DNA strand, and they lack the ability to separate the two DNA strands (Saha, et al., 2006).

4.4.4. Discussion

4.4.4.1. The R-loop maintenance function of *rha-1* might be dependent on SET-2 (or H3K4me3)

The human ortholog of *rha-1* and *rcq5*, RHA and RECQ5, have been shown to resolve R-loop *in vitro* and *in vivo* respectively (Chakraborty & Grosse, 2011; Kanagaraj, et al., 2010).

Here I have shown that the *C. elegans* mutants *rba-1(tm329)* and *rcq-5(ok660)* have increased R-loop accumulation compared to wild-type worms, indicating that the R-loop resolution function of the orthologue is conserved in *C. elegans*. On the contrary, *rcq-5* RNAi in *set-2(bn129)* mutants did not show increased R-loop level, suggesting that its R-loop maintenance function may be dependent on SET-2 (or H3K4me3). Further experiments with an *rcq-5* and *set-2* double mutant would be required to validate this result. The *rba-1* RNAi experiment on *set-2(bn129)* background is inconclusive and would require more replicates. It must be considered, however, that either or both of these RNAi bacteria could have a low knockdown efficiency in *set-2(bn129)* mutants resulting in false-negative results. One way to circumvent this problem is to use double mutants.

4.4.4.2. COMPASS complex mutants are mildly resistant to specific RNAi bacteria

Both *set-2(bn129)* and *cfp-1(tm6369)* mutants show a minor resistance to *dpy-10* and *unc-15* RNAi respectively. This result indicates that the resistance of the COMPASS mutants towards RNAi is not uniform and depends on the specific RNAi bacteria or gene. However, although the phenotype strength deviates from the wild-type control, the difference is very marginal and mostly not be significant. Care must be taken to evaluate such statistical output, as phenotype strength is a continuous spectrum and cannot be precisely characterized using a binary “yes or no” rating. Statistical hypothesis testing does not return any information regarding the extent of the phenotypic difference. For example, a small effect size (difference in phenotypic strength) with a large enough sample size can have a similar significant p-value as a large effect size with low sample size. Therefore, statistical hypothesis testing cannot be considered alone for judging RNAi sensitivity. Overall, the relative RNAi resistance of *set-2(bn129)* and *cfp-1(tm6369)* mutants is very small, making them suitable for the helicase screen. However, this small resistance should still be taken into account when comparing RNAi phenotypes between COMPASS mutants and wild-type worms.

4.4.4.3. Seven helicases affect R-loop formation

Out of the seven helicases found by the suppressor screen, four are chromatin remodelers belonging to the SWI/SNF (*rad-54*, *Y116A8C.13b* and *F54E12.2*) and NURF (*ism-1*) families (McAndrew & McManus, 2017; Jiang, et al., 2004; Andersen, et al., 2006). They are negative regulators of R-loops and are required to maintain low R-loop levels. This indicates that

chromatin architecture is an important factor that determines the formation or resolution of R-loops, perhaps through the control of transcriptional activity and the chromatin landscape.

rad-54 and Y116A8C.13b

Human *RAD54L* and *RAD54B* (orthologs of *rad-54* and *Y116A8C.13b*, respectively) have both been identified to play a role in the DNA damage response and disrupting protein-DNA interactions (i.e. histone-DNA interaction) to make the chromatin more accessible (McAndrew & McManus, 2017). As chromatin accessibility is directly linked with transcriptional activity, it could imply that *RAD54L* and *RAD54B* enhance R-loop formation via increased transcription and vice versa their deactivation would reduce R-loop formation, which is the opposite of what the RNAi results show. However, they could have other functions that enhance the accumulation of R-loops. For example, through their involvement in the DNA damage response.

In *C. elegans*, *rad-54* (RADiation sensitivity abnormal) is required for strand invasion in HR. Deficiency in RAD-54 protein makes the nematode more sensitive to DSB generated after γ -radiation (Ryu, et al., 2013). Ohle, et al. (2016) suggest that DSB repair by HR includes a step where RNA polymerase II binds to the ssDNA generated by the MRN (Mre11, Rad50 and Nbs1) complex and starts transcription. The resulting RNA transcript then competes with RPA, an essential protein that binds the ssDNA and protects it from degrading during HR, for binding to ssDNA and forms R-loops. Before the strand invasion step can occur, R-loops need to be degraded and replaced by RPA. *rad-54* mutations have been observed to result in inaccurate DSB repair (Lemmens & Tijsterman, 2011), which might be related to R-loops not being replaced by RPA in a *rad-54* dependent mechanism.

Y116A8C.13b is classified as a RAD-54 related protein and animals exposed to *Y116A8C.13b* RNAi show sensitivity to radiation phenotype (Boulton, et al., 2004). Both *Y116A8C.13b* and RAD-54 interact with RAD-51 (Boulton, et al., 2002). In HR, RAD-51 replaces RPA proteins on the ssDNA, then seeks and bind to the homologous DNA. While not much is known about *Y116A8C.13b*, its similarity to *rad-54* (radiation sensitivity phenotype, R-loop accumulation, RAD-51 binding) imply that it works in the same pathway or complex.

The current knowledge of the function of *rad-54* and *Y116A8C.13b* infer that they indirectly increase R-loop formation via increased transcription. However, results from the RNAi suppressor screen shows that they negatively affect R-loops. The slot blot is not able to distinguish between transcriptional R-loop and the HR-dependent R-loops. Thus it might well be that they enhance transcriptional R-loop formation while reducing HR-dependent R-

loop accumulation. The recruitment of specific methyltransferases and demethyltransferase has been suggested to determine whether HR or Non-Homologous End Joining is activated to repair DSBs. However, studies do not agree on how H3K4 methylation influences this choice (Wei, et al., 2018). In this regard, SET-2 might influence which DSB repair pathway is utilized, thereby also determining if *rad-54* and *Y116A8C.13b* are needed to prevent HR-dependent R-loop accumulation.

F54E12.2

F54E12.2 is an ortholog of human transcription termination factor 2 (TTF2) and Helicase Like Transcription Factor (HLTF) (Kim, et al., 2018). TTF2 is part of the SWI2/SNF2 family of proteins and acts to dissociate RNA polymerase II and the nascent RNA from the DNA template. It is also suggested that TTF2 is a negative regulator of transcription by terminating the early elongation complex (Jiang, et al., 2004). Hypothetically, without TTF2, the nascent RNA and RNA polymerase II might stick with the template ssDNA long enough for the RNA to anneal with the ssDNA to form R-loops. This could also prolong the unwound state of the DNA, making it more prone to *trans* R-loops.

HLTF belongs to the SWI/SNF family involved in chromatin remodelling and is implicated in DNA repair. Its function as a transcription factor is involved in many pathways, mainly related to genetic stability (Dhont, et al., 2016). As such, its transcription factor function might be indirectly involved in R-loop maintenance by regulating the transcription of genes that maintain genetic stability, including the resolution of R-loops. As a chromatin remodeler, HLTF uses its translocase activity to facilitate fork regression at DNA lesion sites (a mechanism to circumvent damaged nucleotides and continue replication) (Dhont, et al., 2016). DNA lesions can also be bypassed with a strand invasion mechanism similar to HR. Instead of using a homologous chromosome as a template, HLTF uses the reverse complement daughter strand as a template (Dhont, et al., 2016). Since HLTF plays a crucial role in circumventing DNA lesions during replication, mutations in HLTF could destabilize the DNA strand at the site of the lesion, promoting DNA damage and R-loop formation.

Both HLTF and TTF2 are suggested to be homologs of *F54E12.2*. Due to the very different function of both homologs, it is unclear what role *F54E12.2* has in *C. elegans*. The increased R-loop signal in the RNAi experiment could be a result of failed transcriptional termination in case of a TTF2-like function, or R-loop formed during replication at DNA lesions as a result of reduced HLTF-like function. There has been no research on this gene so far, making it an exciting target to investigate.

isw-1

ISWI, the *Drosophila melanogaster* ortholog of *isw-1*, is the ATPase component of the NURF complex and has been shown to be crucial to remodel the chromatin landscape in a way that promotes the recruitment of the transcriptional machinery (Badenhorst, et al., 2002). Furthermore, NURF activity is dependent on histone tails. The NURF complex can directly associate with H3K4me3 through its PHD finger in the largest subunit (NURF301) and depletion of H3K4me3 results in the partial release of the NURF subunit from chromatin (Wysocka, et al., 2006). Since *set-2(bn129)* mutants are depleted of H3K4me3, the NURF complex might have difficulty maintaining a strong bond to the chromatin and thus unable to open up the chromatin to make it accessible for the transcriptional machinery. This could explain the reduced R-loop levels in *set-2(bn129)* mutants. However, the strong R-loop signal in the *isw-1* RNAi suppressor screen would suggest that ISW-1 has a function, that may be NURF independent and prevents R-loop accumulation.

isw-1 in *C. elegans* has been shown to be active during a wide range of stresses, including mitochondrial and histone stress, and is required to regulate normal lifespan during development and in adulthood (Matilainen, et al., 2017). The importance of *isw-1* in stress regulation might reflect the accumulation of R-loops in *isw-1* RNAi worms, as stress can have a positive effect on R-loop accumulation (Allison & Wang, 2019). The reduction of *isw-1* function weakens the nematode ability to defend against stressors, resulting in the accumulation of various damages, including R-loop accumulation. Outside of stress-induced *isw-1* expression, constitutively active *isw-1* is associated with longevity. Matilainen and co-workers hypothesize that *isw-1* controls longevity through the regulation of the epigenetic landscape and the promotion of protein homeostasis (proteostasis) via the mediation of heat shock proteins (Matilainen, et al., 2017). This suggests that *isw-1* is required to maintain cellular integrity and the knockdown of this gene reduces the ability to regulate the accumulation of toxic agents such as misfolded proteins and R-loops.

mog-5

mog-5 encodes an essential RNA helicase for *C. elegans* that is required for worm and germline development (Wormbase, 2019). Mutation in this gene changes the sex of the worms from hermaphrodite to males, losing the ability to produce oocytes, thus unable to self-fertilize and produce offspring (Puoti & Kimble, 2000) (which could be the reason why worms fed on *mog-5* RNAi bacteria produced very few progenies).

The ortholog of *mog-5* in *Saccharomyces cerevisiae* is PRP22. PRP22 is a component of the spliceosome machinery that removes introns from the pre-mRNA (Puoti & Kimble, 2000). Spliceosome assembly and splicing occur during and alongside transcription (Pandya-Jones, 2011). The assembly of the spliceosome machinery occurs stepwise with the binding of U1, U2, U4/U7 and U5 building blocks. Prp22 acts towards the later stages of the spliceosome in RNA-RNA rearrangement and ribonucleoprotein remodelling events and also proofreads mRNA before releasing it from the spliceosome machinery (Wahl, et al., 2009). It is unknown how PRP22 could affect R-loop formation. Since it functions towards the end of the splicing event, it is unlikely that it acts as a steric hindrance by occupying the nascent RNA. PRP22 mutants accumulate pre-mRNA and intron in the spliceosome (Company, et al., 1991), both of which are single-stranded. The inability to dissociate the spliceosome and intron from the mRNA could allow both the exon and introns to associate with the DNA. Furthermore, it is unknown how this affects splicing events of other introns downstream. Will each splicing event need to be completed before the next one downstream can initiate? If so, then R-loops can be formed downstream, of the stalled splicing event. Finally, there is always the possibility that *mog-5* could have different or additional functions compared to the PRP22 ortholog.

vbb-1

vbb-1 encodes a Vasa and Belle like RNA helicase, that is mainly associated with germline and embryonic development (Wormbase, 2019). Its closest orthologs are the Vasa helicase (DDX4) and Belle Helicase (DDX3) that have been found in many organisms such as humans and *Drosophila melanogaster* (Paz-Gomez, et al., 2014).

Somatic expression (and potentially germline expression) of *vbb-1* is essential for stress survival against heat shock and oxidative stress. Similar to *ism-1*, heat shock proteins are positively regulated by *vbb-1* (Paz-Gomez, et al., 2014). Not much is known about the function and biochemical mechanism of *vbb-1* outside the germline. Apart from its role in stress resistance, *vbb-1* could have other functions yet to be determined that affect R-loop accumulation.

DDX3 has been shown to regulate different steps of RNA metabolism and in various biological processes. These include mRNA export, mRNA splicing and stress response and transcription (He, et al., 2018). Due to the vast array of involvement of DDX3, its effect in R-loop accumulation could be the result of multiple mechanisms. In mRNA export, DDX3 acts in the late stage of cytoplasmic export (Yekdavalli, et al., 2004), indicating that it doesn't act as a steric hindrance for the nascent RNA to anneal to ssDNA during transcription. The

export function could contribute to reducing *trans* R-loop formation through the prevention of mRNA accumulation in the nucleus that could otherwise spontaneously anneal to unwound DNA. DDX3 has been found associated with the exon junction complex (Merz, et al., 2007). The exon junction complex is a complex of proteins that binds to the exon-exon junction post splicing and stays on the mRNA up until the mRNA is being translated by the ribosomes. The attachment of proteins to the mature mRNA increases the stability and could also prevent it from reannealing to unwound DNA, specifically in the case when the mRNA is not transported out of the nucleus immediately. DDX3 is associated with transcription and stress response. It acts as both an enhancer as well as a repressor of specific promoters (Ariumi, 2014) and contributes to the formation of stress granules to halt translation under stress (Oh, et al., 2016). In these regards, the contribution to R-loop formation would depend on the extent to which DDX3 promotes transcription and how important DDX3 is in stress response.

F33H12.6

The *C. elegans* gene *F33H12.6* encodes a PIF-like ATP-dependent DNA helicase (The Uniprot Consortium, 2019) which has not yet been researched. A protein BLAST search found a 29% query cover and 27% identity with the human PIF1 protein (Altschul, et al., 1990). PIF1 has been associated with telomeres and chromosome maintenance during DNA replication. *In vitro* studies elaborated its function in inhibiting telomerase, unwinding and rewinding DNA. The PIF1 homologs in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are required for genome maintenance after stress (George, et al., 2009). Human PIF1 shares significant homology with *E. coli* RecD helicase, which is required to process the DNA:RNA hybrid Okazaki fragment and to rescue stalled replication forks (George, et al., 2009).

Not much is known about PIF1. Its association with genomic maintenance during various stresses could indicate a relationship towards R-loop maintenance. Although RecD in *E. coli* could have a more direct role in R-loop resolution due to its ability to process Okazaki fragments, it is not similar to *F33H12.6* at all, with only a protein BLAST query cover of 5% (Altschul, et al., 1990). Due to the low similarity to well-researched homologs, it is difficult to assess how similar the function is. Ultimately, the function of *F33H12.6* needs to be determined, before further hypothesis about its mechanism on R-loop accumulation prevention can be postulated.

4.5. Conclusion and future work of the epigenetic study

A central question being worked on by the study of epigenetics is how the expression of genes is controlled without changing the DNA sequence. Multiple epigenetic modifications have been described. Two modifications that are directly associated with active transcription are the histone modification H3K4me3 and the DNA:RNA hybrid R-loops, both of which accumulate at the promoter region of a gene. This study aimed to identify how H3K4me3 levels affect R-loop levels and screen for helicases that can control R-loop aggregation.

The *set-2(bn129)* and *cfp-1(tm6369)* mutants that were used throughout the study show an overall drastic reduction of H3K4me3 levels, mainly at constitutively active genes, supporting the observations that SET-2 and CFP-1 are required to maintain H3K4me3 levels and that this histone modification is associated with active genes. Additionally, the *de novo* motif discovery found the T-block motif enriched, which has also been associated with active genes. Furthermore, the SL1 motif was also identified by the motif discovery software, suggesting that SL1 may also be associated with active genes. A small number of developmental genes and chromatin genes have slightly elevated levels of H3K4me3 in the two COMPASS mutants, inferring the existence of another H3K4 trimethyltransferase. The H3K4me3 depleted *set-2(bn129)* mutant had reduced R-loop levels compared to wild-type *C. elegans* worms. This result suggests that H3K4me3 or SET-2 (or both) is required for maintaining wild-type R-loop levels. The helicase suppressor screen identified seven helicases (four of which are related to chromatin remodelling complexes) that rescues the R-loop signal in *set-2(bn129)* mutants, indicating their role in reducing R-loop aggregation. Furthermore, the findings imply a link between chromatin remodeler and R-loop aggregation.

The novel antibody-independent method proposed here offers a unique approach to purify R-loops but requires further optimization for practical application. Purified R-loops from this method, could be used to analyze their structure and test the specificity of S9.6. Furthermore, purified R-loops may act as a better control compared to synthesized DNA:RNA hybrids, as they may have a physically distinct structure that is biologically more accurate.

To improve the reliability of the results presented here, additional experiments could be included. To test the specificity of the S9.6 antibody, a positive control using synthetic DNA:RNA hybrids, as well as, negative controls of dsDNA and dsRNA could be

incorporated. The positive control is not only useful in slot blots as a marker of absolute R-loop quantity, making comparisons between different blots more reliable, but also to test the ability of nucleases to degrade hybrids during the R-loop purification method development. To confirm the effect of COMPASS mutants on R-loop levels, a rescue experiment could be performed. A positive control for the RNAi sensitivity assay by using RNAi sensitive strains could be added to verify that the RNAi bacteria are working as intended and confirm that the COMPASS mutants are indeed not sensitive against RNAi. For the bioinformatics work, a biological replicate of the COMPASS mutant H3K4me3 ChIP-seq would be important to validate the results.

The next step in this research is to test the seven candidate helicase hits from the RNAi suppressor screen on wild-type (N2) worms to distinguish whether the increase (rescue) of the R-loop signal is *set-2* dependent or independent. Helicase candidates that were inconclusive could be repeated to potentially find further hits. Afterwards, R-loop levels in the helicase candidate of interest could be measured in loss-of-function mutants, as RNAi is not guaranteed to completely silence the gene. This could be paired with a rescue experiment as an extra layer of verification. Further research could compare the absolute R-loop signal between OP50 and EV (HT115) food to identify whether the R-loop accumulation between the two standard food diets are similar or not.

Part 2

Bioinformatic analysis to investigate the association between the innate immune response and heat shock response in *C. elegans*

Chapter 5: Introduction of stress response and innate immunity in *C. elegans*

Cellular stress responses are essential biological reactions in response to unfavourable internal and external environmental conditions that disturb cellular homeostasis (cellular steady-state condition) and proteostasis (protein homeostasis) (Welch, 1993). This response is universally preserved as an essential defence mechanism that plays a significant role in the cell's health. The health implications that the cellular stress response is involved with have been extensively studied, which includes proteopathic diseases, such as Alzheimer's and Parkinson's Diseases, as well as, pathogenic infections (Soto & Estrada, 2008; Huang, et al., 2011; Kourtis & Tavernarakis, 2011). The cues that activate the cellular stress response are regarded as stressors, which come in various forms and can be classified according to their properties. Abiotic stressors make up the physical (e.g. heat) and chemical (e.g. reactive oxygen species) stressor. Biological stressors originate from viral, bacterial and eukaryotic pathogens (e.g. fungi). The response against pathogenic microbial invaders, including viruses and bacteria, is known as the immune response (Chaplin, 2010).

Like the various stressors, there are multiple stress responses. Classically the cellular stress responses are classified into cytosolic heat shock response (HSR), the unfolded protein response (UPR) of the endoplasmic reticulum (ER) and mitochondria, the oxidative stress response and the DNA damage response (DDR) (Fulda, et al., 2010). Depending on the stressor, one or multiple stress responses are activated to maintain cellular homeostasis. Indeed, the different stress responses are interconnected and share common elements to solve related macromolecular damages (Kültz, 2005). Proteins that work as part of the stress response are classified as “stress proteins”, which are universally conserved and include molecular chaperones and DNA repair enzymes (Kültz, 2005). Their function is to sense and resolve macromolecular damages but can also include the control of the cell cycle and metabolism (Kültz, 2005).

The immune response, although not classically considered as a cellular stress response, shares many overlapping features and functions with the various cellular stress responses. The HSR, DDR and endoplasmic reticulum UPR (UPR^{ER}) interact and work with the immune response to tackle various stressors (Muralidharan & Mandrekar, 2019). This relationship makes the immune response a relevant inclusion in research regarding cellular stress response.

In the following sections, I provide an overview of various stress responses and their signalling pathways with a focus on the model organism *C. elegans*, specifically emphasizing

the innate immune response and HSR. Furthermore, as part of the immune response, I summarize the effect of various pathogen infections on *C. elegans*.

5.1. Oxidative stress response

The oxidative stress response manages the accumulation of reactive oxygen species (ROS) and free radicals and neutralizes these highly reactive molecules. ROS are oxygen-containing chemicals with high oxidizing strength, such as peroxides and superoxides (Fulda, et al., 2010). Both ROS and free radicals are naturally found in organisms as they are byproducts of various chemical reactions such as metabolism and auto-oxidation of various molecules, such as ascorbic acid and thiols. Owing to their strong chemical reactivity, they can react with and damage all major classes of macromolecules including nucleic acids, proteins and carbohydrates (Fulda, et al., 2010; Rodriguez, et al., 2013). Due to their inevitable production, an organism needs to be able to equilibrate the generation and elimination of these ROS and free radicals. The imbalance between pro-oxidants (ROS and free radicals) and anti-oxidizing agents (e.g. glutathione and superoxide dismutases) in favour of the pro-oxidants triggers the oxidative stress response (Scandalios, 2002; Fulda, et al., 2010).

Over the years, a large variety of pro-oxidants and anti-oxidants have been identified generating a redox proteome, consisting of proteins that undergo oxidation-reduction (redox) reactions. Currently, a central issue is the identification of mechanisms governing the expression of the redox proteome. Transcription factors have been identified to be master regulators of the oxidative stress response, including Nrf2 and NF- κ B (Sies, et al., 2017). Nrf2 and the *C. elegans* homolog SKN-1 have been identified to activate the expression of phase II detoxification genes (mainly defence genes against oxidative stress) required for oxidative stress resistance (An & Blackwell, 2003), such as Cytochrome-P450 and glutathione S-transferases (Ma, 2013). The *C. elegans* ortholog of the human forkhead box protein O (FOXO), DAF-16, which upregulates SKN-1, has also been shown to be activated during oxidative stress and is vital in maintaining normal stress resistance (Senchuk, et al., 2018). The transcription factor PQM-1 initially identified to be responsive to paraquat-induced oxidative stress (Tawe, et al., 1998), has been implied to complement DAF-16 in stress regulation (Tepper, et al., 2013). Heat shock transcription factor 1 (HSF-1), which is the master regulator of the cytosolic heat shock response, has been shown to induce the expression of antioxidants in yeast (Yamamoto, et al., 2007).

Oxidative stress also plays an important role in the innate immune response. The production of ROS is an important defence mechanism that has been proposed to contribute to

pathogen killing by, for example, *Pseudomonas aeruginosa* and *Enterococcus faecalis*. The amount of ROS production requires a careful balance between the necessary function in pathogen defence and destructive effect of oxidative stress (Kim & Ewbank, 2018; King, et al., 2018; Liu, et al., 2019). As such, it is reasonable to suggest that both the innate immune response and oxidative stress response are active during pathogen infection. Indeed, one of the primary innate immune pathways, the p38 MAPK pathway (refer **Section 5.4.2.1**), also regulates the oxidative stress response pathway through the transcription factor SKN-1 (Inoue, et al., 2005), which is required for survival against various pathogens such as *Streptococcus gordonii* (Naji, et al., 2018).

The relationship between the oxidative stress response and the HSR is described in section 5.3.1.1.

5.2. Unfolded protein response

The UPR defends against the accumulation of unfolded or misfolded protein when it exceeds the folding capacity of the ER or mitochondria. This response adjusts the protein folding capacity to maintain proteostasis. The UPR is separated into two types, depending on the location of the response. The endoplasmic reticulum unfolded protein response (UPR^{ER}) and mitochondrial unfolded protein response (UPR^{mt}). While in both cases, the response is activated due to the accumulation of misfolded proteins, the pathway and regulation of each of these responses are different (Pellegrino, et al., 2013).

The UPR^{ER} is triggered by three signalling branches that sense unfolded protein accumulation within the ER lumen: the inositol-requiring protein 1 α (IRE1 α), the Protein kinase RNA-like Endoplasmic Reticulum Kinase (PERK) pathway and the Activating Transcription Factor 6 (ATF6) (Ron & Walter, 2007) (See **Figure 5.1**). The IRE1 α pathway splices and activates the transcription factor (TF) X box-binding protein 1 (XBP-1), which transcribes genes related to ER proteostasis maintenance. PERK activates the eukaryotic translation initiation factor 2 α (eIF2 α) through phosphorylation, thereby contributing to activation of UPR^{ER} proteins. ATF6 exists as an inactive form tethered to the ER membrane and moves to the Golgi apparatus upon sensing ER stress where it is cleaved before moving into the nucleus. It then binds to DNA and activates gene expression of UPR target genes (Ron & Walter, 2007).

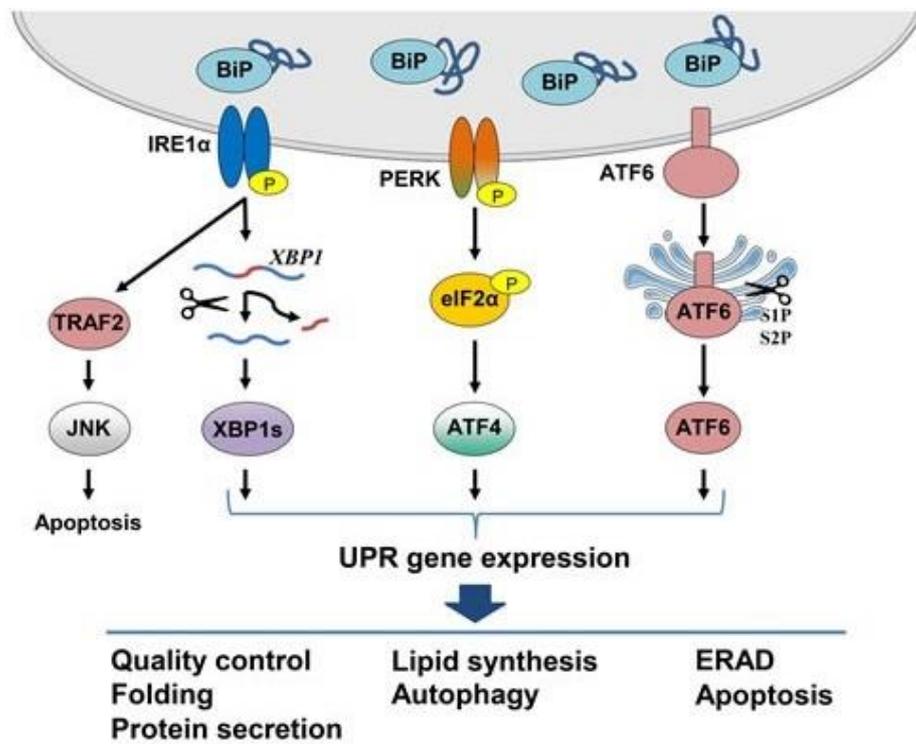


Figure 5.1 Signalling pathway of all three branches of the UPR^{ER}. Binding immunoglobulin protein (BiP) dissociates from the sensory proteins IRE1 α , PERK and ATF6 and binds to misfolded proteins upon activation of the UPR. This dissociation activates IRE1 α and PERK oligomerization and their phosphorylation. IRE1 signalling uses the TF XBP-1, while PERK utilizes ATF4 to transcribe stress response genes. ATF6, on the other hand, acts as an information carrier, and itself moves away from the ER into the nucleus to act as a TF. Image taken from Storm, et al. (2016).

The UPR^{mt} is initiated through the sensing of unfolded protein by the quality control protease ClpP (Pellegrino, et al., 2013). This protease degrades the proteins into peptides, which activates the Activating Transcription Factor associated with Stress (ATFS-1). ATFS-1 and another transcription factor downstream of ClpP, DVE-1, are both required for the up-regulation of mitochondrial molecular chaperone genes, including HSP-60 and HSP-10 (Pellegrino, et al., 2013).

The UPR has been associated with the innate immune response. This is because misfolded proteins are increased as a side effect of pathogen infection, as the cell responds by increasing the production of innate immune peptides (Ermolaeva & Schumacher, 2014). Like the oxidative stress, the UPR will need to be active to counteract the detrimental effects of misfolded protein aggregation. XBP-1 is a key TF of the UPR^{ER}, and *xbp-1* loss-of-function *C. elegans* mutants show reduced survivability against *P. aeruginosa* (Richardson, et al., 2010). The UPR^{mt} also helps in maintaining proteostasis following the innate immune response and improves survival against pathogens. The bZIP TF ATFS-1, which mediates the UPR^{mt}, is essential for resistance against *P. aeruginosa* (Pellegrino, et al., 2014). At the same time, the knockdown of the bZIP TF ZIP-3, which is a negative regulator of the UPR^{mt}, confers resistance against *P. aeruginosa* infection (Deng, et al., 2019).

The crosstalk between the unfolded protein response and the HSR is described in section 5.3.1.3.

5.3. Heat shock response

The HSR is one of the most ancient transcriptional programs in eukaryotes that is highly conserved from yeast to plants and mammals. This response is activated upon heat stress that facilitates the expression of heat shock proteins (HSPs), which are molecular chaperones (Mathew & Morimoto, 2006). Molecular chaperones target damaged or misfolded proteins as a direct consequence of increased temperature and refold them or target them to cellular degradation (Mathew & Morimoto, 2006). Different to what the name suggests, the HSR is not limited to heat stress but is involved in a large variety of cellular stresses that would lead to protein misfolding in the cytosol (Verghese, et al., 2012; Brunquell, et al., 2016).

The heat shock factor 1 (HSF-1) is widely conserved in eukaryotes and plays a central role in the HSR (Vihervaara & Sistonen, 2014). While this protein is called heat shock factor, owing to its initial identification, it also functions beyond the HSR. HSF-1 is crucial in other stress response pathways and is involved in development, metabolism, gametogenesis and ageing (Vihervaara & Sistonen, 2014). It has also been found essential in cancer cells (Mendillo, et al., 2012). In the inactive form, the HSF-1 monomer is bound by certain HSPs, including HSP-70 and HSP-90, in the cytoplasm. During heat stress, HSF-1 dissociates from the HSPs, becomes trimerized, and moves into the nucleus where it binds to specific DNA sequences known as heat-shock elements (HSE) to transcribe heat shock genes including HSPs and to minimize misfolded and damaged proteins (**Figure 5.2**) (Brunquell, et al., 2016; Anckar & Sistonen, 2007; O'Brien & van Oosten-Hawle, 2016; Prahlad & Morimoto, 2009). The HSE consists of three adjacent and inverted nGAAn pentamers, to accommodate the three HSF-1 proteins that make up the trimer. Heat shock genes have clusters of HSEs, while developmental genes typically have only one HSE (Li, et al., 2016). Certain HSPs inhibit HSF-1 activity, thereby creating an autoregulatory cycle that adjusts the intensity of response according to the extent of the stress (Shi, et al., 1998; Zou, et al., 1998). In addition, HSF-1 activity is also controlled by various post-translational modifications such as phosphorylation, acetylation and sumoylation. Apart from phosphorylation at S230 and S326 that enhance the activity of HSF-1, all other known modifications negatively affect its activity (Anckar & Sistonen, 2007; Vihervaara & Sistonen, 2014).

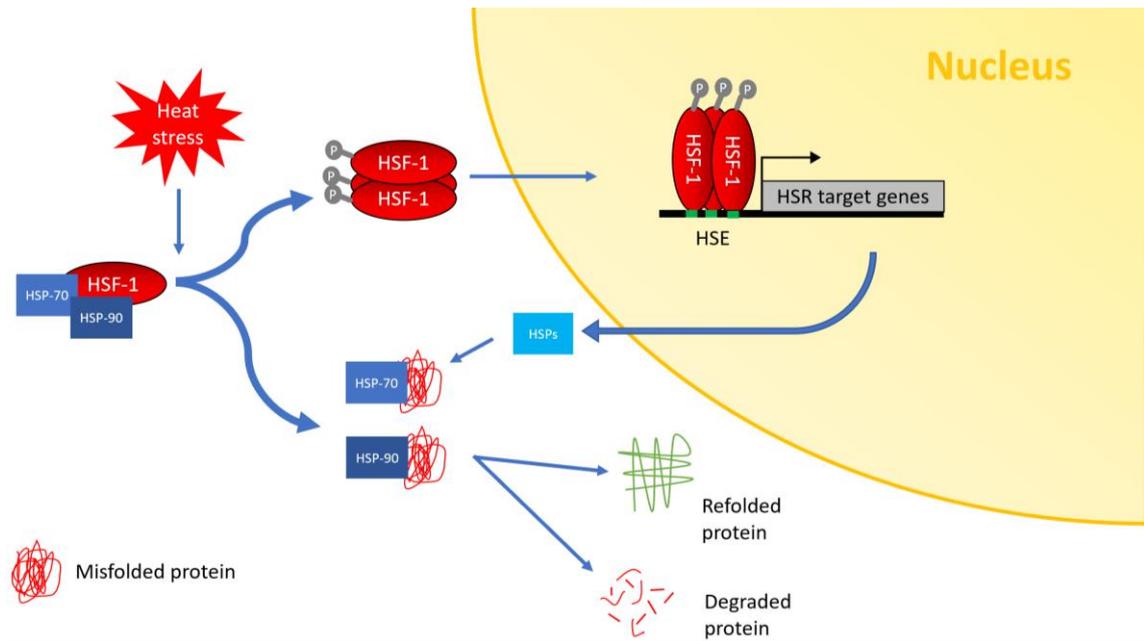


Figure 5.2 Heat shock response pathway. Monomeric HSF-1 is inactive and bound by specific HSPs, including HSP-70 and HSP-90. Upon heat stress (or other signals), HSF-1 dissociates from the HSPs, trimerizes and becomes phosphorylated at specific residues. The HSF-1 trimer moves into the nucleus where it binds to HSE and transcribes HSR target genes. Dissociated and newly produced HSPs refold misfolded protein or target these for degradation.

HSF-1 is also regulated by other pathways such as the insulin-like signalling (ILS) pathway that is involved in lifespan extension and the cyclic guanosine monophosphate (cGMP) signalling associated with development (Barna, et al., 2012).

In *C. elegans*, the two thermosensory AFD neurons that sense temperature changes can also stimulate HSF-1 activity cell-non-autonomously by a guanylate cyclase 8 (*gcy-8*) and serotonin dependent mechanism (Prahlad, et al., 2008; Tatum, et al., 2015). Further experiments using *C. elegans* showed that overexpression of HSF-1 increases DAF-16-dependent longevity and stress resistance (Hsu, et al., 2003; Vihervaara & Sistonen, 2014; Brunquell, et al., 2016; Kumsta, et al., 2017).

5.3.1. Crosstalk between the HSR and other stress response pathways

5.3.1.1. Oxidative stress response and the HSR

The response to oxidative stress results in the upregulation of a large set of genes in *C. elegans*, which includes six heat shock proteins, such as the small heat shock proteins *hsp-16.1* and *hsp-16.2* that are also controlled by DAF-16 (Park, et al., 2009). In human cell lines, ROS induces the expression of *hsp72* and *hsp27*, while antioxidants reduce their expression

(Gorman, et al., 1999). The human NRF2 transcription factor, which regulates the defence against oxidative stress (Ma, 2013), is found to activate HSF1 under oxidative stress (Paul, et al., 2018) but also shares many target genes with HSF1 (Naidu, et al., 2015), indicating that the HSR and oxidative stress responses are connected. However, SKN-1 was shown not to overlap with the heat shock response pathway, as knockdown of *skn-1* via RNAi did not affect the expression of oxidative stress-induced heat-shock proteins (Park, et al., 2009), pointing out distinctive features exclusive to the oxidative stress response.

5.3.1.2. Insulin-like signalling pathway and the HSR

The ILS pathway shares some key regulators with the heat shock response. In *C. elegans*, the ILS pathway is associated with many fundamental properties, including longevity and stress resistance (Murphy & Hu, 2013). FOXO/ DAF-16, a TF negatively regulated by ILS, is required for longevity and the proper upregulation of HSPs during the HSR (Hsu, et al., 2003). Incidentally, like HSF-1, DAF-16 has been observed to move into the nucleus during heat shock. The upstream insulin-like receptor DAF-2 of the ILS pathway controls the phosphorylation state of both DAF-16 and HSF-1. Under normal conditions, hyperphosphorylation keeps the two transcription factors in the cytoplasm. Upon stress, this pathway facilitates the dephosphorylation and allows them to enter the nucleus (Rodriguez, et al., 2013). As previously mentioned, HSF-1 plays a role in the ILS-dependent lifespan extension. The lifespan extension observed in *daf-2* mutants is dependent on both DAF-16 and HSF-1 (Hsu, et al., 2003). Hsu et al, (2003) hypothesized that this lifespan extension is the result of increased expression of small HSPs, as the mutation in *daf-2* allows for higher DAF-16 and HSF-1 activity (Hsu, et al., 2003).

5.3.1.3. Unfolded protein response of the endoplasmic reticulum and the HSR

The UPR^{ER} is a response against misfolded proteins in the endoplasmic reticulum (see **section 5.1** for more detail). While the HSR and UPR^{ER} might work in different locations of the cell, they have some shared characteristics. Overexpression of Hsf1 in *Saccharomyces cerevisiae* can relieve ER stress of UPR-deficient *ire1* mutants (including growth and protein transport defects) (Liu & Chang, 2008). Furthermore, the ER specific oxidoreductin ERO1, required for the formation of disulphide bonds in protein folding, is activated by Hsf1 under heat, ethanol and oxidative stress (Takemori, et al., 2006). In humans, Hsp72 and Hsp90

have been shown to increase IRE1 activity by binding to its cytosolic domain to promote the UPR^{ER} (Gupta, et al., 2010).

5.3.1.4. Innate immune response and the HSR

Multiple pieces of evidence show the involvement of heat shock proteins in the innate immune response in vertebrates and invertebrates (Barna, et al., 2018). For example, hyperthermia (also known as fever) in mammals is observed during pathogen infection and increases HSF1 activity, which in turn negatively regulates the production of cytokines (Barna, et al., 2018). Furthermore, the heat-induced ion channel Transient Receptor Potential Vanilloid 1 (TRPV1), responsible for controlling body temperature, is itself regulated by HSF1 (Barna, et al., 2018). In humans, Hsp60 has been shown to activate the MAPK pathway of the innate immune response through the Toll-like receptor TLR2 and TLR4 (Vabulas, et al., 2001). Experiments using *C. elegans* have identified HSF-1 to be essential for normal resistance against various pathogens (see **Section 5.4.2.5**) and have found that *hsp-90* knockdown by RNAi, which activates the HSF-1 mediated HSR, also induces the expression of specific innate immune response genes (Eckl, et al., 2017).

5.3.2. Transcellular chaperone signalling

Transcellular signalling is a necessary function whereby the different cell-types and tissues within a multicellular organism can communicate with one another to regulate the expression of genes at an organismal level. The transcellular chaperone signalling (TCS) is a more specific case to describe the upregulation of molecular chaperones of the heat shock protein family through cross-tissue communication in a neuronal as well as non-neuronal cell non-autonomous manner (van Oosten-Hawle & Morimoto, 2014; O'Brien & van Oosten-Hawle, 2016). In *C. elegans*, *hsp-90* and *hsp-70* have been observed to be regulated via TCS cell non-autonomously. It has been shown that tissue-specific knockdown or overexpression of *hsp-90* affects the expression of molecular chaperones in distal tissues (the receiver tissue) in a FOXA/PHA-4 (transcription factor involved many processes) dependent manner (van Oosten-Hawle, et al., 2013). The expression of *hsp-90* and *hsp-70* is carefully balanced so that expressional changes in one will be compensated by the opposite change in the other. For example, knockdown of *hsp-90* in neurons or intestine causes upregulation of *hsp-70* in distal tissues (van Oosten-Hawle, et al., 2013). A recent discovery by the van Oosten-Hawle group identified PQM-1 as a mediator of TCS of *hsp-90* via the innate immunity-associated

transmembrane protein CLEC-41, reinforcing the relationship between HSR and the innate immune response (O'Brien, et al., 2018).

5.4. Innate immunity in *C. elegans*

The innate immune response is highly conserved in all animals and plants, while the adaptive immune system is specific to vertebrates (Reece, et al., 2011). Unlike the adaptive immune response that has been designed to recognize previously encountered pathogen, the innate immune response depends on recognizing conserved features of pathogens. Therefore, it is the first line of defence against infections from novel pathogen and the only one in non-vertebrates (Alberts, et al., 2002; Reece, et al., 2011).

The innate immune response involves the recognition of infection or/and pathogen via a wide array of receptors, followed by a change in gene expression through the activation/deactivation of transcription factors and finally the production of antimicrobial peptides and proteins to serve as the defence response against the pathogenic infection. Furthermore, cells can also secrete signalling peptides (cytokines in vertebrates) that coordinates the immune response across different cells and tissues, so that the production of the proteins/peptides are at the correct location (Kim & Ewbank, 2018).

The bacterivore *C. elegans* is affected by a wide range of pathogens that activates the nematodes innate immune response. Even before a pathogen infects *C. elegans*, it has various adaptations such as avoidance behaviour that actively reduce exposure to the pathogens (Anderson & McMullan, 2018). Unavoidable pathogens often aggregate in “hot spots”, which are the intestine/rectum and cuticle/epidermis (Kim & Ewbank, 2018). Since *C. elegans* main diet is bacteria, they are susceptible to bacterial infection that can establish inside the intestine. However, this requires the bacteria to survive through the pharyngeal grinder, which is very effective in breaking up microbial cells (Gravato-Nobre & Hodgkin, 2005) or penetrate through the cuticle, which acts as a physical barrier that separates the *C. elegans* tissues/organs from the pathogen filled external environment. When pathogens are detected or overcome these barriers and successfully infect the nematode, they trigger the innate immune response.

5.4.1. Physical defence against pathogens

5.4.1.1. *Physical avoidance behaviour of pathogenic bacteria*

The avoidance behaviour and innate immune response have overlapping pathways, which suggests that the immune response can, to some extent, influence the bacterial avoidance behaviour. The Toll-Interleukin-1 Receptor (TIR-1), Neural Symmetry (NSY-1) and SAPK/ERK kinase (SEK-1) signalling cascade are part of the p38 MAPK pathway (more detail in **Section 5.4.2.1**). This pathway is a major part of the innate immune response and also regulates serotonin-dependent avoidance behaviour against *P. aeruginosa* (Shivers, et al., 2009). Similarly, the *tol-1* gene associated with the innate immune response against *Salmonella enterica* (Tenor & Aballay, 2008) is required for the avoidance behaviour of *Serratia marcescens* (Pujol, et al., 2001).

It is unknown how *C. elegans* distinguishes between harmless and harmful bacteria, but the chemosensory ability has been shown to play a role in identifying chemicals released by bacteria. For example, the presence of Serrawettin W2 secreted by *S. marcescens* is sensed by the AWB neuron that elicits an avoidance behaviour (Pradel, et al., 2007). Dodecanoic acid from *Streptomyces* is perceived through the chemosensory neurons ASH, ADL, ADF or AWA (Tran, et al., 2017). The Cry6A toxin produced by *Bacillus thuringiensis* evokes an aversion behaviour dependent on neuropeptides and the ILS pathway (Luo, et al., 2013).

5.4.1.2. *Cuticle and Pharynx: primary protection against pathogen infection*

The cuticle and pharynx play a vital role in the defence against pathogens. The cuticle of *C. elegans*, analogous to the external layer of the skin of vertebrates, is an exoskeleton secreted by the epidermis, which is the interface between the nematode's organs and its environment. The cuticle is composed of multiple layers of tough collagen that act as a physical barrier against pathogen invasion (Taffoni & Pujol, 2015). It is covered by a surface coat of negatively charged glycoproteins that prevent the adhesion of bacteria and fungi (Blaxter & Bird, 1997; Page & Johnstone, 2007). However, some pathogens are able to adhere to the surface, mainly targeting regions with natural openings (e.g. mouth/vulva and anal opening), such as the fungus *Drechmeria coniospora* and the bacteria *Yersinia pestis* and *Microbacterium nematophilum*, (Gravato-Nobre & Hodgkin, 2005; Page & Johnstone, 2007). Mutations of the cuticle can affect the efficiency of pathogen adhesion. The mutations of the Bacterially Un-Swollen (*bus*) genes *bus-2*, *bus-4*, *bus-12* and *bus-17* reduce the attachment of *M. nematophilum*

and *Yersinia pseudotuberculosis* but increase the attachment and susceptibility to *D. coniospora* (Gravato-Nobre, et al., 2011; Höflich, et al., 2004; Drace, et al., 2009; Rouger, et al., 2014). Reduced attachment for the *bus* mutants is due to the reduced recognition of surface moieties, while increased susceptibility to *D. coniospora* is explained by increased attachment efficiency (Kim & Ewbank, 2018). Differential expression analysis of infected *C. elegans* often identifies the enrichment of genes associated with the cuticle (Yang, et al., 2015). Likewise, physical damage to the cuticle can lead to the expression of innate immune genes such as the antimicrobial peptide *nlp-29* (Taffoni & Pujol, 2015).

The pharynx has similar defensive importance as the cuticle but acts more like a gate to control outside sources to be transported into the nematode. Therefore, it is essential for the pharyngeal grinder to break down potentially harmful bacteria and neutralize them before allowing these to arrive at the intestine. While the grinding mechanism is highly effective, some bacteria can still survive this process and start to form colonies in the gut (Kim & Ewbank, 2018). This becomes more apparent as the worm ages and the pharynx efficiency decreases, due to structural changes that make the pharynx more swollen and disorganized (Wolkow, et al., 2017). Mutations that lead to defects in the pharynx increase the nematodes susceptibility to pathogens, as more pathogens are able to survive and accumulate in the pharynx and intestine. Mutants of the Pharyngeal Muscle 2 (*phm-2*) gene have a defective pharyngeal grinder, making them less resistant against *P. aeruginosa* and *S. enterica* (Gravato-Nobre & Hodgkin, 2005). The *hpx-22* gene is a peroxidase associated with the production of cuticle material in the hypodermis and pharynx. Mutation in this gene results in a more penetrable cuticle of the epidermis and the pharyngeal lumen, as well as a higher susceptibility to some pathogens such as *E. faecalis* (Liu, et al., 2019).

5.4.2. Innate immune response

When the pathogen successfully infects *C. elegans*, the nematode's innate immune response is activated. This response counteracts pathogenic infections through a wide array of signalling pathways. The known innate immune signalling pathways are the MAP kinase pathways, the Insulin-like receptor signalling pathway, the Transforming Growth Factor β -like pathway (TGF- β) and the Toll-like receptor pathway. The MAP kinase pathway is further divided into three sub-pathways: p38 MAP kinase, ERK MAP kinase and C-Jun amino-terminal kinase pathways (Gravato-Nobre & Hodgkin, 2005). The pathways are pathogen-specific, and multiple pathways can work together against common infections.

5.4.2.1. Mitogen-activated Protein Kinases (MAPK) pathway

p38 MAPK pathway

The p38 MAP kinase pathway is a central player in the innate immune response and is important against many pathogens. The p38 MAP kinase 1 (PMK-1) (ortholog of human MAPK12, MAPK13 and MAPK14) is the central kinase of this pathway and is essential for resistance against various pathogens, such as *Candida spp.* (Pukkila-Worley, et al., 2011; Souza, et al., 2018), *Y. pestis* (Bolz, et al., 2010), *D. coniospora* (Pujol, et al., 2008a), *P. aeruginosa* (Cheesman, et al., 2016), *E. faecium*, *E. faecalis* (Yuen & Ausubel, 2018), *S. enterica* (Tenor & Aballay, 2008), *Proteus spp.* (JebaMercy, et al., 2013), *M. marinum* (Galbadage, et al., 2016) and *Coxiella burnetii* (Battisti, et al., 2017). This pathway is ineffective against the intracellular pathogen *Nematocida parisii* (Bakowski, et al., 2014). The activation of the TF ATF-7 by PMK-1 is one aspect that gives resistance against *P. aeruginosa* and *S. marcescens* but does not affect resistance against *E. faecalis* (Shivers, et al., 2010; Fletcher, et al., 2019), indicating that PMK-1 confers pathogenic resistance through multiple mechanisms.

The upstream proteins in this pathway are the receptor TIR-1, which activates the kinase NSY-1, that in turn phosphorylates SEK-1, which then phosphorylates PMK-1 (**Figure 5.3**). Mutations of any of the upstream kinases show similar resistance changes to the pathogens as loss of PMK-1 (Shivers, et al., 2009; Shivers, et al., 2010; JebaMercy, et al., 2013; Cheesman, et al., 2016; Pujol, et al., 2008b). PMK-1 is also involved in the response to oxidative stress via the phosphorylation of the TF SKN-1, which then accumulates in the intestinal nuclei and transcribes phase II detoxification enzymes such as *gcs-1* (Inoue, et al., 2005). Since SKN-1 requires PMK-1 kinase activity, its effect also depends on the upstream SEK-1, NSY-1 and TIR-1 proteins (van der Hoeven, et al., 2011). PMK-1 also mediates resistance to osmotic stress and depends on the same three upstream proteins (Solomon, et al., 2004).

Hyperactivation of the p38 MAPK pathway is like a double-edged sword. The *nsy-1* gain-of-function mutant, although conferring higher resistance against *P. aeruginosa* infection, also results in developmental delay (Cheesman, et al., 2016), which suggests that there is a trade-off between enhancing defence function and maintaining developmental functions. Interestingly, the *nsy-1* loss-of-function mutant did not reduce the survivability against *P. aeruginosa* significantly (Cheesman, et al., 2016).

PMK-1 is required for normal life span (Pujol, et al., 2008a), and partially required (same as SKN-1) for the extended life span of *daf-2* mutants, indicating that the p38 MAPK pathway plays a role during reduced insulin signalling. This is further supported by the observation

that PMK-1 and DAF-16 essentially do not upregulate the same group of genes (Troemel, et al., 2006). SKN-1, which is downstream of PMK-1, also affects lifespan (Tang & Choe, 2015). Since SKN-1 and PMK-1 are involved in both longevity and immune response, it is not surprising that these two proteins are also associated with immunosenescence (gradual deterioration of the immune response due to age). Mutants of either of the two genes show an earlier decline in immune response gene expression or higher susceptibility to pathogen compared to wild-type animals with age (Youngman, et al., 2011; Papp, et al., 2012), and declining PMK-1 activity with age (Pukkila-Worley & Ausubel, 2012). However, it is arguable, how much of this observed decline in immunosenescence is due to the naturally shorter lifespan of the mutants in the first place.

Finally, PMK-1 has temperature-dependent activities, whereby it localizes to the nucleus beyond 33°C and helps in the expression of constitutive Hsp70 (*hsp-1*) but not heat-inducible *hsp-70* and *hsp-16.2* chaperones. It is hypothesized that PMK-1 could activate HSF-1 by phosphorylating its serine residue and therefore play an essential part in the HSR (Mertenskötter, et al., 2013). Furthermore, PMK-1 expression can also be cell non-autonomous. While intestinal PMK-1 is expressed cell-autonomously as a response to intestinal infection, other tissues can also regulate the p38 MAPK intestinal response through the nervous system (Bolz, et al., 2010; Cao & Aballay, 2016). This mechanism may be beneficial for activating the intestinal innate immune response before an infection can be established or enhance the immune response due to infection in other parts of the worm.

Extracellular-signal-regulated Kinase (ERK) MAPK pathway

Another well-studied MAPK pathway is the ERK MAPK pathway, with MPK-1 being the central kinase. The signalling cascade consists of LET-60, LIN-45, MEK-2 and MPK-1 (**Figure 5.3**). Mutation in any of the components of this signalling cascade results in decreased survivability against various pathogens. The severity of the anal bus phenotype and constipation as a result of *M. nematophilum* infection is increased following the RNAi-mediated knockdown of *lin-45*, *mek-2* and *mpk-1* (Nicholas & Hodgkin, 2004). Similarly, the extent of anal swelling observed during *S. aureus* infection is dependent on *mpk-1*. The transcriptional response in the intestine, however, is not affected by *mpk-1* mutation (Irazoqui, et al., 2010). This pathway also controls a specific type of autophagy, that is an important defence mechanism against *P. aeruginosa* infection by neutralizing pathogen-imposed necrosis (Zou, et al., 2014).

The ERK MAPK pathway also affects longevity similar to PMK-1, and mutation in any of the signalling cascade kinases results in a reduced life span of the worm (Okuyama, et al.,

2010). SKN-1 is a vital component of this longevity phenotype and is phosphorylated by MPK-1, as well as, mediated through the ILS pathway. DAF-16 can increase expression of *skn-1*, and conversely, DAF-2 negatively regulates *skn-1* (Okuyama, et al., 2010; Tullet, et al., 2017). *skn-1* mutants have a shorter lifespan and are hypersensitive to oxidative stress. Overexpression of DAF-16 can rescue the short-lived *skn-1* mutants. However, it cannot rescue the hypersensitivity to oxidative stress. This indicates that SKN-1 promotes life span and oxidative stress resistance via different mechanisms/pathways (Tullet, et al., 2017).

MPK-1 is also associated with other stress responses such as UV induced developmental arrest (Bianco & Schumacher, 2018). While knockdown of *mpk-1* on its own does not affect UV resistance; in combination with *csb-1* mutants, this enhances the UV resistance, while overactivation of the ERK MAPK pathway results in the opposite effect. Interestingly, this effect is dependent on DAF-16 (Bianco & Schumacher, 2018).

The ERK MAPK pathway can also be induced cell non-autonomously. For example, the downstream lysozyme *ihys-3* is required in the pharynx to enable pharyngeal grinder function and for pathogen defence in the intestine. The pharyngeal *mpk-1* expression is required to induce expression of *ihys-3* in the intestine, while intestinal *mpk-1* expression does not affect *ihys-3* expression (Gravato-Nobre, et al., 2016).

C-Jun Amino-terminal Kinase (JNK) MAPK pathway

The JNK pathway is perhaps the least researched signalling cascades among the three MAPK pathways with regards to innate immune response. The mitogen activated protein kinase analogous to MPK-1 and PMK-1 in the other two MAPK pathways is KGB-1. Similar to the other MAPKs, KGB-1 phosphorylates and activates TFs such as the bZIP TF FOS-1 and JUN-1 (Gerke, et al., 2014; Zhang, et al., 2017). Upstream of KGB-1 is MIG-2, MAX-2, MLK-1 and MEK-1 as well as the negative regulator VHP-1, a MAPK phosphatase (**Figure 5.3**) (Mizuno, et al., 2004; Fujiki, et al., 2010). KGB-1 knockdown reduces pathogen resistance of *C. elegans* fed on *P. aeruginosa*. However, this reduction is less compared to the knockdown of its upstream kinase MEK-1. It was subsequently found that MEK-1 further confers pathogen resistance by activating PMK-1 and thus also plays a role in the p38 MAPK pathway (Kim, et al., 2004). KGB-1 is a key component for defence against the pore-forming toxin Cry5B secreted by the pathogen *B. thuringiensis*. Around ~50% of the Cry5B-responsive genes are dependent on *kgb-1* including *jun-1*. Interestingly, *fos-1* is not dependent on *kgb-1*, and neither of the two bZIP TFs is dependent on *sek-1*. The JNK MAPK pathway works in parallel with the p38 MAPK pathway against Cry5B, but controls more of the response, including the expression of p38-dependent genes (Kao, et al., 2011).

KGB-1 is also associated with other stress responses. While it is essential for the activation of the ROS-dependent UPR^{mt} (Runkel, et al., 2013) and resistance to heavy metals (Mizuno, et al., 2004), it reduces the resistance to osmotic stress (Gerke, et al., 2014). Similar to the other two MAPK pathways, *kgb-1* is also required for the *daf-16*-dependent lifespan extension. KGB-1 helps DAF-16 to localize to the nucleus in larvae, but it reduces the nuclear localization in adults, indicating an age-dependent role for KGB-1. The overall net benefit is still positive, i.e. The beneficial contribution at the larval stage more than compensates for the detrimental effects in the adult stage (Twumasi-Boateng, et al., 2012).

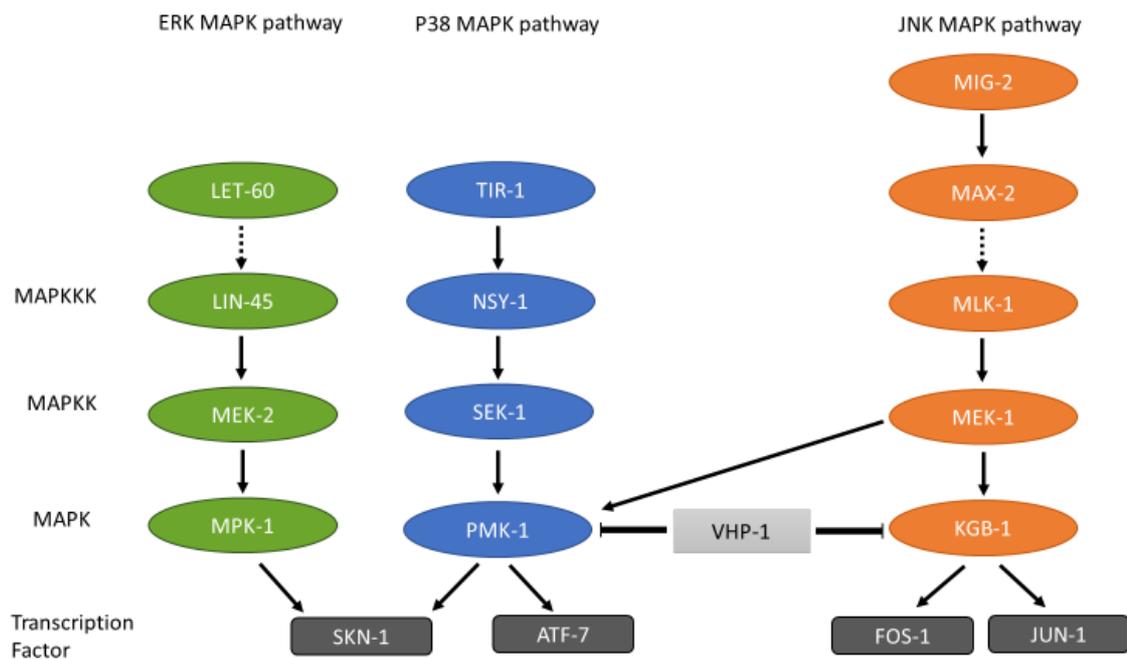


Figure 5.3 Diagram depicting the three MAPK pathways associated with the innate immune response. Dashed arrow indicates a partial requirement or redundant role in the signalling cascade. Figure based on published information (Fujiki, et al., 2010; Mizuno, et al., 2004; Nicholas & Hodgkin, 2004; Shivers, et al., 2009; Shivers, et al., 2010; Zhang, et al., 2017; van der Hoeven, et al., 2011; Okuyama, et al., 2010).

5.4.2.2. Insulin-like Signalling (ILS) pathway

The insulin-like signalling pathway affects *C. elegans* lifespan but is also associated with stress responses pathways and innate immunity. The most studied proteins in this pathway are the DAF-2 insulin-like receptor and the downstream transcription factor DAF-16, which DAF-2 negatively regulates (Shapira, et al., 2006; Nag, et al., 2017; Kim & Ewbank, 2018). *daf-16* mutants have been shown to reduce survivability against *Proteus spp.* (JebaMercy, et al., 2013) but do not affect *Mycobacterium marinum* resistance (Galbadage, et al., 2016). Since *daf-16* is negatively regulated by *daf-2*, either the loss of *daf-2* or increased *daf-16* expression confers enhanced resistance against *E. faecalis*, *S. aureus*, *P. aeruginosa*, *S. enterica*, *Y. pestis*, *Cryptococcus*

neoformans, *B. thuringiensis* (Wang, et al., 2012) and *Coxiella burnetii* (Battisti, et al., 2017). The JCMS strain of *S. maltophilia*, on the other hand, shows normal pathogenicity in *daf-2* mutants (White, et al., 2016). The functionally unknown *pud* genes (*pud-1.2*, *pud-2.1*, and *pud-3*) that are among the highest up-regulated genes following *Ochrobactrum spp.* infection are also up-regulated in *daf-2* mutants (Cassidy, et al., 2018), further supporting the involvement of the ILS pathway in the innate immune response.

5.4.2.3. Transforming Growth Factor β -like (TGF- β) pathway

The TGF- β pathway is another pathway associated with innate immune response. The TGF- β ligand DBL-1 has been found to control the defence response against infection by *D. coniospora* and *S. marcescens* (Kim & Ewbank, 2018), and is vital for survival against *S. enterica* and *E. faecalis* (Tenor & Aballay, 2008). Responses to pro-biotic lifespan-enhancing bacteria such as *Butyrivibrio pullicaecorum* and *Megasphaera elsdenii* also depended on this pathway (Kwon, et al., 2018). This pathway is also connected with other pathways described above. Infection with *P. luminescens* enhances the expression of many p38 MAPK and TGF- β target genes (Wong, et al., 2007) and DAF-16 has been shown to mediate the TGF- β pathway during infection (Nag, et al., 2017).

5.4.2.4. Toll-like receptor (TLR) signalling pathway

The Toll-like receptor (TLR) plays a surveilling role in the innate immune response. It senses the invasion of microbial pathogens. In mammals, the 12 TLR as well as the downstream signalling pathway, have been extensively studied with regards to their role in the innate immune response. *C. elegans* only has a single such receptor, *tol-1*, and lacks many of the other components of the TLR signalling pathway (Pukkila-Worley & Ausubel, 2012; Battisti, et al., 2017). Involvement of the *tol-1* receptor in *C. elegans* immune response is weak, as it did not show any effect on *D. coniospora* spore attachment, *P. aeruginosa* resistance and *M. nematophilum* tail swelling, and was only involved in the avoidance behaviour against *S. marcescens* (Pujol, et al., 2001). Furthermore, the bacteria *C. burnetii* normally identified by the TLR in vertebrates is not affected by mutation of *tol-1* in *C. elegans* (Battisti, et al., 2017). *tol-1* is also dispensable for activating the p38 MAPK pathway (Pukkila-Worley & Ausubel, 2012). To date, only *S. enterica* resistance is affected by *tol-1* (Tenor & Aballay, 2008). The involvement of *tol-1* in pathogenic avoidance behaviour is likely at the developmental level, where it promotes the development and function of the chemosensory BAG neuron (Brandt & Ringstad, 2015).

5.4.2.5. *Other transcription factors and transcriptional co-regulators involved in the innate immune response*

HSF-1 in pathogen resistance

Inactivation of *hsf-1* results in increased sensitivity of *C. elegans* to *P. aeruginosa* and prior heat-shock treatment improves the survival against *P. aeruginosa* infection (Singh & Aballay, 2006). The heat-shock induced pathogenic resistance has also been observed vice versa, where *C. elegans* grown on the pathogenic *E. coli* strain 536 conferred higher heat-shock resistance, which was correlated with higher chaperone gene expression (Leroy, et al., 2012), indicating a hormetic effect. The gene *pals-22*, associated with reduced heat shock survival and increased polyQ aggregation, has been associated as a repressor of the intracellular pathogen (*N. parisii* and Orsay virus) response. The heat shock protein *hsp-60*, which is induced by HSF-1 is important for *P. aeruginosa* resistance through its role in the upregulation of the p38 MAPK signalling (Jeong, et al., 2017).

MDT-15

The mediator complex subunit MDT-15 works with various transcription factors to regulate gene expression, including DAF-16, PMK-1 and SKN-1. Upregulation of cell non-autonomous DAF-16 target genes (e.g. *dod-11*) requires MDT-15 (Zhang, et al., 2014). MDT-15 itself can act from a distance, as the *dod-11* expression can be seen in tissues that lack *mdt-15* expression, which could indicate signalling by other molecules such as lipids (Zhang, et al., 2014). Owing to this dependence, the ILS-dependent lifespan extension phenotype also requires MDT-15. However, the extent to which the ILS pathway plays a major role is difficult to assess as *mdt-15* itself is already required for normal lifespan of wild-type worms and enhanced lifespan of other long-lived mutants (Grants, et al., 2015). Furthermore, like *daf-16* mutants, *mdt-15* mutants are also sterile to some degree and sensitive to pathogen infection (Pukkila-Worley, et al., 2014). While sterility can directly enhance pathogen resistance through a DAF-16 dependent induction of stress response gene (Miyata, et al., 2008), this is not the case for the *mdt-15* mutants against *P. aeruginosa* (Pukkila-Worley, et al., 2014). This indicates that the gene *mdt-15* itself and not the sterility of *mdt-15* mutants is involved with the innate immune response. MDT-15 is required for the induction of PMK-1 dependent gene expression to confer resistance against the Phenazine toxins produced by *P. aeruginosa* (Pukkila-Worley, et al., 2014), perhaps via the transcription factor SKN-1, since MDT-15 physically associates with SKN-1 and is required to induce SKN-1 target genes (Goh, et al., 2014).

Since MDT-15 induces expression of SKN-1 target genes, it also suggests some link to oxidative stress. Indeed, *mdt-15* mutants are more susceptible to oxidation agents arsenite and t-BOOH (tert-butyl hydroperoxide). Resistance against t-BOOH, however, is SKN-1 independent, indicating that *mdt-15* is also involved with SKN-1 independent oxidative stress response pathways (Goh, et al., 2014).

ELT-2

The GATA transcription factor ELT-2 works close together with p38 MAPK pathway to induce pathogenic resistance. It is a major transcriptional regulator in the intestine (Yang, et al., 2016). ELT-2 is required for resistance against *P. aeruginosa* (Shapira, et al., 2006; Head, et al., 2017), *S. typhimurium*, *E. faecalis* and *C. neoformans* (Yang, et al., 2016). Apart from functioning in the innate immune response, p38 MAPK signalling and ELT-2 have also been shown to mediate osmotic stress (Gravato-Nobre & Hodgkin, 2005; Yang, et al., 2016). On the other side, ELT-2 does not affect resistance to cadmium, heat and oxidative stress (Shapira, et al., 2006).

PQM-1

The transcription factor PQM-1 is up-regulated in *C. elegans* under *S. marcescens* and *Xenorhabdus nematophila* infection (Sinha, et al., 2012). Mutation in PQM-1 results in reduced survivability against *P. aeruginosa* (Shapira, et al., 2006). A close but inverse relationship has been observed between PQM-1 and DAF-16, where they do not localize at the nucleus simultaneously and appear to be oppositely regulated by insulin signalling (Tepper, et al., 2013). The observation that DAF-16 does not localize to the nucleus following *C. albicans* infection but genes negatively regulated by DAF-16 are still down-regulated (Pukkila-Worley, et al., 2011), could potentially point towards PQM-1 taking over the role of suppressing the DAF-16 dependent genes. Furthermore, PQM-1 could play a role in the signalling of the immune response across tissues, as it is important for transcellular chaperone signalling of the heat shock protein HSP-90, which has been associated with the innate immune response (O'Brien, et al., 2018).

5.5. *C. elegans* pathogens investigated in this study

Like many organisms, *C. elegans* can be infected by a wide range of pathogens from fungi and bacteria to viruses and lead to the activation of the innate immune response and associated cellular stress responses. Due to the complexity of multicellular organisms, of which *C. elegans*

is no exception, pathogens can exploit various mechanical and biological mechanisms to work in their advantage. This allows for a diverse range of niches to which pathogens have evolved to occupy. With regards to the metaphor of the “arms race” between host and pathogen, each pathogen only needs to successfully overcome the defence at one exploitable niche/mechanism, while the host needs to defend each and every aspect of its complex working. Here, various *C. elegans* pathogens are described, that are directly relevant to the datasets used in this study.

5.5.1. Gram-positive bacteria

5.5.1.1. *Bacillus thuringiensis*

B. thuringiensis is a spore-forming bacterium that produces the commercially important Crystal toxins (Cry) used as an insecticide. The Cry toxins are classified as pore-forming toxins that function by lysing epithelial cells through osmotic shock, thereby disrupting the intestine epithelium and killing the insect (Wan, et al., 2019). While *B. thuringiensis* is a soil-dwelling bacterium, it can complete its full lifecycle within various invertebrates, including *C. elegans*. The bacterium produces a diverse range of related crystal toxins, but only a limited number of these affect nematodes. Wei, et al. (2003) found that only Cry5B was toxic to all their tested nematodes (*C. elegans*, *Pristionchus pacificus*, *Panagrellus redivivus*, *Acrobeloides* sp., *Distolabrellus veechi* and *Nippostrongylus brasiliensis*). Cry14A is the most potent toxin against *C. elegans* but does not affect other nematodes including *Pristionchus pacificus* and *Acrobeloides* sp. (Wei, et al., 2003). The *B. thuringiensis* strain DB27 is very lethal to *C. elegans*, killing the worm within 16 hours, while not affecting *P. pacificus* (Sinha, et al., 2012). The chemical Cry5B itself is lethal for *C. elegans* starting from a concentration of 8 µg/mL. Cry5B requires specific glycolipid receptors and mutants deficient in these (*bre-4*, *bre-5*) show no infection and lethality (Hu, et al., 2010; Kho, et al., 2011).

B. thuringiensis infects *C. elegans* through the oral route, accumulating in the intestine where it releases toxins that form pores at the intestinal epithelial junction. The pores act as a groove for the bacterial spores to stick in and germinate. The nematode dies from the toxins, and the germinated spores use the body to survive necrotrophically and sporulate (Wan, et al., 2019), giving rise to the bag of bacteria phenotype (Bob) (See **Figure 5.4**) (Kho, et al., 2011).

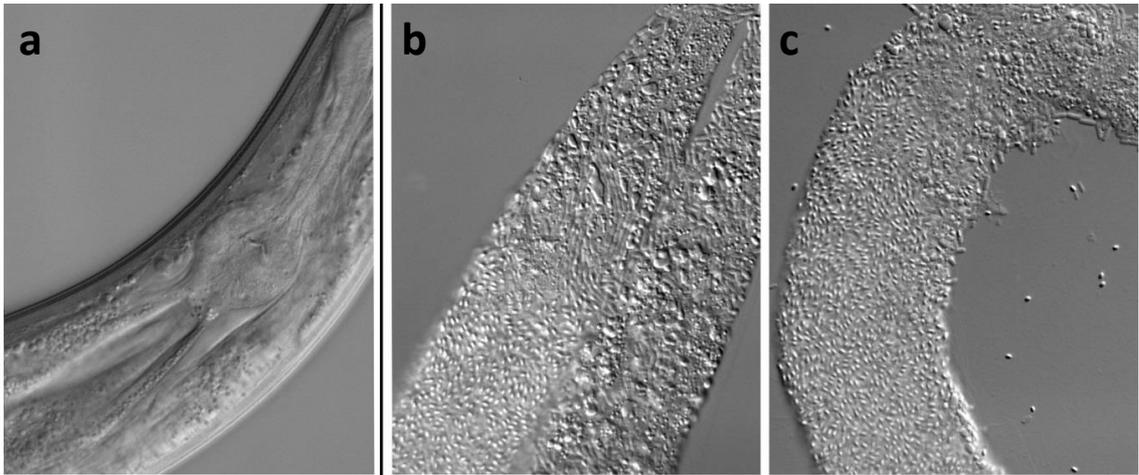


Figure 5.4 *C. elegans* infected with *Bacillus thuringiensis*. a) Infection with *B. thuringiensis* strain that does not produce Cry5B. *C. elegans* does not show any health-compromising symptoms. b) & c) *B. thuringiensis* infection with Cry5B toxin. Animals show Bob phenotype and are nearly completely digested with sporulating bacterial spore. Image taken from Kho, et al. (2011).

5.5.1.2. *Bacillus subtilis*

B. subtilis also belongs to the same genus as *B. thuringiensis*, but its effect on *C. elegans* is drastically different. Under normal conditions, infection by *B. subtilis* induces a longevity phenotype partly dependent on reduced insulin signalling, as *daf-2* mutant longevity is not significantly enhanced by *B. subtilis* (Donato, et al., 2017). However, when Cry5B is available, *B. subtilis* becomes infectious to *C. elegans* and shows the same phenotype, but at reduced levels, as *B. thuringiensis* (similar to **Figure 5.4b & c**). Tests with other *Bacillus* species, *Bacillus sphaericus* and *Bacillus megaterium*, showed a barely detectable level of infection (Kho, et al., 2011), indicating that specifically *B. subtilis*, and not the other two *Bacillus* species, could be an opportunistic pathogen in the context of *C. elegans*.

Compared to the standard *E. coli* (OP50) diet, *C. elegans* fed on *B. subtilis* produce fewer offspring under standard conditions. However, under heat shock stress, a *B. subtilis* diet results in more offspring compared to the heat-shocked worms fed with OP50 (Hoang, et al., 2019). This effect may be related to *B. subtilis* conferring higher heat shock lifespan as a result of nitric oxide production and biofilm formation in the nematode gut (Donato, et al., 2017). These two observations may be due to the hormetic effect discussed in section 5.4.2.5, where a diet of pathogenic *E. coli* strain conferred higher heat shock resistance survival (Leroy, et al., 2012).

5.5.1.3. *Enterococcus faecalis*

Enterococcus spp. are spherical (cocci) bacteria found in the intestine (enteric) of many organisms, including mammals, reptiles and insects. They are also found in the soil and water bodies, as well as in dairy products and fermented food products. *E. faecalis* is the most prevalent species in the *Enterococcus* genus and accounts for 80-90% of Enterococcus-associated infections in humans (H. M. S. Goh, et al., 2017). In *C. elegans*, live *E. faecalis* can kill the nematode adults, as well as, eggs and hatchlings in a mechanistically distinct manner. In the adult worms, the ingested bacteria that survived the pharyngeal grinder can form colonies in the intestine and accumulate to high titres thereby grossly distending the intestinal lumen (**Figure 5.5B**) (Garsin, et al., 2001; Yuen & Ausubel, 2018).

The immune response against *E. faecalis* depends partly on the p38 MAPK pathway (PMK-1) as well as FSHR-1 and BAR-1 dependent pathways (Yuen & Ausubel, 2018). Although heat-killed *E. faecalis* does not kill *C. elegans*, it still activates the innate immune response, indicating that recognition of this pathogen is through certain heat-stable microbe-associated molecular patterns (MAMPs) (Yuen & Ausubel, 2018). The ILS pathway plays a significant role in *E. faecalis* resistance, as mutations in *daf-2* drastically increase the survival rate (four to fivefold increase compared to wild-type) of the worm (Garsin, et al., 2003). Furthermore, oxidative stress plays a vital role during *E. faecalis* infection as the oxidative stress response transcription factor SKN-1 (Papp, et al., 2012) and DAF-16-regulated antioxidant enzymes SOD-3 and CTL-2 contribute to *E. faecalis* resistance (Chavez, et al., 2007).

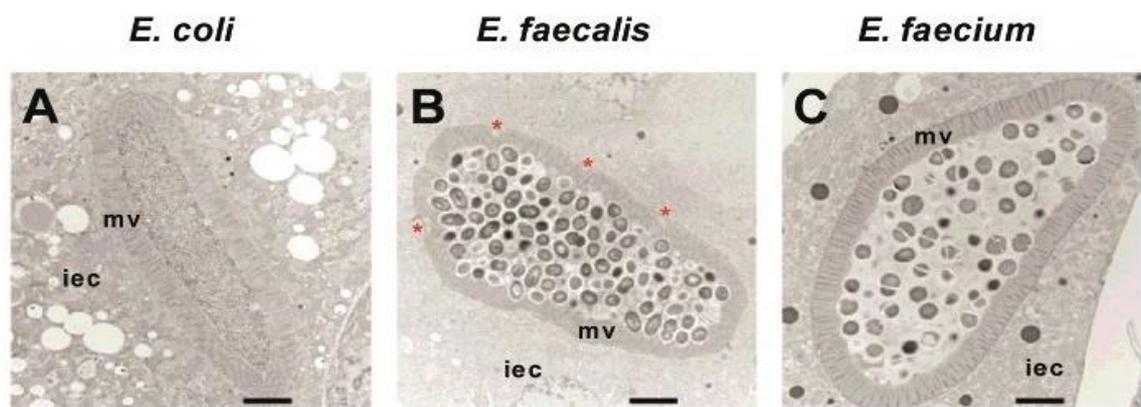


Figure 5.5 Cross-section of *C. elegans* fed on *Enterococcus* bacteria after 8 hours. A) Healthy intestine of wild-type worms fed on *E. coli*. B) *E. faecalis* and C) *E. faecium* fed worms show significant distending of the intestine and accumulation of the bacteria compared to standard *E. coli* bacterial diet (A). mv = microvilli. iec = intestinal epithelial cell. Red stars marks site of dehiscence (separation) between the terminal web and the luminal membrane. Image taken from Yuen & Ausubel (2018).

5.5.1.4. *Enterococcus faecium*

E. faecium is the second most common *Enterococcus*-species infecting humans, accounting for 10-15% of *Enterococcus*-associated infections (H. M. S. Goh, et al., 2017). While it shares many aspects of infection with *E. faecalis*, such as high titre accumulation in the intestine (**Figure 5.5C**) and the ability to kill *C. elegans* eggs and hatchlings, there are some differences as well. For example, *E. faecium* is unable to kill *C. elegans* adult worms. As such, the distention of the intestine due to *Enterococcus* bacteria accumulation is unlikely the cause of death in infected *C. elegans* (Garsin, et al., 2001).

Although *E. faecium* does not kill wild-type adult worms, immunocompromised *C. elegans* with mutations in the p38 MAPK pathway (*pmk-1*) and *fsbr-1* and *bar-1* genes show reduced survival rate against *E. faecium* in the same manner as observed in *E. faecalis*, indicating that *E. faecium* acts as an opportunistic pathogen whose pathogenicity depends on the host innate immune system (Yuen & Ausubel, 2018). In addition, the transcriptomic analysis revealed that there is a significant overlap in differentially expressed gene signature of *C. elegans* infected with *E. faecalis* and *E. faecium*. Mutant worms of *pmk-1*, *fsbr-1* and *bar-1* fed on live or dead *Enterococcus* species also show a high degree of similarity in gene expression signatures (Yuen & Ausubel, 2018). These indicate that the *C. elegans* immune system recognizes *E. faecium* and *E. faecalis* using similar cues.

5.5.1.5. *Microbacterium nematophilum*

M. nematophilum is a rod-shaped bacteria that attaches itself tightly to the anal opening of the worm which results in the *dar* (deformed anal region) phenotype. This phenotype is characterized by the worms having a swollen tail (**Figure 5.6**). The bacteria do not penetrate the cuticle, and their invasion does not become any more severe, making its pathogenicity different from other pathogens (Hodgkin, et al., 2000). Hodgkin, et al., 2000 hypothesized that the bacterial attachment to the anal region has survival benefits for the bacteria. Firstly, the location is the safest place to avoid being eaten. Secondly, defecation and leakage of gut contents are nutritious for the bacteria. Thirdly, the bacteria can use the nematode as a vehicle to disperse itself to new sites behind the worm's tail. No benefit was identified for *C. elegans*, which led the authors to classify *M. nematophilum* as pathogenic rather than symbiotic (Hodgkin, et al., 2000). Upon further investigation, the *dar* phenotype has been identified as a defensive response by the nematode through the ERK MAPK pathway. Mutation of the ERK MAPK pathway that abrogates the *dar* phenotype resulting in the *bus* phenotype.

Although this phenotype has less anal swelling, it shows much stronger constipation that would result in developmental arrest and cause sterility (Nicholas & Hodgkin, 2004).

M. nematophilum has difficulties attaching to *C. elegans* whose cuticle surface has been altered via mutations (*stf-3* and *bus-2*, *bus-3*, *bus-12* and *bus-17*), leading to the hypothesis that *M. nematophilum* depends on surface composition to identify and attach to *C. elegans* (Gravato-Nobre, et al., 2011; Höflich, et al., 2004).

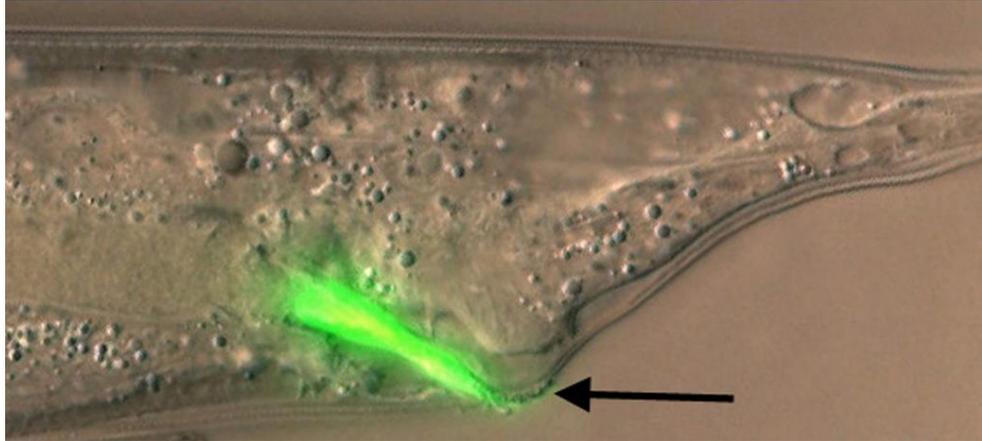


Figure 5.6 *C. elegans* infected by *M. nematophilum*. Arrow indicates the deformed anal region (dar) phenotype. Bacteria are stained with Syto 13 dye. Photo taken by Hannah Nicholas, Delia O'Rourke and Jonathan Hodgkin (Darby, 2005).

5.5.1.6. *Staphylococcus aureus*

S. aureus is an opportunistic pathogen that is clinically important as it infects humans, causing diseases from mild skin infections to severe life-threatening conditions. It is also economically relevant as it affects cattle, resulting in mastitis. In *C. elegans*, most of the *S. aureus* strains can kill the nematode within 5 days post-infection, provided there is a prolonged accumulation of the bacteria in the intestine (**Figure 5.7**). *C. elegans* exposed to the bacteria shorter than 8 hours can recover back to normal lifespan, as the bacterium is unable to colonize and persist in the worm's intestine. After 8 hours, the length of exposure is inversely related to lifespan, with more prolonged exposure leading to a shorter lifespan, indicating that accumulation of the bacteria beyond a certain threshold becomes lethal (Sifri, et al., 2003). *S. aureus* damages the intestinal cells by lysing the epithelial cells, after which they invade and degrade the rest of the body. Furthermore, infected worms show a variety of phenotypes including slower movement, fewer egg production and deformed anal region (Irazoqui, et al., 2010).

Transcriptomic analysis of *S. aureus* infected *C. elegans* shows upregulation of epithelial detoxifying and antimicrobial peptides. Furthermore, a small subset of the differentially

expressed genes correlates with those enriched through *B. thuringiensis* and Cry5B toxin, indicating a potential commonality towards intestinal cell membrane rupture (Irazoqui, et al., 2010).

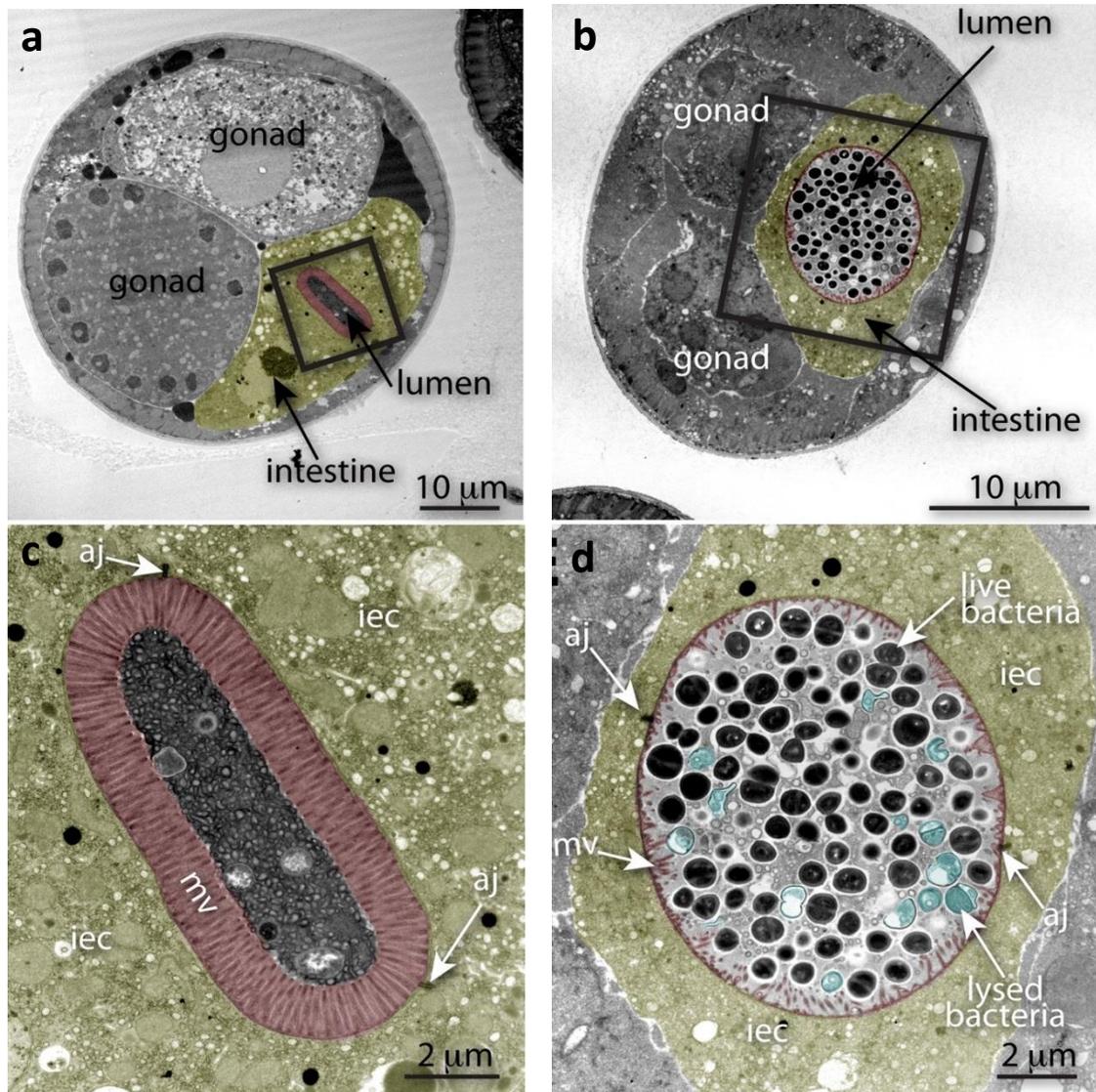


Figure 5.7 *S. aureus* accumulation inside *C. elegans* gut. a) & c) *C. elegans* fed with non-pathogenic *E. coli* bacteria after 12 hours. b) & d) *C. elegans* fed on pathogenic *S. aureus* bacteria after 12 hours. mv = microvilli. iec = intestinal epithelial cell. aj = apical junction. Image adjusted from Irazoqui, et al. (2010).

5.5.2. Gram-negative bacteria

5.5.2.1. *Salmonella enterica* (serovar *Typhimurium*)

S. enterica is an opportunistic enteric bacterium infecting a wide range of species. Once it has infected the host, the bacterium acts as an intracellular pathogen, infecting macrophages where they can proliferate and replicate. *S. enterica* has many variants (serovar) that occupy

different ranges of hosts and cause various diseases. In humans, *S. enterica* infection is frequently encountered as a food-borne illness, causing salmonellosis with a wide range of symptoms (Jantsch, et al., 2011). In *C. elegans*, many serovars are infectious, the most commonly researched one being *S. enterica* serovar Typhimurium (henceforth *S. enterica*). Larval stage 4 (L4) and adult *C. elegans* show distended intestine and die much quicker when fed on *S. enterica* compared to OP50 fed worms (**Figure 5.8**). Even short exposure of 5 hours and very diluted bacterial concentration is sufficient for the bacteria to infect and persist in the nematode intestine, accumulating to high titre for the rest of its life. The killing of the worm requires live bacteria, as feeding on heat-killed *S. enterica*, show normal lifespan (Aballay, et al., 2000; Labrousse, et al., 2000).

The programmed cell death pathway is necessary for the innate immune response against *S. enterica* and has been shown to be dependent on the p38 MAPK pathway (Aballay & Ausubel, 2001; Aballay, et al., 2003). Furthermore, functional bacterial lipopolysaccharides are required to elicit intestinal persistence and programmed cell death (Aballay, et al., 2003).

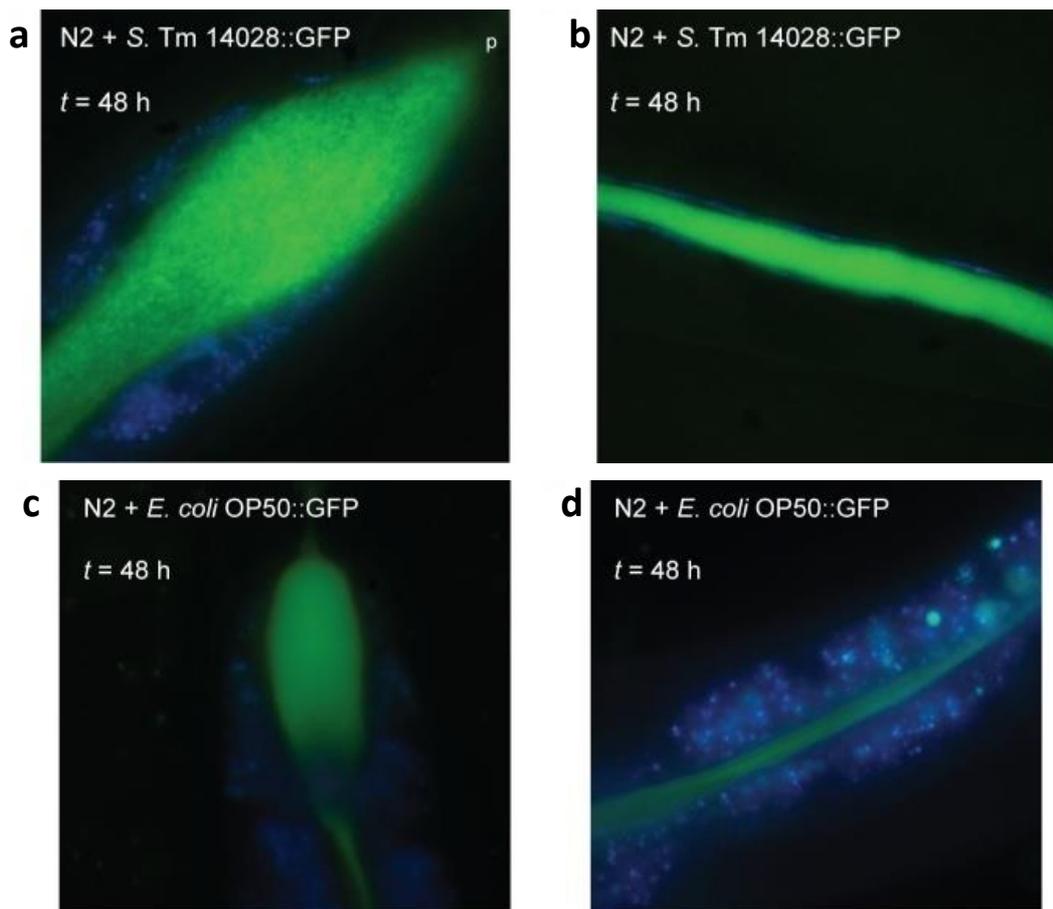


Figure 5.8 *S. enterica* (Typhimurium) accumulation inside *C. elegans* gut. a) & b) *C. elegans* fed on pathogenic *S. enterica* bacteria after 48 hours. c) & d) *C. elegans* fed with non-pathogenic *E. coli* bacteria after 48 hours. Green fluorescence shows GFP-tagged bacteria. Blue fluorescence is the intestinal autofluorescence (from intestinal lysosome-related gut granules). Image taken from Sem & Rhen (2012).

5.5.2.2. *Stenotrophomonas maltophilia*

S. maltophilia is a nosocomial (originating from a hospital) opportunistic pathogen that has been associated with many diseases in humans such as pneumonia and meningitis. In the wild, *S. maltophilia* is found ubiquitously from soil to water (White & Herman, 2018). The *S. maltophilia* strain JCMS is pathogenic in *C. elegans*, able to kill the nematode in 5 days on average. It is suggested that the accumulation of live bacteria in the intestine (**Figure 5.9**) is the cause of *C. elegans* mortality rather than secreted toxins (White, et al., 2016). Some *S. maltophilia* strains can be avirulent (e.g. K279a) or less potent than JCMS (e.g. R551-3) (White, et al., 2016).

Multiple stress response pathways play a role in *S. maltophilia* resistance, including the UPR and the innate immune response pathways p38 MAPK and TGF β . The ILS pathway, on the other hand, does not have any observable effects on the susceptibility to *S. maltophilia* (White, et al., 2016).

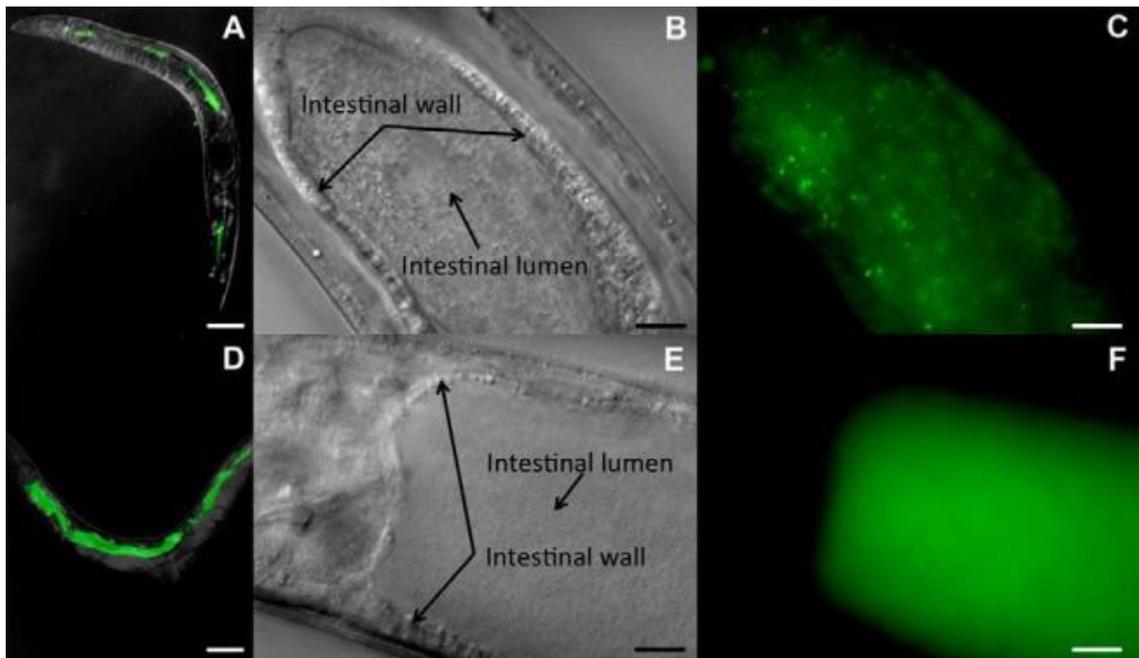


Figure 5.9 *S. maltophilia* accumulation inside *C. elegans* gut. A) - C) *C. elegans* fed with non-pathogenic *E. coli* bacteria. D) - F) *C. elegans* fed on pathogenic *S. maltophilia* bacteria. A) & D) fluorescence images on day 6 after the worms were exposed to the bacteria. B) & E) images taken on day 4. C) & F) fluorescence images of panel B) and E) respectively. Image taken from White, et al. (2016).

5.5.2.3. *Photorhabdus luminescens*

P. luminescens is an enteric bacterium that lives within the nematode *Heterorhabditis bacteriophora* symbiotically but acts as an insecticide and nematicide in other organisms including *C. elegans*. *H. bacteriophora* invades insects, where it then regurgitates *P. luminescens* that kills the host. *H.*

bacteriophora then uses the cadaver to survive. Various strains of *P. luminescence* kills most *C. elegans* within 5 days post-infection (Sato, et al., 2016) while causing reduced developmental and reproductive rate (Sicard, et al., 2007). Although the bacterium does not proliferate in the intestine of *C. elegans*, the nematode pharynx is relatively inefficient in grinding up this bacterium, allowing it to arrive in the intestine mostly alive. Infection after exposure starts rapidly after 2 hours of feeding, where crystal-like structures begin forming within the intestinal lumen (**Figure 5.10**). Removing *C. elegans* from *P. luminescence* exposure after as much as 12 hours allows the worms to survive with a healthy morphology and reproduction, but does not remove or reduce the crystals in the intestine. (Sato, et al., 2014). The p38 MAPK pathway is required for host defence, while the ILS pathway is deactivated by *P. luminescence* (Sato, et al., 2014).

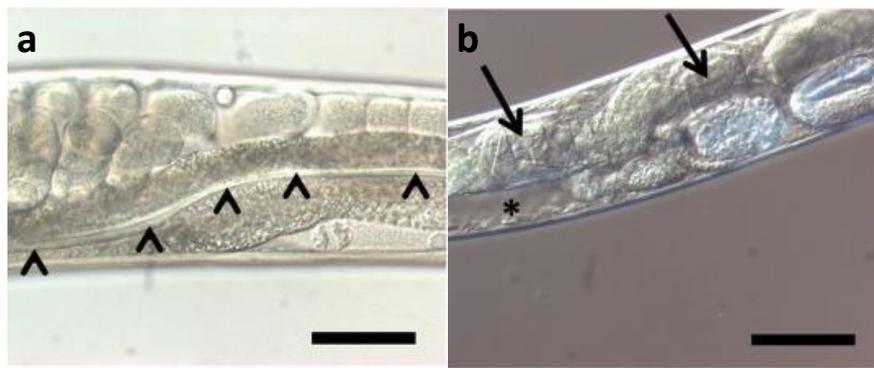


Figure 5.10 Crystal structure formation inside *C. elegans* intestine following *P. luminescence* infection. a) *C. elegans* fed on non-pathogenic *E. coli* bacteria for 44 hours. Arrowhead points to the healthy intestine. b) *C. elegans* fed on pathogenic *P. luminescence* bacteria after 44 hours. Arrows indicate crystal-like structures in the intestinal lumen. Image taken from Sato, et al. (2014).

5.5.2.4. *Yersinia pestis*

Y. pestis is a well-known pathogen for its devastating role as the black death in Eurasia (Perry & Fetherston, 1997). In *C. elegans*, *Y. pestis* and its relative *Yersinia pseudotuberculosis* attaches to the head of the nematode and forms a biofilm around the mouth, thereby preventing the worms from feeding and starving them to death (**Figure 5.11**) (Darby, et al., 2002). However, biofilm formation only happens in adult worms as eggs hatched on *Y. pestis* lawns can still reach adulthood (Styler, et al., 2005). The binding doesn't seem to be affected by the cuticle structure, as *C. elegans* with mutations in the collagen genes *dpy-5*, *dpy-9*, *dpy-17* and *rol-6* as well as the blistered *bli-6* mutant show normal *Y. pestis* binding. Mutation of the surface coat on the other hand such as the *bus* mutants *bus-2*, *bus-3*, *bus-12* and *bus-17*, as well as *srf-2* and *srf-5* mutants reduce the biofilm formation, causing the *bab* (Biofilm absent on head) phenotype (Drace, et al., 2009).

Y. pestis mutants that lack the biofilm-forming gene *bmsHFRS* can also kill *C. elegans* in a biofilm-independent manner through the accumulation and colony formation in the intestine (Styler, et al., 2005).

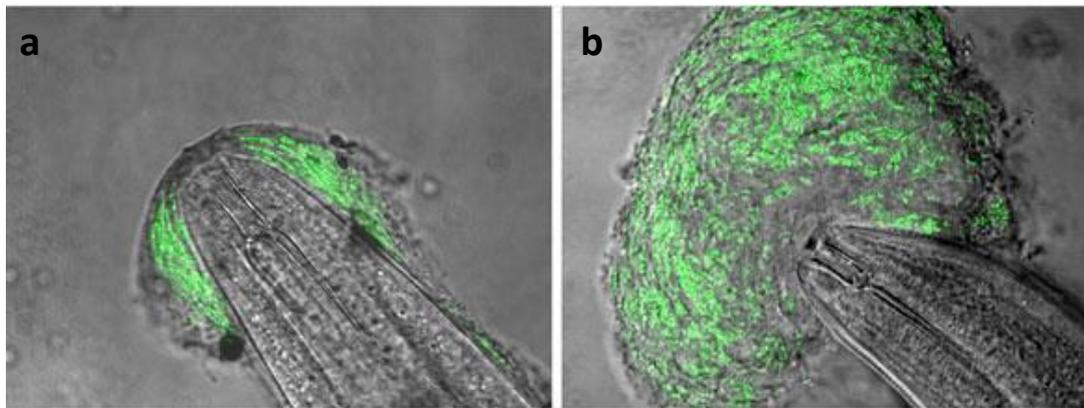


Figure 5.11 *Y. tuberculosis* accumulation on *C. elegans* head. Accumulation of bacteria on *C. elegans* head after a) 1 hour exposure and b) overnight incubation on a bacterial lawn. Green fluorescence shows GFP tagged *Y. tuberculosis*. Image taken from Darby (2005).

5.5.2.5. *Pseudomonas aeruginosa*

P. aeruginosa is a common soil and water bacterium that is also an opportunistic human pathogen, causing diseases in immunocompromised hosts. It is one of the most studied *C. elegans* pathogens (Tan, et al., 1999; Ermolaeva & Schumacher, 2014). The PA14 strain can kill *C. elegans* in a “slow killing” mechanism or “fast killing” mechanism, depending on the condition in which the bacteria grow. “Slow killing” requires live bacteria, while “fast killing” can be achieved even in heat-killed bacteria (Tan, et al., 1999). The slow-killing mechanism resembles a *bona fide* infection process which involves the accumulation of live bacteria in the intestine of the worm (**Figure 5.12**), leading to the death of the worm within 3 days. One toxin believed to be responsible for slow killing is pyoverdine, which extracts iron from the mitochondria, damaging it, for use by the bacteria for growth and biofilm formation (Kang & Kirienko, 2017; Kang, et al., 2018). The fast-killing mechanism, on the other hand, relies on other toxins such as phenazine, which rapidly kills most worms within hours after exposure (Tan, et al., 1999). The main attribute of phenazines is their ability to generate reactive oxygen species (ROS) in tissues of the host and damage them via oxidative stress (Pierson III & Pierson, 2010; King, et al., 2018). Fast killing is not counteracted by the intestinal innate defence mechanism and depends on the Ethanol and Stress Response Element (ESRE) which is mediated by the bZIP protein family (Tjahjono & Kirienko, 2017). Phenazine and other toxins disrupt the oxidative phosphorylation in *C. elegans* which would result in the activation of the UPR^{mt}. However, *P. aeruginosa* exploits the negative regulator

of UPR^{mt}, the bZIP protein ZIP-3, to repress the UPR^{mt} (Deng, et al., 2019). The p38 MAPK pathway plays a role in defence against *P. aeruginosa* intestinal colonization (slow killing). Loss-of-function or knockdown of any of the kinases in this pathway (PMK-1, SEK-1 and NSY-1) reduces survivability against *P. aeruginosa* (Kim, et al., 2002). Furthermore, the p38 MAPK pathway receptor TIR-1 is necessary to activate SKN-1 under PA14 infection, which is necessary for resistance against this pathogen (Papp, et al., 2012). The ILS pathway, on the other hand, doesn't affect resistance as *daf-16* RNAi knockdown did not affect survivability significantly (Sun, et al., 2011). The oxidative stress associated transcription factor PQM-1 does not affect *C. elegans* survival against *P. aeruginosa* when knocked-out but is required for TCS-dependent resistance (O'Brien, et al., 2018).

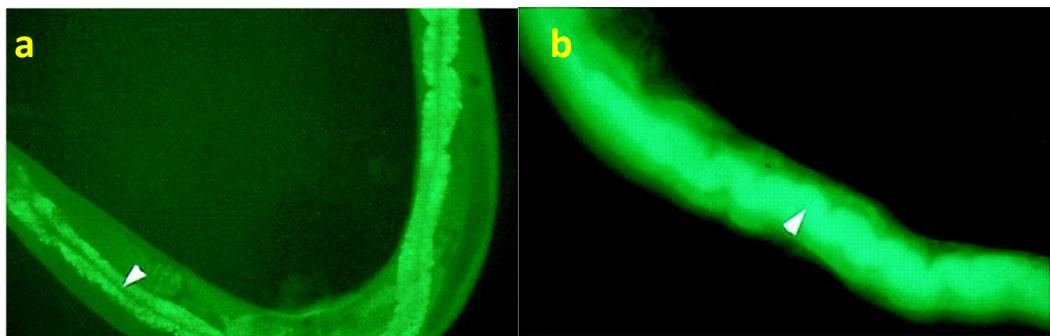


Figure 5.12 *P. aeruginosa* accumulation inside *C. elegans*. a) *C. elegans* fed on non-pathogenic *E. coli* bacteria for 48 hours. b) *C. elegans* fed on pathogenic *P. aeruginosa* bacteria after 48 hours. Bacteria are tagged with GFP. *P. aeruginosa* accumulates in the lumen at a much higher level compared to *E. coli*. Arrows point towards the intestinal lumen. Image taken and adapted from Tan, et al. (1999).

5.5.2.6. *Vibrio cholerae*

V. cholerae is a clinically relevant bacteria that causes the diarrheal disease cholera, prevalent in many parts of Africa and Asia. This bacterium is found in aquatic reservoirs, adapted to different environmental conditions (e.g. temperature variation and osmotic stress). The microbivore *C. elegans* represents a natural predator to *V. cholerae*, to which the bacterium has developed protective responses (List, et al., 2018). *V. cholerae* can accumulate in the pharynx and intestine (**Figure 5.13**). The killing of *C. elegans* by *V. cholerae* takes up to 5 days after exposure and requires continuous exposure to life *V. cholerae* (Vaitkevicius, et al., 2006). The exact mechanism of killing is not known; interestingly, biofilm formation or secretion of Cholera toxin, which is important for pathogenesis in other organisms, is not required for killing in *C. elegans* (Vaitkevicius, et al., 2006). Vacuolization of the *C. elegans* intestine could play a role in the pathogenesis as *V. cholerae hlyA* mutants which do not produce vacuoles in the nematode are less lethal (Cinar, et al., 2010). Furthermore, transcriptome analysis found

that *blyA* is important to induce the differential expression of DAF-16 target genes in *C. elegans*, indicating a connection with the ILS pathway (Sahu, et al., 2012).

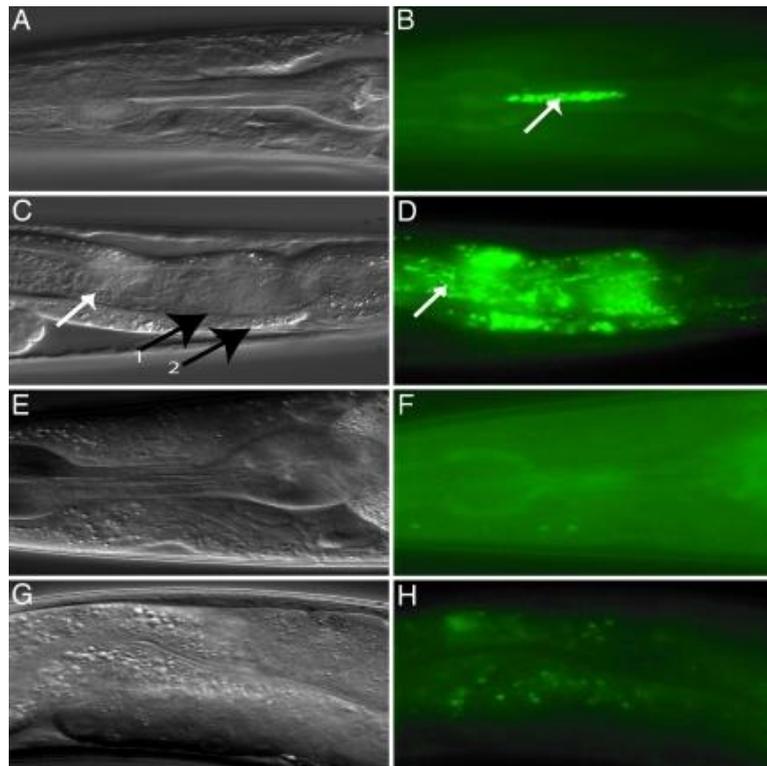


Figure 5.13 *V. cholerae* accumulation inside *C. elegans* pharynx and intestine. A) - D) *C. elegans* fed with *V. cholerae* for 12 hours. White arrows point to *V. cholerae* bacilli and black arrows show the apical membrane. E) - H) *C. elegans* fed on non-pathogenic *E. coli* bacteria. Green fluorescence indicates GFP-tagged bacteria. Image taken from Vaitkevicius, et al. (2006).

5.5.2.7. *Serratia marcescens*

S. marcescens is a soil-dwelling human opportunistic pathogen that causes various diseases such as meningitis and endocarditis. Over the past decade, nosocomial infection and antibiotic resistance have been increasing, making it a clinically more significant bacterium (Kurz & Ewbank, 2000). In *C. elegans*, *S. marcescens* infection requires live bacteria reaching the intestine, after which it is able to kill the host within 6 days (Kurz, et al., 2003). Infection of larval stage 4 (L4) worms starts within 6 hours of exposure to *S. marcescens* followed by rapid proliferation, leading to intestinal distention and progressive vacuolation of intestinal cells. After 48 hours, damage to the intestinal epithelium and germline is observed accompanied by a reduction in egg-laying. After 72 hours, worms start to die. Worms hatched directly on *S. marcescens* bacterial environment (lawn) however are more resistant to the pathogen (Mallo, et al., 2002; Kurz, et al., 2003).

Many genes have been implicated to play a role in the *C. elegans* – *S. marcescens* relationship. *C. elegans* recognition and avoidance of *S. marcescens* depend on the worms AWB olfactory

sensory neuron to identify the bacterial cyclic pentapeptide surfactant Serrawettin W2 (Pradel, et al., 2007). The TGF- β pathway has been found to be important for resistance against this pathogen (Mallo, et al., 2002).

5.5.3. Fungi

5.5.3.1. *Drechmeria coniospora*

D. coniospora is an endoparasitic nematophagous fungus that infects various nematodes and uses them for its own reproduction. Infection by *D. coniospora* starts when the conidia (non-motile spore) attaches itself to the nematode cuticle at the head (sensory amphid) and vulva (inner labial papillae) via adhesive knobs. The penetration tube of the fungus then pierces the cuticle of the nematode, using the combination of enzymatic action and mechanical force. Trophic hyphae then grow into the nematode and spread throughout the epidermis until the whole worm has been taken over (**Figure 5.14**). New spores grow from the bulbs that develop from the trophic hyphae while the fungi absorb the remaining nutrients from the nematode corpse (Jansson, 1994; Pujol, et al., 2008a; Zhang, et al., 2016). Similar to the other cuticle-dependent pathogens *M. nematophilum* and *Y. pestis*, the tested bus genes (*bus-2*, *bus-3*, *bus-12* and *bus-17*) also affected *D. coniospora* attachment, however with opposite effects. Mutations of these genes enhance fungal attachment to the nematode body, which the authors hypothesize might be related to the increased level of α -linked L-fucose-specific lectin on the surface (Rouger, et al., 2014).

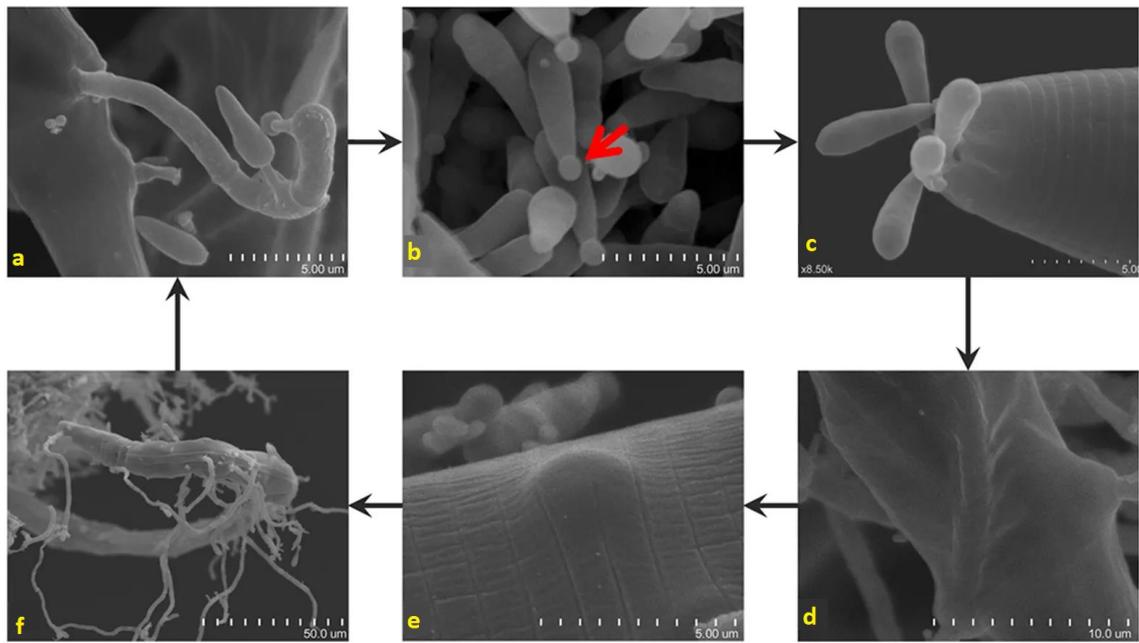


Figure 5.14 *D. coniospora* infection cycle in *C. elegans*. a) *D. coniospora* spores grow from the bulbs. b) Spores mature by developing adhesive knobs (red arrow). c) Spores attached to the head of *C. elegans*. d) Trophic hyphae penetrate and grow inside *C. elegans*. e) Trophic hyphae develop bulbs that are pressed against the insides of the cuticle. f) Conidiophores sprout out of the *C. elegans* cadaver. Image taken from Zhang, et al. (2016).

5.5.3.2. *Nematocida parisii*

N. parisii belongs to the Microsporidia phylum that contains many species of obligate intracellular pathogens (which depend on host cells to replicate) with a wide variety of animal hosts, including humans. Microsporidia exist as spores outside the host and inject their nuclei and sporoplasm into the host via a “syringe” called polar tube. In *C. elegans*, *N. parisii* infects the nematode from its intestine, likely by an oral route. After infecting intestinal cells, the fungi replicates in a cell-wall deficient form called a meront, which then differentiates into spores (**Figure 5.15**). The spores are then believed to escape the worm through rupturing the terminal web (a cytoskeletal structure beneath the base of the villi lining the intestinal wall) (Troemel, et al., 2008). The size and number of the spores affect horizontal infection, as *C. elegans* with only few small-sized spores were able to infect other worms. *N. parisii* is one of two intracellular pathogens, the other being the Orsay virus, that has been found to naturally infect *C. elegans* (Bakowski, et al., 2014). Although it is a fungus, the gene expression response is distinct from other fungal (and bacterial) pathogens and most closely resemble the Orsay virus (Troemel, et al., 2008; Bakowski, et al., 2014), indicating that intracellular pathogens trigger a similar response. The p38 MAPK and ILS pathway do not play a significant role in resistance against *N. parisii* (Troemel, et al., 2008).

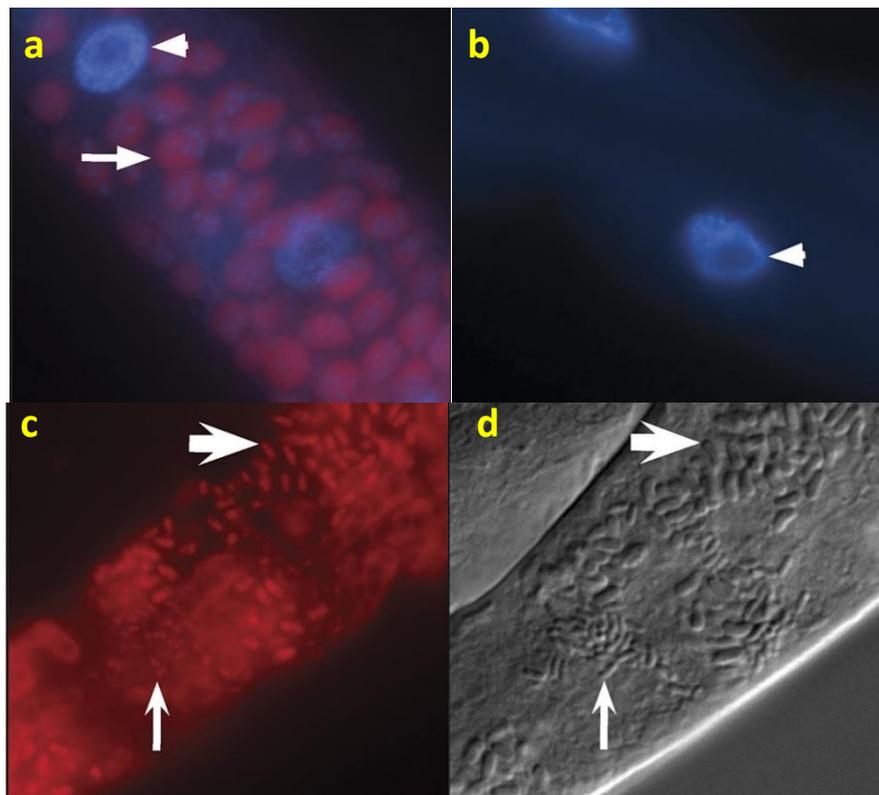


Figure 5.15 *C. elegans* infected by *N. parisii*. a) & b) FISH staining using *N. parisii* rRNA specific probes visualized in red and DAPI staining in blue of a) infected and b) uninfected *C. elegans*. Arrowheads point to host nucleus and arrows indicate meronts. c) & d) Rod-shaped *N. parisii* spores inside *C. elegans* host visualized using c) FISH with *N. parisii* rRNA specific probes and d) Normaski microscopy. Small arrows indicate small spores and large arrows indicate large spores. Image adjusted from Troemel, et al. (2008).

5.5.3.3. *Harposporium* sp.

Species belonging to *Harposporium* are endoparasitic and, for the majority, nematophagous fungi that infect a broad range of nematodes. They are characterized by sickle-shaped conidia that infect through ingestion (except *Harposporium subuliforme*) and germinate and colonize within the host's body (Wang, et al., 2007). The *Harposporium* species whose infection transcriptomic data is used in this study (JUf27) was isolated relatively recently (October 2008) from *C. elegans* collected in France and never properly classified. This nematophagous fungus has been observed to infect the worm through the intestine after ingestion of the conidia, where it produces hyphae that then invade the whole body and penetrates the epidermis to form spores on the surface of the dead worm (**Figure 5.16**). *C. elegans* infected by this fungus die within six to eight days (Engelmann, et al., 2011; Felix & Duvéau, 2012).

Harposporium infection results in a differentially expressed gene signature overlapping with that of *D. coniospora* infection and various bacterial infections (*E. faecalis*, *P. luminescens* and *S. marcescens*) (Engelmann, et al., 2011).

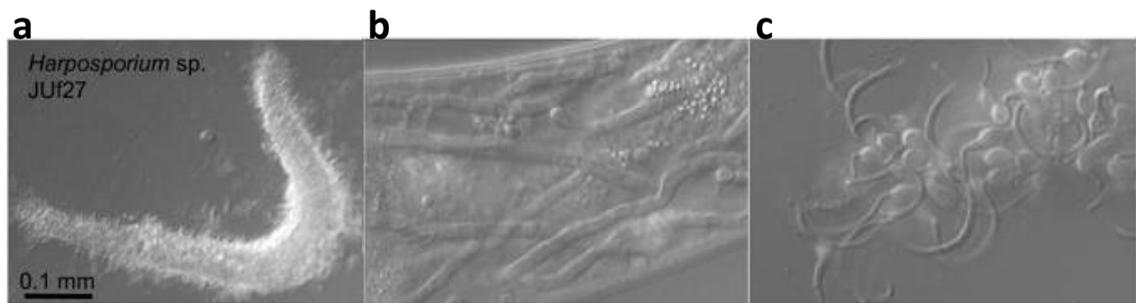


Figure 5.16 *C. elegans* infected with *Harposporium sp.* (JUf27). a) Infected *C. elegans* corpse with fungal hyphae protruding out of the host. b) Hyphae are invading the body of the worm. c) *Harposporium sp.* sickle-shaped conidia spore. Image taken from Felix & Duveau (2012).

5.5.3.4. *Candida albicans*

C. albicans is an opportunistic fungus that can grow vegetatively as yeast or hyphae, both of which play different roles in pathogenesis. The yeast form colonizes mucosal surfaces and spreads through the host bloodstream, while the hyphae form invades the host and destroys tissues (Pukkila-Worley, et al., 2011). It is the most common human fungal pathogen accounting for 70-90% of all invasive mycoses (Pukkila-Worley, et al., 2009). In *C. elegans*, the yeast form of *C. albicans* can be grown on solid agar media and ingested by the worms. The yeast form seems to easily survive the pharyngeal grinding since as little as 5 minutes of exposure to the yeast is enough to lead to infection. Transferring infected worms to (pathogen-free) liquid medium quickly leads to the death of the worms as fungal hyphae can be seen piercing through the worm's cuticle (**Figure 5.17**). The solid-liquid interface is crucial for the yeast form to develop filaments (filamentation) that then differentiate into hyphae. Filamentation is not observed on solid media alone and submerging the yeast itself into liquid media also does not lead to filamentation (Breger, et al., 2007; Pukkila-Worley, et al., 2009; Pukkila-Worley, et al., 2011).

The p38 MAPK pathway seems to play an important role in *C. albicans* resistance as *sek-1* mutants are more susceptible to the fungi (Breger, et al., 2007). Genetic expression analysis following *C. elegans* infection by *C. albicans* found many immune response genes down-regulated, that are commonly up-regulated by bacterial pathogen *P. aeruginosa* and *S. aureus*. Furthermore, both live and dead *C. albicans* elicit similar gene expression response, indicating that *C. elegans* immune response involves the recognition of heat-stable chemicals (Pukkila-Worley, et al., 2011).

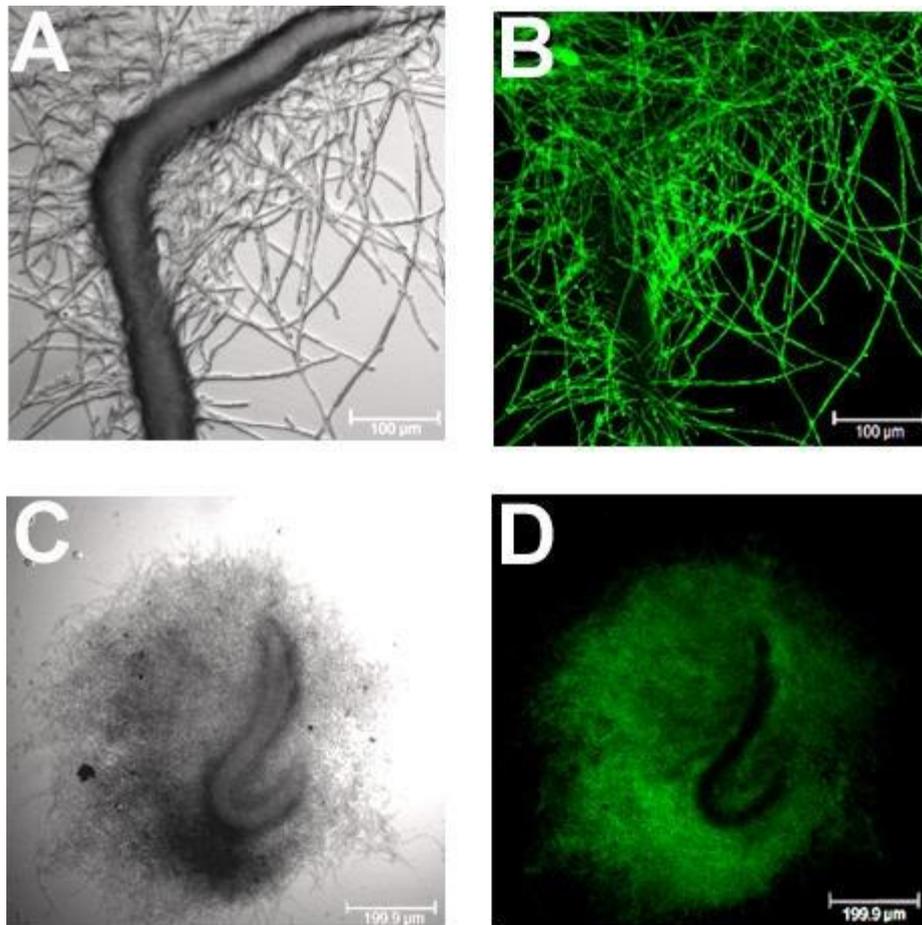


Figure 5.17 *C. albicans* emerging from dead *C. elegans* hosts. *glp-4;sek-1* double mutant worms were left on *C. albicans* populated solid media for 2 hours and then moved to pathogen-free liquid media. Images showing the growth progression of *C. albicans* after 6 days A) & B) and 8 days C) and D). B) & D) are fluorescent images with the fungal filaments stained with Concanavalin A-Alexafluor. Image taken from Breger, et al. (2007).

5.5.4. Orsay Virus

The Orsay virus is the first and only virus found that naturally infects *C. elegans* and has been isolated from *C. elegans* samples collected in 2011 in Orsay (France). The Orsay virus is a small positive-stranded RNA virus, most closely resembling Nodaviruses, that is specific to *C. elegans* and does not infect *Caenorhabditis briggsae*. The Virus is transmitted horizontally, probably via feeding and defecation. Viral RNA is mainly located in intestinal tissues, but has also been observed in the somatic gonad (Felix, et al., 2011; Franz, et al., 2014).

The innate immune response against the Orsay Virus includes antiviral defences such as the *rde* argonaute gene family responsible for RNA interference and breaking down viral RNA (Felix, et al., 2011). Transcriptomic analysis has further shown that the transcriptional response to Orsay virus is most similar to *N. parisii* (Yang, et al., 2016) and both pathogens are affected by the ubiquitin-proteasome response (Bakowski, et al., 2014). STA-1, the *C.*

elegans homolog of mammalian STAT (which is part of the JAK/STAT-pathway associated with immunity) has been inferred to repress resistance against Orsay virus, as infected *sta-1* mutants contain less viral RNA (Tanguy, et al., 2017).

5.6. Aim and objective

The aim of this study is to investigate the transcriptional relationship between the HSR and the innate immune response to identify potential overlapping or unique gene expression profiles and transcriptional regulators between these two responses. For this, I chose to perform a systematic analysis of publicly available high throughput gene expression datasets from the model organism *C. elegans* exposed to 22 different pathogens and heat stress. This organism shares conserved stress response pathways and organs (cuticle and intestine) with humans (Rodriguez, et al., 2013). The fact that *C. elegans* only possesses the more ancient and conserved innate immune system has the advantage that the results are only due to the innate immune response and not a combination of the innate and adaptive immune systems (Ermolaeva & Schumacher, 2014), thus allowing a direct comparison between the heat shock response and the innate immune response.

Chapter 6: Methods for the bioinformatic analysis of high-throughput datasets

6.1. Data selection

Datasets used in this part of the thesis were found through searching the NCBI database for publications regarding *C. elegans* pathogen or heat stress response that have publicly available raw data from transcriptome profiling technologies (RNA-seq and single-channel microarrays). In addition, the Gene Expression Omnibus database (Edgar, et al., 2002) and modENCODE database (Celniker, et al., 2009) were searched for deposited data not found through NCBI. Data with no adequate control or no replicates were not used. The complete list of datasets used can be found in **Appendix 5**.

6.2. Data processing

6.2.1. Affymetrix Microarray

The microarray data were analysed using the R package *limma* (Linear Models for Microarray and RNA-seq Data) (version 3.38.3) according to chapter 17 of the *limma* user guide (Ritchie, et al., 2015). The raw data was read into R using `rma()` and `ReadAffy()` from the *affy* package (version 1.60.0) (Gautier, et al., 2004), which simultaneously corrects for background intensity and normalizes the expression. The quality of the data was assessed using `qc()` from the *simpleaffy* package (version 2.58.0) (Miller, 2018). Data uniformity was checked with histograms and box plots in the R base package. Affymetrix IDs were converted to gene symbols using the `getSYMBOL()` function from the *annotate* package (version 1.50.1) (Gentleman, 2019). Affymetrix IDs that do not have a corresponding gene symbol were removed.

In cases where the publication did not mention whether the samples were collected in series or in parallel, a PCA plot was drawn using *limma* function `plotMDS()` to check how similar the samples were and help determine whether a paired analysis was required. A design matrix was constructed using `model.matrix()` from the *stats* package (version 3.5.2) (R Core Team, 2014) with treatment as a factor. In the case of a paired analysis, both the treatment and the experiment/replicate number were given as factors. For the differential expression

analysis using *limma*, the data were fit to a linear model with `lmFit()`, and statistical significance was computed using `eBayes()`. An MA plot using `plotMD()` was produced to visualize the fold change and average expression of significant genes.

6.2.2. Agilent Microarray

For Agilent microarray data, the data was read into R using `read.maimages()` from the *limma* package. The data uniformity was checked with a density plot and boxplot. The Agilent IDs were converted to gene symbols using a conversion list provided by the platform (Agilent Technologies, 2016). The data was then corrected for background intensity using `backgroundCorrect(,method="minimum")` and were normalized between the different arrays using `normalizeBetweenArrays(,method="quantile")`. A PCA plot was drawn using *limma* function `plotMDS()` to assess whether a paired analysis was required. Probes were filtered out if they did not have a gene symbol, Entrez ID, or if the signal intensity was below the background intensity.

The differential expression analysis was then done the same way as for Affymetrix data.

6.2.3. Nimblegen Microarray

Nimblegen format (`.calls` and `.pair`) is not supported by *limma*. The `.pair` data was converted to `.xys` datafile using a custom script (Carvalho, 2013). An annotation package was created using the `makepdInfoPackage()` function from the *pdInfoBuilder* package (version 1.46.0) (Falon, et al., 2018) by providing the `.xys` and `.ndf` data (the `.ndf` data is available from NCBI under the accession GPL16196), that was then installed using `install.package()`. The `.xys` files were read into R using `read.xysfiles()` from the *oligo* package (version 3.10) (Carvalho & Irizarry, 2010) and the newly installed annotation package. The data were then normalized using `rma()`.

The resulting expression set was then analysed the same way as for Affymetrix data.

6.2.4. RNA-seq Data

Raw RNA-seq data was downloaded in `.sra` format using `getSRAfile()` from the *SRADB* package (version 1.44.0) (Zhu, et al., 2013). These `.sra` data files were then converted to `.fastq` format using NCBI's `fastq-dump` (version 2.8.0). For paired-end RNA-seq data, the option `--split-files` was used. The raw data was checked for quality and adapter content

using FastQC (version 0.11.8) (Andrews, 2010) and adapter trimming was done, if necessary, using TrimGalore, a wrapper of Cutadapt (version 0.5.0) (Krueger, 2012). Reads (sequenced strands of nucleic acids) were aligned using STAR (version 2.6) (Dobin, et al., 2013), which required a pre-generated genome index. The index was generated using *C. elegans* genome and annotated transcripts from Ensembl release 95 (WBcel235) (Ensembl, 2019) with default `sJdbOverhang` of 100 (for read length 100bp or more). For read length below 100bp, a new genome index was generated corresponding to `sJdbOverhang` of 1 - read length. After Alignment, the sorted .bam files were filtered using SAMtools (Li, et al., 2009), flagging unmapped reads and reads that failed the quality control. A quality check using BamQC was also performed to check for uniform alignment of the reads (Andrews, 2013). A count matrix for the reads was then generated using featureCounts (Liao, et al., 2014). MultiQC (Ewels, et al., 2016) was used to summarize the results from various tools used in order to compare the quality of alignment within and between samples.

Differential expression analysis on the count matrix was done using the R package *DESeq2* (version 1.22.2) (Love, et al., 2014). The design formula for the function `DESeqDataSetFromMatrix()` included all factors initially. The resulting `DESeqDataSet` was then analysed using a PCA plot (`plotPCA()`) after transforming using `vst()`. Similarly to the microarray analysis, this plot is used to help determine whether a paired analysis was required. If it was not required, the `DESeqDataSetFromMatrix()` function was rerun, without including the samples/replicates as a factor. The wormbase ID of the `DESeqDataSet` was converted to Gene symbols using a conversion table generated from Ensembl Biomart (Ensembl, 2019). Wormbase ID without a gene symbol and very low count number (less than the number of samples) were removed. The datasets were then analysed using the `DESeq()` function. The results for the comparison of interest was extracted and subsequently transformed using `lfcShrink()` function with the “apeglm” method (Zhu, et al., 2018). An MA plot was generated using `plotMD()` to visualize the expression pattern of significant genes.

6.2.5. CHIP-seq Data

ChIP-seq data were downloaded from the GEO (Edgar, et al., 2002) and modENCODE databases (Celniker, et al., 2009) in bedgraph or bigwig form. These were visualized using IGV (Integrative Genomics Viewer) (Robinson, et al., 2011). If the input control track was relatively clean compared to the treatment track, then the treatment track was normalized against the control track using MACS2 `bdgcmp` function with the log-likelihood method

(Zhang, et al., 2008). On the other hand, if the input signal was stronger than the treatment track, then the input signal was not considered in the downstream analysis. MACS2 (Zhang, et al., 2008) peak call was then done on the tracks to find enriched binding sites. Min-length was set to 200, and the max-gap was set to 30. The cut-off was set to a threshold that produces an adequate number of hits, which was determined by manually sampling sections at various thresholds. The cut-off for each ChIP-seq dataset is given in the caption of **Appendix 14**.

In order to find the genes associated with the enriched binding site, the transcript start sites of all genes were acquired from Biomart (Ensembl, 2019) and 500 bp were added upstream and downstream of it, to generate a 1 kb window. The peak-called regions were intersected with the 1 kb transcript start site regions using BEDtools `intersect` function (Quinlan & Hall, 2010), to assign genes to the enriched binding site.

6.3. Meta-analysis/systematic review

6.3.1. Identifying common differentially expressed genes

The gene expression data from all pathogen response studies and HSR studies were compared separately to find commonly differentiated genes before finding the genes that respond to both pathogen infection and heat shock.

6.3.1.1. Venn diagrams

Venn diagrams were generated to intersect differentially expressed genes from different datasets, which were defined as having a p-value smaller than 0.05 and a $|\log_2|$ fold change greater than 0.6 (≈ 1.5 fold change). A maximum of 5 samples could be compared in a Venn diagram from the *VennDiagram* package. Since there were many more datasets than could fit in a single Venn diagram, datasets were compared to each other based on the evolutionary distance of the pathogens used to infect *C. elegans* (heat shock datasets were compared independent of the pathogen response datasets). The common tree from NCBI (NCBI, n.d.) was used to determine the evolutionary distance between pathogens.

6.3.1.2. Heatmap

The gene expression datasets from the pathogen studies were combined into a large matrix. Genes with an incomplete set of values (i.e. some datasets do not have a value for the gene)

were removed if they had 14 or more values missing. The reason why 14 was chosen was that out of all the datasets used, 14 came from the same microarray platform (Affymetrix). Since microarrays depend on the design of probes from known genes, older microarray chip designs may not have probes for genes identified recently, which were identified in the RNA-seq data. The remaining missing values were imputed. Imputation was done for each gene, by replacing the missing values with the average of the known values of the gene. Next, genes were filtered to retain only those that can be considered generally responsive (i.e. the gene is not just responsive to few specific datasets). Firstly, genes with overall low differential expression across all datasets were filtered out. For each gene, the log₂ fold change from each dataset was added together, and if the sum was below a certain threshold, the gene was removed (see the relevant section for the specific threshold). Secondly, only genes that were found to be significantly differentially expressed ($|\log_2|$ fold change larger than 0.6) in at least two datasets were kept. The requirement of at least two datasets would avoid false positives, where otherwise only one “outlier” would have been enough to consider a gene to be “commonly differentiated”. The last filtering criterion retained only the genes where the majority of datasets (see the relevant section for what is defined as “majority”) had a minimum log₂ fold change of 0.1 (or 0.3 for the heat shock dataset). This retained genes that were consistently differentially expressed, albeit to a low degree, while being robust against outlier data that would otherwise dismiss potentially interesting genes. Due to the large number of datasets used, the effect size can be small (i.e. log₂ fold change of 0.1) to still obtain a significant p-value (see **Appendix 8**). After all the filtering criterion to reduce the number of genes to a computationally feasible size, a heatmap was generated using `heatmap.2()` from the *gplot* package (version 3.0.1.1) (Warnes, et al., 2019). The same method with different filtering threshold was used for the heat shock datasets.

6.3.2. K-means clustering

Datasets were grouped⁴ based on expression similarity, using the K-means clustering function `Kmeans()` from the *amap* package (version 0.8-16) (Lucas, 2018). The number of groups to separate the datasets into was determined by calculating the within-groups sum-of-square for all possible number of groups and choosing the optimal number of groups using the elbow method (Ketchen & Shook, 1996). This method determines the best number based on: lowering the number of groups starts to see a comparatively larger loss in raw

⁴ Here the word group is used to refer to clusters, since the word cluster is used to refer to a different list of genes in the results and discussion section.

information (larger within-groups sum-of-square), indicating that dissimilar datasets have been merged while increasing the number of groups does not model the data much better as the marginal gain in precision decreases (within-groups sum-of-square decreases less).

6.3.3. Determining p-value of filtering Criteria

The matrix of all differential expression values of all datasets was converted to a list, which was then used to find the median value (μ), mean deviation from the median (τ), mean and standard deviation. A histogram was drawn using all the values in the list, and the Laplace and normal density distribution were superimposed on the histogram to identify which probability distribution fits better to the actual data. The Laplace density curve was drawn using `dlaplace()` from the *jmuOutlier* package (version 2.2) (Garren, 2019) with the variables μ and τ . The normal distribution was drawn using `dnorm()` from the *stats* package (version 3.5.2) (R Core Team, 2014) with the mean and standard deviation.

The p-value (for the filtering criteria) was calculated based on finding the probability of picking at least X number of differential expression values above the threshold log2 fold change value Y using Z number of trials. The probability of successfully picking a value at random above the threshold log2 fold change value was calculated using `plaplace()` from the *jmuOutlier* package (version 2.2) (Garren, 2019) with the variables μ and τ .

For example, the probability of picking a log2 fold change value greater than 0.1 from the Laplace distribution with μ and τ is 0.56. The probability of this occurring at least 23 times out of 25 trials follows the binomial distribution with the formula:

$$\sum_{k=23}^{25} \binom{25}{k} \times 0.56^k \times (1 - 0.56)^{25-k} = 0.0001$$

In this example, the probability of genes passing the filtering criteria would be 0.0001 or smaller.

6.3.4. Gene Enrichment Analysis

GO term analysis was done initially on the whole list of differentially expressed genes as well as smaller cluster of genes within the list. The analysis incorporates a number of tools. Significant GOterms, pathways, protein domains and other commonalities were identified using Wormbase GSEA (Angeles-Albores, et al., 2016), g:profiler (Raimand, et al., 2007) and

StringDB (Szkarczyk, et al., 2019). Protein-protein interaction networks were generated using StringDB. Enriched GOterm networks were generated using BiNGO (Maere, et al., 2005) and ClueGO (Bindea, et al., 2009).

The smaller clusters of genes were generated by using the dendrogram generated as part of the `heatmap.2()` function using the default method of complete linkage using the Euclidean distance measure. The dendrogram was cut-off at the point that would result in at least three large clusters representing up-regulated, down-regulated and mixed-regulated genes. These clusters were also individually analysed.

6.3.5. TF binding around the transcript Start Site (SeqPlot)

The average ChIP-seq signal around transcript start sites was visualized using SeqPlot (Stempor & Ahringer, 2016). The tracks used were the ChIP-seq data of PQM-1 (Niu, et al., 2011) and DAF-16 (modENCODE ID: 591). The features dataset used were the .bed file of all *C. elegans* protein-coding transcript start site from the annotation version WBcel215 and WBcel235 for ce10 and ce11, respectively. The type of plot was set to “point feature”, and the plotting distance was set to 1000 bp up- and downstream.

6.3.6. Hypergeometric test

For comparison between two gene lists, the hypergeometric distribution test was done in R using the `phyper()` function from the *stats* package (R Core Team, 2014), with `lower.tail` set to `FALSE`:

*phyper(size of overlap – 1, size of list 1, total number of genes
– size of list 1, size of list 2, lower.tail = FALSE, log.p = FALSE)*

The total number of genes is the number of unique protein-coding genes among all datasets used (union of genes in all datasets).

For comparison and visualization of multiple gene lists and their overlaps, the `supertest()` function from the *SuperExactTest* package (version 1.0.7) (Wang, et al., 2015) was used.

6.4. Motif enrichment

Motif and TF enrichment analysis was done using HOMER (Heinz, et al., 2010) with the worm promoter data v5.5 and *C. elegans* genome version ce11.

de novo motif discovery was done using BMM (Siebert & Söding, 2016), DREME (Bailey, 2011) and Trawler (Ettwiller, et al., 2007). The input sequences were the promoter regions, defined as 500bp upstream and downstream of the transcript start site. The list of transcript start site was downloaded from Ensembl BioMart, (Ensembl, 2019), filtered for the genes of interest and 500bp were subtracted or added to the start and end position respectively. The file was converted to bed format and sequences were extracted using *BEDTools* `getfasta` (Quinlan & Hall, 2010) and the ce11 reference genome fasta file. For DREME, the sequences were submitted under default settings. For BMM, the MMcompare Motif Database was set to JASPAR2018. For Trawler, the organism was changed to *Caenorhabditis elegans* (ce10), Motif database set to nematodes and analysis was done for both the single and double-strand option.

Chapter 7: The association of the heat shock and innate immune response

Stress responses have been classically separated and studied by the nature of the stressor such as temperature (physical), reactive oxygen species (chemical) and pathogens (biological). Rather than viewing the response against different stressors as separate mechanisms, they might be interconnected to an extent previously unknown. To this end, the work presented here aims to analyse the relationship between the two distinct stress responses, the temperature-dependent stress response (HSR) and response to pathogen infection (innate immune response) in *C. elegans*. The approach concentrates on the change in the transcriptional landscape following the exposure to the stressor to identify key differentially expressed genes. For this, many high-throughput screening data are analysed to identify genes responsive to pathogen infection and genes responsive to heat shock. Finally, the genes responsive to each of the stressor are analysed and compared to determine the connection between the response to heat shock and pathogen infection.

A systematic review of the available high-throughput screening data is an important step in validating and answering various scientific questions. The large number of available *C. elegans* datasets, especially with regards to pathogen infection, can provide strong statistical power and a better overview of the question at hand. A comparison of the differentially expressed genes between various published pathogen datasets has previously been carried out by Yang et al. (2016). However, here for the first time, the published datasets are reanalysed with the same software and same conditions, creating a unified and consistent database of differentially expressed genes, which makes direct comparisons between different datasets more reliable.

7.1. Data used and quality control

The data used in this study came from 23 publications with 32 datasets as well as one dataset generated internally by Dr Laura Jones. Pathogen infection response data is comprised of 29 datasets, 19 of which are microarray data and ten are RNA-seq data. For the HSR, three publicly available RNA-seq datasets were used in addition to the dataset generated by Dr Laura Jones. The full list of datasets used here, with experimental details and platform can be found in **Appendix 5**.

7.1.1. Microarray data

Quality control was performed for each of the 19 microarray datasets. An example is shown in **Figure 7.1** and includes density estimates (**Figure 7.1a & b**), principal component analysis (PCA) plot (**Figure 7.1d**), mean difference plot (MD plot) (**Figure 7.1e**). For Affymetrix data, an additional QC stats plot from the *simpleaffy* package was done (**Figure 7.1c**).

Many publications do not state whether the samples were collected from the same or different experiments. This is important in order to determine whether the data should be analysed in pairs or not. In some cases, a PCA plot was helpful in determining the statistical model, when samples with the same naming scheme are separated along one of the principal components axis. In the example, this is seen where the replicates are separated along the y-axis, while the control (OP50) and treatment (PA14) groups are separated along the x-axis (**Figure 7.1d**). Separation of replicates could indicate a batch effect, inferring inconsistencies in the execution of the experiment (E.g. experiments done by different people or on different days). Cases where a PCA plot was insufficient to determine the statistical model were analysed by comparing the number of significantly differentially expressed genes between a paired and unpaired analysis. The p-value of a differentially expressed gene is dependent on both the fold change and the statistical variation. When analysing in pairs, the samples within each pair are compared among each other before the results from all pairs are pooled together (W. W. B. Goh, et al., 2017; Smyth, et al., 2019)⁵. As such, when a large variation between replicates exists, a paired analysis tends to result in stronger statistical power, as the absolute variation between samples in different pairs is avoided and only the relative gene expression change (fold change) between each pair is compared. In cases where replicates do not vary much, the statistical power does not differ much between a paired and unpaired analysis, as the inter-pair variability does not differ much. Thus, when the MA plot shows a much larger number significantly differentially expressed genes in a paired analysis compared to an unpaired analysis, the result of the paired analysis is used.

⁵ Here the word “pair” is used instead of “batch” or “block”, as each batch/block is comprised of a pair of samples: one treatment sample and one control sample.

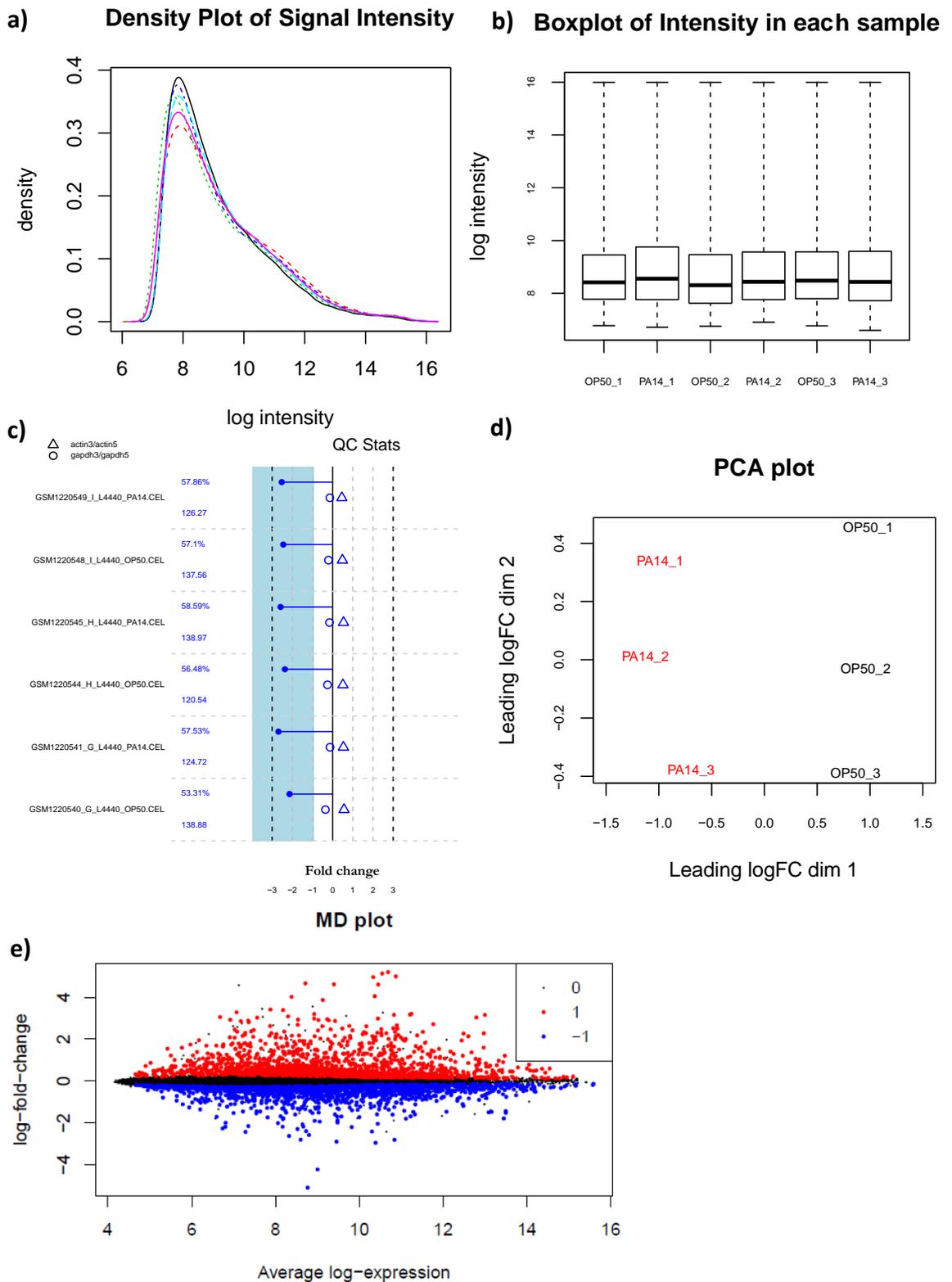


Figure 7.1 Representative quality control using the dataset from Estes et al. 2010. a) & b) Plot of signal intensity distribution and average intensity as a density plot and a boxplot, respectively. c) QC stat plot from the *simpleaffy* package. Good arrays are represented in blue, while bad arrays would be coloured red. The line emerging from the 0-fold line should end up in the light blue area to indicate compatible scale-factors. Actin control (triangle) should be within 1-fold-change, and GAPDH (open circle) should be within 3-fold-change (Miller, 2018). d) PCA plot of the dataset. e) MD plot of the differential expression analysis, using a paired statistical model. Red dots are significantly up-regulated genes (denoted “1” in the legend), and blue dots are significantly down-regulated genes (denoted “-1” in the legend). Black dots are not significantly differentially expressed (denoted “0” in the legend).

7.1.2. RNA-seq data

The quality control for the 14 RNA-seq datasets includes the *featureCounts* summary, STAR alignment score, *Cutadapt* report and *FastQC* report (an example of such a quality control can be found in **Appendix 6**). A PCA plot was often enough to determine whether a paired analysis was required. The paired analysis result for Osman, et al. (2018), K. Chen, et al. (2017) and ModENCODE *H. sphaerosporum* datasets were used for further analysis due to the stronger statistical power and a higher number of differentially expressed genes.

Due to the extremely low differential expression of some of the datasets (**Table 7.1** and **Figure 7.2**), further quality controls were conducted. Histogram plots of the differential expression of all genes showed that the global distribution of log₂ fold changes in some datasets did not show a continuous probability distribution, but rather concentrate at one point (**Figure 7.2a**). I was able to identify the cause to be the function `lfcshrink()` in the DEseq2 package, which shrinks the log₂ fold change based on the statistical power (such as the standard error) so that the final log₂ fold change value is adjusted to the consistency of the data. The datasets most strongly affected by this shrinkage were Osman, et al. (2018), K. Chen, et al. (2017) and ModENCODE *H. sphaerosporum*, mainly due to their relatively large log fold change standard error (lfcSE), indicating that the replicates have large variations (see **Figure 7.2b-d** for an example). These datasets were not omitted at this stage, as they still contain a small number of differentially expressed genes that may be important.

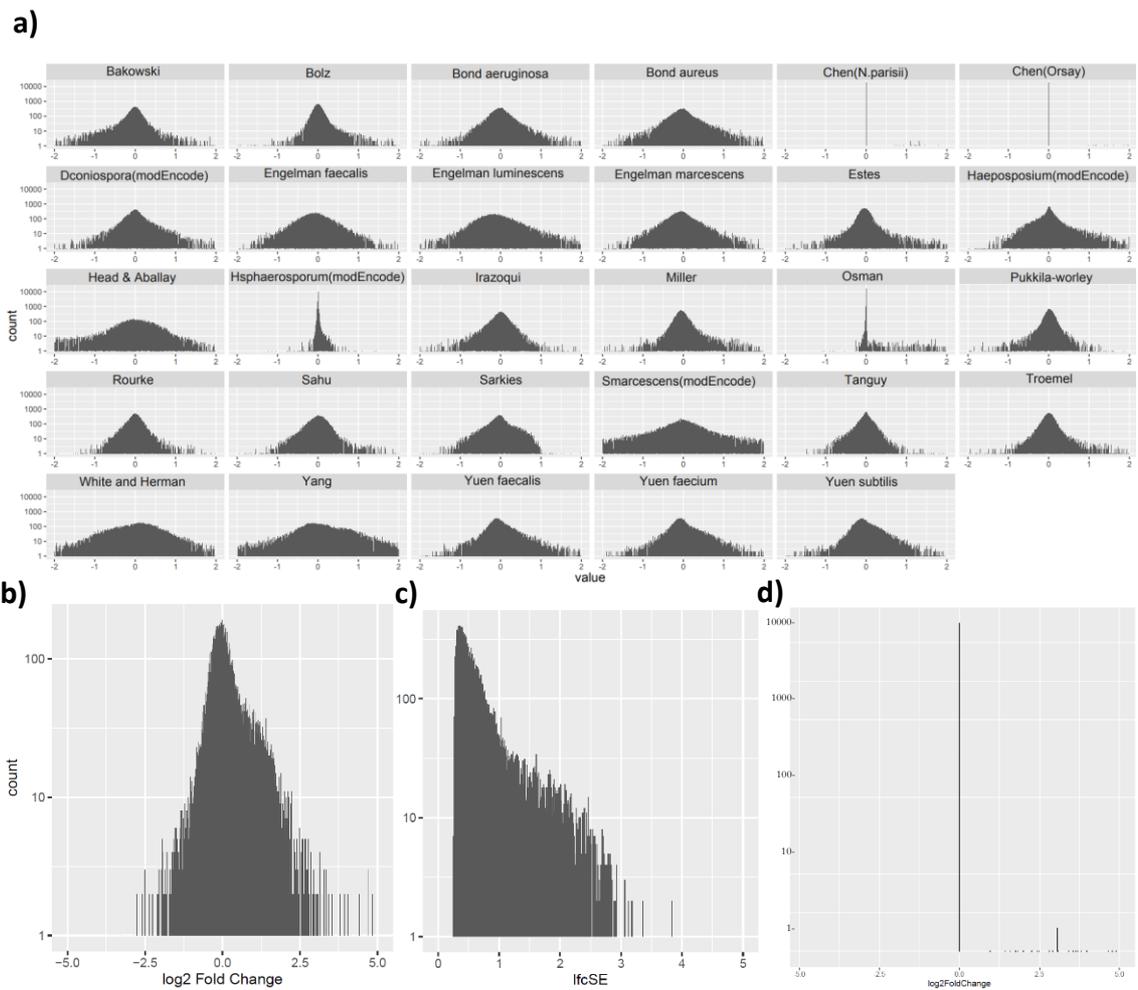


Figure 7.2 Histogram plot of the log₂ Fold Change distribution for all genes in each dataset. a) log₂ Fold Change distribution for all genes in each of the 29 pathogen datasets. b) - d) Chen (*N. parisii*) dataset b) log₂ Fold Change distribution before applying `lfcshrink()`, c) log fold change standard error (`lfcSE`), and d) log₂ Fold Change distribution after `lfcshrink()`.

7.2. Change of the gene expression landscape as a result of pathogen infection

The resulting number of significantly differentially expressed genes for each of the pathogen response datasets are summarized in the table below. For microarray data, the results were generated using *limma*, and RNA-seq data were analysed using *DEseq2* (Table 7.1).

Pathogen	Reference	Up-regulated genes	Down-regulated genes
<i>Pseudomonas aeruginosa</i>	Troemel, et al. (2006)	421	192
	Bond, et al. (2014)	483	371
	Estes, et al. (2010)	526	354
	Miller, et al. (2015)	491	189
<i>Staphylococcus aureus</i>	Irazoqui, et al. (2010)	89	156
	Bond, et al. (2014)	877	457
<i>Enterococcus faecalis</i>	Engelmann, et al. (2011)	15	12
	Yuen & Ausubel (2018)	435	220
<i>Microbacterium nematophilum</i>	O'Rourke, et al. (2006)	199	200
<i>Yersinia pestis</i>	Bolz, et al. (2010)	408	88
<i>Salmonella enterica</i>	Head & Aballay (2014)	1249	1467
<i>Bacillus subtilis</i>	Yuen & Ausubel (2018)	258	180
<i>Enterococcus faecium</i>	Yuen & Ausubel (2018)	355	208
<i>Stenotrophomonas maltophilia</i>	White & Herman (2018)	61	66
<i>Vibrio cholerae</i>	Sahu, et al. (2012)	324	462
<i>Photorhabdus luminescens</i>	Engelmann, et al. (2011)	125	136
<i>Serratia marcescens</i>	Engelmann, et al. (2011)	84	13
<i>Orsay Virus</i>	Sarkies, et al. (2013)	29	4
<i>Candida albicans</i>	Pukkila-Worley, et al. (2011)	1551	239
<i>Nematocida parisii</i>	Bakowski, et al. (2014)	475	784
	K. Chen, et al. (2017)	24	0
<i>Orsay virus</i>	Tanguy, et al. (2017)	159	227
	K. Chen, et al. (2017)	32	0
<i>Bacillus thuringiensis</i>	Yang, et al. (2015)	4036	1941
<i>Myzocytiopsis humicola</i>	Osman, et al. (2018)	356	5
<i>Drechmeria coniospora</i>	ModENCODE	1061	556
<i>Harposporium sp.</i>	ModENCODE	1069	697
<i>Serratia marcescens</i>	ModENCODE	2135	2180
<i>Haptocillium sphaerosporum</i>	ModENCODE	7	3

Table 7.1 Summary of the number of significantly differentially expressed genes of all 29 pathogen infection datasets. Significantly differentially expressed genes are defined as genes having a $|\log_2FC| > 0.6$ and an adjusted p-value < 0.05 . The non-shaded area is microarray data. Grey shaded area indicates RNA-seq data.

From the 29 datasets analysed here, the microarray data show a relatively consistent number of differentially expressed genes, especially with the same platform, mainly in the range of hundreds. RNA-seq data, on the other hand, vary significantly from single digits to thousands of differentially expressed genes (**Table 7.1**). While microarrays show higher consistency, their disadvantage lies in the dependence of probes. Many of the microarray chips are relatively old and do not encompass all the currently known genes to date. Furthermore, the probes depend on specific sequences of the gene and might not be able to detect the expression of transcripts with alternative splicing or different transcriptional start sites. This becomes biologically significant when different transcripts of the same gene are expressed under different conditions. RNA-seq, on the other hand, sequences each transcript of a given size and through alignment software, tries to identify each of the transcripts. Therefore, RNA-seq can capture a larger range of different transcripts and generally paints a more accurate picture of the gene expression landscape, but at the cost of larger variations. Such large variation was significant in four datasets (the ModENCODE *H. sphaerosporum*, Osman et al. (2018) and the two Chen et al. (2017) datasets) where the log fold change standard error (lfcSE) was larger than the log fold change itself, resulting in the shrinkage of most of the signal to zero after using the `lfcshrink()` function (**Figure 7.2**).

7.2.1. Pathogen infection datasets show low overlapping up- and down-regulated genes

Concentrating on the microarray datasets first we can see that there is a relatively large variation in the number of significantly differentially expressed genes between different pathogens as well as within the same pathogen (but conducted by different research groups). For example, *Staphylococcus aureus* only has 89 up-regulated genes in the dataset from Irazoqui et al. (2010), while Bond et al. (2014) has 877 up-regulated genes. This difference in results could be due to differences in experimental design such as *C. elegans* strains used, temperature and infection efficiency. Irazoqui et al. (2010) used a temperature-sensitive sterile strain and conducted the experiment at 25°C on young adult worms, while Bond et al. (2014) used L4 stage wild-type worms at 18°C. It is difficult to determine which experimental design is more precise, as both have their advantages and disadvantages and were designed to look at specific aspects of the pathogen infection response.

Comparison of the up- and down-regulated genes within the same pathogen shows some variation. The four *P. aeruginosa* datasets, for example, have a relatively consistent number of up- and down-regulated genes, but only 87 up-regulated (**Figure 7.3a & b**) and 31 down-

regulated genes are in common between the four datasets (**Figure 7.3c & d**). *S. aureus* has 59 and 42 commonly up- and down-regulated genes, respectively (**Figure 7.3e & f**). While these numbers may appear small, hypergeometric testing shows that the overlap is very significant.

When comparing between different pathogens, the overlap between up- and down-regulated genes becomes drastically reduced. Starting with the comparison of most closely related pathogens (based on the Entrez Taxonomy Database) in the Terrabacteria phylum (*S. aureus*, *M. nematophilum*, *E. subtilis*, *E. faecalis* and *E. faecium*) only two up-regulated (**Figure 7.4a**) and one down-regulated gene are shared (**Figure 7.4b**). For the pathogens of the Enterobacteriales order (*Y. pestis*, *S. enterica*, *S. marcescens* and *P. luminescens*), six up-regulated (**Figure 7.4c**) and no down-regulated genes are shared (**Figure 7.4d**). The remaining three pathogens: *S. maltophilia*, *P. aeruginosa* and *V. cholerae* are very distantly related and share eight up-regulated (**Figure 7.4e**) and no down-regulated genes (**Figure 7.4f**). Hypergeometric testing for each of the comparisons (**Appendix 7**) shows that the observed overlap is higher than the expected overlap for all instances (except for comparisons with zero overlapping genes). This indicates that a small group of genes are enriched under various pathogenic infection.

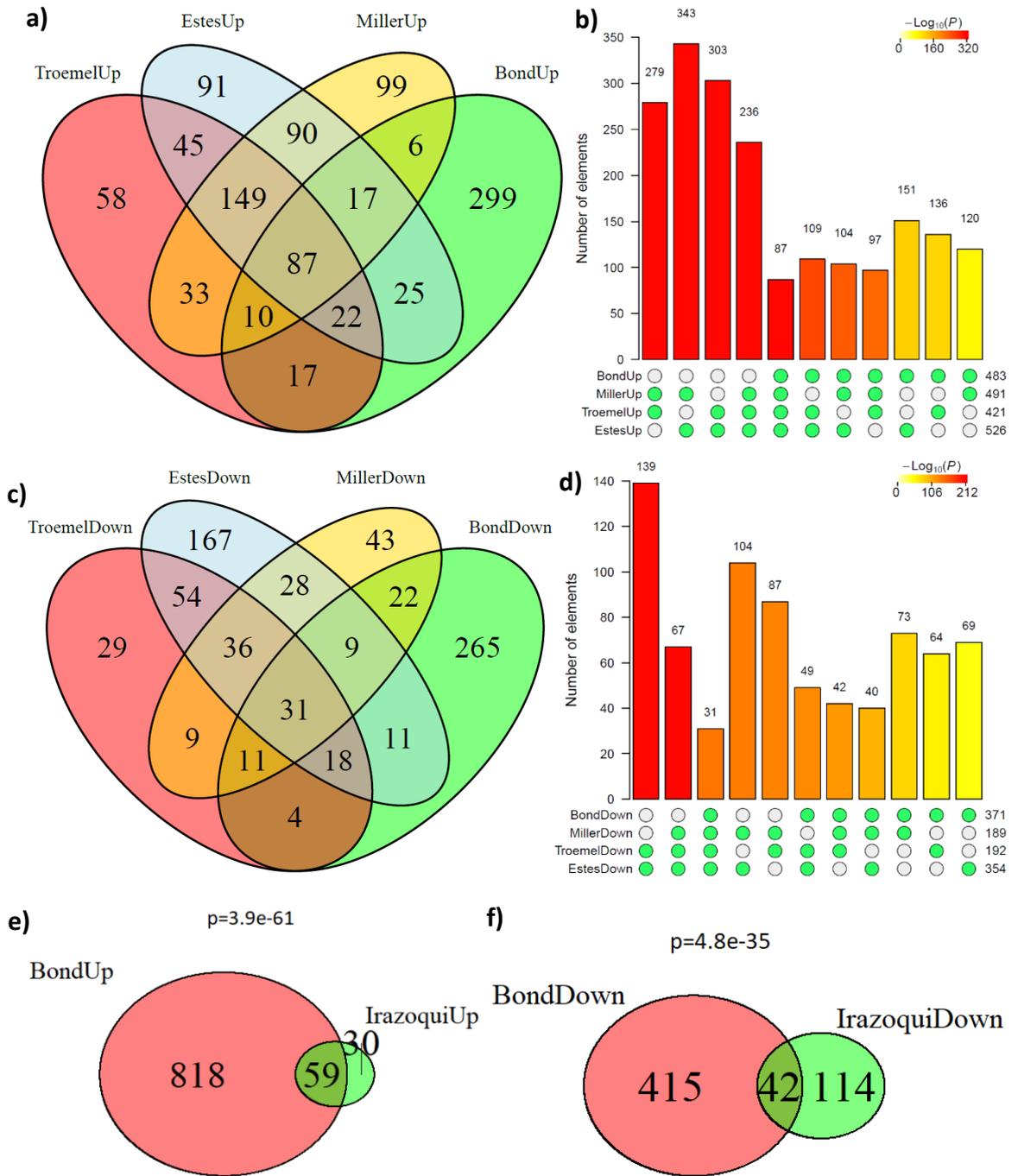


Figure 7.3 Comparison of significantly differentially expressed genes from microarray datasets using the same pathogens. a) - d) Venn diagram and hypergeometric test of *P. aeruginosa* dataset of up- and down-regulated genes respectively. The legend in b) & d) indicate the p-value (larger number means more significant). The number above each column shows the number of shared genes. The redder the colour of the column, the more significant the overlap of the genes are. e) & f) *S. aureus* infection datasets up- and down-regulated genes, respectively.

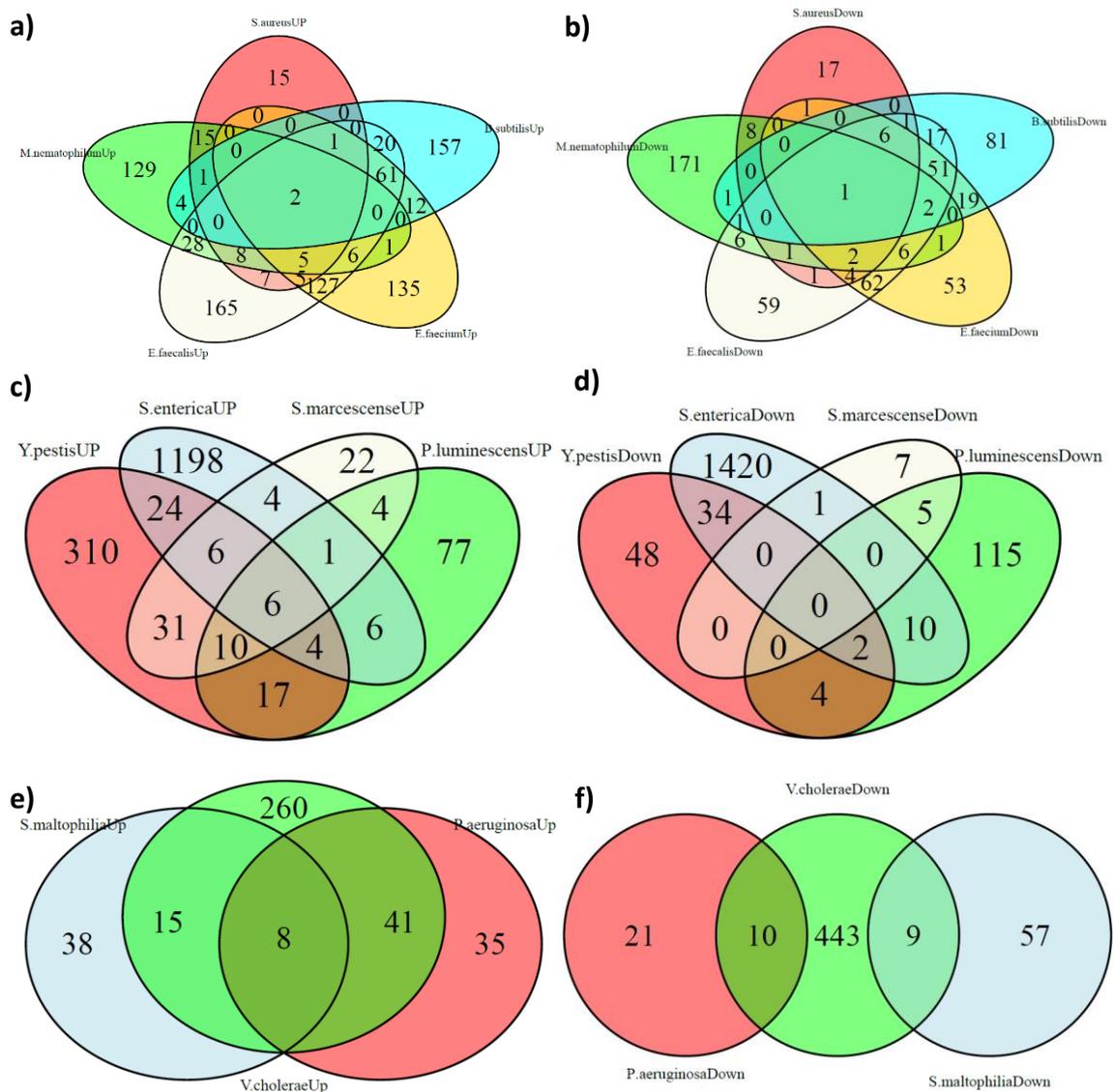


Figure 7.4 Venn diagram of significantly differentially expressed genes of closely related pathogens from the microarray datasets. a) & b) Venn diagram of pathogens from the Terrabacteria phylum: *S. aureus*, *M. nematophilum*, *E. subtilis*, *E. faecalis* and *E. faecium* up-regulated and down-regulated genes respectively. c) & d) Venn diagram of pathogens from the Enterobacteriales order: *Y. pestis*, *S. enterica*, *S. marcescens* and *P. luminescens* up-regulated and down-regulated genes, respectively. e) & f) Venn diagram of remaining pathogens *S. maltophilia*, *P. aeruginosa* and *V. cholerae* up-regulated and down-regulated genes respectively. Hypergeometric test for each of the comparisons can be found in Appendix 7.

RNA-seq datasets (except K. Chen, et al. (2017) due to the low number of differentially expressed genes) were also compared in the same manner as the microarray datasets by generating Venn diagrams comparing bacterial pathogens (**Figure 7.5a & b**) and fungal pathogens (**Figure 7.5c & d**). Like the microarray datasets, not many of the up-regulated and down-regulated genes are shared between different pathogens.

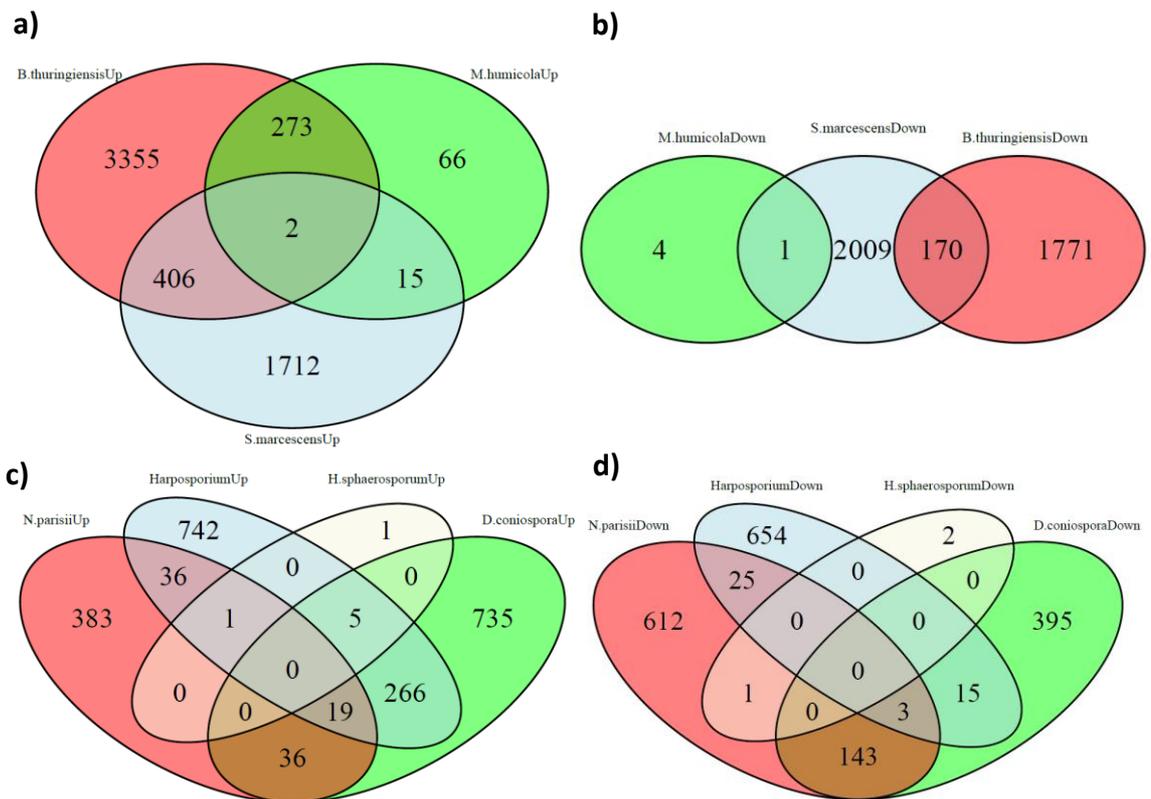


Figure 7.5 Venn diagram of significantly differentially expressed genes of pathogens in the same domain from the RNA-seq datasets. a) & b) Venn diagram of bacterial pathogens up-regulated and down-regulated genes, respectively. c) & d) Venn diagram of fungal pathogens up-regulated and down-regulated genes, respectively.

The method of finding common differentially expressed gene lists is very stringent, as it requires the gene of all intersecting groups to have a minimum fold change and p-value in order to count as a pathogen response gene, which due to the inherent variability of biological experiments is difficult to keep consistent (i.e. this method is not robust against outlier). Especially the RNA-seq datasets suffer from this due to the much larger variation in the number of differentially expressed genes compared to the microarray datasets. To overcome this weakness, a different method of finding commonly expressed genes needs to be used that can make use of the many datasets to preserve high statistical power while reducing the stringency of the cut-off threshold. Such a method is described in more detail in the next section (**Section 7.2.2**).

7.2.2. A small set of genes is consistently differentially expressed under various pathogens

The Venn diagram intersection method of finding shared genes was not able to return a list of genes that would be considered “general” pathogen response genes. Furthermore, Venn diagrams are not able to visualize all 29 datasets at once and become highly complex beyond

five datasets. Therefore, a new way of visualizing the results and a more dynamic filtering approach is required that makes use of the large sample size (29 datasets) and is more robust against outliers, i.e. it does not miss a gene due to only one or a few “outlier datasets”.

Heatmaps are a suitable way to visualize the differential gene expression landscape for a relatively large number of genes and datasets. However, heatmap generation is computationally intensive for a large number of data points and cannot feasibly visualize the whole *C. elegans* genome. Therefore, the number of genes to be included in the heatmap had to be reduced by filtering out genes whose expression did not change much overall. To achieve this, multiple filtering criteria were used to reduce the number of genes.

The first filtering criteria removed genes for which more than half (15 or more) of the datasets do not have a measurement. This number was chosen because 14 of the experiments were done on the same Affymetrix microarray chip, which at the time it was designed (2002), did not have all the currently annotated genes (WBcel235). As such, if a gene were to be missing on these Affymetrix chips, it would automatically result in 14 missing measurements for that gene, which was still measured by RNA-seq and other microarray platforms. The remaining missing values for each gene were imputed by using the average of values from existing measurements.

The second filtering criterion removed genes that on average show a $|\log(2)|$ fold change less than 0.1. In this case, genes were removed when the *absolute sum of log2FC* < 2.9 across all datasets (since there are 29 datasets, the absolute sum needs to be greater than 2.9 to reach the average 0.1 fold change threshold). Genes that were filtered out by this method showed very little differential expression in response to pathogen infection in general.

The third filtering criterion required the gene to be strongly differentially expressed in at least two samples. The threshold for strongly differentially expressed was defined as $|\log2FC| > 0.6$. This criterion filtered out genes that would only be strongly differentially expressed in none or only one of the 29 datasets, which could indicate an outlier. Having a gene significantly differentially expressed in multiple datasets would indicate more of a general pathogen response gene.

The fourth filtering criterion kept only the genes that had $|\log2FC| > 0.1$ in at least 24 datasets. The value for this criterion was determined by the number of genes left after the previous three filtering steps, as this criterion is very stringent. This criterion makes use of pooling the 29 experiments together to act as a large sample size, which can return significant genes even if the effect size is small. While the threshold of 0.1 may seem small, reaching

this threshold in 24 out of the 29 has a low probability of 0.002, based on hypergeometric testing using Laplace distribution (**Appendix 8**).

After these filtering criteria, the number of genes was reduced from 25952 to 331 (see **Appendix 20** for the gene list). A heatmap was drawn that clusters the genes and experiments based on complete linkage using Euclidean distance measure (**Figure 7.6**).

The heatmap shows a mixture of signals with no clear clustering of genes. However, the left side of the heatmap seems to contain genes that are generally stronger differential expressed (both up-regulated (red) and down-regulated (blue)) compared to the rest. The RNA-seq samples that previously showed a very low number of differentially expressed genes (Osman et al. (2018), Chen et al. (2017) and ModENCODE *H. sphaerosporum*) (**Table 7.1**), also show low differential expression in here, represented as a mostly continuous yellow horizontal bar in the middle of the heatmap.

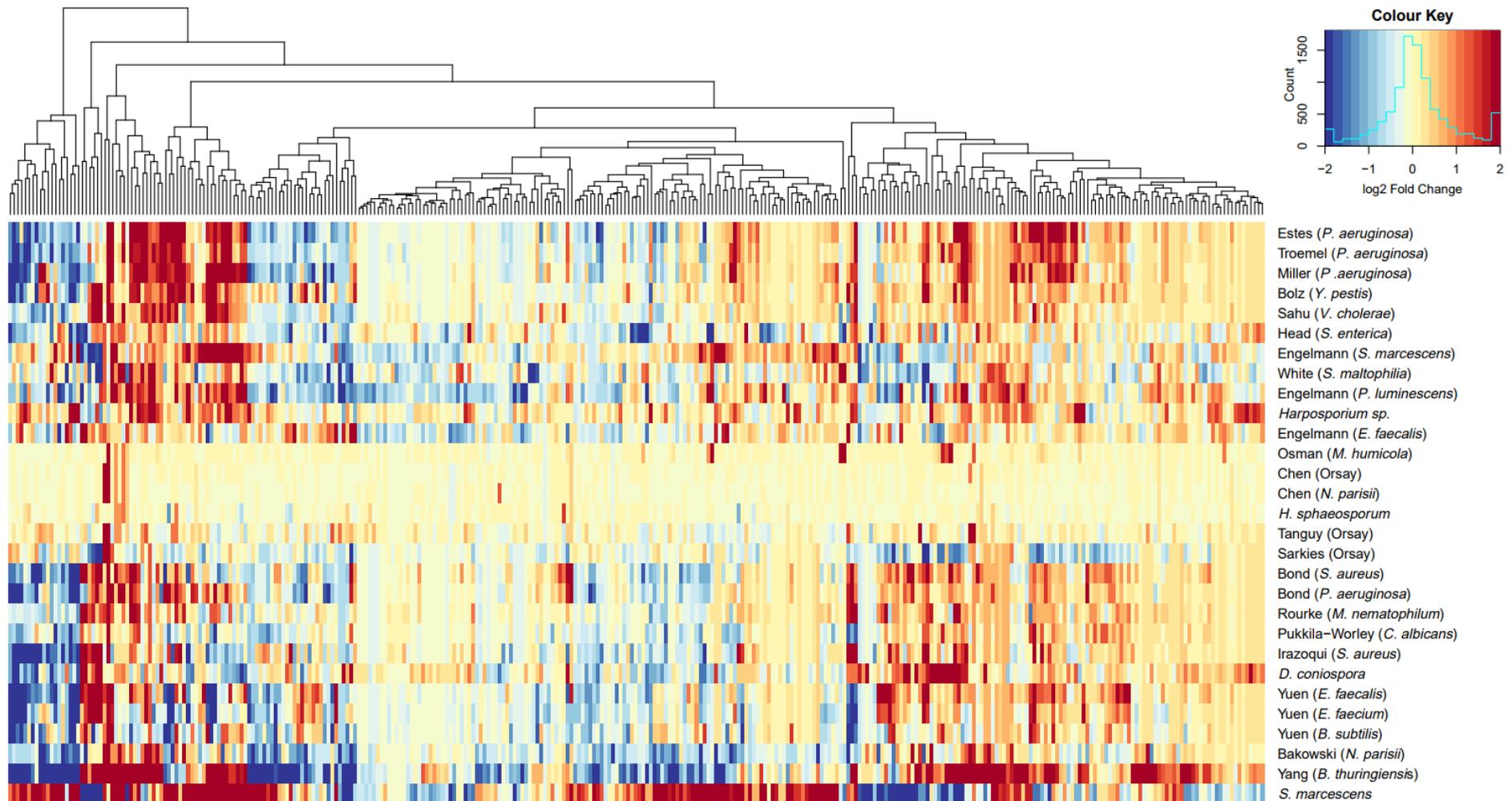


Figure 7.6 Heatmap of the 331 filtered genes from all 29 datasets. The genes are placed along the x-axis while the datasets are on the y-axis. The datasets are named based on their first author publication and the pathogen name, or in case of modENCODE data, only the pathogen name. The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes in the dataset, while blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The Colour key also shows a density plot.

Due to the mixture of signals and absence of clear clustering, datasets were grouped together based on their gene expression response towards pathogen infection, with the aim of reducing the number of datasets (rows) to make the heatmap clearer and better to analyse. The grouping was done using K-means clustering on the 331 filtered genes, where the 29 datasets were reduced to 7 groups. This number of groups was chosen because further reducing the number of groups starts to see a steeper loss in raw information (increased within groups sum of squares) (**Figure 7.7**), indicating that dissimilar datasets have been merged. On the other side, increasing the number of clusters does not model the data much better as the marginal gain in precision (loss of within groups sum of squares) decreases. This method of interpretation is known as the elbow method (Ketchen & Shook, 1996).

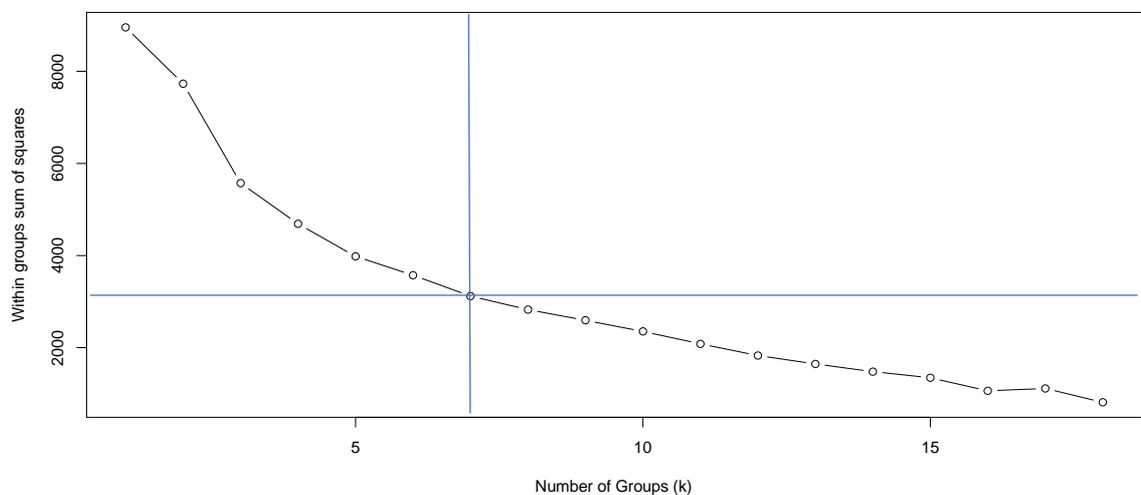


Figure 7.7 Within-groups sum of squares of the datasets by k-means clustering. The k-means clustering was run on the 29 datasets for a range of k (number of groups) from k = 2 to k = 18. The within-group sum of square was calculated for each value of k. The blue lines intersect at k = 7 and are guidelines to compare the steepness of the line before and after this point.

The groupings by k-means clustering can be seen in **Table 7.2**. Most of the time, the same pathogens are grouped together, except *N. parisii* and Orsay virus, which are both found in group 5 but also group 1 and 6, respectively. This can be explained by the fact that group 5 contains datasets with low differential expression (**Table 7.1**), in part due to the `lfcshrink()` function (refer back to **Section 7.1.2**), thus grouping them together. The six fungal pathogen datasets do not cluster together and are spread into five different groups indicating that, although classified into the same domain, they infect *C. elegans* through a wide array of mechanisms eliciting different gene expression response.

The clustering of group 5 indicates that the presence of datasets with very low differential expression may negatively affect the clustering algorithm as the clustering may not reflect biological similarity due to the absence of distinct signal patterns. One way to resolve this

issue is to avoid using the `lfcshrink()` function from DEseq2; however, this would make the analysis less accurate as the data would contain more false-positive gene hits. A more conservative solution is to omit the dataset with very low differential expressed genes compared to the other datasets as observed from the histograms in **Figure 7.2**.

1	Bakowski <u>Nematocida parisi</u>	Pukkila-Worley <u>Candida albicans</u>	Yang <i>Bacillus thuringiensis</i>		
2	Engelman <i>Enterococcus faecalis</i>	Yuen <i>Enterococcus faecalis</i>	Head <i>Salmonella enterica</i>	Yuen <i>Enterococcus faecium</i>	
3	Troemel <i>Pseudomonas aeruginosa</i>	Bond <i>Pseudomonas aeruginosa</i>	Miller <i>Pseudomonas aeruginosa</i>	Estes <i>Pseudomonas aeruginosa</i>	Bolz <i>Yersinia pestis</i>
	Irazoqui <i>Staphylococcus aureus</i>	Bond <i>Staphylococcus aureus</i>	Rourke <i>Microbacterium nematophilum</i>	Sahu <i>Vibrio cholerae</i>	modENCODE <u>Drechmeria coniospora</u>
4	Engelman <i>Serratia marcescens</i>	modENCODE <i>Serratia marcescens</i>	modENCODE <u>Harposporium sp.</u>	Engelman <i>Photorhabdus luminescens</i>	
5	Chen <u>Nematocida parisi</u>	Chen Orsay virus	Tanguy Orsay virus	Osman <u>Myzocytiopsis humicola</u>	
6	Sarkies Orsay virus	White <i>Stenotrophomonas maltophilia</i>			
7	modENCODE <u>Haptocillium sphaerosporum</u>	Yuen <i>Bacillus subtilis</i>			

Table 7.2 Grouping of the 29 datasets into 7 groups via k-means clustering. Fungal pathogens are underlined, and viral pathogens are written in bold. The dataset name (first author of publication or modENCODE) is written in red. Pathogens are colour coded when it exists more than once.

The data were reanalysed the same way as previously, but excluding the Osman et al. (2018), modENCODE *Haptocillium sphaerosporum* and the two Chen et al. (2017) datasets. This leaves 25 datasets with a large distribution of differentially expressed genes. The fourth filtering criterion was altered to accommodate this change (23 samples must have $|\log_2FC| > 0.1$). This criterion is more stringent with an equivalent p-value cut-off of 0.0001 (**Appendix 8**). A final filtering criterion was included which removes genes for which the adjusted p-value is larger than 0.05. As each differentially expressed gene in each dataset has its own adjusted p-value, Fisher's method was used to combine the adjusted p-values to calculate a combined adjusted p-value for each gene. If the combined adjusted p-value for the gene is larger than 0.05, this gene is removed. This criterion removed genes for which there is weak statistical support, which could be due to a low base expression where small variations result in large log₂ fold changes or the variation between replicates/duplicates is relatively large compared

to the log₂ fold change. This reduces the list of genes to 585 (see **Appendix 20** for the gene list) which is then used for k-means clustering the datasets into 6 groups (instead of 7) (**Table 7.3**).

1	Sarkies Orsay virus	White <i>Stenotrophomonas maltophilia</i>			
2	Irazoqui <i>Staphylococcus aureus</i>	Bond <i>Staphylococcus aureus</i>	modENCODE <u><i>Drechmeria coniospora</i></u>	Rourke <i>Microbacterium nematophilum</i>	Pukkila-Worley <u><i>Candida albicans</i></u>
3	Engelman <i>Enterococcus faecalis</i>	Yuen <i>Enterococcus faecalis</i>	Yuen <i>Enterococcus faecium</i>	Yuen <i>Bacillus subtilis</i>	
4	Engelman <i>Serratia marcescens</i>	modENCODE <i>Serratia marcescens</i>	modENCODE <u><i>Harposporium sp.</i></u>		
5	Tanguy Orsay virus				
6	Bond <i>Pseudomonas aeruginosa</i>	Troemel <i>Pseudomonas aeruginosa</i>	Estes <i>Pseudomonas aeruginosa</i>	Miller <i>Pseudomonas aeruginosa</i>	Sahu <i>Vibrio cholerae</i>
	Bolz <i>Yersinia pestis</i>	Bakowski <u><i>Nematocida parisii</i></u>	Engelman <i>Photorhabdus luminescens</i>	Head <i>Salmonella enterica</i>	Yang <i>Bacillus thuringiensis</i>

Table 7.3 Grouping of the 585 differentially expressed genes from the 25 datasets into 6 groups via k-means clustering. Fungal pathogens are underlined, and viral pathogens are written in bold. The dataset name (first author of publication or modENCODE) is written in red. Pathogens are colour coded when it exists more than once.

This K-means clustering (**Table 7.3**) is relatively consistent with the previous clustering of all 29 datasets. All pathogens, except the Orsay virus, were grouped in the same groups. The heatmap for the 585 genes from the six K-mean clustered group can be seen in **Figure 7.8**.

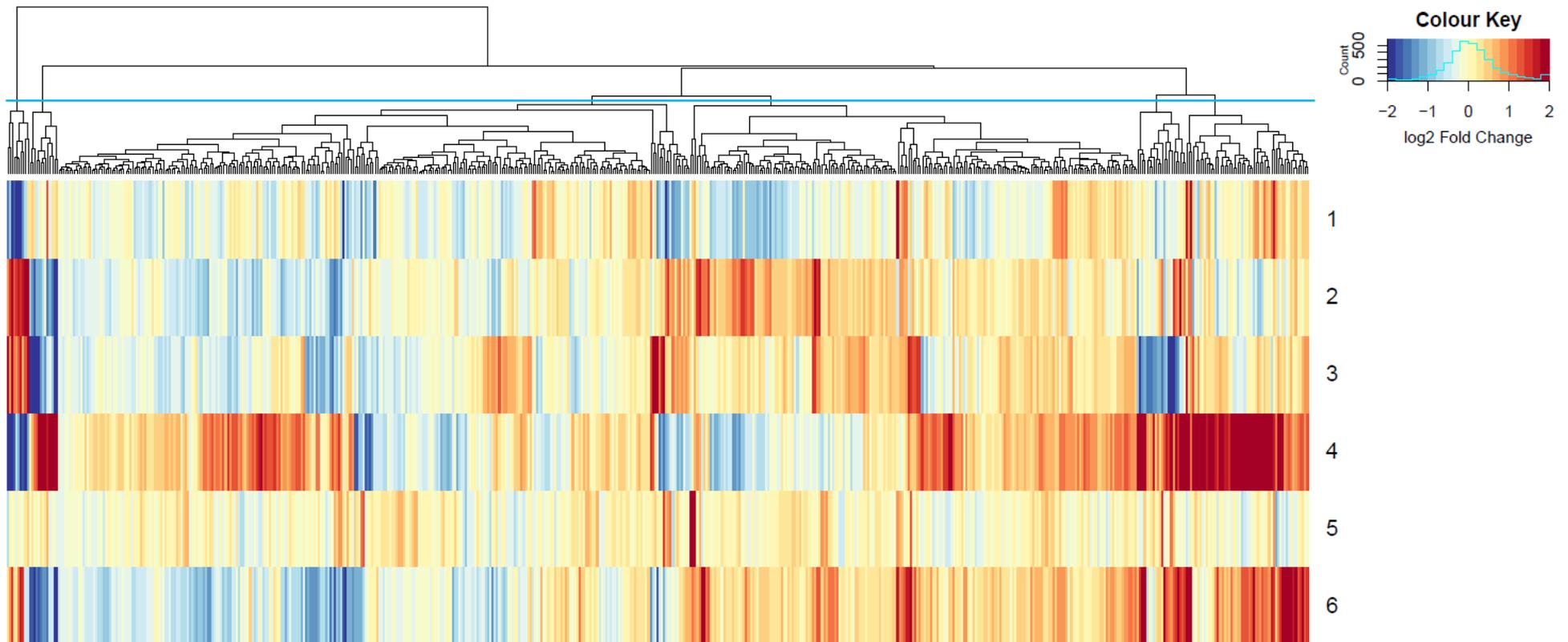


Figure 7.8 Heatmap of the 585 filtered genes from the 25 datasets. The genes are placed along the x-axis while the grouped datasets are on the y-axis. The numbers on the left side of the y-axis denote the K-mean clustering group (Table 7.3). The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes in the particular group, while blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The horizontal blue line at the dendrogram indicates the cut-off level used to group genes into clusters.

The heatmap shows a more distinct patterning on both the left side and the right side. The left-most genes show strong differential expression in all groups except for group number 5, while many genes on the right show up-regulation in group 6 and 4, which makes up more than half of the datasets. The middle section of the heatmap shows weakly differentially expressed genes, represented by lighter shades of blue and yellow. Compared to the initial heatmap that included all 29 datasets, the result of clustering reduced the log₂ fold change of the genes, due to averaging these across each group, making the heatmap look more “faded” (**Figure 7.8**).

Some patterning on the heatmap started to appear. However, the larger number of genes left after filtering in addition to the clustering that reduces the signal intensity of the heatmap makes this result less ideal than preferred. As such, I chose to increase the stringency of the filtering criteria further to reduce the number of genes to a similar level as the initial heatmap containing all 29 datasets (**Figure 7.6**).

The new method used the same initial 3 filtering steps, albeit with more stringent cut-offs (the *absolute sum of log₂FC* > 7.5 across all 25 datasets for the gene. At least six datasets must have a $|\log_2FC| > 0.6$. At least 13 datasets must have a $|\log_2FC| > 0.1$ and $p < 0.05$). 1671 genes remained, which were then used by the k-means clustering function to reduce the 25 datasets into 7 groups (**Table 7.4**). Two more filtering steps were then done to reduce the number of genes further. At least 3 groups must have a $|\log_2FC| > 0.6$ and at least 5 groups must have a $|\log_2FC| > 0.3$. By clustering the k-means at an earlier stage, it provides the function with more data points (1671 genes), thus allowing a more comprehensive comparison between datasets, while also avoiding noisy background data (genes that are not differentially expressed, but show technical/biological variation) that would negatively impact the accuracy. The addition of filtering after clustering has the benefit that pathogens with more datasets do not account for more. For example, the four *P. aeruginosa* datasets would favour *P. aeruginosa* specific genes, while pathogens where only one dataset exists, such as *V. cholerae* would see their weighting reduced. Clustering would put similar pathogens together, and the groups will have the same weightings.

1	Tanguy Orsay virus	Head <i>Salmonella enterica</i>		
2	Sarkies Orsay virus	White <i>Stenotrophomonas maltophilia</i>		
3	Irazoqui <i>Staphylococcus aureus</i>	Bond <i>Staphylococcus aureus</i>	Rourke <i>Microbacterium nematophilum</i>	modENCODE <u><i>Drechmeria coniospora</i></u>
	Engelman <i>Photorhabdus luminescens</i>	Yang <i>Bacillus thuringiensis</i>	Bakowski <u><i>Nematocida parisii</i></u>	
4	Troemel <i>Pseudomonas aeruginosa</i>	Bond <i>Pseudomonas aeruginosa</i>	Estes <i>Pseudomonas aeruginosa</i>	Miller <i>Pseudomonas aeruginosa</i>
	Bolz <i>Yersinia pestis</i>	Sahu <i>Vibrio cholerae</i>		
5	Engelman <i>Serratia marcescens</i>	modENCODE <i>Serratia marcescens</i>	modENCODE <u><i>Harposporium sp.</i></u>	
6	Engelman <i>Enterococcus faecalis</i>	Yuen <i>Enterococcus faecalis</i>	Yuen <i>Enterococcus faecium</i>	Pukkila-Worley <u><i>Candida albicans</i></u>
7	Yuen <i>Bacillus subtilis</i>			

Table 7.4 Grouping of the 1671 genes from the 25 datasets into seven groups via k-means clustering. Fungal pathogens are underlined, and viral pathogens are written in bold. The dataset name (first author of publication or modENCODE) is written in red. Pathogens are colour coded when it exists more than once.

The new list of differentially expressed genes consists of 383 genes (see **Appendix 20** for the gene list) and shows more distinct heatmap pattern and stronger colouring indicative of larger log₂ Fold Change (**Figure 7.9**). The left side of the heatmap is a region of predominantly down-regulated genes (except group 5), while the right side consists of mainly up-regulated genes. The middle region shows genes that are both up-regulated and down-regulated in a group-specific manner.

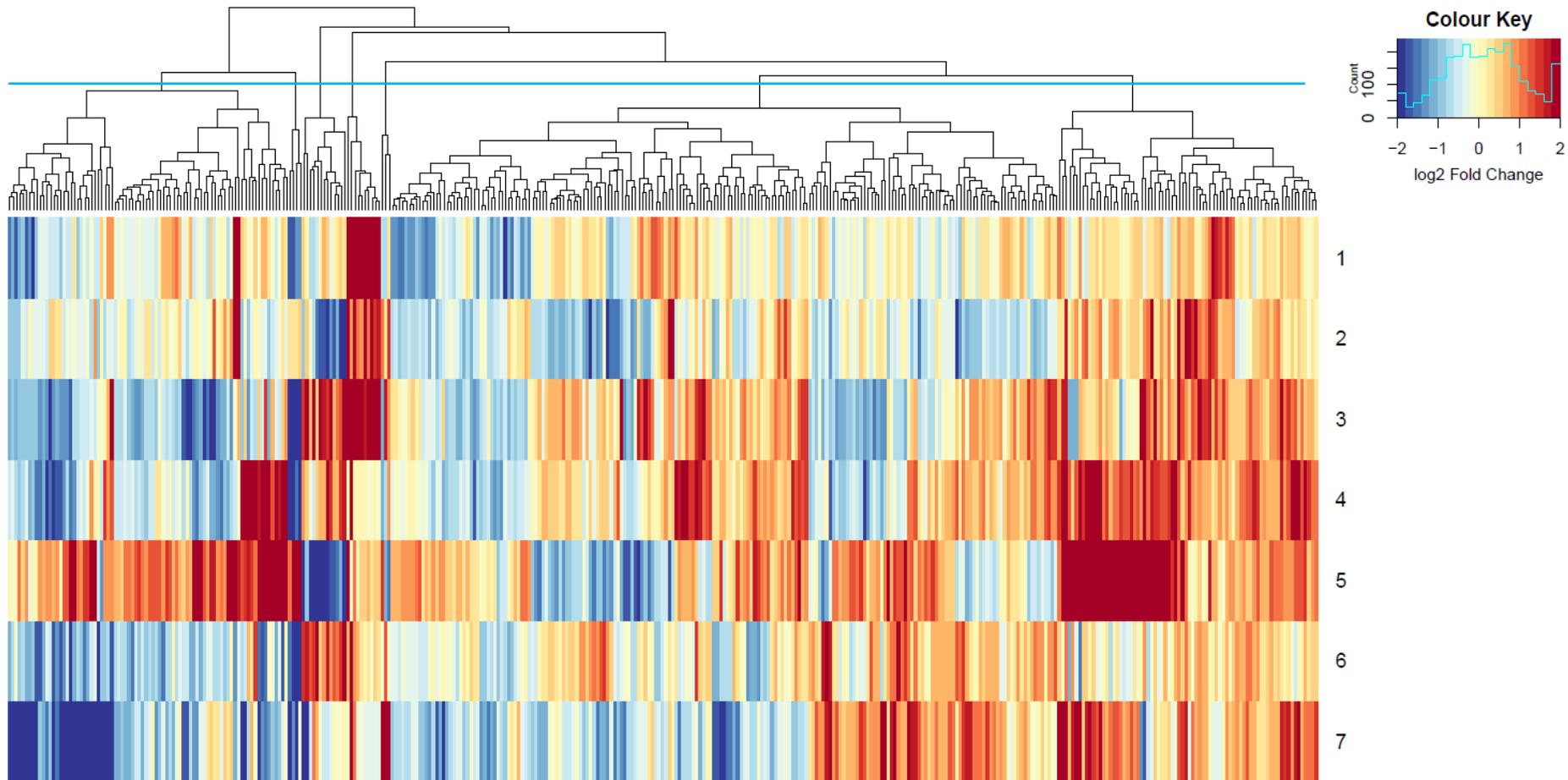


Figure 7.9 Heatmap of the 383 genes from the 25 datasets using the more stringent filtering criteria. The genes are placed along the x-axis while the datasets are on the y-axis. The numbers on the left side of the y-axis denote the K-mean clustering group (see Table 7.4). The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes in the particular group, while blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The horizontal blue line at the dendrogram indicates the cut-off level used to group genes into clusters.

Comparing the list of genes which were used to generate the three heatmaps (**Figure 7.6**, **Figure 7.8** & **Figure 7.9**), it shows that the different filtering methodology and criteria do affect the final composition of genes to a larger degree than anticipated (**Figure 7.10a**). However, the hypergeometric distribution test shows that the similarities are still very significant (**Figure 7.10b**), indicating that the important biological signal from the datasets is robust.

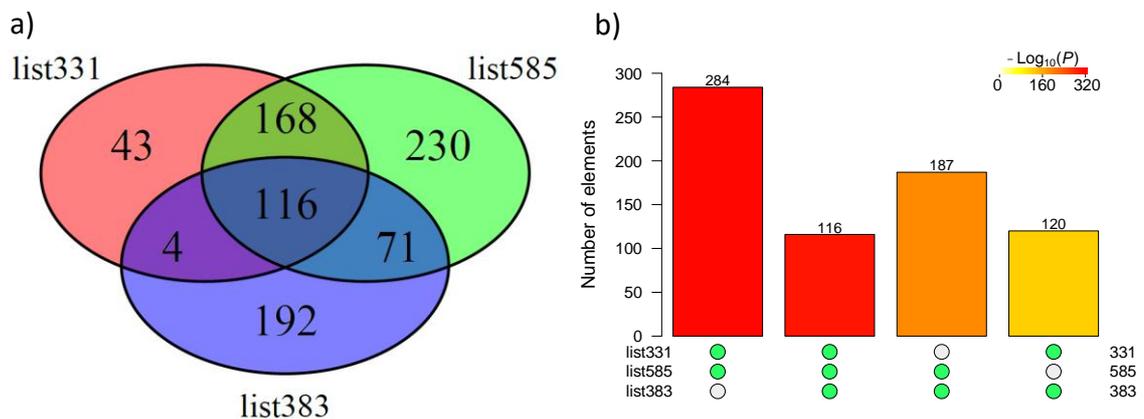


Figure 7.10 Comparison of the three filtered gene lists. a) Venn diagram of the three lists containing 331, 585 and 383 genes as a result of the different filtering methodology. b) Hypergeometric distribution test of the intersecting genes. The green dots below the bar chart indicate the intersecting lists. The legend indicates the p-value (larger number means more significant). The number above each column shows the number of shared genes. The redder the colour of the column, the more significant the overlap of the genes are. list331 = the initial list of genes (Figure 7.6). list585 = the second list of genes omitting the four outlier datasets (Figure 7.8). list383 = the final list of genes (Figure 7.9).

Both list585 (the second list of genes, omitting the four outlier datasets) and list383 (the final list of genes) were analysed for gene set enrichment using Wormbase Enrichment Analysis software (**Figure 7.11**) and g:Profiler (**Appendix 9**) web-based software. The analysis on list331 (initial list of genes) was not included here, as the majority of its genes are shared with list585.

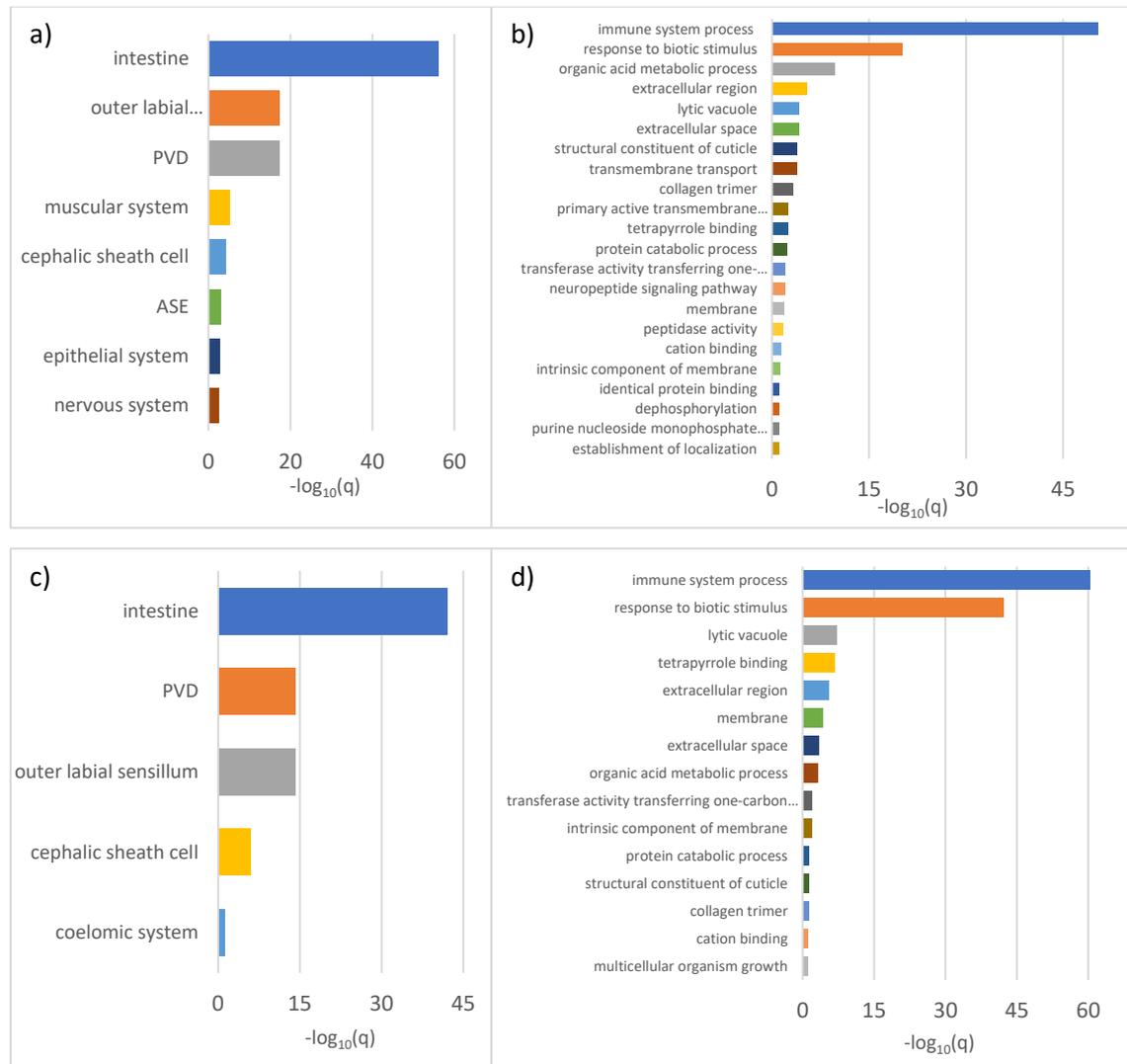


Figure 7.11 Wormbase Enrichment Analysis on list585 and list383. a) & b) List585 results for Tissue Enrichment Analysis and Gene Ontology (GO) Enrichment Analysis, respectively. c) & d) List383 results for Tissue Enrichment Analysis and Gene Ontology (GO) Enrichment Analysis, respectively. All terms have a q-value lower than 0.05.

The results from the Wormbase Enrichment Analysis of the two gene lists do not differ much, especially at the high confidence terms, which includes immune system processes, response to biotic stimuli, intestine, defence response and membrane raft. list585 has more GO terms toward the lower end of the significance cut-off that are not found in list383, mainly related to metabolic and synthetic processes. The next step was to determine which gene set enrichment terms corresponds to which cluster (section) of the heatmap, i.e. which terms are correlated with up-regulated genes and which terms are associated with down-regulated genes. The method of clustering the genes of the heatmap together is by using a line of cut-off on the dendrogram which then groups all the genes below the line together (**Figure 7.8** and **Figure 7.9** blue line).

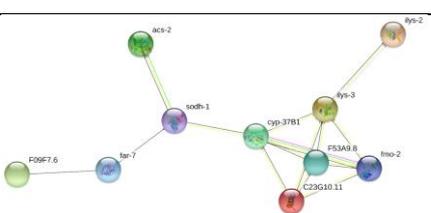
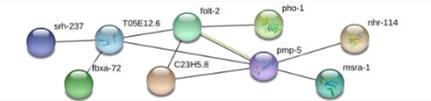
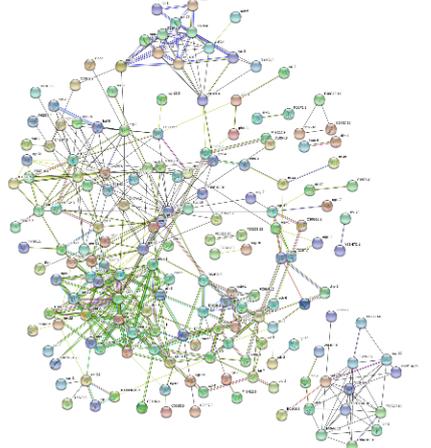
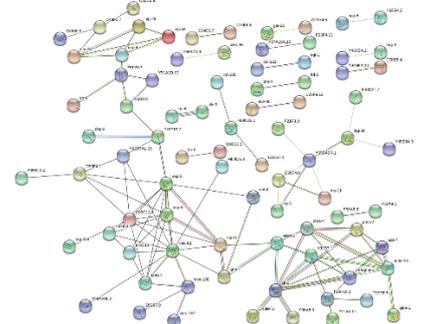
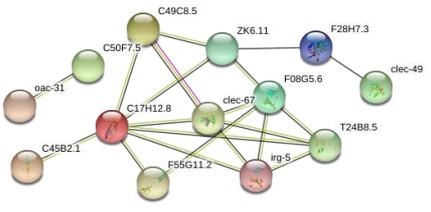
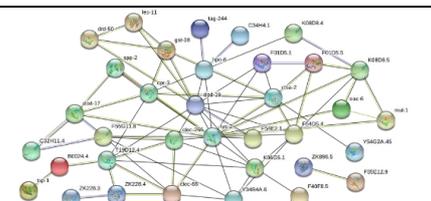
list585		STRING	# of genes		
Wormbase Enrichment Analysis					
Term	Q value				
pm7 WBbt:0003721	0.074		10		
pm3 WBbt:0003740	0.074				
pm5 WBbt:0003737	0.074				
excretory gland cell WBbt:0005776	0.074				
response to biotic stimulus GO:0009607	7.30E-06				
lytic vacuole GO:0000323	0.054				
iron ion binding GO:0005506	0.054				
tetrapyrrole binding GO:0046906	0.056				
No enrichment					
			13		
intestine WBbt:0005772	4.90E-26		284		
organic acid metabolic process GO:0006082	2.80E-07				
transmembrane transport GO:0055085	0.0002				
structural constituent of cuticle GO:0042302	0.0002				
lipid catabolic process GO:0016042	0.00021				
response to biotic stimulus GO:0009607	0.0023				
immune system process GO:0002376	0.0026				
collagen trimer GO:0005581	0.0026				
extracellular region GO:0005576	0.0049				
purine nucleoside monophosphate metabolic process GO:0009126	0.0081				
extracellular space GO:0005615	0.0084				
purine nucleotide metabolic process GO:0006163	0.0086				
lytic vacuole GO:0000323	0.011				
nucleoside phosphate metabolic process GO:0006753	0.019				
transferase activity transferring one-carbon groups GO:0016741	0.03				
ribose phosphate metabolic process GO:0019693	0.042				
intestine WBbt:0005772	2.90E-11		201		
outer labial sensillum WBbt:0005501	3.10E-07				
PVD WBbt:0006831	3.10E-07				
immune system process GO:0002376	5.80E-15				
response to biotic stimulus GO:0009607	7.40E-06				
extracellular region GO:0005576	8.30E-05				
lipid catabolic process GO:0016042	0.00041				
extracellular space GO:0005615	0.001				
tetrapyrrole binding GO:0046906	0.0017				
organic acid metabolic process GO:0006082	0.0042				
neuropeptide signalling pathway GO:0007218	0.0042				
iron ion binding GO:0005506	0.018				
response to topologically incorrect protein GO:0035966	0.027				
intestine WBbt:0005772	0.0033				
cephalic sheath cell WBbt:0008406	0.0057		22		
PVD WBbt:0006831	0.024				
outer labial sensillum WBbt:0005501	0.024				
immune system process GO:0002376	1.80E-13				
lipid catabolic process GO:0016042	0.0011				
response to biotic stimulus GO:0009607	0.0038				
lytic vacuole GO:0000323	0.029				
identical protein binding GO:0042802	0.035				
intestine WBbt:0005772	1.40E-12				
PVD WBbt:0006831	2.30E-09				55
outer labial sensillum WBbt:0005501	2.30E-09				
immune system process GO:0002376	2.00E-43				
response to biotic stimulus GO:0009607	3.70E-06				
membrane GO:0016020	0.015				

Figure 7.12 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of Figure 7.8. The order of cluster from top to bottom corresponds to the cluster on the heatmap from left to right. The number of genes in each cluster are shown in the right-most column, and the colour corresponds to the colouring in Appendix 20. Wormbase Enrichment Analysis results include tissue and GO term enrichment analysis. Tissue enrichment terms have the ID prefix “WBbt” and GO terms have the prefix “GO”. STRING excludes disconnected nodes.

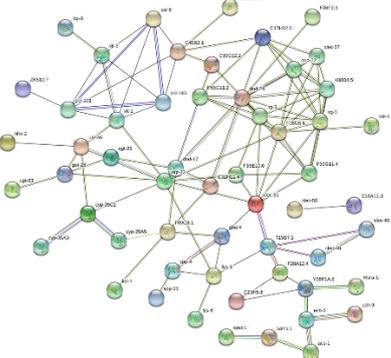
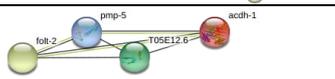
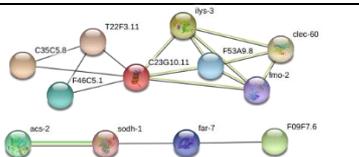
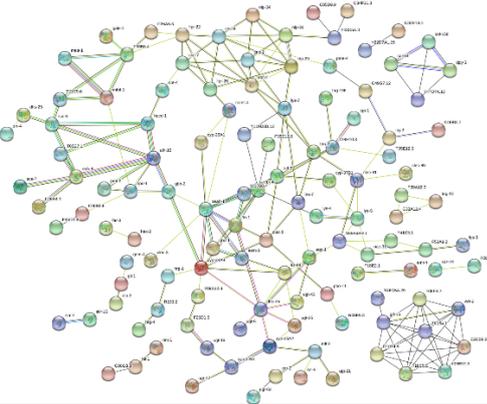
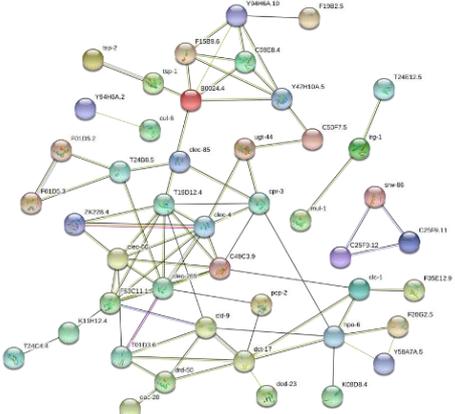
list383		STRING	# of genes
Wormbase Enrichment Analysis			
Term	Q value		
intestine WBbt:0005772	1.00E-12		82
cephalic sheath cell WBbt:0008406	5.10E-06		
PVD WBbt:0006831	0.016		
outer labial sensillum WBbt:0005501	0.016		
immune system process GO:0002376	2.40E-11		
response to biotic stimulus GO:0009607	4.50E-06		
lipid catabolic process GO:0016042	0.00064		
iron ion binding GO:0005506	0.00064		
organic acid metabolic process GO:0006082	0.00072		
tetrapyrrole binding GO:0046906	0.0012		
lytic vacuole GO:0000323	0.0029		
structural constituent of cuticle GO:0042302	0.017		
collagen trimer GO:0005581	0.017		
primary active transmembrane transporter activity GO:0015399	0.043		4
response to biotic stimulus GO:0009607	1.80E-07		13
cephalic sheath cell WBbt:0008406	0.041		10
No enrichment		No interaction	3
intestine WBbt:0005772	7.60E-14		195
outer labial sensillum WBbt:0005501	3.60E-05		
PVD WBbt:0006831	3.60E-05		
coelomic system WBbt:0005749	0.039		
response to biotic stimulus GO:0009607	1.40E-14		
immune system process GO:0002376	4.10E-13		
tetrapyrrole binding GO:0046906	3.30E-06		
iron ion binding GO:0005506	3.30E-06		
extracellular region GO:0005576	8.20E-06		
lytic vacuole GO:0000323	0.00011		
extracellular space GO:0005615	0.00011		
lipid catabolic process GO:0016042	0.00018		
organic acid metabolic process GO:0006082	0.00064		
transferase activity transferring one-carbon groups GO:0016741	0.02		
intestine WBbt:0005772	6.40E-16		76
PVD WBbt:0006831	5.10E-11		
outer labial sensillum WBbt:0005501	5.30E-11		
immune system process GO:0002376	5.90E-38		
response to biotic stimulus GO:0009607	2.10E-09		
membrane GO:0016020	5.30E-05		
intrinsic component of membrane GO:0031224	0.0039		

Figure 7.13 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of Figure 7.9. The order of cluster from top to bottom corresponds to the cluster on the heatmap from left to right. The number of genes in each cluster are shown in the right-most column, and the colour corresponds to the colouring in Appendix 20. Wormbase Enrichment Analysis results include tissue and GO term enrichment analysis. Tissue enrichment terms have the ID prefix “WBbt” and GO terms have the prefix “GO”. STRING excludes disconnected nodes.

Analysis of gene clusters from list585 and list383 show similar trends (**Figure 7.12** and **Figure 7.13**). “Immune system process” is found at the clusters further down in the table (corresponding to the right side of the heatmaps), indicating that this process sees an upregulation of its corresponding genes during pathogen infection. The “Intestine” term comes up in most clusters which imply that most of the pathogen response is located at the intestine. Terms corresponding to metabolic processes are found mainly at the top of the table, indicating that genes controlling these are down-regulated. STRING protein-protein interaction analysis illustrates a large protein interaction web, suggesting that *C. elegans* pathogen response is not controlled by a few relatively isolated pathways, but a large network of proteins.

In order to proceed further, I concentrated the next analysis on list383. I chose this list over list585 because list585 has a very large difference in cluster size, as most of the genes are contained in only 2 clusters (284 and 201 genes making up 82.9% of all the differentially expressed genes). Coincidentally, these two clusters contain genes with relatively low log₂ Fold Change and mixture of up- and down-regulated genes. This makes it difficult to determine whether the clusters correspond to up-regulated or down-regulated genes as a result of pathogen infection. Furthermore, the q-values for the list383 Wormbase Enrichment Analysis results are on average statistically stronger and are thus less likely to be false positive.

A GO-term network analysis was done for list383 using ClueGO (Bindea, et al., 2009) and BiNGO (Maere, et al., 2005) apps on the Cytoscape platform. Both the ClueGO (**Figure 7.14**) and BiNGO (**Appendix 10**) results show that the significant GO-terms all belong to a few large groups of distinct biological processes. The first very distinct group is related to defence responses and include significant parent and child terms related to more general ‘response to stimulus’ as well as more specific terms like ‘response to fungus’. Next, there is a diverse group of metabolic processes that are related to ‘lipid/fatty acid metabolic process’, ‘oxidation/reduction’ and ‘carbohydrate catabolic/metabolic process’. Finally, there are two small groups related to lifespan and transport (of molecules) (**Figure 7.14**).

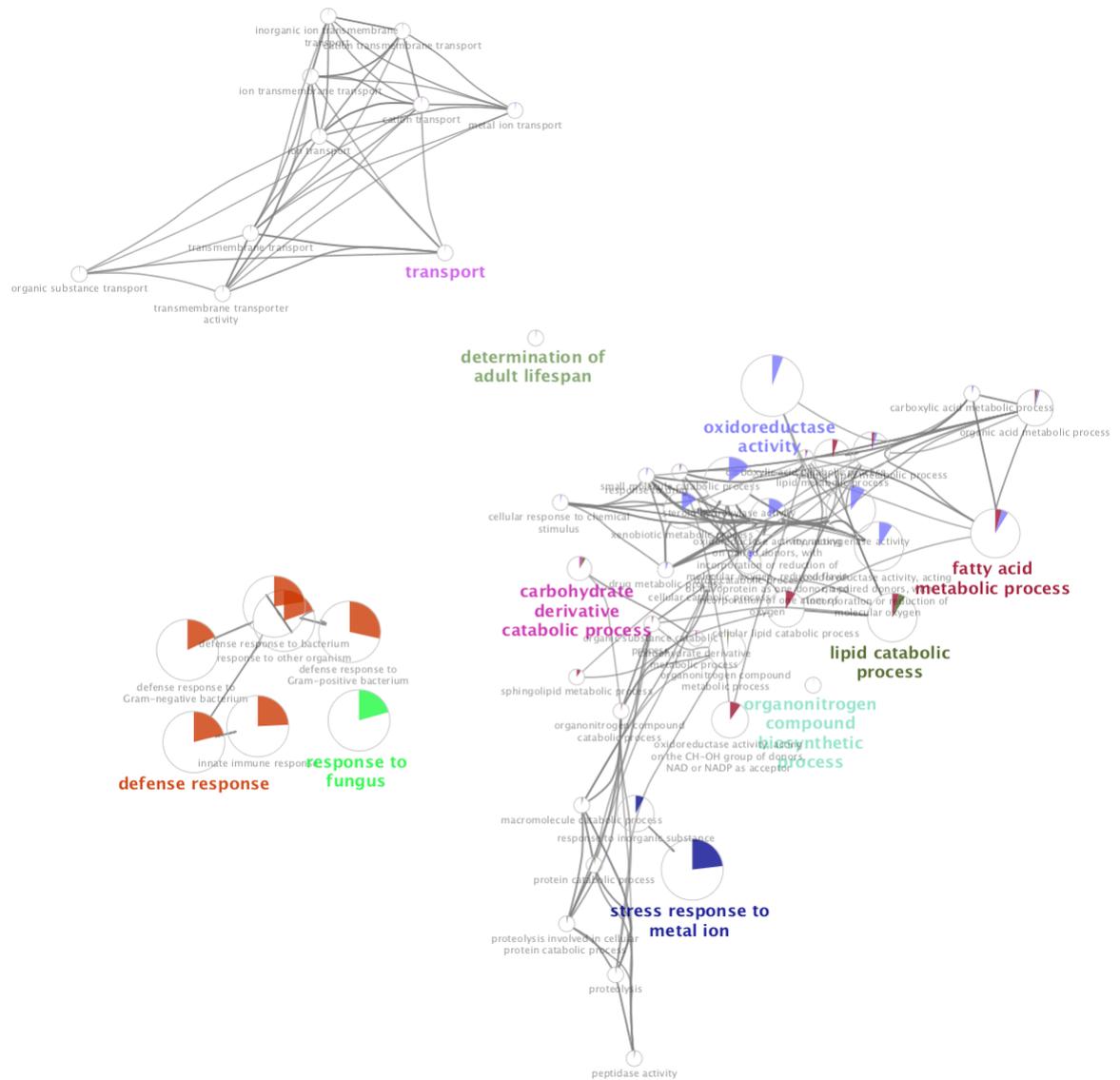


Figure 7.14 ClueGO network analysis for 'GO:biological processes' on all genes from list383. GO terms are indicated by nodes. The pie-chart indicates the proportion of genes from the query list relative to the whole set of genes associated with the term (See Appendix 11 for more details). Terms written in bold are the most significant terms within the group. Edges indicate parent/child term relationship. Minimum of five genes per node and GO hierarchy level range from 3-6. GO Term Fusion was enabled.

By concentrating on only the three largest clusters of genes from list383, that make up 92% of the genes in the list, we can better determine their function within the network. The three clusters correspond to genes that are mainly down-regulated (**Figure 7.13** blue), genes that are both up- and down-regulated in a group-specific manner (**Figure 7.13** yellow), and genes that are mainly up-regulated (**Figure 7.13** red). In the heatmap (**Figure 7.9**) this corresponds to the left, middle and right clusters with 82, 195 and 76 genes (**Figure 7.13**) respectively. The ClueGO analysis was repeated, this time by specifying three lists of genes corresponding to three clusters (**Figure 7.15a**).

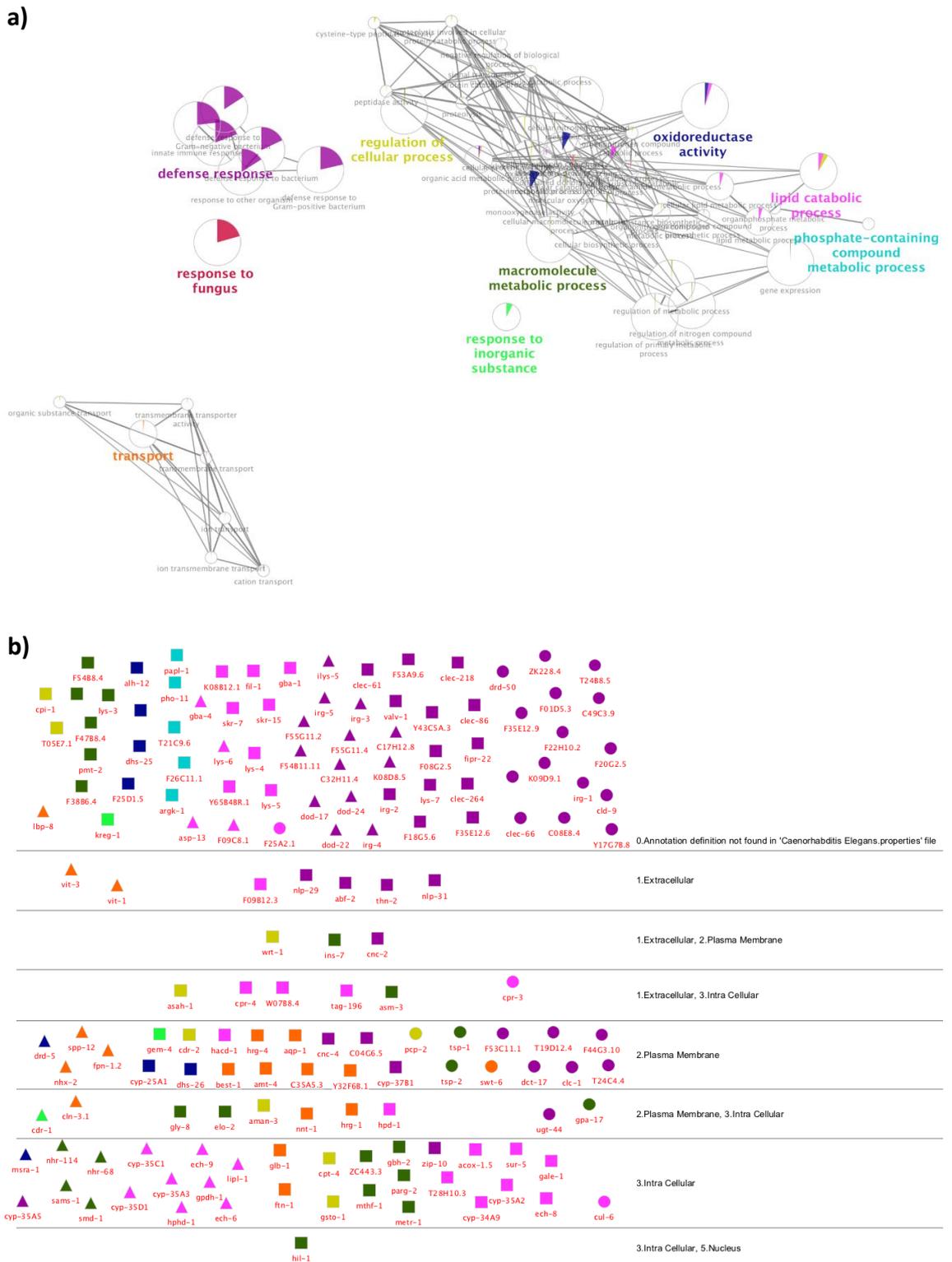


Figure 7.15 ClueGO analysis for GO:biological processes for the three largest groups of list383. The groups are determined based on hierarchical clustering and dendrogram cut-off. a) network analysis with a minimum of five genes per node per cluster and GO hierarchy level range from 3-8 (see Appendix 12 for more detailed list). b) Cerebral View of the network showing all genes that are part of any of the GO terms and are colour coded respectively. Each shape corresponds to one gene and depending on the shape, it belongs to the cluster of mainly down-regulated genes (triangle), group-specific up- and down-regulated genes (square) and mainly up-regulated genes (circles). The genes are also ordered according to where their protein locates in the organism.

From the ClueGO Cerebral View (**Figure 7.15b**), there is a clear distinction between the three gene clusters. Proteins located extracellularly are nearly exclusively squares, which are group dependent up- and down-regulation genes. The cluster of genes that are mainly up-regulated (circles) are found to be mostly associated with defence response (dark purple) and found mostly at the plasma membrane. The genes corresponding to the down-regulated cluster (triangle) are predominantly found intracellularly and are associated with metabolic processes.

7.2.3. Expression of immune effector protein after pathogen infection

Many proteins have been classified as immune effectors. Families of immune effector proteins are ABF (antibacterial factor related), CNC (caenacin), LEC (lectin) & CLEC (c-type lectin), LYS (lysozyme) & ILYS (invertebrate lysozyme), NLP (neuropeptide-like protein) and SPP (saposin-like protein) (Kim & Ewbank, 2018). I was interested to see how the genetic expression of these protein families changes upon infection. As such, I generated heatmaps for each protein family (**Figure 7.16** and **Appendix 13a** for LEC & CLEC).

Based on the gene expression changes, not all protein families mentioned above are responsive to all types of pathogens. The five ABFs are only differentially expressed in a small number of pathogen infections (**Figure 7.16a**). The CNC genes seem to be specific to certain pathogens. *D. coniospora* leads to upregulation of a large number of CNC genes, while *S. marcescens* infection reduces their expression (**Figure 7.16b**). The ILYS and LYS genes (except *ihys-1* & *ihys-6*) are overall differentially expressed in most pathogens (except Orsay Virus). Some of these genes tend towards down-regulation (*hys-4,5,6,7* & *ihys-5,10*), while others are predominantly up-regulated (*hys-1,2,3,8,9* & *ihys-4*). *ihys-2* and *ihys-3* show a strong but pathogen-dependent direction of regulation (**Figure 7.16c**). Around half of the SPP genes are consistently differentially expressed in all datasets, and the direction of expressional change varies depending on the pathogen (**Figure 7.16d**). The NLP genes are mostly weakly differential expression, but there seems to be a consistent but weak pathogen-specific down-regulation (*S. marcescens*) or up-regulation (*B. thuringiensis*) (**Figure 7.16e**). The CLEC & LEC genes are for the most part not differentially expressed (**Appendix 13a**).

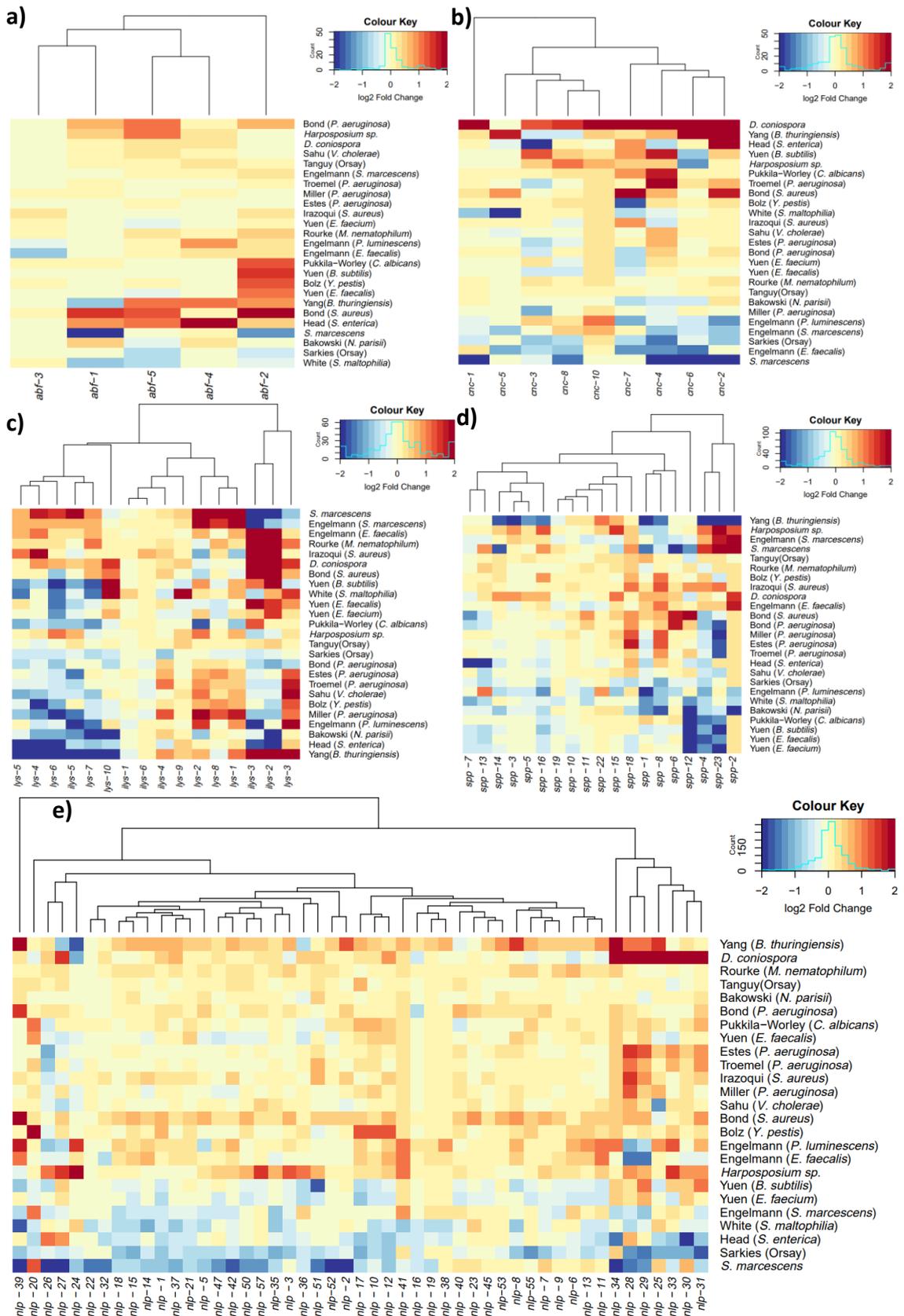


Figure 7.16 Heatmaps of gene expression of protein families considered to be immune effectors in each pathogen dataset: a) ABF, b) CNC, c) LYS & ILYS, d) SPP and e) NLP. The genes are placed along the x-axis while the datasets are on the y-axis. The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes in the particular dataset, while blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The Colour key also shows a density plot.

7.2.4. Discussion

From the results presented in section 7.2.1 - 7.2.3, we can see that during pathogen infection, *C. elegans* responds by changing the expression landscape. The extent to which the gene expression landscape changes varies between different pathogens. The genes that see a significant change in expression differ across the different pathogens, and no gene consistently pass the differential expression significance threshold in all pathogen datasets (**Figure 7.4** and **Figure 7.5**). Visualizing the data as a heatmap then showed more clearly that the gene differential expression across all pathogen datasets varies significantly with no clear clustering (**Figure 7.6**). Furthermore, heatmaps of known immune effector protein families, that were identified to be important for certain pathogens, show that their expressions differ across all the pathogen datasets. Some of these families, such as ABF and NLP generally show low differential expression in most datasets, while other families such as LYS & ILYS are relatively strongly differentially expressed (**Figure 7.16** and **Appendix 13a**). Using an optimized multi-layer filtering criterion, I was able to identify high confidence genes, with consistent expressional changes, that can be labelled as ‘general pathogen responsive genes’, with a more distinct clustering of genes (**Figure 7.9**). These results demonstrate that responses to pathogen infections vary widely and reflect the various mechanisms that different pathogens use to infect the host. However, a relatively small set of genes are consistently differentially expressed in most if not all datasets and may be up-regulated under some pathogen infection and down-regulated in other infections.

Enrichment Analysis on all the genes in list383 showed that a significant portion of the genes is active in the intestine, which is known to be a hot spot for infections (Kim & Ewbank, 2018). Enrichment in the ‘PVD’ and ‘outer labial sensillum’ sensory neurons (Altun & Hall, 2010) implies that neurological sensory activity is involved (**Figure 7.11a & c**), perhaps as a defence/avoidance mechanism by moving away from stressful environments. ‘Immune system process’ and ‘response to biotic stimuli’ is also found strongly enriched (**Figure 7.11b & d**), which is not unexpected, since pathogens are biotic stimuli and induce the activity of the immune system. All the aforementioned enrichment terms become more significant towards the up-regulated gene cluster (**Figure 7.12** and **Figure 7.13**), indicating that up-regulated genes drive more of the response against pathogen infections. Down-regulated gene cluster, on the other hand, show enrichment for terms associated with various metabolic processes (**Figure 7.12** and **Figure 7.13**), which implies a reduction in metabolic activity.

Analysis of the cellular location of the genes in list383 shows that the up-regulated genes tend to be located at the plasma membrane, while down-regulated genes are predominantly found intracellularly. The clusters containing genes with a pathogen-specific up- and down-regulation are found in all locations, including extracellularly (**Figure 7.15**). The up-regulated genes that are exclusively located at the plasma membrane could be membrane-bound proteins such as transport proteins and membrane receptors (**Figure 7.15b** circle). I would hypothesize that these are membrane receptors that received important extracellular cues and signals corresponding to the presence of pathogen infection. Increasing the expression of receptors in general improves the ability to sense a wide variety of different cues. These genes could also act as membrane transport proteins that increase transmembrane transport of chemicals and proteins related to more general stress response function, such as control of metabolism and signalling.

Genes that are up- or down-regulated in a pathogen dependent manner could be pathogen-specific genes (**Figure 7.15b** square). The ones located extracellularly could be pathogen-specific immune response proteins or transcellular signalling proteins that are secreted into the extracellular matrix after the pathogen has been identified, to fight the infection or alert neighbouring cells. Intracellular genes, on the other hand, might have a defensive mechanism to protect the intracellular environment from pathogen-induced stressor or toxins, such as neutralizing or maintaining stable concentrations of chemicals/proteins. At the plasma membrane, their functions could be to act as transport proteins to facilitate the movement of immune response and signalling proteins out of the cell, potentially as part of transcellular signalling, or increase the influx and outflow of chemicals related to cellular maintenance. As membrane receptors, they could help increasing signalling cascades to elevate the immune response signal intensity.

The down-regulated genes are mainly located intracellularly and are associated with metabolic processes (**Figure 7.15b** triangle). This down-regulation may result in reduced metabolic activity and may be necessary to allocate more resources towards the expression of crucial life-preserving genes that are immediately required. A sleep-like (quiescence) response has been reported during heat stress when the nematode show reduced activity (locomotion and feeding) during and after heat-shock (Hill, et al., 2014), arguing in favour of a stress-dependent reduction in metabolic activity. This suggests that during pathogen infection (or other stress), the cell switches from a passive cell maintenance orientated transcriptional instruction towards an active defence focused one to increase survivability (analogously to the sympathetic and parasympathetic nervous system).

From the points above I hypothesize that *C. elegans* increases sensing activity in the presence of pathogens, both at the organismal level through increased neuron activity as well as at the cellular level with the expression of increased membrane receptors. These are pathogen unspecific as the nature of the pathogen will be initially unknown. After identification of the pathogen, the pathogen-specific genes see a change in their expression levels, and at the same time, metabolic genes see a reduction in expression to free more resources to fight the immediate life-endangering threat.

7.3. Transcription factors related to general pathogen response

Similar to the analysis of the immune effector protein family (**Section 7.2.3**), I wondered how the expression of transcription factors (TFs) associated with the cellular stress response and the innate immune response is affected by pathogen infection. Most of the analysed TFs (taken from Kim & Ewbank (2018)) do not show a change in expression under any of the pathogen infections (**Figure 7.17**). This is not unusual as TF activity is often not dependent on transcription, but rather on post-translational modifications (such as phosphorylation), which activates the TF and promote movement from the cytoplasm into the nucleus (Whiteside & Goodbourn, 1993). Transcription for *pqm-1* (and to a lesser degree *zfp-2*) however shows a strong up-regulation in around half of the datasets (**Figure 7.17**).

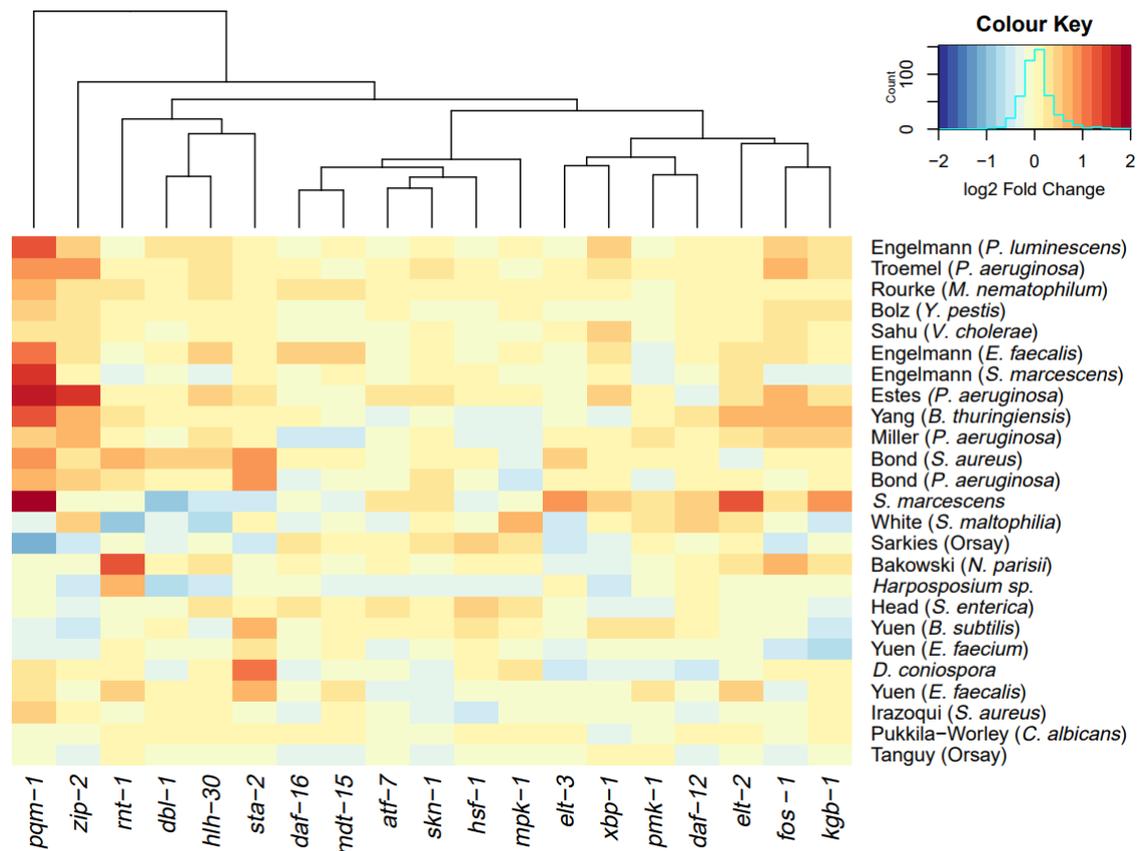


Figure 7.17 Heatmap of transcription factor and co-factor expression changes in each pathogen dataset. The genes are placed along the x-axis while the datasets are on the y-axis. The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes in the particular dataset, while blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The Colour key also shows a density plot.

While the heatmap shows strong support that *pqm-1* could be related to the immune response, I wanted to validate this finding further and potentially identify further transcription factors that do not see a change in expression but might play an important role in the immune response. Therefore, I used *de novo* motif discovery method to find immune response-related TFs and analyse published TF ChIP-seq data, to identify the TF binding sites to then compare them with the genes in list383.

In order to identify TF binding targets, it is important to identify the region where such TF binding would be associated with the respective gene. In general, the promoter region is the main area where TFs are expected to bind and can be identified by a transcription start site (TSS). However, *C. elegans* transcription is affected by trans-splicing events and operons, making it difficult to pinpoint the exact TSS location and promoter region. Trans-splicing is the event where the pre-mRNA has part of its 5'-end (including the TSS) replaced by a splice-leader. Thus conventional methods of assessing the TSS based on mRNA sequencing is not able to pinpoint the exact TSS location. Around 70% of all *C. elegans* genes are trans-spliced

(Blumenthal, 2012). Operons are a cluster of genes located downstream of each other that are transcribed at the same time and controlled by a single promoter. Hence, if the gene in the operon is not the first immediate downstream of the promoter, it is difficult to find the TSS. Around 15% of all *C. elegans* genes are within operons (Blumenthal, 2012). Current public databases such as Ensembl BioMart (Ensembl, 2019) only contain the “transcript start site” (not to be confused with “transcription start site”) of mature mRNA, i.e. for which the 5'-end has already been replaced by the splice leader.

By identifying where the transcription factors bind relative to the transcript start site, the rough position of the transcription start site and promoter can be identified. For this, I used published ChIP-seq data of PQM-1 (Niu, et al., 2011) and DAF-16 (modENCODE ID: 591) from modENCODE. Both of these transcription factors are related to the immune response (Tepper, et al., 2013; O'Brien, et al., 2018; Kim & Ewbank, 2018). MACS2 peak calling (Zhang, et al., 2008) (for threshold cut-off, refer to **Appendix 14**) was performed, and the resulting peaks were assigned to the genes if they are located within 500bp upstream or downstream (total of 1 kb window) of the transcript start site. Seqplot was then utilized to visualize the average ChIP-seq signal around all the TF target gene transcript start site (**Figure 7.18**). Both DAF-16 and PQM-1 ChIP-seq peak starts at around 500 bp upstream and end at around 200 bp downstream, peaking at roughly 150 bp upstream, showing that the 1 kb window is a suitable choice. This is similar to the estimations inferred from H3K4me3 peaks (Kolasinska-Zwierz, et al., 2009). Although the modENCODE ChIP-seq data have been aligned to the ce10 reference assembly, using the most recent gene annotation version based on ce11, more defined peaks were obtained compared to using the older gene annotation version based on ce10 (**Figure 7.18**).

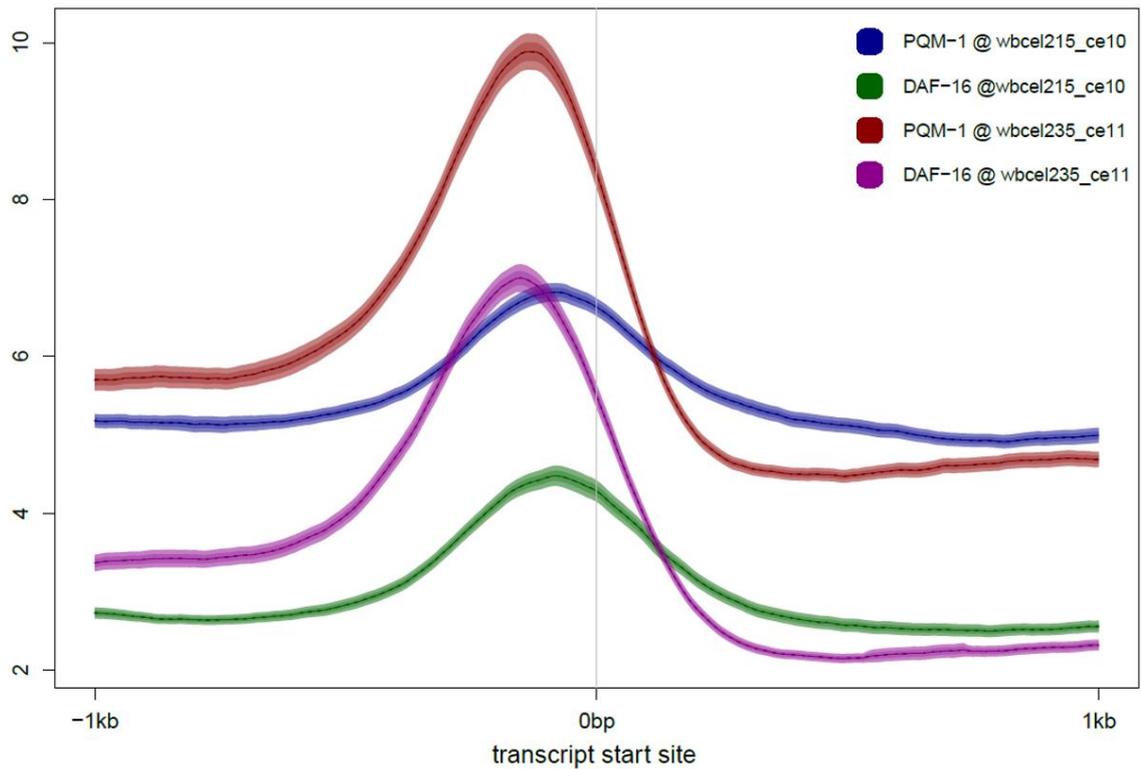


Figure 7.18 Seqplot of the ChIP-seq signal around the transcript start site. The blue (PQM-1) and green (DAF-16) track use the older gene annotation based on ce10. The red (PQM-1) and purple (DAF-16) tracks use the newest gene annotation version based on ce11. ce11 is on average 55bp larger than ce10 per chromosome. The dashed line denotes the mean, the dark area is the standard error, and the light area indicates a 95% confidence interval.

After having identified the area (500bp upstream and downstream of the transcript start site) where TF binding is expected, I was then able to use *de novo* motif discovery software to identify enriched motifs in the area of all genes in list383. For this, HOMER has been used to find motifs and their associated TFs (**Figure 7.19a**). The top hit from HOMER is PQM-1, with a p-value much smaller than any other hit. Interestingly, the PQM-1 motif has the same sequence as that of ELT-3 (**Figure 7.19b**). ELT-3 but not PQM-1 was also identified using other *de novo* motif discovery software: Trawler, BAMB motif and DREME (**Appendix 15**). The reason for this is that the motif data for PQM-1 is only included in the database exclusive to HOMER. The other software programs refer to existing databases, mostly JASPAR, which does not have an entry for PQM-1 (at the time when the analysis was conducted).

a)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-36	-8.445e+01	46.53%	10.75%	77.3bp (104.5bp)	PQM-1(7)/cElegans-L3-ChIP-Seq(modEncode)/Homer(0.898) More Information Similar Motifs Found
2		1e-13	-3.193e+01	5.45%	0.13%	61.5bp (99.7bp)	POL012.1_TATA-Box/Jaspar(0.808) More Information Similar Motifs Found
3		1e-12	-2.847e+01	9.41%	0.99%	213.4bp (353.5bp)	POL010.1_DCE_S_III/Jaspar(0.584) More Information Similar Motifs Found
4*		1e-11	-2.639e+01	6.93%	0.48%	78.7bp (113.6bp)	unc-62/MA0918.1/Jaspar(0.599) More Information Similar Motifs Found
5*		1e-11	-2.621e+01	11.39%	1.79%	315.5bp (327.0bp)	PL0005.1_hlh-30/Jaspar(0.544) More Information Similar Motifs Found
6*		1e-11	-2.602e+01	21.29%	6.47%	262.6bp (318.3bp)	HLH-1(bHLH)/cElegans-Embryo-HLH1-ChIP-Seq(modEncode)/Homer(0.599) More Information Similar Motifs Found
7*		1e-10	-2.491e+01	9.41%	1.22%	259.8bp (320.9bp)	MF0004.1_Nuclear_Receptor_class/Jaspar(0.610) More Information Similar Motifs Found
8*		1e-10	-2.482e+01	5.45%	0.26%	86.2bp (84.8bp)	POL012.1_TATA-Box/Jaspar(0.753) More Information Similar Motifs Found
9*		1e-10	-2.453e+01	7.92%	0.81%	102.9bp (101.1bp)	PL0008.1_hlh-29/Jaspar(0.564) More Information Similar Motifs Found
10*		1e-10	-2.415e+01	23.76%	8.33%	95.7bp (110.5bp)	skn-1/MA0547.1/Jaspar(0.815) More Information Similar Motifs Found

b)

PQM-1(?)/cElegans-L3-ChIP-Seq(modEncode)/Homer	
Match Rank: 1	
Score: 0.90	
Offset: 1	
Orientation: reverse strand	
Alignment: KTCTTATCAGKT -TCTTATCAGT-	
elt-3/MA0542.1/Jaspar	
Match Rank: 2	
Score: 0.88	
Offset: 1	
Orientation: forward strand	
Alignment: KTCTTATCAGKT -TCTTATCA---	
ELT-3(Gata)/cElegans-L1-ELT3-ChIP-Seq(modEncode)/Homer	
Match Rank: 3	
Score: 0.87	
Offset: 1	
Orientation: reverse strand	
Alignment: KTCTTATCAGKT -TCTTATCAWT-	

Figure 7.19 de novo Motif discovery result using HOMER on list383. a) Top 10 most enriched motifs and the best fitting known TF motif. A red asterisk denotes possible false positives. b) Known TF motifs that match the top enriched motif hit.

Using the published ChIP-seq data for PQM-1, DAF-16 and ELT-3, as well as HSF-1 (since it is associated with a wide variety of stress responses), I compared the TF binding targets with one another to see how similar the TFs set of target genes are. PQM-1, DAF-16 and ELT-3 ChIP-seq datasets are taken from modENCODE, while the HSF-1 dataset is from Li, et al. (2016).

From **Figure 7.20**, the HSF-1 ChIP-seq did not find many target genes compared to the other datasets, which is due to the low signal of the ChIP-seq datasets. Proportionally, all TFs share a significant portion of target genes. Between PQM-1 and ELT-3, 1911 genes are shared which makes up 27% and 82% of their total number of target genes, respectively.

This high overlap is not unreasonable, given the similarity of their motif. When comparing the genes in list383 to the PQM-1 ChIP-seq data, 249 genes are in common ($p = 1.1 \times 10^{-36}$).

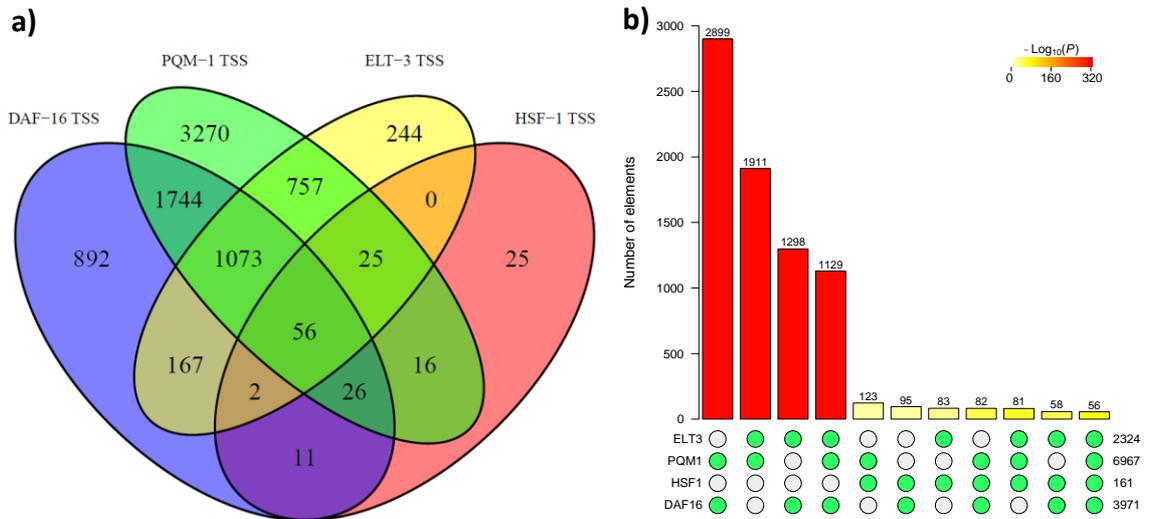


Figure 7.20 Comparison of the target genes of different TFs associated with immune response. a) Venn diagram showing the overlap of different TF target genes. b) Hypergeometric distribution of each overlap combination. Green dots below the graph denotes which groups were overlapped. The legend indicates the p-value (larger number means more significant). The number above each column shows the number of shared genes. The redder the colour of the column, the more significant the overlap of the genes are.

Due to the extremely low P-value, I tested the program that calculates the hypergeometric distribution (*superexacttest* (Wang, et al., 2015)) on unrelated TF ChIP-seq data. The tested ChIP-seq data all have many binding targets with a significant number of overlapping genes (**Appendix 14**). The results further show a positive correlation between the number of binding targets and the significance of overlaps, i.e. ChIP-seq data with a larger number of binding targets also show more significant overlap. However, when choosing genes at random, the program returns high p-values (low $-\log_{10}(P)$), which would be expected under completely random distribution. This shows that ChIP-seq data or TF binding might have a bias associated to them that prefers binding to the same target genes, but not necessarily affecting their expression (e.g. High-occupancy target (HOT) regions). Some genes might also never be bound by TFs in proximity, such as downstream operon genes. In this case, the number of “background” genes (used for hypergeometric testing) is smaller than conventionally assumed, and the p-value becomes inflated. Ways to account for this would be to create a reference background gene list of all the currently known genes that have been found bound by TFs. HOT regions are difficult to assess as the threshold for “high occupancy” is difficult to define. However, various groups such as Wreczycka, et al. (2019) have compiled lists for what they define as HOT regions for various organisms, including *C. elegans* with 422 regions. Such lists could be used to blacklist regions during the ChIP-seq analysis (Wreczycka, et al., 2019).

With the number of TF binding targets in the thousands, it became important to distinguish between functional and non-functional TF binding, to reduce false-positive hits. For this, I combine the ChIP-seq data that identifies TF binding targets with RNA-seq data (of samples where the TF activity is inhibited) which helps in filtering for the genes that also see a change in transcriptional activity. There are no published RNA-seq data for ELT-3. For the DAF-16 and PQM-1, I used the RNA-seq data generated by Dr Laura Jones using the loss of function mutant *C. elegans* mutants *daf-16(mu86)* and *pqm-1(ok485)*. In addition to the standard temperature of 20°C, another set of RNA-seq data from experiments conducted at 35°C (heat shock) was also analysed. This heat shock dataset was included here because both TFs are responsive to environmental stress and might be more active during stress such as heat shock, potentially identifying a more complete set of downstream genes (Tepper, et al., 2013; Laura Jones et al., unpublished). The data was processed and analysed the same way as the pathogen response RNA-seq datasets. **Table 7.5** summarizes the number of genes that are up- and down-regulated as a result of the mutation in 20°C and 35°C (heat shock). The temperature seems to affect the up- and down-regulation in both mutants differently. In *daf-16(mu86)*, the heat-shocked dataset has more significantly down-regulated genes than the dataset at 20°C. On the other hand, the *pqm-1(ok485)* 35°C dataset has more significantly up-regulated genes than its 20°C dataset.

	<i>daf-16(mu86)</i>		<i>pqm-1(ok485)</i>	
	20°C	35°C	20°C	35°C
Up-regulation	66	41	33	84
Down-Regulation	63	75	25	20

Table 7.5 Number of up and down-regulated genes in the *daf-16(mu86)* and *pqm-1(ok485)* mutants at different temperatures. Significance is $|\log_2FC| > 0.6$ and p-value < 0.05 .

Comparisons of the up- and down-regulated genes between the two mutants at both temperatures show that there is a small but significant overlap of genes (**Figure 7.21b & d**: 3rd and 4th column from the left), indicating that PQM-1 and DAF-16 share a number of downstream genes.

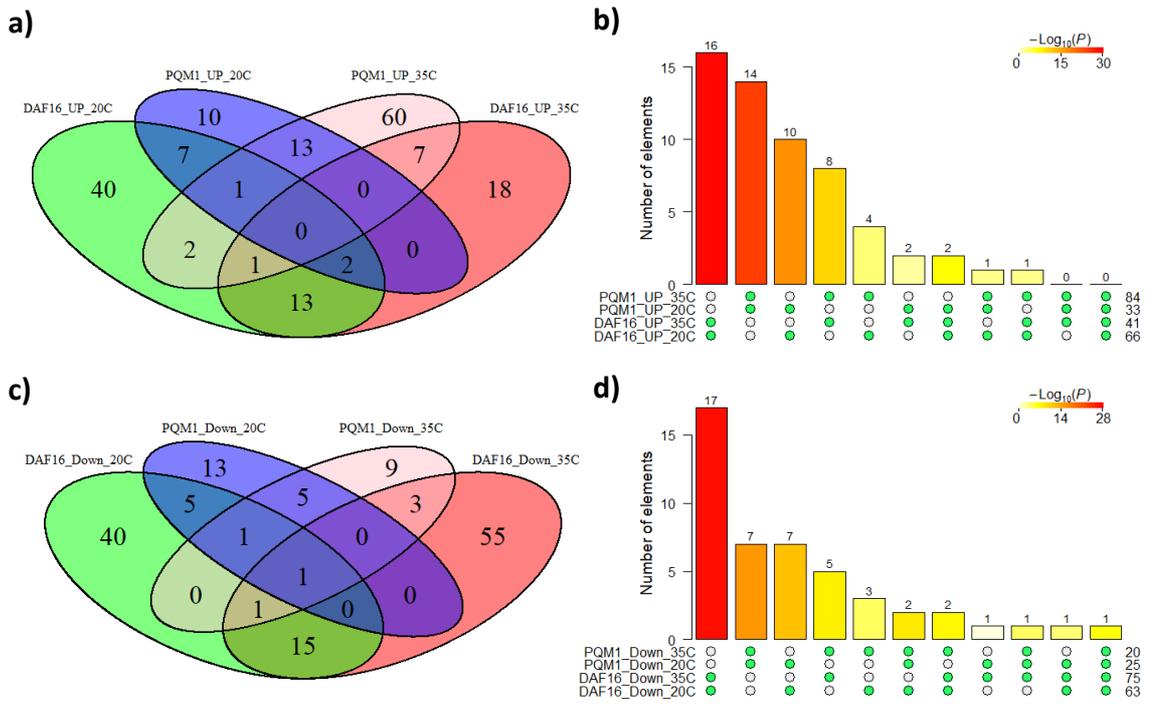


Figure 7.21 Comparison of the differentially expressed genes between *daf-16(mu86)* and *pqm-1(ok485)* at different temperatures. a) & b) Comparison of the up-regulated genes in *daf-16(mu86)* and *pqm-1(ok485)* mutants. c) & d) Comparison of the down-regulated genes in *daf-16(mu86)* and *pqm-1(ok485)* mutants.

Next, I compared the TF ChIP-seq target gene with their respective RNA-seq differentially expressed genes at both temperatures (**Figure 7.22**).

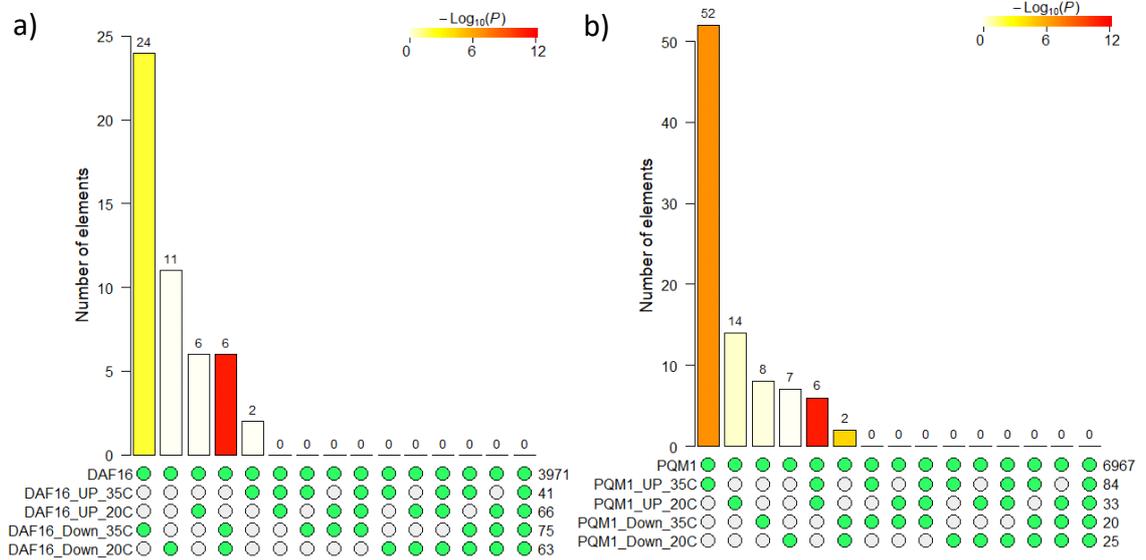


Figure 7.22 Overlap between TF ChIP-seq binding data and the RNA-seq differential expression data of DAF-16 and PQM-1. a) DAF-16: ChIP-seq samples were extracted from L4-Young Adult stages at 20°C b) PQM-1: ChIP-seq samples were extracted from L3 stage worms at 20°C. RNA-seq samples were extracted from the L4 stage at 20°C with and without 35°C heat shock. Green dots below the graph denotes which groups were overlapped.

Since TFs can act as transcriptional activators or repressors, inhibition of a TF should result in a decreased or increased transcription of their direct target genes, respectively. DAF-16 is suggested to be a transcriptional activator as the number of down-regulated genes (at 35°C)

are shared more significantly with the ChIP-seq data compared to the up-regulated genes (**Figure 7.22a** 1st column is more significant than 3rd and 5th column). PQM-1, on the other hand, shows the opposite, where the significant overlap is with the up-regulated genes (**Figure 7.22b** 1st column is more significant than 3rd and 5th column), indicating that PQM-1 could be a transcriptional repressor. Interestingly, the strong significance is only observed for the samples at 35°C (heat shock), which would infer that the TFs bind, but may not be active under normal condition and only activate under stress conditions such as heat shock.

Results from three different analysis (RNA-seq, *de novo* motif discovery and ChIP-seq) all identified *pqm-1* to be a key TF responsive to pathogen infection. These results support the idea that PQM-1 is associated with the innate immune response (Shapira, et al., 2006; O'Brien, et al., 2018). ChIP-seq data for PQM-1 and RNA-seq data using the *pqm-1(ok485)* mutant were only able to identify few direct targets of PQM-1, due to the significantly lower number of differential expressed genes in the RNA-seq dataset (**Figure 7.22b**). This may not be the complete set of genes regulated by PQM-1. Within the 383 pathogen response genes, a significant number of 249 genes were identified as PQM-1 binding targets in the ChIP-seq data. Furthermore, around half of the significantly differentially expressed genes in the *pqm-1(ok485)* RNA-seq data at 35°C are also relatively differentially expressed in a number of the pathogen datasets (Error! Reference source not found.). Both comparisons support the idea that PQM-1 dependent genes may play key roles in pathogen response. It must be noted that the genes identified here may not be the full list of PQM-1 target genes which may be due to PQM-1 not being fully active under the experimental conditions.

One explanation why *pqm-1* expression is up-regulated and not the other TFs (**Figure 7.17**) could be that this gene might have roles as a co-factor for other TFs during general stress conditions. Its role may be in the negative regulation (transcriptional suppressor) of metabolism. However, Gene Enrichment analysis on the 52 PQM-1 target genes (**Figure 7.22b** 1st column) does not return an enrichment in metabolism-related terms (**Appendix 17**). This may be due to the timing of sample collection, as the worms were left to recover for two hours after heat shock, rather than collected immediately, potentially missing the timepoint where metabolism is reduced and allowing it to return to normal levels. For example, the quiescence (a reduced activity which may be related to metabolism) observed during heat-shock is short-lived, with most wild-type worms (N2) showing normal locomotion and feeding behaviour 1-hour after heat shock (Hill, et al., 2014)

7.4. Change of the gene expression landscape as a result of heat shock

After obtaining the list of general pathogen responsive genes (list383 and list585), only the list of heat shock responsive genes is missing in order to be able to compare the genes responsive to the two different types of stressors (innate immunity and heat stress) and getting a step closer in answering the question of how biologically interconnected the response to these different types of stressors is.

Contrary to the popularity of research into heat stress, there are surprisingly few publicly available high-throughput heat shock datasets from *C. elegans*. In order to obtain the list of heat shock responsive genes, RNA-seq data from three publications: Brunquell, et al. (2016), Li, et al. (2016) and Haas, et al. (2018), as well as a dataset generated internally by Dr Laura Jones were used. Data processing and analysis were done the same way as the pathogen response (**Section 7.1.2**).

7.4.1. Differentially expressed genes are consistent across the heat shock datasets

The resulting number of differentially expressed genes for each of the heat shock datasets are summarized in the table below (**Table 7.6**). The differentially expressed genes from all 4 datasets were compared to identify how similar the datasets are (**Figure 7.23**).

Condition	Reference	Up-regulated Genes	Down-regulated Genes
L4, EV food, 33°C (30 min)	Brunquell, et al. (2016)	1366	868
L2, OP50 food, 34°C (30 min)	Li, et al. (2016)	1109	1853
L4, OP50 food, 34°C (75 min) + Recovery period	Haas, et al. (2018)	5203	2586
L4, OP50 food, 35°C (60 min) + Recovery period	Laura Jones	810	969

Table 7.6 Number of significantly differentially expressed genes for each of the 4 heat shock datasets. The conditions correspond to the developmental stage of the animal, bacterial diet and heat shock experimental setup (temperature and time). Haas, et al. (2018) included a recovery period of 20 minutes at room temperature followed by 2 hours at 20°C. Laura Jones heat shock experiment included a recovery period of 2 hours at 20°C post-heat-shock. Significantly differentially expressed genes are defined as genes having a $|\log_2FC| > 0.6$ and an adjusted p-value < 0.05 .

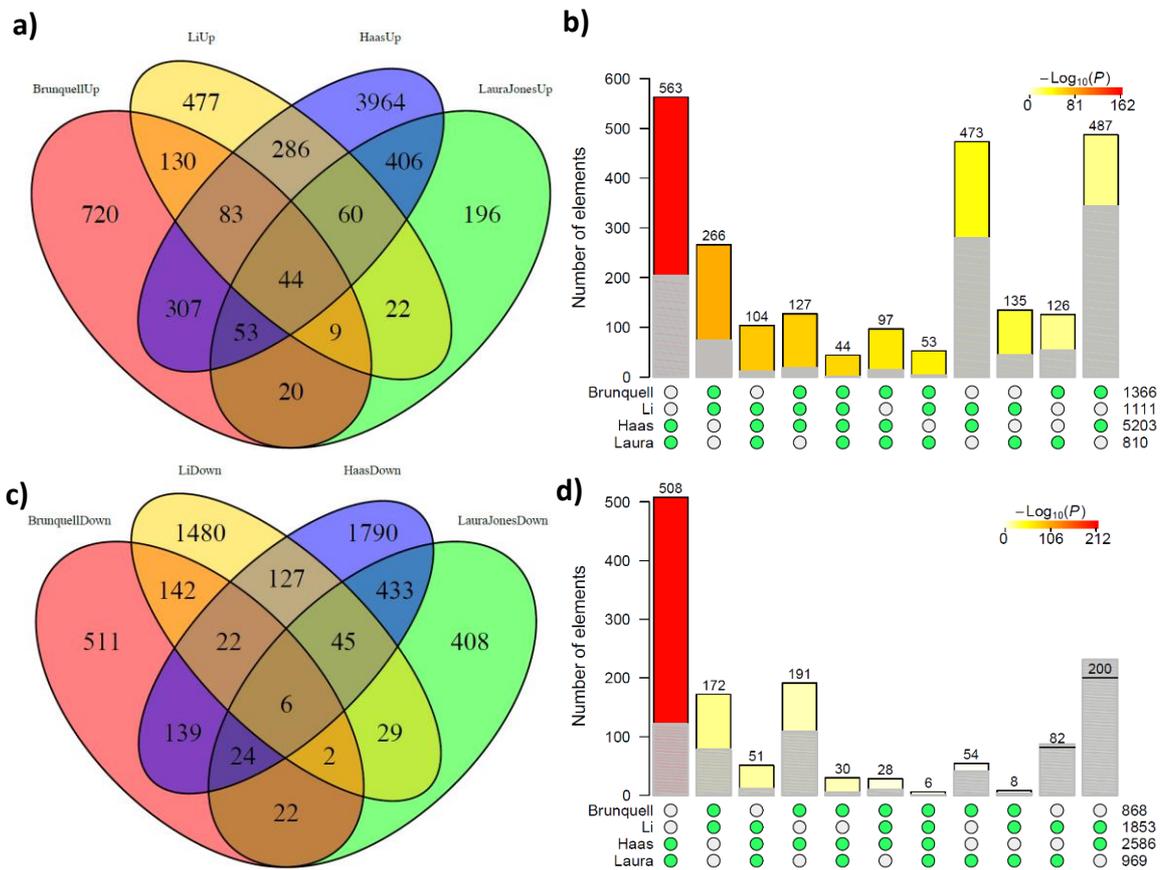


Figure 7.23 Comparison of up and down-regulated genes in the heat shock datasets. Venn diagram a) & c) and hypergeometric distribution testing b) & d) of all four heat shock RNA-seq datasets. a) & b) Up-regulated genes. c) & d) Down-regulated genes. Significance cut-off threshold is $p\text{-value} < 0.05$ and $|\log_2FC| > 0.6$. The grey shading overlaying each column represents the expected overlap.

There is some overlap between the up- and down-regulated genes, much higher than for the pathogen datasets (compared to **Figure 7.4** and **Figure 7.5**). Considering the similarity of the heat-shock regimes, they ought to deliver better correspondence in gene sets compared to the pathogen comparisons. Interestingly, the observed overlap of down-regulated genes in two comparisons (Li/Laura and Li/Haas) are lower than the expected overlap (**Figure 7.23d**), which might be due to Li, et al. (2016) using L2 stage worms.

The multi-filtering method and heatmap generation was also done for these heat shock datasets, to identify a potentially larger set of heat shock responsive genes (**Figure 7.24**). Different filtering cut-offs have been chosen to accommodate the smaller number but more consistent datasets. The first filtering criteria removes all genes where two or more datasets do not have a measurement for it. The second filtering criterion removed genes that on average shows a \log_2 fold change less than 0.5. In this case, genes were removed when the *absolute sum of $\log_2FC < 2$* across all experiments. At least 3 experiments must have a $|\log_2FC| > 0.6$ for the gene and all experiments need to have $|\log_2FC| > 0.3$ for the gene. The remaining genes were filtered by the p-value generated using Fisher's method. The final number of remaining genes was reduced from 21392 to 255.

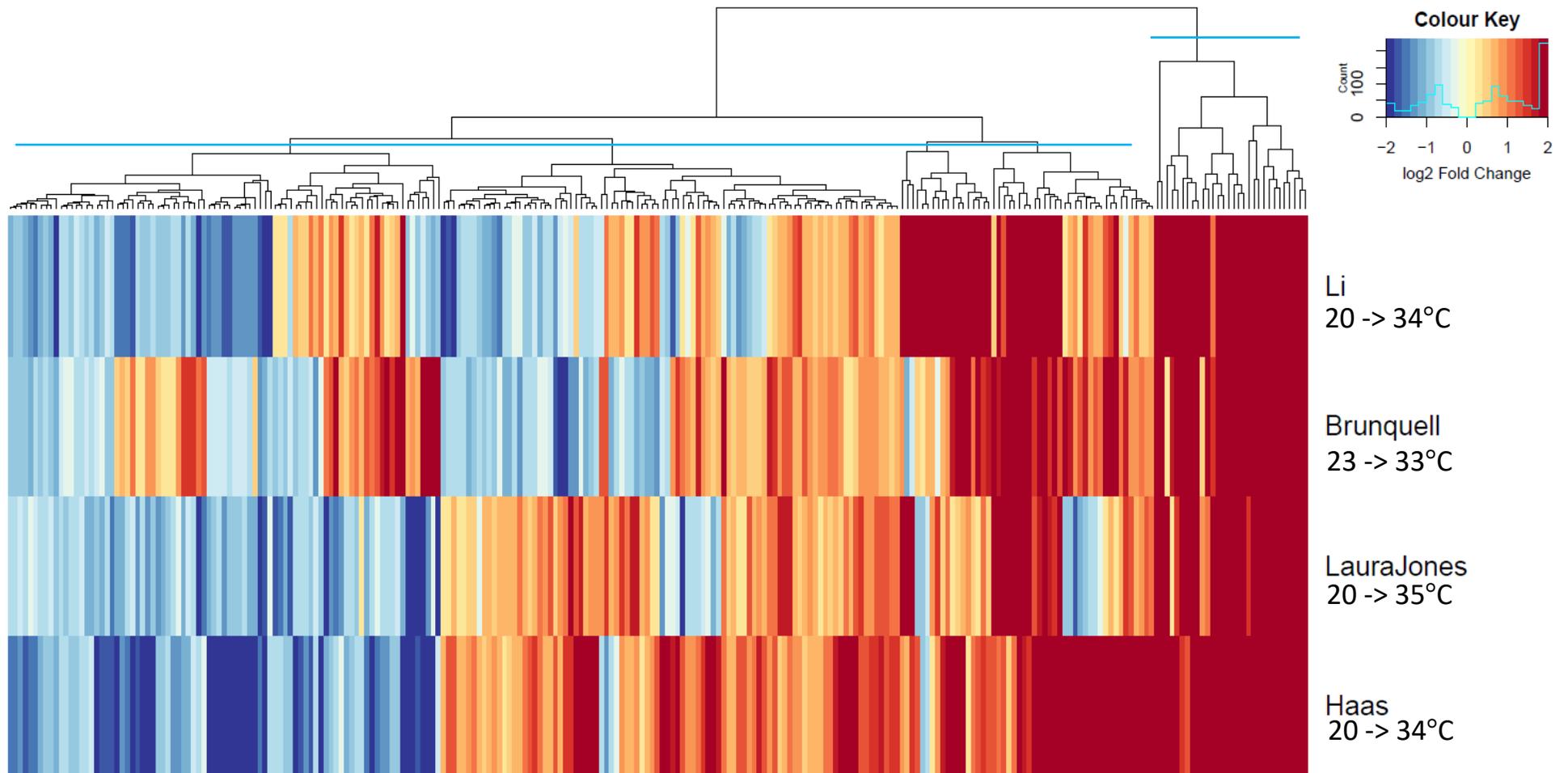


Figure 7.24 Heatmap of the 255 genes from the 4 heat shock datasets after filtering. The genes are placed along the x-axis while the datasets are on the y-axis. Datasets are named based on the first author or the person generating the dataset. The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes, and blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change. The horizontal blue line at the dendrogram indicates the cut-off level used to group genes into clusters.

The heatmap shows a clear clustering of up- and down-regulated genes. A small number of genes in certain datasets show differential expression patterns that do not agree with the other datasets, mainly in the middle region of the heatmap (**Figure 7.24**). This difference could be due to different conditions (e.g. temperature, developmental stage and food) (**Appendix 5**) or biological and technical variations. The 255 heat shock responding genes were analysed for gene set enrichment. The only significant tissue enrichment is the ‘epithelial system’. Enriched GO terms include terms related to cuticle structure and responses, such as ‘collagen trimer’ and ‘response to biotic stimuli’ (**Figure 7.25**). Interestingly, the immune system process GO term is also found in the result for HSR gene, indicating that perhaps some of the genes are general stress response genes or that heat shock and pathogen response have a relatively close relationship.

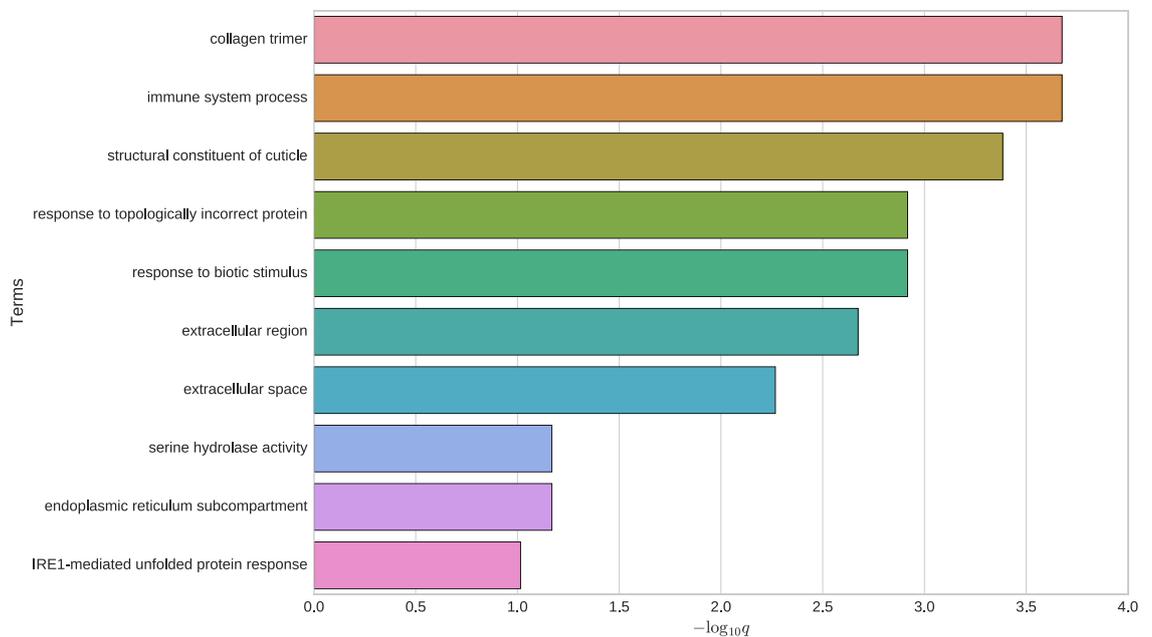


Figure 7.25 Wormbase Enrichment Analysis of the 255 HSR genes. Results only show Gene Ontology (GO) Enrichment Analysis. All terms have a q-value of lower than 0.05.

I next separated genes into clusters (**Figure 7.24** blue line), in the same way as for the pathogen response data and analysed for gene set enrichment and protein-protein interaction for each cluster of genes (**Figure 7.26**). Analysis with g:Profiler was also done; however, it showed fewer unique terms, but more levels of child terms. This was informative only for the cluster with the highly up-regulated genes, as it shows heat stress-related terms only present at higher GO levels, which Wormbase Enrichment Analysis did not include (**Figure 7.26** red).

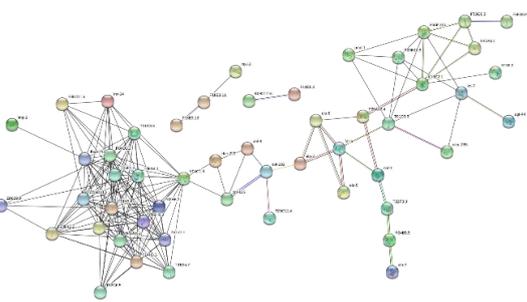
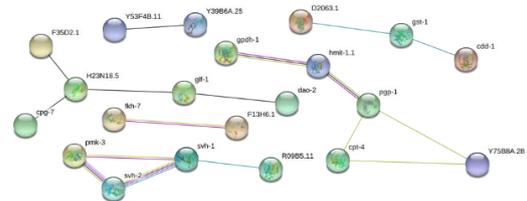
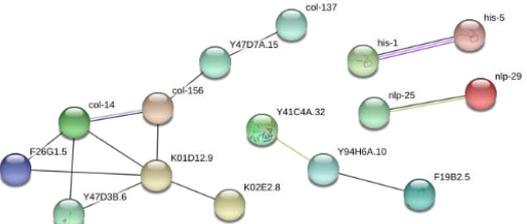
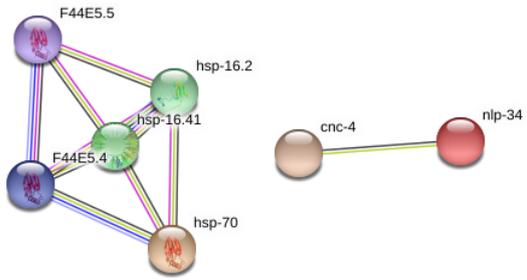
Heatshock255		STRING	# of genes	
Wormbase Enrichment Analysis				
Term	Q value			
intestine WBbt:0005772	0.0026		85	
cephalic sheath cell WBbt:0008406	0.034			
epithelial system WBbt:0005730	0.045			
AB WBbt:0004015	0.088			
peptidase activity GO:0008233	0.0095			
endoplasmic reticulum subcompartment GO:0098827	0.0095			
serine hydrolase activity GO:0017171	0.012			
ion homeostasis GO:0050801	0.017			
protein catabolic process GO:0030163	0.017			
immune system process GO:0002376	0.018			
collagen trimer GO:0005581	0.03			
organic acid metabolic process GO:0006082	0.033			
transmembrane transport GO:0055085	0.036			
response to biotic stimulus GO:0009607	0.036			
lytic vacuole GO:0000323	0.039			
multicellular organism growth GO:0035264	0.046			
outer labial sensillum WBbt:0005501	0.0069		90	
PVD WBbt:0006831	0.0069			
No enrichment			18	
structural constituent of cuticle GO:0042302	1.70E-08		32	
collagen trimer GO:0005581	1.70E-08			
protein heterodimerization activity GO:0046982	0.0044			
extracellular region GO:0005576	0.0044			
extracellular space GO:0005615	0.0044			
neuropeptide signalling pathway GO:0007218	0.0044			
response to topologically incorrect protein GO:0035966	0.0027		30	
IRE1-mediated unfolded protein response GO:0036498	0.0083			
response to biotic stimulus GO:0009607	0.025			
immune system process GO:0002376	0.035			
ATPase activity GO:0016887	0.038			
hydrolase activity acting on acid anhydrides GO:0016817	0.038			
GO:BP	stats			
Term name	Term ID			P _{adj}
response to heat	GO:0009408	2.042×10 ⁻⁴		
response to temperature stimulus	GO:0009266	8.571×10 ⁻⁴		
response to stress	GO:0006950	9.447×10 ⁻³		
cellular response to unfolded protein	GO:0034620	3.078×10 ⁻²		
response to abiotic stimulus	GO:0009628	3.166×10 ⁻²		
response to unfolded protein	GO:0006986	3.683×10 ⁻²		
cellular response to topologically incorrect protein	GO:0035967	4.219×10 ⁻²		
protein refolding	GO:0042026	4.982×10 ⁻²		

Figure 7.26 Wormbase Enrichment Analysis and STRING analysis for each cluster of genes based on the heatmap of the heat shock datasets (Figure 7.24). The order of cluster from top to bottom corresponds to the cluster on the heatmap from left to right. The number of genes in each cluster are shown in the right-most column, and the colour corresponds to the colouring in Appendix 20. Wormbase Enrichment Analysis results include tissue and GO term enrichment analysis. Tissue enrichment terms have the ID prefix “WBbt” and GO terms have the prefix “GO”. STRING excludes disconnected nodes. g:Profiler results for GO: Biological Processes were included only for the last gene cluster.

The enrichment analysis on separate clusters returned more significant GO terms compared to the analysis on the whole set of 255 heat shock responsive genes. Of surprising interest are the GO terms that were also found enriched in pathogen response such as ‘Intestine’, ‘Immune system process’ and ‘response to biotic stimulus’. In the up-regulated cluster, ‘response to heat’ and ‘response to topologically incorrect protein’ is found, which correlates with the expected response to heat shock in *C. elegans*. The STRING analysis returned relatively low protein-protein interaction when compared to the pathogen response data (**Figure 7.12 & Figure 7.13**).

7.4.2. Heat shock responsive genes are related to the immune system

The analysis on the four heat-shock datasets shows a relative consistent expression of the same set of genes. In total, 255 genes were identified to be consistently differentially expressed under heat shock (**Figure 7.24**). Enrichment analysis identified high enrichment in the immune system process (**Figure 7.25**). This could indicate that some genes classified as immune system process are in fact general stress response genes rather than specific to the immune system, or it could mean that both the HSR and the immune response share overlapping pathways and mechanisms. Terms related to the cuticle structure (‘collagen timer’ and ‘structural constituent of cuticle’) were also among the highest-ranking terms. However, their function with regards to heat-shock is unknown. Brunquell, et al. (2016) also found such terms in their heat-shock study and commented on these genes having signal transduction function and may relay signals to stress-specific TFs.

Detailed analysis of the up- and down-regulated clusters of the heatmap identified tissue enrichment of ‘PVD’ and ‘outer labial sensillum’ (**Figure 7.26**), that was also enriched in the pathogen datasets, implying increased sensory activity to defend or avoid dangerous environments. Within the down-regulated set of genes, we find metabolic processes that may function the same way as explained for the pathogen response: to change from a passive metabolic orientated transcriptional instruction to an active life-preserving one. The cluster of up-regulated genes sees an increase in both the ‘immune system process’ as well as heat shock-related terms, demonstrating that both the immune and HSR pathways work simultaneously against heat stress.

STRING analysis returned relatively low protein-protein interaction compared to the pathogen response genes, which suggests that the HSR might be relatively small and

concentrated, compared to pathogen response. It seems logical when considering that temperature can only either increase or decrease and does not act through a large variety of mechanism, which is the case for pathogen infections. There is a concentrated network of co-expressed proteins in the down-regulated gene cluster (**Figure 7.26** blue). Unfortunately, many of those genes have not been researched, and it remains to be determined what their role is.

Generating heatmaps for the Immune Effectors proteins show in general very little differential expression in the immune effector proteins (**Appendix 18**), except for caenacins, especially *cnc-4* which is strongly up-regulated in all heat shock datasets (**Figure 7.27**). While caenacins were found up-regulated in various *C. elegans* pathogen studies, especially *D.coniospora*, little is known about their functions. This family of protein is closely related to NLPs with signal peptides at their N-terminus and confers resistance against *D. coniospora* (Couillault, et al., 2004; Dierking, et al., 2016). There has been no emphasis on caenacins in other stress-related studies, making *cnc-4* an interesting gene to do further research on. Other individual genes strongly up-regulated in all heat shock datasets are *nlp-25*, *nlp-30*, *nlp-34* and *clec-196* (**Appendix 18**), all of which are not well researched as well.

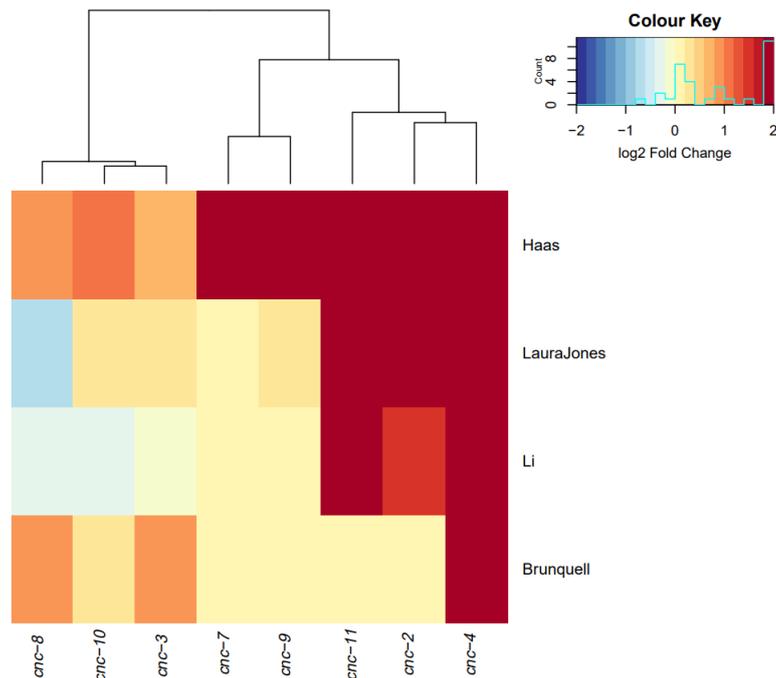


Figure 7.27 Heatmap of the heat shock datasets for caenacin (CNC). The genes are placed along the x-axis while the datasets are on the y-axis. Datasets are named based on the first author or the person generating the dataset. The data is grouped based on hierarchical clustering using complete linkage and Euclidean distance measure. Red cells are up-regulated genes, and blue cells are down-regulated genes. The colour coding is capped at 2 and -2 log₂ Fold Change.

7.5. Comparison of pathogen response genes and heat shock response genes

After identifying the general pathogen responsive genes (list331, list585, list383) and heat shock responsive genes (heatshock255) from section 7.2.2 and 7.4.2, respectively, all the required information is in place to answer the question as to how related the HSR and the pathogen response is. In the following section, I compare the pathogen responsive genes with the heat shock responsive genes to find shared genes and analyse potential links.

I started with comparing all the three pathogen responsive gene lists with the heat shock responsive gene list (**Figure 7.28**).

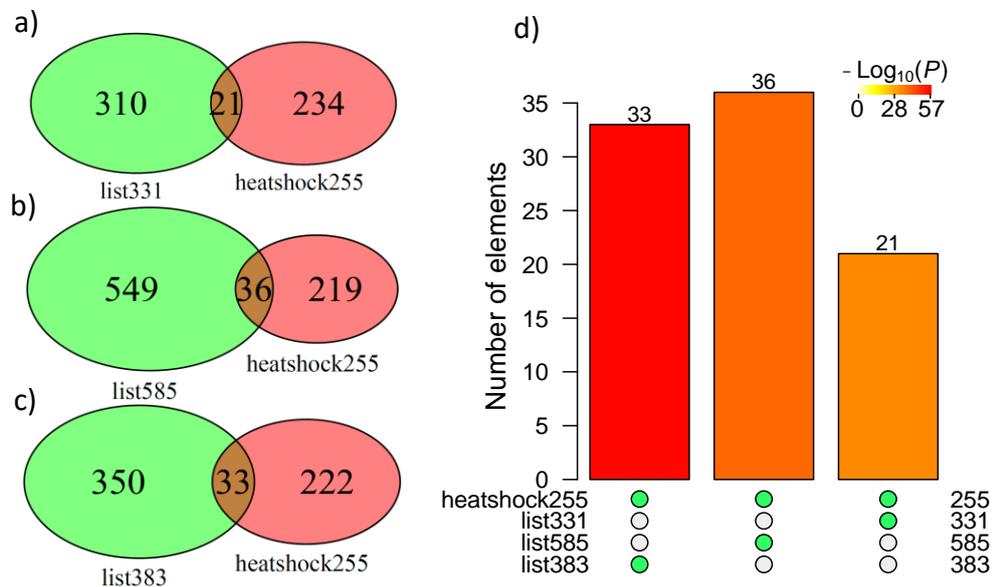


Figure 7.28 Comparison of pathogen responsive gene lists with the heat shock gene list. a) – c) Venn diagram showing the number of overlapping genes between pathogen and heat shock responsive gene list. d) Hypergeometric test of the intersecting gene lists. Green dots below the graph denotes which groups were overlapped.

In each of the pathogen responsive gene lists 6.2-8.6% of the genes are also heat shock responsive (list331 = 6.3%, list585 = 6.2%, list383 = 8.6%). While these numbers might be similar, the actual genes in each of the intersects could be different, similar to how each pathogen responsive gene list is different to some extent (**Figure 7.10**). The total number of unique genes in all three intersects are 50 genes (**Figure 7.29a** and **Appendix 19** for the gene list), which can be considered genes that are responsive to both pathogens and heat stress. Comparing each of the intersects, it can be seen that the 21 heat shock responsive genes in list331 (heat331) are a subset of the 36 heat shock responsive genes in list585 (heat585). Comparison of the heat shock responsive genes between list585 (heat585) and list383

(heat383) shows that around 1/2 of the genes are in common, with 14 genes being exclusive to list383, while 17 are missing from it (**Figure 7.29a**).

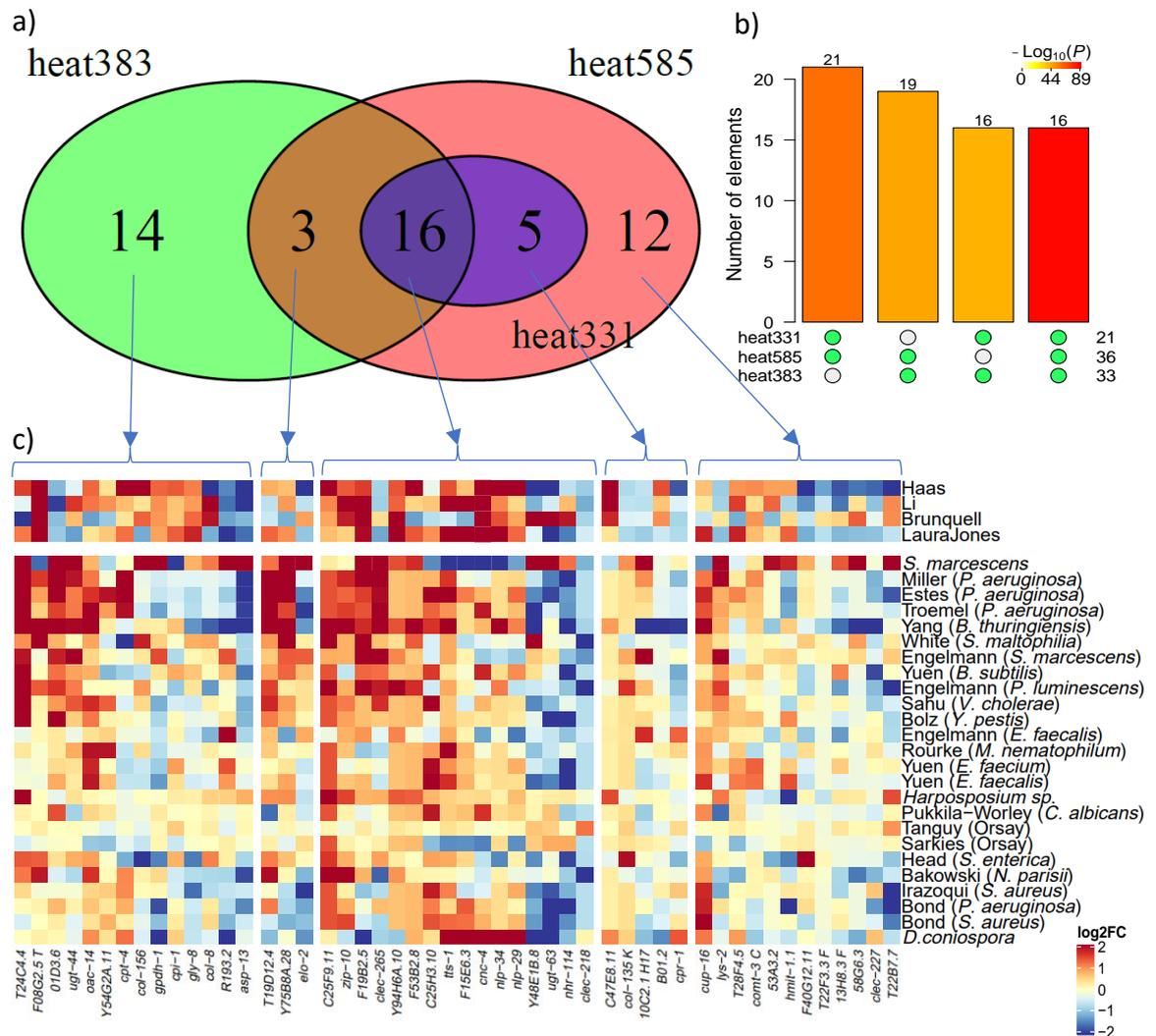


Figure 7.29 Comparison of the genes in each pathogen responsive gene list that are also in the heat shock responsive gene list. a) Venn diagram showing the similarity of the heat shock responsive genes in each of the pathogen responsive list. b) Hypergeometric test of the intersecting gene lists from a). Green dots below the graph denotes which groups were overlapped. c) Heatmap visualization of the genes that are both pathogen and heat shock responsive corresponding to each list and intersections (blue arrow) for each dataset. Datasets are named based on the first author or the person generating the dataset and if applicable, the pathogen used. Heat331 = intersection of list331 and heatshock255. Heat585 = intersection of list585 and heatshock255. Heat383 = intersection of list383 and heatshock255.

The 16 “high confidence” genes are common in all three lists (heat331, heat585 and heat383) show overall stronger differential expression (darker red and blue shades) compared to the other sets (**Figure 7.29c**). However, the other sets cannot be disregarded as there is still a relative consistent degree of differential expression.

Enrichment analysis on all 50 genes resulted in the immune response related terms such as ‘immune system process’, ‘response to biotic stimulus’ and ‘collagen trimer’ (**Figure 7.30a**). It must be noted however that the GO-term database is not complete and is updated very

frequently, as such, some genes may be placed in specific GO-terms due to the lack of information. I propose that the genes found here could be more accurately defined as heat & pathogen stress (or general stress) genes rather than specific to the immune response. The only significant tissue enrichment was the ‘intestine’.

Interestingly, STRINGs functional enrichment tool identified enrichment in the ‘von Willebrand factor type A domain’. This domain is found in various plasma proteins and is associated with haemostasis and various disease (InterPro, n.d.). This domain was also reported by Wong, et al., (2007). However, its function in *C. elegans* is unknown.

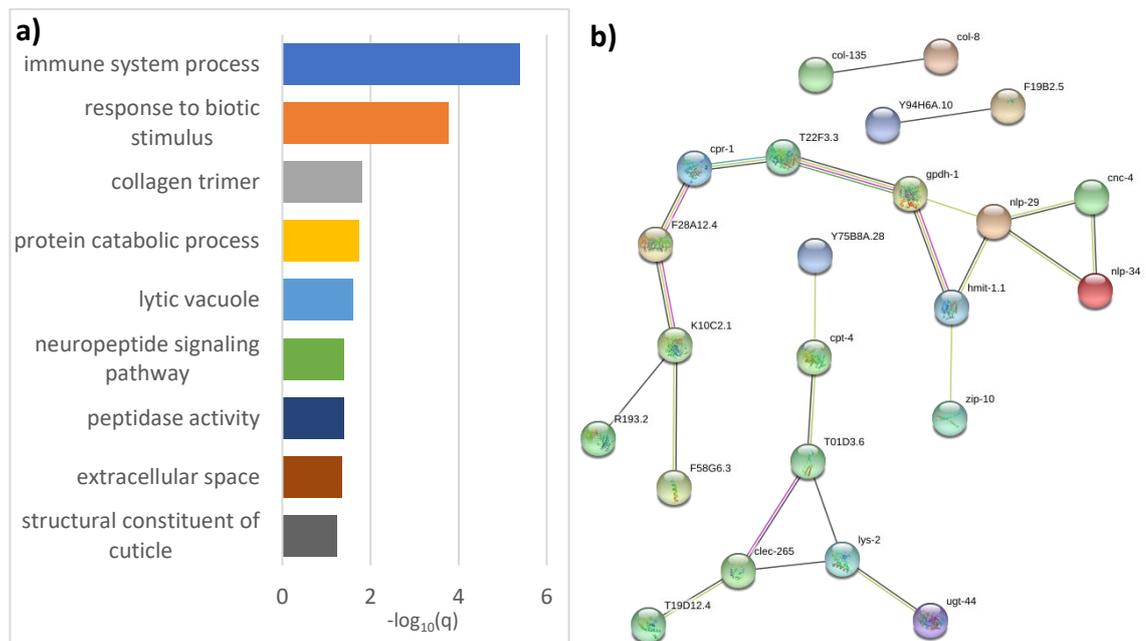


Figure 7.30 Wormbase Enrichment Analysis and STRING analysis for the 50 genes that are responsive to both pathogen infection and heat shock. a) Wormbase Enrichment Analysis including only the GO term enrichment results. b) STRING analysis. Disconnected nodes are excluded.

The STRING analysis found some protein-protein interaction in the form of two networks that look like a signalling cascade or a chain reaction due to their linear shape (**Figure 7.30b**). Each stressor may trigger multiple different signalling cascades, and some of these may be shared between different stressors, which may be the case for these two. It would be interesting to see how each of these proteins affect different stress responses and whether these networks are indeed signalling cascade pathways.

Within the 50 heat and pathogen responsive genes, the top 3 up-regulated genes are *C25F9.11*, *zip-10*, *F19B2.5* and the top 3 down-regulated genes are *nbr-114*, *clec-218*, *ugt-63* (based on the absolute mean expression and standard deviation (**Appendix 19**)).

C25F9.11 is enriched in multiple neurons including PVD and OLL and is affected by multiple TFs, including the two stress-related TFs *daf-16* and *hsf-1*. This gene is also affected

by various chemicals such as ethanol (Peltonen, et al., 2013). Considering that this gene is expressed in sensory neurons, I would hypothesize that its function is related to sensing environmental stressors. Its consistent upregulation in all the pathogen and heat stress implies that the nematodes sensory function is enhanced to avoid further stressors or find safer environments. This sensory role is further supported as *C25F9.11* expression is enhanced upon contact with various chemicals.

F19B2.5 is similarly enriched in multiple neurons, including PVD and OLL. It also shows changes in expression under osmotic stress and cadmium treatment (oxidative stress). A decreased expression is observed under various antibiotic treatments (Admasu, et al., 2018; Koh, et al., 2018). These expressional changes imply that *F19B2.5* could have a similar role as *C25F9.11* in sensing environmental stressors.

zip-10 is a bZIP-transcription factor whose expression depends on temperature changes and is controlled by *mir-60* and *isy-1*. It is observed to enhance phenoptosis (programmed organismal death) under cold shock (Jiang, et al., 2018). My results support a temperature-dependent expression, but also extends the expressional response further to include pathogen infection. Whether the phenoptosis function also extends towards pathogenic stress remains to be determined.

nbr-114 is part of the nuclear hormone receptor family and is predicted to have transcription factor activity (MacNeil, et al., 2015). It has a role in fertility that is dependent on the bacterial diet with respect to the availability of the amino acid tryptophan. This suggests that *nbr-114* has a role in buffering against dietary metabolites (Gracida & Eckmann, 2013). Based on its role related to metabolism, its down-regulation might be associated with a reduction in metabolism, which correlates with the hypothesis that under stress conditions, *C. elegans* reduces metabolism to allocate more resources towards the production of stress proteins.

lec-218 is a relatively unknown c-type lectin with limited research. Expression profiling shows that *lec-218* is expressed in many, if not all, neurons in the early stages of development. Furthermore, it is also expressed in the intestine and various muscles such as the pharyngeal muscle (Spencer, et al., 2011). Lectins (and c-type lectins) are carbohydrate-binding proteins and have been observed to function in a wide variety of biological processes (Drickamer, 1993). Specifically, *lec-218* may play a similar role as the other consistently down-regulated genes in reducing metabolism. The expression in the muscles and neurons could imply a behavioural response against stress, such as the sleep-like quiescence program observed following various stress conditions (Hill, et al., 2014).

ugt-63 is part of the family of UDP-Glucuronosyl transferases and plays a role in phase II detoxification (Ladage, et al., 2016). It has been shown that RNAi targeting *ugt-63* increases anoxia survivability (Ladage, et al., 2016). This observation combined with the down-regulation of it in most of the datasets in this study indicates that *ugt-63* has a negative impact on stress resistance. While this may be contradictory as *ugt-63* plays a role in phase II detoxification, its transferase role might be limited to endobiotic (chemicals originating from the organism) toxins rather than xenobiotic (chemicals found in the organism that is not naturally produced by it) ones, returning back to the argument of reduced metabolism under stress.

Another interesting finding is that heat shock protein, normally associated with the HSR, are also differentially expressed during pathogen infections. Specifically, the small heat shock proteins (*hsp-12.3*, *hsp-16.2*, *hsp-16.41* and *hsp-17*) show strong up-regulation in around a quarter to half of the pathogen datasets (**Appendix 13b**). The response to pathogen infections increases the number of misfolded proteins due to the large surge of stress protein production (Zügel & Kaufmann, 1999). As different pathogens induce the expression of different genes, the variety of proteins become relatively large. Since small heat shock proteins bind a wide range of proteins (Haslbeck, et al., 2005), it is not surprising that these proteins are up-regulated to manage the large variety of proteins that are produced as a result of different pathogen infections.

7.6. Conclusion and future work of the stress resistance study

The cellular stress response and the immune response play a vital part in providing protection and preserving well-being for organisms against external and internal stressors. Understanding these response pathways can open up the potential to utilize and enhance these natural defence mechanisms for medical purposes. The aim of this study is to investigate the association between the innate immune response and the HSR in *C. elegans*. By systematically reviewing published data and literature using a bioinformatics approach, this study enhances our knowledge about the innate immune response and the HSR.

In summary, the findings of this study suggest that highly differentially expressed genes as a result of pathogen infection is largely pathogen-specific. Genes that were differentially expressed in most of the pathogen datasets were less strongly differentially expressed. For these general pathogen responsive genes, the up-regulated genes tended to be associated with defence response, while down-regulated genes were often associated with metabolic processes. This suggests that under pathogenic influences, *C. elegans* responds by changing their transcriptional instruction from a passive cell maintenance orientated one to an active defence focused one. The cellular defence mechanisms include the upregulation of membrane proteins likely to function as receptors to sense pathogenic or signalling molecules, or as transport proteins to transport signalling and immune response proteins out of the cell and supplies into the cell. For the heat shock data, the analysis found 255 differentially expressed genes across the four datasets, which show enrichment for both the heat shock and immune response GO terms. Overall, 50 genes are shared among the heat shock and pathogen infection datasets, indicating that these two responses overlap significantly. Additionally, an interesting observation was made, that showed that PQM-1 is the only pathogen response related TF that is significantly up-regulated in multiple pathogen datasets.

The degree to which the conclusion from this bioinformatic investigation reflects the “real” biology is difficult to assess. Statistical testing, for example, assumes complete randomness, which may not be an accurate reflection of the reality. Possible improvements would be to increase the p-value threshold but at the expense of missing true positive hits, or cross-validating the results from one software by other software (e.g. use EdgeR and limma to validate the DEseq2 results). Another point to consider is that the data used here comes from high throughput screening experiments done by different research groups and were designed differently, which likely negatively impacted the signal-to-noise ratio. However, it

can be argued that any signal that comes through the noise would be very robust as it is not affected by differences in experimental design and conduct. Future research would combine systematically designed and consistent high throughput screening experiments with computational analysis and emphasize other variables such as treatment time (e.g. length of exposure to pathogen or heat). This would allow the differentiation of genes into fast and slow response genes and improve the confidence of the gene hits. The next steps would then be to validate the gene hits identified by the computational analysis through biological “wet-lab” experiments. Here, the top differentially expressed genes (**Appendix 19**) provide a good starting point, to take forward for experimental laboratory testing.

References

- Aballay, A. & Ausubel, F. M., 2001. Programmed Cell Death mediated by *ced-3* and *ced-4* protects *Caenorhabditis elegans* from *Salmonella typhimurium*-mediated Killing. *Proceedings of the National Academy of Science*, 98(5), pp. 2735-2739.
- Aballay, A., Drenkard, E., Hilbun, L. R. & Ausubel, F. M., 2003. *Caenorhabditis elegans* Innate Immune Response triggered by *Salmonella enterica* requires intact LPS and is mediated by a MAPK Signaling Pathway. *Current Biology*, 13(1), pp. 47-52.
- Aballay, A., Yorgey, P. & Ausubel, F. M., 2000. *Salmonella typhimurium* proliferates and establishes a persistent Infection in the Intestine of *Caenorhabditis elegans*. *Current Biology*, 10(23), pp. 1539-1542.
- Admasu, T. D. et al., 2018. Drug Synergy slows Aging and improves Healthspan through IGF and SREBP Lipid Signaling. *Developmental Cell*, 47(1), pp. 67-79.
- Agilent Technologies, 2016. *GPL11346*. [Online]
Available at: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL11346>
[Accessed 2 October 2018].
- Aguilera, A. & Garcia-Muse, T., 2012. R loops. From Transcription Byproducts to Threats to Genome Stability. *Molecular Cell*, 46(2), pp. 115-124.
- Aguilera, A. & Huertas, P., 2003. Cotranscriptionally formed DNA:RNA Hybrids mediate Transcription Elongation Impairment and Transcription-Associated Recombination. *Molecular Cell*, 12(3), pp. 711-721.
- Ahringer, J., 2005. *Reverse genetics*. [Online]
Available at:
http://www.wormbook.org/chapters/www_introreversegenetics/introreversegenetics.html
[Accessed 8 July 2019].
- Alberts, B. et al., 2002. *Molecular Biology of the Cell*. 4th ed. New York: Garland Science.
- Al-Hadid, Q. & Yang, Y., 2016. R-loop: an emerging Regulator of Chromatin Dynamics. *Acta Biochimica et Biophysica Sinica*, 48(7), pp. 623-631.
- Allison, D. F. & Wang, G. G., 2019. R-loops: Formation, Function, and Relevance to Cell Stress. *Cell Stress*, 3(2), pp. 38-47.
- Altschul, S. F. et al., 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), pp. 403-410.
- Altun, Z. & Hall, D., 2009. *WormAtlas: Introduction to C. elegans Anatomy*. [Online]
Available at: <http://www.wormatlas.org/hermaphrodite/introduction/Introframeset.html>
[Accessed 20 March 2019].
- Altun, Z. & Hall, D., 2010. *Nervous System, Neuronal Support Cells*. [Online]
Available at: <https://www.wormatlas.org/hermaphrodite/neuronalsupport/mainframe.htm>
[Accessed 16 August 2019].
- Ankar, J. & Sistonen, L., 2007. Heat Shock Factor 1 as a Coordinator of Stress and Developmental Pathways. In: P. Csermely & L. Vigh, eds. *Molecular Aspects of the Stress Response: Chaperones, Membranes and Networks*. New York: Springer, pp. 78-88.
- Andersen, E. C., Lu, X. & Horvitz, R., 2006. *C. elegans* ISWI and NURF301 antagonize an Rb-like Pathway in the determination of multiple Cell Fates. *Development*, 133, pp. 2695-2704.
- Anderson, A. & McMullan, R., 2018. Neuronal and non-Neuronal Signals regulate *Caenorhabditis elegans* Avoidance of contaminated Food. *Philosophical Transactions of the Royal Society B*, 373(1571), article no: 20170255.
- Andrews, S., 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. [Online]
Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
[Accessed 20 November 2018].

- Andrews, S., 2013. *BamQC*. [Online]
Available at: <https://github.com/s-andrews/BamQC>
[Accessed 21 November 2018].
- Angeles-Albores, D., Lee, R. Y., Chan, J. & Sternberg, P. W., 2016. Tissue Enrichment Analysis for *C. elegans* Genomics. *BMC Bioinformatics*, 17, article no: 366.
- An, J. H. & Blackwell, T. K., 2003. SKN-1 links *C. elegans* Mesendodermal Specification to a conserved Oxidative Stress Response. *Genes & Development*, 17(15), pp. 1882-1893.
- An, W., 2007. Histone Acetylation and Methylation. In: T. K. Kundu, et al. eds. *Chromatin and Disease. Subcellular Biochemistry, vol 41*. Dordrecht: Springer, pp. 355-374.
- Arab, K. et al., 2019. GADD45A binds R-loops and recruits TET1 to CpG Island Promoters. *Nature Genetics*, 51(2), pp. 217-223.
- Ardehali, M. B. et al., 2011. *Drosophila* Set1 is the major Histone H3 Lysine 4 Trimethyltransferase with Role in Transcription. *The EMBO Journal*, 30(14), pp. 2817-2828.
- Ariumi, Y., 2014. Multiple Functions of DDX3 RNA Helicase in Gene Regulation, Tumorigenesis, and Viral Infection. *Frontiers in Genetics*, 5, article no: 423.
- Baaklini, I. et al., 2004. RNase HI overproduction is required for efficient Full-Length RNA Synthesis in the absence of Topoisomerase I in *Escherichia coli*. *Molecular Microbiology*, 54(1), pp. 198-211.
- Bachman, K. E. et al., 2003. Histone Modifications and Silencing prior to DNA Methylation of a Tumor Suppressor Gene. *Cancer Cell*, 3(1), pp. 89-95.
- Badenhorst, P., Voas, M., Rebay, I. & Wu, C., 2002. Biological Functions of the ISWI Chromatin Remodeling Complex NURF. *Genes & Development*, 16(24), pp. 3186-3198.
- Bailey, T. L., 2011. DREME: Motif Discovery in Transcription Factor ChIP-seq Data. *Bioinformatics*, 27(12), pp. 1653-1659.
- Bailey, T. L. & Machanick, P., 2012. Inferring direct DNA Binding from ChIP-seq. *Nucleic Acids Research*, 40(17), article no: e128.
- Bakowski, M. A. et al., 2014. Ubiquitin-Mediated Response to Microsporidia and Virus Infection in *C. elegans*. *PLoS Pathogens*, 10(6), article no: e1004200.
- Bannister, A. J. & Kouzarides, T., 2011. Regulation of Chromatin by Histone Modifications. *Cell Research*, 21(3), pp. 381-395.
- Barna, J., Csermely, P. & Vellai, T., 2018. Roles of Heat Shock Factor 1 beyond the Heat Shock Response. *Cellular and Molecular Life Sciences*, 75(16), pp. 2897-2916.
- Barna, J. et al., 2012. Heat Shock Factor-1 intertwines Insulin/IGF-1, TGF- β and cGMP Signaling to Control Development and Aging. *BMC Developmental Biology*, 12(32).
- Barreto, G. et al., 2007. Gadd45a promotes Epigenetic Gene Activation by Repair-Mediated DNA Demethylation. *Nature*, 445, pp. 671-675.
- Barski, A. et al., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4), pp. 823-837.
- Battisti, J. M. et al., 2017. Analysis of the *Caenorhabditis elegans* Innate Immune Response to *Coxiella burnetii*. *Innate Immunity*, 23(2), pp. 111-127.
- Baugh, L. R., 2013. To grow or not to grow: Nutritional Control of Development during *Caenorhabditis elegans* L1 Arrest. *Genetics*, 194(3), pp. 539-555.
- Becherel, O. J. et al., 2013. Senataxin plays an essential Role with DNA Damage Response Proteins in Meiotic Recombination and Gene Silencing. *PLoS Genetics*, 9(4), article no: e1003435.
- Beletskii, A. & Bhagwat, A. S., 1996. Transcription-induced Mutations: Increase in C to T Mutations in the nontranscribed Strand during transcription in *Escherichia coli*. *Proceedings of the National Academy of Science*, 93(24), pp. 13919-13924.

- Benedik, M. J. & Styrch, U., 1998. *Serratia marcescens* and its Extracellular Nuclease. *FEMS Microbiology Letters*, 165(1), pp. 1-13.
- Beurton, F. et al., 2019. Physical and functional Interaction between SET1/COMPASS Complex Component CFP-1 and a Sin3S HDAC Complex in *C. elegans*. *Nucleic Acids Research*, 47(21), pp. 11164-11180.
- Beyer, D. C., Ghoneim, M. K. & Spies, M., 2013. Structure and Mechanisms of SF2 DNA Helicases. In: M. Spies, ed. *DNA Helicases and DNA Motor*. New York: Springer, pp. 47-73.
- Bhati, V. et al., 2014. BRCA2 prevents R-loop Accumulation and associates with TREX-2 mRNA Export Factor PCID2. *Nature*, 511(7509), pp. 362-365.
- Bianco, J. N. & Schumacher, B., 2018. MPK-1/ERK Pathway regulates DNA Damage Response during Development through DAF-16/FOXO. *Nucleic Acids Research*, 46(12), pp. 6129-6139.
- Bindea, G. et al., 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped Gene Ontology and Pathway Annotation Networks. *Bioinformatics*, 25(8), pp. 1091-1093.
- Bird, A., 2002. DNA Methylation Patterns and Epigenetic Memory. *Genes & Development*, 16(1), pp. 6-21.
- Biswas, S. & Rao, C. M., 2018. Epigenetic Tools (the Writers, the Readers and the Erasers) and their implications in Cancer Therapy. *European Journal of Pharmacology*, 837, pp. 8-24.
- Black, J. C. & Whetstone, J. R., 2011. Chromatin Landscape: Methylation beyond Transcription. *Epigenetics*, 6(1), pp. 13-19.
- Blaxter, M. & Bird, D., 1997. Chapter 30. Parasitic Nematodes. Section IV the Nematode Surface. In: D. L. Riddle, T. Blumenthal, B. J. Meyer & J. R. Priess, eds. *C. elegans II, 2nd edition*. New York: Cold Spring Harbor.
- Bledau, A. S. et al., 2014. The H3K4 Methyltransferase Setd1a is first required at the Epiblast Stage, whereas Setd1b becomes essential after Gastrulation. *Development*, 141(5), pp. 1022-1035.
- Blumenthal, T., 2012. *Trans-splicing and Operons*. [Online] Available at: http://www.wormbook.org/chapters/www_transsplicingoperons/transsplicingoperons.html [Accessed 29 July 2019].
- Bochman, M., Sabouri, N. & Zakian, V. A., 2010. Unwinding the Functions of the Pif1 Family Helicases. *DNA Repair*, 9(3), pp. 237-249.
- Boguslawski, S. J. et al., 1986. Characterization of Monoclonal Antibody to DNA · RNA and its application to Immunodetection of Hybrids. *Journal of Immunological Methods*, 89(1), pp. 123-130.
- Bolz, D. D., Tenor, J. L. & Aballay, A., 2010. A Conserved PMK-1/p38 MAPK is required in *Caenorhabditis elegans* Tissue-specific Immune Response to *Yersinia pestis* Infection. *The Journal of Biological Chemistry*, 285(14), pp. 10832-10840.
- Bond, M. R., Ghosh, S. K., Wang, P. & Hanover, J. A., 2014. Conserved Nutrient Sensor O-GlcNAc Transferase is integral to *C. elegans* Pathogen-Specific Immunity. *PLoS One*, 9(12), article no: e113231.
- Boque-Sastre, R. et al., 2015. Head-to-Head Antisense Transcription and R-loop Formation promotes Transcriptional Activation. *Proceedings of the National Academy of Science*, 112(18), pp. 5785-5790.
- Boule, J.-B. & Zakian, V. A., 2007. The Yeast Pif1p DNA Helicase preferentially unwinds RNA-DNA Substrates. *Nucleic Acids Research*, 35(17), pp. 5809-5818.
- Boulton, S. J. et al., 2002. Combined Functional Genomic Maps of the *C. elegans* DNA Damage Response. *Science*, 295(5552), pp. 127-131.
- Boulton, S. J. et al., 2004. BRCA1/BARD1 Orthologs required for DNA Repair in *Caenorhabditis elegans*. *Current Biology*, 14(1), pp. 33-39.

- Brandt, J. P. & Ringstad, N., 2015. Toll-like Receptor Signaling promotes the Development and Function of Sensory Neurons required for a *C. elegans* Pathogen-avoidance Behavior. *Current Biology*, 25(17), pp. 2228-2237.
- Breger, J. et al., 2007. Antifungal Chemical Compounds identified using a *C. elegans* Pathogenicity Assay. *PLoS Pathogens*, 3(2), article no: e18.
- Brown, D. A. et al., 2017. The SET1 Complex selects actively transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Reports*, 20(10), pp. 2313-2327.
- Brunquell, J. et al., 2016. The Genome-wide Role of HSF-1 in the Regulation of Gene Expression in *Caenorhabditis elegans*. *BMC Genomics*, 17, article no: 559.
- Bryd, A. K. & Raney, K. D., 2012. Superfamily 2 Helicases. *Frontiers in Bioscience*, 17, pp. 2070-2088.
- Cao, X. & Aballay, A., 2016. Neural Inhibition of dopaminergic Signaling Enhances Immunity in a Cell-non-autonomous Manner. *Current Biology*, 26(17), pp. 2329-2334.
- Carles-Kinch, K., George, J. & Kreuzer, K., 1997. Bacteriophage T4 UvsW Protein is a Helicase involved in Recombination, Repair and the Regulation of DNA Replication Origins. *The EMBO Journal*, 16(13), pp. 4142-4151.
- Caruthers, J. M. & McKay, D. B., 2002. Helicase Structure and Mechanism. *Current Opinion in Structural Biology*, 12(1), pp. 123-133.
- Carvalho, B. & Irizarry, R., 2010. A Framework for Oligonucleotide Microarray Preprocessing. *Bioinformatics*, 26(19), pp. 2363–2367.
- Carvalho, B., 2013. *Bioconductor*. [Online]
Available at: <https://stat.ethz.ch/pipermail/bioconductor/2013-June/053382.html>
[Accessed 14 November 2018].
- Cassidy, L. et al., 2018. The *Caenorhabditis elegans* Proteome Response to naturally associated Microbiome Members of the Genus *Ochrobactrum*. *Proteomics*, 18(8), article no: 1700426.
- Castellano-Pozo, M. et al., 2013. R Loops are linked to Histone H3 S10 Phosphorylation and Chromatin Condensation. *Molecular Cell*, 52(4), pp. 583-590.
- Celniker, S. E. et al., 2009. Unlocking the Secrets of the Genome. *Nature*, 459(7249), pp. 927-930.
- Cerritelli, S. M. & Crouch, R. J., 2009. Ribonuclease H: the Enzymes in Eukaryotes. *FEBS Journal*, 276(6), pp. 1494-1505.
- Chakraborty, P. & Grosse, F., 2011. Human DHX9 Helicase preferentially unwinds RNA-containing Displacement Loops (R-loops) and G-quadruplexes. *DNA repair*, 10(6), pp. 654-665.
- Chakraborty, P., Huang, J. T. & Hiom, K., 2018. DHX9 Helicase promotes R-loop Formation in Cells with impaired RNA splicing. *Nature Communications*, 9, article no: 4346.
- Chang, E. Y.-C. et al., 2017. RECQ-like Helicases Sgs1 and BLM regulate R-loop-associated Genome Instability. *Journal of Cell Biology*, 216(12), pp. 3991-4005.
- Chang, Y.-F., Imam, .. S. & Wilkinson, M. F., 2007. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry*, 76, pp. 51-74.
- Chaplin, D. D., 2010. Overview of the Immune Response. *The Journal of Allergy and Clinical Immunology*, 125(2 Suppl 2), pp. S3-23.
- Chavez, V. et al., 2007. Oxidative Stress Enzymes are required for DAF-16-mediated Immunity due to generation of Reactive Oxygen Species by *Caenorhabditis elegans*. *Genetics*, 176(3), pp. 1567-1577.
- Cheesman, H. K. et al., 2016. Aberrant activation of p38 MAP Kinase-dependent Innate Immune Responses is toxic to *Caenorhabditis elegans*. *G3 Genes | Genomes | Genetics*, 6(3), pp. 541-549.
- Chen, K. et al., 2017. An evolutionarily conserved Transcriptional Response to Viral Infection in *Caenorhabditis* Nematodes. *BMC Genomics*, 18, article no: 303.
- Chen, L. et al., 2017. R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at Gene Promoters. *Molecular Cell*, 68(4), pp. 745-757.

- Chen, R. A.-J. et al., 2014. Extreme HOT Regions are CpG-dense Promoters in *C. elegans* and Humans. *Genome Research*, 24(7), pp. 1138-1146.
- Cinar, H. N. et al., 2010. *Vibrio cholerae* Hemolysin is required for Lethality, Developmental Delay, and Intestinal Vacuolation in *Caenorhabditis elegans*. *PLoS One*, 5(7), article no: e11558.
- Clouaire, T., Webb, S. & Bird, A., 2014. Cfp-1 is required for Gene Expression-dependent H3K4 Trimethylation and H3K9 Acetylation in Embryonic Stem Cells. *Genome Biology*, 15(451), pp. 1-16.
- Clouaire, T. et al., 2012. Cfp1 integrates both CpG content and Gene Activity for accurate H3K4me3 Deposition in Embryonic Stem Cells. *Genes & Development*, 26(15), pp. 1714-1728.
- Coleman, R. et al., 2017. p53 dynamically directs TFIID Assembly on Target Gene Promoters. *Molecular and Cellular Biology*, 37(13), article no: e00085-17.
- Company, M., Arenas, J. & Abelson, J., 1991. Requirement of the RNA helicase-like Protein PRP22 for Release of Messenger RNA from Spliceosomes. *Nature*, 349(6309), pp. 487-493.
- Conte, D. J., MacNeil, L. T., Walhout, A. J. & Mello, C. C., 2015. RNA Interference in *Caenorhabditis elegans*. *Current Protocols in Microbiology*, 109(1), pp. 26.3.1-26.330.
- Corsi, A. K., Wightman, B. & Chalfie, M., 2015. *A Transparent Window into Biology: A Primer on Caenorhabditis elegans*. [Online]
Available at: http://www.wormbook.org/chapters/www_celegansintro/celegansintro.html [Accessed 11 July 2019].
- Couillault, C. et al., 2004. TLR-independent Control of Innate Immunity in *Caenorhabditis elegans* by the TIR Domain Adaptor Protein TIR-1, an Ortholog of Human SARM. *Nature Immunology*, 5(5), pp. 488-494.
- Cross, S. H. & Birds, A. P., 1995. CpG Islands and Genes. *Current Opinion in Genetics & Development*, 5(3), pp. 309-314.
- Curradi, M., Izzo, A., Badaracco, G. & Landsberger, N., 2002. Molecular Mechanisms of Gene Silencing mediated by DNA Methylation. *Molecular and Cellular Biology*, 22(9), pp. 3157-3173.
- Darby, C., 2005. *Wormbook: Interactions with Microbial Pathogens*. [Online]
Available at: http://www.wormbook.org/chapters/www_intermicrobpath/intermicrobpath.html [Accessed 12 12 2019].
- Darby, C., Hsu, J. W., Ghori, N. & Falkow, S., 2002. Plague Bacteria Biofilm blocks Food intake. *Nature*, 417(6886), pp. 243-244.
- Dehe, P.-M. et al., 2006. Protein Interactions with the Set1 Complex and their Roles in the regulation of Histone 3 Lysine 4 Methylation. *Journal of Biological Chemistry*, 281(46), pp. 35404-35412.
- Deng, P. et al., 2019. Mitochondrial UPR repression during *Pseudomonas aeruginosa* infection requires the bZIP protein ZIP-3. *Proceedings of the National Academy of Science*, 116(13), pp. 6146-6151.
- Dhont, L., Mascaux, C. & Belayew, A., 2016. The Helicase-like Transcription Factor (HLTF) in Cancer: Loss of Function or Oncomorphic Conversion of a Tumor Suppressor? *Cellular and Molecular Life Science*, 73(1), pp. 129-145.
- Di Noia, J. & Neuberger, M. S., 2002. Altering the Pathway of Immunoglobulin hypermutation by inhibiting Uracil-DNA Glycosylase. *Nature*, 419(6902), pp. 43-48.
- Dierking, K., Yang, W. & Schulenburg, H., 2016. Antimicrobial Effectors in the Nematode *Caenorhabditis elegans*: an Outgroup to the Arthropoda. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1695), article no: 20150299.
- Dillon, S. C., Zhang, X., Trievel, R. C. & Cheng, X., 2005. The SET-Domain Protein Superfamily: Protein Lysine Methyltransferases. *Genome Biology*, 6(227), pp. 1-10.
- Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq Aligner. *Bioinformatics*, 29(1), pp. 15-21.
- Dominguez-Sanchez, M. S. et al., 2011. Genome Instability and Transcription Elongation Impairment in Human Cells depleted of THO/TREX. *PLoS Genetics*, 7(12), article no: e1002386.

- Donato, V. et al., 2017. *Bacillus subtilis* Biofilm extends *Caenorhabditis elegans* Longevity through downregulation of the Insulin-like Signalling Pathway. *Nature Communications*, 8, article no: 14332.
- Drace, K., McLaughlin, S. & Darby, C., 2009. *Caenorhabditis elegans* BAH-1 is a DUF23 Protein expressed in Seam Cells and required for Microbial Biofilm binding to the Cuticle. *PLoS One*, 4(8), article no: e6741.
- Drickamer, K., 1993. Biology of Animal Lectins. *Annual Review of Cell and Developmental Biology*, 9, pp. 237-264.
- Drolet, M., Bi, X. & Liu, L. F., 1994. Hypernegative Supercoiling of the DNA Template during Transcription Elongation *in vitro*. *The Journal of Biological Chemistry*, 269(3), pp. 2068-2074.
- Drolet, M. et al., 2003. The Problem of hypernegative Supercoiling and R-loop formation in Transcription. *Frontiers in Bioscience*, 8, d210-221.
- Drolet, M. et al., 1995. Overexpression of RNase H partially complements the Growth Defect of an *Escherichia coli* delta *topA* Mutant: R-loop formation is a major Problem in the absence of DNA Topoisomerase I. *Proceedings of the National Academy of Science*, 92(8), pp. 3526-3530.
- Dudas, k. C. & Kreuzer, k. N., 2001. UvsW Protein regulates Bacteriophage T4 Origin-Dependent Replication by unwinding R-loops. *Molecular and Cellular Biology*, 21(8), pp. 2706-2715.
- Dunn, K. & Griffith, J. D., 1980. The Presence of RNA in a Double Helix inhibits its Interaction with Histone Protein. *Nucleic Acids Research*, 8(3), pp. 555-566.
- Dupont, C., Armant, D. R. & Brenner, C. A., 2009. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Seminars in Reproductive Medicine*, 27(5), pp. 351-357.
- Duquette, M. L. et al., 2004. Intracellular Transcription of G-rich DNAs induces formation of G-loops, novel Structures containing G4 DNA. *Genes & Development*, 18(13), pp. 1618-1629.
- Eckl, J. et al., 2017. Hsp90-downregulation influences the Heat-Shock Response, Innate Immune Response and Onset of Oocyte Development in Nematodes. *PLoS one*, 12(10), article no: e0186386.
- Edgar, R., Domrachev, M. & Lash, A. E., 2002. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research*, 30(1), pp. 207-210.
- Engelmann, I. et al., 2011. A Comprehensive Analysis of Gene Expression Changes provoked by Bacterial and Fungal Infection in *C. elegans*. *PLoS One*, 6(5), article no: e19055.
- Ensembl, 2019. *BioMart*. [Online]
Available at: <https://www.ensembl.org/biomart/martview/>
[Accessed 15 October 2018].
- Ensembl, 2019. *Caenorhabditis elegans (WBcel235)*. [Online]
Available at: https://www.ensembl.org/Caenorhabditis_elegans/Info/Index
[Accessed 21 January 2019].
- Epicentre, n.d. *DNA-Specific Nucleases*. [Online]
Available at: <http://www.epibio.com/docs/default-source/forum-archive/forum-02-1---dna-specific-nucleases.pdf>
[Accessed 1 march 2018].
- EpiGenie, n.d. *epigenie Informally Informative*. [Online]
Available at: <https://epigenie.com/key-epigenetic-players/histone-proteins-and-modifications/histone-h3k4/>
[Accessed 26 October 2018].
- Ermolaeva, M. A. & Schumacher, B., 2014. Insights from the Worm: The *C. elegans* Model for Innate Immunity. *Seminars in Immunology*, 26(4), pp. 303-309.
- Erzberger, J. P. & Berger, J. M., 2006. Evolutionary Relationship and Structural Mechanisms of AAA+ Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 35, pp. 93-114.
- Estes, K. A. et al., 2010. bZIP Transcription Factor *zip-2* mediates an early Response to *Pseudomonas aeruginosa* Infection in *Caenorhabditis elegans*. *Proceedings of the National Academy of Science*, 107(5), pp. 2153-2158.

- Ettwiller, L. et al., 2007. Trawler: *de novo* regulatory Motif Discovery Pipeline for Chromatin Immunoprecipitation. *Nature Methods*, 4(7), pp. 563-565.
- Ewels, P., Magnusson, M., Lundin, S. & Källér, M., 2016. MultiQC: Summarize Analysis Results for multiple Tools and Samples in a Single Report. *Bioinformatics*, 32(19), pp. 3047-3048.
- Fairman-Williams, M. E., Guenther, U.-P. & Jankowsky, E., 2010. SF1 and SF2 Helicases: Family Matters. *Current Opinion in Structural Biology*, 20(3), pp. 313-324.
- Falkenberg, K. J. & Johnstone, R. W., 2014. Histone Deacetylases and their Inhibitors in Cancer, Neurological Diseases and Immune Disorders. *Nature Review Drug Discovery*, 13(9), pp. 673-691.
- Falon, S. et al., 2018. *pdInfoBuilder: Platform Design Information Package Builder*. [Online] Available at: <https://www.bioconductor.org/packages/release/bioc/html/pdInfoBuilder.html> [Accessed 14 November 2018].
- Felix, M.-A. et al., 2011. Natural and experimental Infection of *Caenorhabditis* Nematodes by novel Viruses related to Nodaviruses. *PLoS Biology*, 9(1), article no: e1000586.
- Felix, M.-A. & Duveau, F., 2012. Population Dynamics and Habitat Sharing of Natural Populations of *Caenorhabditis elegans* and *C. briggsae*. *BMC Biology*, 10, article no: 59.
- Ferguson-Smith, A. C. & Bourc'his, D., 2018. 2018 Gairdner Awards: The Discovery and Importance of Genomic Imprinting. *eLife*, 7, article no: e42368.
- Fischer, S. E. et al., 2013. Multiple small RNA Pathways regulate the Silencing of repeated and foreign Genes in *C. elegans*. *Genes & Development*, 27(24), pp. 2678-2695.
- Fletcher, M. et al., 2019. Global Transcriptional Regulation of Innate Immunity by ATF-7 in *C. elegans*. *PLoS Genetics*, 15(2), article no: e1007830.
- Franz, C. J. et al., 2014. Orsay, Santeuil and Le Blanc Viruses primarily infect Intestinal Cells in *Caenorhabditis* Nematodes. *Virology*, 448, pp. 255-264.
- Frederic, C., 2016. Nascent Connections: R-loops and Chromatin Patterning. *Trends in Genetics*, 32(12), pp. 828-838.
- Frick, D. N. & Richardson, C. C., 2001. DNA Primases. *Annual Review of Biochemistry*, 70, pp. 39-80.
- Fujiki, K., Mizuno, T., Hisamoto, N. & Matsumoto, K., 2010. The *Caenorhabditis elegans* Ste20-related Kinase and Rac-Type Small GTPase regulate the c-Jun N-Terminal Kinase Signaling Pathway mediating the Stress Response. *Molecular and Cellular Biology*, 30(4), pp. 995-1003.
- Fulda, S., Gorman, A. M., Hori, O. & Samali, A., 2010. Cellular Stress Responses: Cell Survival and Cell Death. *International Journal of Cell Biology*, 2010, article no: 214074.
- Galbadage, T. et al., 2016. The *Caenorhabditis elegans* p38 MAPK Gene plays a Key Role in Protection from Mycobacteria. *Microbiology Open*, 5(3), pp. 436-452.
- Gan, W. et al., 2011. R-loop-mediated Genomic Instability is caused by Impairment of Replication Fork Progression. *Genes & Development*, 25(19), pp. 2041-2056.
- Garcia-Benitez, F., Gaillard, H. & Aguilera, A., 2017. Physical proximity of Chromatin to Nuclear Pores prevents harmful R loop Accumulation contributing to maintain Genome Stability. *Proceedings of the National Academy of Science*, 114(41), pp. 10942-10947.
- Garcia-Picardo, D. et al., 2017. Histone Mutants separate R loop Formation from Genome Instability Induction. *Molecular Cell*, 66(51), pp. 597-609.
- Garcia-Rubio, M. L. et al., 2015. The Fanconi Anemia Pathway protects Genome Integrity from R-loops. *PLoS Genetics*, 11(11), article no: e1005674.
- Garigan, D. et al., 2002. Genetic analysis of Tissue Aging in *Caenorhabditis elegans*: a Role for Heat-Shock Factor and Bacterial Proliferation. *Genetics*, 161(3), pp. 1101-1112.
- Garren, S. T., 2019. *jmuOutlier: Permutation Tests for Nonparametric Statistics*. [Online] Available at: <https://cran.r-project.org/web/packages/jmuOutlier/index.html> [Accessed 5 August 2019].

- Garsin, D. A. et al., 2001. A simple Model Host for identifying Gram-positive Virulence Factors. *Proceedings of the National Academy of Science*, 98(19), pp. 10892-10897.
- Garsin, D. A. et al., 2003. Long-Lived *C. elegans* *daf-2* Mutants are resistant to Bacterial Pathogens. *Science*, 300(5627), p. 1921.
- Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A., 2004. affy—Analysis of Affymetrix GeneChip Data at the Probe Level. *Bioinformatics*, 20(3), pp. 307-315.
- GE Healthcare Life Sciences, 2019. *Amersham Hybond-N+*. [Online] Available at: <https://www.gelifesciences.com/en/us/shop/protein-analysis/blotting-and-detection/nylon-membranes/amersham-hybond-n-p-05398> [Accessed 17 July 2019].
- Gentleman, R., 2019. *annotate: Annotation for Microarrays.* [Online] Available at: <https://www.bioconductor.org/packages/release/bioc/html/annotate.html> [Accessed 28 November 2018].
- George, T. et al., 2009. Human Pif1 Helicase unwinds Synthetic DNA Structures resembling Stalled DNA Replication Forks. *Nucleic Acids Research*, 37(19), pp. 6491-6502.
- Gerke, P., Keshet, A., Mertenskötter, A. & Paul, R., 2014. The JNK-Like MAPK KGB-1 of *Caenorhabditis elegans* promotes Reproduction, Lifespan, and Gene Expressions for Protein Biosynthesis and Germline Homeostasis but interferes with Hyperosmotic Stress Tolerance. *Cellular Physiology and Biochemistry*, 34(6), pp. 1951-1973.
- Ge, S. X. & Jung, D., 2018. *ShinyGO: A Graphical Enrichment Tool for Animals and Plants.* [Online] Available at: <https://www.biorxiv.org/content/10.1101/315150v1> [Accessed 19 July 2019].
- Gilhooly, N. S., Gwynn, E. J. & Dillingham, M. S., 2013. Superfamily 1 Helicases. *Frontiers in Bioscience*, 5(1), pp. 206-216.
- Ginno, P. A. et al., 2013. GC Skew at the 5' and 3' Ends of the Human Genes links R-loop Formation to Epigenetic Regulation and Transcription Termination. *Genome Research*, 23(10), pp. 1590-1600.
- Ginno, P. A. et al., 2012. R-loop Formation is a distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Molecular Cell*, 45(6), pp. 814-825.
- Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G., 2008. RNA-binding Proteins and Post-transcriptional Gene Regulation. *FEBS Letters*, 582(14), pp. 1977-1986.
- Gogol, E. P., Seifried, S. E. & von Hippel, P. H., 1991. Structure and Assembly of the *Escherichia coli* Transcription Termination Factor rho and its Interactions with RNA I. Cryoelectron Microscopic Studies. *Journal of Molecular Biology*, 221(4), pp. 1127-1138.
- Goh, G. Y. et al., 2014. The conserved Mediator subunit MDT-15 is required for Oxidative Stress Responses in *Caenorhabditis elegans*. *Aging Cell*, 13(1), pp. 70-79.
- Goh, H. M. S., Yong, M. H. A., Chong, K. K. L. & Kline, K. A., 2017. Model Systems for the Study of Enterococcal Colonization and Infection. *Virulence*, 8(8), pp. 1525-1562.
- Goh, W. W. B., Wang, W. & Wong, L., 2017. Why Batch Effects matter in Omics Data, and how to avoid them. *Trends in Biotechnology*, 35(6), pp. 498-507.
- Gorbalenya, A. E. & Koonin, E. V., 1993. Helicases: Amino Acid Sequence Comparisons and Structure-Function Relationships. *Current Opinion in Structural Biology*, 3(3), pp. 419-429.
- Gorman, A. M. et al., 1999. Antioxidant-mediated Inhibition of the Heat Shock Response leads to Apoptosis. *FEBS Letters*, 445(1), pp. 98-102.
- Gracida, X. & Eckmann, C. R., 2013. Fertility and Germline Stem Cell Maintenance under Different Diets requires *nhr-114/HNF4* in *C. elegans*. *Current Biology*, 23(7), pp. 607-613.
- Grant, C. E., Bailey, T. L. & Stafford Noble, W., 2011. FIMO: scanning for occurrences of a given Motif. *Bioinformatics*, 27(7), pp. 1017-1018.

- Grants, J. M., Goh, G. Y. & Taubert, S., 2015. The Mediator Complex of *Caenorhabditis elegans*: Insights into the Developmental and Physiological Roles of a conserved Transcriptional Coregulator. *Nucleic Acids Research*, 43(4), pp. 2442-2453.
- Gravato-Nobre, M. J. & Hodgkin, J., 2005. *Caenorhabditis elegans* as a Model for Innate Immunity to Pathogens. *Cellular Microbiology*, 7(6), pp. 741-751.
- Gravato-Nobre, M. J. et al., 2011. Glycosylation Genes expressed in Seam Cells determine Complex Surface Properties and Bacterial Adhesion to the Cuticle of *Caenorhabditis elegans*. *Genetics*, 187(1), pp. 141-155.
- Gravato-Nobre, M. J. et al., 2016. The Invertebrate Lysozyme Effector ILYS-3 is systemically activated in Response to Danger Signals and confers Antimicrobial Protection in *C. elegans*. *PLoS Pathogens*, 12(8), article no: e1005826.
- Greer, E. L. et al., 2015. DNA Methylation on N6-Adenine in *C. elegans*. *Cell*, 161(4), pp. 868-878.
- Grierson, P. M., Acharya, S. & Groden, J., 2013. Collaborating Functions of BLM and DNA Topoisomerase I in regulating Human rDNA Transcription. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 743-744(Mar-Apr), pp. 89-96.
- Grishkevich, V., Hashimshony, T. & Yanai, I., 2011. Core Promoter T-blocks correlate with Gene Expression Levels in *C. elegans*. *Genome Research*, 21(5), pp. 707-717.
- Groh, M., Lufino, M. M., Wade-Martins, R. & Gromak, N., 2014. R-loops associated with Triplet Repeat Expansions promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLoS Genetics*, 10(5), article no: e1004318.
- Guillemette, B. et al., 2011. H3 Lysine 4 is acetylated at Active Gene Promoters and is regulated by H3 Lysine 4 Methylation. *PLoS Genetics*, 7(3), article no: e1001354.
- Gupta, S. et al., 2010. HSP72 protects Cells from ER Stress-induced Apoptosis via Enhancement of IRE1 α -XBP1 Signaling through a Physical Interaction. *PLOS Biology*, 8(7), article no: e1000410.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Stafford Noble, W., 2007. Quantifying similarity between Motifs. *Genome Biology*, 8(2), article no: R24
- Gwack, Y., Kim, D. W., Han, j. H. & Choe, J., 1997. DNA Helicase activity of the Hepatitis C Virus Nonstructural Protein 3. *European Journal of Biochemistry*, 250(1), pp. 47-54.
- Haas, R. et al., 2018. A-to-I RNA Editing affects lncRNAs Expression after Heat Shock. *Genes*, 9(12), article no: 0627.
- Hamperl, S. & Cimprich, K. A., 2014. The Contribution of co-transcriptional RNA:DNA Hybrid Structures to DNA Damage and Genome Instability. *DNA repair*, 19(Jul), pp. 84-94.
- Handy, D. E., Castro, R. & Loscalzo, J., 2011. Epigenetic Modifications: Basic Mechanisms and Role in Cardiovascular Disease. *Circulation*, 123(19), pp. 2145-2156.
- Han, K. et al., 2019. Genome-Wide Identification of Histone Modifications involved in Placental Development in Pigs. *Frontiers in Genetics*, 10, article no: 277.
- Han, S. et al., 2017. Mono-unsaturated Fatty Acids link H3K4me3 Modifiers to *C. elegans* Lifespan. *Nature*, 544(7649), pp. 185-190.
- Hartono, S. R. et al., 2018. The Affinity of the S9.6 Antibody for double-Stranded RNAs impacts the accurate mapping of R-Loops in Fission Yeast. *Journal of Molecular Biology*, 430(3), pp. 272-284.
- Haslbeck, M., Franzmann, T., Weinfurter, D. & Buchner, J., 2005. Some like it hot: the Structure and Function of Small Heat-Shock Proteins. *Nature Structural & Molecular Biology*, 12(10), pp. 842-846.
- Head, B. & Aballay, A., 2014. Recovery from an Acute Infection in *C. elegans* requires the GATA Transcription Factor ELT-2. *PLoS Genetics*, 10(10), article no: e1004609.
- Head, B. P., Olaitan, A. O. & Aballay, A., 2017. Role of GATA Transcription Factor ELT-2 and p38 MAPK PMK-1 in Recovery from acute *P. aeruginosa* Infection in *C. elegans*. *Virulence*, 8(3), pp. 261-274.

- Heintzman, N. D. et al., 2007. Distinct and predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome. *Nature Genetics*, 39(3), pp. 311-318.
- Heinz, S. et al., 2010. Simple Combinations of Lineage-determining Transcription Factors prime cis-regulatory Elements required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4), pp. 576-589.
- Hergeth, S. P. & Schneider, R., 2015. The H1 Linker Histones: Multifunctional Proteins beyond the Nucleosomal Core Particle. *EMBO Reports*, 16(11), pp. 1439-1453.
- He, Y. et al., 2018. A double-edged Function of DDX3, as an Oncogene or Tumor Suppressor, in Cancer Progression. *Oncology Reports*, 39(3), pp. 883-892.
- Hickman, A. B. & Dyad, F., 2005. Binding and unwinding: SF3 Viral Helicases. *Current Opinion in Structural Biology*, 15(1), pp. 77-85.
- Hill, A. J. et al., 2014. Cellular Stress induces a protective sleep-like State in *C. elegans*. *Current Biology*, 24(20), pp. 2399-2405.
- Hoang, K. L., Gerardo, N. M. & Morran, L. T., 2019. The effects of *Bacillus subtilis* on *Caenorhabditis elegans* Fitness after Heat Stress. *Ecology and Evolution*, 9(6), pp. 3491-3499.
- Hodgkin, J., Kuwabara, P. E. & Corneliusen, B., 2000. A novel Bacterial Pathogen, *Microbacterium nematophilum*, induces morphological change in the Nematode *C. elegans*. *Current Biology*, 10(24), pp. 1615-1618.
- Hodroj, D. et al., 2017a. An ATR-dependent Function for the Ddx19 RNA Helicase in Nuclear R-loop Metabolism. *The EMBO Journal*, 36(9), pp. 1182-1198.
- Hodroj, D., Serhal, K. & Maiorano, D., 2017b. Ddx19 links mRNA Nuclear Export with Progression of Transcription and Replication and suppresses Genomic Instability upon DNA Damage in proliferating Cells. *Nucleus*, 8(5), pp. 489-495.
- Höflich, J. et al., 2004. Loss of srf-3-encoded Nucleotide Sugar Transporter Activity in *Caenorhabditis elegans* alters Surface Antigenicity and prevents Bacterial Adherence. *Journal of Biological Chemistry*, 279(29), pp. 30440-30448.
- Hong, X., Cadwell, G. W. & Kogoma, T., 1995. *Escherichia coli* RecG and RecA Proteins in R-loop Formation. *The EMBO Journal*, 14(19), pp. 2385-2392.
- Howard, J. A., Delmas, S., Ivancic-Bace, I. & Bolt, E. L., 2011. Helicase Dissociation and Annealing of RNA-DNA Hybrids by *Escherichia coli* Cas3 Protein. *Biochemical Journal*, 439(1), pp. 85-95.
- Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J., 2017. Is H3K4me3 instructive for Transcription Activation?. *BioEssays*, 39(1), pp. 1-12.
- Hsu, A.-L., Murphy, C. T. & Kenyon, C., 2003. Regulation of Aging and Age-Related Disease by DAF-16 and Heat-Shock Factor. *Science*, 300(5622), pp. 1142-1145.
- Huang, W. et al., 2011. The Influence of Stress Factors on the Reactivation of Latent Herpes Simplex Virus Type 1 in infected Mice. *Cell Biochemistry and Biophysics*, 61(1), pp. 115-122.
- Huang, Y. et al., 2006. Recognition of Histone H3 Lysine-4 Methylation by the Double Tudor Domain of JMJD2A. *Science*, 312(5774), pp. 748-751.
- Huisinga, K. L., Brower-Toland, B. & Elgin, S. C., 2006. The contradictory Definitions of Heterochromatin: Transcription and Silencing. *Chromosoma*, 115(2), pp. 110-122.
- Hung, S.-H., Yu, Q., Gray, D. M. & Ratliff, R. L., 1994. Evidence from CD Spectra that d(purine).r(pyrimidine) and r(purine).d(pyrimidine) Hybrids are in different Structural Classes. *Nucleic Acids Research*, 22(20), pp. 4326-4334.
- Hung, T. et al., 2010. ING4 mediates Crosstalk between Histone H3 K4 Trimethylation and H3 Acetylation to attenuate Cellular Transformation. *Molecular Cell*, 33(2), pp. 248-256.

- Hu, P. J., 2007. *Dauer*. [Online]
Available at: http://www.wormbook.org/chapters/www_dauer/dauer.html
[Accessed 4 July 2019].
- Huppert, J. L., Bugaut, A., Kumari, S. & Balasubramanian, S., 2008. G-quadruplexes: the Beginning and End of UTRs. *Nucleic Acids Research*, 36(19), pp. 6260-6268.
- Hu, Y., Platzer, E. G., Bellier, A. & Aroian, R. V., 2010. Discovery of a highly synergistic Anthelmintic Combination that shows mutual Hypersusceptibility. *Proceedings of the National Academy of Science*, 107(13), pp. 5955-5960.
- Hyun, M., Bohr, V. A. & Ahn, B., 2008. Biochemical Characterization of the WRN-1 RecQ Helicase of *Caenorhabditis elegans*. *Biochemistry*, 47(28), pp. 7583-7593.
- Illingworth, R. S. & Bird, A. P., 2009. CpG Islands - 'A rough Guide'. *FEBS Letters*, 583(11), pp. 1713-1720.
- Inoue, H. et al., 2005. The *C. elegans* p38 MAPK Pathway regulates Nuclear Localization of the Transcription Factor SKN-1 in Oxidative Stress Response. *Genes & Development*, 19(19), pp. 2278-2283.
- InterPro, n.d. *von Willebrand Factor, Type A*. [Online]
Available at: <https://www.ebi.ac.uk/interpro/beta/entry/InterPro/IPR002035/>
[Accessed 15 August 2019].
- Iraozqui, J. E. et al., 2010. Distinct Pathogenesis and Host Responses during Infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. *PLoS Pathogens*, 6(7), article no: e1000982.
- Jankowsky, A., Guenther, U.-P. & Jankowsky, E., 2011. The RNA Helicase Database. *Nucleic Acids Research*, 39, pp. D338-D341.
- Jankowsky, E. & Fairman-Williams, M. E., 2010. An Introduction to RNA Helicases: Superfamilies, Families, and major Themes. In: E. Jankowsky, ed. *RNA Helicases*. Cambridge: RSC Publishing, pp. 1-31.
- Jansson, H.-B., 1994. Adhesion of Conidia of *Drechmeria coniospora* to *Caenorhabditis elegans* Wild Type and Mutants. *Journal of Nematology*, 26(4), pp. 430-435.
- Jantsch, J., Chikkaballi, D. & Hensel, M., 2011. Cellular Aspects of Immunity to Intracellular *Salmonella enterica*. *Immunological Reviews*, 240(1), pp. 185-195.
- JebaMercy, G., Vigneshwari, L. & Balamurugan, K., 2013. A MAP Kinase Pathway in *Caenorhabditis elegans* is required for Defense against Infection by opportunistic Proteus Species. *Microbes and Infection*, 15(8-9), pp. 550-568.
- Jenuwein, T. & Allis, D. C., 2001. Translating the Histone Code. *Science*, 293(5532), pp. 1074-1080.
- Jeong, D.-E. et al., 2017. Mitochondrial Chaperone HSP-60 regulates Anti-Bacterial Immunity via p38 MAP Kinase Signaling. *EMB Journal*, 36(8), pp. 1046-1065.
- Jiang, W. et al., 2018. A Genetic Program mediates Cold-Warming Response and promotes Stress-induced Phenoptosis in *C. elegans*. *eLIFE*, 7, article no: e35037.
- Jiang, Y., Liu, M., Spencer, C. A. & Price, D. H., 2004. Involvement of Transcription Termination Factor 2 in Mitotic Repression of Transcription Elongation. *Molecular Cell*, 14(3), pp. 375-386.
- Jin, B., Li, Y. & Robertson, K. D., 2011. DNA Methylation: superior or subordinate in the Epigenetic Hierarchy?. *Genes & Cancer*, 2(6), pp. 607-617.
- Joseph, J. W. & Kolodner, R., 1983. Exonuclease VIII of *Escherichia coli*. *The Journal of Biological Chemistry*, 258(17), pp. 10418-10424.
- Kamath, R. S. & Ahringer, J., 2003. Genome-wide RNAi Screening in *Caenorhabditis elegans*. *Methods*, 30(4), pp. 313-321.
- Kamath, R. S. et al., 2000. Effectiveness of specific RNA-mediated Interference through ingested double-stranded RNA in *Caenorhabditis elegans*. *Genome Biology*, 2(1), article no: research0002

- Kanagaraj, R. et al., 2010. RECQ5 Helicase associates with the C-terminal Repeat Domain of RNA Polymerase II during productive Elongation Phase of Transcription. *Nucleic Acids Research*, 38(22), pp. 8131-8140.
- Kang, D. et al., 2018. Pyoverdine, a Siderophore from *Pseudomonas aeruginosa*, translocates into *C. elegans*, removes Iron, and activates a distinct Host Response. *Virulence*, 9(1), pp. 804-817.
- Kang, D. & Kirienko, N. V., 2017. High-Throughput Genetic Screen reveals that Early Attachment and Biofilm Formation are necessary for full Pyoverdine Production by *Pseudomonas aeruginosa*. *Frontiers in Microbiology*, 8, article no: 1707.
- Kao, C.-Y. et al., 2011. Global Functional Analyses of Cellular Responses to Pore-Forming Toxins. *PLoS Pathogens*, 7(3), article no: e1001314.
- Kasahara, M., Clikeman, J. A., Bates, D. B. & Kogoma, T., 2000. RecA Protein-dependent R-loop Formation *in vitro*. *Genes & Development*, 14(3), pp. 360-365.
- Keller, W. & Crouch, R., 1972. Degradation of DNA RNA Hybrids by Ribonuclease H and DNA Polymerases of Cellular and Viral Origin. *Proceedings of the National Academy of Sciences of the United States of America*, 69(11), pp. 3360-3364.
- Ketchen, D. J. & Shook, C. L., 1996. The Application of Cluster Analysis in Strategic Management Research: an Analysis and Critique. *Strategic Management Journal*, 17(6), pp. 441-458.
- Kho, M. F. et al., 2011. The Pore-forming Protein Cry5B elicits the Pathogenicity of *Bacillus sp.* against *Caenorhabditis elegans*. *PLoS One*, 6(12), article no: e29122.
- Kim, D. H. & Ewbank, J. J., 2018. *Signaling in the Innate Immune Response*. [Online] Available at: http://www.wormbook.org/chapters/www_signalingimmuneresponse.2/signalingimmuneresponse.2.html [Accessed 6 May 2019].
- Kim, D. H. et al., 2002. A Conserved p38 MAP Kinase Pathway in *Caenorhabditis elegans* Innate Immunity. *Science*, 297(5581), pp. 623-626.
- Kim, D. H. et al., 2004. Integration of *Caenorhabditis elegans* MAPK Pathways mediating Immunity and Stress Resistance by MEK-1 MAPK Kinase and VHP-1 MAPK Phosphatase. *Proceedings of the National Academy of Science*, 101(30), pp. 10990-10994.
- Kim, H.-D., Choe, J. & Seo, Y.-S., 1999. The sen1+ Gene of *Schizosaccharomyces pombe*, a Homologue of Budding Yeast SEN1, encodes an RNA and DNA Helicase. *Biochemistry*, 38(44), pp. 14697-14710.
- Kim, J. et al., 2010. RAD6-Mediated Transcription-coupled H2B Ubiquitylation directly stimulates H3K4 Methylation in Human Cells. *Cell*, 137(3), pp. 459-471.
- Kim, W., Underwood, R. S., Greenwald, I. & Shaye, D. D., 2018. OrthoList 2: A New Comparative Genomic Analysis of Human and *Caenorhabditis elegans* Genes. *Genetics*, 210(2), pp. 445-461.
- King, C. D. et al., 2018. Proteomic Identification of virulence-related Factors in young and aging *C. elegans* infected with *Pseudomonas aeruginosa*. *Journal of Proteomics*, 181, pp. 92-103.
- Kogoma, T., 1997. Stable DNA Replication: Interplay between DNA Replication, Homologous Recombination, and Transcription. *Microbiology and Molecular Biology Reviews*, 61(2), pp. 212-238.
- Koh, J. H., Wang, L., Beaudoin-Chabot, C. & Thibault, G., 2018. Lipid bilayer stress-activated IRE-1 modulates Autophagy during Endoplasmic Reticulum Stress. *Journal of Cell Science*, 131(22), article no: jcs217992.
- Kolasinska-Zwierz, P. et al., 2009. Differential Chromatin Marking of Introns and expressed Exons by H3K36me3. *Nature Genetics*, 41(3), pp. 376-381.
- König, F., Schubert, T. & Längst, G., 2017. The monoclonal S9.6 Antibody exhibits highly variable Binding Affinities towards different R-loop Sequences. *PLoS One*, 12(6), article no: e0178875.
- Kornberg, R. D., 1974. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science*, 184(4139), pp. 868-871.

- Kornberg, R. D. & Lorch, Y., 1999. Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell*, 98(3), pp. 285-294.
- Kourtis, N. & Tavernarakis, N., 2011. Cellular Stress Response Pathways and Ageing: intricate Molecular Relationships. *EMBO Journal*, 30(13), pp. 2520-2531.
- Krajewski, W. A., Nakamura, T., Mazo, A. & Canaani, E., 2005. A Motif within SET-Domain Proteins binds single-stranded Nucleic Acids and transcribed and supercoiled DNAs and can interfere with Assembly of Nucleosomes. *Molecular and Cellular Biology*, 25(5), pp. 1891-1899.
- Krueger, F., 2012. *Trim Galore*. [Online]
Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[Accessed 20 November 2018].
- Kültz, D., 2005. Molecular and Evolutionary Basis of the Cellular Stress Response. *Annual Review of Physiology*, 67, pp. 225-257.
- Kumsta, C., Chang, J. T., Schmalz, J. & Hansen, M., 2017. Hormetic Heat Stress and HSF-1 induce Autophagy to improve Survival and Proteostasis in *C. elegans*. *Nature Communications*, 8, article no: 14337.
- Kurz, C. L. et al., 2003. Virulence Factors of the Human opportunistic Pathogen *Serratia marcescens* identified by *in vivo* Screening. *The EMBO Journal*, 22(7), pp. 1451-1460.
- Kurz, C. L. & Ewbank, J. J., 2000. *Caenorhabditis elegans* for the Study of Host-Pathogen Interactions. *Trends in Microbiology*, 8(3), pp. 142-144.
- Kusch, T., 2012. Histone H3 Lysine 4 Methylation revisited. *Transcription*, 3(6), pp. 310-314.
- Kutscher, L. M. & Shaham, S., 2014. Forward and Reverse Mutagenesis in *C. elegans*. *Wormbook*, pp. 1-26.
- Kuznetsov, V. A. et al., 2018. Toward predictive R-loop Computational Biology: Genome-scale Prediction of R-loops reveals their Association with complex Promoter Structures, G-quadruplexes and transcriptionally active Enhancers. *Nucleic Acids Research*, 46(15), pp. 7566-7585.
- Kwon, G., Lee, J., Koh, J.-H. & Lim, Y.-H., 2018. Lifespan Extension of *Caenorhabditis elegans* by *Butyrivibrio pullicaecorum* and *Megasphaera elsdenii* with probiotic Potential. *Current Microbiology*, 75(5), pp. 557-564.
- Labrousse, A. et al., 2000. *Caenorhabditis elegans* is a Model Host for *Salmonella typhimurium*. *Current Biology*, 10(23), pp. 1543-1545.
- Ladage, M. L. et al., 2016. Glucose or altered Ceramide Biosynthesis mediate Oxygen Deprivation Sensitivity through novel Pathways revealed by Transcriptome Analysis in *Caenorhabditis elegans*. *G3 (Bethesda)*, 6(10), pp. 3149-3160.
- Lane, A. N., Chaires, J. B., Gray, R. D. & Trent, J. O., 2008. Stability and Kinetics of G-quadruplex Structures. *Nucleic Acids Research*, 36(17), pp. 5482-5515.
- Lang, K. S. et al., 2017. Replication-Transcription Conflicts generate R-loops that orchestrate Bacterial Stress Survival and Pathogenesis. *Cell*, 170(4), pp. 787-799.
- Längst, G. & Manlyte, L., 2015. Chromatin Remodelers: From Function to Dysfunction. *Genes*, 6(2), pp. 299-324.
- Lauberth, S. M. et al., 2014. H3K4me3 Interactions with TAF3 regulate Preinitiation Complex Assembly and Selective Gene Activation. *Cell*, 152(5), pp. 1021-1036.
- Lee, B. B. et al., 2018. Rpd3L HDAC links H3K4me3 to transcriptional repression Memory. *Nucleic Acids Research*, 46(16), pp. 8261-8274.
- Lee, J.-H. & Skalnik, D. G., 2005. CpG-binding Protein (CXXC Finger Protein 1) is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex. *Journal of Biological Chemistry*, 280(50), pp. 41725-41731.

- Leela, J. K., Syeda, A. H., Anupama, K. & Gowrishankar, J., 2013. Rho-dependent Transcription Termination is essential to prevent excessive Genome-wide R-loops in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 110(1), pp. 258-263.
- Lemmens, B. B. & Tijsterman, M., 2011. DNA Double-Strand Break Repair in *Caenorhabditis elegans*. *Chromosoma*, 120(1), pp. 1-21.
- Leroy, M. et al., 2012. Pathogen-induced *Caenorhabditis elegans* Developmental Plasticity has a hormetic Effect on the Resistance to Biotic and Abiotic Stresses. *BMC Evolutionary Biology*, 12, article no: 187.
- Liang, Z. et al., 2019. Binding of FANCI-FANCD2 Complex to RNA and R-loops stimulates robust FANCD2 Monoubiquitination. *Cell Report*, 26(3), pp. 564-572.
- Liao, Y., Smyth, G. k. & Shi, W., 2014. featureCounts: an efficient general purpose Program for assigning Sequence Reads to Genomic Features. *Bioinformatics*, 30(7), pp. 923-930.
- Li, B., Carey, M. & Workman, J. L., 2007. The Role of Chromatin during Transcription. *Cell*, 128(4), pp. 707-719.
- Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp. 2078-2079.
- Li, J. et al., 2016. E2F coregulates an essential HSF Developmental Program that is distinct from the Heat-shock Response. *Genes & Development*, 30(18), pp. 2062-2075.
- Lim, Y. W. et al., 2015. Genome-wide DNA Hypomethylation and RNA:DNA Hybrid Accumulation in Aicardi-Goutières Syndrome. *eLife*, 4, article no: e08007.
- Lindahl, T., 1993. Instability and Decay of Primary Structure of DNA. *Nature*, 362(6422), pp. 709-715.
- List, C. et al., 2018. Genes activated by *Vibrio cholerae* upon exposure to *Caenorhabditis elegans* reveal the Mannose-Sensitive Hemagglutinin to be Essential for Colonization. *mSphere*, 3(3), article no: e00238-18.
- Liu, N., Balliano, A. & Hayes, J. J., 2011. Mechanism(s) of SWI/SNF-induced Nucleosome Mobilization. *ChemBioChem*, 12(2), pp. 196-204.
- Liu, Y. & Chang, A., 2008. Heat shock Response relieves ER Stress. *EMBO Journal*, 27(7), pp. 1049-1059.
- Liu, Y., Kaval, K. G., van Hoof, A. & Garsin, D. A., 2019. Heme Peroxidase HPX-2 protects *Caenorhabditis elegans* from Pathogens. *PLoS genetics*, 15(1), article no: e1007944.
- Long, H. K., Blackledge, N. P. & Klose, R. J., 2013. ZF-CxxC Domain-containing Proteins, CpG Islands and the Chromatin Connection. *Biochemical Society Transactions*, 41(3), pp. 727-740.
- Loomis, E. W., Sanz, L. A., Chedin, F. & Hagerman, P. J., 2014. Transcription-associated R-loop Formation across the Human FMR1 CGG-Repeat Region. *PLoS Genetics*, 10(4), article no: e1004294.
- Lorch, Y., Maier-Davis, B. & Kornberg, R. D., 2010. Mechanism of Chromatin Remodeling. *Proceedings of the National Academy of Sciences*, 107(8), pp. 3458-3462.
- Love, M. I., Huber, W. & Anders, S., 2014. Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. *Genome Biology*, 15(12), article no: 550.
- Lovett, S. T., 2011. The DNA Exonucleases of *Escherichia coli*. *EcoSal Plus*, 4(2), pp. 1-45.
- Lucas, A., 2018. *amap: Another Multidimensional Analysis Package*. [Online] Available at: <https://cran.r-project.org/web/packages/amap/index.html> [Accessed 11 March 2019].
- Luo, H. et al., 2013. The Effects of *Bacillus thuringiensis* Cry6A on the Survival, Growth, Reproduction, Locomotion, and Behavioral Response of *Caenorhabditis elegans*. *Applied Microbiology and Biotechnology*, 97(23), pp. 10135-10142.

- MacNeil, L. T. et al., 2015. Transcription Factor Activity Mapping of a Tissue-specific *in vivo* Gene Regulatory Network. *Cell Systems*, 1(2), pp. 152-162.
- Maere, S., Heymans, K. & Kuiper, M., 2005. BiNGO: a Cytoscape Plugin to assess Overrepresentation of Gene Ontology Categories in Biological Networks. *Bioinformatics*, 21(16), pp. 3448-3449.
- Mallo, G. V. et al., 2002. Inducible Antibacterial Defense System in *C. elegans*. *Current Biology*, 12(14), pp. 1209-1214.
- Manosas, M. et al., 2013. RecG and UvsW catalyse robust DNA Rewinding critical for stalled DNA Replication Fork Rescue. *Nature Communications*, 4, article no: 2368.
- Manosas, M., Xi, X. G., Bensimon, D. & Croquette, V., 2010. Active and passive Mechanisms of Helicases. *Nucleic Acids Research*, 38(16), pp. 5518-5526.
- Ma, Q., 2013. Role of Nrf2 in Oxidative Stress and Toxicity. *Annual Review of Pharmacology and Toxicology*, 53, pp. 401-426.
- Margaritis, T. et al., 2012. Two distinct Repressive Mechanisms for Histone 3 Lysine 4 Methylation through promoting 3'-End Antisense Transcription. *PLoS Genetics*, 8(9), article no: e1002952.
- Masse, E. & Drolet, M., 1999. *Escherichia coli* DNA Topoisomerase I inhibits R-loop Formation by relaxing Transcription-induced negative Supercoiling. *Journal of Biological Chemistry*, 274(23), pp. 16659-16664.
- Masumoto, H., Hawke, D., Kobayashi, R. & Verreault, A., 2005. A Role for Cell-cycle-regulated Histone H3 Lysine 56 Acetylation in the DNA Damage Response. *Nature*, 436(7048), pp. 294-298.
- Mathew, A. & Morimoto, R. I., 2006. Role of the Heat-shock Response in the Life and Death of Proteins. *Annals of the New York Academy of Sciences*, 851(1), pp. 99-111.
- Matilainen, O. et al., 2017. The Chromatin Remodeling Factor ISW-1 integrates organismal Responses against Nuclear and Mitochondrial Stress. *Nature Communications*, 8, article no: 1818.
- McAndrew, E. N. & McManus, K. J., 2017. The enigmatic Oncogene and Tumor suppressor-like Properties of RAD54B: Insights into Genome Instability and Cancer. *Genes Chromosomes Cancer*, 56(7), pp. 513-523.
- Mendillo, M. L. et al., 2012. HSF1 drives a Transcriptional Program distinct from Heat Shock to support highly malignant Human Cancers. *Cell*, 150(3), pp. 549-562.
- Mertenskötter, A., Keshet, A., Gerke, P. & Paul, R. J., 2013. The p38 MAPK PMK-1 shows heat-induced Nuclear Translocation, supports Chaperone Expression, and affects the Heat Tolerance of *Caenorhabditis elegans*. *Cell Stress & Chaperones*, 18(3), pp. 293-306.
- Merz, C., Urlaub, H., Will, C. L. & Lührmann, R., 2007. Protein Composition of Human mRNPs spliced *in vitro* and differential Requirements for mRNP Protein Recruitment. *RNA*, 13(1), pp. 116-128.
- Miller, C. J., 2018. *simpleaffy: Very simple high level Analysis of Affymetrix Data*. [Online] Available at: <https://www.bioconductor.org/packages/release/bioc/html/simpleaffy.html> [Accessed 31 October 2018].
- Miller, E. V. et al., 2015. The conserved G-Protein Coupled Receptor FSHR-1 regulates protective Host Responses to Infection and Oxidative Stress. *PLoS One*, 10(9), article no: e0137403.
- Miller, T. et al., 2001. COMPASS: A Complex of Protein Associated with a Trithorax-related SET Domain Protein. *Proceedings of the National Academy of Sciences*, 98(23), pp. 12902-12907.
- Mischo, H. E. et al., 2011. Yeast Sen1 Helicase protects the Genome from Transcription-associated Instability. *Molecular Cell*, 41(1), pp. 21-32.
- Miyata, S., Begun, J., Troemel, E. R. & Ausubel, F. M., 2008. DAF-16-dependent Suppression of Immunity during Reproduction in *Caenorhabditis elegans*. *Genetics*, 178(2), pp. 903-918.
- Mizuno, T. et al., 2004. The *Caenorhabditis elegans* MAPK Phosphatase VHP-1 mediates a novel JNK-like Signaling Pathway in Stress Response. *The EMBO Journal*, 23(11), pp. 2226-2234.

- Moazed, D., 2011. Mechanisms for the Inheritance of Chromatin States. *Cell*, 146(4), pp. 510-518.
- Molly, P. L. & Symons, R. H., 1980. Cleavage of DNA:RNA Hybrids by Type II Restriction Enzymes. *Nucleic Acids Research*, 8(13), pp. 2939-2946.
- Mosesson, Y., Voichek, Y. & Barkai, N., 2014. Divergence and Selectivity of Expression-coupled Histone Modifications in Budding Yeasts. *PLoS One*, 9(7), article no: e101538.
- Mosig, G. et al., 1995. Multiple Initiation Mechanisms adapt Phage T4 DNA Replication to physiological Changes during T4's Development. *FEMS Microbiology Reviews*, 17(1-2), pp. 83-98.
- Muralidharan, S. & Mandrekar, P., 2019. Cellular Stress Response and Innate Immune Signaling: Integrating Pathways in Host Defense and Inflammation. *Journal of Leukocyte Biology*, 94(6), pp. 1167-1184.
- Muramatsu, M. et al., 2000. Class Switch Recombination and Hypermutation require Activation-induced Cytidine Deaminase (AID), a potential RNA Editing Enzyme. *Cell*, 102(5), pp. 553-563.
- Murphy, C. T. & Hu, P. J., 2013. *Insulin/Insulin-like Growth Factor Signaling in C. elegans*. [Online] Available at: http://www.wormbook.org/chapters/www_insulingrowthsignal/insulingrowthsignal.pdf [Accessed 7 11 2019].
- Murray, I. A., Stickel, S. K. & Roberts, R. J., 2010. Sequence-specific cleavage of RNA by Type II Restriction Enzymes. *Nucleic Acids Research*, 38(22), pp. 8257-8268.
- Nadel, J. et al., 2015. RNA:DNA Hybrids in the Human Genome have distinctive Nucleotide Characteristics, Chromatin Composition, and Transcriptional Relationships. *Epigenetics & Chromatin*, 8, article no: 46.
- Nag, P. et al., 2017. Interplay of neuronal and non-neuronal Genes regulates intestinal DAF-16-mediated Immune Response during Fusarium Infection of *Caenorhabditis elegans*. *Cell Death Discovery*, 3, article no: 17073.
- Naidu, S. D., Kostov, R. V. & Dinkova-Kostova, A. T., 2015. Transcription Factors Hsf1 and Nrf2 engage in Crosstalk for Cytoprotection. *Trends in Pharmacological Sciences*, 36(1), pp. 6-14.
- Naji, A. et al., 2018. The Activation of the Oxidative Stress Response Transcription Factor SKN-1 in *Caenorhabditis elegans* by Mitis Group Streptococci. *PLoS One*, 13(8), article no: e0202233.
- NCBI, n.d. *Taxonomy Browser*. [Online] Available at: <https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi> [Accessed 14 February 2019].
- Nelson, S. W. & Benkovic, S. J., 2006. The T4 Phage UvsW Protein contains both DNA Unwinding and Strand Annealing Activities. *Journal of Biological Chemistry*, 282(1), pp. 407-416.
- Neumann, B. & Barlow, D. P., 1996. Multiple Roles for DNA Methylation in gametic Imprinting. *Current Opinion in Genetics & Development*, 6(2), pp. 159-163.
- New England Biolabs, 2019. *CutSmart® Buffer*. [Online] Available at: <https://www.neb.com/products/b7204-cutsmart-buffer#Product%20Information> [Accessed 8 10 2019].
- New England Biolabs, 2019. *Monarch® PCR & DNA Cleanup Kit (5 µg)*. [Online] Available at: <https://international.neb.com/products/t1030-monarch-pcr-dna-cleanup-kit-5-ug> [Accessed 18 July 2019].
- New England Biolabs, 2019. *Star Activity*. [Online] Available at: <https://international.neb.com/tools-and-resources/usage-guidelines/star-activity> [Accessed 17 July 2019].
- New England Biolabs, no date. *RNase H*. [Online] Available at: <https://www.neb.com/products/m0297-rnase-h#Product%20Information> [Accessed 27 November 2018].
- Nicholas, H. R. & Hodgkin, J., 2004. The ERK MAP Kinase Cascade mediates Tail Swelling and a Protective Response to rectal Infection in *C. elegans*. *Current Biology*, 14(14), pp. 1256-1261.

- Niu, W. et al., 2011. Diverse Transcription Factor Binding Features revealed by Genome-wide ChIP-seq in *C. elegans*. *Genome Research*, 21(2), pp. 245-254.
- O'Brien, D. et al., 2018. A PQM-1-mediated Response triggers Transcellular Chaperone Signaling and regulates organismal Proteostasis. *Cell Reports*, 23(13), pp. 3905-3919.
- O'Brien, D. & van Oosten-Hawle, P., 2016. Regulation of Cell-non-autonomous Proteostasis in Metazoans. *Essays in Biochemistry*, 60(2), pp. 133-142.
- Ohle, C. et al., 2016. Transient RNA-DNA Hybrids are required for efficient Double-Strand Break Repair. *Cell*, 167(4), pp. 1001-1013.
- Oh, S. et al., 2016. Medulloblastoma-associated DDX3 Variant selectively alters the Translational Response to Stress. *Oncotarget*, 7(19), pp. 28169-28182.
- Okuyama, T. et al., 2010. The ERK-MAPK Pathway regulates Longevity through SKN-1 and Insulin-like Signaling in *Caenorhabditis elegans*. *Journal of Biological Chemistry*, 285(39), pp. 30274-30281.
- O'Rourke, D. et al., 2006. Genomic Clusters, putative Pathogen Recognition Molecules, and Antimicrobial Genes are induced by Infection of *C. elegans* with *M. nematophilum*. *Genome Research*, 16(8), pp. 1005-1016.
- Orphanides, G. & Reinberg, D., 2000. RNA polymerase II elongation through chromatin. *Nature*, 407(6803), pp. 471-475.
- Osman, G. A. et al., 2018. Natural Infection of *C. elegans* by an Oomycete Reveals a New Pathogen-Specific Immune Response. *Current Biology*, 28(4), pp. 640-648.
- Owen-Hughes, T. et al., 1996. Persistent Site-specific Remodeling of a Nucleosome Array by transient Action of the SWI/SNF Complex. *Science*, 273(5274), pp. 513-516.
- Pacri, R., 2017. R loops in the Regulation of Antibody Gene Diversification. *Genes*, 8(6), article no: 154.
- Page, A. P. & Johnstone, I. L., 2007. *The Cuticle*. [Online]
Available at: http://www.wormbook.org/chapters/www_cuticle/cuticle.html
[Accessed 11 June 2019].
- Pall, G. S. et al., 2007. Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Research*, 35(8), article no: e60.
- Pandya-Jones, A., 2011. Pre-mRNA Splicing during Transcription in the Mammalian System. *Wiley Interdisciplinary Review: RNA*, 2(5), pp. 700-717.
- Pang, P. S., Jankowsky, E., Planet, P. J. & Pyle, A. M., 2002. The Hepatitis C viral NS3 Protein is a processive DNA Helicase with Cofactor enhanced RNA Unwinding. *The EMBO Journal*, 21(5), pp. 1168-1176.
- Papp, D., Csermely, P. & Soti, C., 2012. A Role for SKN-1/Nrf in Pathogen Resistance and Immunosenescence in *Caenorhabditis elegans*. *PLoS Pathogens*, 8(4), article no: e1002673.
- Park, S.-K., Tedesco, P. M. & Johnson, T. E., 2009. Oxidative Stress and Longevity in *C. elegans* as mediated by SKN-1. *Aging Cell*, 8(3), pp. 258-269.
- Patel, S. S. & Donmez, I., 2006. Mechanisms of Helicases. *Journal of Biological Chemistry*, 281(27), pp. 18265-18268.
- Paul, S. et al., 2018. NRF2 transcriptionally activates the Heat Shock Factor 1 Promoter under Oxidative Stress and affects Survival and Migration Potential of MCF7 Cells. *Journal of Biological Chemistry*, 293(50), pp. 19303-19316.
- Paz-Gomez, D., Villanueva-Chimal, E. & Navarro, R. E., 2014. The DEAD Box RNA Helicase VBH-1 is a new Player in the Stress Response in *C. elegans*. *PLoS one*, 9(5), article no: e97924.
- Pedersen, M. T. et al., 2014. The Demethylase JMJD2C localizes to H3K4me3-positive Transcription Start Sites and is dispensable for Embryonic Development. *Molecular and Cellular Biology*, 34(6), pp. 1031-1045.

- Pellegrino, M. W., Nargund, A. M. & Haynes, C. M., 2013. Signaling the Mitochondrial Unfolded Protein Response. *Biochimica et Biophysica Acta*, 1833(2), pp. 410-416.
- Pellegrino, M. W. et al., 2014. Mitochondrial UPR-regulated Innate Immunity provides Resistance to Pathogen Infection. *Nature*, 516(7531), pp. 414-417.
- Peltonen, J. et al., 2013. Chronic Ethanol Exposure increases Cytochrome P-450 and decreases activated in blocked Unfolded Protein Response Gene Family Transcripts in *Caenorhabditis elegans*. *Journal of Biochemical and Molecular Toxicology*, 27(3), pp. 219-228.
- Perry, R. D. & Fetherston, J. D., 1997. *Yersinia pestis*--etiologic Agent of Plague. *Clinical Microbiology Reviews*, 10(1), pp. 35-66.
- Petty, E. & Pillus, L., 2013. Balancing Chromatin Remodeling and Histone Modifications in Transcription. *Trends in Genetics*, 29(11), pp. 621-629.
- Pierson III, L. S. & Pierson, E. A., 2010. Metabolism and Function of Phenazines in Bacteria: Impacts on the Behavior of Bacteria in the Environment and Biotechnological Processes. *Applied Microbiology and Biotechnology*, 86(6), pp. 1659-1670.
- Pingoud, A., 1985. Spermidine increases the Accuracy of Type II Restriction Endonucleases. *European Journal of Biochemistry*, 147(1), pp. 105-109.
- Pokhrel, B., 2019. *The novel Function of CFP-1 in Caenorhabditis elegans Development*. Leeds: University of Leeds.
- Pokhrel, B., Chen, Y. & Biro, J. J., 2019. CFP-1 interacts with HDAC1/2 Complexes in *C. elegans* Development. *The FEBS Journal*, 286(13), pp. 2490-2504.
- Portal-Celhay, C., Bradley, E. R. & Blaser, M. J., 2012. Control of Intestinal Bacterial Proliferation in Regulation of Lifespan in *Caenorhabditis elegans*. *BMC Microbiology*, 12, article no: 49.
- Pradel, E. et al., 2007. Detection and Avoidance of a Natural Product from the pathogenic Bacterium *Serratia marcescens* by *Caenorhabditis elegans*. *Proceedings of the National Academy of Science*, 104(7), pp. 2295-2300.
- Prahlad, V., Cornelius, T. & Morimoto, R. I., 2008. Regulation of the Cellular Heat Shock Response in *Caenorhabditis elegans* by Thermosensory Neurons. *Science*, 320(5877), pp. 811-814.
- Prahlad, V. & Morimoto, R. I., 2009. Integrating the Stress Response: Lessons for Neurodegenerative Diseases from *C. elegans*. *Trends in Cell Biology*, 19(2), pp. 52-61.
- Promega, 2016. *Mung Bean Nuclease*. [Online] Available at: <https://www.promega.co.uk/products/cloning-and-dna-markers/molecular-biology-enzymes-and-reagents/mung-bean-nuclease/?catNum=M4311> [Accessed 1 March 2018].
- Pujol, N. et al., 2008a. Distinct Innate Immune Responses to Infection and Wounding in the *C. elegans* Epidermis. *Current Biology*, 18(7), pp. 481-489.
- Pujol, N. et al., 2001. A Reverse Genetic Analysis of Components of the Toll Signaling Pathway in *Caenorhabditis elegans*. *Current Biology*, 11(11), pp. 809-82.
- Pujol, N. et al., 2008b. Anti-Fungal Innate Immunity in *C. elegans* is enhanced by evolutionary Diversification of Antimicrobial Peptides. *PLoS Pathogens*, 4(7), article no: e1000105.
- Pukkila-Worley, R. & Ausubel, F. M., 2012. Immune Defense Mechanisms in the *Caenorhabditis elegans* Intestinal Epithelium. *Current Opinion in Immunology*, 24(1), pp. 3-9.
- Pukkila-Worley, R., Ausubel, F. M. & Mylonakis, E., 2011. *Candida albicans* Infection of *Caenorhabditis elegans* induces Antifungal Immune Defenses. *PLoS Pathogens*, 7(6), article no: e1002074.
- Pukkila-Worley, R. et al., 2014. The evolutionarily conserved Mediator Subunit MDT-15/MED15 links protective Innate Immune Responses and Xenobiotic Detoxification. *PLoS Pathogens*, 10(5), article no: e1004143.

- Pukkila-Worley, R., Peleg, A. Y., Tampakakis, E. & Mylonakis, E., 2009. *Candida albicans* Hyphal Formation and Virulence assessed using a *Caenorhabditis elegans* Infection Model. *Eukaryotic Cell*, 8(11), pp. 1750-1758.
- Puoti, A. & Kimble, J., 2000. The Hermaphrodite Sperm/Oocyte Switch requires the *Caenorhabditis elegans* Homologs of PRP2 and PRP22. *Proceedings of the National Academy of Sciences of the United States of America*, 97(7), pp. 3276-3281.
- Qi, B., Kniazeva, M. & Han, M., 2017. A Vitamin-B2-sensing Mechanism that regulates Gut Protease activity to impact Animal's Food Behavior and Growth. *eLIFE*, 6, article no: e26243.
- Quinlan, A. R. & Hall, I. M., 2010. BEDTools: a flexible Suite of Utilities for comparing Genomic Features. *Bioinformatics*, 26(6), pp. 841-842.
- R Core Team, 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Raimand, J. et al., 2007. g:Profiler—a Web-based Toolset for functional Profiling of Gene Lists from large-scale Experiments. *Nucleic Acids Research*, 35, pp. W193-W200.
- Raney, K. D., Byrd, A. k. & Aarattuthodiyil, S., 2013. Structure and Mechanisms of SF1 DNA Helicases. *Advances in Experimental Medicine and Biology*, 767, pp. 17-46.
- Ratmeyer, L. et al., 1994. Sequence Specific Thermodynamic and Structural Properties for DNA:RNA Duplexes. *Biochemistry*, 33(17), pp. 5298-5304.
- Reddy, K. et al., 2011. Determinants of R-loop formation at convergent bidirectionally transcribed Trinucleotide Repeats. *Nucleic Acids Research*, 39(5), pp. 1749-1762.
- Reece, J. B. et al., 2011. *Campbell Biology*. 9th ed. San Francisco: Pearson Education.
- Richardson, C. E., Kooistra, T. & Kim, D. H., 2010. An Essential Role for XBP-1 in Host Protection against Immune Activation in *C. elegans*. *Nature*, 463(7284), pp. 1092-1095.
- Riggs, A. D. & Pfeifer, G. P., 1992. X-chromosome Inactivation and Cell Memory. *Trends in Genetics*, 8(5), pp. 169-174.
- Ritchie, M. E. et al., 2015. limma powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Research*, 43(7), article no: e47.
- Robinson, J. T. et al., 2011. Integrative Genomics Viewer. *Nature Biotechnology*, 29(1), pp. 24-26.
- Rodriguez, M., Snoek, L. B., De Bono, M. & Kammenga, J. E., 2013. Worms under Stress: *C. elegans* Stress Response and its Relevance to complex Human Disease and Aging. *Trends in Genetics*, 29(6), pp. 367-374.
- Rogakou, E. P., Pilch, D. R., Ivanova, V. S. & Bonner, W. M., 1998. DNA double-stranded Breaks induce Histone H2AX Phosphorylation on Serine 139. *Journal of Biological Chemistry*, 273(10), pp. 5858-5868.
- Ron, D. & Walter, P., 2007. Signal Integration in the Endoplasmic Reticulum Unfolded Protein Response. *Nature Reviews Molecular Cell Biology*, 8(7), pp. 519-529.
- Rosenfeld, J. A. et al., 2009. Determination of enriched Histone Modifications in non-genic Portions of the Human Genome. *BMC Genomics*, 10, article no: 143.
- Rouger, V. et al., 2014. Independent synchronized Control and Visualization of Interactions between living Cells and Organisms. *Biophysical Journal*, 106(10), pp. 1096-2104.
- Roy, D. & Lieber, M. R., 2009. G Clustering is important for the Initiation of Transcription-Induced R-loops *in vitro*, whereas high G Density without Clustering is sufficient thereafter. *Molecular and Cellular Biology*, 29(11), pp. 3124-3133.
- Rudolph, C. J., Upton, A. L., Briggs, G. S. & Lloyd, R. G., 2010. Is RecG a general Guardian of the Bacterial Genome? *DNA Repair*, 9(3), pp. 210-223.
- Runkel, E. D., Liu, S., Baumeister, R. & Schulze, E., 2013. Surveillance-activated Defenses block the ROS-induced Mitochondrial Unfolded Protein Response. *PLoS Genetics*, 9(3), article no: e1003346.

- Ryu, J.-S., Kang, S. J. & Koo, H.-S., 2013. The 53BP1 Homolog in *C. elegans* influences DNA Repair and promotes Apoptosis in response to Ionizing Radiation. *PLoS One*, 8(5), article no: e64028.
- Saha, A., Wittmeyer, J. & Cairns, B. R., 2006. Chromatin Remodelling: The Industrial Revolution of DNA around Histones. *Nature Reviews Molecular Cell Biology*, 7(6), pp. 437-447.
- Sahu, S. N. et al., 2012. Genomic Analysis of Immune Response against *Vibrio cholerae* Hemolysin in *Caenorhabditis elegans*. *PLoS One*, 7(5), article no: e38200.
- Saito, T. L. et al., 2013. The Transcription Start Site Landscape of *C. elegans*. *Genome Research*, 23(8), pp. 1348-1361.
- Santos-Pereira, J. M. & Aguilera, A., 2015. R loops: new Modulators of Genome Dynamics and Function. *Nature Reviews Genetics*, 16(10), pp. 583-597.
- Sanz, L. A. et al., 2016. Prevalent, dynamic, and conserved R-loop Structures associate with specific Epigenomic Signatures in Mammals. *Molecular Cell*, 63(1), pp. 167-178.
- Sarkies, P. et al., 2013. Competition between Virus-derived and endogenous Small RNAs regulates Gene Expression in *Caenorhabditis elegans*. *Genome Research*, 23(8), pp. 1258-1270.
- Sato, K., Yoshiga, T. & Hasegawa, K., 2014. Activated and inactivated Immune Responses in *Caenorhabditis elegans* against *Photobacterium luminescens* T101. *SpringerPlus*, 3, article no: 274.
- Sato, K., Yoshiga, T. & Hasegawa, K., 2016. Involvement of Vitamin B6 Biosynthesis Pathways in the insecticidal Activity of *Photobacterium luminescens*. *Applied and Environmental Microbiology*, 82(12), pp. 3546-3553.
- Saxonov, S., Berg, P. & Brutlag, D. L., 2006. A Genome-wide Analysis of CpG Dinucleotide in the Human Genome distinguishes two distinct Classes of Promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 1412-1417(2006), pp. 1414-1417.
- Scandalios, J. G., 2002. Oxidative Stress Responses - what have Genome-scale Studies taught us? *Genome Biology*, 3(7), pp. 1019.1-1019.6.
- Schafer, W. R., 2005. *Egg-laying*. [Online]
Available at: http://www.wormbook.org/chapters/www_egglaying/egglaying.html
[Accessed 5 July 2019].
- Schwab, R. A. et al., 2015. The Fanconi Anemia Pathway maintains Genome Stability by coordinating Replication and Transcription. *Molecular Cell*, 60(3), pp. 351-361.
- Sem, X. & Rhen, M., 2012. Pathogenicity of *Salmonella enterica* in *Caenorhabditis elegans* relies on disseminated Oxidative Stress in the infected Host. *PLoS One*, 7(9), article no: e45417.
- Senchuk, M. M. et al., 2018. Activation of DAF-16/FOXO by Reactive Oxygen Species contributes to Longevity in Long-lived Mitochondrial Mutants in *Caenorhabditis elegans*. *PLoS Genetics*, 14(3), article no: e1007268.
- Seo, C., Jang, D., Chae, J. & Shin, S., 2017. Altering the Coffee-Ring Effect by adding a Surfactant-like viscous Polymer Solution. *Scientific Reports*, 7, article no: 500.
- Shapira, M. et al., 2006. A conserved Role for a GATA Transcription Factor in regulating Epithelial Innate Immune Responses. *Proceedings of the National Academy of Science*, 103(38), pp. 14086-14091.
- Sharif, J. & Koseki, H., 2018. Hemimethylation: DNA's lasting odd Couple. *Science*, 359(6380), pp. 1102-1103.
- Shaw, N. N. & Arya, D. P., 2008. Recognition of the unique Structure of DNA:RNA Hybrids. *Biochimie*, 90(7), pp. 1026-1039.
- Shilatifard, A., 2012. The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. *Annual Review of Biochemistry*, 81, pp. 65-95.
- Shivers, R. P. et al., 2009. Tissue-specific activities of an Immune Signaling Module regulate physiological Responses to Pathogenic and nutritional Bacteria in *C. elegans*. *Cell Host Microbe*, 6(4), pp. 321-330.

- Shivers, R. P. et al., 2010. Phosphorylation of the conserved Transcription Factor ATF-7 by PMK-1 p38 MAPK regulates Innate Immunity in *Caenorhabditis elegans*. *PLoS Genetics*, 6(4), article no: e1000892.
- Shi, X. et al., 2011. ING2 PHD Domain links Histone H3 Lysine 4 Methylation to Active Gene Repression. *Nature*, 442(7098), pp. 96-99.
- Shi, X. et al., 2009. Proteome-wide Analysis in *Saccharomyces cerevisiae* identifies several PHD Fingers as novel direct and selective Binding Modules of Histone H3 Methylated at either Lysine 4 or Lysine 36. *Journal of Biological Chemistry*, 282(4), pp. 2450-2455.
- Shi, Y., Mosser, D. D. & Morimoto, R. I., 1998. Molecular Chaperones as HSF1-specific Transcriptional Repressors. *Genes & Development*, 12(5), pp. 654-666.
- Sicard, M. et al., 2007. The Effect of *Photobhabdus luminescens* (Enterobacteriaceae) on the Survival, Development, Reproduction and Behaviour of *Caenorhabditis elegans* (Nematoda: Rhabditidae). *Environmental Microbiology*, 9(1), pp. 12-25.
- Siebert, M. & Söding, J., 2016. Bayesian Markov Models consistently outperform PWMs at predicting Motifs in Nucleotide Sequences. *Nucleic Acids Research*, 44(13), pp. 6055-6069.
- Siegfried, Z. & Cedar, H., 1997. DNA Methylation: A Molecular Lock. *Current Biology*, 7(5), pp. R305-R307.
- Sies, H., Berndt, C. & Jones, D. P., 2017. Oxidative Stress. *Annual Review of Biochemistry*, 86, pp. 715-748.
- Sifri, C. D., Begun, J., Ausubel, F. M. & Calderwood, S. B., 2003. *Caenorhabditis elegans* as a Model Host for *Staphylococcus aureus* Pathogenesis. *Infection and Immunity*, 71(4), pp. 2208-2217.
- Simonet, T., Dulermo, R., Schott, S. & Palladino, F., 2007. Antagonistic Functions of SET-2/SET1 and HPL/HP1 Proteins in *C. elegans* Development. *Developmental Biology*, 312(1), pp. 367-383.
- Sims III, R. J. et al., 2007. Recognition of trimethylated Histone H3 Lysine 4 facilitates the Recruitment of Transcription Post-initiation Factors and pre-mRNA Splicing. *Molecular Cell*, 28(4), pp. 665-676.
- Singh, V. & Aballay, A., 2006. Heat-shock Transcription Factor (HSF)-1 Pathway required for *Caenorhabditis elegans* Immunity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35), pp. 13092-13097.
- Singleton, M. R., Dillingham, M. S. & Wigley, D. B., 2007. Structure and Mechanism of Helicases and Nucleic Acid Translocases. *Annual Review of Biochemistry*, 76, pp. 23-50.
- Singleton, M. R. & Wigley, D. B., 2002. Modularity and Specialization in Superfamily 1 and 2 Helicases. *Journal of Bacteriology*, 184(7), pp. 1819-1826.
- Sinha, A., Rae, R., Iatsenko, I. & Sommer, R. J., 2012. System Wide Analysis of the Evolution of Innate Immunity in the Nematode Model Species *Caenorhabditis elegans* and *Pristionchus pacificus*. *PLoS One*, 7(9), article no: e44255.
- Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J., 2014. R-loops induce Repressive Chromatin Marks over Mammalian Gene Terminators. *Nature*, 516(7531), pp. 436-439.
- Skourti-Stathaki, K. & Proudfoot, N. J., 2014. A Double-edged Sword: R loops as Threats to Genome Integrity and powerful Regulators of Gene Expression. *Genes & Development*, 28(13), pp. 1384-1396.
- Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N., 2011. Human Senataxin resolves RNA/DNA Hybrids formed at Transcriptional Pause Sites to promote Xrn2-dependent Termination. *Molecular Cell*, 42(6), pp. 794-805.
- Smyth, G. K. et al., 2019. *limma*. [Online]
Available at:
<https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>
[Accessed 1 May 2019].
- Sollier, J. & Cimprich, K. A., 2015. R-loops Breaking Bad. *Trends in Cell Biology*, 25(9), pp. 514-522.

- Sollier, J. et al., 2014. Transcription-coupled Nucleotide Excision Repair Factors promote R-loop-induced Genome Instability. *Molecular Cell*, 56(6), pp. 777-785.
- Solomon, A. et al., 2004. *Caenorhabditis elegans* OSR-1 regulates behavioral and physiological Responses to Hyperosmotic Environments. *Genetics*, 167(1), pp. 161-170.
- Song, C., Hotz-Wagenblatt, A., Voit, R. & Grummt, I., 2017. SIRT7 and the DEAD-box Helicase DDX21 cooperate to resolve genomic R loops and safeguard Genome Stability. *Genes & Development*, 31(13), pp. 1370-1381.
- Soto, C. & Estrada, L. D., 2008. Protein Misfolding and Neurodegeneration. *Neurological Review*, 65(2), pp. 184-189.
- Souza, A. C. R. et al., 2018. Pathogenesis of the *Candida Parapsilosis* Complex in the Model Host *Caenorhabditis elegans*. *Genes*, 9(8), article no: 401.
- Spencer, W. C. et al., 2011. A spatial and temporal Map of *C. elegans* Gene Expression. *Genome Research*, 21(2), pp. 325-341.
- Stasser, M. J. et al., 1995. The *Drosophila Trithorax* Proteins contain a novel Variant of the Nuclear Receptor Type DNA Binding Domain and an ancient conserved Motif found in other Chromosomal Proteins. *Mechanism of Development*, 52(2-3), pp. 209-223.
- Stempor, P. & Ahringer, J., 2016. SeqPlots - Interactive Software for exploratory Data Analyses, Pattern Discovery and Visualization in Genomics. *Wellcome Open Research*, 1, article no: 14.
- Stiernagle, T., 2006. *Wormbook*. [Online]
Available at: http://www.wormbook.org/chapters/www_strainmaintain/strainmaintain.pdf
[Accessed 7 June 2017].
- Storm, M., Sheng, X., Arnoldussen, Y. J. & Saaticioglu, F., 2016. Prostate Cancer and the Unfolded Protein Response. *Oncotarget*, 7(33), pp. 54051-54066.
- Styler, K. L. et al., 2005. *Yersinia pestis* kills *Caenorhabditis elegans* by a Biofilm-independent Process that involves novel Virulence Factors. *EMBO reports*, 6(10), pp. 992-997.
- Sun, H., Yabuki, A. & Maizels, N., 2001. A Human Nuclease Specific for G4 DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22), pp. 12444-12449.
- Sun, J., Singh, V., Kajino-Sakamoto, R. & Aballay, A., 2011. Neuronal GPCR controls Innate Immunity by regulating non-canonical Unfolded Protein Response Genes. *Science*, 332(6030), pp. 729-732.
- Szklarczyk, D. et al., 2019. STRING v11: Protein-Protein Association Networks with increased Coverage, supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Research*, 47(D1), pp. D607-D613.
- Taffoni, C. & Pujol, N., 2015. Mechanisms of Innate Immunity in *C. elegans* Epidermis. *Tissue Barriers*, 3(4), article no: e1078432.
- Tajima, k. et al., 2015. SETD1A modulates Cell Cycle Progression through a miRNA Network that regulates p53 Target Genes. *Nature Communications*, 6, article no: 8257.
- Takara, n.d. *Mung bean nuclease*. [Online]
Available at: <https://www.takarabio.com/products/cloning/modifying-enzymes/nucleases/mung-bean-nuclease>
[Accessed 24 January 2019].
- Takemori, Y. et al., 2006. Stress-induced Transcription of the Endoplasmic Reticulum Oxidoreductin Gene ERO1 in the Yeast *Saccharomyces cerevisiae*. *Molecular Genetics and Genomics*, 275(1), pp. 89-96.
- Takeshima, H. et al., 2009. The presence of RNA Polymerase II, active or stalled, predicts Epigenetic Fate of Promoter CpG Islands. *Genome Research*, 19(11), pp. 1974-1982.
- Tang, L. & Choe, K. P., 2015. Characterization of *skn-1/wdr-23* Phenotypes in *Caenorhabditis elegans*; Pleiotrophy, Aging, Glutathione, and Interactions with other Longevity Pathways. *Mechanisms of Ageing and Development*, 149, pp. 88-89.

- Tanguy, M. et al., 2017. An alternative STAT Signaling Pathway acts in Viral Immunity in *Caenorhabditis elegans*. *mBio*, 8(5), article no: e00924-17.
- Tan, M.-W., Mahajan-Miklos, S. & Ausubel, F. M., 1999. Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model Mammalian Bacterial Pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(2), pp. 715-720.
- Tanner, N. & Linder, P., 2001. DExD/H Box RNA Helicases: from generic Motors to specific Dissociation Functions. *Molecular Cell*, 8(2), pp. 251-262.
- Tatum, M. C. et al., 2015. Neuronal Serotonin Release triggers the Heat Shock Response in *C. elegans* in the absence of Temperature Increase. *Current Biology*, 25(2), pp. 163-174.
- Taverna, S. D. et al., 2015. Yng1 PHD Finger binding to H3 trimethylated at K4 promotes NuA3 HAT Activity at K14 of H3 and Transcription at a Subset of targeted ORFs. *Molecular Cell*, 24(5), pp. 785-796.
- Tawe, W. N., Eschbach, M.-L., Walter, R. D. & Henkle-Dührsen, K., 1998. Identification of Stress-responsive Genes in *Caenorhabditis elegans* using RT-PCR differential Display. *Nucleic Acids Research*, 26(7), pp. 1621-1627.
- Tenor, J. L. & Aballay, A., 2008. A conserved Toll-like Receptor is required for *Caenorhabditis elegans* Innate Immunity. *EMBO Reports*, 9(1), pp. 103-109.
- Tepper, R. G. et al., 2013. PQM-1 complements DAF-16 as a Key Transcriptional Regulator of DAF-2-mediated Development and Longevity. *Cell*, 154(3), pp. 676-690.
- The *C. elegans* Deletion Mutant Consortium, 2012. Large-Scale Screening for Targeted Knockouts in the *Caenorhabditis elegans* Genome. *G3 (Bethesda)*, 2(11), pp. 1415-1425.
- The Uniprot Consortium, 2019. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Research*, 47(D1), pp. D506-D515.
- Thermo Fisher Scientific, 2019. *dsDNase*. [Online]
Available at: <https://www.thermofisher.com/order/catalog/product/EN0771>
[Accessed 17 July 2019].
- Thomas, M., White, R. & David, R., 1976. Hybridization of RNA to double-stranded DNA: Formation of R-loops. *Proceedings of the National Academy of Sciences of the United States of America*, 73(7), pp. 2294-2298.
- Tjahjono, E. & Kirienko, N. V., 2017. A conserved Mitochondrial Surveillance Pathway is required for Defense against *Pseudomonas aeruginosa*. *PLoS Genetics*, 13(6), article no: e1006876.
- Torres, I. O. & Fujimori, D. G., 2015. Functional coupling between Writers, Erasers and Readers of Histone and DNA Methylation. *Current Opinion in Structural Biology*, 35, pp. 68-75.
- Tous, C. & Aguilera, A., 2007. Impairment of Transcription Elongation by R-loops *in vitro*. *Biochemical and Biophysical Research Communications*, 360(2), pp. 428-432.
- Tran, A. et al., 2017. *C. elegans* avoids Toxin-producing *Streptomyces* using a seven Transmembrane Domain Chemosensory Receptor. *eLife*, 6, article no: e23770.
- Troemel, E. R. et al., 2006. p38 MAPK regulates Expression of Immune Response Genes and contributes to Longevity in *C. elegans*. *PLoS Genetics*, 2(11), article no: e183.
- Troemel, E. R. et al., 2008. Microsporidia are natural intracellular Parasites of the Nematode *Caenorhabditis elegans*. *PLoS Biology*, 6(12), article no: e309.
- Trojer, P. & Reinberg, D., 2007. Facultative Heterochromatin: is there a distinctive Molecular Signature? *Molecular Cell*, 28(1), pp. 1-13.
- Tsompana, M. & Buck, M. J., 2014. Chromatin Accessibility: A Window into the Genome. *Epigenetics & Chromatin*, 7(1), article no: 33.
- Tuduri, S. et al., 2009. Topoisomerase I suppresses genomic Instability by preventing Interference between Replication and Transcription. *Nature Cell Biology*, 11(11), pp. 1315-1324.

- Tullet, J. M. et al., 2017. The SKN-1/Nrf2 Transcription Factor can protect against Oxidative Stress and increase Lifespan in *C. elegans* by distinct Mechanisms. *Aging Cell*, 16(5), pp. 1191-1194.
- Tuma, R., 2010. Hexameric viral RNA Helicases. In: E. Jankowsky, ed. *RNA helicases*. Cambridge: RSC Publishing, pp. 213-242.
- Twumasi-Boateng, K. et al., 2012. An age-dependent Reversal in the protective Capacities of JNK Signaling shortens *Caenorhabditis elegans* Lifespan. *Aging Cell*, 11(4), pp. 659-667.
- Umate, P., Tuteja, N. & Tuteja, R., 2011. Genome-wide comprehensive Analysis of Human Helicases. *Communicative & Integrative Biology*, 4(1), pp. 118-137.
- Uribe, S. & Sampedro, J. G., 2003. Measuring Solution Viscosity and its Effect on Enzyme Activity. *Biological Procedures Online*, 5, pp. 108-115.
- Vabulas, R. M. et al., 2001. Endocytosed HSP60s use Toll-like Receptor 2 (TLR2) and TLR4 to activate the Toll/Interleukin-1 Receptor Signaling Pathway in Innate Immune Cells. *Journal of Biological Chemistry*, 276(33), pp. 31332-31339.
- Vaitkevicius, K. et al., 2006. A *Vibrio cholerae* Protease needed for killing of *Caenorhabditis elegans* has a Role in Protection from natural Predator grazing. *Proceedings of the National Academy of Sciences of the United States of America*, 103(24), pp. 9280-9285.
- Valsala, G. & Sugathan, S., 2017. Enzymes as Molecular Tools. In: S. Sugathan, N. S. Pradeep & S. Abdulhameed, eds. *Bioresources and Bioprocess in Biotechnology: Volume 2: Exploring Potential Biomolecules*. s.l.:Springer Singapore, pp. 99-128.
- van der Hoeven, R., McCallum, K. C., Cruz, M. R. & Garsin, D. A., 2011. Ce-Duox1/BLI-3 generated Reactive Oxygen Species trigger protective SKN-1 Activity via p38 MAPK Signaling during Infection in *C. elegans*. *PLoS Pathogens*, 7(12), article no: e1002453.
- van Ingen, H. et al., 2008. Structural Insight into the Recognition of the H3K4me3 Mark by the TFIID Subunit TAF3. *Structure*, 16(8), pp. 1245-1256.
- van Oosten-Hawle, P. & Morimoto, R. I., 2014. Transcellular Chaperone Signaling: An organismal Strategy for integrated Cell Stress Responses. *Journal of Experimental Biology*, 217, pp. 129-136.
- van Oosten-Hawle, P., Porter, R. S. & Morimoto, R. I., 2013. Regulation of Organismal Proteostasis by Transcellular Chaperone Signaling. *Cell*, 153(6), pp. 1366-1378.
- Vanoosthuyse, V., 2018. Strengths and Weaknesses of the current Strategies to map and characterize R-loops. *Non-coding RNA*, 4(2), article no: E9.
- Vardevanyan, P. et al., 2013. Mechanisms for Binding between Methylene Blue and DNA. *Journal of Applied Spectroscopy*, 80(4), pp. 595-599.
- Varga-Weisz, P. D. et al., 1997. Chromatin-remodelling Factor CHRAC contains the ATPases ISWI and Topoisomerase II. *Nature*, 388(6642), pp. 598-602.
- Varshavsky, A., 1997. The N-end Rule Pathway of Protein Degradation. *Genes to Cells*, 2(1), pp. 13-28.
- Varshavsky, A., 2019. N-degron and C-degron Pathways of Protein Degradation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(2), pp. 358-366.
- Verghese, J., Abrams, J., Wang, Y. & Morano, K. A., 2012. Biology of the Heat Shock Response and Protein Chaperones: Budding Yeast (*Saccharomyces cerevisiae*) as a Model System. *Microbiology and Molecular Biology Reviews*, 76(2), pp. 115-158.
- Vihervaara, A. & Sistonen, L., 2014. HSF1 at a Glance. *Journal of Cell Science*, 127, pp. 261-266.
- Waddington, C. H., 1942. The Epigenotype. *Endeavour*, 1, pp. 18-20.
- Wahl, M. C., Will, C. L. & Lührmann, R., 2009. The Spliceosome: Design Principles of a dynamic RNP Machine. *Cell*, 136(4), pp. 701-718.
- Walstrom, K. M., Schmidt, D., Bean, C. J. & Kelly, W. G., 2005. RNA Helicase A is important for Germline Transcriptional Control, Proliferation, and Meiosis in *C. elegans*. *Mechanism of Development*, 122(5), pp. 707-720.

- Wang, C. et al., 2007. Morphological Characteristics and Infection Processes of nematophagous *Harposporium* with Reference to two new Species. *Fungal Diversity*, 26(1), pp. 287-304.
- Wang, J., Nakad, R. & Schulenburg, H., 2012. Activation of the *Caenorhabditis elegans* FOXO Family Transcription Factor DAF-16 by pathogenic *Bacillus thuringiensis*. *Developmental & Comparative Immunology*, 37(1), pp. 193-201.
- Wang, M., Zhao, Y. & Zhang, B., 2015. Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports*, 5, article no: 16923.
- Wang, X., Lim, H. J. & Son, A., 2014. Characterization of Denaturation and Renaturation of DNA for DNA Hybridization. *Environmental Health and Toxicology*, 29, article no: e2014007.
- Wan, L. et al., 2019. *Bacillus thuringiensis* targets the Host Intestinal Epithelial Junctions for successful Infection of *Caenorhabditis elegans*. *Environmental Microbiology*, 21(3), pp. 1086-1098.
- Warnes, G. R. et al., 2019. *gplots: Various R Programming Tools for plotting Data*. [Online] Available at: <https://cran.r-project.org/web/packages/gplots/index.html> [Accessed 27 01 2019].
- Wei, J.-Z. et al., 2003. *Bacillus thuringiensis* Crystal Proteins that target Nematodes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5), pp. 2760-2765.
- Wei, J.-W., Huang, K., Yang, C. & Kang, C.-S., 2016. Non-coding RNAs as Regulators in Epigenetics (Review). *Oncology Reports*, 37(1), pp. 3-9.
- Weiner, A. et al., 2012. Systematic dissection of Roles for Chromatin Regulators in a Yeast Stress Response. *PLoS Biology*, 10(7), article no: e1001369.
- Wei, S. et al., 2018. Histone Methylation in DNA Repair and clinical Practice: new Findings during the past 5-years. *Journal of Cancer*, 9(12), pp. 2072-2081.
- Welch, W., 1993. How Cells respond to Stress. *Scientific American*, 268(5), pp. 56-64.
- White, C. V., Darby, B. J., Breeden, R. J. & Herman, M. A., 2016. A *Stenotrophomonas maltophilia* Strain evades a major *Caenorhabditis elegans* Defense Pathway. *Infection and Immunity*, 84(2), pp. 524-536.
- White, C. V. & Herman, M. A., 2018. Transcriptomic, Functional, and Network Analyses reveal novel Genes involved in the Interaction between *Caenorhabditis elegans* and *Stenotrophomonas maltophilia*. *Frontiers in Cellular and Infection Microbiology*, 8, article no: 266.
- Whitehouse, I. & Smith, D. J., 2013. Chromatin Dynamics at the Replication Fork: there's more to Life than Histones. *Current Opinion in Genetics & Development*, 23(2), pp. 120-146.
- Whiteside, S. T. & Goodbourn, S., 1993. Signal Transduction and Nuclear Targeting: Regulation of Transcription Factor Activity by Subcellular Localisation. *Journal of Cell Sciences*, 104, pp. 949-955.
- Wolkow, C., Derndon, L. & Hall, D., 2017. *The Aging Pharynx*. [Online] Available at: <https://www.wormatlas.org/aging/pharynx/APhaframeset.html> [Accessed 11 June 2019].
- Wong, D. et al., 2007. Genome-wide Investigation reveals Pathogen-specific and shared Signatures in the Response of *Caenorhabditis elegans* to Infection. *Genome Biology*, 8(9), article no: R194.
- Wormbase, 2019. *WormBase*. [Online] Available at: <https://wormbase.org/#012-34-5> [Accessed 9 July 2019].
- Wreczycka, K. et al., 2019. HOT or not: examining the Basis of high-occupancy Target Regions. *Nucleic Acids Research*, 47(11), pp. 5735-5745.
- Wu, C.-t. & Morris, J., 2001. Genes, Genetics, and Epigenetics: A Correspondence. *Science*, 293(5532), pp. 1103-1105.
- Wu, Y., 2012. Unwinding and Rewinding: Double Faces of Helicase? *Journal of Nucleic Acids*, 2012, article no: 140601.

- Wysocka, J. et al., 2006. A PHD Finger of NURF couples Histone H3 Lysine 4 Trimethylation with Chromatin Remodelling. *Nature*, 442(7098), pp. 86-90.
- Xiao, Y. et al., 2011. *Caenorhabditis elegans* Chromatin-associated Proteins SET-2 and ASH-2 are differentially required for Histone H3 Lys 4 Methylation in Embryos and Adult Germ Cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(20), pp. 8305-8310.
- Yamamoto, A. et al., 2007. Role of Heat Shock Transcription Factor in *Saccharomyces cerevisiae* Oxidative Stress Response. *Eukaryotic Cell*, 6(8), pp. 1373-1379.
- Yang, W. et al., 2015. Overlapping and unique Signatures in the Proteomic and Transcriptomic Responses of the Nematode *Caenorhabditis elegans* toward pathogenic *Bacillus thuringiensis*. *Developmental & Comparative Immunology*, 51(1), pp. 1-9.
- Yang, W., Dierking, K., Rosenstiel, P. C. & Schulenburg, H., 2016. GATA Transcription Factor as a likely key Regulator of the *Caenorhabditis elegans* Innate Immune Response against Gut Pathogens. *Zoology*, 119(4), pp. 244-253.
- Yang, Y. et al., 2014. Arginine Methylation facilitates the Recruitment of TOP3B to Chromatin to prevent R-loop Accumulation. *Molecular Cell*, 53(3), pp. 484-497.
- Yekdavalli, V. S. et al., 2004. Requirement of DDX3 DEAD Box RNA Helicase for HIV-1 Rev-RRE Export Function. *Cell*, 119(3), pp. 381-392.
- Youngman, M. J., Rogers, Z. N. & Kim, D. H., 2011. A Decline in p38 MAPK Signaling underlies Immunosenescence in *Caenorhabditis elegans*. *PLoS Genetics*, 7(5), article no: e1002082.
- Yuen, G. J. & Ausubel, F. M., 2018. Both live and dead *Enterococci* activate *Caenorhabditis elegans* Host Defense via Immune and Stress Pathways. *Virulence*, 9(1), pp. 683-699.
- Yu, N.-K., Baek, S. H. & Kaang, B.-K., 2011. DNA Methylation-mediated Control of Learning and Memory. *Molecular Brain*, 4, article no: 5.
- Yunker, P. J., Still, T., Lohr, M. A. & Yodh, A., 2011. Suppression of the Coffee-Ring Effect by Shape-dependent Capillary Interactions. *Nature*, 476(7360), pp. 308-311.
- Zeller, P. et al., 2016. Histone H3K9 Methylation is dispensable for *Caenorhabditis elegans* Development but suppresses RNA:DNA Hybrid-associated Repeat Instability. *Nature Genetics*, 48(11), pp. 1385-1395.
- Zhang, L. et al., 2016. Insights into adaptations to a near-obligate Nematode Endoparasitic Lifestyle from the finished Genome of *Drechmeria coniospora*. *Scientific Reports*, 6, article no: 23122.
- Zhang, P., Judy, M., Lee, S.-J. & Kenyon, C., 2014. Direct and indirect Gene Regulation by a Life-extending FOXO Protein in *C. elegans*: Roles for GATA Factors and Lipid Gene Regulators. *Cell Metabolism*, 17(1), pp. 85-100.
- Zhang, T., Cooper, S. & Brockdorff, N., 2015. The Interplay of Histone Modifications - Writers that read. *EMBO Reports*, 16(11), pp. 1467-1481.
- Zhang, Y. et al., 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), article no: R137.
- Zhang, Y. et al., 2006. DNA Translocation and Loop Formation Mechanism of Chromatin Remodeling by SWI/SNF and RSC. *Molecular Cell*, 24(4), pp. 559-568.
- Zhang, Z. et al., 2017. FOS-1 functions as a Transcriptional Activator downstream of the *C. elegans* JNK Homolog KGB-1. *Cellular Signalling*, 30, pp. 1-8.
- Zhao, Y. & Garcia, B. A., 2015. Comprehensive Catalog of currently documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology*, 7(9), article no: a025064.
- Zhu, A., Ibrahim, J. G. & Love, M. I., 2018. Heavy-tailed prior Distributions for Sequence Count Data: removing the Noise and preserving large Differences. *Bioinformatics*, 35(12), pp. 2084-2092.
- Zhu, Y., Stephens, R. M., Meltzer, P. S. & Davis, S. R., 2013. SRadb: query and use public next-generation Sequencing Data from within R. *BMC Bioinformatics*, 14, article no: 19.

Zou, C.-G., Ma, Y.-C., Dai, L.-L. & Zhang, K.-Q., 2014. Autophagy protects *C. elegans* against Necrosis during *Pseudomonas aeruginosa* Infection. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), pp. 12480-1285.

Zou, J. et al., 1998. Repression of Heat Shock Transcription Factor HSF1 Activation by HSP90 (HSP90 Complex) that forms a Stress-sensitive Complex with HSF1. *Cell*, 94(4), pp. 471-480.

Zügel, U. & Kaufmann, S. H., 1999. Role of Heat Shock Proteins in Protection from and Pathogenesis of Infectious Diseases. *Clinical Microbiology Reviews*, 12(1), pp. 19-39.

Zuryn, S. & Jarriault, S., 2013. Deep Sequencing Strategies for Mapping and Identifying Mutations from Genetic Screens. *Worm*, 2(3), article no: e25081.

Appendix 1

a)

Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	T	Q	T ₁ Q	U
nervous system development	GO:0007399	1.666×10 ⁻⁶		372	100	19	10256
plasma membrane bounded cell projection organization	GO:0120036	1.761×10 ⁻⁵		298	100	16	10256
anatomical structure morphogenesis	GO:0009653	2.124×10 ⁻⁵		691	100	24	10256
cell projection organization	GO:0030030	3.063×10 ⁻⁵		310	100	16	10256
positive regulation of developmental process	GO:0051094	4.612×10 ⁻⁵		238	100	14	10256
generation of neurons	GO:0048699	5.434×10 ⁻⁵		281	100	15	10256
negative regulation of chromatin organization	GO:1905268	6.603×10 ⁻⁵		13	100	5	10256
neurogenesis	GO:0022008	8.595×10 ⁻⁵		291	100	15	10256
regulation of nervous system development	GO:0051960	8.842×10 ⁻⁵		143	100	11	10256
neuron differentiation	GO:0030182	1.197×10 ⁻⁴		257	100	14	10256
regulation of plasma membrane bounded cell projection...	GO:0120035	1.268×10 ⁻⁴		117	100	10	10256
regulation of cell projection organization	GO:0031344	1.883×10 ⁻⁴		122	100	10	10256
nucleosome positioning	GO:0016584	1.995×10 ⁻⁴		7	100	4	10256
regulation of neurogenesis	GO:0050767	2.033×10 ⁻⁴		123	100	10	10256
regulation of neuron projection development	GO:0010975	3.463×10 ⁻⁴		100	100	9	10256
regulation of cell differentiation	GO:0045595	3.651×10 ⁻⁴		240	100	13	10256
negative regulation of chromatin silencing	GO:0031936	3.960×10 ⁻⁴		8	100	4	10256
multicellular organism development	GO:0007275	4.066×10 ⁻⁴		1621	100	36	10256
cell morphogenesis involved in differentiation	GO:0000904	4.258×10 ⁻⁴		167	100	11	10256
tissue development	GO:0009888	4.537×10 ⁻⁴		205	100	12	10256
cell differentiation	GO:0030154	5.772×10 ⁻⁴		825	100	24	10256
developmental process	GO:0032502	8.075×10 ⁻⁴		1975	100	40	10256
system development	GO:0048731	8.127×10 ⁻⁴		720	100	22	10256
regulation of cell morphogenesis involved in differentiat...	GO:0010769	9.108×10 ⁻⁴		83	100	8	10256
regulation of neuron differentiation	GO:0045664	9.792×10 ⁻⁴		113	100	9	10256
anatomical structure development	GO:0048856	1.097×10 ⁻³		1841	100	38	10256
positive regulation of gene expression, epigenetic	GO:0045815	1.170×10 ⁻³		10	100	4	10256
positive regulation of multicellular organismal process	GO:0051240	1.378×10 ⁻³		270	100	13	10256
cellular developmental process	GO:0048869	1.621×10 ⁻³		938	100	25	10256
cellular component organization	GO:0016043	2.005×10 ⁻³		2125	100	41	10256
positive regulation of nervous system development	GO:0051962	2.035×10 ⁻³		65	100	7	10256
cell morphogenesis involved in neuron differentiation	GO:0048667	2.421×10 ⁻³		161	100	10	10256
regulation of cell development	GO:0060284	2.560×10 ⁻³		162	100	10	10256
regulation of developmental process	GO:0050793	2.624×10 ⁻³		595	100	19	10256
positive regulation of neuron projection development	GO:0010976	2.647×10 ⁻³		44	100	6	10256
negative regulation of gene silencing	GO:0060969	2.717×10 ⁻³		12	100	4	10256
regulation of chromatin silencing	GO:0031935	2.717×10 ⁻³		12	100	4	10256
negative regulation of cellular process	GO:0048523	2.737×10 ⁻³		714	100	21	10256
neuron projection development	GO:0031175	3.189×10 ⁻³		205	100	11	10256
positive regulation of neuron differentiation	GO:0045666	3.458×10 ⁻³		46	100	6	10256
neuron projection morphogenesis	GO:0048812	3.546×10 ⁻³		168	100	10	10256
plasma membrane bounded cell projection morphogene...	GO:0120039	3.546×10 ⁻³		168	100	10	10256
cell projection morphogenesis	GO:0048858	3.940×10 ⁻³		170	100	10	10256
regulation of chromatin organization	GO:1902275	4.490×10 ⁻³		28	100	5	10256
cell part morphogenesis	GO:0032990	4.845×10 ⁻³		174	100	10	10256

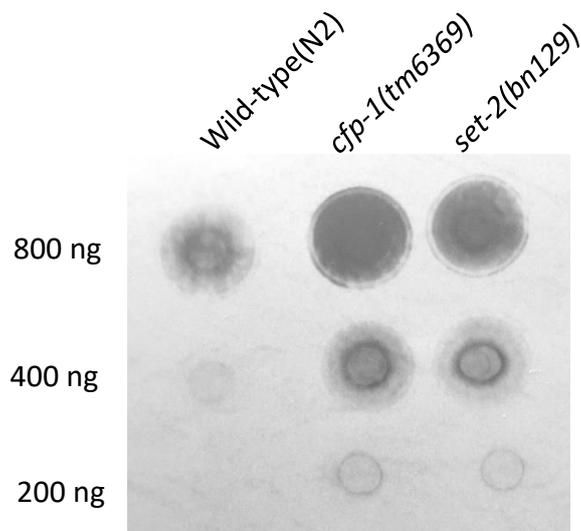
b)

Term	Expected	Observed	Enrichment Fold Change	P-value	Q value
development of primary sexual characteristics GO:0045137	0.34	4	12	2.6e-05	0.0032
reproductive system development GO:0061458	0.41	4	9.8	6.1e-05	0.0037
protein heterodimerization activity GO:0046982	0.34	3	8.8	0.0004	0.016

Appendix 1 Gene set enrichment analysis for H3K4me3 enriched genes in *cfp-1(tm6369)* and *set-2(bn129)* mutants.

a) g:Profiler analysis of the 195 H3K4me3 enriched genes unique to *set-2(bn129)*. The p-value threshold was kept at 0.005, and only biological processes are shown to reduce the size of the list. Analysis using ShinyGO software only returned 4 terms all of which were related to chromatin and nucleosome. b) Wormbase GSEA software result for biological processes of the 53 H3K4me3 enriched genes unique to *cfp-1(tm6369)*. Both g:Profiler and ShinyGO were unable to find any enriched GO term.

Appendix 2



Appendix 2 Dot blot result showing inconsistent when compared to Figure 4.1. Genetic samples used in this blot were collected from young adult wild-type(N2), *cfp-1(tm6369)* and *set-2(bn129)* worms fed on the standard laboratory *E. coli* strain OP50. The quantity of gDNA spotted is shown on the left. Here, wild-type shows the lowest R-loop signal, while *set-2(bn129)* signal is relatively strong, which is the opposite result of Figure 4.1.

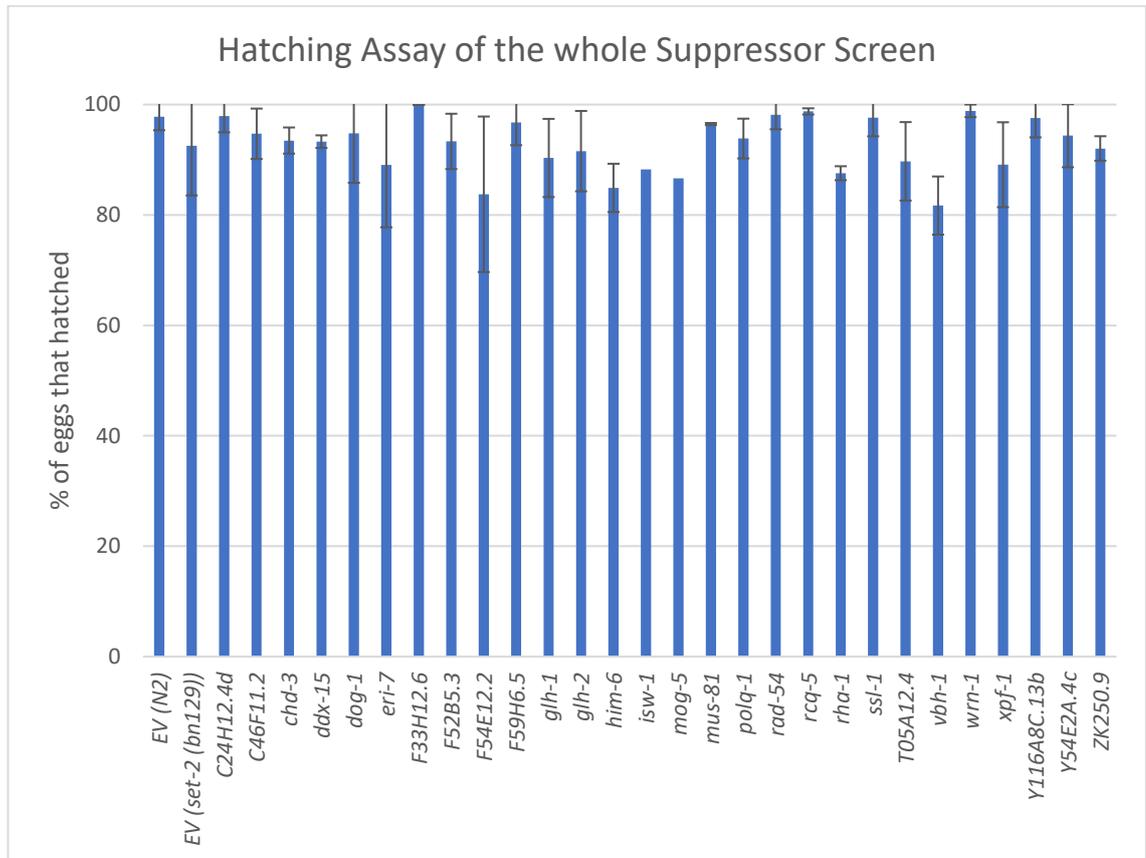
Appendix 3

Sequence Name	Gene Symbol
<i>T12F5.3.2</i>	<i>glh-4</i>
<i>C41D11.7</i>	<i>eri-7</i>
<i>C43E11.2b</i>	<i>mus-81</i>
<i>F55A12.8</i>	<i>nath-10</i>
<i>F33D11.10</i>	
<i>F55F8.2b</i>	
<i>T05E8.3</i>	<i>let-355</i>
<i>T23H2.3</i>	
<i>C55B7.1</i>	<i>glh-2</i>
<i>T21G5.3</i>	<i>glh-1</i>
<i>F52B5.3</i>	
<i>W06D4.6</i>	<i>rad-54</i>
<i>B0511.6</i>	
<i>F33H2.1.2</i>	<i>dog-1</i>
<i>C24H12.4d</i>	
<i>F59H6.5</i>	
<i>ZK250.9</i>	
<i>F33H12.6</i>	
<i>EEED8.5</i>	<i>mog-5</i>
<i>F18C5.2</i>	<i>wrn-1</i>
<i>T07D4.3</i>	<i>rha-1</i>
<i>C47D12.8</i>	<i>xpf-1</i>
<i>Y17G7B.5b</i>	<i>mcm-2</i>
<i>F43G6.1b</i>	<i>dna-2</i>
<i>Y54E2A.4c</i>	
<i>C46F11.4</i>	
<i>E03A3.2</i>	<i>rcq-5</i>
<i>R10E4.4.2</i>	<i>mcm-5</i>
<i>F56D2.6b</i>	<i>ddx-15</i>
<i>W03A3.2</i>	<i>polq-1</i>
<i>F01F1.7</i>	<i>ddx-23</i>
<i>F57B9.3</i>	
<i>ZK686.2</i>	

Sequence Name	Gene Symbol
<i>C07H6.5</i>	<i>cgh-1</i>
<i>C06E1.10</i>	<i>rha-2</i>
<i>T26G10.1</i>	
<i>K03H1.2</i>	<i>mog-1</i>
<i>M03C11.2</i>	<i>chl-1</i>
<i>M03C11.8</i>	
<i>Y111B2A.22d</i>	<i>ssl-1</i>
<i>F53H1.1e</i>	
<i>T05A12.4</i>	
<i>C27B7.4</i>	<i>rad-26</i>
<i>W08D2.7</i>	<i>mtr-4</i>
<i>C04H5.6b</i>	<i>mog-4</i>
<i>F01G4.1</i>	<i>swsn-4</i>
<i>C08F8.2c.2</i>	
<i>F54E12.2</i>	
<i>T04A11.6</i>	<i>him-6</i>
<i>Y116A8C.13b</i>	
<i>T06A10.1c</i>	<i>mel-46</i>
<i>ZC317.1</i>	
<i>F26F12.7</i>	<i>let-418</i>
<i>T14G8.1</i>	<i>chd-3</i>
<i>Y65B4A.6b</i>	
<i>T04D1.4</i>	<i>chd-7</i>
<i>H27M09.1</i>	<i>sacy-1</i>
<i>Y23H5B.6c</i>	
<i>Y54E10A.9c.2</i>	<i>vbh-1</i>
<i>Y71H2AM.19b.2</i>	<i>laf-1</i>
<i>ZK512.2b</i>	
<i>Y55B1AL.3b</i>	<i>helq-1</i>
<i>Y66D12A.15</i>	<i>xpb-1</i>
<i>Y50D7A.2</i>	<i>xpd-1</i>
<i>Y50D7A.11</i>	
<i>F37A4.8</i>	<i>isw-1</i>

Appendix 3 Table showing all candidate helicases after filtering to be used for the helicase suppressor screen. Genes in bold have been attempted in the study.

Appendix 4



Appendix 4 Hatching assay of all genes in Table 4.1 (except *mtr-4* and *xpb-1*). Genes are presented in alphanumerical order. At least two samples have been counted for each RNAi except for *mog-5* and *isw-1* (as they have very few eggs). Error bars represent the standard deviation.

Appendix 5

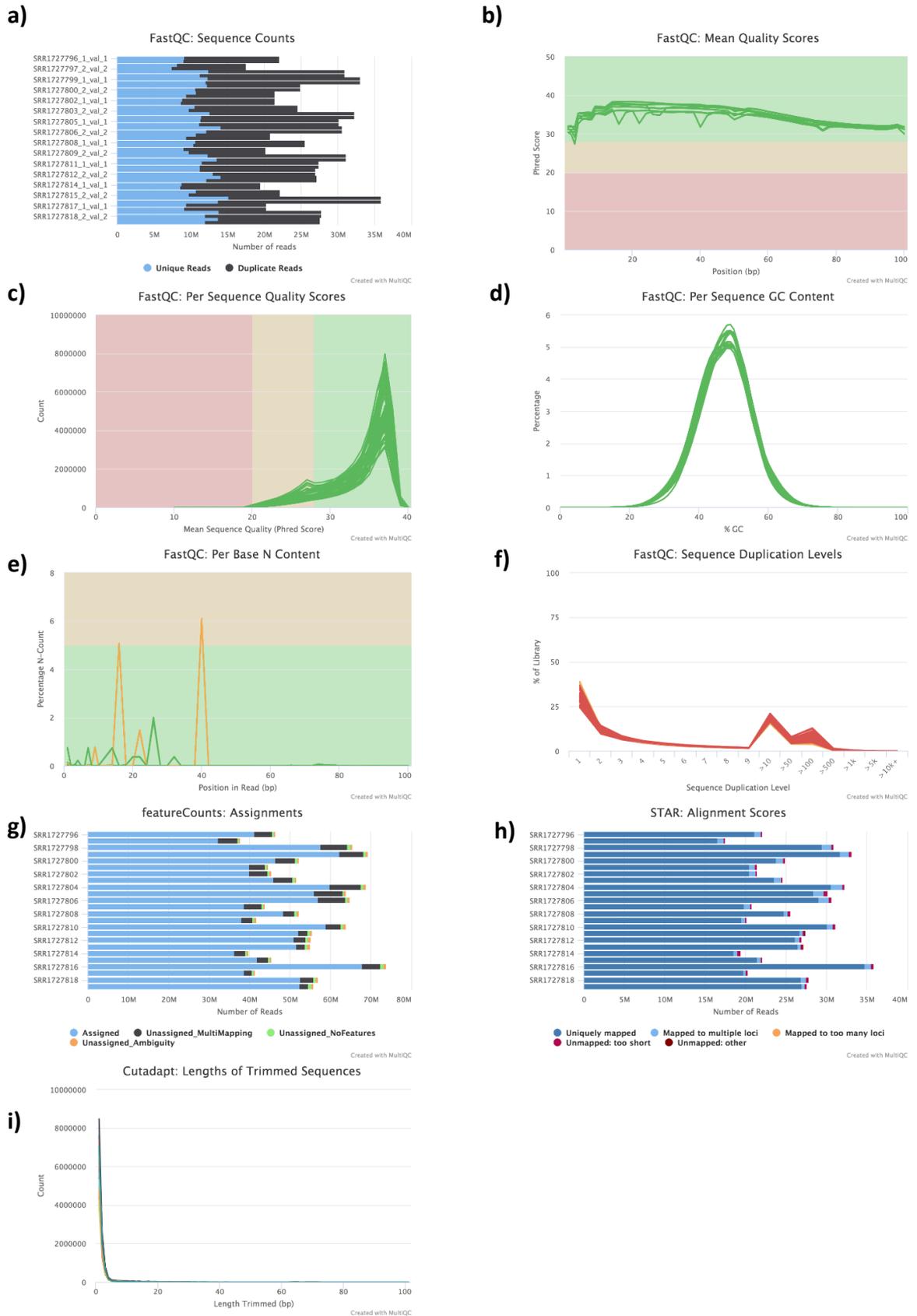
Source	Pathogen	Conditions	Type/Platform	Deposit
(Engelmann, et al., 2011)	<i>Serratia marcescens</i>	YA (25°C)	Microarray (Affymetrix)	GSE23275 & GSE23277
	<i>Enterococcus faecalis</i>	YA (25°C)	Microarray (Affymetrix)	GSE23273 & GSE23277
	<i>Photorhabdus luminescens</i>	YA (25°C)	Microarray (Affymetrix)	GSE23272 & GSE23277
ModENCODE	<i>Drechmeria coniospora</i>	<i>fer-15</i> L4 (25°C)	RNA-seq (Genome Analyzer II)	SRS150472 & SRS150966
	<i>Harposporium sp.</i>	<i>fer-15</i> L4 (25°C)	RNA-seq (Genome Analyzer II)	SRS008265 & SRS008266
	<i>Serratia marcescens</i>	YA (25°C)	RNA-seq (Genome Analyzer II)	SRS008267 & SRS008268
	<i>Haptocillium sphaerosporum</i>	<i>fer-15</i> L4 (25°C)	RNA-seq (Illumina HiSeq 2000)	SRS266743 & SRS266741
(Yang, et al., 2015)	<i>Bacillus thuringiensis</i>	L4 (20°C)	RNA-seq (Illumina HiSeq 2000)	GSE64401
(Pukkila-Worley, et al., 2011)	<i>Candida albicans</i>	YA (20°C)	Microarray (Affymetrix)	GSE27401
(Bakowski, et al., 2014)	<i>Nematocida parisii</i>	<i>fer-15(b26);fem-1(hc17)</i> L1 (25°C)	RNA-seq (Illumina HiSeq 2000)	SRP013019
(Head & Aballay, 2014)	<i>Salmonella enterica</i>	<i>fer-1(b232ts)</i> L1 (25°C)	Microarray (Agilent)	GSE54212
(Yuen & Ausubel, 2018)	<i>Enterococcus faecalis</i>	YA (20°C)	Microarray (Affymetrix)	GSE95636
	<i>Enterococcus faecium</i>	YA (20°C)	Microarray (Affymetrix)	GSE95636
	<i>Bacillus subtilis</i>	YA (20°C)	Microarray (Affymetrix)	GSE95636
(Bond, et al., 2014)	<i>Staphylococcus aureus</i>	L4 (18°C)	Microarray (Affymetrix)	GSE53732
	<i>Pseudomonas aeruginosa</i>	L4 (18°C)	Microarray (Affymetrix)	GSE53732
(Bolz, et al., 2010)	<i>Yersinia pestis</i>	L4 (25°C)	Microarray (Affymetrix)	GSE20053

(Estes, et al., 2010)	<i>Pseudomonas aeruginosa</i>	YA (25°C)	Microarray (Affymetrix)	GSE50513
(Irazoqui, et al., 2010)	<i>Staphylococcus aureus</i>	fer-15(b26)ts;fem-1(hc17) YA (25°C)	Microarray (Affymetrix)	GSE21819
(Miller, et al., 2015)	<i>Pseudomonas aeruginosa</i>	L4 (25°C)	Microarray (Affymetrix)	GSE72029
(O'Rourke, et al., 2006)	<i>Microbacterium nematophilum</i>	L2/L3 (25°C) in liquid	Microarray (Affymetrix)	E-MEXP-696
(Sahu, et al., 2012)	<i>Vibrio cholerae</i>	L2/L3 (22°C)	Microarray (Affymetrix)	GSE34026
(Troemel, et al., 2006)	<i>Pseudomonas aeruginosa</i>	YA (25°C)	Microarray (Affymetrix)	GSE5793
(K. Chen, et al., 2017)	<i>Orsay Virus</i>	L3 (20°C)	RNA-seq (Illumina HiSeq 2500)	SRP100798
	<i>Nematocida parisii</i>	L3 (20°C)	RNA-seq (Illumina HiSeq 2500)	SRP100798
(Osman, et al., 2018)	<i>Myzocytiopsis humicola</i>	L4 (25°C)	RNA-seq (Illumina HiSeq 2500)	GSE101647
(Tanguy, et al., 2017)	Orsay Virus	L4 (20°C)	RNA-seq (Illumina HiSeq 1500)	GSE95230
(Sarkies, et al., 2013)	Orsay Virus	Mixed-stage (23°C)	Microarray (Affymetrix)	GSE41056
(White & Herman, 2018)	<i>Stenotrophomonas maltophilia</i>	L4 (20°C)	microarray (NimbleGem)	GSE107272

Source	Condition	Type/Platform	Deposit
(Brunquell, et al., 2016)	L4, EV food, 33°C (30 min)	RNA-seq (Illumina HiSeq 2000)	PRJNA311958
(Li, et al., 2016)	L2, OP50 food, 34°C (30 min)	RNA-seq (Illumina HiSeq 2000)	GSE81520
(Haas, et al., 2018)	L4, OP50 food, 34°C (75 min) + Recovery period	RNA-seq (Illumina HiSeq 2500)	GSE122015
Laura Jones	L4, OP50 food, 35°C (60 min) + Recovery period	RNA-seq (NextSeq 500)	

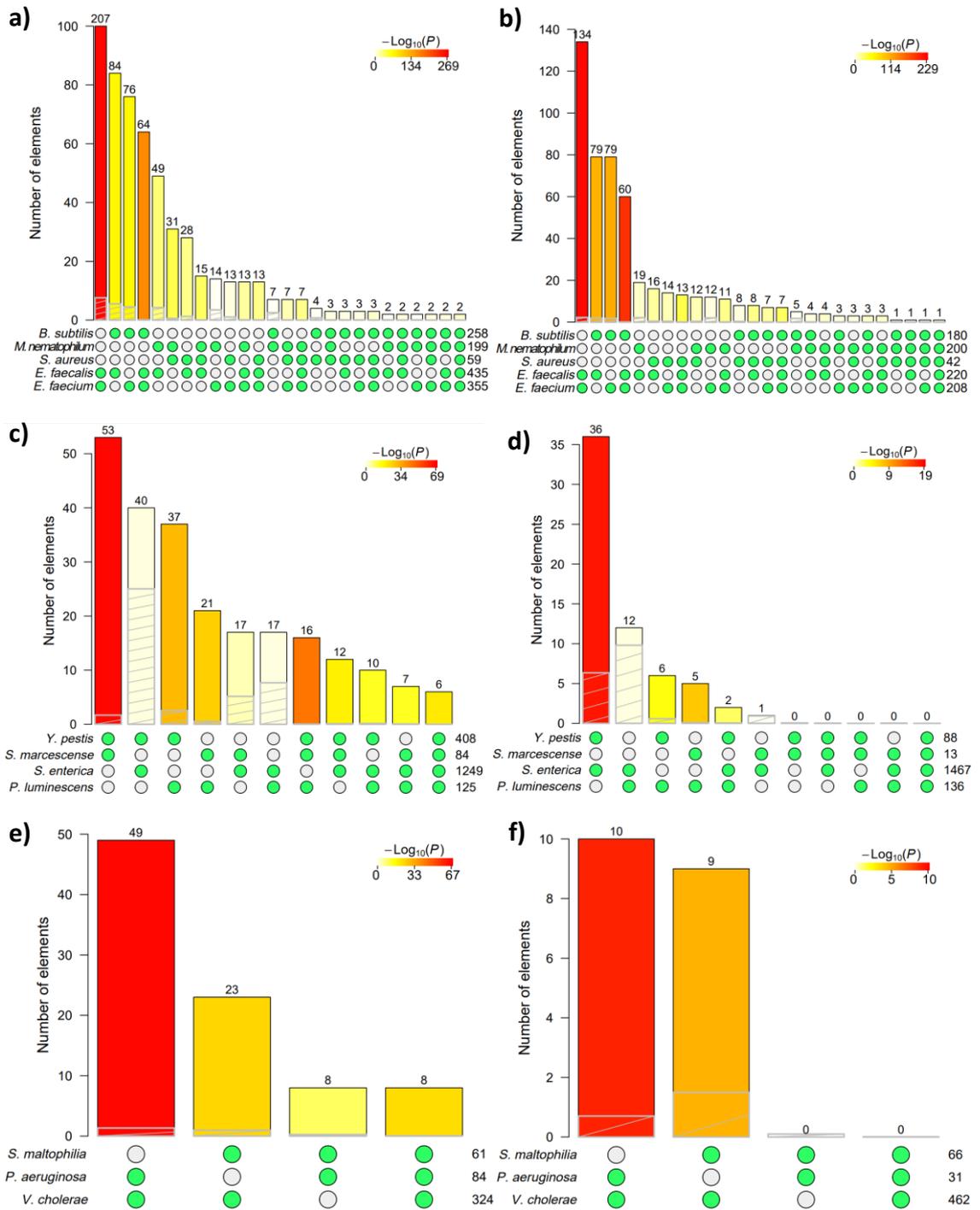
Appendix 5 Summary of the datasets used in this study, with details about the condition (pathogen), type and platform, source and the deposit.

Appendix 6



Appendix 6 Example of RNA-seq quality control. Here the data from Yang et al. (2015) is used. a) - f) FastQC reports. g) featureCounts Summary. h) STAR alignment score. i) Cutadapt report. MultiQC was used to summarize these reports.

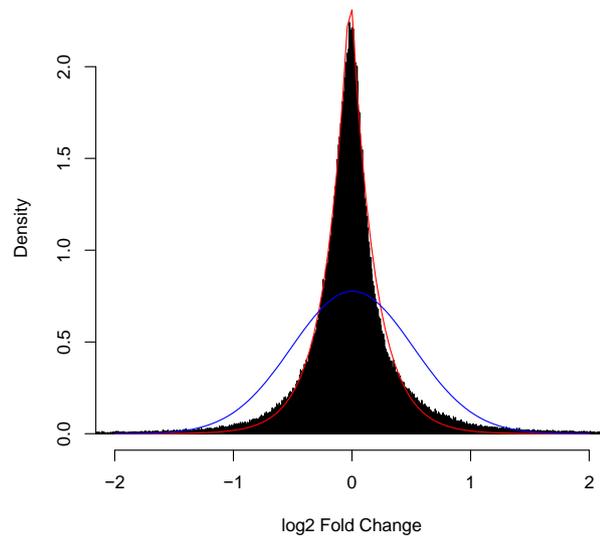
Appendix 7



Appendix 7 Hypergeometric distribution test for each of the Venn diagrams in Figure 7.4. a) & b) Pathogens from the Terrabacteria phylum: *S. aureus*, *M. nematophilum*, *E. subtilis*, *E. faecalis* and *E. faecium* up-regulated and down-regulated genes respectively. c) & d) Pathogens from the Enterobacteriales order: *Y. pestis*, *S. enterica*, *S. marcescens* and *P. luminescens* up-regulated and down-regulated genes, respectively. e) & f) Remaining pathogens *S. maltophilia*, *P. aeruginosa* and *V. cholerae* up-regulated and down-regulated genes respectively. The grey shading overlaying each column represents the expected overlap.

Appendix 8

The multi-layer filtering methods p-value was determined using hypergeometric testing. The probability distribution was modelled after Laplace distribution as this distribution fits better to the actual data compared to the normal distribution. This may be due to the `lfcshrink()` function.



Appendix 8 Graph showing the distribution of log2 Fold Change for each gene in each of the 25 pathogen response datasets that show a continuous distribution. The blue line shows the normal distribution with the data average (0.004) and standard deviation (0.51). The red line shows the Laplace distribution with μ (median) = -0.016 and b (mean deviation from the median) = 0.28

Taking the most stringent criterion, the fourth filtering criteria for list331, where the genes are filtered out when they do not have at least 0.1 log2 Fold change in at least 24 of the 29 datasets. From the Laplace distribution, the probability of drawing a measurement at random that is greater than $|0.1|$ is 0.56. Thus, the probability of drawing it 24 times out of 29 trials is:

$$\sum_{k=24}^{29} \binom{29}{k} \times 0.56^k \times (1 - 0.56)^{29-k} = 0.0023$$

For list585 (**Figure 7.8**), the filtering criterion was changed to: at least 0.1 log2 Fold change in at least 23 of the 25 datasets. This changes the probability to:

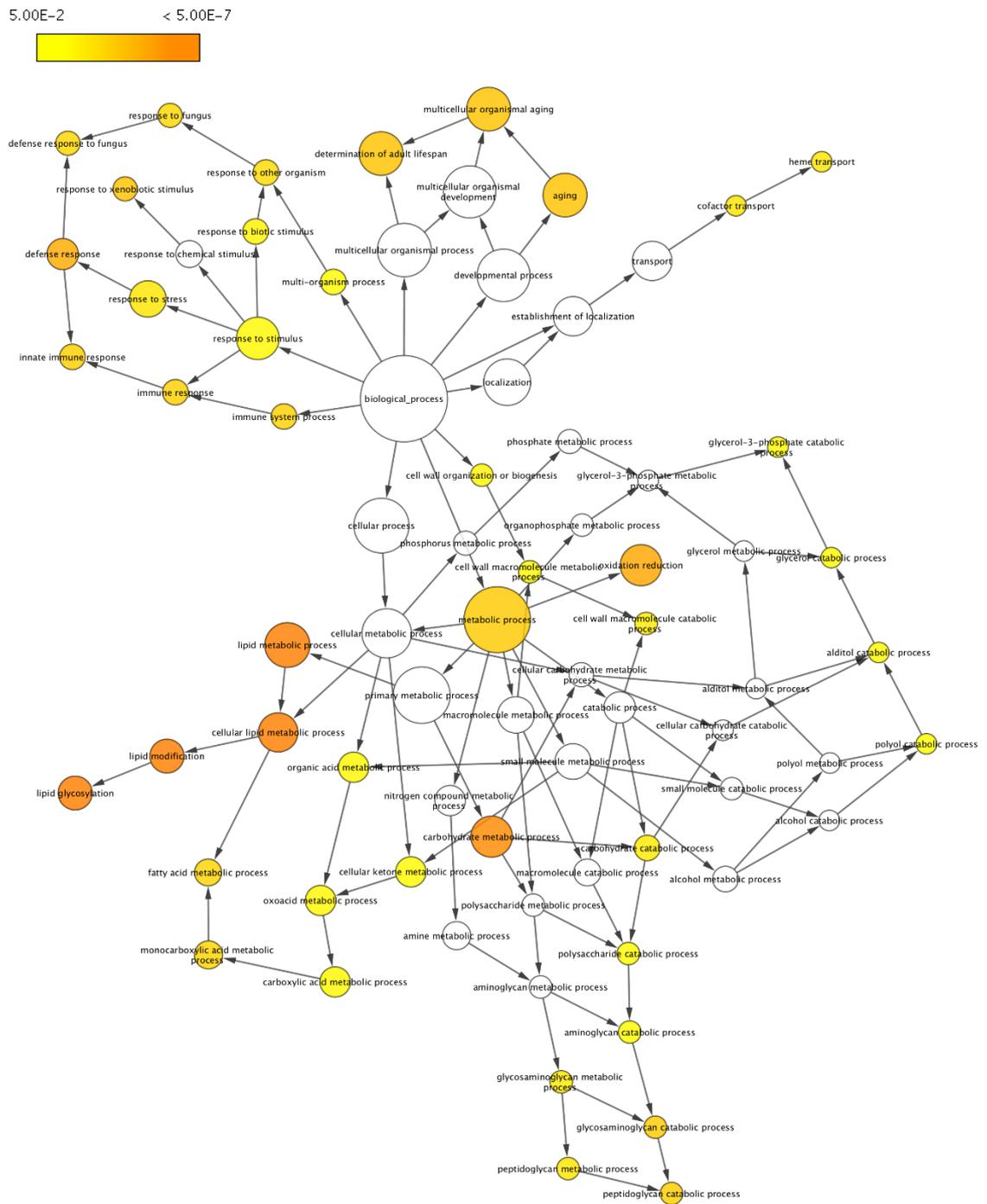
$$\sum_{k=23}^{25} \binom{25}{k} \times 0.56^k \times (1 - 0.56)^{25-k} = 0.0001$$

Appendix 9

List585					List383				
Legend					Legend				
GO:MF					GO:MF				
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
catalytic activity	GO:0003824	6.236e-09	8.797		glucuronosyltransferase activity	GO:0015020	4.782e-07	6.897	
lysozyme activity	GO:0003796	2.501e-05	4.601		peptidoglycan muralytic activity	GO:0061783	1.639e-06	5.797	
peptidoglycan muralytic activity	GO:0061783	2.501e-05	4.601		lysozyme activity	GO:0003796	1.639e-06	5.797	
oxidoreductase activity	GO:0016491	6.668e-05	4.377		oxidoreductase activity	GO:0016491	6.441e-06	5.601	
organic acid binding	GO:0043177	1.680e-04	3.877		carbohydrate binding	GO:0030246	2.350e-05	4.601	
carboxylic acid binding	GO:0031406	1.680e-04	3.877		hydrolase activity, acting on glycosyl bonds	GO:0016798	2.383e-05	4.597	
cofactor binding	GO:0048037	2.010e-04	3.697		catalytic activity	GO:0003824	2.931e-05	4.437	
hydrolase activity, acting on glycosyl bonds	GO:0016798	3.706e-03	2.727		hydrolase activity, hydrolyzing O-glycosyl compounds	GO:0004553	6.442e-05	4.397	
glucuronosyltransferase activity	GO:0015020	4.203e-03	2.577		iron ion binding	GO:0005506	1.069e-04	3.997	
monocarboxylic acid binding	GO:0033293	5.073e-03	2.497		UDP-glycosyltransferase activity	GO:0008184	1.548e-04	3.917	
coenzyme binding	GO:0005692	6.638e-03	2.377		tetraspyrrole binding	GO:0046906	1.204e-03	2.917	
hydroxymethyl-, formyl-, and related transferase activity	GO:0016742	2.417e-02	1.617		oxidoreductase activity, acting on paired donors, with...	GO:0016705	1.464e-03	2.797	
monooxygenase activity	GO:0004497	3.061e-02	1.517		monooxygenase activity	GO:0004497	3.184e-03	2.497	
carbohydrate binding	GO:0030246	3.428e-02	1.457		cofactor binding	GO:0048037	5.401e-03	2.257	
fatty acid binding	GO:0005504	4.838e-02	1.327		heme binding	GO:0020037	5.996e-03	2.217	
					steroid hydroxylase activity	GO:0008395	1.909e-02	1.717	
					oxidoreductase activity, acting on paired donors, with...	GO:0016712	2.203e-02	1.657	
					transferase activity, transferring hexosyl groups	GO:0016798	2.545e-02	1.577	
					transferase activity, transferring glycosyl groups	GO:0016757	3.104e-02	1.507	
GO:BP					GO:BP				
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
defense response	GO:0006952	2.157e-10	10.197		defense response	GO:0006952	2.027e-10	10.197	
immune system process	GO:0002376	4.210e-10	10.197		innate immune response	GO:0045087	3.098e-10	10.197	
innate immune response	GO:0045087	2.267e-10	10.197		immune system process	GO:0002376	3.674e-10	10.197	
immune response	GO:0006955	2.821e-10	10.197		immune response	GO:0006955	3.826e-10	10.197	
response to stress	GO:0006950	1.009e-10	10.197		response to stress	GO:0006950	2.582e-10	10.197	
defense response to other organism	GO:0098542	1.685e-10	10.197		response to other organism	GO:0051707	5.162e-10	10.197	
response to biotic stimulus	GO:0009607	2.830e-10	10.197		response to biotic stimulus	GO:0009607	5.162e-10	10.197	
response to other organism	GO:0051707	2.830e-10	10.197		response to external biotic stimulus	GO:0043207	5.162e-10	10.197	
response to external biotic stimulus	GO:0043207	2.830e-10	10.197		defense response to other organism	GO:0098542	4.074e-10	10.197	
defense response to bacterium	GO:0042742	2.850e-10	10.197		defense response to bacterium	GO:0042742	8.034e-10	10.197	
response to bacterium	GO:0009617	1.038e-10	10.197		response to fungus	GO:0009620	6.531e-10	10.197	
defense response to Gram-positive bacterium	GO:0009113	6.780e-10	10.197		oxidation-reduction process	GO:0005114	7.522e-10	10.197	
defense response to Gram-negative bacterium	GO:0050829	7.234e-10	10.197		interspecies interaction between organisms	GO:0044419	1.576e-09	10.197	
organic acid metabolic process	GO:0006082	1.924e-10	10.197		peptidoglycan catabolic process	GO:0009223	3.717e-09	10.197	
carboxylic acid metabolic process	GO:0019752	3.959e-10	10.197		peptidoglycan metabolic process	GO:0002070	3.717e-09	10.197	
lipid catabolic process	GO:0016042	4.744e-10	10.197		defense response to fungus	GO:0050832	4.383e-09	10.197	
oxidation-reduction process	GO:0005114	5.041e-10	10.197						
nucleobase biosynthetic process	GO:0046112	7.811e-10	10.197		GO:CC				
purine nucleobase metabolic process	GO:0006144	1.412e-09	10.197		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
oxoacid metabolic process	GO:0043436	1.565e-09	10.197		membrane raft	GO:0045121	4.176e-10	10.197	
fatty acid oxidation	GO:0019395	3.465e-09	10.197		membrane microdomain	GO:0098857	4.176e-10	10.197	
lipid oxidation	GO:0034440	3.465e-09	10.197		membrane region	GO:0098589	4.176e-10	10.197	
cellular lipid catabolic process	GO:0044242	3.851e-09	10.197		whole membrane	GO:0098805	1.432e-09	10.197	
multi-organism process	GO:0051704	4.665e-09	10.197		lysosome	GO:0005764	1.419e-09	10.197	
small molecule metabolic process	GO:0044281	5.122e-09	10.197		lytic vacuole	GO:0000323	1.893e-09	10.197	
response to external stimulus	GO:0009605	5.126e-09	10.197		vacuole	GO:0005773	1.884e-09	10.197	
fatty acid beta-oxidation	GO:0006635	8.560e-09	10.197		extracellular region	GO:0005576	1.611e-09	10.197	
fatty acid metabolic process	GO:0006831	2.214e-09	10.197						
lipid modification	GO:0030258	2.617e-09	10.197		KEGG				
fatty acid catabolic process	GO:0009602	2.829e-09	10.197		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
IMP biosynthetic process	GO:0006188	3.249e-09	10.197		Lysosome	KEGG:04142	1.869e-08	8.227	
purine nucleoside monophosphate biosynthetic process	GO:0009127	3.293e-09	10.197		Fatty acid degradation	KEGG:00071	4.490e-08	7.847	
purine ribonucleoside monophosphate biosynthetic pro...	GO:0009168	3.293e-09	10.197		Metabolic pathways	KEGG:01100	1.181e-04	3.901	
organic acid catabolic process	GO:0016054	4.104e-09	10.197		Fatty acid metabolism	KEGG:01212	2.224e-04	3.647	
carboxylic acid catabolic process	GO:0046395	4.104e-09	10.197		Drug metabolism - cytochrome P450	KEGG:00982	2.941e-04	3.527	
catabolic process	GO:0009056	4.275e-09	10.197		Butanoate metabolism	KEGG:00650	4.595e-04	3.341	
					Ascorbate and aldarate metabolism	KEGG:00053	7.821e-04	3.107	
					Valine, leucine and isoleucine degradation	KEGG:00280	8.057e-04	3.097	
					Retinol metabolism	KEGG:00830	9.982e-04	2.901	
					Lysine degradation	KEGG:00310	1.403e-03	2.851	
					Metabolism of xenobiotics by cytochrome P450	KEGG:00980	1.635e-03	2.791	
					Tryptophan metabolism	KEGG:00380	7.723e-03	2.107	
					Porphyryr and chlorophyll metabolism	KEGG:00860	1.346e-02	1.877	
					Pentose and glucuronate interconversions	KEGG:00040	1.623e-02	1.791	
					Riboflavin metabolism	KEGG:00740	1.795e-02	1.741	
					Fatty acid elongation	KEGG:00062	1.939e-02	1.701	
					One carbon pool by folate	KEGG:00670	3.206e-02	1.491	
WP					WP				
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
Metapathway UDP-glucuronosyltransferases	WP:WP1431	1.561e-04	3.807		Metapathway UDP-glucuronosyltransferases	WP:WP1431	6.064e-09	8.407	
Valine, leucine and isoleucine degradation	WP:WP1451	9.457e-03	2.021		Regulation of immune response in the intestine	WP:WP2234	3.070e-02	1.507	
TF					TF				
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
Factor: elt-3; motif: TCTTATCA; match class: 0	TF:M07154_0	6.058e-20	19.707		Factor: elt-3; motif: TCTTATCA; match class: 0	TF:M07154_0	1.463e-19	19.707	
Factor: elt-3; motif: TCTTATCA; match class: 1	TF:M07154_1	5.067e-04	3.801		Factor: elt-3; motif: TCTTATCA; match class: 1	TF:M07154_1	2.672e-04	3.697	
HP					HP				
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$		Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	
Hyperhomocystinemia	HP:0002190	1.872e-02	1.741		Hyperhomocystinemia	HP:0002190	1.872e-02	1.741	
Homocystinuria	HP:0002156	4.425e-02	1.351		Homocystinuria	HP:0002156	4.425e-02	1.351	

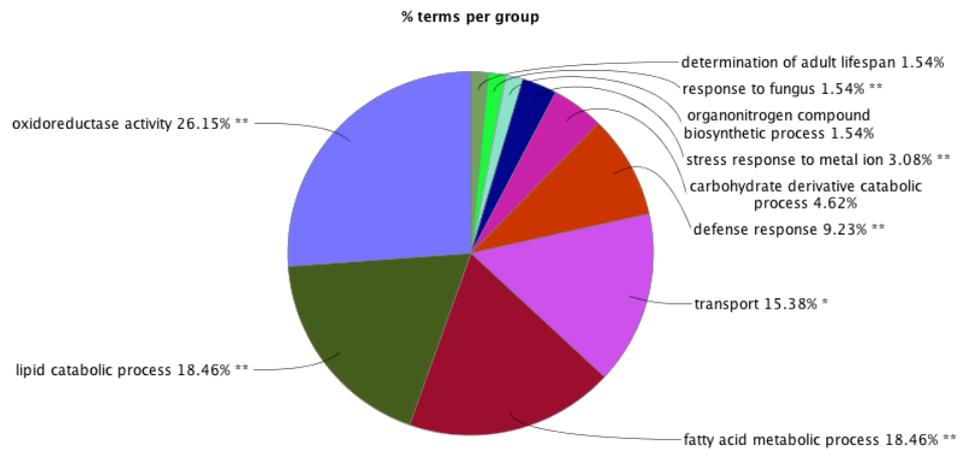
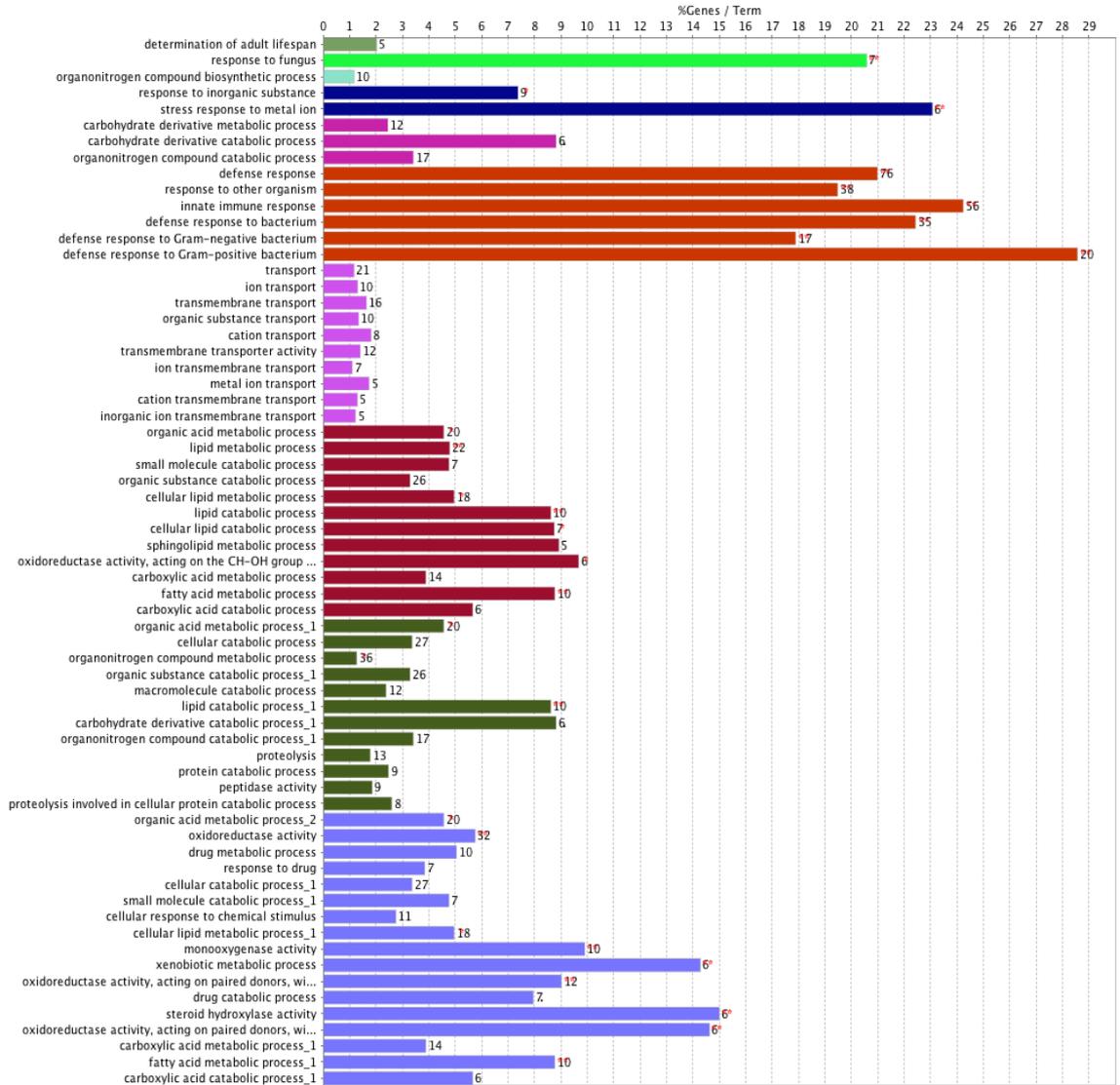
Appendix 9 g: Profiler analysis results for the genes in list585 and list383

Appendix 10



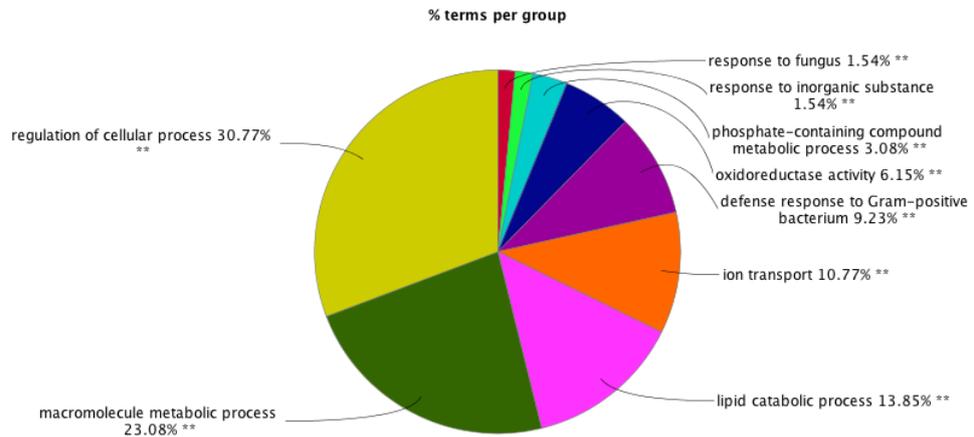
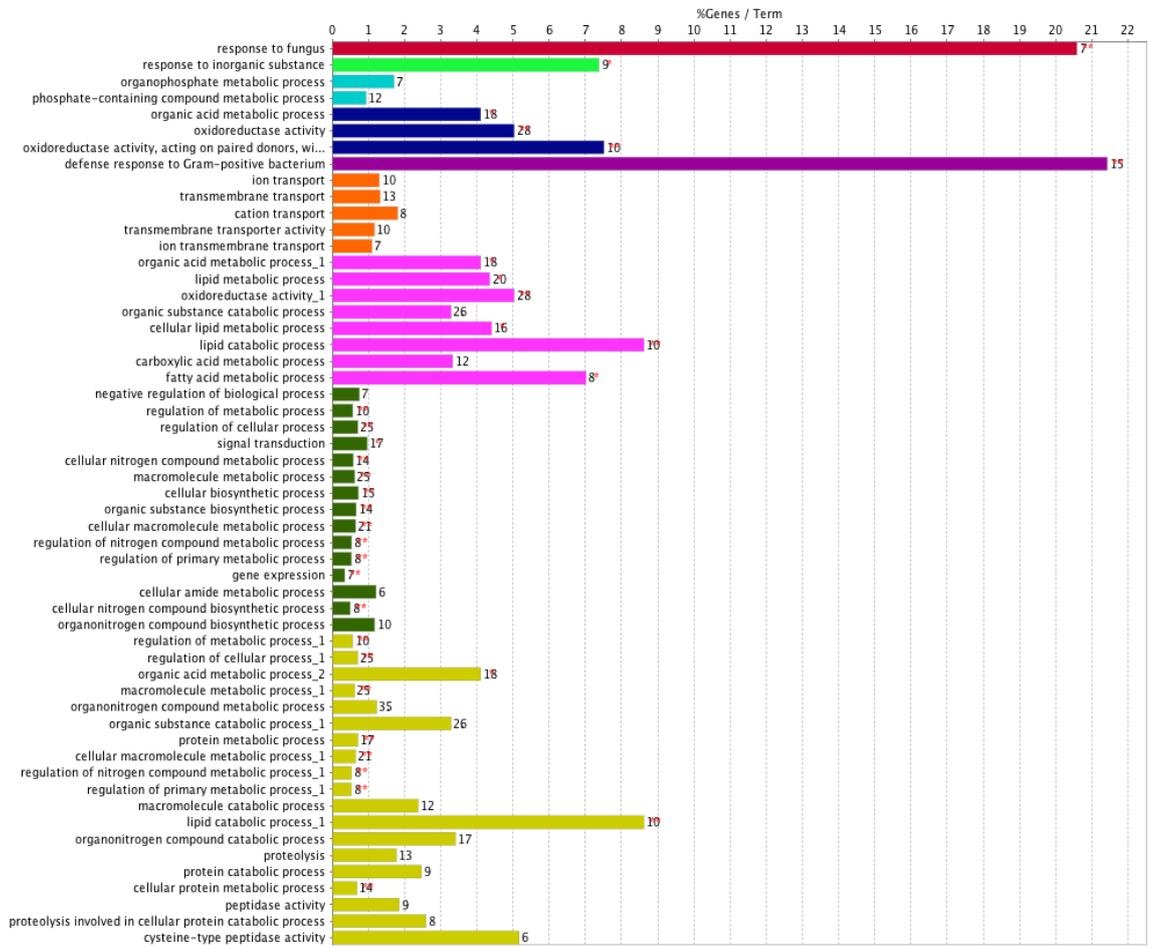
Appendix 10 BinGO network of the genes in list383. This network shows connections between various Biological Process GO terms. The colouring in each node corresponds to the p-value (legend).

Appendix 11



Appendix 11 ClueGO detailed report for Figure 7.14 showing the exact number of genes (and % from the input) in each term and the proportion of these terms relative to each other. Single asterisk (*) denotes p-value < 0.05 and double asterisk (**) denotes p-value < 0.01

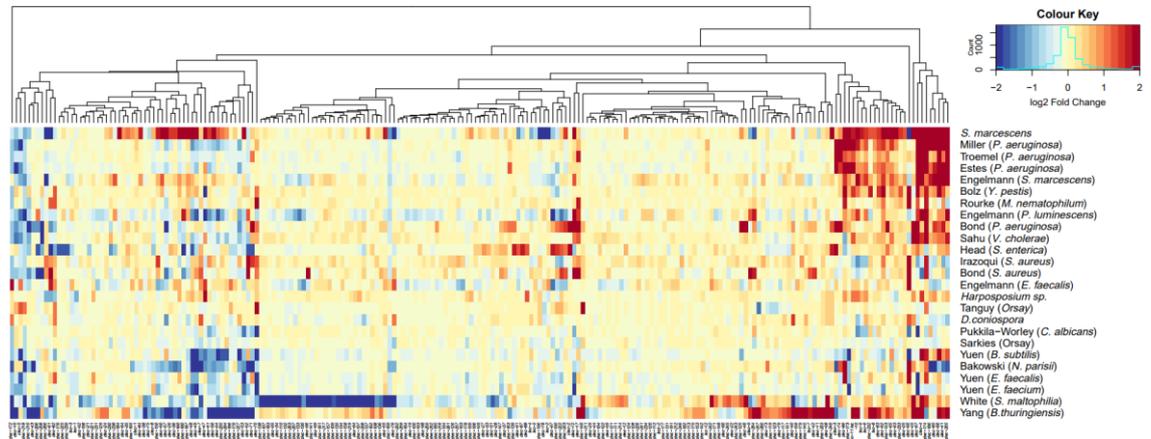
Appendix 12



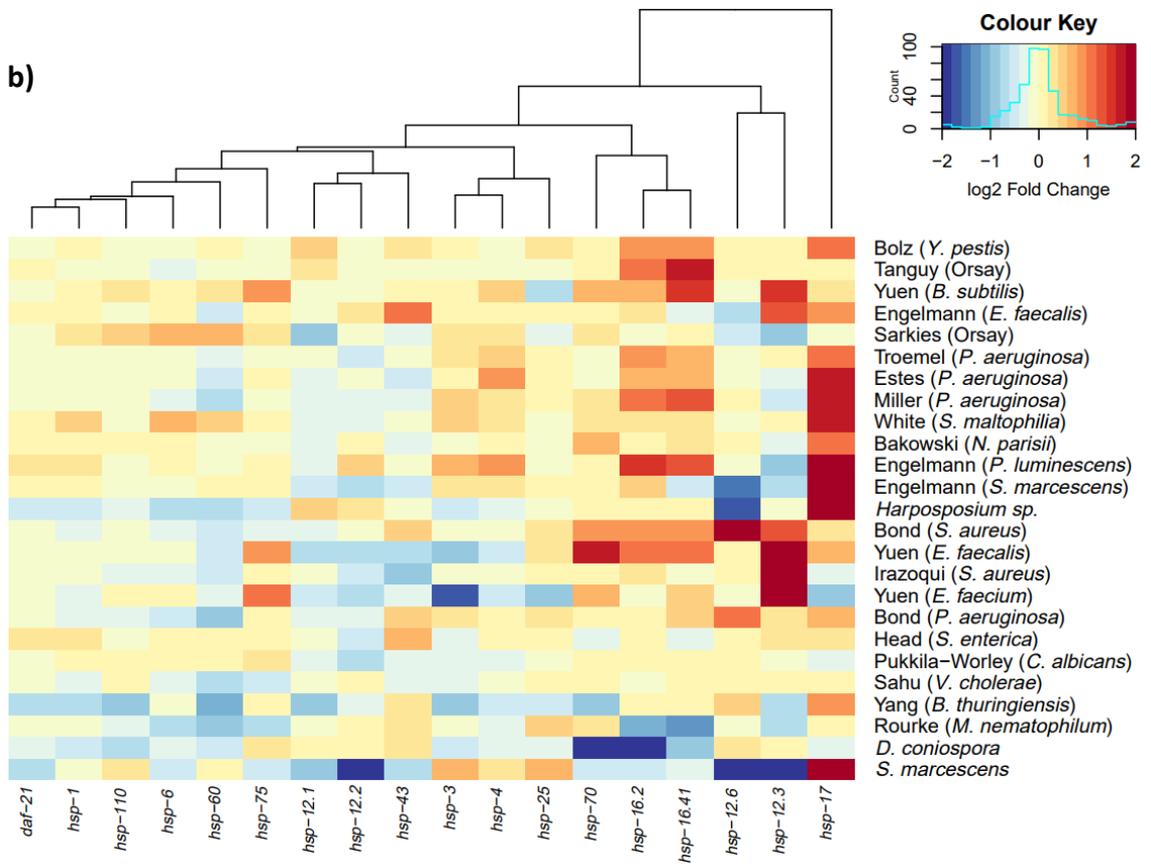
Appendix 12 ClueGO detailed report for Figure 7.15 showing the exact number of genes (and % from the input) in each term and the proportion of the terms associated to the analysis of the 3 main clusters from list383. Single asterisk (*) denotes p-value < 0.05 and double asterisk (**) denotes p-value < 0.01

Appendix 13

a)



b)



Appendix 13 Heatmap of the pathogen dataset for various protein families. a) heatmap of lectins (LEC and CLEC) proteins. b) heatmap of heat shock proteins.

Appendix 15

a)

Motif	PWM	Hits against known TFBS databases
family_1		
family_2		ELT-3,
family_5		
family_3		

b)

#	IUPAC	PWM	reverse Comp.	motif AvRec	frac. occurrence
1	AACTGATAAGAA			0.564	0.610
3	GAKAAGAARA			0.337	0.784
2	GGGMGRAGG			0.326	0.617
4	CCTCCWCC			0.659	0.246

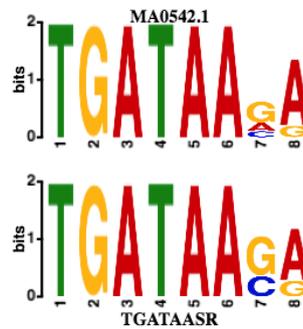
name	e-value	query motif	database PWM	reverse Comp.
elt-3	2.0E-05			
GAT1	1.5E-04			
Gata1	2.8E-04			
Gata4	4.5E-04			
GATA6	1.1E-03			

c)

	Motif ?	Logo ?	RC Logo ?	E-value ?	Unersased E-value ?
1.	AWTTYCAG			2.6e-029	2.6e-029
2.	TGATAASR			2.3e-025	2.3e-025
3.	CWCCDCC			2.3e-023	2.3e-023
4.	CTGNAAA			2.4e-012	2.2e-015
5.	TACBGTA			3.9e-012	4.6e-010

Summary [?](#)

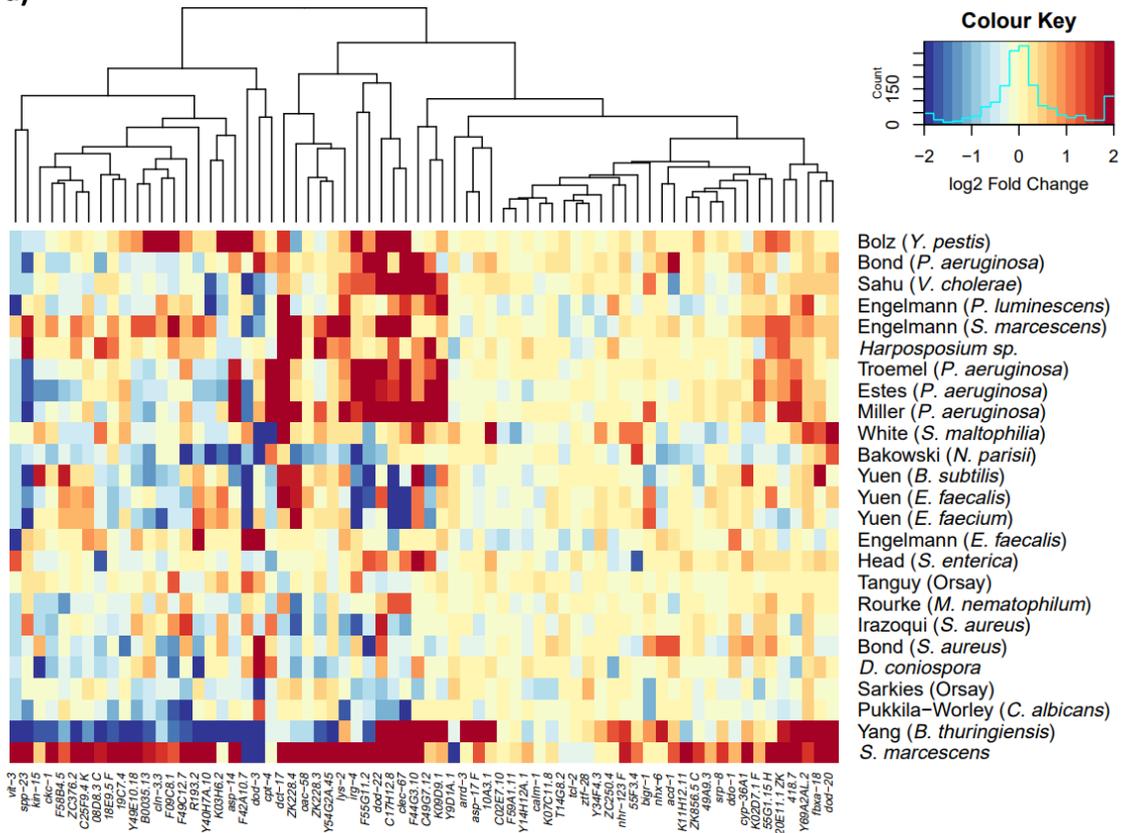
Name	MA0542.1 (elt-3)
Database	JASPAR2018_CORE_nematodes_non-redundant
p-value	1.52e-06
E-value	3.95e-05
q-value	7.89e-05
Overlap	8
Offset	0
Orientation	Reverse Complement
	Show logo download options

Optimal Alignment [?](#)

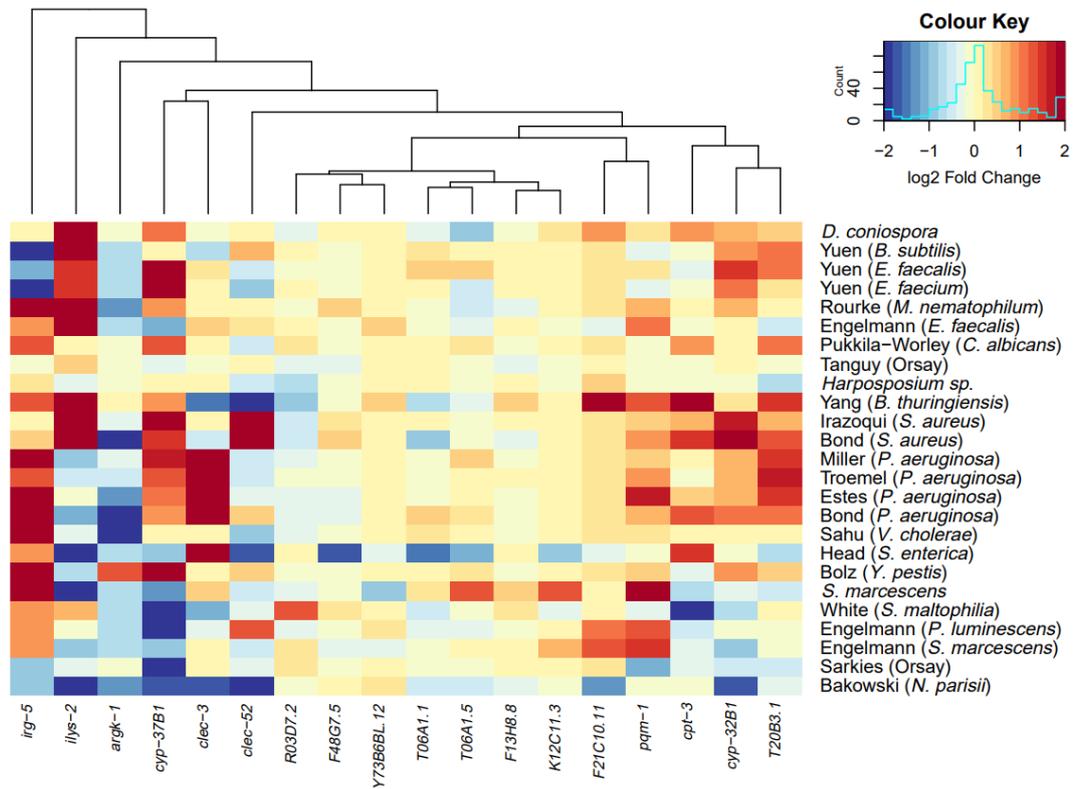
Appendix 15 *de novo* motif discovery on the 383 pathogen responsive genes (list383) using various software. a) Trawler. b) BaMM motif and identification of the best matching transcription factor to the top enriched motif. c) DREME & Tomtom. Tomtom was used to find known transcription factors that best matches the enriched motif from DREME.

Appendix 16

a)



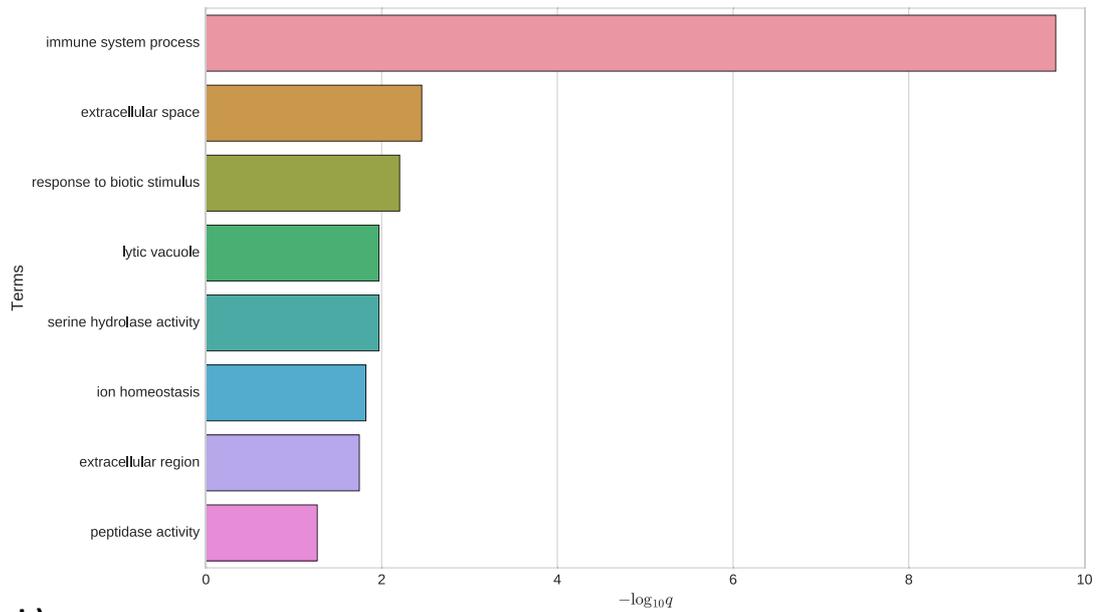
b)



Appendix 16 Heatmap of the pathogen datasets showing the list of genes that are differentially expressed in the *pqm-1(ok485)* mutant at 35°C. a) heatmap of genes that are up-regulated in the *pqm-1(ok485)* mutant at 35°C. b) heatmap of genes that are down-regulated in the *pqm-1(ok485)* mutant at 35°C.

Appendix 17

a)

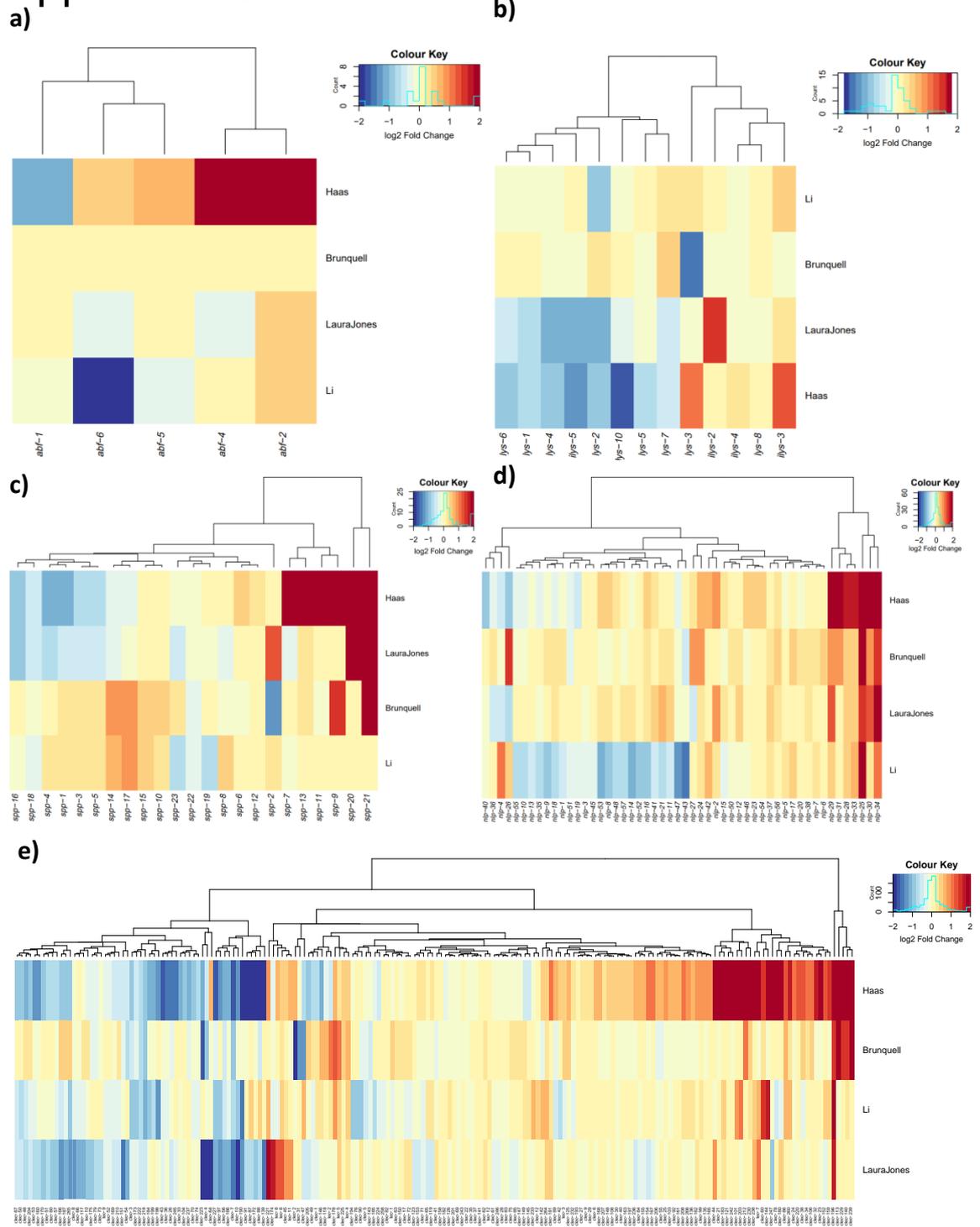


b)

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
innate immune response	GO:0045087	9.303×10^{-7}	~6.9
immune response	GO:0006955	1.345×10^{-6}	~6.6
immune system process	GO:0002376	1.428×10^{-6}	~6.5
defense response to other organism	GO:0098542	2.328×10^{-5}	~4.6
response to biotic stimulus	GO:0009607	2.490×10^{-5}	~4.6
response to external biotic stimulus	GO:0043207	2.490×10^{-5}	~4.6
response to other organism	GO:0051707	2.490×10^{-5}	~4.6
defense response	GO:0006952	2.781×10^{-5}	~4.6
response to external stimulus	GO:0009605	4.065×10^{-3}	~2.4
multi-organism process	GO:0051704	5.401×10^{-3}	~2.3
defense response to Gram-negative bacterium	GO:0050829	1.097×10^{-2}	~1.9

Appendix 17 Gene Enrichment analysis on the 52 PQM-1 target genes that are up-regulated at 35°C in *pqm-1(mu86)* mutants (**Figure 7.22b** 1st column). a) Wormbase Enrichment Analysis showing only the Gene Ontology Enrichment Analysis Results. b) g:Profiler analysis showing the GO:Biological Processes results.

Appendix 18



Appendix 18 Heatmap of the heat shock datasets for specific protein families. a) antibacterial factors (ABF). b) lysozyme (LYS and ILYS). c) sapsin-like protein (SPP). d) neuropeptide-like protein (NPL). e) lectins (LEC and CLEC).

Appendix 19

	Sum of log2FC	Mean	absSum	absMean	sds	SE
C25F9.11	45.784424	1.57877324	46.227197	1.5940413	1.3257362	0.24618304
T24C4.4	40.616012	1.40055212	48.498806	1.6723726	1.9807224	0.36781091
tts-1	34.353489	1.18460307	43.566285	1.5022857	1.6498256	0.30636492
zip-10	29.584740	1.02016344	33.001995	1.1379998	1.2794279	0.23758379
F19B2.5	28.846898	0.99472063	39.080043	1.3475877	1.3958376	0.25920053
T19D12.4	27.806680	0.95885102	37.234019	1.2839317	1.4963004	0.27785600
F08G2.5	27.733223	0.95631803	32.849293	1.1327343	1.6242647	0.30161837
oac-14	26.174176	0.90255779	30.840440	1.0634635	0.9763452	0.18130275
Y94H6A.10	26.005087	0.89672714	27.894680	0.9618855	0.6899931	0.12812851
cup-16	25.671244	0.88521531	31.633950	1.0908259	1.0260872	0.19053962
F53B2.8	24.174010	0.83358655	27.028293	0.9320101	0.8672203	0.16103877
cnc-4	22.310150	0.76931553	40.447952	1.3947570	1.9760644	0.36694594
nlp-34	21.153848	0.72944305	30.608272	1.0554577	1.6390874	0.30437088
T01D3.6	20.341170	0.70141966	32.257049	1.1123120	1.3342464	0.24776334
Y75B8A.28	20.286745	0.69954292	31.497336	1.0861150	1.2757003	0.23689160
C25H3.10	19.584898	0.67534132	30.707265	1.0588712	1.0990706	0.20409229
clec-265	18.747998	0.64648270	36.966130	1.2746941	1.6873013	0.31332398
C47E8.11	17.956605	0.61919326	20.472491	0.7059480	1.5014432	0.27881099
ugt-44	17.444407	0.60153127	26.325105	0.9077623	1.1614556	0.21567689
lys-2	13.427486	0.46301677	29.864739	1.0298186	1.4614586	0.27138604
nlp-29	12.633170	0.43562657	25.555243	0.8812153	1.1646560	0.21627120
F15E6.3	9.844870	0.33947827	26.898748	0.9275430	1.3036619	0.24208393
col-135	8.758611	0.30202106	16.342817	0.5635454	0.6125446	0.11374667
T28F4.5	7.208626	0.24857332	17.945815	0.6188212	0.7365321	0.13677058
cpt-4	7.107400	0.24508275	24.776997	0.8543792	1.1635343	0.21606289
comt-3	7.080165	0.24414362	14.838867	0.5116851	0.5855541	0.10873467
Y54G2A.11	7.044682	0.24292008	16.941277	0.5841820	0.7435549	0.13807469
col-156	2.878239	0.09924962	21.363250	0.7366638	1.1383455	0.21138545
cpi-1	1.902796	0.06561365	16.576936	0.5716185	0.7200633	0.13371240
K10C2.1	-1.199657	-0.04136749	21.185556	0.7305364	0.9993286	0.18557066
col-8	-1.215938	-0.04192891	20.977422	0.7233594	0.9941819	0.18461495
C53A3.2	-1.540499	-0.05312065	11.990901	0.4134793	0.6355838	0.11802494
gpdh-1	-2.655315	-0.09156257	27.337682	0.9426787	1.3291574	0.24681833
gly-8	-3.524475	-0.12153361	19.186813	0.6616142	0.8164611	0.15161302
hmit-1.1	-3.543449	-0.12218791	22.795013	0.7860349	0.9772416	0.18146921
F40G12.11	-4.498811	-0.15513141	17.333019	0.5976903	0.8591688	0.15954363
R193.2	-5.590966	-0.19279192	32.597381	1.1240476	1.5977620	0.29669696
H17B01.2	-5.732420	-0.19766966	13.105948	0.4519293	0.6625890	0.12303969
T22F3.3	-6.609824	-0.22792495	8.100469	0.2793265	0.3183935	0.05912418
F58G6.3	-7.367309	-0.25404513	19.799331	0.6827356	1.1518617	0.21389535
F13H8.3	-9.614563	-0.33153664	18.925419	0.6526007	0.7786028	0.14458291
cpr-1	-9.723780	-0.33530275	18.974107	0.6542796	0.7627858	0.14164577
elo-2	-10.485013	-0.36155218	25.244008	0.8704830	1.0156735	0.18860584
T22B7.7	-11.015784	-0.37985461	27.123148	0.9352810	1.1638954	0.21612994
ugt-63	-18.304823	-0.63120081	32.783943	1.1304808	1.2974712	0.24093435
Y48E1B.8	-18.377662	-0.63371247	42.103165	1.4518333	1.9100507	0.35468751
clec-227	-18.448116	-0.63614192	22.269012	0.7678970	1.0075274	0.18709314
clec-218	-24.083416	-0.83046261	26.663764	0.9194401	0.7109114	0.13201293
nhr-114	-28.384504	-0.97877601	38.050936	1.3121012	1.2045933	0.22368736
asp-13	-31.970063	-1.10241597	38.837148	1.3392120	1.4577935	0.27070546

Appendix 19 Table showing all 50 genes that are consistently differentially expressed in both the pathogen and heat shock datasets. Statistical analysis was performed to show the expression pattern of the gene across all datasets.

Appendix 20

List331

<i>lys-4</i>	<i>asp-12</i>	<i>mrp-3</i>	<i>W07G4.5</i>	<i>F40F8.5</i>	<i>F53A9.1</i>	<i>droe-4</i>
<i>clec-56</i>	<i>ugt-18</i>	<i>Y38H6C.19</i>	<i>K02D7.1</i>	<i>abt-4</i>	<i>oac-31</i>	<i>Y46H3A.5</i>
<i>ilys-5</i>	<i>irg-4</i>	<i>K01C8.1</i>	<i>F45D3.3</i>	<i>col-162</i>	<i>H02F09.3</i>	<i>skr-5</i>
<i>pud-4</i>	<i>ech-9</i>	<i>Y32F6B.1</i>	<i>egl-15</i>	<i>Y54G2A.49</i>	<i>F22H10.2</i>	<i>F09F9.3</i>
<i>spp-4</i>	<i>srh-237</i>	<i>C02B4.4</i>	<i>ptr-23</i>	<i>asp-6</i>	<i>K11H12.4</i>	<i>C54D10.13</i>
<i>pud-3</i>	<i>T05E12.6</i>	<i>T23E7.6</i>	<i>C10C5.5</i>	<i>gpx-1</i>	<i>C25F9.11</i>	<i>Y102A11A.9</i>
<i>vit-3</i>	<i>C05D12.3</i>	<i>fjpr-22</i>	<i>ZK550.6</i>	<i>Y51A2D.13</i>	<i>B0024.4</i>	<i>Y47D3B.1</i>
<i>papl-1</i>	<i>folt-2</i>	<i>gln-3</i>	<i>tyr-1</i>	<i>H34I24.2</i>	<i>T27C5.8</i>	<i>C18E9.9</i>
<i>C46F2.1</i>	<i>clec-172</i>	<i>cpr-1</i>	<i>nlp-29</i>	<i>lact-1</i>	<i>sri-36</i>	<i>F46B3.1</i>
<i>fkf-3</i>	<i>vit-1</i>	<i>asah-1</i>	<i>cnc-4</i>	<i>R07C12.2</i>	<i>F49H6.13</i>	<i>F21C10.11</i>
<i>DH11.2</i>	<i>nspa-9</i>	<i>T12B5.15</i>	<i>ugt-54</i>	<i>C02F5.14</i>	<i>C54C8.12</i>	<i>ZK1290.14</i>
<i>metr-1</i>	<i>srh-216</i>	<i>Y43C5A.7</i>	<i>C06B3.6</i>	<i>C05C12.4</i>	<i>ins-11</i>	<i>lec-11</i>
<i>Y51H7C.1</i>	<i>clec-257</i>	<i>C40A11.4</i>	<i>C32F10.4</i>	<i>Y73B6BL.31</i>	<i>F53A9.6</i>	<i>K01F9.2</i>
<i>clec-170</i>	<i>C47F8.6</i>	<i>W08F4.11</i>	<i>F45D3.4</i>	<i>F48C5.2</i>	<i>C50F4.9</i>	<i>cyp-33C8</i>
<i>math-45</i>	<i>R02F2.8</i>	<i>Y38H6A.1</i>	<i>ZK970.7</i>	<i>nspe-3</i>	<i>ZC196.1</i>	<i>C49A9.9</i>
<i>F10F2.2</i>	<i>Y69E1A.5</i>	<i>Y20C6A.4</i>	<i>sdz-24</i>	<i>T11F8.2</i>	<i>F39G3.4</i>	<i>C38H2.3</i>
<i>lipl-5</i>	<i>C34E11.4</i>	<i>gst-24</i>	<i>E01G6.3</i>	<i>F10A3.17</i>	<i>Y46G5A.20</i>	<i>C47E8.11</i>
<i>alh-12</i>	<i>F46G11.6</i>	<i>T13F3.8</i>	<i>tag-10</i>	<i>clec-84</i>	<i>C04G6.5</i>	<i>Y60A3A.24</i>
<i>E04F6.15</i>	<i>ZK622.4</i>	<i>scl-21</i>	<i>pcp-3</i>	<i>F57B1.9</i>	<i>zip-10</i>	<i>slc-17.9</i>
<i>Y49E10.18</i>	<i>T06D8.3</i>	<i>C29G2.2</i>	<i>gsto-1</i>	<i>W02G9.4</i>	<i>kreg-1</i>	<i>Y46H3A.4</i>
<i>ugt-5</i>	<i>T01C3.11</i>	<i>clec-5</i>	<i>mct-4</i>	<i>K10C2.1</i>	<i>W02A2.9</i>	<i>nhr-6</i>
<i>best-7</i>	<i>C04E6.13</i>	<i>srh-214</i>	<i>mel-32</i>	<i>Y54G2A.45</i>	<i>srw-86</i>	<i>C41G7.8</i>
<i>D1086.3</i>	<i>H32K21.1</i>	<i>scl-15</i>	<i>ZK899.2</i>	<i>spp-2</i>	<i>Y58A7A.5</i>	<i>F10E9.12</i>
<i>T15B7.1</i>	<i>F48D6.4</i>	<i>F42H10.6</i>	<i>ZK512.7</i>	<i>ctsa-2</i>	<i>F16H6.10</i>	<i>clec-179</i>
<i>C23H5.8</i>	<i>F25B4.8</i>	<i>mtl-2</i>	<i>hpo-34</i>	<i>cyp-33C1</i>	<i>ifd-2</i>	<i>pgp-6</i>
<i>gba-4</i>	<i>C30F2.5</i>	<i>clec-218</i>	<i>F49H12.5</i>	<i>F54E2.1</i>	<i>B0348.1</i>	<i>C36B1.14</i>
<i>pho-1</i>	<i>H36L18.2</i>	<i>thn-2</i>	<i>T24D3.2</i>	<i>T16G1.6</i>	<i>cul-6</i>	<i>ZK1320.13</i>
<i>F38B6.4</i>	<i>ZK287.3</i>	<i>dod-3</i>	<i>lon-3</i>	<i>R08F11.4</i>	<i>Y41D4B.15</i>	<i>F47B7.3</i>
<i>ugt-62</i>	<i>F30A10.13</i>	<i>ZC395.5</i>	<i>col-145</i>	<i>sodh-1</i>	<i>C18H7.11</i>	<i>C39H7.4</i>
<i>F46F2.3</i>	<i>elc-1</i>	<i>aqp-1</i>	<i>aqp-11</i>	<i>acs-2</i>	<i>Y69A2AL.2</i>	<i>nlp-41</i>
<i>msra-1</i>	<i>fbxa-21</i>	<i>far-7</i>	<i>Y62H9A.4</i>	<i>ilys-3</i>	<i>F53B2.8</i>	<i>B0252.8</i>
<i>acs-1</i>	<i>F30A10.14</i>	<i>F53A9.8</i>	<i>srv-1</i>	<i>C23G10.11</i>	<i>T28H10.3</i>	<i>flp-26</i>
<i>clec-49</i>	<i>skr-12</i>	<i>C25H3.10</i>	<i>Y51H4A.24</i>	<i>fmo-2</i>	<i>acs-7</i>	<i>fbxa-58</i>
<i>nhr-114</i>	<i>C02F12.5</i>	<i>tts-1</i>	<i>math-4</i>	<i>pals-6</i>	<i>clec-86</i>	<i>F13E9.15</i>
<i>fbxa-72</i>	<i>ttr-42</i>	<i>ilys-2</i>	<i>C01H6.4</i>	<i>B0507.10</i>	<i>hrg-3</i>	<i>F14F9.8</i>
<i>sams-1</i>	<i>D1086.19</i>	<i>F15E6.3</i>	<i>cdr-2</i>	<i>clec-174</i>	<i>best-5</i>	<i>H32K16.2</i>
<i>T13F3.6</i>	<i>F49E8.2</i>	<i>C06G3.3</i>	<i>ppat-1</i>	<i>dod-19</i>	<i>col-135</i>	<i>srap-1</i>
<i>Y48E1B.8</i>	<i>Y49E10.29</i>	<i>F43C11.7</i>	<i>aat-4</i>	<i>H43E16.1</i>	<i>dhrs-4</i>	<i>F10E7.11</i>
<i>ugt-63</i>	<i>col-178</i>	<i>F21C10.10</i>	<i>C17E7.12</i>	<i>Y94H6A.10</i>	<i>F59C6.16</i>	<i>C30G7.4</i>
<i>sur-5</i>	<i>ifb-1</i>	<i>cyp-37B1</i>	<i>cyp-36A1</i>	<i>drd-50</i>	<i>C50F4.1</i>	<i>H29C22.1</i>
<i>ugt-26</i>	<i>pyc-1</i>	<i>hsp-12.3</i>	<i>W02B12.4</i>	<i>clec-265</i>	<i>nlp-34</i>	<i>hpo-15</i>
<i>ZK180.6</i>	<i>Y47G6A.33</i>	<i>F11A5.9</i>	<i>col-179</i>	<i>F19B2.5</i>	<i>Y22D7AL.15</i>	<i>F27D4.8</i>
<i>clec-187</i>	<i>H17B01.2</i>	<i>C46H11.2</i>	<i>far-3</i>	<i>K08D8.4</i>	<i>gem-4</i>	<i>F36A2.12</i>
<i>sqt-1</i>	<i>scl-20</i>	<i>F22B7.9</i>	<i>acox-1.5</i>	<i>T24B8.5</i>	<i>sqst-1</i>	
<i>rol-1</i>	<i>C39D10.8</i>	<i>T28B8.1</i>	<i>gba-1</i>	<i>ZK6.11</i>	<i>F46A8.7</i>	
<i>col-175</i>	<i>tag-147</i>	<i>C49F5.7</i>	<i>endu-2</i>	<i>C17H12.8</i>	<i>fbxa-164</i>	
<i>lipl-1</i>	<i>ttr-49</i>	<i>C27B7.9</i>	<i>M28.10</i>	<i>clec-67</i>	<i>F08B12.4</i>	
<i>F55G11.2</i>	<i>btb-16</i>	<i>ZC204.12</i>	<i>K06G5.1</i>	<i>ZK896.5</i>	<i>Y95B8A.6</i>	

List585

<i>acs-2</i>	<i>tsp-10</i>	<i>T16G12.1</i>	<i>T19B4.3</i>	<i>ZK550.2</i>	<i>F21C10.11</i>	<i>nlp-28</i>
<i>sodh-1</i>	<i>clec-170</i>	<i>E04F6.15</i>	<i>col-178</i>	<i>hum-6</i>	<i>C41G7.8</i>	<i>coel-1</i>
<i>fmo-2</i>	<i>gpx-7</i>	<i>Y49E10.18</i>	<i>Y48A5A.3</i>	<i>col-10</i>	<i>K04C2.8</i>	<i>T12D8.5</i>
<i>F09F7.6</i>	<i>math-45</i>	<i>F32H5.1</i>	<i>F15B10.3</i>	<i>haf-9</i>	<i>C47E8.11</i>	<i>M7.8</i>
<i>cyp-37B1</i>	<i>F13H8.3</i>	<i>acbp-3</i>	<i>skr-12</i>	<i>math-4</i>	<i>C54C6.7</i>	<i>F13D11.3</i>
<i>far-7</i>	<i>ugt-23</i>	<i>hpd-1</i>	<i>Y22D7AR.10</i>	<i>F55H12.3</i>	<i>fat-7</i>	<i>Y95B8A.6</i>
<i>C23G10.11</i>	<i>ZK185.5</i>	<i>W02G9.4</i>	<i>C02F12.5</i>	<i>aman-1</i>	<i>gst-1</i>	<i>F43G6.8</i>
<i>F53A9.8</i>	<i>nhr-144</i>	<i>cdr-2</i>	<i>F30A10.14</i>	<i>comt-3</i>	<i>W03D2.6</i>	<i>D1086.5</i>
<i>ilys-3</i>	<i>F48D6.4</i>	<i>F13B6.2</i>	<i>tni-3</i>	<i>ech-8</i>	<i>ttr-44</i>	<i>T28F4.5</i>
<i>ilys-2</i>	<i>srv-1</i>	<i>acox-1.2</i>	<i>Y69E1A.5</i>	<i>rol-1</i>	<i>aqp-1</i>	<i>ZK550.6</i>
<i>nhr-114</i>	<i>C14C11.4</i>	<i>elo-2</i>	<i>F14B8.4</i>	<i>col-175</i>	<i>hsp-12.3</i>	<i>T04A6.1</i>
<i>fbxa-72</i>	<i>Y51H4A.24</i>	<i>F38B6.4</i>	<i>ttr-42</i>	<i>odc-1</i>	<i>asm-3</i>	<i>Y39B6A.5</i>
<i>pmp-5</i>	<i>Y62H9A.4</i>	<i>ppat-1</i>	<i>F30A10.13</i>	<i>C29G2.2</i>	<i>dod-3</i>	<i>F10E9.12</i>
<i>msra-1</i>	<i>ugt-5</i>	<i>C01H6.4</i>	<i>F09E10.1</i>	<i>Y69A2AR.25</i>	<i>scl-2</i>	<i>E04F6.8</i>
<i>srh-237</i>	<i>cpg-9</i>	<i>F58G6.3</i>	<i>T01C3.11</i>	<i>D2045.2</i>	<i>clec-82</i>	<i>C49F5.7</i>
<i>T13F3.6</i>	<i>ent-7</i>	<i>F58G6.7</i>	<i>C04E6.13</i>	<i>2RSSE.1</i>	<i>nnt-1</i>	<i>C27B7.9</i>
<i>C23H5.8</i>	<i>ZK899.2</i>	<i>ugt-47</i>	<i>R02F2.8</i>	<i>C38H2.3</i>	<i>ZC204.12</i>	<i>ttr-8</i>
<i>spp-23</i>	<i>zip-3</i>	<i>aat-4</i>	<i>vap-1</i>	<i>Y105E8A.28</i>	<i>ZC395.5</i>	<i>F46C3.6</i>
<i>Y48E1B.8</i>	<i>Y43F8C.13</i>	<i>amt-4</i>	<i>H36L18.2</i>	<i>Y102A5C.36</i>	<i>F21C10.10</i>	<i>F45E10.2</i>
<i>pho-1</i>	<i>D1086.3</i>	<i>drd-1</i>	<i>ZK287.3</i>	<i>nhr-6</i>	<i>F11A5.9</i>	<i>fbxa-157</i>
<i>C05D12.3</i>	<i>C44B7.7</i>	<i>vit-3</i>	<i>F25B4.8</i>	<i>slc-17.9</i>	<i>C32F10.4</i>	<i>Y53F4B.45</i>
<i>T05E12.6</i>	<i>Y47G6A.33</i>	<i>nhr-68</i>	<i>H32K21.1</i>	<i>Y46H3A.4</i>	<i>C06B3.6</i>	<i>C25H3.10</i>
<i>folt-2</i>	<i>ZK512.7</i>	<i>drd-5</i>	<i>elc-1</i>	<i>Y37D8A.6</i>	<i>F18G5.6</i>	<i>Y65B4BR.1</i>
<i>H17B01.2</i>	<i>dhs-26</i>	<i>F44A6.5</i>	<i>B0250.4</i>	<i>C42D8.1</i>	<i>skr-5</i>	<i>cup-16</i>
<i>ram-2</i>	<i>pud-4</i>	<i>DH11.2</i>	<i>T26C12.3</i>	<i>C54D10.13</i>	<i>Y46G5A.20</i>	<i>tts-1</i>
<i>aldo-1</i>	<i>cth-1</i>	<i>lys-6</i>	<i>C53H9.3</i>	<i>T08A9.13</i>	<i>ZC196.1</i>	<i>T23F11.6</i>
<i>T28D9.3</i>	<i>F10A3.4</i>	<i>spp-4</i>	<i>BE0003N10.</i>	<i>F20G2.2</i>	<i>K04F1.9</i>	<i>C23H4.6</i>
<i>haf-4</i>	<i>atic-1</i>	<i>metr-1</i>	<i>3</i>	<i>aqp-11</i>	<i>ins-11</i>	<i>Y102A11A.9</i>
<i>T22F3.3</i>	<i>ech-7</i>	<i>mthf-1</i>	<i>fipr-29</i>	<i>fbxc-32</i>	<i>F53A9.6</i>	<i>Y22D7AL.15</i>
<i>Y49E10.29</i>	<i>C46F2.1</i>	<i>Y51H7C.1</i>	<i>H01G02.1</i>	<i>F40G12.11</i>	<i>F53A9.1</i>	<i>C33H5.13</i>
<i>gpx-5</i>	<i>papl-1</i>	<i>R08E5.3</i>	<i>clec-227</i>	<i>ugt-31</i>	<i>endu-2</i>	<i>gem-4</i>
<i>mct-3</i>	<i>D1054.8</i>	<i>sams-1</i>	<i>cpr-1</i>	<i>gpx-1</i>	<i>C50F4.9</i>	<i>Y37D8A.16</i>
<i>nstp-2</i>	<i>gst-26</i>	<i>acs-1</i>	<i>ZK673.1</i>	<i>asp-14</i>	<i>F10D7.3</i>	<i>sqst-1</i>
<i>C49G7.3</i>	<i>F46F2.3</i>	<i>C30G12.2</i>	<i>C53C9.2</i>	<i>cpt-5</i>	<i>T28B8.1</i>	<i>ptr-23</i>
<i>dim-1</i>	<i>ilys-5</i>	<i>ugt-63</i>	<i>gln-3</i>	<i>fbxa-51</i>	<i>C35C5.9</i>	<i>srap-1</i>
<i>mct-4</i>	<i>gst-27</i>	<i>T22B7.7</i>	<i>Y38H6C.19</i>	<i>glb-1</i>	<i>aqp-2</i>	<i>tep-1</i>
<i>ahcy-1</i>	<i>grd-14</i>	<i>pud-3</i>	<i>T06D8.3</i>	<i>ZK669.2</i>	<i>dct-1</i>	<i>icmt-1</i>
<i>mel-32</i>	<i>C26B9.5</i>	<i>clec-172</i>	<i>ifb-1</i>	<i>spp-22</i>	<i>asah-1</i>	<i>fbxa-140</i>
<i>col-145</i>	<i>T19C3.2</i>	<i>ugt-17</i>	<i>T03D8.6</i>	<i>F59A6.12</i>	<i>F08B12.4</i>	<i>K07E3.4</i>
<i>col-180</i>	<i>F49C12.14</i>	<i>sdz-24</i>	<i>mrp-3</i>	<i>col-104</i>	<i>ttr-17</i>	<i>F07C3.9</i>
<i>F10F2.2</i>	<i>gba-4</i>	<i>lips-14</i>	<i>aco-1</i>	<i>T21H3.1</i>	<i>droe-4</i>	<i>F56D5.6</i>
<i>K11G9.2</i>	<i>clec-56</i>	<i>lys-4</i>	<i>K01C8.1</i>	<i>col-77</i>	<i>M03A1.8</i>	<i>F13E9.15</i>
<i>lipl-5</i>	<i>T15B7.1</i>	<i>mtl-2</i>	<i>E01G4.3</i>	<i>T24D3.2</i>	<i>C06G3.3</i>	<i>C46H11.2</i>
<i>F26C11.1</i>	<i>ech-6</i>	<i>thn-2</i>	<i>pyc-1</i>	<i>F49H12.5</i>	<i>hil-1</i>	<i>Y54G2A.36</i>
<i>F28A12.3</i>	<i>C35A11.4</i>	<i>F15E6.4</i>	<i>C03B1.13</i>	<i>grd-4</i>	<i>F15E6.3</i>	<i>best-5</i>
<i>alh-12</i>	<i>F53F1.4</i>	<i>clec-218</i>	<i>hmit-1.1</i>	<i>ttr-41</i>	<i>F43C11.7</i>	<i>del-5</i>
<i>fkf-3</i>	<i>ugt-26</i>	<i>clec-5</i>	<i>tag-10</i>	<i>ttr-49</i>	<i>cnc-4</i>	<i>F54C9.3</i>
<i>C39D10.8</i>	<i>lon-3</i>	<i>fipr-22</i>	<i>epi-1</i>	<i>btb-16</i>	<i>nlp-34</i>	<i>C49G9.2</i>
<i>tag-147</i>	<i>ZK180.6</i>	<i>C02B4.4</i>	<i>cyp-35A2</i>	<i>cdr-6</i>	<i>Y43C5A.3</i>	<i>glct-6</i>
<i>hprt-1</i>	<i>pgp-9</i>	<i>T23E7.6</i>	<i>gst-4</i>	<i>nas-20</i>	<i>F45D3.4</i>	<i>H29C22.1</i>
<i>slc-17.4</i>	<i>hpo-34</i>	<i>cav-1</i>	<i>pho-11</i>	<i>ZK1320.13</i>	<i>nlp-29</i>	<i>flp-26</i>
<i>C53A3.2</i>	<i>ent-4</i>	<i>C29F7.2</i>	<i>pcp-3</i>	<i>C36B1.14</i>	<i>ZK970.7</i>	<i>C30G7.4</i>
<i>acl-4</i>	<i>mth-1</i>	<i>gst-20</i>	<i>E01G6.3</i>	<i>C17E7.12</i>	<i>nhr-19</i>	<i>tyr-1</i>
<i>Y54E10A.17</i>	<i>ugt-62</i>	<i>F22F4.5</i>	<i>dhs-2</i>	<i>Y60A3A.24</i>	<i>spp-8</i>	<i>H32K16.2</i>
<i>F49E8.2</i>	<i>F09B12.3</i>	<i>T11B7.2</i>	<i>Y32F6B.1</i>	<i>D1007.19</i>	<i>nhr-21</i>	<i>F14F9.8</i>
<i>T12B5.14</i>	<i>sur-5</i>	<i>T12B5.15</i>	<i>F41C3.2</i>	<i>fbxa-37</i>	<i>ZK669.3</i>	<i>hpo-15</i>
<i>best-7</i>	<i>K10C2.1</i>	<i>gst-24</i>	<i>W07G4.5</i>	<i>F27D4.8</i>	<i>C33A12.4</i>	<i>F21A3.11</i>

<i>K09E9.4</i>	<i>F53F4.13</i>	<i>C39H7.4</i>	<i>Y38C1AA.6</i>	<i>clec-67</i>	<i>spp-2</i>	<i>hsp-17</i>
<i>igeg-1</i>	<i>Y51A2D.13</i>	<i>F14H3.12</i>	<i>C33A12.19</i>	<i>T24B8.5</i>	<i>col-179</i>	<i>drd-50</i>
<i>Y58A7A.5</i>	<i>cyp-33C7</i>	<i>cbp-2</i>	<i>C55B7.3</i>	<i>clec-49</i>	<i>Y54G2A.49</i>	<i>F22H10.2</i>
<i>T26H5.9</i>	<i>C05C12.4</i>	<i>pgp-6</i>	<i>F10A3.17</i>	<i>F31D4.8</i>	<i>F40F8.5</i>	<i>sri-36</i>
<i>W02A2.9</i>	<i>cyp-13B1</i>	<i>math-24</i>	<i>F57B1.9</i>	<i>C45B2.1</i>	<i>Y34B4A.6</i>	<i>B0205.13</i>
<i>srw-86</i>	<i>srr-6</i>	<i>nlp-41</i>	<i>col-162</i>	<i>F28H7.3</i>	<i>C32H11.4</i>	<i>H43E16.1</i>
<i>C25F9.12</i>	<i>F45D3.3</i>	<i>dhps-4</i>	<i>tag-38</i>	<i>F42H10.6</i>	<i>dod-17</i>	<i>clec-66</i>
<i>irg-3</i>	<i>ugt-25</i>	<i>col-135</i>	<i>ZK1290.14</i>	<i>clec-72</i>	<i>F01D5.1</i>	<i>hpo-6</i>
<i>C04G6.5</i>	<i>C08F11.13</i>	<i>F59C6.16</i>	<i>B0403.5</i>	<i>far-3</i>	<i>M02H5.8</i>	<i>T19D12.4</i>
<i>T28H10.3</i>	<i>C10C5.5</i>	<i>C50F4.1</i>	<i>B0457.6</i>	<i>lipl-1</i>	<i>F01D5.5</i>	<i>gst-38</i>
<i>C08E8.4</i>	<i>ctl-2</i>	<i>Y34F4.4</i>	<i>F40F12.7</i>	<i>F55G11.2</i>	<i>ctsa-2</i>	<i>cpr-3</i>
<i>hrg-3</i>	<i>asp-6</i>	<i>K10G4.3</i>	<i>clec-25</i>	<i>K08D8.6</i>	<i>K06G5.1</i>	<i>tsp-1</i>
<i>acs-7</i>	<i>egl-15</i>	<i>Y67A10A.10</i>	<i>cyp-33C8</i>	<i>asp-12</i>	<i>tag-244</i>	<i>zip-10</i>
<i>gba-1</i>	<i>M01A8.1</i>	<i>F22D6.15</i>	<i>ifo-1</i>	<i>irg-4</i>	<i>H20E11.1</i>	<i>greg-1</i>
<i>M28.10</i>	<i>K02D7.1</i>	<i>C33G8.2</i>	<i>ZK418.7</i>	<i>ugt-18</i>	<i>F54D5.4</i>	<i>lec-11</i>
<i>gale-1</i>	<i>C07H4.1</i>	<i>C49A9.9</i>	<i>F16H6.10</i>	<i>ech-9</i>	<i>F54E2.1</i>	<i>F35E12.9</i>
<i>F55B11.4</i>	<i>R07C12.4</i>	<i>cul-6</i>	<i>ifd-2</i>	<i>irg-5</i>	<i>K08D8.5</i>	<i>mul-1</i>
<i>clec-62</i>	<i>M04D5.3</i>	<i>Y113G7B.14</i>	<i>F10D2.10</i>	<i>C50F7.5</i>	<i>ZK896.5</i>	<i>M28.8</i>
<i>acox-1.5</i>	<i>nhr-231</i>	<i>F35E12.4</i>	<i>fbxa-59</i>	<i>C49C8.5</i>	<i>C34H4.1</i>	<i>Y41D4B.17</i>
<i>H34I24.2</i>	<i>cysl-2</i>	<i>Y41D4B.15</i>	<i>K01F9.2</i>	<i>oac-31</i>	<i>oac-6</i>	
<i>C18E9.9</i>	<i>dhs-9</i>	<i>F53B2.8</i>	<i>clec-86</i>	<i>B0024.4</i>	<i>dod-19</i>	
<i>asp-5</i>	<i>acds-10</i>	<i>C18H7.11</i>	<i>Y94H6A.10</i>	<i>H02F09.3</i>	<i>F55G11.8</i>	
<i>Y47D3B.1</i>	<i>cyp-36A1</i>	<i>Y69A2AL.2</i>	<i>swt-6</i>	<i>C25F9.11</i>	<i>K11H12.4</i>	
<i>clec-187</i>	<i>pmp-2</i>	<i>W02B12.4</i>	<i>Y17G7B.8</i>	<i>ZK228.4</i>	<i>Y75B8A.28</i>	
<i>sqt-1</i>	<i>acox-3</i>	<i>H03A11.2</i>	<i>tag-234</i>	<i>T16G1.6</i>	<i>F19B2.5</i>	
<i>abt-4</i>	<i>R07C12.2</i>	<i>Y73B6BL.31</i>	<i>gst-22</i>	<i>R08F11.4</i>	<i>lys-2</i>	
<i>clec-186</i>	<i>B0252.8</i>	<i>clec-84</i>	<i>ZK6.11</i>	<i>ZK228.3</i>	<i>clec-265</i>	
<i>ttr-23</i>	<i>C27D6.12</i>	<i>ZK287.9</i>	<i>C17H12.8</i>	<i>Y54G2A.45</i>	<i>K08D8.4</i>	

List383

<i>cyp-35A5</i>	<i>ugt-22</i>	<i>F09F9.2</i>	<i>F53F1.6</i>	<i>H02F09.2</i>	<i>cnc-4</i>	<i>H43E16.1</i>
<i>cyp-35D1</i>	<i>lbp-8</i>	<i>F18C5.5</i>	<i>dod-3</i>	<i>F13B6.1</i>	<i>nlp-29</i>	<i>F20G2.5</i>
<i>cyp-35A3</i>	<i>vit-1</i>	<i>K08B12.1</i>	<i>aqp-1</i>	<i>ifd-2</i>	<i>nlp-31</i>	<i>irg-1</i>
<i>lys-6</i>	<i>Y48E1B.8</i>	<i>Y47D7A.13</i>	<i>scl-2</i>	<i>K11H12.3</i>	<i>C06B3.6</i>	<i>hpo-6</i>
<i>F54B11.11</i>	<i>F46F2.3</i>	<i>col-54</i>	<i>F21C10.10</i>	<i>Y41D4B.15</i>	<i>gem-4</i>	<i>C25F9.11</i>
<i>cyp-35C1</i>	<i>ugt-63</i>	<i>D1014.5</i>	<i>comt-4</i>	<i>F35E12.6</i>	<i>abf-2</i>	<i>Y47H9C.1</i>
<i>F21C10.9</i>	<i>R07E5.4</i>	<i>dpy-5</i>	<i>cyp-34A9</i>	<i>clec-86</i>	<i>cyp-32B1</i>	<i>F15B9.6</i>
<i>Y34F4.2</i>	<i>gpdh-1</i>	<i>ZK154.1</i>	<i>clec-61</i>	<i>oac-31</i>	<i>F47B8.2</i>	<i>Y58A7A.5</i>
<i>ZK593.3</i>	<i>smd-1</i>	<i>wrt-1</i>	<i>C29F7.2</i>	<i>zip-10</i>	<i>tag-196</i>	<i>Y47H10A.5</i>
<i>hphd-1</i>	<i>sdz-6</i>	<i>lys-5</i>	<i>cpt-4</i>	<i>kreg-1</i>	<i>cnc-7</i>	<i>F43C1.7</i>
<i>nhr-68</i>	<i>C43D7.7</i>	<i>lys-4</i>	<i>W07B8.4</i>	<i>tag-10</i>	<i>glb-1</i>	<i>C25F9.12</i>
<i>asp-13</i>	<i>F55G11.2</i>	<i>lys-7</i>	<i>mtl-2</i>	<i>Y32F6B.1</i>	<i>F59C6.16</i>	<i>W02A2.9</i>
<i>C30G12.2</i>	<i>irg-5</i>	<i>F49C12.14</i>	<i>thn-2</i>	<i>E01G6.3</i>	<i>C05D9.9</i>	<i>srw-86</i>
<i>fbxa-72</i>	<i>C32H11.4</i>	<i>cpr-4</i>	<i>clec-218</i>	<i>cpi-1</i>	<i>F47B8.4</i>	<i>sri-36</i>
<i>sams-1</i>	<i>dod-17</i>	<i>alh-12</i>	<i>fipr-22</i>	<i>asm-3</i>	<i>gbh-2</i>	<i>Y94H6A.2</i>
<i>nhr-114</i>	<i>irg-3</i>	<i>F26C11.1</i>	<i>R07C12.1</i>	<i>hrg-1</i>	<i>F10E9.12</i>	<i>T24E12.5</i>
<i>msra-1</i>	<i>C17H12.8</i>	<i>F28A12.3</i>	<i>W09G12.7</i>	<i>ttr-44</i>	<i>C34F11.8</i>	<i>Y46G5A.20</i>
<i>acs-1</i>	<i>irg-4</i>	<i>papl-1</i>	<i>C54C8.12</i>	<i>ugt-6</i>	<i>nlp-34</i>	<i>B0348.2</i>
<i>srh-237</i>	<i>ech-9</i>	<i>dhs-26</i>	<i>fipr-26</i>	<i>lips-14</i>	<i>F45D3.4</i>	<i>ZC196.1</i>
<i>spp-23</i>	<i>dod-24</i>	<i>F25D1.5</i>	<i>cnc-2</i>	<i>clec-57</i>	<i>ZC443.3</i>	<i>dod-23</i>
<i>clec-48</i>	<i>ZK896.5</i>	<i>Y119D3B.1</i>	<i>skr-7</i>	<i>F38B6.4</i>	<i>F18G5.6</i>	<i>fbxa-182</i>
<i>clec-49</i>	<i>F55G11.4</i>	<i>3</i>	<i>skr-15</i>	<i>elo-2</i>	<i>C32F10.4</i>	<i>gpa-17</i>
<i>cln-3.1</i>	<i>K08D8.5</i>	<i>sdz-24</i>	<i>Y69A2AR.2</i>	<i>ugt-62</i>	<i>Y43C5A.3</i>	<i>C50F4.1</i>
<i>C18A11.3</i>	<i>clec-67</i>	<i>ugt-17</i>	<i>5</i>	<i>M03B6.1</i>	<i>tts-1</i>	<i>tag-234</i>
<i>cdr-1</i>	<i>dod-22</i>	<i>cyp-35A2</i>	<i>Y51B9A.8</i>	<i>F09B12.3</i>	<i>best-1</i>	<i>C08E8.4</i>
<i>Y39B6A.1</i>	<i>pmp-5</i>	<i>metr-1</i>	<i>fbxa-163</i>	<i>sur-5</i>	<i>Y65B4BR.1</i>	<i>oac-14</i>
<i>spp-12</i>	<i>acdh-1</i>	<i>mthf-1</i>	<i>F08G2.5</i>	<i>C35A5.3</i>	<i>F44G3.10</i>	<i>M28.8</i>
<i>F42A10.7</i>	<i>T05E12.6</i>	<i>srr-4</i>	<i>clec-3</i>	<i>hacd-1</i>	<i>T24C4.4</i>	<i>Y41D4B.17</i>
<i>C18H9.6</i>	<i>folt-2</i>	<i>F44A6.5</i>	<i>C49G7.12</i>	<i>ugt-21</i>	<i>B0024.4</i>	<i>Y51H4A.25</i>
<i>lipl-1</i>	<i>sodh-1</i>	<i>DH11.2</i>	<i>C06B3.7</i>	<i>pho-11</i>	<i>pcp-2</i>	<i>tba-7</i>
<i>ugt-18</i>	<i>clec-60</i>	<i>F15E6.4</i>	<i>F54B8.4</i>	<i>amt-4</i>	<i>ZK228.4</i>	<i>cul-6</i>
<i>clec-51</i>	<i>thn-1</i>	<i>argk-1</i>	<i>irg-2</i>	<i>dhs-25</i>	<i>F01D5.3</i>	<i>Y94H6A.10</i>
<i>clec-56</i>	<i>ilys-3</i>	<i>Y38H6C.21</i>	<i>ZK896.4</i>	<i>cyp-25A1</i>	<i>T24B8.5</i>	<i>C18H7.11</i>
<i>ZK512.7</i>	<i>acs-2</i>	<i>Y51H7C.1</i>	<i>C04G6.5</i>	<i>R08F11.4</i>	<i>K08D8.4</i>	<i>F53B2.8</i>
<i>nhx-2</i>	<i>T22F3.11</i>	<i>R08E5.3</i>	<i>F53A9.6</i>	<i>F29C6.1</i>	<i>dct-17</i>	<i>tsp-1</i>
<i>spp-4</i>	<i>F46C5.1</i>	<i>pmt-2</i>	<i>lys-3</i>	<i>T05E7.1</i>	<i>clec-66</i>	<i>K04F1.9</i>
<i>clec-53</i>	<i>C23G10.11</i>	<i>gly-8</i>	<i>Y22D7AL.1</i>	<i>gst-1</i>	<i>cld-9</i>	<i>tsp-2</i>
<i>ugt-46</i>	<i>F53A9.8</i>	<i>Y46G5A.29</i>	<i>5</i>	<i>hrg-4</i>	<i>K11H12.4</i>	<i>T19D12.4</i>
<i>T15B7.1</i>	<i>F09F7.6</i>	<i>K08D8.3</i>	<i>valv-1</i>	<i>R193.2</i>	<i>clec-4</i>	<i>cpr-3</i>
<i>ilys-5</i>	<i>C35C5.8</i>	<i>col-156</i>	<i>F42H10.6</i>	<i>acs-7</i>	<i>ugt-44</i>	<i>mul-1</i>
<i>ech-6</i>	<i>far-7</i>	<i>T05H10.3</i>	<i>F49F1.5</i>	<i>ttr-23</i>	<i>clc-1</i>	<i>F35E12.9</i>
<i>F53F1.4</i>	<i>fmo-2</i>	<i>F41E7.7</i>	<i>gba-1</i>	<i>T28H10.3</i>	<i>F49F1.7</i>	<i>T01D3.6</i>
<i>gba-4</i>	<i>pals-11</i>	<i>Y77E11A.1</i>	<i>fil-1</i>	<i>ftn-1</i>	<i>Y75B8A.28</i>	<i>K09D9.1</i>
<i>ugt-26</i>	<i>B0507.10</i>	<i>4</i>	<i>Y47D7A.7</i>	<i>T21C9.6</i>	<i>F19B2.5</i>	<i>Y17G7B.8</i>
<i>gst-28</i>	<i>pals-6</i>	<i>asah-1</i>	<i>ins-7</i>	<i>hpd-1</i>	<i>F01D5.2</i>	<i>swt-6</i>
<i>C45B2.1</i>	<i>pals-32</i>	<i>Y54G2A.11</i>	<i>ugt-25</i>	<i>cdr-2</i>	<i>clec-85</i>	
<i>col-101</i>	<i>pals-37</i>	<i>K01A2.10</i>	<i>W02G9.4</i>	<i>ugt-43</i>	<i>F53C11.1</i>	
<i>F09C8.1</i>	<i>C43D7.4</i>	<i>F15E6.3</i>	<i>T02B11.4</i>	<i>aman-3</i>	<i>clec-265</i>	
<i>fpn-1.2</i>	<i>pals-2</i>	<i>hil-1</i>	<i>srr-6</i>	<i>F55H12.3</i>	<i>oac-20</i>	
<i>col-143</i>	<i>T05A8.2</i>	<i>T28A11.19</i>	<i>clec-72</i>	<i>gtl-1</i>	<i>drd-50</i>	
<i>col-8</i>	<i>pals-29</i>	<i>C06G3.3</i>	<i>far-3</i>	<i>F54C9.3</i>	<i>C50F7.5</i>	
<i>drd-5</i>	<i>pals-28</i>	<i>F36F2.2</i>	<i>K08D8.6</i>	<i>hex-2</i>	<i>H02F09.3</i>	
<i>pud-3</i>	<i>fat-7</i>	<i>Y43F8B.9</i>	<i>Y46D2A.2</i>	<i>ech-8</i>	<i>F22H10.2</i>	
<i>vit-3</i>	<i>ilys-2</i>	<i>C33A12.4</i>	<i>gale-1</i>	<i>cyp-37B1</i>	<i>clec-45</i>	
<i>T13F3.6</i>	<i>W03F9.4</i>	<i>clec-264</i>	<i>ugt-16</i>	<i>nnt-1</i>	<i>Y37H2A.14</i>	
<i>pho-13</i>	<i>grl-15</i>	<i>F41C3.1</i>	<i>acox-1.5</i>	<i>C25H3.10</i>	<i>C49C3.9</i>	
<i>C23H5.8</i>	<i>C26B9.3</i>	<i>T05E12.3</i>	<i>parg-2</i>	<i>ZC395.5</i>	<i>F25A2.1</i>	

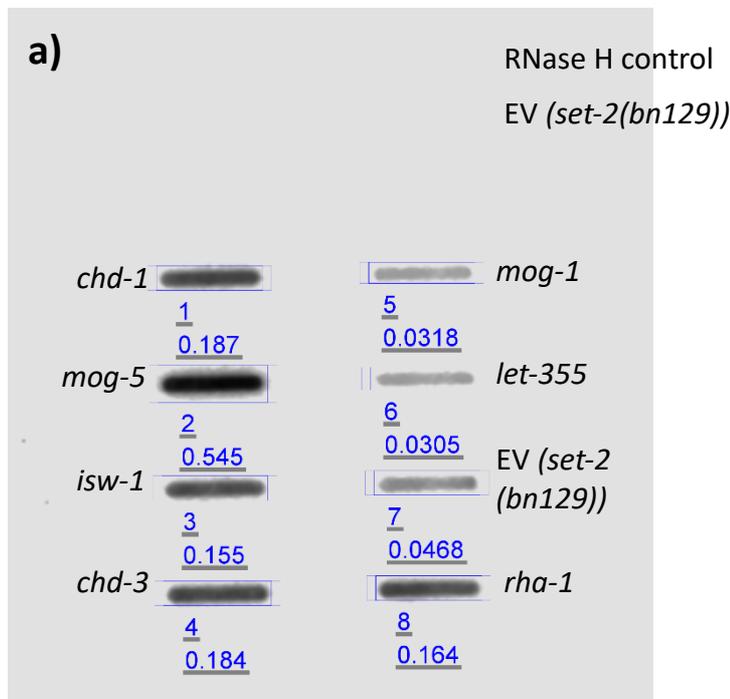
Heatshock255

F02H6.2	F18E3.12	F42A8.1	F53B2.8	ttr-7	C30E1.9	ZK909.3
R04D3.4	F22E5.1	nlp-26	ZK1307.7	fbxa-54	scp-1	nlp-29
T02G6.5	ZK177.1	col-181	ipla-2	T20D4.10	Y43F8B.2	C05B5.8
F02H6.3	R04D3.3	col-8	nhr-63	gst-1	hsp-110	hsp-16.2
imp-1	T01G5.7	T28D6.3	lin-12	Y53H1B.2	aip-1	hsp-16.41
F35C11.5	F23D12.2	W07A12.8	B0507.3	gly-8	F29B9.5	dct-10
ZK822.5	K04C1.5	elo-6	R09E12.9	his-41	Y56A3A.33	C47E8.11
Y32F6A.5	F14H3.4	ugt-63	K10D3.6	cls-2	zip-10	F01D5.7
clec-218	W06D11.3	K08D12.6	C25H3.10	cdd-1	srx-76	R10D12.19
T01D3.6	F14H3.3	Y48E1B.8	tos-1	ttr-6	ttll-12	R09E10.13
col-135	F40G12.11	F18E3.11	Y75B8A.32	C03G6.5	Y94H6A.10	linc-122
K10C2.1	Y47D7A.6	sox-2	cup-16	fmo-3	col-40	fbxb-72
F56F10.1	R193.2	cutl-24	T28F4.5	comt-3	col-84	R02E4.3
prg-2	asp-13	R09B5.11	Y54G2A.11	tni-1	col-44	C45B2.8
clec-222	K06A4.7	svh-2	Y75B8A.28	dao-2	linc-6	nlp-34
nrf-6	oac-54	eat-16	C53A3.2	cpg-7	F19B2.5	F08G2.5
R09H10.5	F55H12.2	rca-1	oac-14	cpt-4	hsp-4	cnc-4
T06D4.1	F49C12.7	bcl-11	zig-3	linc-7	F09E10.15	Y17D7C.2
clec-227	pept-1	olrn-1	F35D2.1	Y39B6A.25	F59C12.4	col-36
cpr-1	fbxa-215	ztf-2	C08H9.15	F32A5.4	clec-196	Y38H6C.8
clec-7	T01C3.3	aex-3	pgp-1	gpdh-1	nlp-25	Y38H6C.25
F13H8.3	inx-14	peb-1	K12B6.11	Y53F4B.11	his-1	srt-42
rhr-1	T05F1.2	F43G6.4	cnc-8	F37C4.5	K02E2.8	mir-239.1
F35C12.3	clec-47	K03A11.5	msp-42	F13H10.6	K01D12.9	nhr-241
W02B3.4	ZK829.9	sup-26	ins-33	cpi-1	col-137	hsp-70
asp-8	F58G6.9	pmk-3	H17B01.2	T09F5.12	col-156	F44E5.4
Y37D8A.4	fat-6	F34H10.3	R11H6.7	D20G3.1	grl-27	F44E5.5
elo-5	nhr-114	gei-1	glf-1	K06H6.2	C04G6.2	R11A5.3
elo-2	ugt-44	svh-1	H23N18.5	F15E6.3	Y47D7A.15	ZC21.10
clec-265	R04B5.5	K10B4.3	dgat-2	tts-1	T01B7.13	M03F4.12
lys-2	Y53C10A.5	F21G4.1	T13H5.6	gln-4	Y47D3B.6	F33H12.6
cth-2	ZK185.3	nhr-213	glb-11	msd-4	T27F6.8	Y53F4B.8
T22F3.3	K09H9.5	T19D12.4	F21A3.3	C31H2.14	his-5	F26D10.23
twk-16	F58G6.3	pho-9	src-2	M05D6.3	col-14	
C40H1.7	T22B7.7	fkh-7	hmit-1.1	linc-84	Y41C4A.32	
F11E6.3	F23D12.11	T24C4.4	F46G11.2	pqn-44	F26G1.5	
mxl-3	T28C12.4	wht-1	cdh-7	Y50E8A.12	C25F9.11	

Appendix 20 Table showing all the genes in each of the lists: list331, list585, list383 and Heatmap255. The colour coding shows individual clusters as determined by the dendrogram cut-off. The order of the genes from top to bottom and left to right corresponds to the heatmap from left to right.

Appendix 21

All the images shown here are the data corresponding to the RNAi helicase screen on the *set-2(bn129)* mutant background (**Table 4.1**). Results of a few additional helicase candidates are shown here but were not included in **Table 4.1** as these lack a replicate. All slot blots were visualized as explained in section 2.8.2, except for the first slot blot, which was visualized using the LI-COR Odyssey® FC imaging system. One slot blot experiment is presented per page. The amount of DNA loaded is dependent on the sample with the least amount of available sample. When comparing the signal intensity between samples, the “quantification” from G:BOX is always preferred over ImageJ quantification, because G:BOX uses the raw data, while ImageJ uses the compressed and processed image.

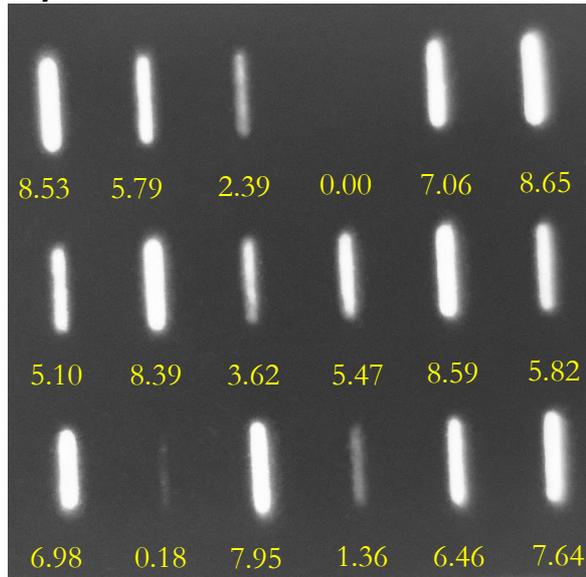


Appendix 21.1 Slot Blot 1 of the RNAi helicase screen. a) The LI-COR Odyssey® FC Imaging System was used to visualize the first slot blot. Similar to the G:BOX machine, the membrane was placed into the chamber of the LI-COR Odyssey® FC Imaging System while submerged under ECL. Unlike the G:BOX machine, this system has a build-in quantification function that quantifies the signal strength (blue number with 3 decimal places). The RNAi bacteria used are labelled next to each blot. 400ng of DNA was loaded onto each slot.

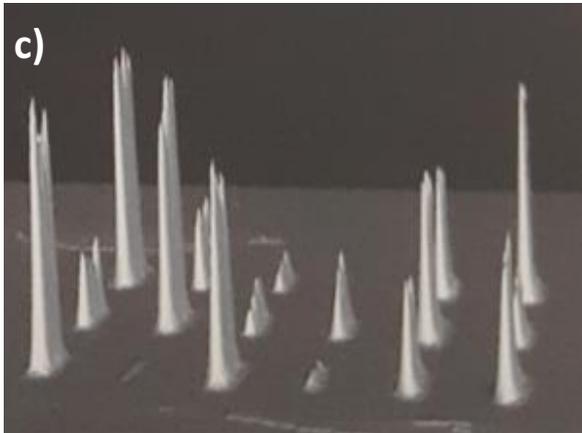
a)

EV (N2)	<i>chd-3</i>	<i>mus-81</i>	<i>mus-81*</i>	<i>rad-54</i>	<i>rad-54*</i>
EV (<i>set-2(bn129)</i>)	<i>isw-1</i>	<i>eri-7</i>	<i>eri-7*</i>	<i>F59H6.5</i>	<i>F59H6.5*</i>
<i>mog-5</i>	<i>rha-1</i>	<i>him-6</i>	<i>him-6*</i>	<i>dog-1</i>	<i>dog-1*</i>

b)



c)

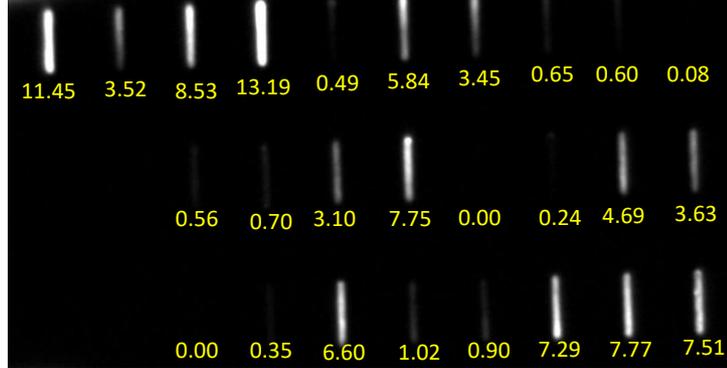


Appendix 21.2 Slot Blot 2 of the RNAi helicase screen. a) Schema showing the position where the samples were loaded onto. Stars indicate biological replicates. The *mog-5* and *isw-1* is the combination of two biological replicates, due to the low recovery of samples. b) R-loop signal as captured by the G:BOX machine. Yellow numbers show the signal intensity as determined by ImageJ. c) Signal intensity “quantification” by the G:BOX system. 400ng of DNA was loaded onto each slot.

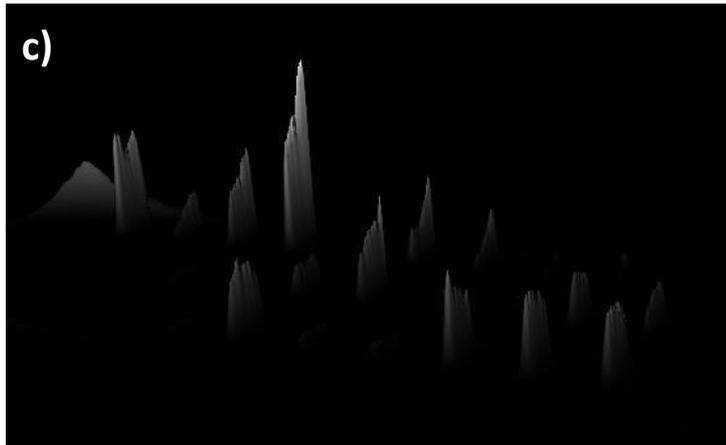
a)

EV (N2)	EV (set-2 (bn129))	ZK250.9	ZK250.9*	<i>xpf-1</i>	<i>xpf-1</i> *	C46F11.4	C46F11.04*	<i>xpb-1</i>	<i>xpb-1</i> *
		<i>rcq-5</i>	<i>rcq-5</i> *	<i>ddx-15</i>	<i>ddx-15</i> *	<i>mtr-4</i>	<i>mtr-4</i> *	Y116A8C.13b	Y116A8C.13b*
		Y54E2A 4.c	Y54E2A 4.c*	<i>polq-1</i>	<i>polq-1</i> *	<i>ssl-1</i>	<i>ssl-1</i> *	F54E12.2	F54E12.2*

b)



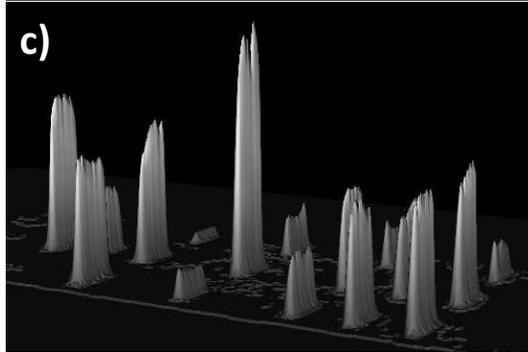
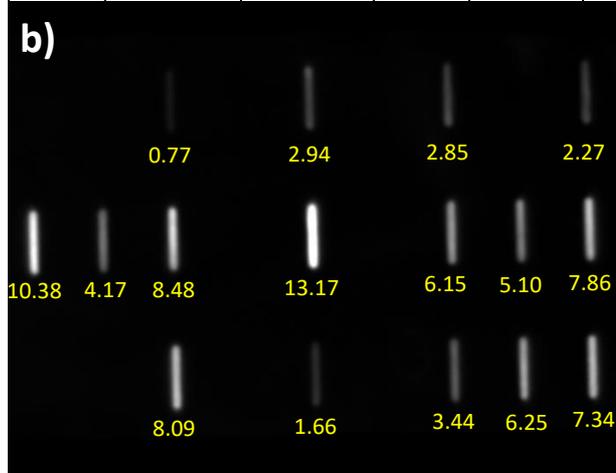
c)



Appendix 21.3 Slot Blot 3 of the RNAi helicase screen. a) Schema showing the position where the samples were loaded onto. Stars indicate biological replicates. The two EV control samples came from the same sample as appendix 21.2. b) R-loop signal as captured by the G:BOX machine. Yellow numbers show the signal intensity as determined by ImageJ. d) Signal intensity “quantification” by the G:BOX system. 300ng of DNA was loaded onto each slot.

a)

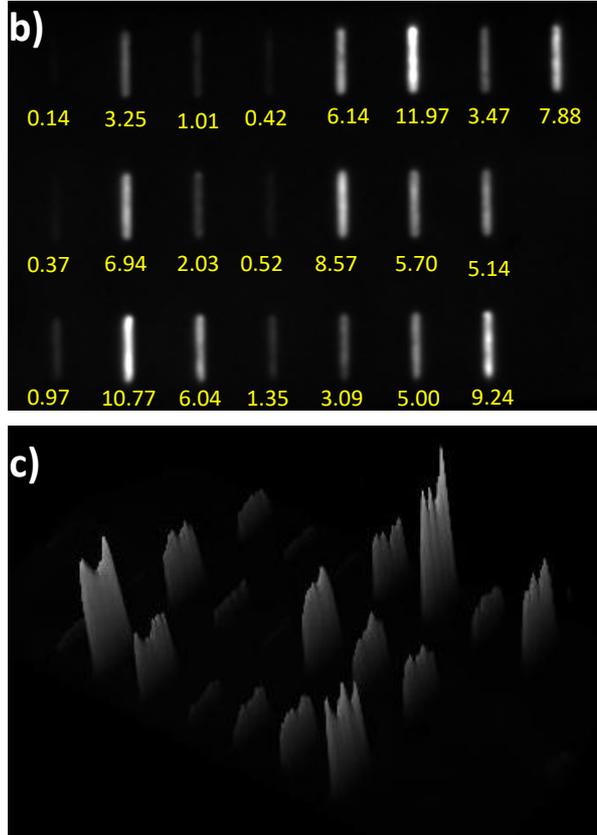
		<i>T23H2.3</i>		<i>rha-2</i>		<i>T05A12.4</i>		<i>helq-1</i>
EV (N2)	EV(set-2 (bn129))	<i>glh-1</i>		<i>wrn-1</i>		<i>F33H12.6</i>	<i>F33H12.6*</i>	<i>vbh-1</i>
		<i>glh-2</i>		<i>dna-2</i>		<i>C24H12.4d</i>	<i>C24H12.4d*</i>	<i>F52B5.3</i>



Appendix 21.4 Slot Blot 4 of the RNAi helicase screen. a) Schema showing the position where the samples were loaded onto. Stars indicate biological replicates. The two EV control samples came from the same sample as appendix 21.2. b) R-loop signal as captured by the G:BOX machine. Yellow numbers show the signal intensity as determined by ImageJ. c) Signal intensity “quantification” by the G:BOX system. 400ng of DNA was loaded onto each slot.

a)

EV (<i>set-2(bn129)</i>) (600ng)	EV (<i>set-2(bn129)</i>) 600ng	EV (<i>set-2(bn129)</i>) 300ng	EV (<i>set-2(bn129)</i>) 150ng	<i>T05A12.4*</i>	<i>T05A12.4*</i>	<i>glh-2</i>	<i>glh-2*</i>
EV (<i>cfp-1(tm6369)</i>) (600ng)	EV (<i>cfp-1(tm6369)</i>) 600ng	EV (<i>cfp-1(tm6369)</i>) 300ng	EV (<i>cfp-1(tm6369)</i>) 150ng	<i>F52B5.3</i>	<i>F52B5.3*</i>	<i>wrn-1</i>	
EV (<i>N2</i>) (600ng)	EV (<i>N2</i>) 600ng	EV (<i>N2</i>) 300ng	EV (<i>N2</i>) 150ng	<i>glh-1</i>	<i>glh-1*</i>	<i>vbh-1</i>	

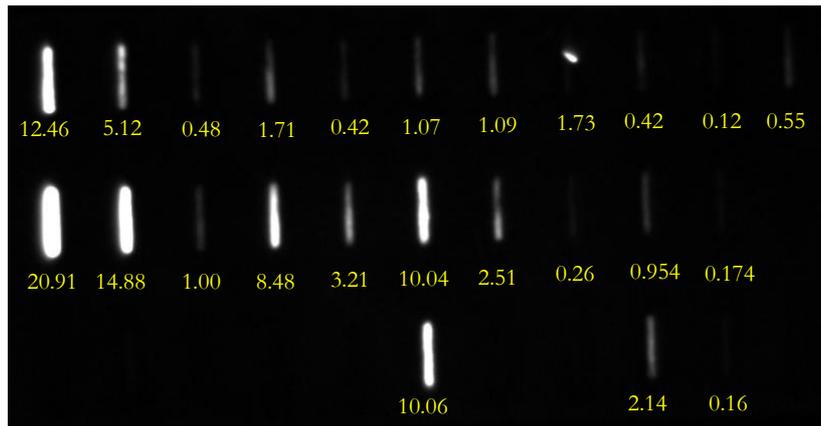


Appendix 21.5 Slot Blot 5 of the RNAi helicase screen. a) Schema showing the position where the samples were loaded onto. Stars indicate biological replicates. New EV controls were made as the previous samples were insufficient for the experiment. RNaseH treated samples are highlighted in grey. b) R-loop signal as captured by the G:BOX machine. Yellow numbers show the signal intensity as determined by ImageJ. c) Signal intensity “quantification” by the G:BOX system. 300ng of DNA was loaded onto each slot unless otherwise stated.

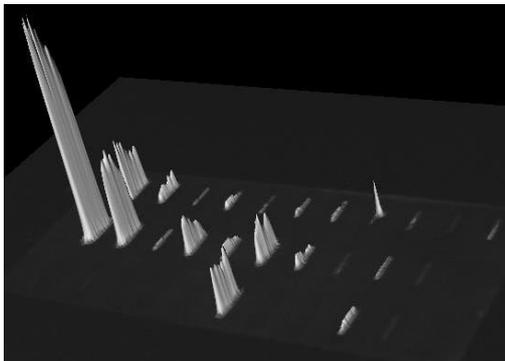
a)

EV (N2) 300ng	EV (N2)* 300ng	EV (<i>cfp-1</i> (<i>tm6369</i>) 300ng	EV (<i>set-2</i> (<i>bn129</i>) 300ng	EV (<i>set-2</i> (<i>bn129</i>))* 150ng	<i>vbh-1</i>	<i>wrn-1</i>	<i>F52B5.3</i>	<i>T05A12.4</i>	<i>glh-2</i>
EV (N2) 600ng	EV (N2)* 600ng	EV (<i>cfp-1</i> (<i>tm6369</i>) 600ng	EV (<i>set-2</i> (<i>bn129</i>) 600ng	EV (<i>set-2</i> (<i>bn129</i>))* 600ng	<i>vbh-1</i> *	<i>wrn-1</i> *	<i>F52B5.3</i> *	<i>T05A12.4</i> *	
EV (N2) (600ng)	EV (N2)* (150ng)	EV (<i>cfp-1</i> (<i>tm6369</i>) (600ng)	EV (<i>set-2</i> (<i>bn129</i>) (600ng)	EV (<i>set-2</i> (<i>bn129</i>))* (600ng)	<i>vbh-1</i> **			<i>T05A12.4</i> **	

b)



c)



Appendix 21.6 Slot Blot 6 of the RNAi helicase screen. a) Schema showing the position where the samples were loaded onto. Stars indicate biological replicates. New EV controls were made as the previous samples were insufficient for the experiment. RNaseH treated samples are shaded grey. b) R-loop signal as captured by the G:BOX machine. Yellow numbers show the signal intensity as determined by ImageJ. c) Signal intensity “quantification” by the G:BOX system. 300ng of DNA was loaded onto each slot.