UNIVERSITE D'AIX-MARSEILLE

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Faculté des Sciences de Luminy

THESE DE DOCTORAT

Spécialité : Bioinformatics

En vue d'obtenir le titre de

DOCTEUR DE L'UNIVERSITE D'AIX-MARSEILLE

Présentée et soutenue publiquement par :

Nishant Thakur

Integrated signaling networks in *C. elegans* innate immunity

Réseaux de signalisation intégrés dans l'immunité innée chez *C. elegans*

Thèse soutenue le 8/09/2016 devant le jury composé de:

Pr. FÉLIX Marie-Anne	Rapporteur
Dr. LEHNER Ben	Rapporteur
Dr. THIERRY-MIEG Nicolas	Examinateur
Pr. VAN HELDEN Jacques	Examinateur
Dr. EWBANK Jonathan	Directeur de thèse
Dr. TICHIT Laurent	Co-directeur de thèse

I dedicate this Ph.D to my parents and relatives.

I thank you all for standing behind me and for supporting me at different phases of my scientific and non-scientific life.

ACKNOWLEDGMENTS

This journey would not have been possible without my family, professors, mentors and friends, a big thanks to everyone for encouraging me in all of my pursuits and inspiring me to follow my dreams. I dedicate this thesis to all of them.

I must thank all the people with whom I have worked in the past 3 years for showing me what it means to be dedicated, each of you have shared your time, energy and expertise with me and I'm richer now.

First and the foremost, I would like to thank my supervisor & mentor **Dr. Jonathan Ewbank** for giving me the opportunity to conduct my Ph.D research in his lab. I am very grateful for his openness to ideas, his insightful suggestions, and expert guidance and for being a perfect mentor. I am thankful for his persistent assistance, encouragement, guidance and the immense support and opportunities he provided throughout my PhD studies. His support and guidance made my thesis work possible.

I thank my 'co-directeur de thèse' **Dr. Laurent Tichit** for his unconditional support and all the encouragement that he gave me during my Ph.D. I am very thankful to him for the fruitful conversations we had throughout my Ph.D. He has been actively interested in my work and has always been available (with a smiling face).

I owe a debt of gratitude to **Dr. Nathalie Pujol** for her time and support throughout my Ph.D. It has been a great honor working with her and I gained a lot biological knowledge through her teachings, both consciously and unconsciously. I also thank you for teaching me different ways to see the biological problems. I appreciate her devoted discussions on different aspects of biological and experimental work that she designed for various projects of my Ph.D work.

Prof. Jacques Van Helden for all the statistical and bioinformatics discussion we had, I got to learn a lot from you & your experiences. You have been an encouraging factor for the betterment of my work and I'm also thankful to you for giving me several unasked opportunities to discuss my naive ideas.

I am especially grateful for the fun group of **JElab**, It has been like a family. I had an excellent time in the lab, both, scientifically and non-scientifically because of the kind help and weird jokes :)

The work I conducted during my Ph.D would not have been possible without the help of my CIML colleagues-cum-friends, who all supported me in different ways: **Ceca** for the delicious vegetarian cake. **LE and Clara** for being excellent siblings away from home, **Shizue** for being a great teacher of experimental techniques and for helping me on experimental part of RNA-Seq project, **Annie** and **Maxime** for doing various fluidigm experiments for the validation of my bioinformatics analysis (Merci Maxime for intensive reading of my thesis), **SongHua**'s management lessons will be useful for rest of my life, **Chris** for making complicated statistical discussion funny and friendly, **Jola** for explaining various biological concepts and being my lab desk neighbor, **Jerome** for smiling all the time, even though I could not understand reason of your everlasting smile, I always enjoyed talking to you.

Additionally special thanks to my other CIML gang: **Lulu, Marci, Stephi, Jesica, Alexis, Lisiena, Laufey, Elli** and **Song yu** who have shared a part of themselves with me through their thoughts and presence.

I also extend my thanks to all the Indian friends, with whom I spent memorable time in Marseille celebrating festivals and getting home coziness.
Thank you **Ankush** for your support & wonderful fun and adventurous time we spent during my stay, **Vikas, Sourabh, Neha & Bala** for your beautiful company and last but not the least **Prashanth** for always being a guiding star and big brother.

List of Abbreviations

ABF	Antibacterial factor
AMP	Antimicrobial peptides
BLAST	Basic local alignment search tool
BN	Bayesian networks
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP-sequencing
CNC	CaeNaCin
CPT	Conditional probability table
CRE	Cis-regulatory elements
CRISPR	Clustered regularly interspaced short palindromic repeats
CRM	Cis-regulatory modules
DAG	Directed acyclic graph
DAPK	Death-associated protein kinase
DBD	DNA-binding domains
dcpm	Depth of coverage per million mapped reads
DNA	Deoxyribonucleic acid
DRSC	Drosophila RNAi Screening Center
DSCAM	Downs syndrome cell adhesion molecule
dsRNA	Double-stranded RNA
ERK	Extracellular-signal-regulated kinases
GCN	general control non-derepressible
GEO	Gene expression omnibus
GFP	Green fluorescent protein
GO	Gene ontology
GPCR	G-protein-coupled receptors

GRN	Gene regulatory network
ILR	InsulinLike receptor
JAK	Janus kinase
JNK	Jun N-terminal kinase
JPD	Joint probability distribution
LPS	Lipopolysaccharide
LRR	Leucine rich repeat
MAMP	Microorganism-associated molecular patterns
MAPK	Mitogen-activated protein kinase
mirRNA	Micro Ribonucleic acid
modENCODE	model organism ENCyclopedia Of DNA Elements
mRNA	Messenger ribonucleic acid
mtDNA	Mitochondrial deoxyribonucleic acid
NGD	No-go Mediated mRNA decay pathway
NLP	Neuropeptidelike proteins
NMD	Nonsense-mediated mRNA decay pathway
NO	Nitric oxide
NSD	Nonstop Mediated mRNA decay pathways
PAMP	Pathogen-associated molecular patterns
PBM	Protein-binding microarray
PCR	Polymerase chain reaction
PGRP	Peptidoglycan recognition proteins
PKC	Protein kinase C
PLC	Phospholipase C
PPI	Protein-protein interaction
PRR	Pattern-recognition receptors
RAG	Recombination-activating genes

RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal ribonucleic acid
RSAT	Regulatory sequence analysis tools
SAPK	Stress-Activated Protein kinase
SARM	Sterile alpha and HEAT/Armadillo motif
SNR	Signal-to-noise ratio
SPELL	Serial Pattern of Expression Levels Locator
STAT	Signal transducers and activators of transcription
TCR	T-cell receptors
TF	transcription factors
TFBS	Transcription factor binding site
TGF-β	Transforming growth factor
TIR	Toll-Interleukin-1 Receptor
TLR	Toll-like receptors
TOF	Time of flight
TRG	Taxonomically-restricted genes
tRNA	Transfer RNA
TSS	Transcription start sites
UPR	Unfolded protein response
UPRmt	Mitochondrial Unfolded protein response
USCO	Universal single-copy orthologs
YAAT	Yet another analysis tool

Abstract

Integrated signaling networks in innate immunity in C.elegans.

For more than 15 years, *C. elegans* has been successfully used as a model organism for studying innate immunity. *C. elegans* is infected by diverse pathogens, including bacteria, fungi and viruses. Upon fungal infection, *C. elegans* up-regulates the expression of many antimicrobial peptide (AMP) genes. These AMPs provide direct protection against pathogen attack. The main aim of my thesis was to build an integrated gene regulatory network representing the induction of these AMP genes upon infection. To find the main/backbone components of the regulatory network, through a genome-wide RNAi screen (Zugasti et al. 2016), we identified 278 Nipi (for "no induction of antimicrobial peptides after infection") clones that abrogate AMP induction. Using "CloneMapper" (Thakur et al. 2014), we identified 338 target genes for these 278 Nipi clones. We showed that MAPK pathways are central to the induction of AMPs. Bioinformatics analysis revealed a role for the mitochondrial unfolded response (mtUPR) in AMP induction. We subsequently validated the involvement of mtUPR experimentally. We also identified various protein complexes including the major mRNA deadenylases CCR4-NOT as being involved in the induction of AMPs. In parallel, we characterized the transcriptional changes provoked by infection using RNA-Sequencing and identified more than 300 genes that are dynamically up-regulated after infection, including 13 AMPs. Among 50 arbitrary selected up-regulated genes, we validated 48 (96%) genes using Fludigm.

Interestingly, the up-regulation of 7 of these genes was independent of MAPK signaling. We also assayed the expression of these genes upon osmotic stress and found that most of the genes are regulated also by osmotic stress. We are interested in identifying the regulators of these up-regulated genes. To assign functions to genes identified in these high-throughput studies, we developed a functional enrichment tool for *C. elegans* community (MS in preparation). We used this tool to analyse the genome-wide RNAi screen targets and other pathogen-related datasets. Further, with this tool, we did functional enrichment analysis of ChIPseq targets of CEBP-1, a transcription factor linked to the regulation of the innate immune response (Kim et al., submitted). Enriched classes clustered into two groups, one related to development and the other to stress. Finally, to understand better the interaction between host and pathogen, we sequenced, assembled, annotated and analysed the *D. coniospora* genome (Lebrigand et al. 2016). We identified various potential virulence factors in the fungal genome. Overall, through analysis of different kind of data, we have identified many novel components of the *C. elegans* innate immunity pathways that have been assembled into a putative signalling network. This analysis broaden the understanding of the innate immunity of *C. elegans* while at the same time providing various useful tools/resources for the *C. elegans* community.

Réseaux de signalisation intégrés dans l'immunité innée chez C. elegans.

Depuis plus de 15 ans, *C. elegans* a été utilisé avec succès comme un organisme modèle pour l'étude de l'immunité innée. *C. elegans* est infecté par divers agents pathogènes, y compris les bactéries, les champignons et les virus. Lors d'une infection fongique, *C. elegans* régule à la hausse l'expression de nombreux peptides antimicrobiens (AMP) gènes. Ces amplificateurs offrent une protection directe contre les attaques pathogènes. Le principal objectif de ma thèse était de construire un réseau de régulation génique intégré représentant l'induction de ces gènes AMP sur l'infection. Pour trouver les principales composantes / épine dorsale du réseau de réglementation, à travers un écran d'ARNi du génome entier (Zugasti et al. 2016), nous avons identifié 278 Nipi (pour "pas d'induction de peptides antimicrobiens après l'infection") clones qui abrogent AMP induction. Utilisation de "CloneMapper" (Thakur et al. 2014), nous avons identifié 338 gènes cibles pour ces 278 clones Nipi. Nous avons montré que les voies de MAPK sont au cœur de l'induction de SAP. l'analyse bioinformatique a révélé un rôle pour la réponse déplié mitochondrial (mtUPR) en AMP induction. Nous avons ensuite validé la participation de mtUPR expérimentalement.Nous avons également identifié divers complexes de protéines, y compris le principal ARNm deadenylases CCR4-NOT comme étant impliqués dans l'induction de SAP. En parallèle, nous avons caractérisé les changements de transcription provoquées par une infection à l'aide de l'ARN-séquençage et identifié plus de 300 gènes qui sont dynamiquement régulés à la hausse après l'infection, y compris 13 ampères. Parmi les 50 gènes arbitraires sélectionnés régulés à la hausse, nous avons validé 48 (96%) des gènes en utilisant Fludigm. Fait intéressant, la régulation positive de 7 de ces gènes est indépendante de la signalisation MAPK. Nous avons également dosé l'expression de ces gènes sur le stress osmotique et constaté que la plupart des gènes sont régulés aussi par le stress osmotique. Nous sommes intéressés à identifier les régulateurs de ces gènes régulés à la hausse. Pour attribuer des fonctions aux gènes identifiés dans ces études à haut débit, nous avons développé un outil d'enrichissement fonctionnel pour la communauté *C. elegans* (MS en préparation).Nous avons utilisé cet outil pour analyser les cibles de l'écran d'ARNi du génome entier et d'autres ensembles de données liées à des agents pathogènes. En outre, avec cet outil, nous avons fait une analyse de l'enrichissement fonctionnel des cibles ChIPseq de CEBP-1, un facteur de transcription lié à la régulation de la réponse immunitaire innée (Kim et al., Soumis). des classes enrichies regroupées en deux groupes, l'un lié au développement et à l'autre au stress. Enfin, pour mieux comprendre l'interaction entre l'hôte et l'agent pathogène, nous avons séquencé, assemblé, annoté et analysé le génome *coniospora D.* (Lebrigand et al. 2016). Nous avons identifié plusieurs facteurs de virulence potentiels dans le génome fongique.

Dans l'ensemble, grâce à l'analyse de différents types de données, nous avons identifié de nombreux nouveaux composants des *C. elegans* voies de l'immunité innée qui ont été assemblés en un réseau de signalisation putatif. Cette analyse élargir la compréhension de l'immunité innée de *C. elegans* tout en fournissant en même temps divers utiles outils / ressources pour la communauté *C. elegans*.

Contents

List of Figures

8

Chapter 1

Introduction

In this chapter, I will start by briefly discussing the innate and the adaptive immunity in general and then will discuss innate immunity in more details, focusing mainly on the invertebrates, particularly on the current understanding of innate immune defences in the model animal *Caenorhabditis elegans*. I will then go on to describe specifically how upon fungal infection, antimicrobial peptides (AMP)s are up-regulated in *C. elegans* Figure 1.1. The aim of my thesis was to build an integrated regulatory network from multiple kinds of functional and phenotypic data to explain the induction of these defence AMPs Figure 1.2. I will therefore also introduce the basics of gene regulation, the concept of RNAi and genome-wide RNAi screens, genome-wide transcriptome analysis, ChIP-seq and protein-protein interactions. I will also talk about the gene regulatory network and high- throughput data analysis. I will end by introducing the 5 publications that form the results section.

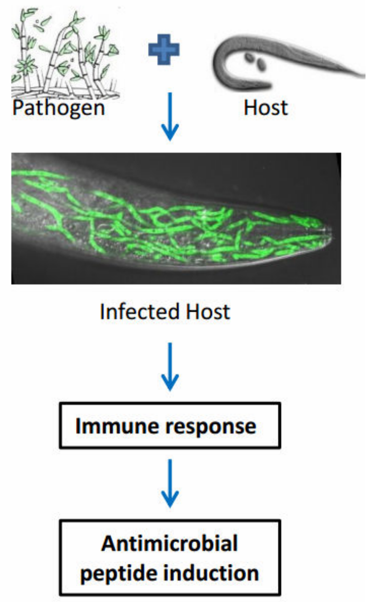
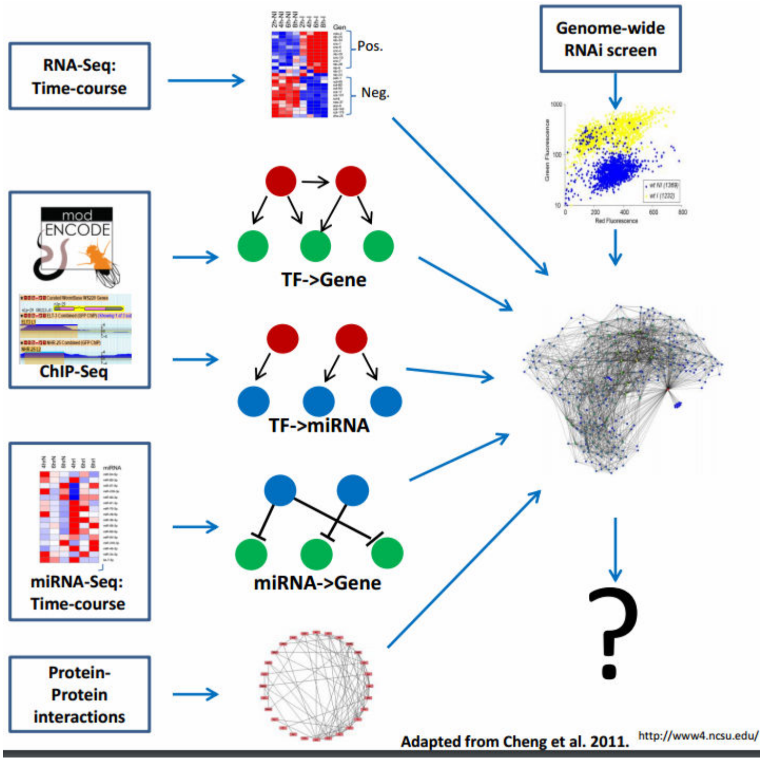


Figure 1.1: Immune response of *C. elegans* against a pathogen. Infection of the host (here *C. elegans*) by a fungal pathogen (*Drechmeria coniospora*), provokes the production of antimicrobial peptides.



Adapted from Cheng et al. 2011. http://www4.ncsu.edu/

Figure 1.2: Workflow of integrated regulatory network construction from functional and phenotypic data.

1.1 Vertebrate and invertebrate immune systems

Since the emergence of life on earth, organisms have needed to react to abiotic and biotic insults that they encounter in the environment. The defensive mechanisms an organism uses to protect itself from infection are collectively called immunity (Hoffmann et al., 1999). Immune defences have been identified in bacteria and very recently even in viruses (Levasseur et al., 2016). In bacteria, the restriction enzymes and the CRISPR/Cas system provide resistance to foreign genetic elements (Arber and Linn, 1969; Barrangou et al., 2007). Similar CRISPR/Cas type systems have been recently identified in Mimiviruses that infect amoebae (Levasseur et al., 2016). Both systems provide sequence-specific recognition of foreign nucleic acids, using very few protein and/or nucleic acid components. At the opposite end of the scale, as far as complexity is concerned, jawed vertebrates possess the adaptive immune system. These are characterized by their slow but highly specific response to invading pathogens. The specificity of the adaptive immune response is mediated by selection from a large repertoire of lymphocytes bearing antigen-specific receptors (T-cell receptors (TCRs) on T-cells and immunoglobulins (Ig) on B-cells) that are generated by somatic gene rearrangement in a process called V(D)J recombination. Lymphocytes specific enzymes (recombination-activating genes (RAGs)), encoded by rag-1 and rag-2 help in the rearrangement and recombination of the immunoglobulin and T-cell receptor (TCR) genes (Jones and Gellert, 2004; Huang et al., 2016). Vertebrates can generate as many as 10 billion different antibodies against different antigens. This diverse plethora of antibodies is generated by shuffling, cutting and recombining a few hundred genes to create a huge number of permutations by VDJ recombination using the RAGs genes. The adaptive immune response becomes more intense over time through expansion of B- and T-cell clones expressing high-affinity antibodies and TCRs. Even when an infection has been cleared, a subset of the pathogen-specific cells remain. This provides an immunological memory and the capacity to mount an immune response more rapidly if the host is infected by the same pathogen a second time. Invertebrates lack adaptive immunity; major players such as RAGs, B and T lymphocytes are absent. On the other hand, a number of novel adaptive and memory-like responses have been proposed in invertebrates based on a variety of molecular mechanisms including, 1) Fibrinogen-related proteins: fibrinogen and one or two immunoglobulin superfamily domain-containing proteins which responds to trematode parasites in snails (Zhang et al., 2004; Gordy et al., 2015). 2) Downs syndrome cell adhesion molecule (DSCAM): a single gene can give rise to as many as 38000 protein isoforms through alternative splicing (Neves et al., 2004) and has been associated with phagocytosis of bacteria (Watson et al., 2005). 3) Immune priming, in certain arthropods, memory immune responses have been reported: a higher immune response is found when a host encounters the same pathogen second time (Arala-Chaves and Sequeira, 2000; Hildemann et al., 1980b,a; Bigger et al., 1983; Salter-Cid and Bigger, 1991; Hartman and Karp, 1989). These alternate immune mechanisms are not found universally. Two constant aspects of invertebrate innate immunity, however, are the role of pattern recognition receptors (PRRs), and the production of antimicrobial factors.

1.1.1 Pattern recognition

Hosts recognize different molecules/components including peptidoglycans, lipopolysaccharides, or flagellins that are expressed by a wide range of bacteria. Janway coined the term pathogen-associated molecular patterns (PAMPs) for these different molecules (Janeway, 1989). This term was popularized by (Medzhitov and Janeway, 1997). Later, it was replaced by microorganism-associated molecular patterns (MAMPs) (Koropatnick et al., 2004; Ausubel, 2005), because these molecules are not limited to the pathogenic microbes and are also found in non-pathogenic species. These PAMPs/MAMPs are recognised by pattern-recognition receptors (PRRs), including the best-known examples, Toll-like receptors (TLRs) in vertebrates and peptidoglycan recognition proteins (PGRPs)s in insects (Akira and Takeda, 2004; Lemaitre et al., 1996). They activate different signalling pathways, leading to phagocytosis, or induction of humoral immune responses (see below). Because most animal hosts encounter MAMPs from both pathogenic and non-pathogenic microorganisms, a more advanced level of regulation are required to discriminate beneficial and pathogenic microorganisms (Gewirtz et al., 2001; Vance et al., 2009a). For example, many PAMPs/MAMPs are recognized by multiple different sensors, often in different contexts, Figure 1.3. To cite one case, flagellin in human intestinal epithelium, the component of bacterial flagella, is a virulence factor that is recognized by the innate immune system. Flagellin, the principal component of bacterial flagella, is a virulence factor in as much as it is required for bacterial movement, but it is found in both pathogenic and commensal bacteria. It is recognised by the innate immune system but is proinflammatory only if it comes in contact with basolateral epithelial surfaces otherwise there is no effect (Gewirtz et al., 2001). Only pathogenic bacteria like *Salmonella* can translocate flagellin across epithelia, which leads to epithelial proinflammatory gene expression, whereas commensal *E. coli* cannot. Additionally, flagellin is recognised in different cellular compartments by different receptors. At the cell surface, flagellin is a ligand for TLR5, whereas in cytosol it has an entirely distinct sensor, the Naip5/Ipaf inflammasome. In a similar manner, DNA and RNA are recognized by different receptors in different subcellular compartments, allowing the innate immune response to be tailored to different pathogens Figure 1.3.

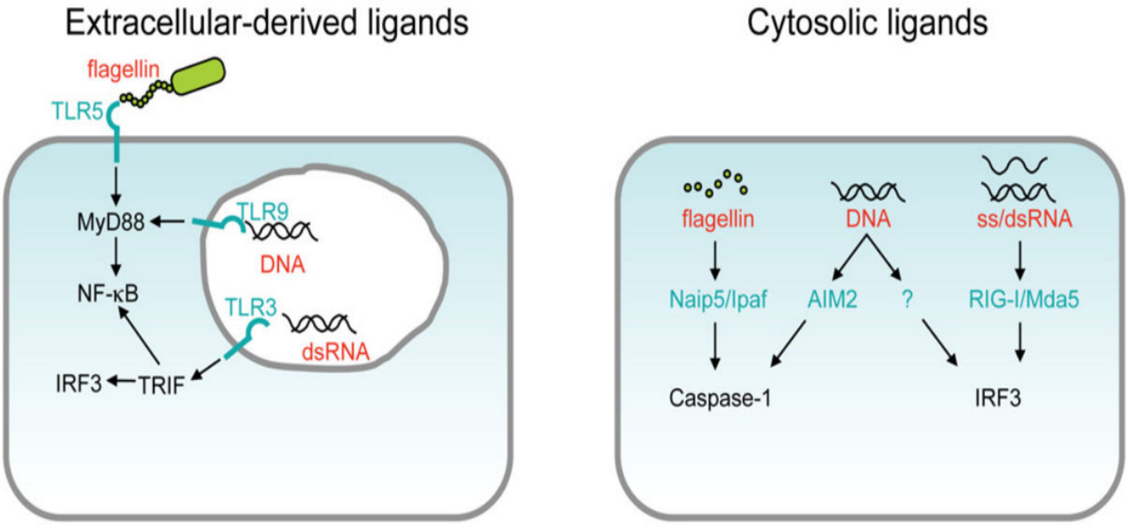


Figure 1.3: Dual recognition of MAMPs: A highly simplified schematic of the responses to three MAMPs is shown. MAMPs are often sensed in different subcellular compartments, or in different cell types, leading to distinct responses. Thus, host cells not only sense whether a MAMP is present but also sense when and where it is. Figure is taken from (Vance et al., 2009b).

1.1.2 Phagocytic cells (cellular immune response)

Most multicellular animals have phagocytic cells which ingest harmful pathogens, foreign particles and dying or dead host cells. This phenomenon of eating is called phagocytosis. In many animals, these cells, macrophages in vertebrates, hemocytes in insects, are circulating and are involved in a broad range of functions. They are key to the immune defences in many species (Stuart and Ezekowitz, 2008). In *C. elegans* there are no professional phagocytic cells; the task of phagocytosis of dying cells from the body is handled by the neighbors surrounding the dying cell. Phagocytosis is not thought to play a role in innate immunity in *C. elegans* (Kim and Ewbank, 2015), and will not be discussed further here.

1.1.3 Humoral immune response

Apart from cellular immune response due to phagocytic cells, animals have a humoral defence system evoked above, which includes the production of AMPs, which lyses cells and the reactive oxygen species (ROS), that oxidizes lipids and proteins. These mechanisms have been highly conserved throughout the evolution; AMPs and ROS are the main immune effectors in both vertebrates and invertebrates (Dowling and Simmons, 2009; Ausubel, 2005). AMP-based innate immunity is therefore a common first line of defense against invading pathogens. AMPs are generally short (<100 amino acids), amphipathic molecules that exhibit broad spectrum antimicrobial activity, against the Gram-positive and Gram-negative bacteria, fungi, and viruses (Bahar and Ren, 2013; Pasupuleti et al., 2012). AMPs are inducible, quickly synthesized, made available shortly after an infection. Either they lyse the microbes or stop the growth of microorganisms. The universal presence of AMPs over the evolutionary scale shows AMPs effectiveness and importance in combating invading pathogens. AMPs are ancient in origin but continue to be effective in killing microbes. This has led to their suggested use as a new class of antibiotics. Indeed, at least 15 peptides are already in clinical trials (Fjell et al., 2012). Many resources for such peptides are available online

CAMP (Waghu et al., 2014), APD (Wang and of Computer Science, 2003), DAMPD (Sundararajan et al., 2012), YADAMP (Piotto et al., 2012), PhytAMP (Hammami et al., 2009), RAPD (Li and Chen, 2008), etc. As described further below, various AMPs have been identified in *C. elegans*.

1.2 *C. elegans* immune system

The genetic tractability of *C. elegans* and the range of known pathogens make it a good model system for investigating innate immunity (Ewbank, 2006). *C. elegans* has been shown to be infected by various pathogens, including natural viruses (Félix et al., 2011; Sarkies et al., 2013) and microsporidia (Troemel et al., 2008), and many bacteria (Ewbank, 2002; Tan, 2002; Alegado et al., 2003; Kurz and Ewbank, 2007; Powell and Ausubel, 2008). *C. elegans* can be infected via the cuticle and epidermis (by fungi; *Drechmeria coniospora* (Jansson, 1994) and species of *Haptocillium* (Barron, 1977)), the uterus (*Leucobacter chromiireducens* Muir and Tan (2008) or the rectum (Hodgkin et al., 2000). Unlike other higher organisms with specific immune cells, no dedicated immune cells are found in *C. elegans*; different tissues play an important role in its immunity.

1.2.1 Strategies/mechanisms of defense

There are various strategies/mechanisms of defense that *C. elegans* deploy against the pathogens, including avoidance behaviour. *C. elegans* can differentiate between pathogenic and non-pathogenic microbes and avoid potential dangerous pathogens (Schulenburg and Ewbank, 2007). This avoidance behaviour involves G protein-coupled chemoreceptors and a G protein-signalling network. There is an indirect role for TLR signaling, since the unique nematode TLR, TOL-1, is required for the terminal differentiation of CO2-sensing neurons that modulate the interaction with pathogens (Brandt and Ringstad, 2015). Another mode of protection in *C. elegans* is physical barriers; the collagen- and the cuticlin-rich cuticle forms an impervious exoskeleton. *C. elegans* is bacteriovorus and eats bacteria from the environment that could be harmful (Labrousse et al., 2000; Kim et al., 2002). To avoid such pathogenic live bacteria getting into the gut, *C. elegans* have a chitin-rich pharyngeal grinder that mechanically disrupts bacteria before they enter the gut. Finally, an important way of protection is inducible defense mechanisms, where *C. elegans* induces the expression of AMP and defense proteins in a pathogen and tissue-specific way.

1.2.2 Pathogen recognition

To fight the pathogen, the first step is to identify or recognize the pathogen. As already mentioned above, this step is carried out by MAMPs. While TOL-1 plays a role in pathogen avoidance behaviour, apart from one study (Tenor and Aballay, 2008), TOL-1 has not been shown to play a role in pathogen resistance or induction of the immune effectors reviewed in (Ewbank and Pujol, 2015). TLRs are characterised by the presence of multiple leucine-rich repeat (LRR) domains in their extracellular part. Other families of PRR that contain LRR domains are involved in pathogen recognition in plant and animals. FSHR-1, an LRR domain containing GPCR in *C. elegans* has been shown to play role in a bacterial recognition (Powell et al., 2009), resistance to oxidative stress (Miller et al., 2015) and acute death from longer cold shock (Robinson and Powell, 2016). This functional pleiotropy suggests that it may not act as a genuine

PRR. Another family of proteins that plays a role in pathogen recognition is C-type lectins which bind carbohydrates. This family of proteins is known to be important in fungal pathogen recognition in vertebrates (Palm and Medzhitov, 2009), but is not known to have such a role in *C. elegans*.

1.2.3 Signaling pathways

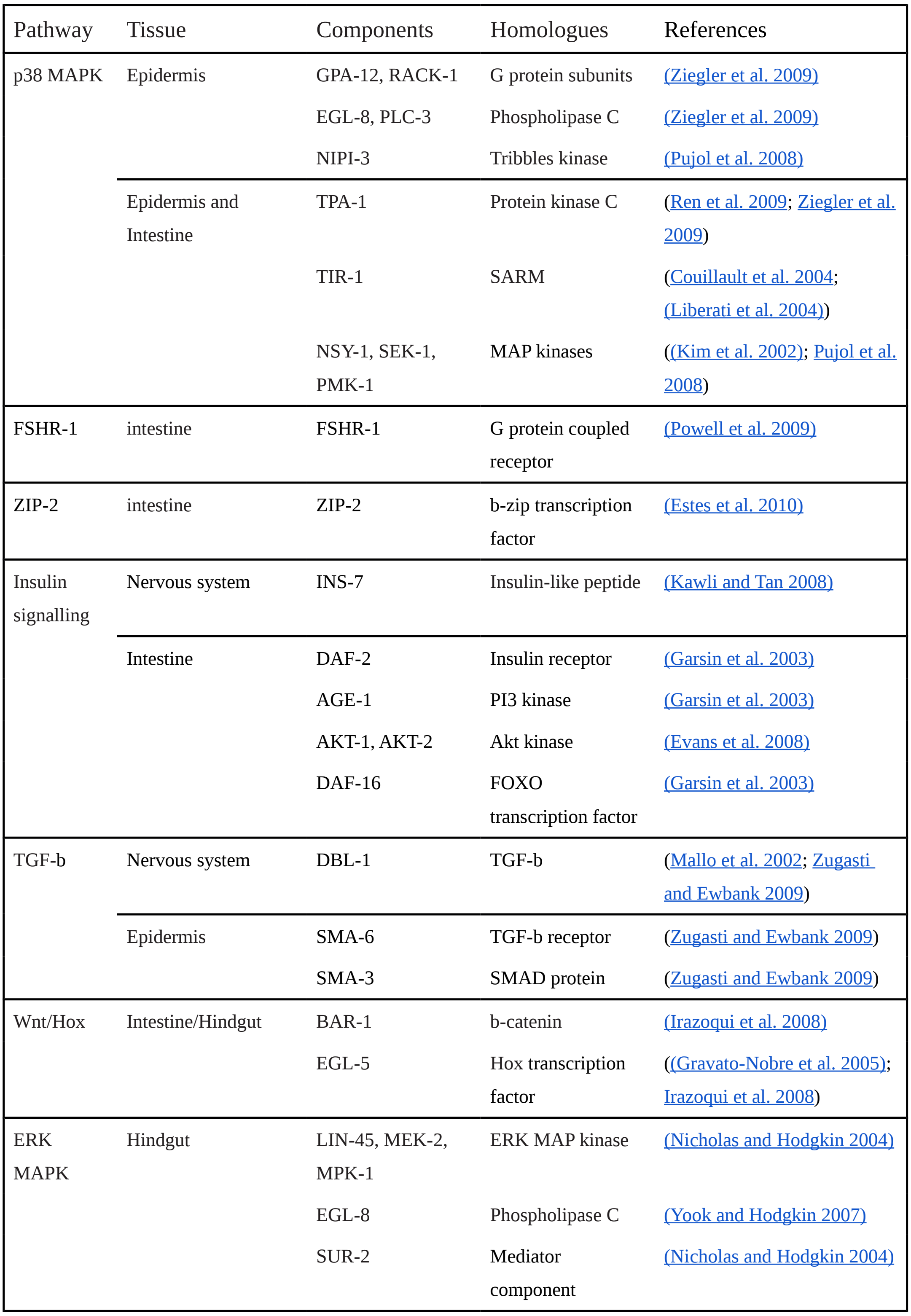
Many signaling pathways have been identified as being important for the immune response in *C. elegans*, Table 1 (reviewed in (Engelmann and Pujol, 2010; Kim and Ewbank, 2015)). Few of them are pathogen-specific but are important for the immune response against many pathogens. Here, we will focus only on the main signaling pathways and specifically on these pathways that are involved in the immune response against the fungal pathogen *Drechmeria coniospora* , see Figure 1.8.

1.2.3.1 The MAP kinase pathway

With the help of molecular genetic methods, various conserved immune signaling pathways have been shown to be important for the immune response, including mitogen-activated protein kinase (MAPK) pathways. This pathway is found to have a central role in resistance to microbial pathogens (Zugasti et al., 2016; Bandyopadhyay et al., 2010). The pathway is similar to the Hog1 (high osmolarity glycerol 1) MAPK pathway that was initially identified in yeast where it is involved in the response to osmotic stress (Brewster et al., 1993). This discovery of Hog1 pathway in yeast was followed by the discovery of the p38 and JNK (c-Jun N-terminal kinase) orthologs of this kinase in mammals (Han et al., 1994; Galcheva-Gargova et al., 1994). These kinase pathways are the main components of the stress-activated protein kinase (SAPK) pathways and play a central role in the stress-activated signaling. p38 and JNK stress-activated MAPKs regulate both innate defenses and responses to abiotic stress (Huang et al., 2009b; Keshet and Seger, 2010). Genetic analysis of *C. elegans* has established ancient, evolutionarily conserved roles for MAPK signaling in innate immunity (Cheesman et al., 2016; Kamaladevi and Balamurugan, 2015; Kim et al., 2002; Zugasti et al., 2016; Ziegler et al., 2009a).

This signaling pathway consists of a cascade of kinases in which the first kinase, a MAPK kinase kinase (MAPKKK or MAP3K) activates a MAPK kinase (MAPKK or MAP2K), and this MAP2K phosphorylates the MAPK. MAPKs includes four subfamilies with distinct upstream kinases to activate the cascade: (1) the extracellular-signal-regulated kinases (ERK) pathway, (2) the ERK5 (BMK1 or MAPK7) pathway, (3) the JUN N-terminal kinases (JNKs) pathway and (4) the p38 kinases pathway (a 38 kDa protein, hence the name), Figure 1.4 (Ashwell, 2006). The ERK pathway is activated by various growth factors and is implicated in regulating the cell cycle. JNK and p38 are linked to various stress responses and can be activated by various environmental factors, such as osmotic shock, oxidative stress and pro-inflammatory cytokines (Hayakawa and Smyth, 2006).

Table 1, Pathways involved in immune response in *C. elegans* (Engelmann and Pujol 2010).



Pathway	Tissue	Components	Homologues	References
p38 MAPK	Epidermis	GPA-12, RACK-1	G protein subunits	(Ziegler et al. 2009)
		EGL-8, PLC-3	Phospholipase C	(Ziegler et al. 2009)
		NIPI-3	Tribbles kinase	(Pujol et al. 2008)
	Epidermis and Intestine	TPA-1	Protein kinase C	(Ren et al. 2009; Ziegler et al. 2009)
		TIR-1	SARM	(Couillault et al. 2004; (Liberati et al. 2004))
		NSY-1, SEK-1, PMK-1	MAP kinases	((Kim et al. 2002); Pujol et al. 2008)
FSHR-1	intestine	FSHR-1	G protein coupled receptor	(Powell et al. 2009)
ZIP-2	intestine	ZIP-2	b-zip transcription factor	(Estes et al. 2010)
Insulin signalling	Nervous system	INS-7	Insulin-like peptide	(Kawli and Tan 2008)
	Intestine	DAF-2	Insulin receptor	(Garsin et al. 2003)
		AGE-1	PI3 kinase	(Garsin et al. 2003)
		AKT-1, AKT-2	Akt kinase	(Evans et al. 2008)
		DAF-16	FOXO transcription factor	(Garsin et al. 2003)
TGF-b	Nervous system	DBL-1	TGF-b	(Mallo et al. 2002; Zugasti and Ewbank 2009)
	Epidermis	SMA-6	TGF-b receptor	(Zugasti and Ewbank 2009)
		SMA-3	SMAD protein	(Zugasti and Ewbank 2009)
Wnt/Hox	Intestine/Hindgut	BAR-1	b-catenin	(Irazoqui et al. 2008)
		EGL-5	Hox transcription factor	((Gravato-Nobre et al. 2005); Irazoqui et al. 2008)
ERK MAPK	Hindgut	LIN-45, MEK-2, MPK-1	ERK MAP kinase	(Nicholas and Hodgkin 2004)
		EGL-8	Phospholipase C	(Yook and Hodgkin 2007)
		SUR-2	Mediator component	(Nicholas and Hodgkin 2004)

The p38 MAPK pathway has been shown to be important in *C. elegans* immunity to various bacterial and fungal pathogens (Kim et al., 2002, 2004; Aballay et al., 2003; Sifri et al., 2003; Pujol et al., 2008b). In these studies, mutants of the p38 MAPK pathway (MAP3K *nsy-1*, MAP2K *sek-1* and MAPK *pmk-1*) have been shown to be more susceptible to various pathogens, Figure 1.5. All studies are consistent with the central role of this pathway in *C. elegans* immunity and numerous other components of *C. elegans* p38 MAPK pathways have been identified. Upon infection, the *C.*

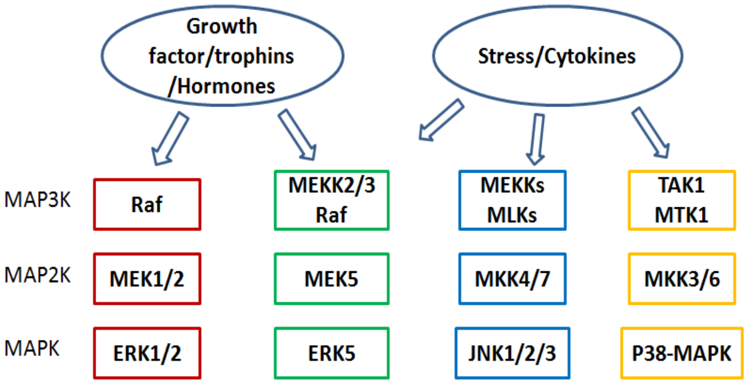


Figure 1.4: Conventional (ERK1/2, JNK, and p38-MAP kinase) and ERK5 MAP kinase signaling pathways in mammalian cells. Only representative signaling molecules are shown. Adapted from (Lu and Xu, 2006).

elegans orthologue of mammalian SARM protein, Toll-Interleukin-1 Receptor (TIR) domain adaptor protein, TIR-1 (Couillault et al., 2004), has been shown to activate the PMK-1 pathway (Liberati et al., 2004). In vertebrates, TIR-domain adaptor proteins act downstream of TOL- 1 TLRs (O'Neill and Bowie, 2007), however, they do not play role in PMK-1 activation, neither in intestinal nor epidermal innate immunity. Other components of the MAPK pathway in *C. elegans* involved in the response to *D. coniospora* include the GPCR DCAR-1 (DihydroCaffeic Acid Receptor). DCAR-1 has been shown to be activated by the tyrosine derivative 4-hydroxyphenyllactic acid (HPLA) (Zugasti et al., 2014). DCAR-1 then activates the G-alpha protein, GPA-12 which acts upstream of the TIR-1/NSY-1/SEK-1/PMK-1 cascade (Zugasti et al., 2014; Ziegler et al., 2009a). This cascade has been shown to activate the STAT-like transcription factor STA-2, which then directly or indirectly induces the Neuropeptide-Like Protein (NLPs) genomic cluster of AMP genes that protect *C. elegans* against *D. coniospora* infection, Figure 1.8.

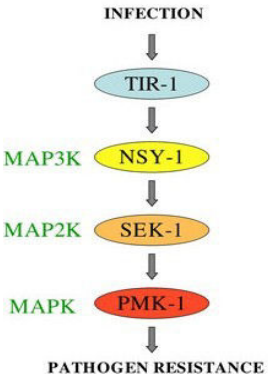


Figure 1.5: The TIR-1/NSY-1/SEK-1/PMK-1 cassette functions in innate immunity. See text for details.

1.2.3.2 TGF-β -like pathway

The transforming growth factor β (TGF-β)-like pathway is another important immune signaling pathway. This pathway is highly conserved in mammals and most of the components of this pathway in *C. elegans* have clear orthologs in mammals. This pathway was originally shown to control developmental processes such as body size, and morphology of the tail of males (Gumienny and Savage-Dunn, 2013)). A TGF-β -like pathway also regulates the expression of certain effector molecules, such as lectins and lysozymes, following infection with *Serratia marcescens* (Mallo et al., 2002) and acts in the MAPK independent regulation of CNC (CaeNaCin (Caenorhabditis bacteriocin)) AMPs upon *D. coniospora* infection (Zugasti and Ewbank, 2009). Upon *D. coniospora* infection, the main TGF-β signaling cascade involves the binding of DBL-1 to the serine/threonine protein kinase receptor SMA-6/DAF-4, which phosphorylates SMA-3, Figure 1.6. The transcription factor STA-2 is also required for the expression of the CNC cluster of AMP genes. SMA-3 is a nuclear protein, known to physically interact with LIN-31, a forkhead transcription factor (Wang and Tissenbaum, 2005). Whether and how SMA-3 and STA-2 interact is not known. In addition to its role in response to *S. marcescens* infection, the TGF-β pathway was shown to be required for resistance to *P. aeruginosa* (Tan, 2001). In contrast to the regulation of cnc genes in the epidermis, which is independent of the SMAD protein genes *sma-2* and *sma-3*, intestinal defences apparently require these 2 genes. Thus, variants of the TGF-β-like pathway represent another conserved signaling pathway required for the response of *C. elegans* bacterial and fungal Figure 1.8 infections.

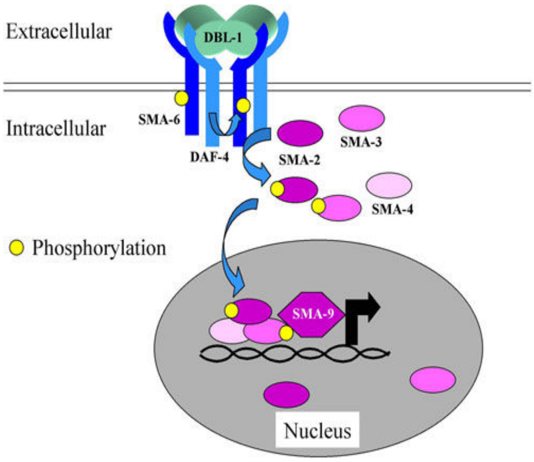


Figure 1.6: TGF- signaling pathway. While in the developmental context, this pathway involves the co-SMADs SMA-2 and SMA-4, as well as the co-activator SMA-9, these are dispensable for the regulation of cnc genes (Zugasti and Ewbank, 2009).

1.2.3.3 DAF-2/InsulinLike Receptor (ILR) pathway

This pathway involves DAF-16, a FOXO family transcription factor that has been shown to be important for the immune response of *C. elegans* to *P. aeruginosa* (Garsin et al., 2003; Singh and Aballay, 2009). When the DAF-2 receptor is active, DAF-16/FOXO is retained in the cytoplasm, whereas, in a *daf-2* mutant, DAF-16 moves from the cytoplasm to the nucleus and regulates DAF-16-dependent gene expression. This promotes longevity and increased resistance to infection by several bacteria. There are four known serine threonine kinases downstream of DAF-2, PDK-1, SGK-1, AKT-1 and AKT-2, Figure 1.7. Only *akt-1* and *akt-2* mutants are reported to be more resistant to infection by *P. aeruginosa.* Although DAF-16 plays a role in modulating the expression of antimicrobial proteins under steady-state conditions (Murphy et al., 2003), so far, there is no direct evidence that DAF-16 is required for AMP gene induction after infection (Engelmann and Pujol, 2010). Other studies, on the other hand, have shown DAF-16 dependent regulation of many stress response genes (Alper et al., 2007). This pathway seems to be involved in general stress responsiveness rather than being a specific regulator of the immune response.

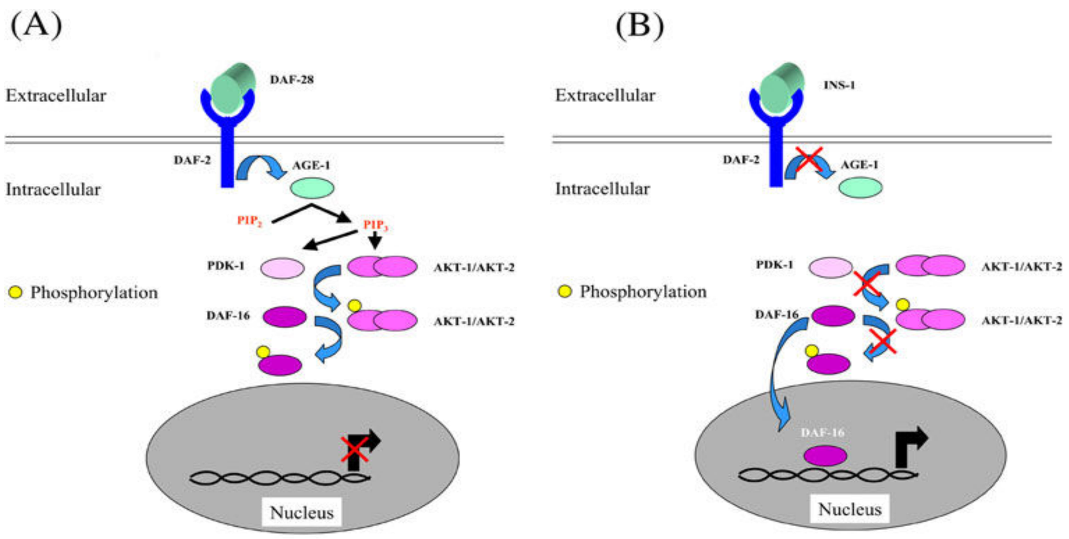


Figure 1.7: The DAF-2/DAF-16 pathway. (A) In the presence of an agonist ligand, such as the insulin-like peptide DAF-28, the DAF-2 receptor is activated and in turn activates the phosphatidylinositol-3 OH kinase AGE-1 that catalyses the conversion of phosphatidylinositol bisphosphate (PIP2) into phosphatidylinositol trisphosphate (PIP3). On one hand, PIP3 binds to the complex AKT-1/AKT-2 and leads to the exposure of two phosphorylation sites. On the other hand, the kinase PDK-1 by binding to PIP3 is recruited to the membrane where it can phosphorylate and activate AKT-1. The kinase AKT, in turn phosphorylates the transcription factor DAF-16 and thereby ensure its cytoplasmic retention. (B) In the presence of an antagonist ligand such as INS-1, (or in a *daf-2* loss of function mutant), the pathway is not active, DAF-16 is not phosphorylated and can be translocated to the nucleus where it regulates the expression of a set of stress response and antimicrobial genes. Figure and legend taken from (Ewbank, 2006).

1.2.3.4 The Unfolded Protein Response (UPR)/UPRmt

The unfolded protein response (UPR) is a conserved cellular stress response that has been found from yeast to mammals (Janssens et al., 2014). This is an adaptive response to the accumulation of unfolded proteins in the endoplasmic reticulum (UPR^{ER}), cytosol or mitochondria (UPRmt). In *C. elegans*, a protective role of the UPR^{ER} against *B. thuringiensis* has been shown (Bischof et al., 2008). Pore-forming toxins activate UPR response through the p38 MAPK pathway. The UPR is also involved in the immune response to *S. typhimurium* and *P. aeruginosa* (Haskins et al., 2008). Recently, (Pellegrino et al., 2014) have shown a role of mitochondrial UPR in innate immunity against *P. aeruginosa*. The key transcription factor for the UPR is ATFS-1. The authors showed that *P. aeruginosa* provokes an ATFS-1-dependent induction of innate immune genes, including those encoding secreted lysozyme and antimicrobial peptides. ATFS-1 mutant animals were found more susceptible to *P. aeruginosa* infection. These studies clearly point out the important role of UPR in innate immunity against microbes.

1.2.3.5 WNK/GCK signaling

Cells are exposed to osmotic stress due to changes in the intracellular solute levels with a flux of solutes and metabolism. To protect cells during osmotic stress, osmotic homeostasis is required to activate the mechanisms to repair or remove stress-induced damage. In *C. elegans*, osmotic homeostasis is achieved by the accumulation of osmolytes through glycerol 3-phosphate dehydrogenases; *gpdh-1* and *gpdh-2* genes (Lamitina et al., 2006). In mutants such as *dpy-9* and *osm-11* that have elevated levels of *gpdh-1* and increased resistance to osmotic stress, osmotic stress resistance returns to normal when either the nematode Wnk or Ste20/GCK kinase genes (*wnk-1* and *gck-3*, respectively) are inactivated, even though the level of *gpdh-1* remains high (Choe and Strange, 2008). This suggests that osmotic resistance involves more than just controlling osmolyte levels. Osmotic stress also induces a subset of the *nlp* AMP genes, including *nlp-29* (Pujol et al., 2008a), Figure 1.8. While *nlp-29* expression is dependent on the p38 MAPK pathway upon *D. coniospora* infection, it is independent of the p38 MAPK pathway (Pujol et al., 2008a) upon osmotic stress. It requires, however, the *wnk-1/gck-3* pathway, together with *fasn-1*, the nematode ortholog of vertebrate fatty acid synthase, which acts upstream of *wnk-1* and *gck-3* in the *C. elegans* epidermis (Lee et al., 2010), Figure 1.8.

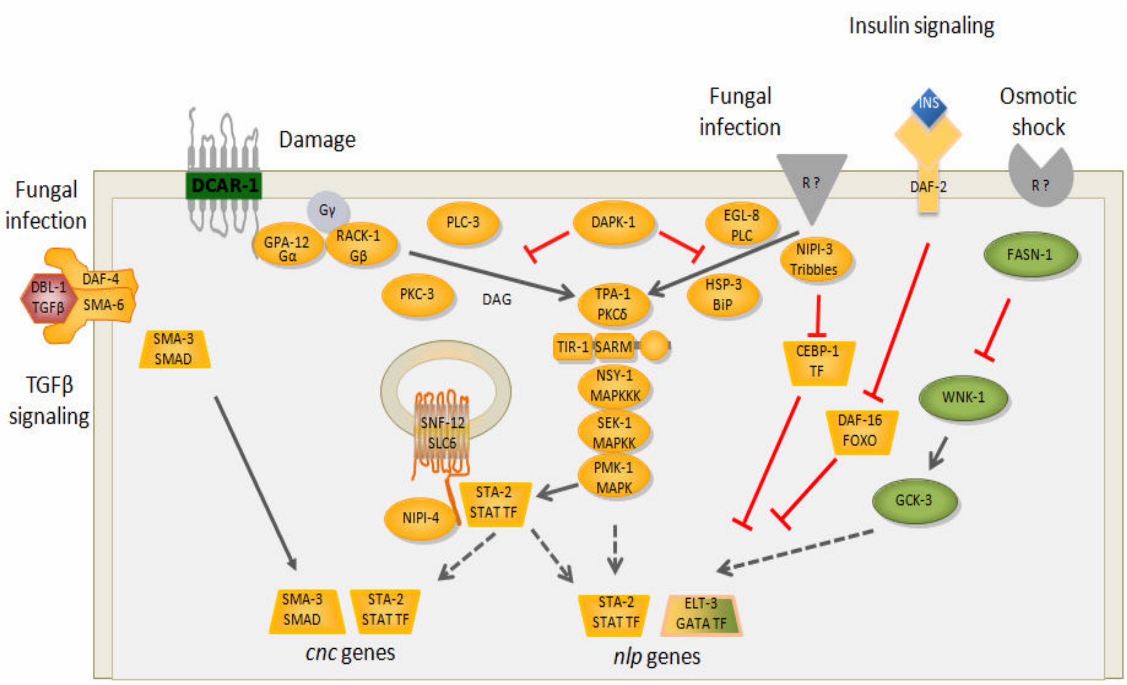


Figure 1.8: Simplified representation of *C. elegans* immune signaling pathways involved in AMP regulation. Although *wnk-1* and *gck-3* are required for the up-regulation of *nlp-29* expression upon infection (Zugasti et al., 2016), the osmotic and infection pathways are separated here for the sake of clarity.

The different signalling pathways mentioned above regulate the expression of diverse effectors that contribute to the immune response to pathogens. In the next section, I will describe in detail the various kind of effectors currently known in *C. elegans*.

1.2.4 *C. elegans* immune effectors

In this section of the introduction, I will discuss about effectors proteins like AMPs, caenopores, lysozymes, lectins, etc., identified in *C. elegans* and potentially or demonstrably important for the host defence against various pathogens. A summary of these effectors, together with their respective regulatory pathways and the tissues in which they are expressed is given in Table 2.

1.2.4.1 Antimicrobial Peptides

As described in humoral immune response section 1.1.3, inducing AMPs is one of the most fundamental defense mechanisms. So far many different families of AMPs have been identified in *C. elegans*. One includes Ascaris suum (antibacterial factor)-type antimicrobial peptide (ABF-1 to ABF-6). These peptides are mollusc defensin/mycitin-like peptides. ABF-1 and 2 and have demonstrated antimicrobial activity against the Gram-positive and Gram-negative bacteria, and yeast (Kato et al., 2002). These peptides are strongly upregulated upon prolonged exposure to *S. typhimurium* (Alegado and Tan, 2008). Another family of AMPs includes those annotated as neuropeptidelike proteins (NLPs) (Pujol et al., 2008b; Zugasti et al., 2014). They are regulated principally upon epidermal fungal infection (e.g. *D. coniospora* or *H. sphaerosporum).* As mentioned above, certain NLPs are also regulated by osmotic stress (Pujol et al., 2008b). Caenacins (CNC) are structurally related to NLPs, but have been described as being more specifically regulated by fungal infection (Zugasti and Ewbank, 2009; Pujol et al., 2008b). The AMP NLP-31 has been shown to have direct antimicrobial activity in-vitro [(Couillault et al., 2004)]. In common with other AMPs it is presumed to act by disrupting cell membrane.

1.2.4.2 Caenopores

Caenopores are a saposin domain containing protein family of 23 members. Many proteins like SPP-1 (Bányai and Patthy, 1998), SPP-5 (Roeder et al., 2010), and SPP-12 (Hoeckendorf et al., 2012) have antibacterial properties and are differentially regulated upon infection. Other member like SPP-3 and SPP-18 are also strongly upregulated upon *P. aeruginosa* infection (Kurz and Tan, 2004). Overall, 19 out of the 23 family members have been found to be upregulated by at least one pathogen (Dierking et al., 2016). Since the caenopores(as well as AMPs)have the potential to damage host tissues, their activity needs to be tightly regulated [(Lebrigand et al., 2016)].

1.2.4.3 Lysozymes

Lysozymes are a class of molecules known to be involved in digestion and immunity in a variety of organisms, ranging from bacteria to vertebrates (Jollès, 1996). In *C. elegans* 16 lysozyme genes are found, ten most closely related to the protists lysozymes (encoded by the *lys* genes), while the remaining six are specific to invertebrate (*ilys*). Animals that overexpress *lys-1* have an increased resistance against pathogenic *S. marcescens* (Mallo et al., 2002). Similarly, *lys-4* and *lys-5* contribute to resistance against *S. aureus* (Irazoqui et al., 2010), *lys-5* and *lys-7* against *B. thuringiensis* (Boehnisch et al., 2011) and *lys-7* against *M. nematophilum* infection (O'Rourke et al., 2006). A total of 11 out of these 16, lysozyme genes were found to be upregulated by at least one pathogen (Dierking et al., 2016).

1.2.4.4 Lectins

Lectins are carbohydrate-binding proteins, macromolecules that are highly specific for sugar moieties. Lectins in general and C-type lectins, in particular, have been implicated in innate immunity of diverse species (van den Berg et al., 2012). Lectins can be involved either in pathogen recognition or immune effector functions. The *C. elegans* genome encodes a total of 276 lectins, 11 galectins and 265 C-type lectins. (Miltsch et al., 2014) showed that recombinant CLEC-39 and CLEC-49 can directly bind *S. marcescens* bacteria. Differential up-regulation of these genes has led to the suggestion that they might be an element conferring specificity to the immune response of *C. elegans* (Schulenburg et al., 2008; Wong et al., 2007; O'Rourke et al., 2006). Although, LEC-8 has been shown to play a role in host defence against *B. thuringiensis* infection by competitively inhibiting the binding of the toxin Cry5B to its host glycolipid receptor (Ideo et al., 2009), the exact function of C-type lectins proteins in *C. elegans* immunity is still unclear.

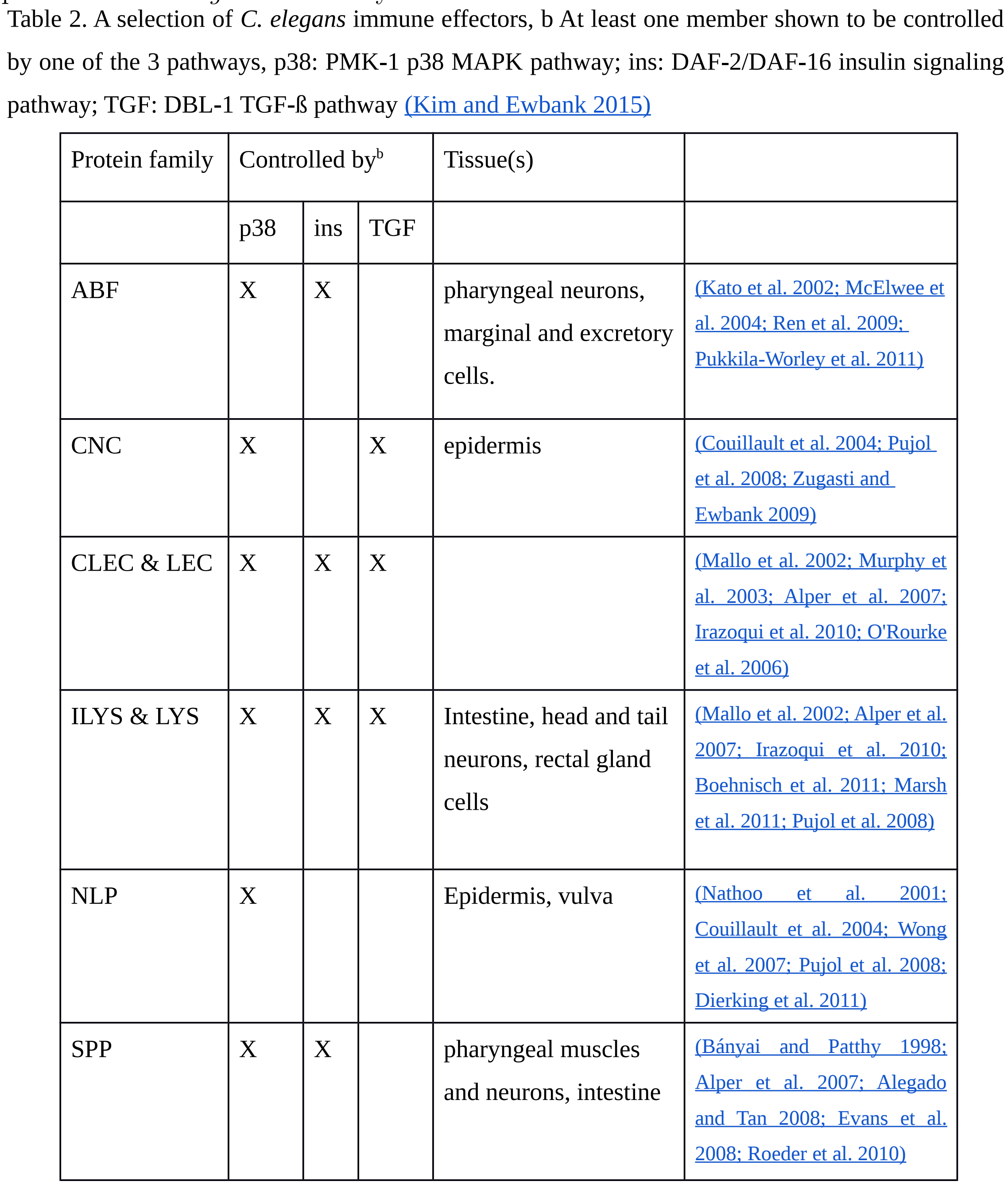


Table 2. A selection of *C. elegans* immune effectors, b At least one member shown to be controlled by one of the 3 pathways, p38: PMK-1 p38 MAPK pathway; ins: DAF-2/DAF-16 insulin signaling pathway; TGF: DBL-1 TGF-ß pathway (Kim and Ewbank 2015)

Protein family	Controlled by^{b}			Tissue(s)	
	p38	ins	TGF		
ABF	X	X		pharyngeal neurons, marginal and excretory cells.	(Kato et al. 2002; McElwee et al. 2004; Ren et al. 2009; Pukkila-Worley et al. 2011)
CNC	X		X	epidermis	(Couillault et al. 2004; Pujol et al. 2008; Zugasti and Ewbank 2009)
CLEC & LEC	X	X	X		(Mallo et al. 2002; Murphy et al. 2003; Alper et al. 2007; Irazoqui et al. 2010; O'Rourke et al. 2006)
ILYS & LYS	X	X	X	Intestine, head and tail neurons, rectal gland cells	(Mallo et al. 2002; Alper et al. 2007; Irazoqui et al. 2010; Boehnisch et al. 2011; Marsh et al. 2011; Pujol et al. 2008)
NLP	X			Epidermis, vulva	(Nathoo et al. 2001; Couillault et al. 2004; Wong et al. 2007; Pujol et al. 2008; Dierking et al. 2011)
SPP	X	X		pharyngeal muscles and neurons, intestine	(Bányai and Patthy 1998; Alper et al. 2007; Alegado and Tan 2008; Evans et al. 2008; Roeder et al. 2010)

In addition to these different classes of demonstrated or putative immune effectors, infection induces the expression of a large number of lineage-specific genes of unknown function. Some may be involved in signaling, other may possess antimicrobial activity. Understanding their function and regulation is a major challenge for the future.

To understand better the regulation of these different effectors, we need an explanation of the basis of gene regulation. So in the next few sections, I will talk about the central dogma and different modes of gene expression regulation.

1.3 Central dogma

The cell is the basic functional and structural unit in any living organism and they can be of two types, eukaryotic or prokaryotic. While prokaryotes are unicellular organisms without a nucleus, eukaryotic cells generally have a more complex structure and have a well-defined nucleus. Within the cell nucleus, the deoxyribonucleic acid (DNA) that carries the genetic instructions is found. A gene corresponds to a region of DNA and is the molecular unit of heredity. Many genes are transcribed into messenger ribonucleic acid (mRNA) that is then translated into protein. This schema of information flow Figure 1.9A is called the central dogma. All the cells in an organism have same DNA which encodes similar genes, but depending on the function of the cell, regulatory mechanisms limit gene expression such that only a subset of protein-coding genes are transcribed into mRNA.

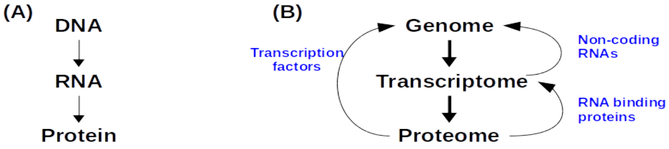


Figure 1.9: Transcription plays a pivotal role in the regulation of gene expression. (A) The central dogma in molecular biology. (B) Different types of networks regulate gene expression: TFs are proteins that control gene expression by interacting with the genome; non-coding RNAs affect both the genome and the transcriptome and RNA binding proteins affect gene expression post-transcriptionally. Adapted from Handbook of systems biology.

1.4 Gene regulation

Gene expression determines the phenotype (observable characteristics) from a given genotype (genetic composition of a cell). Thus two cells with identical genomes can give rise to different cell/tissue phenotypes because of differences in gene expression. Some genes are constitutively expressed, or always ”on,” and these genes are generally important for cell survival. Although their precise level of expression will be modulated, they are also called ”housekeeping genes” as they are needed to keep the main biological processes running. ”Regulated genes”, on the other hand, are condition/stage specific and are needed occasionally, so they are turned on or off conditionally. This modulation of gene expression in multicellular organisms is the primary cause of the diversity that we see at the tissue, organismal or even at the species level. How these genes are regulated has been a significant area of research for several decades. In order to identify the genes that are differentially expressed between two conditions, researchers can use gene expression profiling. A detailed description of different ways to measure the expression profiles is given in section 1.5.4. Beyond the transcriptional regulation, the rate at which the functional proteins are produced is controlled at the translational and post-translational (protein modification, etc.) levels. In what follows, I will deal only with pre-translational regulation.

1.4.1 Transcriptional regulation of gene expression

Transcription starts with the assembly of the pre-initiation complex (RNA polymerase II, transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) at the transcription start site in a step called initiation. This is followed by elongation, where after the recruitment of RNA polymerase (RNA Pol II for eukaryotic protein-coding genes) on DNA, it moves along the DNA to produce an RNA transcript. The step of RNA polymerase II recruitment is the rate-limiting step for transcription (Roeder, 2005). Transcriptional regulation in eukaryotes results from the combined effects of how the transcription factors (TF) interact with each other and on the structural properties of DNA, i.e. how it is ”packaged” at a particular time.

1.4.1.1 Transcription factors

TFs are transcriptional activator and repressor proteins that regulate the expression of target genes generally through binding their promoters, usually non-coding regions 5′of the transcribed gene. TFs contain one or more DNA-binding domains (DBDs) and these domains bind DNA in a sequence-specific manner. Generally, they recognise short DNA motifs adjacent to the genes that they regulate. The number of TFs increases with the size of the genome (van Nimwegen, 2003). TFs comprise almost 5 % (1,000) of the total protein-coding genes (20,000) in the *C. elegans* genome (Reece-Hoyes et al., 2005), whereas in human (Babu et al., 2004) this number is reported to be around 10% (2600). Additional proteins such as coactivators, chromatin remodelers, histone acetylases, deacetylases, kinases, etc. also play crucial roles in gene regulation, but due to their lack of DBDs, they are not classified as transcription factors.

1.4.1.2 DNA chromatin structure/remodelling

In eukaryotes, DNA packaging can control the expression of genes; whether to express or repress a gene is decided by the open or closed state of the chromatin (DNA plus specific proteins). For a gene to be transcribed, chromatin has to be open so that transcriptional regulators can bind to the specific region of the DNA to regulate a specific gene. The basic unit of chromatin structure is the nucleosome, formed by the association of eight histone molecules with about 140 base pairs of DNA repeating units of nucleosomes, collectively called eukaryotic chromatin (Figure 1.10).
Histones are subject to multiple reversible post-translational modifications, including methylation and acetylation on their C-termini. Adding acetyl groups opens up the chromatin to allow gene expression. Conversely, removing acetyl groups from the tails of the histones causes the expression of the respective genes to be shut down. These and other histone modifications that act as epigenetic marks and influence polymerase binding to DNA, are also called the histone code of the DNA. Recent studies have shown that RNA polymerase II pausing represents a key step in the transcriptional regulation. This is a phenomenon in which RNA pol II pauses at the proximal region of the promoter and when released initiates productive elongation of the transcripts (Jonkers and Lis, 2015; Darzacq et al., 2007; Zeitlinger et al., 2007). This phenomenon of RNA Pol II pausing seems to be a general mechanism for stress response regulation and recovery (Maxwell et al., 2014; Liu et al., 2015; Radonjic et al., 2005).

1.4.2 Post-transcriptional regulation of gene expression

Gene expression can be altered in steps that immediately follow mRNA transcription. Once a pre-mRNA transcript is produced, it will be modified by the addition of a 5′m7G cap structure and polyadenylation of its 3′end (Harford and Morris, 1997; Latchman, 2002). These modifications are important for mRNA stability, translation initiation, and nuclear export. The spliceosome complex then removes introns from

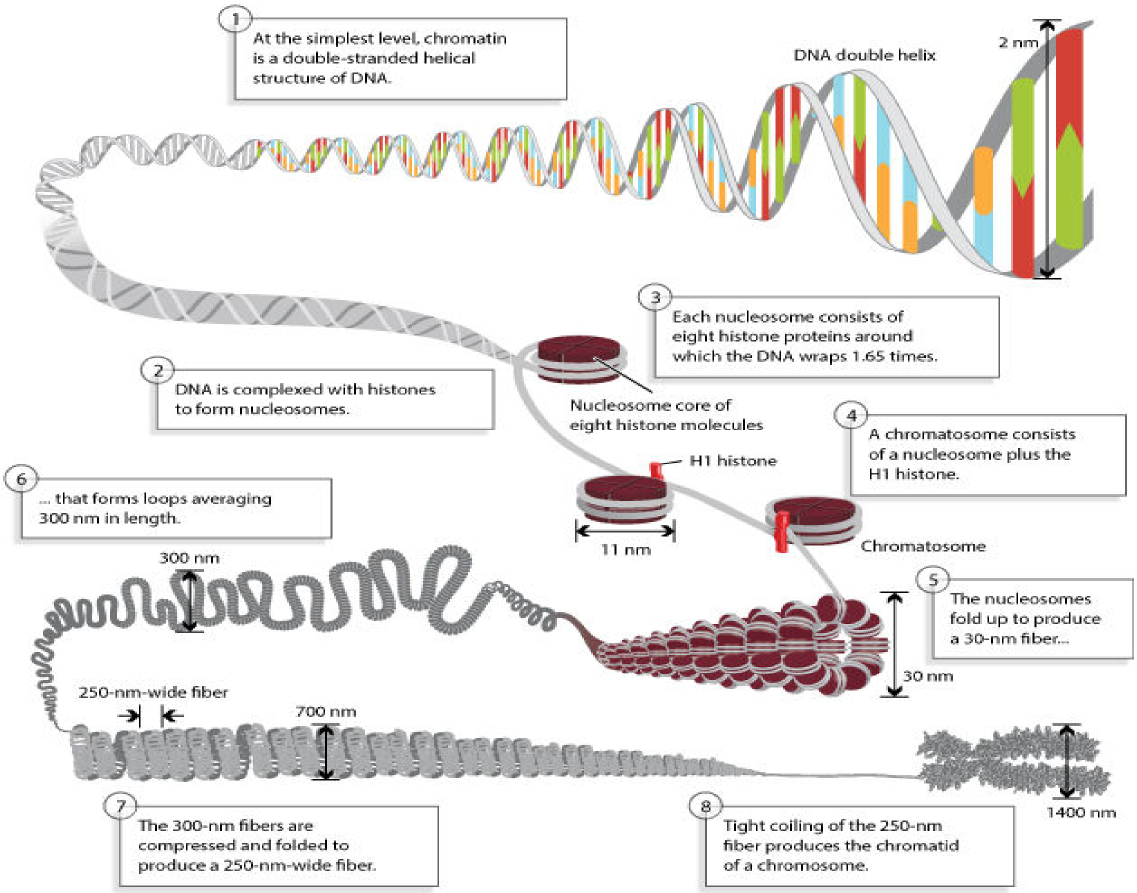

1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
2
DNA is complexed with histones to form nucleosomes.
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
Nucleosome core of eight histone molecules
H1 histone
4
A chromatosome consists of a nucleosome plus the H1 histone.
11 nm
Chromatosome
6
... that forms loops averaging 300 nm in length.
300 nm
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
1400 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

Figure 1.10: Chromosomes are composed of DNA tightly wound around histones. Chromosomal DNA is packaged inside microscopic nuclei with the help of histones. These are positively-charged proteins that strongly adhere to negatively-charged DNA and form complexes called nucleosomes. Each nucleosome is composed of DNA wound 1.65 times around eight histone proteins. Nucleosomes fold up to form a 30-nanometer chromatin fiber, which forms loops averaging 300 nanometers in length. The 300 nm fibers are compressed and folded to produce a 250 nm-wide fiber, which is tightly coiled into the chromatid of a chromosome. Figure and legend are taken from (Annunziato).

modified pre-mRNA. For many genes, exons can be spliced in different combinations, leading to multiple transcripts by the process called alternative splicing (Graveley, 2001; Roy et al., 2013). In higher eukaryotes, the total number of protein isoforms is always higher than the number of genes due to alternative splicing. Other types of post-transcriptional regulation include microRNA-mediated regulation (Friedman et al., 2009) and RNA-editing (Su and Randau, 2011; Maas, 2010).

1.4.2.1 RNA interference

Gene regulation by microRNAs(miRNAs, endogenous) is one form of RNA interference (RNAi). This process of post-transcriptional gene silencing can also be mediated by small interfering RNAs (siRNAs, exogenous). This phenomenon was first observed in plants (Ecker and Davis, 1986) and later in 1998, Andrew Fire and Craig C. Mello identified the pathway in *C. elegans* (Fire et al., 1998) . For this discovery, they were awarded 2006 Nobel Prize in Physiology or Medicine. Later this pathway was found in at least some members of 4 of the 5 supergroups of eukaryotes (Shabalina and Koonin, 2008). Briefly, the endoribonuclease enzyme dicer cleaves double-stranded RNA (dsRNA) into short double-stranded fragments of 21 nucleotide siRNAs. The siRNAs duplex is formed of the guide strand and the passenger strand. The endonuclease Argonaute (Ago) catalyzes the disentanglement of the siRNA duplex. The guide strand is incorporated into the RNA-induced silencing complex (RISC) while the passenger strand is degraded. The guide strand directs the RISC complex to the target mRNA. It then pairs with a complementary sequence in the mRNA molecule and the catalytic Argonaute protein of the RISC complex induces cleavage of the target mRNA (Wilson and Doudna, 2013) see Figure 1.11.

1.4.2.2 Genome-wide RNAi screens

In contrast to many organisms, in worms, plants and a few arthropods RNAi is systemic (Grishok, 2005; Xie and Guo, 2006; Voinnet, 2005; Mlotshwa et al., 2002; May and Plasterk, 2005). In *C. elegans*, due to systemic RNAi, double-stranded RNA (dsRNA) spreads throughout the organism due to two proteins SID-1 and SID-2 (Hunter et al., 2006). This has prompted the development of experimental approaches in several organisms for genome-wide loss-of-function screening (Campeau and Gobeil, 2011; Boutros and Ahringer, 2008). In *Caenorhabditis elegans*, researchers often use a feeding method for RNAi that involves culturing worms on a bacterial clone expressing a double-stranded RNA (dsRNA) that is supposed to target a specific worm gene (Timmons and Fire, 1998; Timmons et al., 2001). Different RNAi screens libraries have been generated for *C. elegans*. Most studies use two libraries, constructed using polymerase chain reaction (PCR)-amplified fragments of genomic DNA (Kamath et al., 2003) (the Ahringer library), or from ORFeome clones (Rual et al., 2004b), which are derived from cDNA (Reboul et al., 2001), (the Vidal library). In *C. elegans* many genome-wide RNAi screens have been performed for various phenotypes (Maia et al., 2015; Hamilton et al., 2005; Boutros and Ahringer, 2008; Poulin et al., 2004; Lehner et al., 2006; Ashrafi et al., 2003; Simmer et al., 2003; Lee et al., 2003). Small population of worms are cultured on different RNAi clones and the consequences are assayed. As mentioned in the following section, almost all of these screens were done manually, so they are prone to human error. At the simplest level, clones can be mixed up. As a consequence, one needs to confirm the identity of each candidate clone′s DNA insert. The problem of clone verification and it′s solution are described in the next section.

1.4.2.3 Clone sequence verification

In common with any large-scale resource, the available bacterial RNAi clone libraries contain errors (e.g. clone positions inverted on 96-well plates). For the Ahringer library, this error rate is estimated to be approximately 7% (http://www2.gurdon.cam.ac.uk/ãhringerlab/pages/rnai.html; (Qu et al., 2011)). These can be compounded by handling errors during a screen, resulting in error rates as high as 15% (Pukkila-Worley et al., 2014). This means that clones need to be checked by sequencing to confirm their identity, typically by using BLAST to compare them with the genome can be laborious when dealing with large numbers of clones. At the start of our project, there was no automated pipeline that could identify clone sequences without manual interpretation existed. So in the Results (section 2.1) I present CloneMapper, a tool designed to overcome these problems (Thakur et al., 2014). Once we know the actual insert for an RNAi clone, we need to identify its potential mRNA target(s). In the next section, I will talk about the problems and solutions for the clone target(s) identification.

1.4.2.4 RNAi targets

In vertebrates short interfering RNA screens, 21nt siRNAs are directly used for silencing the target gene. This makes target identification relatively straightforward. In the case of *Drosophila* and *C. elegans* due to their higher effectiveness, dsRNAs (often 1 kb long) are preferred over direct siRNAs. Each dsRNA can give rise to a multitude of siRNAs, which complicates target identification as described in (Thakur et al., 2014). Although resources like Wormbase (Harris et al., 2014) and UP-TORR (Hu et al., 2013) are available for RNAi clone target identification in *C.elegans*, these tools have a number of limitations (Thakur et al., 2014).

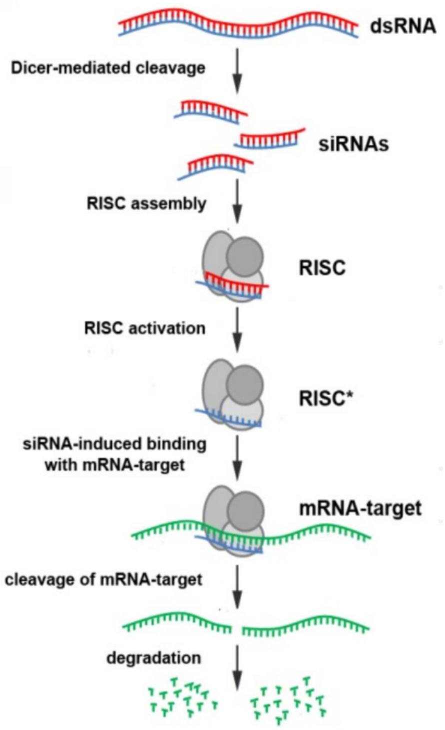


Figure 1.11: Scheme of RNAi. (SPetrova et al., 2013).

I therefore developed a clone target identification algorithm that we incorporated into the CloneMapper tool (Thakur et al., 2014). Details of this tool are presented in section 2.1. Once we identified the potential targets of the RNAi screen clones, the next major hurdle was the analysis of the target genes, to make biological sense out of a list of genes. In the following section, I will discuss about the different types of high throughput data analysis techniques that have been used for RNAi screen analysis.

1.5 High-throughput data list analysis

In the past 2 decades, high-throughput assays for functional genomic, transcriptomic, and proteomic analyses of biological samples have yielded enormous amounts of data. The huge increase in available data and its increasingly complex nature has also increased significantly analytical challenges. Having the lists of significant genes obtained by the high-throughput experiments (e.g. a genome-wide RNAi screen) provides limited mechanistic insights, without additional biological insight. So to gain systems level understanding of a biology, we need to analyse data in a holistic manner. This can be done either by finding clusters of genes which behave in a similar way in a given high-throughput experiment, or by incorporating expert knowledge from pathway databases for focused analysis. I will discuss a few important and common high-throughput analysis techniques like phylogenetic profiles, functional enrichment and clustering in the next sections.

1.5.1 Phylogenetic profiles

Phylogenetic profiling is a technique of inferring/associating meaningful biological function(s) to a set of genes based on the joint presence or absence of either the complete protein or its constituent domains, across the species. It is based on the assumption that different proteins in the same biological pathway will have the same conservation profile, as shown in Figure 1.12. Phylogenetic profile comparison was introduced in 1999 by (Pellegrini et al., 1999). Recently this technique has been used extensively to answer many different biological questions (Jim et al., 2004; Marcotte et al., 2000; Dey et al., 2015; Simonsen et al., 2012; Date and Marcotte, 2003; Tabach et al., 2013b,a). The theory behind phylogenetic profiling is that during evolution, closely-related species are expected to carry out or have similar biological processes, and similar biological process are expected to require similar sets of genes. In a more distant species, if a particular process was totally missing then most of the relevant components would be expected to be missing too. In other words, most of the genes for a given pathway will either be conserved or absent altogether. The overall strategy of phylogenetic profile analysis is shown in Figure 1.12.
Tools available for phylogenetic analysis Numerous tools have been developed to aid phylogenetic profiling. They include:

Clime: clustering by inferred models of evolution (CLIME) (Li et al., 2014).

FunCoup 3.0: a database of genome-wide functional coupling networks (Schmitt et al., 2014).

PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles (Sadreyev et al., 2015).

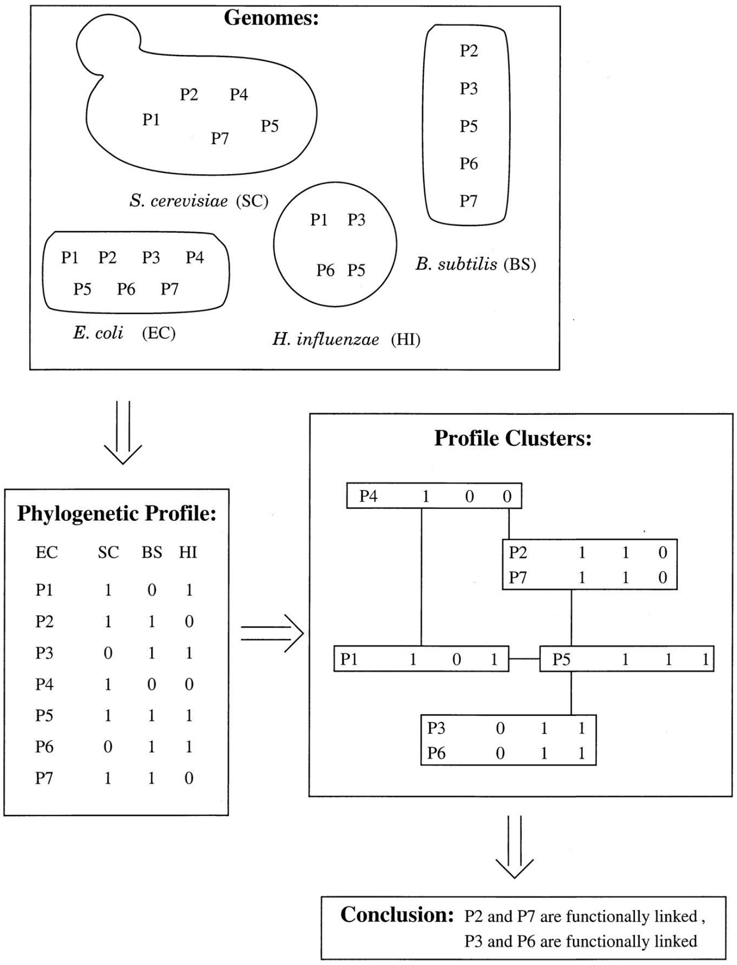
PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya (Cromar et al., 2016).

ProtPhylo: identification of protein-phenotype and protein-protein functional associations via phylogenetic profiling (Cheng and Perocchi, 2015).

SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles (Franceschini et al., 2016).

1.5.2 Functional enrichment

As mentioned above, in recent years, there has been a boom in genomic, transcriptomic and epigenomic studies, largely fuelled by advances in sequencing technologies and the attendant reduction in costs. They often result in the production of large lists of candidate genes. Various publicly-available resources classify genes on the basis of their structure, interactions or function. In any biological process, many genes/proteins are involved and when any high throughput experiment is performed, most likely, one



Phylogenetic Profile:

EC	SC	BS	HI
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0

Conclusion: P2 and P7 are functionally linked , P3 and P6 are functionally linked

Figure 1.12: Pipeline for phylogenetic profile analysis, taken from (Pellegrini et al., 1999). The Pellegrini method of analyzing protein phylogenetic profiles is illustrated schematically for the hypothetical case of four fully sequenced genomes (from E. coli, Saccharomyces cerevisiae, Haemophilus influenzae, and Bacillus subtilis) in which we focus on seven proteins (P1-P7). For each E. coli protein, we construct a profile, indicating which genomes code for homologs of the proteins in question. We next cluster the profiles to determine which proteins share the same profiles. Proteins with identical (or similar) profiles are boxed to indicate that they are likely to be functionally linked. Boxes connected by lines have phylogenetic profiles that differ by one bit and are termed neighbors.

might miss many genes of that particular biological process. It is hard to say whether a particular condition leads to a particular phenotype based on presence or absence of an individual genes. Instead of looking at an individual genes, in functional enrichment analysis we look at a group of genes, which overcomes the problem of these important missing genes. This technique of gene functional enrichment has been used for more than a decade (Huang et al., 2009a) and there are currently dozens of available tools, many listed at http://omictools.com/. These tools suffer from various shortcomings, such as an absence of regular updates of annotations (Wadi et al.). Further, most of these tools use only Gene Ontology (GO) annotations; only very few tools use transcriptome data for functional annotation. Due to their limited knowledgebases, these tools are not so revealing in functional enrichment analysis. In section 2.3 of the Results, I will talk about the importance of integration of various kind of data and the tool that we developed for analysing high throughput data in C. elegans.

1.5.2.1 Hypergeometric Test

Typically, Statistical analysis for finding out the enriched classes for a given gene list is done using hypergeometric tests. These tests use the hypergeometric distribution to calculate the statistical significance of having drawn a specific k successes (out of n total draws) from the aforementioned population. The test is often used to identify which sub-populations are over- or under-represented in a sample. This test has a wide range of applications. For functional enrichment statistics, we used hypergeometric distribution using the R stats package. The p-value is calculated as
phyper(q = x-1, m = m, n = n, k = k, lower.tail = FALSE)
Where, k is the number of genes in the user′s query.
N is a total number of genes annotated in the catalogue of reference classes.
m is a number of genes annotated in the functional class of interest.
x is a number of genes in the user list that are annotated in the functional class of interest.
n= N - m

1.5.3 Clustering

Clustering is an unsupervised machine learning technique of grouping together sets of objects into different clusters or groups based on their similarity. Objects in one group are more similar to each other compared to the objects in other groups. Clusters can be defined as collections of the objects which are similar between themselves and are dissimilar to the objects in other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. An example is shown in Figure 1.13, below.

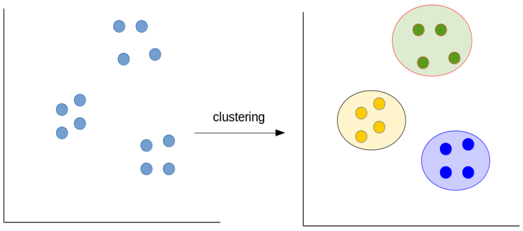


Figure 1.13: Data, associated with values plotted on the 2 axes, have been divided into three clusters based on a distance or similarity criterion.

1.5.3.1 Classification of clustering algorithms

Clustering algorithms may be classified as follows:

- **Exclusive Clustering** (e.g. K-means)
- **Overlapping Clustering** (Fuzzy, C-means)
- **Hierarchical Clustering**
- **Probabilistic Clustering** (Mixture of Gaussian)

In **exclusive clustering**, a data point can only belong to one cluster as shown in the example below Figure 1.14. A straight line separates the points and each point can only be on one side of the bi-dimensional plane. On the other hand, in overlapping clustering, also called **fuzzy clustering**, each point might belong to multiple clusters with different degrees (probability) of membership. The **hierarchical clustering** algorithm is based on the union between the two nearest clusters. At the starting point, each cluster is assigned a separate cluster and after few iterations, it reaches the final clusters. Finally, the last kind of clustering uses a completely **probabilistic approach** (Jain et al., 1999).

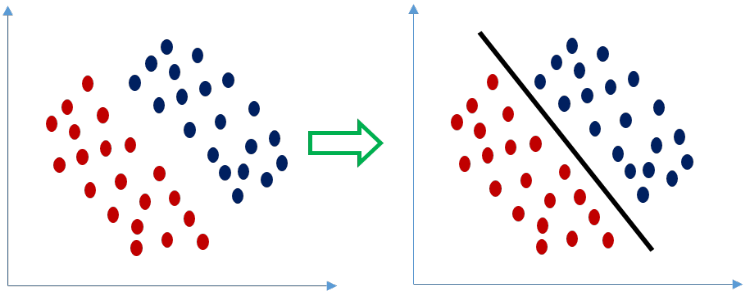


Figure 1.14: Exclusive clustering.

1.5.4 Measuring gene expression

Upon infection, many genes are regulated by the immune system of *C. elegans* and contribute to fighting against the invading pathogen. We tried to identify these effectors by time-course transcriptome analysis as mentioned in the last part of the discussion. In the following few sections I will talk about the techniques to measure the expression of genes, followed by the analysis pipeline. A gene expression profile provides a snapshot of transcriptional activity at the molecular level. It can also represent the collective interactions of many events or phenomena that are difficult to detect. In short, it can be regarded as a proxy for a transcriptional event like, transcription initiation, elongation, termination, splicing, etc.. mRNA abundance can be quantified by RT-qPCR for a small number of genes. With high-throughput transcriptome experiments like microarray and RNA-sequencing, in principle, the expression of all the genes in the genome can be quantified.

1.5.4.1 Whole organism transcriptome analysis

As mentioned above, gene expression changes are at the root of differences in phenotypes for cells and tissues. Monitoring the expression of thousands of genes at once can give a global picture of cellular function. The steps involved in whole organism transcriptome analyses are described below.

1.5.4.2 Measuring gene expression with RNA sequencing

Although many technologies have been developed for measuring gene expression, only hybridisation-based microarray technology and sequencing-based RNA-sequencing (RNA-seq) allow the levels of thousands of genes to be measured simultaneously. In recent years, RNA-seq technologies have overtaken the older microarray technology, due to certain limitations in the latter. Unlike microarrays, species- or transcript-specific probes are not needed in RNA-seq technology; all possible transcripts, including novel ones, not predicted from genome annotations, as well as gene fusions, single nucleotide variants, and indels (small insertions and deletions) can be detected with RNA-seq. Indeed, RNA-seq analysis does not require prior knowledge of gene structure, rather it can be used to predict genes in organisms for which no genome sequence is available. Further RNA-seq is more specific and sensitive in the detection of differential gene expression (Marioni et al., 2008; Wang et al., 2009a).

1.5.4.3 RNA sequencing experiment workflow

Currently, several high-throughput DNA sequencing technologies exist. The most established are those developed by Illumina (Genome Analyzer I/II and Hiseq), Roche (454), and Applied Biosystems (ABI SOLiD) (Wang et al., 2009b; Oshlack et al., 2010). Although most of the technologies have different experimental protocols, the overall idea is more or less the same(Martin and Wang, 2011). Here, I will limit the discussion to how the most used, Illumina′s sequencing protocol, can be applied to RNA-seq. The first step includes extraction/pulldown of only polyadenylated RNA molecules; other classes of RNAs are discarded. These polyadenylated RNA molecules are then fragmented (RNA hydrolysis) into short fragments of 200-500 bp. As with most of the techniques, the Illumina technique sequences only DNA molecules, so these short RNA fragments are then converted into cDNA (complementary DNA) by reverse transcription. To start reverse transcription of these RNA molecules, short primers containing Ts (deoxy-Thymine sequences or oligo-dTs) complementary to the RNA polyA tails hybridize to the RNA sequence or 6 random bases (hexamers) are used to hybridise to random positions on the RNA molecule. After reverse transcription, RNA molecules are removed. Blunt ends are generated after treating these double-stranded cDNA molecules and to both ends, cDNA adapters are ligated and later PCR amplified. Different size of RNA fragments are produced during the fragmentation step, DNA molecules of a specific size range are extracted and purified, so as to keep molecules of a similar length. This purification step is later followed by sequencing. Sequencing of the fragments is either single end (SE) or paired end (PE). Single DNA strands are generated after denaturing double stranded molecules and these are passed over flow cell with oligos complementary to the adapter sequences. Next, sequencing primers are added and the millions of molecules are reverse complemented simultaneously. In each sequencing step, fluorescently labelled reversibly terminated nucleotides compete to bind with the template strands. In each step, only one nucleotide is added to each growing complementary strand. Newly added nucleotides are labelled with a dye (different for each nucleotide type) and newly-added nucleotides are identified by their fluorescence. Terminal groups of fluorescent dye are removed from the newly added nucleotides. Later sequentially-acquired images are analysed using software tools called base callers (Ledergerber and Dessimoz, 2011). The generated files are most commonly in the FASTQ file format which includes reads, which contains a unique read ID, nucleotide sequence and a Phred quality score per base. Base Callers set Phred quality scores Q where Q = -10log10(P), where P is the probability of the base call being incorrect (Ewing et al., 1998). For the paired-end experiments, two files of reads are generated with the first end in one file and the second end in another file.

1.5.5 RNA-seq analysis

There are different RNA-seq pipelines available; most of them follow these common analysis steps, Figure 1.15.

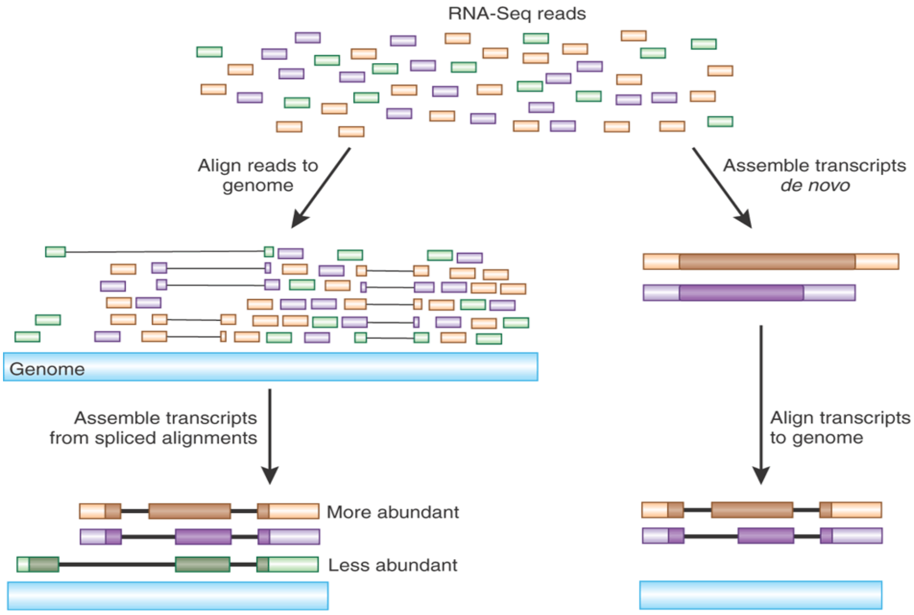

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Figure 1.15: The 'align-then-assemble' approach first aligns short RNA-seq reads to the genome, accounting for possible splicing events, and then reconstructs transcripts from the spliced alignments. The 'assemble-then-align' approach (right) first assembles transcript sequences *de-novo* that is, directly from the RNA-seq reads. These transcripts are then splice-aligned to the genome to delineate intron and exon structures and variations between alternatively spliced transcripts. As *de-novo* assembly is likely to work only for the most abundant transcripts, the align-then-assemble method should be more sensitive, although this warrants further investigation. RNA-seq reads are colored according to the transcript isoform from which they were derived. Protein-coding regions of reconstructed transcript isoforms are depicted in dark colors. Figure and legend are taken from (Haas and Zody, 2010).

1.5.5.1 Read filtering and mapping strategies

The main objective of RNA-seq is to estimate the expression level of particular genomic feature like genes, isoforms, exons, splice junctions etc. In RNA-seq, one obtains reads of good and bad qualities based on the sequencing technique and RNA samples. So the first important step is to filter out these low quality reads and this process is called quality control. There are many tools available, such as FASTQC and FASTX-toolkit. The ends of low-quality reads (generally with a Phred score below 20) are trimmed or the reads are removed. Mapping of filtered reads to the above-mentioned features can be challenging, due to low read number and quality. When a reference genome is available, as described below, reads can be directly aligned, otherwise, de novo gene assembly is performed before mapping of the reads to the predicted genes.

1.5.5.2 Read alignment

If the reference genome is present, the reads are directly aligned to a reference genome or a transcriptome. There are many alignment programs with alternate splicing, TopHat (Trapnell et al., 2012), STAR (Dobin et al.), GSNAP (Wu and Nacu, 2010)), etc., and without spliced alignments, SOAP (Li et al., 2008b) and BWA (Li and Durbin, 2009).

1.5.5.3 Alignment to the genome

The reads are aligned to the genome and different aligners use different algorithms. For example, TopHat, the most used spliced aligner, first tries to align all the reads to the genome with Bowtie and then non-aligned reads are set apart. Other programs try to realign the unaligned reads (local alignment) after trimming their ends (Hillier et al., 2009). The aligned reads form a cluster/islands of expression. Each dense read cluster is followed by decreasing read covered introns region. Alignment to the genome results in a set of one or more genomic coordinates for each aligned read, which may or may not span exon junctions. By itself, this information is of limited use, so alignments can be further matched to known annotated features (for example by searching for overlaps between aligned reads and genes) or they can be used to build gene models *de-novo*. One of the earliest and most well-known programs to achieve the latter is a software application called Cufflinks (Trapnell et al., 2010).

1.5.5.4 Expression Quantification

The next important step in RNA-seq analysis is to compare the expression of genes between different conditions/samples. Ideally, counting the reads that overlap each genomic feature (Gene, CDS, exon, etc.) is straightforward, except for certain experimental approaches (e.g. the study of alternative splicing) that will not be detailed here. There are many tools available that perform the task of read quantification, the most famous being htseq-count (Anders et al., 2014) and featureCounts (Liao et al., 2014). The number of reads mapped to gene depends on many factors like gene expression level, length, the sequencing depth and the expression of all other genes within the sample. For an unbiased comparison between two conditions, one needs to calculate the fraction of the reads actually aligned to each gene relative to the total number of reads in a given RNA-seq sample. There are various normalization tools like DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), etc. Once you have the normalized read counts for a given feature, the next task is to compare the read counts for a given feature between two or more samples. This step is called differential gene expression analysis. For calculating differentially expressed genes (DEG), again many tools are available, for example, DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), Cuffdiff (Trapnell et al., 2013), etc. Selecting a tool for DEG calculation is an important step as none of the tools is perfect (Rapaport et al., 2013; Dillies et al., 2013).

Once we have identified the differentially regulated genes using transcriptome analysis, the next most important thing in defining the regulatory network (discussed later) is to identify the transcription factor (TFs) that directly govern their expression. These TFs bind to specific regions on DNA to regulate nearby genes. I will give a brief introduction to the techniques that are used to identify TF bound regions of the DNA and then explain the pipeline used for analysing data.

1.6 DNA-protein interaction (ChIP-seq)

One direct way to measure or infer gene regulatory networks is by looking for DNA-protein interactions. One in vitro TF-centered high-throughput experimental technique is based on protein-binding microarray (PBM) technology. A purified TF is applied directly to double-stranded DNA microarrays covering a wide range of possible DNA-binding site sequences (Zhu et al., 2009; Badis et al., 2009; Berger et al., 2008; Grove et al., 2009). Many alternative methods have been developed recently for determining protein-DNA interaction based on chromatin immunoprecipitation (ChIP) technologies (O'Neill and Turner, 1996). ChIP involves cross-linking DNA to protein using formaldehyde, followed by DNA sonication and precipitation of the target protein using antibodies. To identify transcription factor binding site (TFBS) in a DNA fragments precipitated with ChIP, microarrays and sequencing techniques can be used. The microarray-based ChiP method (ChIP-on-chip) (Ren et al., 2000) has been largely superseded by methods based on high-throughput sequencing (ChIP-sequencing or ChIP-seq), where ChIP is followed by sequencing (Mikkelsen et al., 2007).

1.6.1 Advantages of ChIP-seq over ChIPchip

As for RNA-seq, ChIP-seq is not limited by array design; organism-specific microarrays need to be designed for ChIP-chip. ChIP arrays are commercially available for only a limited number of species, whereas ChIP-seq can be done for any organism. Compared to ChIP-chip, ChIP-seq has higher spatial resolution, better signal to noise ratio and requires less starting material. ChIP-seq being more advantageous, I will discuss only the ChIP-seq analysis pipeline.

1.6.2 ChIP-seq analysis

The standard ChIP-seq analysis pipeline is straightforward and most of the steps are similar to RNA-seq analysis, Figure 1.16 (Bailey et al., 2013).

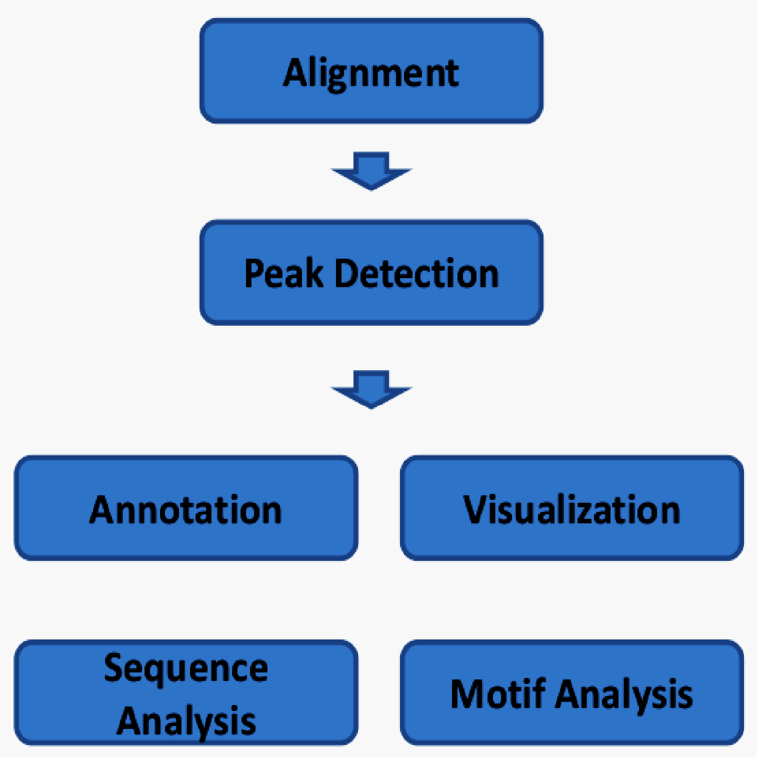


Figure 1.16: Workflow for the computational analysis of ChIP-seq.

1.6.2.1 Read mapping and quality check

Before aligning/mapping reads to the reference genome, the quality of the reads is checked and the bad quality reads are filtered out. Again, filtered reads can be aligned with different aligners, such as Bowtie (Langmead et al., 2009), SOAP (Li et al., 2008b), BWA (Li and Durbin, 2009), or MAQ (Li et al., 2008a). One of the main issues in early ChIP-seq data analysis arise from the short read lengths. This made it difficult to unambiguously map reads to a unique genomic region. Improvement in sequencing have largely eliminated this problem. The signal-to-noise ratio (SNR) of a ChIP-seq experiment is calculated after read mapping. Strand cross-correlation analysis is built into some peak callers, such as SPP (Kharchenko et al., 2008) or MACS (Zhang et al., 2008). This allows more precise identification of TF binding sites.

1.6.2.2 Peak calling

After mapping filtered reads to the reference genome using an NGS aligner as described above, the most important part of the ChIP-seq analysis is to identify the genomic regions that have been significantly and differentially bound by the ChIPed protein. This process of identifying the protein-bound regions is called peak-calling. Peak-calling sensitivity and specificity depend on the peak-calling program. The parameters (Jothi et al., 2008; Guo et al., 2012) are selected for the peak-caller algorithm, the type of protein ChIPed (narrow region binding TFs, broadly bound histone marks factors and RNA Pol II which binds to both wide and narrow region (Pepke et al., 2009)). More details on the parameter and factors that affect peak calling, can be found in an excellent review (Bailey et al., 2013).

1.6.2.3 Peak annotation/motif detection

Once the ChIPed protein binding peaks are defined, one can either look for the enriched binding motif in these peaks or one can assign these peaks to genomic regions, such as gene promoters, transcription start sites (TSS), intergenic regions, etc. Different softwares are available for identifying enriched motif in called peaks, such as RSAT peak-motifs (Thomas-Chollier et al., 2012), MEME-ChIP (Machanick and Bailey, 2011). These tools take binding peak sequences and then find the enriched motif in these sequences. Enriched motifs can be compared to the already known TF binding motifs using tools like RSAT - compare-matrices (Medina-Rivera et al., 2015), or MEME- Tomtom (Bailey et al., 2009). For peak associations to annotated genomic features, various tool can be used, like CEAS (Ji et al., 2006) or the Bioconductor package ChIPpeakAnno (Zhu et al., 2010). Specific tools like GREAT (McLean et al., 2010), or GSEA (Subramanian et al., 2005), etc., are available for relating peaks to Gene ontology.

In any given signaling pathway, at every step, proteins physically interact and these interactions lead to the desired output of the network. I will talk about protein-protein interactions in the following section.

1.7 Protein-protein interactions (PPI)

Transcriptome data provide a picture of how much a gene is transcribed at a particular time point in a cell/tissue/organism. Genes are translated to proteins and these proteins play a major role in cellular processes. Almost always, proteins work together in a hierarchical fashion to perform a certain biological function (Berggård et al., 2007). Frequently, proteins physically bind together to form protein complexes. These complexes further interact with each other to form pathways to carry out cellular processes. As might be expected, (Wan et al., 2015) have shown that most protein complexes are conserved across species. One of the main factors for specificity of protein-protein interaction is the protein′s structural domains (SH2, PDZ, CH, SAM, etc.) that allow sequence-specific interaction between proteins. The composition or activity of the protein complexes is regulated by the amount of expressed proteins, the relative affinity of the proteins, substrate or co-factor concentration, molecular forces, etc. An enormous amount of PPI data has been generated in recent years by high-throughput experimental methods, like two-hybrid analysis, mass spectrometry, etc. This data of individual protein-protein interaction is put together into dedicated resources like STRING (Franceschini et al., 2013), BioGRID (Stark et al., 2006), MINT (Zanzoni et al., 2002), BIND (Bader et al., 2003), IntAct cit (Hermjakob et al., 2004), HPRD (Baolin and Bo, 2007) and common reference databases like Irefindex (Razick et al., 2008). Other resources like PCDq (Kikugawa et al., 2012), CORUM (Ruepp et al., 2010), metazoan macromolecular complexes (Wan et al., 2015) store the protein complex information. In a protein-protein interactions network, when two proteins interact, they either activate or repress the activity of the partner protein. When known, such effects can be represented by ”signs” in the corresponding PPI network (Vinayagam et al., 2014).

1.8 Gene regulatory networks (GRNs)

Non-coding regions of DNA that regulate the expression or transcription of nearby genes are called cis-regulatory elements (CREs) (Wittkopp and Kalay, 2011) CREs are found in the close vicinity of the genes that they regulate. These CREs serve as the DNA-binding sites for specific TFs that either activate or repress expression of multiple genes. CRE can refer either to individual TF-binding sites or to a collection of different TF-binding sites clustered within a broader region of DNA, also called cis-regulatory modules (CRM) (Hardison and Taylor, 2012). CRMs that activate gene expression are called transcriptional enhancers and those that repress gene expression are called transcriptional silencers. A gene regulation system mainly consists of genes, the genes CREs/CRMs (Walhout et al., 2012) and their respective regulators, like TFs, small RNAs or metabolites. TFs bind to cis-elements present in the cis-region of specific genes and control the level of gene expression. These genes, their regulators, together with the regulatory connections between them form gene networks Figure 1.17 and Figure 1.18. These gene regulatory networks determine how genes will be expressed in the cells and what kind of protein will be translated. These expressed genes and proteins can further regulate the expression of other genes, which gives rise to complex network structures.

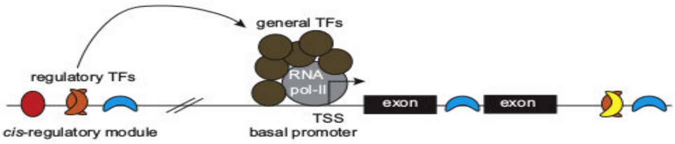


Figure 1.17: Transcription plays a pivotal role in the regulation of gene expression. Cartoon depicting transcriptional regulation for an individual eukaryotic gene. The figure is adapted from ”Chapter 4: Handbook of systems biology”,Bulyk ML, Walhout AJ, 2013.

1.8.1 Modeling gene regulatory networks

In computer models, gene regulatory networks are usually represented as directed graphs, where genes correspond to nodes and interactions between the genes are

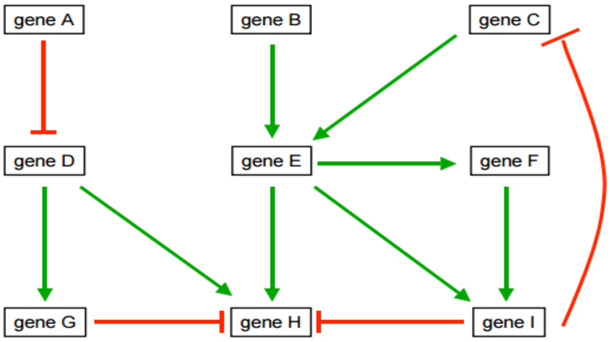
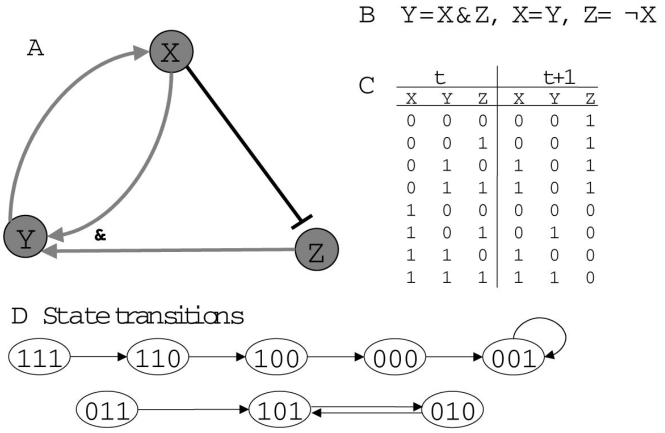


Figure 1.18: Basic concept of Regulatory networks, green edges represent the activator regulatory relation and red represent repressor relation between genes.

represented by edges. Mainly gene regulatory network models can be divided into several different classes, including **Boolean networks**, **Relevance networks**, **Bayesian networks** and **differential equation models** (Kaderali and Radde, 2008; Karlebach and Shamir, 2008). I will not discuss multivalued logic models proposed by Ren Thomas.
Boolean GRN networks, as the name suggests, each gene in the network is considered to be in one of two states, either active (expressed) or inactive (not expressed). In a network with n genes, there will be 2n total possible different states, for instance, a three-gene network can have 8 possible states (0,0,0), (0,0,1), (0,1,0), (1,0,0), (1,0,1), (1,1,0), (0,1,1), (1,1,1). We can follow the succession of states with time and study which states are reached, Figure 1.19. This type of regulatory networks were first presented by Kauffman (Glass and Kauffman, 1973). Boolean logic functions are applied for the interactions between genes, simultaneously, with the state of other gene nodes being updated according to the levels of its regulators at the previous time step. Boolean GRN networks are deterministic and give only a qualitative description of a system (Kauffman et al., 2003). Examples of the application of Boolean network algorithms, established using the tool REVEAL (Liang et al., 1998) can be found in (Akutsu et al., 1999, 2000; Lähdesmäki et al., 2003).



B Y=X&Z, X=Y, Z= ¬X

C

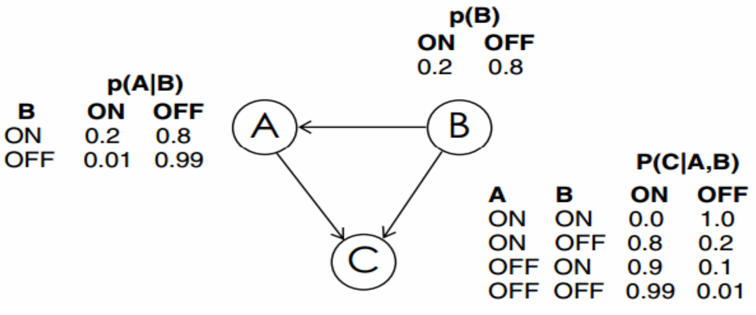
t			t+1		
X	Y	Z	X	Y	Z
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figure 1.19: An example of a small Boolean network consisting of three genes X, Y and Z. There are different ways to represent the network: as (a) a graph, (b) Boolean rules for state transitions, (c) a complete table of all possible states before and after the transition, or (d) a graph representing the state transitions. Figure reproduced from (Schlitt and Brazma, 2006).

In **relevance networks**, unlike Boolean networks, instead of looking for the boolean states of genes, relevance networks look at similarity or dissimilarity between the pairs of genes on a continuous scale (Butte and Kohane, 1999). For relevance network reconstruction, similarity/dissimilarity between all pairs of genes is measured using pairwise gene correlation coefficients or mutual information. Pairwise gene connections are filtered based on a threshold similarity/dissimilarity measure between genes which can be represented in a graphical form. The ARACNe algorithm by Basso et al (Margolin et al., 2006; Basso et al., 2005) is based on relevance model networks. **Bayesian networks** provide a graphical representation of statistical dependencies between variables, but more importantly, they also allow one to visualize independent relations among variables. Bayesian networks are directed, acyclic graphs (DAG) G = (X,A), together with a set of local probability distributions P, where X = X1,...,Xn correspond to variables and are called vertices/nodes, and A are the directed edges which represent probabilistic dependence relations between the X variables. If an arc goes from variable Xi to Xj, then a child node Xj will depend probabilistically on the parent node Xi. If there are no parent nodes, such nodes are called unconditional. P is the local probability distributions of each node Xi conditioned on its parents, p(Xi|parents(Xi)) (Kaderali and Radde, 2008). Once one has a Bayesian network, the joint probability distribution of all the variables in the networks can be easily calculated using joint distribution, given as For a simple Bayesian network such

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | parents(X_i)).$$

as the one shown in Figure 1.20, the joint probability distribution is given by p(A,B,C)=p(B)p(A|B)p(C|A,B) and the joint probability that all nodes are on is p(A = on,B = on, C = on) = p(B = on) p(A = on|B = on) p(C = on|A = on,B = on) = 0.2X0.2X0.0= 0.0. Like this, one can calculate the joint probability distribution of any given Bayesian network.



p(B)

ON	OFF
0.2	0.8

p(A|B)

B	ON	OFF
ON	0.2	0.8
OFF	0.01	0.99

P(C|A,B)

A	B	ON	OFF
ON	ON	0.0	1.0
ON	OFF	0.8	0.2
OFF	ON	0.9	0.1
OFF	OFF	0.99	0.01

Figure 1.20: An example of simple Bayesian network with three nodes A, B and C, each assumed to be in one of two states, either on or off. The conditional probabilities p(A|B), p(C|A,B) and the unconditional probability p(B) in this binary case are easily tabulated, as shown in the figure (Kaderali and Radde, 2008).

Differential equation models provide a quantitative description of gene regulatory networks. They can be simple linear differential equation models or can be very complicated systems of nonlinear partial differential equations and stochastic kinetic approaches.

In the context of a 3-month internship at the Ontario Institute for Cancer Research, I tried to apply a Bayesian approach to model the gene network involved in *nlp-29* regulation. A report of this work is presented in the Results section 2.6.

1.9 Context of publications

My Ph.D. work was mainly focused on the innate immunity gene regulatory networks in *C. elegans*. Upon fungal infection, *C. elegans* up-regulates the expression of many antimicrobial peptide (AMP) genes. These AMPs provide direct protection against the pathogen attack. The aim of my thesis was to build an integrated gene regulatory networks representing the induction of these AMP genes upon infection. Through a genome-wide RNAi screen (Zugasti et al., 2016), we identified 297 Nipi (for no induction of antimicrobial peptides after infection) clones that abrogate AMP induction. In Section 2.2, I describe how from the RNAi screen we identified pathways and complexes that are involved in AMP regulation. In the section 2.1, I will discuss about the shortcoming in the existing RNAi screen analysis tools and I will explain how we overcome these problems by develpoing CloneMapper tool. Using CloneMapper (Thakur et al., 2014) we identified 338 target genes for these 297 Nipi clones. In the section 2.3, I will explain about the YAAT functional enrichment analysis tool that I developed for the analysis of RNAi clone target genes. In this section, firstly, I will talk about the problems with the existing functional enrichment tools, for example, functional annotation in these tools are not updated, etc. and then I will discuss how we overcame these shortcomings by developing the YAAT tool (Thakur et al., MS in preparation). We used this tool to analyse the genome-wide RNAi screen targets and other pathogen-related datasets. Further, with this tool, we did functional enrichment analysis of ChIP-seq targets of CEBP-1, a transcription factor linked to the regulation of the innate immune response (Kim et al., submitted). Enriched classes clustered into two groups, one related to the development and the other to stress. Finally, to gain better understanding of the interaction between host and pathogen, we sequenced, assembled, annotated and analysed the *D. coniospora* genome (Lebrigand et al., 2016). We identified various potential virulence factors in the fungal genome. My contribution to each of the publications is given later in Result Section 2.

1.9.1 Publication 1 (Section 2.1)

Clone Mapper: an online suite of tools for RNAi experiments in *Caenorhabditis elegans*.
Thakur N, Pujol N, Tichit L, Ewbank JJ, G3, 2014

In common with any large-scale resource, the available bacterial RNAi clone libraries contain errors (e.g., clone positions inverted on 96-well plates) . For the Ahringer library, this error rate is estimated to be approximately 7% (http://www2.gurdon.cam.ac.uk/ ahringerlab/ pages/rnai.html; (Qu et al., 2011)). As a consequence, candidate RNAi clones need to be sequenced and the insert sequence compared to the expected sequence. The CloneMapper tool (Thakur et al., 2014) automates the procedure, making it easier and more reliable. In *C. elegans* each dsRNA can give rise to a multitude of siRNAs, which complicates target identification. In the publication, we developed a target identification tool and compared this tool with the existing resources like Wormbase and UP-TORR (Hu et al., 2013). This publication provides an open source, freely accessible web tool for worm community. In near future, this tool will be incorporated into the Wormbase.

1.9.2 Publication 2 (Section 2.2)

A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes.
Zugasti O#, **Thakur N#**, Belougne J, Squiban B, Kurz CL, Soul J, Omi S, Tichit L, Pujol N, Ewbank JJ, BMC Biol, 2016
#=equal contribution

To find the main/backbone components of the regulatory network, we conducted a genome-wide RNAi screen (Zugasti et al., 2016). We identified 278 Nipi (for no induction of antimicrobial peptides after infection) clones that abrogate AMP induction. We analysed 338 target genes for these 297 Nipi clones identified using Clone Mapper (Thakur et al., 2014). We showed that MAPK pathways are central for the induction of AMPs. Bioinformatics analysis revealed a role for the mitochondrial unfolded response (mtUPR) in AMP induction. We subsequently validated the involvement of mtUPR experimentally. We also identified various protein complexes including the major mRNA deadenylases CCR4-NOT as being involved in the induction of AMPs. The results further revealed a cross-tissue signaling, triggered by mitochondrial dysfunction in the intestine, that suppresses antimicrobial peptide gene expression in the nematode epidermis.

1.9.3 Publication 3 (Section 2.3)

Global biological analyses through integrated functional and phylogenetic profiling.
Nishant Thakur, Nathalie Pujol, Jacques van Helden, Laurent Tichit, Jonathan J. Ewbank.

Large-scale studies like genome-wide RNAi screen and omics (genomic, transcriptomic and epigenomics) studies produce large lists of candidate genes. One popular way to analyse such lists is via gene enrichment analysis. There are currently dozens of such tool available, many listed at http://omictools.com/. These tools have the drawback that they are generally not updated and do not contain extensive organism-specific data. Recently, a functional enrichment tool, WormExp, specific for *C. elegans* was developed; it harbours only transcriptome data. For analysis of our genome-wide RNAi screen and other in-house omics data, we developed a *C. elegans*-specific functional enrichment tool called YAAT. In this section, I will talk about YAAT tool and its applications. Finally, we used this tool for the analysis of various datasets and showed that this tool is very powerful compared to other available tools.

1.9.4 Publication 4 (Section 2.4)

Coordinated inhibition of C/EBP by Tribbles in multiple tissues is essential for C. elegans development.
Kyung Won Kim , **Nishant Thakur** , Christopher A. Piggott , Shizue Om , Jolanta Polanowska , Yishi Jin and Nathalie Pujol.

Tribbles proteins are conserved pseudokinases that function to control kinase signaling and transcription in diverse biological processes. In *Caenorhabditis elegans*, the Tribbles orthologue NIPI-3 was previously shown to activate host defense upon infection, via a conserved PMK-1/p38 MAP kinase signaling pathway. In this publication we show that *nipi-3* is essential for larval development and viability. With the help of a genetic screen for suppressor of *nipi-3* null alleles, we found that NIPI-3 negatively controls PMK-1 signaling through transcriptional repression of a C/EBP transcription factor, CEBP-1. We used YAAT for functional enrichment analysis of ChIP-seq targets of CEBP-1 (Kim et al., submitted). Enriched classes clustered into two groups, one related to development and the other to stress.

1.9.5 Publication 5 (Section 2.5)

Comparative Genomic Analysis of Drechmeria coniospora Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes.
Lebrigand K, He LD, **Thakur N**, Arguel MJ, Polanowska J, Henrissat B, Record E, Magdelenat G, Barbe V, Raffaele S, Barbry P, Ewbank JJ, PLoS Genet, 2016

The natural fungal pathogen *Drechmeria coniospora* infects *C. elegans* by adhesion of specialized non-motile spores to the nematode cuticle. To understand better the interaction between the host and the pathogen, we sequenced, assembled, annotated and analysed the *D. coniospora* genome (Lebrigand et al., 2016). Comparative and functional genomic analyses provided insights into how its nematode-destroying lifestyle evolved. We identified various potential virulence factors in the fungal genome. We used them to probe a specific interaction between *D. coniospora* and *C. elegans*, involving the potential interference by the pathogen of a host antimicrobial mechanism.

Chapter 2

RESULTS

Relationship and hierarchization of the result section is little bit complicated due to the reason that first 3 sections is dedicated to the genome-wide RNAi screen (section 2.2) to identify Nipi genes. For the analysis of the RNAi screen results, we had to develop CloneMapper (section 2.1) and YAAT(section 2.3) tools. So, instead of starting with the RNAi screen (section 2.2), I will discuss CloneMapper tool first and then I will present RNAi screen results in section 2.2. This will be followed by YAAT tool and in this section I will talk about the shortcomings of the available functional enrichment analysis tools, why we need to develop this tool and the analysis of Nipi genes with this tool.

2.1 Publication 1

Clone mapper: an online suite of tools for RNAi experiments in *Caenorhabditis elegans*.
Thakur N, Pujol N, Tichit L, Ewbank JJ, G3, 2014

In this publication, I designed and developed an algorithm for automatic RNAi clone sequence verification. I also developed a new algorithm for RNAi clone target identification, followed by comparison with the existing RNAi clone target identification tools like UP-TORR and Wormbase. A dedicated web-server was developed for both the RNAi clone verification and RNAi clone target identification. It is publicly accessible at http://bioinformatics.lif.univ-mrs.fr/RNAiMap/.

Clone Mapper: An Online Suite of Tools for RNAi Experiments in *Caenorhabditis elegans*

Nishant Thakur*,†,‡, Nathalie Pujol*,†,‡, Laurent Tichit§, and Jonathan J. Ewbank*,§,‡,1
*Centre d'Immunologie de Marseille-Luminy, UM2 Aix-Marseille Université, Case 906, 13288 Marseille Cedex 9, France
†INSERM U1104, 13288 Marseille, France ‡CNRS UMR7280, 13288 Marseille, France and §Institut de Mathématiques de Marseille, Site Sud, Campus de Luminy, Case 907 13288 Marseille Cedex 9, France
ORCID ID: 0000-0002-1257-6862 (J.J.E.)

ABSTRACT RNA interference (RNAi), mediated by the introduction of a specific double-stranded RNA, is a powerful method to investigate gene function. It is widely used in the *Caenorhabditis elegans* research community. An expanding number of laboratories conduct genome-wide RNAi screens, using standard libraries of bacterial clones each designed to produce a specific double-stranded RNA. Proper interpretation of results from RNAi experiments requires a series of analytical steps, from the verification of the identity of bacterial clones, to the identification of the clones' potential targets. Despite the popularity of the technique, no user-friendly set of tools allowing these steps to be carried out accurately, automatically, and at a large scale, is currently available. We report here the design and production of Clone Mapper, an online suite of tools specifically adapted to the analysis pipeline typical for RNAi experiments with *C. elegans*. We show that Clone Mapper overcomes the limitations of existing techniques and provide examples illustrating its potential for the identification of biologically relevant genes. The Clone Mapper tools are freely available via http://www.ciml.univ-mrs.fr/EWBANK_jonathan/software.html.

KEYWORDS
database
algorithm
gene discovery
functional genomics
MPscan

RNA interference (RNAi) is a powerful and widely used method to investigate gene function. Researchers using the model nematode *Caenorhabditis elegans* often use a feeding method for RNAi that involves culturing worms on a bacterial clone expressing a double-stranded RNA (dsRNA) that is intended to target a specific worm gene (Timmons *et al.* 2001; Timmons and Fire 1998). Because worms can be handled robotically, screens can be automated and large numbers of clones tested in parallel (Squiban *et al.* 2012). Collections of RNAi clones are available. One made by the Ahringer lab contains polymerase chain reaction (PCR)-amplified fragments of genomic DNA (Kamath *et al.* 2003), whereas the library made by the Vidal lab (Rual *et al.* 2004) was constructed from ORFeome clones, which are derived from cDNA (Reboul *et al.* 2001). Part of the strength of the method arises from the fact that knowledge of the sequence of the dsRNA in principle allows the corresponding target gene(s) to be identified.

In common with any large-scale resource, the available bacterial RNAi clone libraries contain errors (*e.g.*, clone positions inverted on 96-well plates). For the Ahringer library, this error rate is estimated to be approximately 7% (http://www2.gurdon.cam.ac.uk/~ahringerlab/pages/rnai.html; Qu *et al.* 2011). These can be compounded by handling errors during a screen, resulting in error rates as high as 15% (Pukkila-Worley *et al.* 2014). This means that clones need to be checked by sequencing to confirm their identity. Interpreting the sequences, to confirm clone identity, can be laborious when dealing with large numbers of clones.

In *C. elegans* long dsRNAs (often >1 kb) are used, in contrast to the short interfering RNAs (siRNA; typically 19−25 bp long) used in vertebrates. Each dsRNA can thus give rise to a multitude of siRNAs, which complicates target identification. Many published studies have relied on the assignment of targets provided by the community database Wormbase (Yook *et al.* 2012). This currently suffers from a number of limitations (Wormbase release WS242). The first is that target identification is based on empirical criteria. The sequence of a "primary target" is at least 95% identical with the clone insert sequence for at least 100 nucleotides (Fievet *et al.* 2013); for "secondary targets" the

doi: 10.1534/g3.114.013052
Manuscript received July 7, 2014; accepted for publication August 25, 2014; published Early Online September 2, 2014.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.114.013052/-/DC1
[1]Corresponding author: Centre d'Immunologie de Marseille-Luminy, UM2 Aix-Marseille Université, Case 906, 13288 Marseille Cedex 9, France. E-mail: ewbank@ciml.univ-mrs.fr

definition is more than 80% identity for greater than 200 nucleotides (Kamath and Ahringer 2003). These figures are calculated using BLAT (Kent 2002), which is not perfectly adapted to the task for algorithmic reasons (Imelfort 2009). Further, the target(s) of a given clone are predicted assuming that all RNAi clones contain an insert derived from genomic DNA (Figure 1A). This assumption is clearly incorrect when applied to Vidal clones generated from intron-containing genes and can lead to overprediction of clone targets (Figure 1B). At the same time, no secondary targets are predicted for Vidal RNAi clones within Wormbase currently, leading to underprediction of clone targets.

A tool, UP-TORR, has been developed that partially resolves these issues (Hu *et al.* 2013). As discussed herein, it too has some drawbacks. UP-TORR is designed for researchers using RNAi in different model systems (human, mouse, *Drosophila*, *C. elegans*) and so lacks some basic species-specific functions. For example, the standard *C. elegans* RNAi clone names (with prefixes "sjj_" and "sjj2_" or

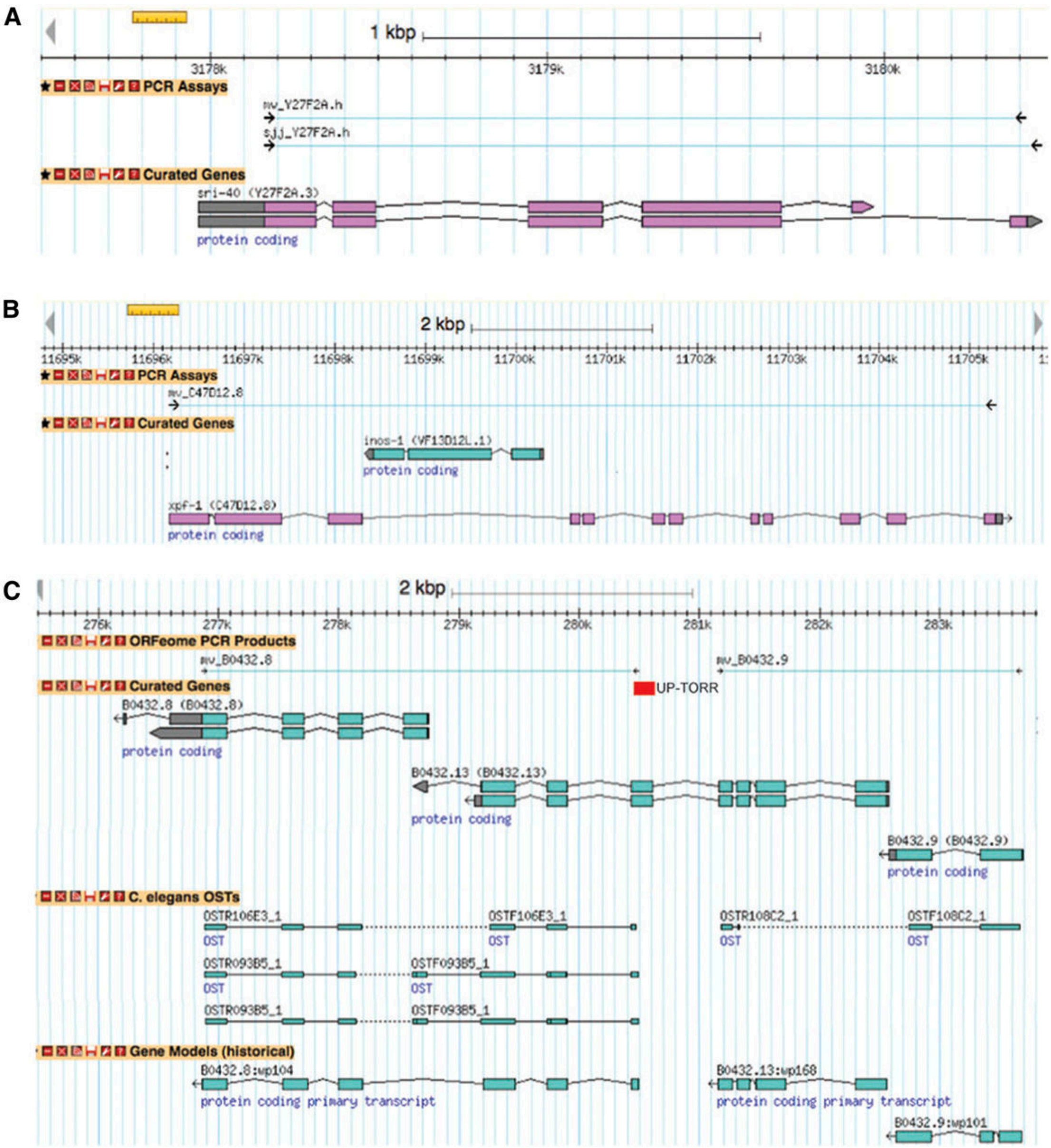


Figure 1 Limitations of current RNAi clone annotation illustrated with edited screen grabs from the Wormbase genome browser (WS242). (A) Wormbase currently reports RNAi clone sequences on the basis of genomic DNA, so that sjj_Y27F2A.h and mv_Y27F2A.h are associated with essentially identical insert sequences. (B) Wormbase consequently erroneously reports intronic genes as cDNA clone targets. In the case shown here, contrary to current Wormbase annotation, *inos-1* cannot be a target of mv_C47D12.8. (C) For certain cDNA-derived clones, the genomic positions of oligonucleotide primer pairs, designed on the basis of a historical gene model, do not correspond to a current gene model. For the left-hand ORFeome polymerase chain reaction (PCR) product, mv_B0432.8, the gene model used when the primer pair was designed is shown (B0432.8:wp168), but for the adjacent mv_B0432.9, the model is unavailable. In some cases, as shown here for the clones mv_B0432.8 and mv_B0432.9, the current gene models may require revision since there is conflicting ORF-sequence tag (OST) evidence. The extent of the PCR product predicted by UP-TORR on the basis of mv_B0432.8 primer sequences is indicated by the red rectangle. The reason for this erroneous prediction is not clear.

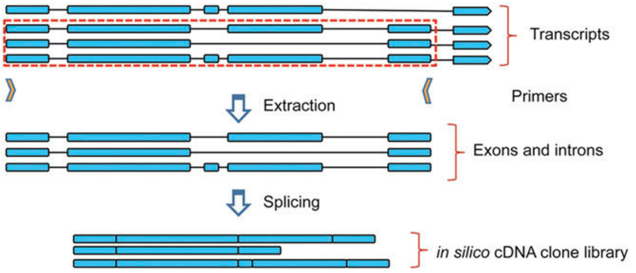


Figure 2 Schematic representation of the *in silico* construction of a library of cDNA-derived RNAi clone inserts. The genomic coordinates of each primer were compared to those of exons in a library of predicted transcripts. For each transcript that could potentially be amplified by a given pair of primers the corresponding sequence was extracted and spliced *in silico*.

"mv_" for the Ahringer or Vidal library clones, respectively) cannot be used as input to UP-TORR. It is also not well adapted to the analysis of large datasets derived from genome-wide screens. We therefore decided to construct a tool specifically for *C. elegans*, basing target identification on matching fragments of sequence generated *in silico* from the predicted inserts of RNAi clones. This is part of a collection of tools, called Clone Mapper, that also allow clone verification and sequence retrieval. It is publically available via http://www.ciml.univ-mrs.fr/EWBANK_jonathan/software.html.

MATERIALS AND METHODS

Data sources

The reference genome sequence and transcript sequences (WS235 and WS240) were downloaded from ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/. Following the Wormbase convention, transcripts corresponding to coding genes were used for the target library; those corresponding to coding genes and pseudogenes were used for the clone insert library. RNAi reagent information was extracted from the GFF3 file at ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/gff/. Since the original ORFeome primer sequences were designed (Reboul *et al.* 2001), there have been changes in the reference sequence of the *C. elegans* genome, most recently for release WS235 (see http://www.wormbase.org/about/wormbase_release_WS235). For some 500 ORFeome products, the original primer sequences no longer match to the genome (K. Howe, personal communication). New (pseudo)-primer sequences designed for these products (incorporating the change present in the WS235 genome sequence) were kindly provided by K. Howe; the relevant file is available on request.

To extract the clone-target gene pairs established by Wormbase (WS235), primary targets were retrieved from ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/annotation/pcr_product2gene; a list of secondary targets was kindly provided by C. Grove.

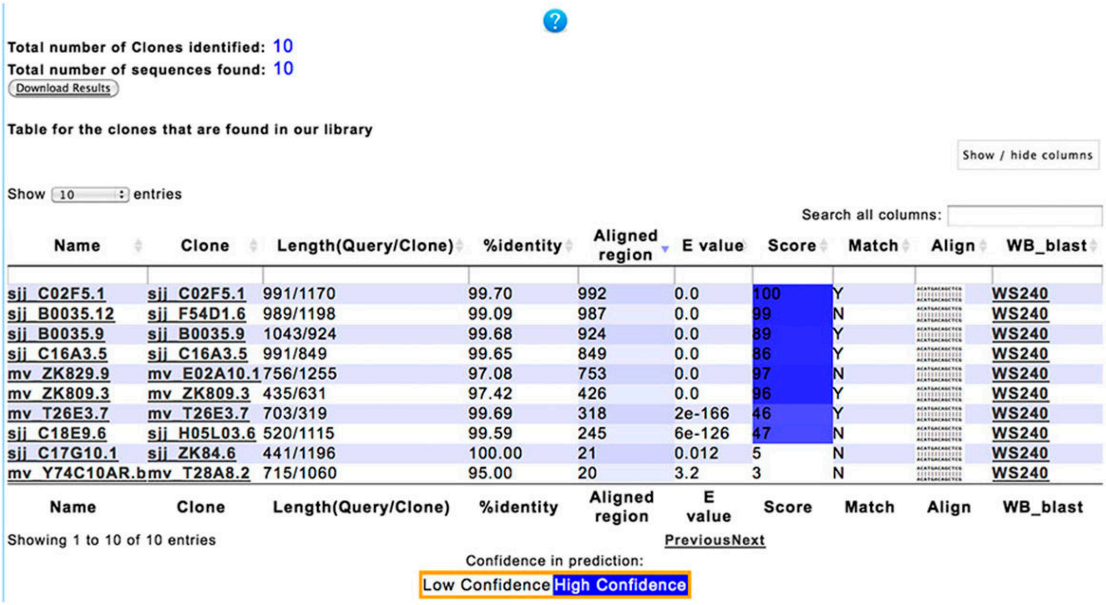
Clone-target identification

To identify potential targets of RNAi clones, we first generated all possible 21 bp fragments from the predicted sequence of each RNAi clone insert and then we searched for matches between these fragments and transcript sequences (see Figure 4B). To rank RNAi clone-target transcript pairs, we calculated a score for each pair using a simple formula:

$$\text{Score}/100 = (\text{MOS}/10)(\text{MOS}/\text{POS})(\text{MNO}/\text{PNO})^2$$

The different parameters are defined as follows:

PNO: Possible nonoverlapping segments; maximum number of nonoverlapping segments of length l that can be generated from the clone insert. By default, l = 21 bp.



Total number of Clones identified: 10
Total number of sequences found: 10

Table for the clones that are found in our library

Show 10 entries

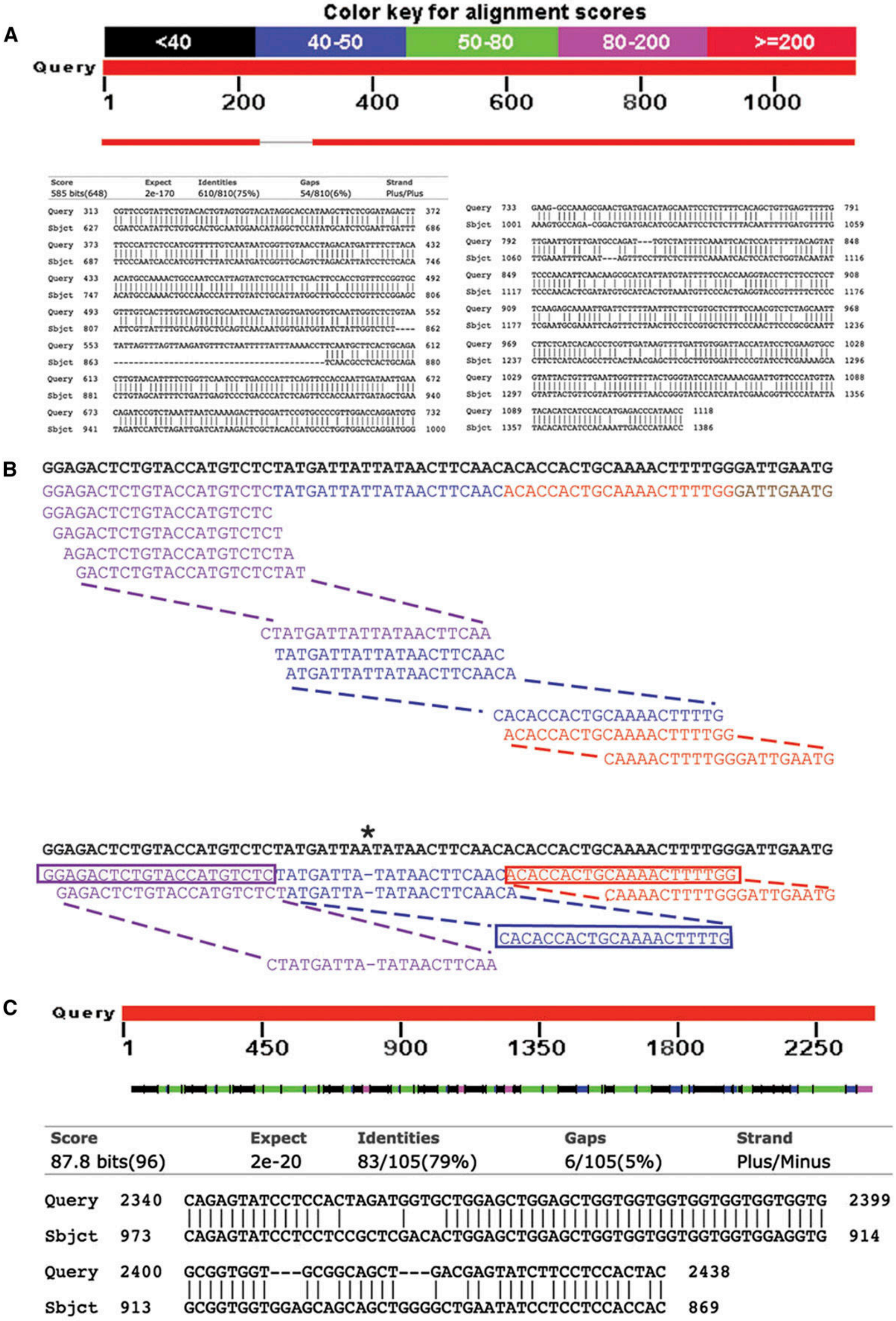
Search all columns:

Name	Clone	Length(Query/Clone)	%identity	Aligned region	E value	Score	Match	Align	WB_blast
sjj_C02F5.1	sjj_C02F5.1	991/1170	99.70	992	0.0	100	Y		WS240
sjj_B0035.12	sjj_F54D1.6	989/1198	99.09	987	0.0	99	N		WS240
sjj_B0035.9	sjj_B0035.9	1043/924	99.68	924	0.0	89	Y		WS240
sjj_C16A3.5	sjj_C16A3.5	991/849	99.65	849	0.0	86	Y		WS240
mv_ZK829.9	mv_E02A10.1	756/1255	97.08	753	0.0	97	N		WS240
mv_ZK809.3	mv_ZK809.3	435/631	97.42	426	0.0	96	Y		WS240
mv_T26E3.7	mv_T26E3.7	703/319	99.69	318	2e-166	46	Y		WS240
sjj_C18E9.6	sjj_H05L03.6	520/1115	99.59	245	6e-126	47	N		WS240
sjj_C17G10.1	sjj_ZK84.6	441/1196	100.00	21	0.012	5	N		WS240
mv_Y74C10AR.b	mv_T28A8.2	715/1060	95.00	20	3.2	3	N		WS240
Name	Clone	Length(Query/Clone)	%identity	Aligned region	E value	Score	Match	Align	WB_blast

Showing 1 to 10 of 10 entries

Confidence in prediction:
Low Confidence High Confidence

Figure 3 An example of RNAi clone identification using Clone Mapper. The DNA sequences obtained upon sequencing of 10 RNAi clones, from (Zugasti *et al.* 2014), were used as input into Clone Mapper. The results obtained, ranked by "Aligned region," are shown in this screen-grab. The leftmost column shows the library name of each clone, the next column the name of the clone that best matches the experimentally determined RNAi clone insert sequence. In this example, half the clones appeared to be what was expected; for 3 of 5 of the others, an alternative identity was assigned with high confidence. For the remaining clones only a very short sequence matches a clone in the *in silico* library. These sequences can be compared directly to the genome of *C. elegans* by clicking the link in the rightmost column. The exact meaning of the different columns and options is explained in the help document, accessible by clicking the question mark at the top of the screen.


A
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
200
400
600
800
1000
B
GGAGACTCTGTACCATGTCTCTATGATTATTATAACTTCAACACACCACTGCAAAACTTTTGGGATTGAATG
GGAGACTCTGTACCATGTCTCTATGATTATTATAACTTCAACACACCACTGCAAAACTTTTGGGATTGAATG
GGAGACTCTGTACCATGTCTC
GAGACTCTGTACCATGTCTCT
AGACTCTGTACCATGTCTCTA
GACTCTGTACCATGTCTCTAT
CTATGATTATTATAACTTCAA
TATGATTATTATAACTTCAAC
ATGATTATTATAACTTCAACA
CACACCACTGCAAAACTTTTG
ACACCACTGCAAAACTTTTGG
CAAAACTTTTGGGATTGAATG
*
GGAGACTCTGTACCATGTCTCTATGATTAATATAACTTCAACACACCACTGCAAAACTTTTGGGATTGAATG
GGAGACTCTGTACCATGTCTCTATGATTA-TATAACTTCAACACACCACTGCAAAACTTTTGG
GAGACTCTGTACCATGTCTCTATGATTA-TATAACTTCAACA
CAAAACTTTTGGGATTGAATG
CACACCACTGCAAAACTTTTG
CTATGATTA-TATAACTTCAA
C
Query
1
450
900
1350
1800
2250

```
Score            Expect    Identities      Gaps         Strand
585 bits(648)    2e-170    610/810(75%)    54/810(6%)   Plus/Plus

Query  313   CGTTCCGTATTCTGTACACTGTAGTGGTACATAGGCACCATAAGCTTCTCGGATAGACTT  372
Sbjct  627   CGATCCATATTCTGTGCACTGCAATGGAACATAGGCTCCATATGCATCTCGAATTGATTT  686

Query  373   TTCCCATTCTCCATCGTTTTTGTCAATAATCGGTTGTAACCTAGACATGATTTTCTTACA  432
Sbjct  687   TTCCCAATCACCATCGTTCTTATCAATGATCGGTTGCAGTCTAGACATTATCCTCTCACA  746

Query  433   ACATGCCAAAACTGCCAATCCATTAGTATCTGCATTCTGACTTCCACCTGTTTCCGGTGC  492
Sbjct  747   ACATGCCAAAACTGCCAACCCATTTGTATCTGCATTATGGCTTGCCCCTGTTTCCGGAGC  806

Query  493   GTTTGTCACTTTGTCAGTGCTGCAATCAACTATGGTGATGGTGTCAATTGGTCTCTGTAA  552
Sbjct  807   ATTCGTTATTTTGTCAGTGCTGCAGTCAACAATGGTGATGGTATCTATTGGTCTCT----  862

Query  553   TATTAGTTTAGTTAAGATGTTTCTAATTTTTATTTAAAACCTTCAATGCTTCACTGCAGA  612
Sbjct  863   ------------------------------------------TCAACGCCTCACTGCAGA  880

Query  613   CTTGTAACATTTTCTGGTTCAATCCTTGACCCATTTCAGTTCCACCAATTGATAATTGAA  672
Sbjct  881   CTTGTAGCATTTTCTGATTGAGTCCCTGACCCATCTCAGTTCCACCAATTGATAGCTGAA  940

Query  673   CAGATCCGTCTAAATTAATCAAAAGACTTGCGATTCCGTGCCCCGTTGGACCAGGATGTG  732
Sbjct  941   TAGATCCATCTAGATTGATCATAAGACTCGCTACACCATGCCCTGGTGGACCAGGATGGG  1000

Query  733   GAAG-GCCAAAGCGAACTGATGACATAGCAATTCCTCTTTTCACAGCTGTTGAGTTTTTG  791
Sbjct  1001  AAAGTGCCAGA-CGGACTGATGACATCGCAATTCCTCTCTTTACAATTTTTGATGTTTTG  1059

Query  792   TTGAATTGTTTGATGCCAGAT---TGTCTATTTTCAAATTCACTCCATTTTTTACAGTAT  848
Sbjct  1060  TTGAAATTTTCAAT---AGTTTCCTTTCTCTTTTCAAAATCACTCCATCTGGTACAATAT  1116

Query  849   TCCCAACATTCAACAAGCGCATCATTATGTATTTTTCCACCAAGGTACCTTCTTCCTCCT  908
Sbjct  1117  TCCCAACACTCGATATGTGCATCACTGTAAATGTTCCCACTGAGGTACCGTTTTTCTCCC  1176

Query  909   TCAAGAGCAAAATTGATTCTTTTAATTTCTTCTGTGCTCTTTCCAACGTCTCTAGCAATT  968
Sbjct  1177  TCGAATGCGAAATTCAGTTTCTTAACTTCCTCCGTGCTCTTCCCAACTTCCCGCGCAATT  1236

Query  969   CTTCTCATCACACCCTCGTTGATAAGTTTTGATTGTGGATTACCATATCCTCGAAGTGCC  1028
Sbjct  1237  CTTCTCATCACGCCTTCACTAACGAGCTTCGCTTGTGGATTCCCGTATCCTCGAAAAGCA  1296

Query  1029  GTATTACTGTTTGAATTGGTTTTTACTGGGTATCCATCAAAACGAATTGTTCCCATGTTA  1088
Sbjct  1297  GTATTACTGTTCGTATTGGTTTTAACCGGGTATCCATCATATCGAACGGTTCCCATATTA  1356

Query  1089  TACACATCATCCACCATGAGACCCATAACC  1118
Sbjct  1357  TACACATCATCCACAAATTGACCCATAACC  1386
```

```
Score            Expect    Identities      Gaps        Strand
87.8 bits(96)    2e-20     83/105(79%)     6/105(5%)   Plus/Minus

Query  2340  CAGAGTATCCTCCACTAGATGGTGCTGGAGCTGGAGCTGGTGGTGGTGGTGGTGGTGGTG  2399
Sbjct  973   CAGAGTATCCTCCTCCGCTCGACACTGGAGCTGGAGCTGGTGGTGGTGGTGGTGGAGGTG  914

Query  2400  GCGGTGGT---GCGGCAGCT---GACGAGTATCTTCCTCCACTAC  2438
Sbjct  913   GCGGTGGTGGAGCAGCAGCTGGGGCTGAATATCCTCCTCCACCAC  869
```

Figure 4 Basis of the target identification strategy. (A) An example of a target identified by Wormbase but not Clone Mapper. The RNAi clone-transcript pair (sjj_B0222.9 - F15E6.6) displays overall high identity, with blocks of >200 nucleotides with >80% identity, but does not contain

POS: Possible overlapping segments; maximum number of overlapping segments of length *l* that can be generated from the clone insert.

MNO: Matched nonoverlapping segments; number of nonoverlapping segments that are found in the targets transcript sequence; with a perfect match MNO = PNO; with no match MNO = 0.

MOS: Matched overlapping segments; number of overlapping segments that are found in the targets transcript sequence; with a perfect match MOS = POS; with no match MOS = 0.

In the score, weight is given to the MOS on the assumption that the absolute number of fragments generated from the RNAi clone insert that perfectly match a target transcript influences the probability that the target transcript will be affected. This value is divided by 10 to compensate for the inappropriate weight that would otherwise be assigned to perfect matches of small transcripts to large RNAi clone inserts. The MOS/POS ratio represents the overall sequence similarity between an RNAi clone insert and its target transcript. The more similar they are, the greater the ratio. The MNO/PNO element derives from the assumption that if different siRNAs produced by a clone insert match sequences within the target transcript, then there will be a greater chance of the target transcript being knocked down compared with when siRNAs produced from a single region of a clone insert match only one or a few sequences within the target transcript. The adjusted weight given to the MNO/PNO ratio reflects the assumption that RNAi will be more efficient when siRNAs are generated from multiple nonoverlapping segments that have the potential to target different nonoverlapping regions of a transcript.

The score was given a constant threshold of 100, so that if the calculated score exceeded 100, it was adjusted to 100. The equation for the score can be rearranged to:

$$\text{Score} = 10(\text{MNO}\times\text{MOS})^2/(\text{POS}\times\text{PNO})^2 \leq 100.$$

Software

For clone mapping, the BLAST program from the National Center for Biotechnology Information (Altschul *et al.* 1990) was locally installed and run with default parameters. Target mapping used MPScan (Rivals *et al.* 2009), with default parameters. For the comparison with published datasets of RNAi screens, when necessary, lists of target genes were updated to WS240 using Wormbase Converter (Engelmann *et al.* 2011). Network analysis used the GeneMania plugin (version 2012-08-02-core; Montojo *et al.* 2010; Saito *et al.* 2012) within Cytoscape (v2.8.1) (Shannon *et al.* 2003; Smoot *et al.* 2011). Programs for the various tools of Clone Mapper were written in Perl and the user interface was developed using HTML, PHP, JavaScript, and MySQL.

RESULTS

Construction of an *in silico* library of RNAi clones

Wormbase is the repository for a wealth of genetic, genomic, and bibliographic information. There are, however, some lacunae, such as the fact that the DNA inserts of cDNA-derived RNAi clones are not available. We therefore first constructed libraries of sequences corresponding to the expected inserts of the clones contained within the Ahringer genomic (Kamath *et al.* 2003) and Vidal cDNA-derived (Rual *et al.* 2004) RNAi collections. For the former, we also included a supplementary set of 3507 clones that recently became available. With the exception of this set, the primers made to amplify clone inserts were designed more than a decade ago. Since then, there have been minor changes in the genome sequence and more extensive changes in gene structure prediction. To correct the former problem, Wormbase calculates pseudo-primer sequences to ensure a perfect alignment between primer and genome sequence (C. Grove, personal communication). Since the Ahringer clones contain genomic inserts, generating insert sequences was relatively straightforward. The relevant coordinates were extracted from the publicly available General Feature Format (gff) file on the Wormbase ftp site and used to retrieve the corresponding genomic sequence for all of the clones. The Vidal RNAi clones are generated from the ORFeome collection. Having extracted the coordinates of the distal end of each mapped oligonucleotide primer pair (kindly provided by K. Howe, Wormbase), we calculated the proximal coordinates using the known length of each primer. The genomic coordinates of each primer were then compared with those of each transcript in an *in silico* transcript library to identify all transcripts that could potentially be amplified by a given pair of primers (Figure 2; see the section *Materials and Methods*).

For close to 15% of the clones in the Vidal library, primer pairs do not match current gene models. In the example shown in Figure 1C, the primer pair mv_B0432.8 was designed on the basis of a single gene model that existed until 2003. The predicted exons of this gene were subsequently assigned to 2 genes (B0432.8 and B0432.13). The insert sequence of clones like mv_B0432.8 that do not correspond to current gene models cannot thus be readily predicted *in silico* and we excluded these clones. This resulted in a library of 18,405 transcripts from 13,792 genes, corresponding to 88.2% of all ORFeome clones and 85.9% of the Vidal collection of 11599 RNAi clones. These sequences are available via Clone Mapper (see the section *A tool for clone verification*).

A tool for clone verification

Given the errors that are intrinsic to any large collection of clones, it is indispensable to verify that RNAi clones selected through screens correspond to what they are supposed to be. This is generally done by resequencing and comparing the obtained sequence to the genome of *C. elegans* and crosschecking the position with that expected for the

a single 21 bp contiguous stretch of identical sequence. (B) The approach implemented in Clone Mapper for defining targets of RNAi clones. The set of possible nonoverlapping 21mer fragments (PNO) are generated starting from the 5′ end of each predicted clone insert sequence (in black). In the example shown, there are 3 complete (purple, blue, red) and one partial (brown) PNOs. All possible overlapping 21mer segments (POS) are generated and assigned to the corresponding PNO; for simplicity only a selection of POS are shown for each PNO. The library of all transcripts is queried with each POS to identify matched overlapping segments (MOS). In the example shown in the lower part of the panel, a transcript (in bold) with a single difference from the clone insert sequence (*) is shown. The number of PNOs that contain at least one POS that exactly matches a given transcript is counted (MNO; here 3). An example of one matching POS for each PNO is boxed. A score is then calculated (see *Materials and Methods*). (C) An example of a target identified by Clone Mapper but not Wormbase. Here the RNAi clone insert sequence has multiple stretch of sequence that have perfect matches over more than 21 nt to a target transcript (upper part of the panel; sjj_Y50E8.g and ZK643.8a), but no contiguous region of 100 nt with 95% identity (lower part of the panel; the longest stretch of identity in the selected alignment of one fragment shown here is 31 nt).

clone. Checking in this way becomes laborious when one needs to sequence-verify tens or hundreds of clones. We therefore made a BLASTN-based tool to match experimentally determined clone sequences with our *in silico* clone sequence libraries. It returns an output showing whether the clone is the expected one, and if not what the clone is most likely to be (Figure 3). This became the first tool in a suite that we have called Clone Mapper and for which we provide a web-based access via www.ciml.univ-mrs.fr/EWBANK_jonathan/software.html. The other functionalities are described below.

Identifying potential targets of RNAi clones

Given the shortcomings of current target prediction (see above; Figure 4A), and given the known molecular basis of RNAi, we next sought to design an alternative approach based on matching short clone-derived sequences against a comprehensive collection of predicted transcript sequences. In *C. elegans*, dsRNA gives rise to siRNAs of different sizes (19−28 bp); 22 bp is the predominant length, approximately 20% are 21mers, and <10% are shorter than 21 bp (Gent *et al.* 2010). In Clone Mapper, therefore, the clone sequence is diced *in silico* into fragments of a predetermined size. By default Clone Mapper uses 21mers, corresponding to >90% of the *in vivo* siRNAs. Increasing the oligomer size would restrict the number of potential targets identified, whereas, as discussed below, decreasing the oligomer size would allow more potential targets to be captured, but at the probable expense of increasing the proportion of false-positives. The number of occurrences of each oligomer within each transcript is then counted, and a score (from 0 to 100, with 100 corresponding to a high confidence target) assigned on the basis of a simple formula (Figure 4B; see the section *Materials and Methods*). The method allows the identification of potential targets that would not otherwise be found (*e.g.*, Figure 4C).

The predicted targets (protein coding genes) for all of the Ahringer and Vidal RNAi clones in our library have been precomputed and can be retrieved by entering a clone name in Clone Mapper. Alternatively, a user can input any sequence and its potential target transcript (protein coding and/or noncoding) will be calculated *de novo*. Conversely, the identity of clones predicted to target a given gene, or set of genes, can be retrieved by entering the relevant identifiers in the query box under the "Find targets" rubric.

A comparative analysis of potential targets

To establish on a genome-wide scale how different the transcript to RNAi clone correspondences obtained with Clone Mapper were from those reported in Wormbase, we conducted a global comparative analysis. We compared the overlap between Wormbase and Clone Mapper predictions across a range of scores for each transcript-RNAi clone pair. With regards the Vidal RNAi clones, even at the greatest scores, Clone Mapper predicted essentially all (98%) of the Wormbase-predicted clone-target pairs (Figure 5A). The missing fraction all falls into the category of overpredicted (Figure 1C). On the other hand, 1865 clone-target pairs not reported in Wormbase were found. Relaxing the stringency (decreasing the cut-off score from the maximum of 100) progressively increased this number; using a cut-off score of ≥1, there were 4482, which represents an increase of 30% over the total

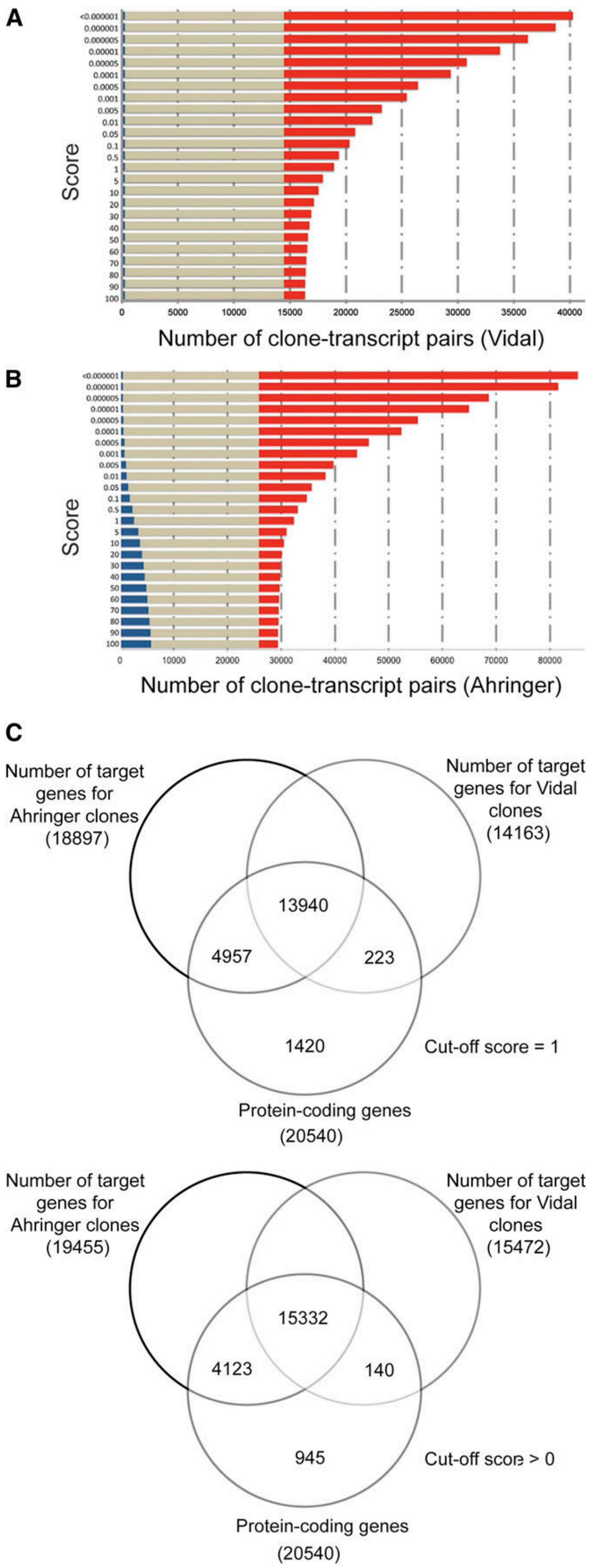


Figure 5 Target coverage with Clone Mapper. Comparison of the coverage of RNAi clone – target pairs for targets identified only with Clone Mapper (red), only by Wormbase (WS235; blue) or both (brown) at different cut-off scores for the Vidal (A) and Ahringer (B) clone collections. (C) Number of protein-coding genes identified by Clone Mapper as potential targets for the Vidal and Ahringer RNAi clones using 2 different scores (1 and the less stringent >0, upper and lower parts of the panel respectively) compared to the total number of predicted protein-coding genes (20540; WS240).

Table 1 Identification of novel target genes with Clone Mapper

	(Pukkila-Worley *et al.* 2014)	(Roy *et al.* 2014)	(Fievet *et al.* 2013)	(Ceron *et al.* 2007)
Original number of target genes	29	102	436	245
False positive (when score >0)	0	18	11	66
False positive (when score >1)	0	19	33	66
New targets with score >0	38	25	400	84
New targets with score >1	0	24	9	73

number of Wormbase predicted clone-target pairs (Figure 5A). For the Ahringer RNAi clones, when the analysis was performed with the maximum cut-off score of 100, 5825 (22.5%) of the Wormbase-predicted clone-target pairs were not found by Clone Mapper, whereas an additional 3552 were found by Clone Mapper alone. In this case, reducing the cut-off score progressively increased both the overlap between the two sets and the number of novel clone-target pairs (Figure 5B). With a cut-off score of ≥1, there were 2581 and 6539 clone-target pairs specific to Wormbase and Clone Mapper, respectively, with 23266 identified by both. This corresponded respectively to 1518, 3137 and 18664 individual RNAi clones. According to Wormbase annotations, half (49.8%) of the 3137 RNAi clones identified by Clone Mapper as potentially targeting a novel transcript (when using a cut-off score of ≥1) were previously predicted to target a single gene. As discussed below, the choice of cut-off is necessarily arbitrary, but our results, taken together with bioinformatic and experimental investigation of on- and off-target effects (Rual *et al.* 2007; Zhou *et al.* 2014), suggest that Clone Mapper can identify a substantial number of novel targets.

We also calculated the number of protein-coding genes targeted by the combined set of Vidal and Ahringer RNAi clones. Using the arbitrary cut-off score of ≥1, the entire set of clones is predicted to target a total of 19,120 of the 20,540 protein coding genes (93.1%; WS240). This figure only increases marginally, to 19,595 (95.4%), when the cut-off score is reduced to include all targets (Figure 5C).

To evaluate the potential impact of these differences in prediction, we compared the list of putative targets in four published data sets with those obtained with Clone Mapper. In the first screen, where just 29 clones were selected (Pukkila-Worley *et al.* 2014), Clone Mapper predicted the same targets as published; no novel targets with high scores were identified. In the second specific case (Ceron *et al.* 2007), 14 of 244 targets were not predicted by Clone Mapper since the insert sequences of the corresponding clones cannot be predicted. On the other hand, Clone Mapper identified 23 new targets with of score >1, 9 of which had a score >50 (Supporting Information, Table S1). Similar results were obtained for the two other studies (Fievet *et al.* 2013; Roy *et al.* 2014) (Table 1, Table S2, and Table S3). In all cases, the novel targets identified with Clone Mapper formed part of a closely linked network (Figure 6). The interconnectivity of the novel RNAi targets suggests that they may be functionally important for the biological process under study. Such a hypothesis requires direct experimental validation, but the results demonstrate the potential utility of Clone Mapper in gene discovery.

A comparison of Clone Mapper with available resources

Most published reports of RNAi experiments in *C. elegans* have relied on Wormbase for target identification. As explained previously, Wormbase has several limitations (Table 2). It does not include predictions for secondary targets for Vidal RNAi clones, and bases target identification on genomic DNA sequence, which is generally inappropriate for open reading frame−derived clones. This limitation has already been addressed in part by the web-based tool UP-TORR (Hu *et al.* 2013) that uses primer sequences to generate *in silico* a potential clone insert and then identify targets for that insert. UP-TORR, however, does not allow easy bulk clone-target mapping, or the use of the names of the Vidal library clones, for example. Furthermore, the current lower limit for stretches of sequence identity when searching for off-target genes with UP-TORR is 15 bp. This can expand the list of potential hits to an unmanageable size, especially since no score is ascribed to each clone-target pair. Clone Mapper addresses these different issues, and as a species-specific tool has been designed to be as simple and intuitive to use as possible.

DISCUSSION

With Clone Mapper, we have attempted to satisfy several unmet needs for *C. elegans* researchers using RNAi. In addition to the central function of identifying potential targets for RNAi clones, it offers tools for clone verification and for the retrieval of RNAi clone and transcript sequences. Clone Mapper complements the tools already available in Wormbase and the web-based tool UP-TORR (Hu *et al.* 2013). It can be used in conjunction with Wormbase Converter (Engelmann *et al.* 2011) (also available via http://www.ciml.univ-mrs.fr/EWBANK_jonathan/software.html) to reanalyze published RNAi datasets. As with any resource, there are certain intrinsic and extrinsic limitations. A total of 1490 Vidal clones that are present in the physical library were purportedly amplified using primers that are not compatible with current gene models. In the example shown in Figure 1C, the mv_B0432.8 primers were used successfully to amplify a cDNA. Sequencing of this PCR product supports the existence of a transcript that spans B0432.8 and B0432.13. For a subset of Vidal library clones, it might thus be possible to reconstruct their insert sequences on the basis of OST data, but the OST coverage is incomplete (see for example, mv_B0432.9 in Figure 1C), and each case would require manual inspection. In common with UP-TORR, we therefore did not attempt to resolve these inconsistencies, nor did we try to evaluate systematically whether the current gene models in question are incorrect.

The ORFeome clones that were used to construct the Vidal RNAi library were generated by amplification of cDNA. Thus, for genes with more than one mRNA isoform, the corresponding clone may contain variants with inserts differing in one or more exon. As a consequence, even when sequence data are available for a given Vidal RNAi clone, one cannot exclude the possibility that multiple different inserts might be present since clones were not always completely sequenced (generally ca. 500 bp from 5′ and 3′ primers) and the prevalence of one splice variant may mask the presence of others (J. Reboul, personal communication).

Although we did not find any inconsistency between the publicly available sequence data and sequence data generated from our in-house library (n > 70; O. Zugasti, unpublished results), to be prudent, when constructing the *in silico* clone insert library, we assumed that each Vidal RNAi clone did contain inserts corresponding to every possible transcript. If in reality not all isoforms are represented in an RNAi clone, then there will be the potential for some over-prediction of off-target genes. When the clone insert sequence is known, it can be used as the input to Clone Mapper, thus avoiding this problem.

Within Clone Mapper, the length used to search for possible matches between clone insert and target transcript can be defined by the user, with a minimum of 6 bp, so that it can be used to identify potential seed regions for miRNAs (Grosswendt *et al.* 2014) in complete *C. elegans* transcripts. It can equally be increased to ensure specificity. The minimal length of sequence identity required to obtain efficient knock-down of green fluorescent protein expression in *C. elegans* has been experimentally determined to be ≥23 bp (Parrish *et al.* 2000). It has also been reported that to observe an efficient RNAi effect, the length may vary from 30 to 50 nucleotides (Rual *et al.* 2007). We set the default oligomer length at 21 bp since this is the size of a substantial proportion of siRNAs in *C. elegans* (Gent *et al.* 2010). Increasing oligomer length will obviously reduce the number of potential targets, whereas decreasing it will broaden the set of potential targets. The different targets are assigned scores that help in the evaluation of whether a transcript is likely to be a high-confidence target. It also permits users to evaluate the consequences of setting different values for these parameters. To be inclusive but selective, one could decrease oligomer length and then set a high cut-off score. There is an element of arbitrariness in choosing oligomer length and cut-off scores, but this reflects a biological reality. The efficiency with which a given transcript is knocked down depends not only on its sequence, but also on the level at which it is expressed, the tissue that it is expressed in, and on the expression of any other transcripts that share sequence with it. Indeed, siRNAs generated from a diced primary target (secondary siRNAs) can knock-down mRNAs that are not a direct target of siRNA derived from an RNAi clone (Zhou *et al.* 2014). We did not implement this level of target identification as part of the tool, but users can search for these indirect hits by inputting the sequence of any target transcript into the *de novo* target prediction utility that is available within Clone Mapper.

The modular architecture of Clone Mapper also allows users to choose the best reagent for specifically knocking down a given gene. The identity of clones predicted to target a given gene, or set of genes, can be retrieved. Then one can check the number of off-target genes predicted for each clone, to identify the most specific clone.

Finally, an *in silico* reanalysis of selected published RNAi datasets identified new target genes. The demonstration of the functional relevance of these targets is beyond the scope of this study, but these results illustrate Clone Mapper's potential for gene discovery.

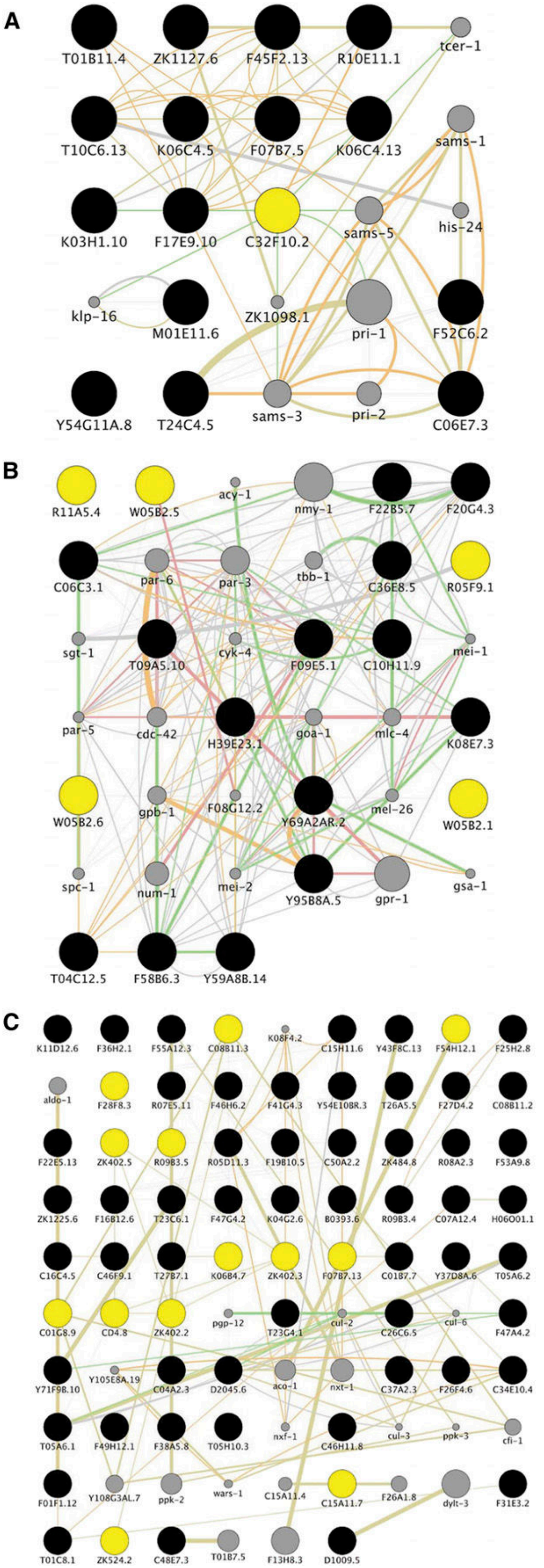


Figure 6 Network analysis of novel RNAi targets. (A) Ceron *et al.* undertook an RNAi screen to identify genes that interact with the *C. elegans* retinoblastoma gene *lin-35* (Ceron *et al.* 2007). The list of novel targets identified with Clone Mapper for the RNAi clones selected by Ceron *et al.* was used as input to GeneMania (black circles), together with *lin-35*/C32F10.2 (highlighted in yellow) as a seed gene. (B) Fievet *et al.* performed RNAi screens for *C. elegans* cell polarity mutants, to generate a polarity network (Fievet *et al.* 2013). A list of novel targets identified with Clone Mapper for the RNAi clones used by Fievet *et al.* was used as input to GeneMania (yellow circles), together with the genes corresponding to the 14 mutant strains used in the study (black circles). (C) Roy *et al.* performed a screen to find components of a regulatory network that promotes developmentally programmed cell-cycle quiescence (Roy *et al.* 2014). Novel targets identified with Clone Mapper for the RNAi clones used by Roy *et al.* (yellow), together with common targets (black) were used as input to GeneMania. The networks were trimmed to retain only direct neighbors; unconnected genes are not shown. Genes that are linked within GeneMania but do not appear on the list of RNAi clone targets are shown as gray circles; their size is proportional to the calculated probability score. Networks were displayed in Cytoscape; green edges represent experimentally-determined genetic interactions, pink edges represent experimentally-determined physical interactions for the corresponding proteins, orange and gray edges interactions predicted on the basis of co-expression or literature mining, respectively.

Table 2 Comparison of tools for RNAi experiments

	Clone Mapper	Wormbase	UP-TORR
Clone verification	Yes	No	Yes
Tool for search	Mpscan	BLAT	Blast
Insert type	Genomic and cDNA	All genomic	Genomic and cDNA
All predicted clone inserts correspond to current Wormbase gene models	Yes	N/A	No
Flexible for match of primers to gene/transcript	Yes (perfect 10 bp match at 5′ or 3′ end sufficient).	N/A (uses pseudo-primers)	No
Secondary targets	Yes	Only sjj clones	Yes
Target score	Yes	No	No
Over-prediction	No	Yes	Yes
Under-prediction	No	Yes	Yes
Batch sequence retrieval	Yes	No	No
Optimal clone search	Yes	No	No

RNAi, RNA interference.

ACKNOWLEDGMENTS

We thank the bioinformatics platforms of the CIML and the Laboratoire d'Informatique Fondamentale de Marseille for providing computing resources, database, and web servers. We are grateful to system administrators Manuel Bertrand and Kai Poutrain for their support and input. We thank the staff at Wormbase, particularly Todd Harris, Chris Grove, and Kevin Howe for their invaluable help, R. Mako Saito and R. Pukkila-Worley for providing clone lists and Ian Hope and Jérôme Reboul for comments. This work was funded by institutional grants from INSERM and CNRS, and a program grant from the ANR (ANR-12-BSV3-0001-01 to J.J.E.).

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Ceron, J., J. F. Rual, A. Chandra, D. Dupuy, M. Vidal *et al.*, 2007 Large-scale RNAi screens identify novel genes that interact with the *C. elegans* retinoblastoma pathway as well as splicing-related components with synMuv B activity. BMC Dev. Biol. 7: 30.

Engelmann, I., A. Griffon, L. Tichit, F. Montanana-Sanchis, G. Wang *et al.*, 2011 A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans*. PLoS ONE 6: e19055.

Fievet, B. T., J. Rodriguez, S. Naganathan, C. Lee, E. Zeiser *et al.*, 2013 Systematic genetic interaction screens uncover cell polarity regulators and functional redundancy. Nat. Cell Biol. 15: 103–112.

Gent, J. I., A. T. Lamm, D. M. Pavelec, J. M. Maniar, P. Parameswaran *et al.*, 2010 Distinct phases of siRNA synthesis in an endogenous RNAi pathway in *C. elegans* soma. Mol. Cell 37: 679–689.

Grosswendt, S., A. Filipchyk, M. Manzano, F. Klironomos, M. Schilling *et al.*, 2014 Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. Mol. Cell 54: 1042–1054.

Hu, Y., C. Roesel, I. Flockhart, L. Perkins, N. Perrimon *et al.*, 2013 UP-TORR: online tool for accurate and up-to-date annotation of RNAi reagents. Genetics 195: 37–45.

Imelfort, M., 2009 Sequence comparison tools, pp. 13–37 in *Bioinformatics*, edited by D. Edwards, J. Stajich, and D. Hansen. Springer, New York.

Kamath, R. S., and J. Ahringer, 2003 Genome-wide RNAi screening in *Caenorhabditis elegans*. Methods 30: 313–321.

Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin *et al.*, 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421: 231–237.

Kent, W. J., 2002 BLAT—the BLAST-like alignment tool. Genome Res. 12: 656–664.

Montojo, J., K. Zuberi, H. Rodriguez, F. Kazi, G. Wright *et al.*, 2010 GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26: 2927–2928.

Parrish, S., J. Fleenor, S. Xu, C. Mello, and A. Fire, 2000 Functional anatomy of a dsRNA trigger. Differential requirement for the two trigger strands in RNA interference. Mol. Cell 6: 1077–1087.

Pukkila-Worley, R., R. L. Feinbaum, D. L. McEwan, A. L. Conery, and F. M. Ausubel, 2014 The evolutionarily conserved mediator subunit MDT-15/MED15 links protective innate immune responses and xenobiotic detoxification. PLoS Pathog. 10: e1004143.

Qu, W., C. Ren, Y. Li, J. Shi, J. Zhang *et al.*, 2011 Reliability analysis of the Ahringer *Caenorhabditis elegans* RNAi feeding library: a guide for genome-wide screens. BMC Genomics 12: 170.

Reboul, J., P. Vaglio, N. Tzellas, N. Thierry-Mieg, T. Moore *et al.*, 2001 Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nat. Genet. 27: 332–336.

Rivals, E., L. Salmela, P. Kiiskinen, P. Kalsi, and J. Tarhio, 2009 mpscan: fast localisation of multiple reads in genomes, pp. 246–260 in *Algorithms in Bioinformatics*, edited by S. Salzberg, and T. Warnow. Springer, Berlin Heidelberg.

Roy, S. H., D. V. Tobin, N. Memar, E. Beltz, J. Holmen *et al.*, 2014 A complex regulatory network coordinating cell cycles during C. elegans development is revealed by a genome-wide RNAi screen. G3 (Bethesda) 4: 795–804.

Rual, J. F., J. Ceron, J. Koreth, T. Hao, A. S. Nicot *et al.*, 2004 Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. Genome Res. 14: 2162–2168.

Rual, J. F., N. Klitgord, and G. Achaz, 2007 Novel insights into RNAi off-target effects using *C. elegans* paralogs. BMC Genomics 8: 106.

Saito, R., M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang *et al.*, 2012 A travel guide to Cytoscape plugins. Nat. Methods 9: 1069–1076.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang *et al.*, 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504.

Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker, 2011 Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.

Squiban, B., J. Belougne, J. Ewbank, and O. Zugasti, 2012 Quantitative and automated high-throughput genome-wide RNAi screens in *C. elegans*. J. Vis. Exp. 60: e3448.

Timmons, L., and A. Fire, 1998 Specific interference by ingested dsRNA. Nature 395: 854.

Timmons, L., D. L. Court, and A. Fire, 2001 Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. Gene 263: 103–112.

Yook, K., T. W. Harris, T. Bieri, A. Cabunoc, J. Chan *et al.*, 2012 WormBase 2012: more genomes, more data, new website. Nucleic Acids Res. 40: D735–D741.

Zhou, X., F. Xu, H. Mao, J. Ji, M. Yin *et al.*, 2014 Nuclear RNAi contributes to the silencing of off-target genes and repetitive sequences in *Caenorhabditis elegans*. Genetics 197: 121–132.

Zugasti, O., N. Bose, B. Squiban, J. Belougne, C. L. Kurz *et al.*, 2014 Activation of a G protein-coupled receptor by its endogenous ligand triggers the innate immune response of *Caenorhabditis elegans*. Nat. Immunol. 15: 833–838.

Communicating editor: M. C. Zetka

2.2 Publication 2

A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes.
Zugasti O#, **Thakur N#**, Belougne J, Squiban B, Kurz CL, Soul J, Omi S, Tichit L, Pujol N, Ewbank JJ, BMC Biol, 2016
#=equal contribution
In this publication, I performed the bioinformatics analyses. My main contributions in this study were 1) providing full and easy access to the data from a whole-genome RNAi screen though the design and deployment of a dedicated web-server application (bioinformatics.lif.univ-mrs.fr/RNAiScreen/). 2) Conducting exploratory bioinformatic analyses of large sets of genes. 3) Identifying potential protein complexes through data mining. 4) Establishing an in-house integrated pipeline to provide functional class enrichment analysis and to generate phylogenetic profiles for gene lists.

Zugasti *et al. BMC Biology* (2016) 14:35
DOI 10.1186/s12915-016-0256-3

BMC Biology

RESEARCH ARTICLE Open Access

A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes

Olivier Zugasti[1,4†], Nishant Thakur[1†], Jérôme Belougne[1], Barbara Squiban[1,3], C. Léopold Kurz[1,4], Julien Soulé[1,5], Shizue Omi[1], Laurent Tichit[2], Nathalie Pujol[1*] and Jonathan J. Ewbank[1*]

Abstract

Background: *Caenorhabditis elegans* has emerged over the last decade as a useful model for the study of innate immunity. Its infection with the pathogenic fungus *Drechmeria coniospora* leads to the rapid up-regulation in the epidermis of genes encoding antimicrobial peptides. The molecular basis of antimicrobial peptide gene regulation has been previously characterized through forward genetic screens. Reverse genetics, based on RNAi, provide a complementary approach to dissect the worm's immune defenses.

Results: We report here the full results of a quantitative whole-genome RNAi screen in *C. elegans* for genes involved in regulating antimicrobial peptide gene expression. The results will be a valuable resource for those contemplating similar RNAi-based screens and also reveal the limitations of such an approach. We present several strategies, including a comprehensive class clustering method, to overcome these limitations and which allowed us to characterize the different steps of the interaction between *C. elegans* and the fungus *D. coniospora*, leading to a complete description of the MAPK pathway central to innate immunity in *C. elegans*. The results further revealed a cross-tissue signaling, triggered by mitochondrial dysfunction in the intestine, that suppresses antimicrobial peptide gene expression in the nematode epidermis.

Conclusions: Overall, our results provide an unprecedented system's level insight into the regulation of *C. elegans* innate immunity. They represent a significant contribution to our understanding of host defenses and will lead to a better comprehension of the function and evolution of animal innate immunity.

Keywords: Fungal pathogen, Functional genomics, High-throughput screening, Signal transduction, Networks, Osmotic stress, Epistasis, Bioinformatics, Databases, Mitochondrial unfolded protein response

Background

Infection of *Caenorhabditis elegans* by its natural fungal pathogen *Drechmeria coniospora* provokes an innate immune response characterized by the expression of antimicrobial peptide (AMP) genes in the worm epidermis [1]. We have focused our attention on the regulation of one group of six AMP genes of the "Neuropeptide-Like Protein" class, *nlp-27–nlp-31* and *nlp-34*, found together in a short genomic interval of less than 12 kb [2], which we call the "*nlp-29* cluster", after the best-studied member of the family. Many genes that play an essential role in controlling *nlp-29* AMP gene expression have been defined, acting together in a relatively complex genetic network. Central to this regulation is a conserved p38 MAPK cascade [3], also required for resistance to intestinal bacterial pathogens [4]. Loss of function of any one of the many genes involved provokes a "No Induction of Peptide after *Drechmeria* Infection" (Nipi) phenotype. After small- and large-scale genetic screens for Nipi mutants [3, 5], our knowledge of anti-fungal innate immunity in *C. elegans* remains, however, fragmentary. Not only are there missing elements from the associated signal transduction pathways, but how these

* Correspondence: pujol@ciml.univ-mrs.fr; ewbank@ciml.univ-mrs.fr
Olivier Zugasti and Nishant Thakur are co-first authors.
†Equal contributors
[1]Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université UM2, Inserm, U1104, CNRS UMR7280, 13288 Marseille, France
Full list of author information is available at the end of the article

pathways cross-talk with each other and with the mechanisms involved in general homeostatic regulation is currently unclear [4]. Another largely unexplored aspect of the worm's antifungal innate defenses relates to the potential for cross-tissue communication. We have demonstrated that a second family of AMP genes, called caenacins (*cnc*), including *cnc-2*, are controlled by a cell non-autonomous signal transduction pathway, wherein the nematode TGF-ß, DBL-1, produced in neurons, modulates *cnc-2* expression in the epidermis following *D. coniospora* infection. This pathway does not, however, influence *nlp-29* expression [6], which up until now has been found to be controlled cell-autonomously in the epidermis [3, 5, 7–9]. On the other hand, in *C. elegans*, the disruption of cellular homeostasis in one tissue can influence a stress response in a distant tissue (reviewed in [10–12]); whether this is also the case for *nlp-29* remains an open question.

To address these lacunae, since *C. elegans* lends itself to large-scale functional genomic analyses [13, 14], we undertook a genome-wide RNAi screen for genes involved in the regulation of the AMP gene *nlp-29*, with a well-characterized reporter gene system used in our previous studies [3]. Many pathogens can infect *C. elegans* when cultured in liquid in 96 or 384-well plates (reviewed in [15–17]). Since *D. coniospora* cannot infect worms in liquid, however, we developed a novel solid-based high-throughput assay, using the COPAS Biosort [18] to obtain a quantitative measure of reporter gene expression [19]. In a previous report, we focused on the large family of worm G-protein coupled receptor (GPCR) genes and defined a key role for DCAR-1 that acts as a "damage-associated molecular pattern" receptor, acting upstream of the p38 MAPK cascade [20]. This clearly validated the experimental approach and illustrated the utility of this large-scale reverse genetic screen for identifying individual genes.

Here, we present the full results of the screen, which led to the identification of more than 250 candidate genes. Perhaps surprisingly for such a well-studied organism, there is a relative paucity of functional information available for nematode genes, which stands as a barrier to the interpretation of large-scale studies in *C. elegans*. For example, in the recent WormBase release (WS250), only a quarter of protein-coding genes (5162/20,362) are associated with a concise description, a similar proportion (27 %) has a UniprotKB gene ontology (GO) annotation, and 57 % of them (11,970/20,362) have any type of GO annotation in WormBase. We have therefore attempted to couple several broad *in silico* analytical methods with targeted secondary screening to define groups of genes that potentially act together. In doing so, we have been able to identify several distinct biological processes that play an important role in the antifungal response and obtain, for the first time, a comprehensive view of the regulation of AMP gene expression.

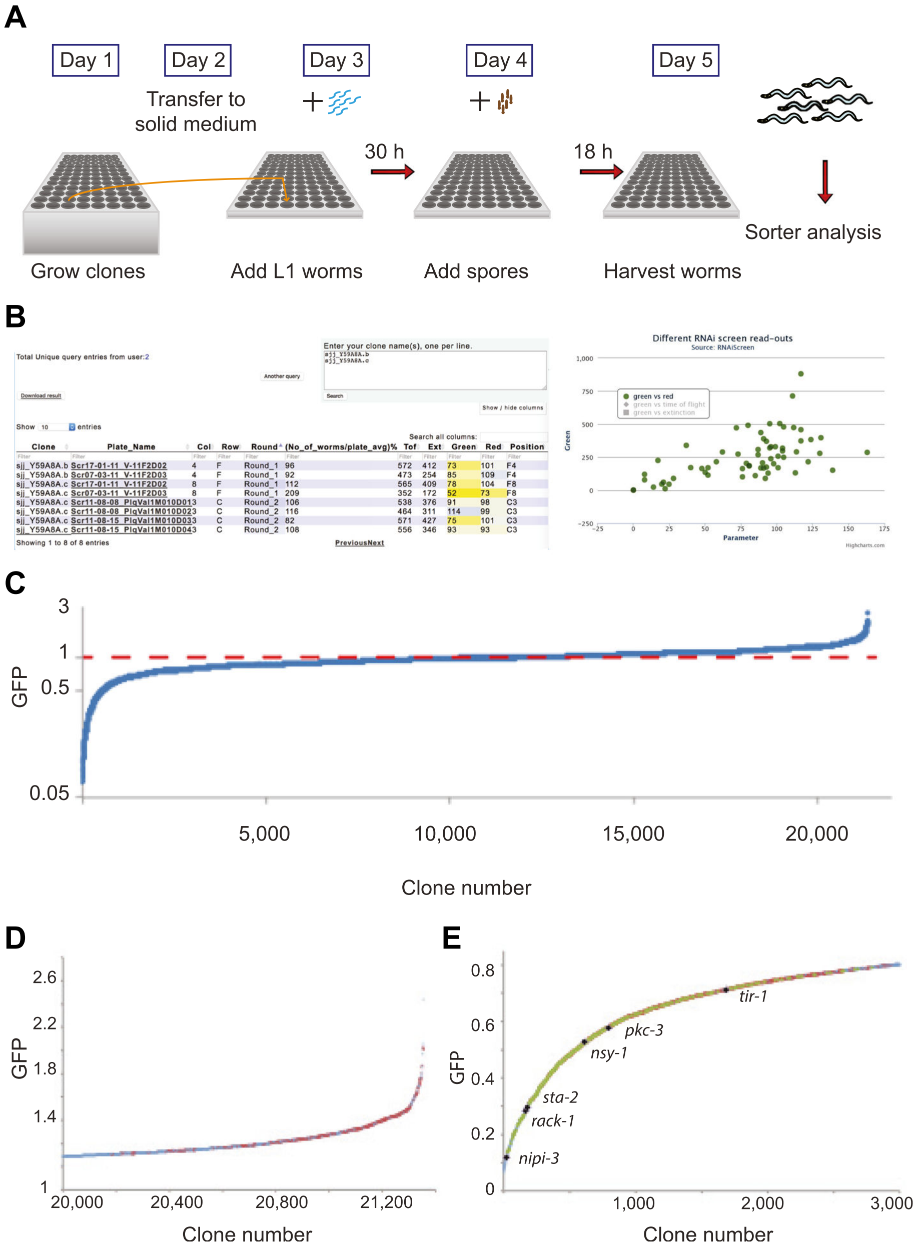
Results

A quantitative genome-wide RNAi screen for innate immunity genes

To identify, in an unbiased way, genes potentially involved in the regulation of the induction of antimicrobial peptide genes, we conducted a genome-wide RNAi screen. We first assembled a non-redundant collection of 21,223 RNAi clones from the Ahringer [21] and Vidal [22] libraries (Additional file 1: Table S1). Together, these clones are predicted to target 85 % of the protein coding genes in *C. elegans*. Using an automated method [19], we screened this library twice and quantified the infection-induced expression of the *nlp-29p::gfp* reporter gene in young adult worms (Fig. 1a). The entire set of results, a total of more than 46.8 million data points, including measures of body length (time of flight; TOF), optical density, and reporter gene expression (green (GFP) and red (dsRed)) from the analysis of more than 3.9 million individual worms, is publicly available and can be queried via a dedicated web interface (http://bioinformatics.lif.univ-mrs.fr/RNAiScreen; Fig. 1b). The overall continuous, but far from normal, distribution of the results from the first round of screening (Fig. 1c) is very much in line with previous quantitative large-scale screens in other organisms [23].

Identification of clones that provoke an exaggerated response

Innate immune responses are limited by negative regulators that contribute to protecting hosts from the collateral damage of their own effector mechanisms [24, 25]. There is emerging evidence that excess NLP-29 can damage host tissue (Dong Yan, Duke University, personal communication). With the aim of identifying negative regulators of the response, in a first step, we retained 295 clones that provoked an average increase of *nlp-29p::gfp* expression of 30 % or more, but that either did not increase the expression of a control transgene, the constitutive epidermal reporter *col-12p::dsRed*, nor the average size of the worms, or if they did, the increase was less than 30 % (Additional file 2: Table S2; Fig. 1d). Inactivation of numerous genes that affect molting, such as *pan-1* [26], the integrity of the cuticle, including *dpy-9*, *osm-11* [2], and *acs-3* [27], or fatty acid metabolism (e.g., *fasn-1*), is known to provoke the "peptide expression no infection", or Peni phenotype: an elevation of *nlp-29p::gfp* expression in the absence of infection [8]. This is associated with an exaggeratedly high expression after infection too [8], which we call here the



Enter your clone name(s), one per line.

```
sjj_Y59A8A.b
sjj_Y59A8A.c
```

Search

Total Unique query entries from user:2

Another query

Download result

Show / hide columns

Show 10 entries

Search all columns:

Clone	Plate_Name	Col	Row	Round	(No_of_worms/plate_avg)%	Tof	Ext	Green	Red	Position
sjj_Y59A8A.b	Scr17-01-11_V-11F2D02	4	F	Round_1	96	572	412	73	101	F4
sjj_Y59A8A.b	Scr07-03-11_V-11F2D03	4	F	Round_1	92	473	254	85	109	F4
sjj_Y59A8A.c	Scr17-01-11_V-11F2D02	8	F	Round_1	112	565	409	78	104	F8
sjj_Y59A8A.c	Scr07-03-11_V-11F2D03	8	F	Round_1	209	352	172	52	73	F8
sjj_Y59A8A.c	Scr11-08-08_PlqVal1M010D01	3	C	Round_2	106	538	376	91	98	C3
sjj_Y59A8A.c	Scr11-08-08_PlqVal1M010D02	3	C	Round_2	116	464	311	114	99	C3
sjj_Y59A8A.c	Scr11-08-15_PlqVal1M010D03	3	C	Round_2	82	571	427	75	101	C3
sjj_Y59A8A.c	Scr11-08-15_PlqVal1M010D04	3	C	Round_2	108	556	346	93	93	C3

Showing 1 to 8 of 8 entries

PreviousNext

Fig. 1 A quantitative genome-wide screen for regulators of AMP gene expression. **a** Simplified overview of the RNAi screen protocol, adapted from [19]. **b** Screenshots from the RNAi screen web interface. Left panel: example of results for two clones (insert at top right: the query box) that target the gene *gck-3*. Contrary to clone sjj_Y59A8A.c that passed the first round of duplicate screening, and for which the results of the second (quadruplicate) round are also displayed, sjj_Y59A8A.b only provoked a 15 % reduction in normalized GFP expression in one of the two first-round tests and so was not retained for the second round. The results for each test are linked to the primary data, which is displayed in the right panel for a single experiment. Users have the option of plotting GFP fluorescence against any or all of three parameters; shown here is GFP versus dsRed expression (in arbitrary units). **c** The ranked averages of the two values for normalized GFP expression for each of the 21,355 RNAi clones tested (21,223 unique clones, 132 present in duplicate), on a log scale. **d** The averages, on a linear scale, of the two values for normalized GFP expression for the last 1,355 RNAi clones. The 295 clones that were retested in a second round are indicated in red. **e** The averages, on a linear scale, of the two values for normalized GFP expression for the first 3000 RNAi clones. The 966 and 360 clones that passed first and second round screening are indicated in red and green, respectively. The results for selected known signaling components are indicated in black

Hipi phenotype (for hyper-induction of peptide expression after infection). To identify clones that caused only a Hipi phenotype, the 295 clones were retested in quadruplicate for their effect on *nlp-29p::gfp* expression, both with and without infection. Using cut-offs that captured all the positive controls (*fasn-1* and *pan-1*) but none of the negative controls (*sta-1* and *K04G11.4* [7]), we removed 21 clones that robustly caused a Peni phenotype (Additional file 3: Table S3). Their characterization will be the subject of a future study.

We then used a simple cut-off to classify 28 clones as being capable of causing a strong Hipi phenotype (termed, "Hipi clones"; Additional file 3: Table S3). We used sequencing and Clone Mapper [28] to verify the identity of the Hipi clones and determine their putative target genes (Additional file 3: Table S3). These included *bus-2* and *bus-12*, which respectively encode a galactosyltransferase and a sugar transporter required for the post-translational modification of surface-exposed proteins [29, 30]. In a detailed analysis, we previously demonstrated that abrogating *bus-2* or *bus-12* function increases spore binding to the nematode cuticle [31]. To address the question of whether the infectious burden of spores affected the strength of reporter gene expression, we exposed wild-type worms carrying *nlp-*

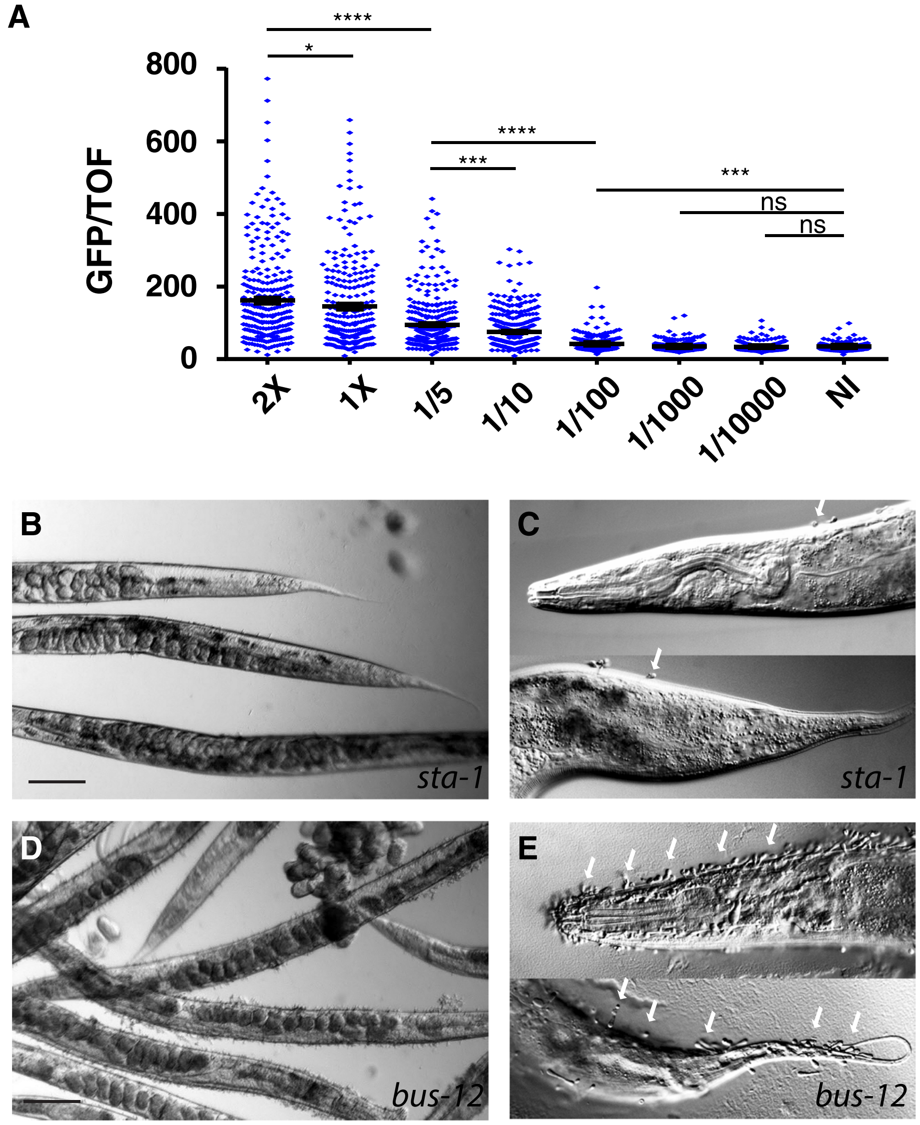


Fig. 2 Infection burden affects the strength of the innate immune response. **a** Normalized fluorescence ratio for worms infected for 18 h with the indicated dilutions of a solution of fresh *D. coniospora* spores, compared to non-infected (NI) worms. In each sample, a minimum of 230 worms was analyzed. The bar indicates the mean value. Since spore virulence depends on the age of the spores and of the plate from which they were harvested [121], the absolute spore concentration is not an informative measure and is not shown here. Comparisons between selected conditions are shown (Mann–Whitney test); ns, not significant; * $P < 0.05$; *** $P < 0.001$; **** $P < 0.0001$. **b**–**e**. Comparison at lower (**b**, **d**) and higher (**c**, **e**) magnification between worms treated with a control RNAi (*sta-1*; **b**, **c**) or RNAi against *bus-12* (**d**, **e**). In contrast to the control worms, *bus-12*(RNAi) animals exhibited a very markedly increased adhesion of spores (white arrows) over the entire body (**c**), prominently at the head and tail (**e**). Scale bar in b and d: 50 µm

29p::gfp to varying doses of *D. coniospora* spores. There was a clear relationship between the concentration of spores and the level of GFP expression (Fig. 2a).

We therefore conducted a third round of screening, this time directly assessing the adhesion of spores to worms treated with the 28 candidate Hipi clones and, in parallel, the degree of expression of *nlp-29p::gfp* relative to worms treated with a control RNAi clone targeting *sta-1*. Half of the clones were again scored as provoking a Hipi phenotype and in each case this was associated with a clear increase in spore binding (Additional file 3: Table S3, Fig. 2b–e). Among the predicted targets of these 14 clones, in addition to *bus-2* and *bus-12*, three other genes are putatively involved in the modification of surface glycans (Table 1). The clone mv_Y38C1AB.5 potentially targets two paralogous genes encoding glycosyltransferases, and is also likely to affect the properties of the cuticle via an effect on surface glycoprotein biosynthesis. Similarly, another predicted target gene, *K08E3.5*, encodes a uridine triphosphate-glucose-1-phosphate uridylyltransferase, expected to be involved in glycoprotein and glycolipid synthesis. While the connection between spore binding and the remaining target genes is less evident and will require further investigation, these results advance our understanding of the interaction between fungal spores and nematode cuticle and emphasize the cardinal importance of the infection burden in determining the strength of the innate immune response.

Table 1 Targets of robust Hipi clones

Gene/sequence name	Brief description
bus-2	Core-1 beta1,3 galactosyltransferase[b]
bus-12	Nucleotide-sugar transporter[b]
cpt-6	Carnitine palmitoyltransferase
sdc-2	Nematode-specific; required for dosage compensation
snf-9	Solute carrier family 6 (SLC6)
tkt-1	Transketolase
ykt-6	v-SNARE
C14H10.3	Pyridoxal-dependent decarboxylase
K08E3.5	Uridine triphosphate:glucose-1-phosphate uridylyltransferase[b]
F35H12.5	Epoxide/serine hydrolase
K06A9.1	Nematode-specific; limited similarity to mucin
T04G9.4	Aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase
Y38C1AB.1[a]	Core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase[b]
Y38C1AB.5[a]	Core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase[b]
C26B9.3	Nematode-specific

[a]Targeted by a single RNAi clone
[b]Involved in carbohydrate metabolism

Identification of clones that abrogate the response

To identify positive regulators of the response, following the first round of screening, we retained clones that reduced the infection-induced expression of *nlp-29p::gfp* by 20 % or more in both of the tests (i.e., provoked a Nipi phenotype), but excluded those that altered the expression of the control *col-12p::dsRed* transgene or reduced the average size of the worms more than when we knocked-down the known signaling component *rack-1* [9]. The selected 966 Nipi clones were then tested in quadruplicate, and 360 clones giving a robust Nipi phenotype (Additional file 4: Supplementary Methods) were chosen for further study (Fig. 1e and Additional file 5: Table S5; full results available at http://bioinformatics. lif.univ-mrs.fr/RNAiScreen). Among them, 314 clones were predicted to target a single gene, 22 to target two genes each, two to target three genes each, and eight clones an average of 10 each (due to targeting of multigene families of very similar sequence, e.g., the *his* histone genes; see below). For the remaining 14 clones, no target could be defined because of sequence ambiguity (Additional file 5: Table S5). A total of 404 genes were thus identified as potential targets for these clones, with 15 genes, not counting the many *his* genes, satisfyingly, being hit by more than one clone. The complete RNAi library contained a further 155 clones that potentially hit the same 404 targets (Additional file 5: Table S5). If the clones we have identified are true positives, then these 155 clones should also have been retained. This gives a measure of the efficiency of the screen (360/(360 + 155), 70 %); in common with other RNAi screens (e.g., [32, 33]), false negatives are thus a notable limitation here. Among the identified targets, in addition to *rack-1*, we noted the presence of *nipi-3*, *nsy-1*, *pkc-3*, *sta-2* and *tir-1*, all previously characterized for their role in regulating antimicrobial peptide gene expression [1, 3, 7, 9]. Further, among the clones, we identified one targeting *dcar-1*, encoding a GPCR that we demonstrated to be required for the induction of the innate immune response upon infection [20]. We consider this to be a validation of the screening and selection method.

Going from a genome-wide approach to a more focused analysis allowed attention to be paid to the phenotypes of the worms that had been treated with each of the 360 RNAi clones. A number of clones provoked developmental delays and/or lethality under our experimental protocol. We used the quantitative data from the second round of screening to identify 63 clones associated with pronounced phenotypes (see Additional file 6: Table S4 for criteria). Of these 63 clones, 42 had been associated with severe developmental phenotypes in previous RNAi studies (Additional file 6: Table S4). While we cannot formally exclude the possibility that they exercise both an essential developmental role and a specific role in regulating innate immune defenses, because of their pleiotropic effects they were not included in subsequent analyses. This left us with a list of 297 clones, predicted to target 338 genes (Additional file 5: Table S5), including all the previously characterized genes mentioned above. This list contains many potentially interesting genes, such as *akir-1*, which encodes the worm ortholog of Akirin, a known regulator of innate immunity in flies and mammals [34, 35], the claudin/calcium channel gamma subunit family gene *nsy-4*, known to act upstream of *nsy-1* during neuronal development [36], as well as several genes encoding transcription factors.

Inspection of this list also revealed a potential confounding factor for subsequent analyses. The prediction of targets for an RNAi clone is based on sequence. Several *C. elegans* gene families contain multiple members with highly similar nucleotide sequences, so that a single RNAi clone can have many potential targets. In addition to the clone sjj_K07F5.1, predicted to hit 15 *msp* genes, this principally concerned clones targeting histone genes; a total of 45 *his* genes were identified as potential targets for just eight RNAi clones (Additional file 5: Table S5). In the absence of functional analyses at the single gene level, it is not possible to ascribe the effect of a given RNAi clone to one or multiple targets. Many published genome-wide RNAi screens in *C. elegans* have reported target genes but not RNAi clones, and therefore potentially suffer from this confounding factor that can lead to biases in analyses. For some of our subsequent analyses, we removed these two gene classes, giving a set of 288 clones potentially targeting 278 (Nipi non-*his* non-*msp*) genes (Tables 2 and 3).

The central role of MAPK signaling

The main signaling pathway known to regulate *nlp*-29 expression has at its core a conserved p38 MAPK cassette involving *tir-1*, *nsy-1*, *sek-1*, and *pmk-1* [2, 3]. The identification in the screen of *tir-1* and *nsy-1* was an important validation of the approach. The failure to identify *sek-1* represents a clear example of a false negative arising because of our deliberate selection strategy. The corresponding clone (sjj_R03G5.2) did not pass the first round of screening since it only abrogated reporter gene expression in one of the two trials. For *pmk-1*, as the corresponding clone (mv_B0218.3) surprisingly did not provoke a Nipi phenotype, we sequenced it. In common with the 54/388 candidate clones that we sequenced (4/28 Hipi and 50/360 Nipi clones, respectively; Additional file 3: Tables S3 and Additional file 5: Table S5), in our cherry-picked library, the clone annotated as mv_B0218.3 did not contain the expected insert. We returned to our original copy of the Vidal RNAi collection; the corresponding clone there was also incorrect. This is another of the known drawbacks in large-scale RNAi screens: the impossibility of being certain of the identity of every clone.

MAPK signaling is central to innate immune defense in many species, including *C. elegans* and vertebrates [25, 37]. Catalogs of proteins and genes involved in the regulation of MAPK pathways are available for yeast [38], flies [39], and human [40]. Using the Drosophila RNAi Screening Center (DRSC) Integrative Ortholog Prediction Tool [41], we compared the worm orthologs of the members of these lists with our hits. More than 1/5 of the candidates (76/338 Nipi genes; 22 %) had previously been associated with MAPK signaling in at least one other species. The constitution of the list of potential components of MAPK signaling was skewed by the inclusion of multiple histone genes (43/76; Additional file 5: Table S5). Nevertheless, the identification of 33/

Table 2 Overview of screen results

Step	Type of clone	Number of clones	Source
Primary screen	Library	21,223	Additional file 1: Table S1
Secondary screen for increased reporter gene expression	Candidate	295	Additional file 2: Table S2
	Peni	21	Additional file 3: Table S3
	Candidate Hipi	28	Additional file 3: Table S3
Tertiary screen for increased reporter gene expression	Hipi	14	Additional file 3: Table S3
Secondary screen for decreased reporter gene expression	Candidate Nipi	966	http://bioinformatics.lif.univ-mrs.fr/RNAiScreen
Tertiary screen for decreased reporter gene expression	Nipi	360	Additional file 6: Table S4
With pronounced developmental phenotype		63	Additional file 6: Table S4
Retained	Nipi	297	Additional file 5: Table S5
Clones targeting *his* or *msp* genes		9	Additional file 5: Table S5
Remainder	Nipi -*his* -*msp*	288	Additional file 5: Table S5

Table 3 Enrichment of functional classes among 278 Nipi genes

Category/Phenotype upon RNAi[a]	Number over-lapping genes	Total number genes in class	Percentage of MAPK signaling genes in class	Probability[b]
1. Conserved in *D. melanogaster*[c] as described in [124]	169	3710	84.8	3.77×10^{-54}
2. Stimulate microbial aversion behavior[c] [26]	65	374	33.3	1.39×10^{-50}
3. Suppress over-expression of *gpdh-1*p::*gfp* seen in *osm-8* mutant[c,d] [72]	52	252	24.2	7.87×10^{-44}
4. Decrease *acdh-1p::gfp* expression [73]	35	146	6.1	3.33×10^{-31}
5. Induce *hsp-6p::gfp* (mitochondrial UPR) [66]	30	95	0.0	1.60×10^{-30}
6. Protein expression [125]	43	446	9.1	1.18×10^{-21}
7. Required for cytoprotective response – four different reporter genes[c] [49]	21	71	24.2	3.08×10^{-20}
8. Required for transgene silencing [57]	50	823	24.2	1.25×10^{-16}
10. Alter RAB-11 sub-cellular localization and transport of the apical membrane protein PEPT-1 [79]	34	426	15.2	3.40×10^{-14}
11. Synthetic phenotype with *lin-35* [126]	27	252	6.1	3.95×10^{-14}
12. Required for mitochondrial surveillance and response [74]	14	45	9.1	2.08×10^{-13}
13. Regulators of *gpdh-1* expression[d] [75]	18	123	6.1	2.45×10^{-11}
15. Aberrant GFP::PGL-1 phenotypes [76]	20	170	12.1	6.87×10^{-11}
16. Suppressors of polyglutamine aggregation [77]	19	173	9.1	9.65×10^{-10}
17. Mitochondrial [125]	17	160	0.0	2.35×10^{-8}
18. Upregulated after 12 h of dauer recovery [127]	44	1175	12.1	3.31×10^{-7}
20. Altered expression in *zpf-1* mutant [128]	21	333	3.0	2.24×10^{-6}
21. Down regulated in an *ogt-1* mutant [129]	30	670	9.1	4.02×10^{-6}
22. Altered expression in *rde-4* mutant [128]	17	232	6.1	7.17×10^{-6}
23. Induce *hsp-70p::gfp* [130]	9	52	9.1	1.03×10^{-5}
24. Genes connected to miRNA function [131, 132]	8	42	9.1	2.79×10^{-5}
25. Increase longevity [133]	9	66	3.0	8.78×10^{-5}
26. Induced after 24 h of *D. coniospora* infection (cDNA microarrays) [2]	21	419	12.1	1.12×10^{-4}
27. Confer hypoxia-resistance [134]	14	198	9.1	1.98×10^{-4}
29. Induce *irg-1*p::*gfp* [59]	10	102	0.0	4.37×10^{-4}
31. Energy generation [125]	10	104	3.0	5.23×10^{-4}

[a] The numbers refer to the order of the classes in the complete analysis; some redundant or similar classes have been removed (see Additional file 7: Table S6 for complete data)

[b] Bonferroni-corrected Fischer exact score

[c] Class significantly enriched in the group of 33 "non-*his* non-*msp* MAPK pathway genes"

[d] Class related to osmotic stress response

278 (non-*his* non-*msp*) MAPK-related genes (Table 4) reinforces the idea that MAPK signaling is central to the regulation of AMP gene expression in *C. elegans* epidermis and underscores the conserved nature of this core signaling process.

To explore the functional relationship of the non-histone candidate genes potentially involved in MAPK signaling, we submitted them to an analysis using WormNet, a phenotype-centric tool that represents known interactions between genes in a list [42]. Of the 33 genes entered, 28 formed a well-connected network (Fig. 3a). On the basis of an input list of genes, WormNet predicts other genes that could be functionally related to them, ranked by probability. Among the top 200 WormNet candidates, there were 53 genes that were found as candidate Nipi genes in our screen but that had not been included in the original query of 33 genes. At first sight, this remarkable enrichment would appear to be a testament to the predictive power of WormNet. Inspection of the results, however, showed that 38 of these 53 genes encode histones (Additional file 7: Table S6), which generally share functional annotations. Nevertheless, 15 non-histone genes that had been found in our screen were identified as potentially linked to the MAPK network (Additional file 7: Table S6), suggesting that other WormNet candidates could also be involved

Table 4 Nipi genes linked to MAPK signaling

Gene/sequence name	Brief description
ccf-1	Subunit 7 of CCR4-NOT transcription complex
cct-3	Gamma subunit of eukaryotic cytosolic ('T complex') chaperonin
cic-1	Cyclin C
cyl-1	Cyclin L
dic-1	DEAD/H BOX 26; mitochondrial
dnc-1	Dynactin complex subunit p150; DNC-1 is located at cortical microtubule attachment sites
ego-2	Bro1 domain-containing protein; positive regulator Notch signaling
exos-9	Exosome component 9
ftt-2	14-3-3 protein
hda-1	Histone deacetylase 1
ima-3	Importin alpha nuclear transport factor
kin-20	Casein kinase
let-92	Catalytic subunit of PP2A (protein phosphatase 2A)
nap-1	NAP (Nucleosome Assembly Protein) family
npp-1	Nucleoporin
npp-10	Nucleoporin
ogdh-1	2-oxoglutarate dehydrogenase, mitochondrial
puf-9	PUMILIO RNA-binding protein
pyp-1	Inorganic pyrophosphatase; predicted to participate in nucleosome remodeling
rack-1	Receptor for Activated C Kinase; homolog of G beta [9]
rbpl-1	E3 ubiquitin-protein ligase RBBP6 (Retinoblastoma binding protein 6)
rps-26	Small ribosomal subunit S26
tba-2	Alpha-tubulin
tba-4	Alpha-tubulin
unc-37	Gro/TLE (Groucho/transducin-like enhancer)
vhp-1	MAP kinase phosphatase
wnk-1	With no lysine kinase [8]
xpo-2	Importin-beta
F19F109	U4/U6U5 tri-snRNP-associated protein 1
K12H44	Signal peptidase complex subunit 3
R1863	Signal recognition particle receptor subunit beta
F20D122	Germinal-center associated nuclear protein; required for mRNA export

References are given for genes previously connected to the regulation of *nlp-29* expression. Each gene is targeted by a different RNAi clone

in the regulation of *nlp-29p::gfp* expression. Further, this analysis underlines the idea that the genes involved in modulating MAPK signaling are embedded within a broader cellular signaling network.

One of the five MAPK-related candidate target genes that was not part of the MAPK network predicted by WormNet (Fig. 3a) was *vhp-1*, which encodes a member of the VH1 dual-specificity phosphatase family. Since *vhp*-1 has been described as a negative regulator of the p38 pathway [43], we would not have expected to have found it as a Nipi gene. The effect of *vhp-1*(RNAi) on *nlp-29p::gfp* expression after *D. coniospora* infection was very pronounced (Fig. 3b). To determine whether this effect was cell-autonomous, we made use of an epidermis-specific RNAi strain, IG1502 [20]. To our surprise, *vhp-1*(RNAi) provoked a substantial ectopic expression of *gfp* in the intestine in this strain, even in the absence of infection (Fig. 3c–f). The intestine of *C. elegans* is functionally regionalized [44]; *vhp-1*(RNAi)-induced *nlp-29p::gfp* expression was strongest in the posterior intestinal cells (Fig. 3g). Thus, reducing the activity, specifically in the epidermis, of a phosphatase previously shown to down-regulate p38 MAPK signaling leads to ectopic gene expression of a p38 MAPK target in a distant tissue.

Global functional analyses

Returning to a more global analysis, when we submitted the list of Nipi genes (except *his* and *msp* genes) to a WormNet analysis, 231 formed an intensely interconnected network with an average of 11.1 edges/node (Fig. 4). To characterize this broader cellular signaling network, we ran the lists of candidate targets through an Expression Analysis Systematic Explorer (EASE) analysis [45], using our in-house database of functional annotations [46]. A number of classes were significantly enriched ($P < 10^{-3}$; Additional file 7: Table S6). Several were derived from early genome-wide ChIP-seq studies produced by the model organism encyclopedia of DNA elements (modENCODE) consortium. We did not exploit this data further since its reliability has recently been questioned by the consortium itself [47].

Most of the genes (245/278) were found in at least one significantly enriched class (Additional file 7: Table S6). The different classes were more or less related (Fig. 5); for example, genes associated with the stability, localization, and function of P granules (class 15 in Table 3) clustered with those associated with Rab11-positive recycling endosome-linked transport (class 10). Many genes belonged to several functional classes; the most frequently found (in 12/34 classes, Additional file 7: Table S6) encodes the E2 ubiquitin-conjugating enzyme LET-70. The most significantly enriched class was for genes defined as being conserved in *Drosophila* (through a pairwise comparison with *C. elegans*), followed by those previously described as stimulating microbial aversion behavior when knocked down by RNAi (classes 1 and 2 in Table 3, respectively). The latter includes genes involved in diverse essential cellular functions [26]. Several other classes linked to stress responses were also highly enriched. One of the most populated classes (50/278) was of genes previously characterized as being necessary for RNAi (class 8 in Table 3).

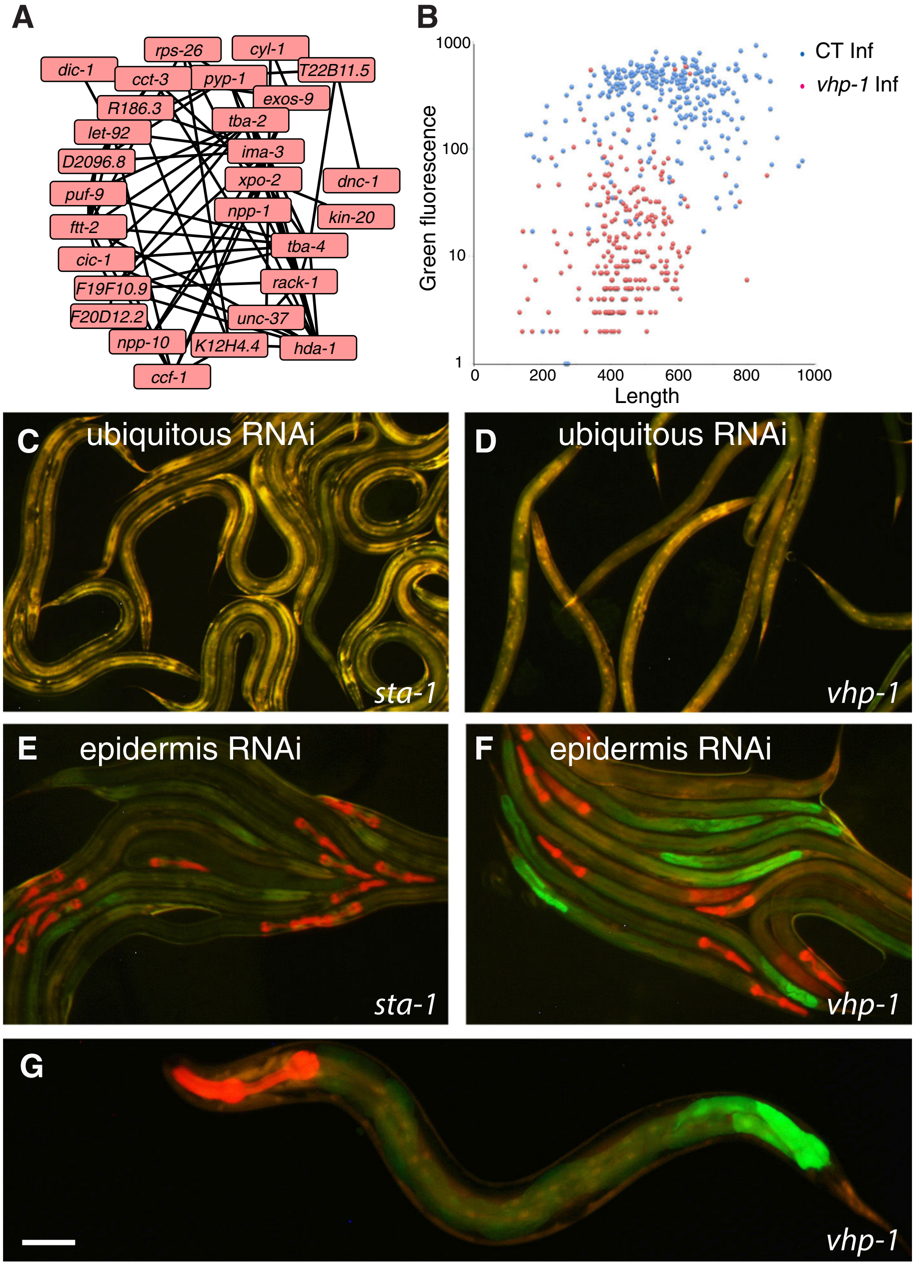


Fig. 3 MAPK pathway genes involved in regulation of AMP expression. **a** Interaction network predicted by WormNet for 33 MAPK pathway-related Nipi genes (Additional file 7: Table S6). The genes *ego-2*, *tag-214*, *vhp-1*, and *wnk-1* are not connected to any other of the genes; Y73B3A.18 is not included in the WormNet set of genes. These five genes are not shown here. The remaining 28 genes are connected by 77 edges. As for all large-scale data mining, there are obvious omissions, due to incomplete coverage in databases. One example is *nsy-1* that encodes a MAP3K [122] but does not appear here. **b** Knocking down *vhp-1* by RNAi provokes a marked reduction of *nlp-29p::gfp* reporter gene expression in IG274 worms carrying the *frIs7* transgene infected by *D. coniospora*. The graph shows the quantification of fluorescence of worms treated with control (CT: K04G11.4; blue; *n* = 296) or *vhp-1* (red; *n* = 258) RNAi. The green fluorescence and length are plotted in arbitrary, but constant units. **c**–**g** A cell-non-autonomous regulatory role for *vhp-1*. Whereas following systemic knock-down of *sta-1* (**c**) or *vhp-1* (**d**), there was no detectable expression of the *nlp-29p::gfp* reporter gene in IG274 worms carrying the *frIs7* transgene in the absence of infection, knocking-down *vhp-1* (**f**) but not *sta-1* (**e**) specifically in the epidermis in strain IG1502 led to ectopic expression of GFP in the nematode intestine. This expression was most pronounced in the posterior intestinal cells (**g**). The red fluorescence in the pharynx in (**e**–**g**) reflects the presence of an additional transgenic marker. Worms were observed at the L4 stage in all cases. Green and red fluorescence are visualized simultaneously with a long pass GFP filter. Scale bar: 50 μm

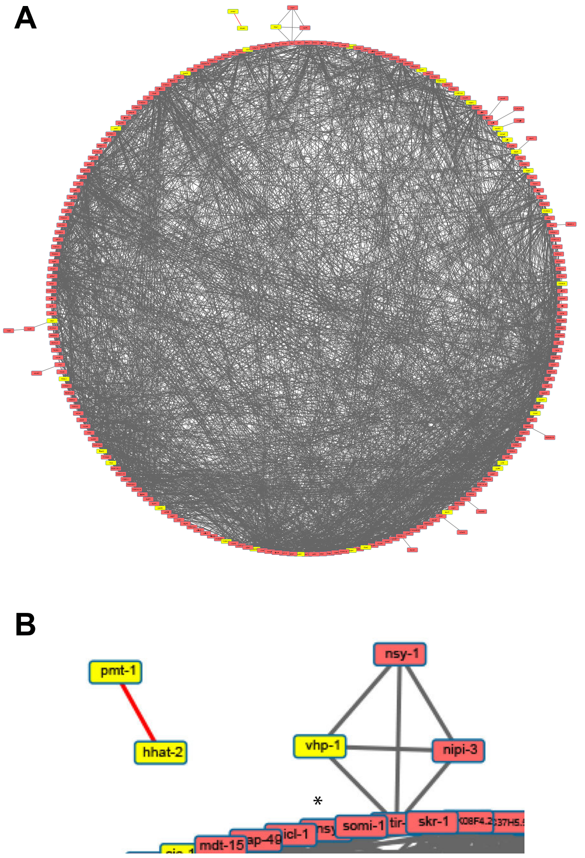


Fig. 4 Nipi genes are connected by a dense network of interactions. **a** Interaction network predicted by WormNet for 233 Nipi genes, with 33 MAPK pathway-related genes shown in yellow. **b** Close-up view of one part of the network shown in b, highlighting two interconnected genes that are apart from the main network and illustrating the relative lack of connections for the known signaling components *nsy-1* and *nipi-3*, which are connected to *tir-1*. The gene *nsy-4* is partially obscured; its position is indicated with an asterisk

This surprising result was corroborated by an analysis of enriched GO terms using GOrilla [48] (Fig. 6; Additional file 7: Table S6), and is discussed below. The proportion of non-histone MAPK signaling genes present in each class varied widely; for "microbial aversion" (class 2) it was 11/33, but was 0/33 for three classes (classes 5, 17 and 29; Table 3). This list of 33 MAPK-related genes overlapped well (8/33) with the list of genes in class 7, reported to be required for multiple cytoprotective responses (i.e., regulators of *gst-4* (detoxification), *hsp-4* (endoplasmic reticulum unfolded protein response (UPR)), *hsp-6* (mitochondrial UPR,UPR^{mt}) and *sod-3* (reactive oxygen species (ROS) response) [49]). There was an equivalent overlap (8/33) with the targets of clones able to suppress over-expression of *gpdh-1p::gfp* seen in *osm-8* mutant worms (class 3). As discussed further below, this gives a further indication of the degree of imbrication of MAPK signaling with different cellular homeostatic processes.

We previously reported a potential role for endocytosis in the induction of *nlp-29* expression provoked by fungal infection [7]. Consistent with this, the GOrilla analysis (Fig. 6a; Additional file 7: Table S6), in common with EASE, also highlighted the role of endocytosis in the regulation of *nlp-29p::gfp* reporter gene expression. They also both drew attention to the potential role of mitochondria in regulating the innate immune response (Fig. 6 and Table 3). For example, all the genes present in at least three of the top four EASE functional classes encode mitochondrial proteins (Additional file 7: Table S6). A Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis [50] of the targets of the 297 Nipi clones (Additional file 5: Table S5) assigned 22 to the category oxidative phosphorylation, corresponding to proteins present in four of the five complexes of the mitochondrial electron transport chain. Extending the analysis to include the targets of the 63 clones that provoked a severe developmental phenotype increased the total number of mitochondrial proteins to 27, covering all five electron transport chain complexes (Fig. 7). The role of mitochondria in the regulation of *nlp-29* expression is explored further below.

Foundling and orphan genes

A total of 33 genes were not found to be associated with any EASE class (Table 5; Additional file 7: Table S6). A WormNet analysis failed to reveal any significant connection between the members of this group (area under the curve = 0.5; $P = 0.3$). Inspection of this list, however, revealed it to include two genes known to play specific and important roles in the regulation of antimicrobial peptide gene expression, namely *dcar-1* and *sta-2*, which encode, respectively, a DAMP receptor [20] and a STAT-like transcription factor [7]. Our EASE database [46], which currently contains more than 500 classes, has been built up by manual annotation and is necessarily biased to categories that we expect to be of interest in our studies. In an attempt to overcome this limitation, we assembled a far more complete and unbiased collection of functional classes, extracting data automatically from multiple sources including WormBase, FlyBase [51], KEGG [52], and the relevant RNAi databases [53, 54], and combined this with our EASE database to give a collection of more than 3700 classes of genes. Even using this collection, we failed to find any significant enrichment for the group of 33 genes.

Genes for which homologues are found only in a specific taxonomic group, irrespective of the level (e.g., animals, nematodes, or *Caenorhabditis*), are called taxonomically-restricted genes (TRGs). TRGs that are restricted to a very narrow taxonomic group, generally a

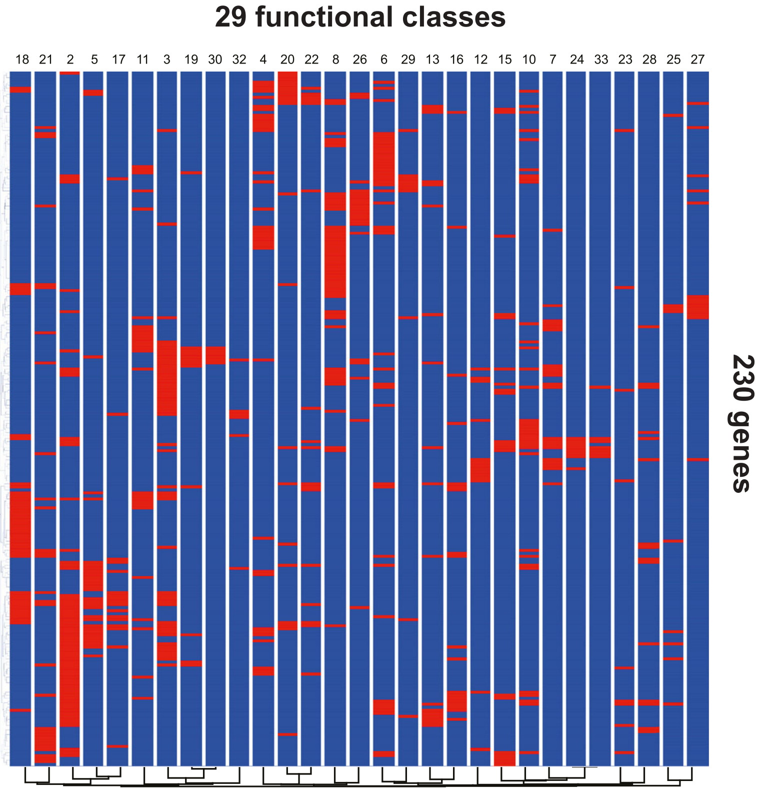


Fig. 5 Relationship between different functional classes. Hierarchical clustering of genes and functional classes (see Table 3 for class labels; full data in Additional file 7: Table S6; 15 genes from the 245 candidates, present only in one or more of the classes 1, 9, 14, and 31 are not shown). The presence of a gene in a class is represented by a red rectangle, its absence in blue

species, can be called orphan genes [55]. Here, we have described a group of genes for which there is no pertinent functional data in a wide range of publically available databases. In a sense, it is as if these genes have been abandoned. By analogy with the term orphan gene, we apply here the term “foundling gene” to them.

The failure to connect the foundling genes could be because they are not in reality linked to each other in any way, or, the explanation we favor, they collectively play specific roles in nematode epidermal defense against fungal infection, which has not hitherto been sufficiently completely described. Such an idea is in line with the pattern of conservation of the 33 foundling genes; more than half are TRGs and encode proteins that are essentially restricted to nematodes. Others are present in a broad range of invertebrate and vertebrate species, while two (WBGene00018063 and WBGene00018670) are currently orphan genes, with no homologs outside *C. elegans* (Fig. 8; Table S7). Genes with similar phylogenetic profiles are more likely to function together in a common biological process [56, 57]. Thus, these diverse patterns of conservation will contribute to elucidating the function of these foundling genes.

Epistasis and functional analyses with candidate clones

The enriched gene classes included four related to osmotic stress, corresponding to a total of 71 genes (Table 3; Additional file 7: Table S6), consistent with the previously established connection between osmotic stress and antimicrobial peptide gene expression [2, 8, 58]. To investigate this link further, we complemented our *in silico* analyses with direct assays to test the capacity of the 297 RNAi clones (Table 2) to block the increase in *nlp-29p::gfp* expression provoked by osmotic stress. Another enriched class was of genes that, when knocked down, provoke the expression of an *irg-1p::gfp* reporter in the nematode intestine [59]. This category is linked to innate immunity since *irg-1* encodes a putative antibacterial effector protein, induced in the intestine upon infection with pathogenic *Pseudomonas aeruginosa* by *zip-2*, which promotes defense [59]. The 297 clones were therefore assayed for their capacity to induce the expression of *irg-1p::gfp*. They were also used in a test of epistasis, by quantitating their potential for abrogation of the elevated *nlp-29p::gfp* expression associated with a constitutively active form of GPA-12 (GPA-12*), the alpha subunit of a heterotrimeric G protein that acts between DCAR-1 and TIR-1 [5].

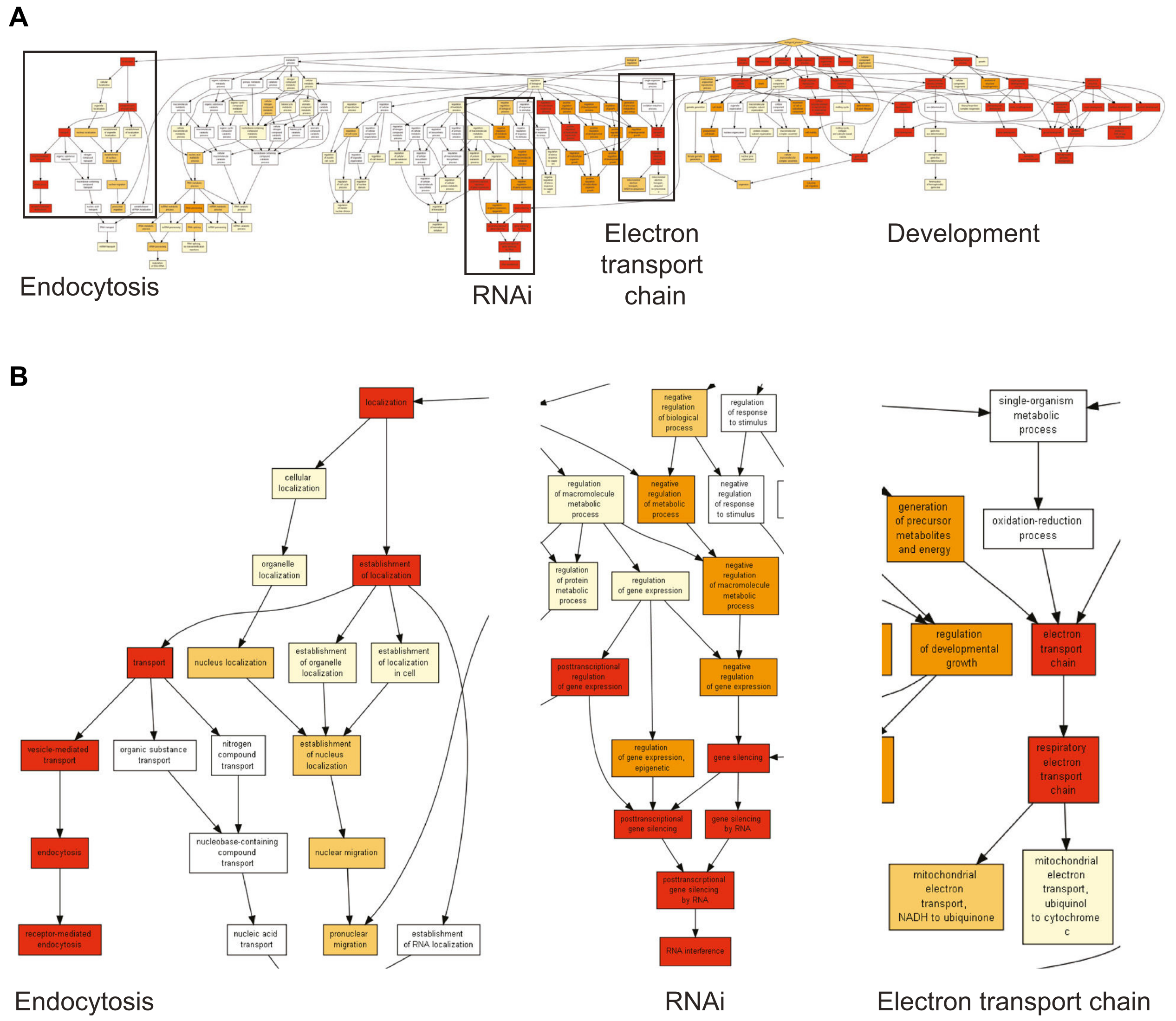

A
Endocytosis
RNAi
Electron transport chain
Development
B
localization
cellular localization
organelle localization
establishment of localization
transport
nucleus localization
establishment of organelle localization
establishment of localization in cell
vesicle-mediated transport
organic substance transport
nitrogen compound transport
establishment of nucleus localization
endocytosis
nucleobase-containing compound transport
nuclear migration
receptor-mediated endocytosis
nucleic acid transport
pronuclear migration
establishment of RNA localization
negative regulation of biological process
regulation of response to stimulus
regulation of macromolecule metabolic process
negative regulation of metabolic process
negative regulation of response to stimulus
regulation of protein metabolic process
regulation of gene expression
negative regulation of macromolecule metabolic process
posttranscriptional regulation of gene expression
negative regulation of gene expression
regulation of gene expression, epigenetic
gene silencing
posttranscriptional gene silencing
gene silencing by RNA
posttranscriptional gene silencing by RNA
RNA interference
single-organism metabolic process
generation of precursor metabolites and energy
oxidation-reduction process
regulation of developmental growth
electron transport chain
respiratory electron transport chain
mitochondrial electron transport, NADH to ubiquinone
mitochondrial electron transport, ubiquinol to cytochrome c
Endocytosis
RNAi
Electron transport chain

Fig. 6 Enrichment of gene ontology terms for Nipi genes. Overview (**a**) and close-up views (**b**) of GOrilla analysis of Nipi genes. The color indicates the degree of enrichment, from red (very significantly enriched) to white (not enriched). The regions enlarged in b are indicated by the boxes in a. In addition to the three categories shown in b, in a, the other very significantly enriched processes, between "RNAi" and "Electron transport chain" are generic ("regulation of multicellular organismal process" and "positive regulation of multicellular organismal process"); those on the right are related to development. See also Additional file 7: Table S6

More than one third of the clones tested (100/297, 34 %; termed here, "I-clones") provoked a marked increase in *irg-1p::gfp*. When compared to the results of Dunbar *et al.* [59], there was a very satisfactory overlap, with identification of 31 of a possible 44 genes previously associated with the phenotype. This figure is similar to that reported by others when comparing screens performed in different laboratories (75 %; [60]). Importantly, many additional candidate negative regulators of *irg-1p::gfp* expression were identified (Additional file 5: Table S5). This is one of the positive consequences of our having conducted a quantitative screen. These results illustrate how the targets of the Nipi clones can have pleiotropic roles, being positive regulators of an epidermal AMP gene, but negative regulators of an intestinal defense gene. They also suggest that the reciprocal relationship between gene regulation in these two tissues that we found for *vhp-1* may reflect a more general phenomenon.

A high proportion (153/297; 51.5 %) of clones abrogated the high constitutive expression of *nlp-29p::gfp* seen in a strain expressing GPA-12* ("G-clones"). Since the p38 MAPK cassette functions downstream of *gpa-12* in the regulation of *nlp-29p::gfp* [7, 9], clones targeting components of the MAPK signaling cascade would be expected to be especially well-represented in this

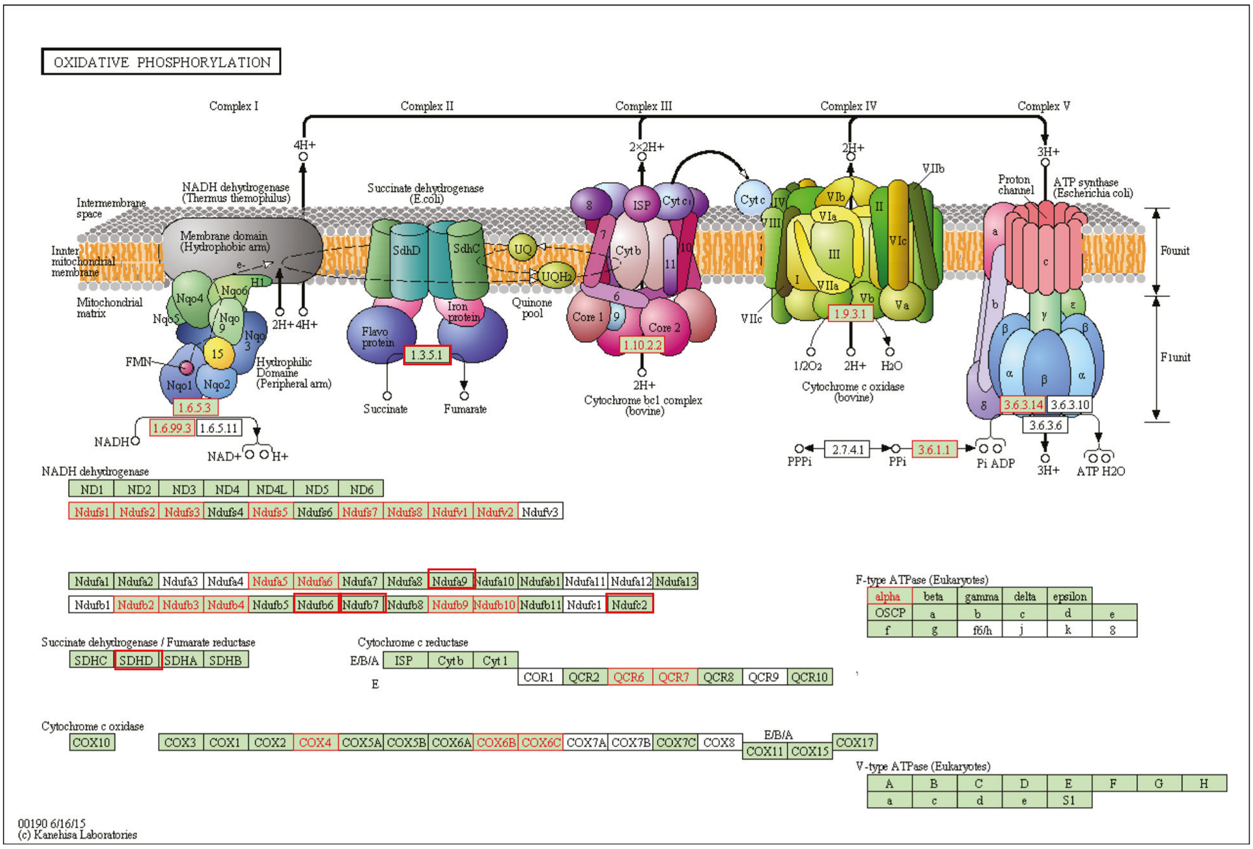

OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
Complex V
4H+
2×2H+
2H+
3H+
Intermembrane space
Inner mitochondrial membrane
Mitochondrial matrix
NADH dehydrogenase (Thermus themophilus)
Membrane domain (Hydrophobic arm)
e-
H1
Nqo4
Nqo6
Nqo5
Nqo 9
Nqo 3
15
FMN
Nqo1
Nqo2
Hydrophilic Domaine (Peripheral arm)
2H+ 4H+
1.6.5.3
1.6.99.3
1.6.5.11
NADH
NAD+
H+
Succinate dehydrogenase (E.coli)
SdhD
SdhC
UQ
UQH2
Flavo protein
Iron protein
Quinone pool
1.3.5.1
Succinate
Fumarate
8
ISP
Cyt c
Cyt b
7
10
11
6
9
Core 1
Core 2
1.10.2.2
2H+
Cytochrome bc1 complex (bovine)
Cyt c
IV
VIII
VIb
VIa
III
II
VIc
VIIb
I
VIIa
Vb
Va
VIIc
1.9.3.1
1/2O2
2H+
H2O
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
a
c
b
γ
ε
β
α
β
α
δ
Fo unit
F1 unit
3.6.3.14
3.6.3.10
3.6.3.6
2.7.4.1
3.6.1.1
PPPi
PPi
Pi ADP
3H+
ATP H2O
NADH dehydrogenase
ND1 ND2 ND3 ND4 ND4L ND5 ND6
Ndufs1 Ndufs2 Ndufs3 Ndufs4 Ndufs5 Ndufs6 Ndufs7 Ndufs8 Ndufv1 Ndufv2 Ndufv3
Ndufa1 Ndufa2 Ndufa3 Ndufa4 Ndufa5 Ndufa6 Ndufa7 Ndufa8 Ndufa9 Ndufa10 Ndufab1 Ndufa11 Ndufa12 Ndufa13
Ndufb1 Ndufb2 Ndufb3 Ndufb4 Ndufb5 Ndufb6 Ndufb7 Ndufb8 Ndufb9 Ndufb10 Ndufb11 Ndufc1 Ndufc2
Succinate dehydrogenase / Fumarate reductase
SDHC SDHD SDHA SDHB
Cytochrome c reductase
E/B/A ISP Cyt b Cyt 1
E
COR1 QCR2 QCR6 QCR7 QCR8 QCR9 QCR10
Cytochrome c oxidase
COX10
COX3 COX1 COX2 COX4 COX5A COX5B COX6A COX6B COX6C COX7A COX7B COX7C COX8
E/B/A
COX11 COX15
COX17
F-type ATPase (Eukaryotes)
alpha beta gamma delta epsilon
OSCP a b c d e
f g f6/h j k 8
V-type ATPase (Eukaryotes)
A B C D E F G H
a c d e S1
00190 6/16/15
(c) Kanehisa Laboratories

Fig. 7 Participation of Nipi gene products in oxidative phosphorylation. The upper part of the figure shows a KEGG-derived schematic representation of the successive complexes that make up the mitochondrial electron transport chain (KEGG pathway cel00190) on which the complexes that include any proteins corresponding to Nipi clone targets are highlighted. For any targets of the 297 clones that did not provoke a strong developmental phenotype, the E.C. name for each complex is boxed in red with red text; for the targets of the remaining 63 clones (Table 2; Additional file 5: Table S5) there is one additional complex (E.C. 1.3.5.1, succinate dehydrogenase), boxed in red with black text. The lower part of the figure shows the individual protein components of the different complexes, annotated as above. Proteins for which there is no KEGG-assigned ortholog in *C. elegans* are uncolored. A full explanation of the symbols can be found at http://www.kegg.jp/kegg/document/help_pathway.html

category. In fact, this was not the case, as only 18 of the 43 clones (42 %) were G-clones (Additional file 5: Table S5). This could be interpreted to mean that a substantial number of G-clones provoked the phenotype for relatively non-specific reasons. Indeed, of the 153 G-clones, 104 (68 %) were found to block the induction of *nlp-29p::gfp* normally provoked by osmotic stress, a markedly higher number than expected, since, in the complete set of 297 clones, there were 131 (44 %) in this category ("O-clones"; Additional file 5: Table S5). Further, 21 of the G-clones targeted genes required for the expression of an *acdh-1p::gfp* reporter gene (class 4, Table 3); *acdh-1* encodes a key enzyme in fatty acid metabolism. Finally, a quarter of the targets of the O-clones (37/147) had previously been associated with the response of *C. elegans* to osmotic stress (Additional file 7: Table S6). The significance of these different overlaps is discussed below.

As described above, an alteration of fungal spore adhesion can lead to a change in defense gene expression. Assaying spore adhesion to worms cultured on RNAi clones proved experimentally challenging because of the variable phenotypes routinely seen with RNAi and especially since there were so many clones to test. We did, however, identify 12 clones that appeared to affect, to a greater or lesser degree, this initial step of the infection process. We were surprised to discover that six were G-clones and five were O-clones, suggesting that the target genes might well play an additional role in governing *nlp-29p::gfp* expression (Additional file 9: Table S8). As a consequence, we did not remove these clones (representing < 5 % of the total) from our lists.

Conserved protein complexes

Functional modules frequently correspond to physical protein complexes. Several studies have defined a variety

Table 5 Foundling Nipi genes

Gene/sequence name	Brief description
dcar-1	GPCR receptor; activated by HPLA [20]
elo-2	Palmitic acid elongase; ELO-2 is required with ELO-1 for 20-carbon PUFA production
frpr-11	GPCR of the FMRFamide Peptide Receptor family; unlike *dcar-1*, acts downstream of *gpa-12* and blocks osmotic induction of *nlp-29* [20]
ins-6	Predicted type-beta insulin-like peptide
mltn-12	MLt-TeN (*mlt-10*) related; MLT-10 is a nematode-specific protein required for ecdysis
nas-37	Astacin-class metalloprotease required for ecdysis; N-terminal signal sequence followed by an Astacin protease domain and three protein-binding domains (EGF-like, CUB, and thrombospondin)
srsx-25	GPCR of the serpentine receptor class SX
srv-21	GPCR of the serpentine receptor class V; unlike *dcar-1*, acts downstream of *gpa-12* and blocks osmotic induction of *nlp-29* [20]
sta-2	STAT family of transcription factor [7]
F56A8.5	Protein containing an F-box
K08C9.5	Protein containing an F-box
C33D9.3	Nematode-specific
F27C8.2[a]	Nematode-specific
F27C8.3[a]	Nematode-specific
F34H10.1	Ubiquitin/40S ribosomal protein S27a fusion protein
K08C9.7	Ubiquitin/40S ribosomal protein S27a fusion protein
R186.8	39S ribosomal protein L33, mitochondrial
T08D2.2[b]	Similar to C-terminal half of UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosaminephosphotransferase
T08D2.6[b]	YIPF4-like; YIPF4, poorly characterized membrane spanning protein; in yeast, interacts with Rab GTPases
Y60A3A.19	YIPF4-like.;YIPF4, poorly characterized membrane spanning protein; in yeast, interacts with Rab GTPases
C42C1.3	Nematode-specific
F35F10.14	Nematode-specific
F41H8.1[c]	Nematode-specific domain: GPCR of the serpentine receptor class BC
K09C6.6[c]	Nematode-specific domain: GPCR of the serpentine receptor class BC
K09C6.10[c]	Nematode-specific
F42C5.9	Actin
F45E4.5	Nematode-specific
F49H12.5	Thioredoxin domain-containing protein 12-like
F52C6.13	Nematode-specific
tsen-54	N-terminal half similar to that of tRNA-splicing endonuclease subunit Sen54
Y39G10AR.7	Nematode-specific
Y51H7BR.7	Contains Spec3 domain, like *Drosophila* Stumbled
Y67D8C.22	Clarin-like

[a,b]Targeted by a single mv clone

[c]Potential targets of the same sjj clone

References are given for genes previous connected to the regulation of *nlp-29* expression; see Additional file 7: Table S6

of protein complexes from different species. One recent report provided more than one million putative high-confidence co-complex interactions present broadly across animal species [61]. Combining this with data from yeast [62–64], and having identified the *C. elegans* orthologues of the component proteins when necessary, we compiled a collection of 1925 predicted *C. elegans* protein complexes (Additional file 10: Table S9). We then associated each of the predicted targets of the Nipi RNAi clones with the different complexes. We focused on complexes with at least three components for which we had picked up more than half of the components in our screen (Additional file 10: Table S9). There was an over-representation of the eukaryotic translation initiation factor (eIF) 2B complex and 66S pre-ribosomal particles, suggesting an important role for protein translation. There was also enrichment for components of the carbon catabolite repression 4-negative regulator of transcription (CCR4-NOT) complex, which is a major mRNA deadenylase, linked to mRNA degradation and general transcriptional regulation, among other functions [65]. We discuss these observations below. The analysis also indicated that there was enrichment in several mitochondrial complexes (Fig. 9). This is consistent with the KEGG analysis described above (Fig. 7), and we focused our attention on this class of gene.

Intestinal UPR$^{\text{mt}}$ inhibits epidermal AMP expression

A total of 30 genes identified in our screen have been shown to induce a mitochondrial UPR (UPR$^{\text{mt}}$) when inactivated [66]. For example, the well-characterized *spg-7* that encodes a mitochondrial metalloprotease, was picked up with two independent RNAi clones in our screen (Additional file 5: Table S5). This suggested that activation of the UPR$^{\text{mt}}$ could block the expression of antimicrobial peptide genes in the epidermis. At the same time, in contrast to intestinal infection with *P. aeruginosa* [67], infection of young adult *C. elegans* by *D. coniospora* does not provoke the UPR$^{\text{mt}}$ since the expression of the hallmark genes *hsp-6* and *hsp-60* is unchanged [2, 46].

Given the links that exist between the UPR$^{\text{mt}}$ and antibacterial defenses in *C. elegans* [68], we decided to explore in more depth the relationship between the response to fungal infection and the UPR$^{\text{mt}}$. As a first step, to validate the results of the screen obtained with the reporter construct, we used qRT-PCR to assay the level of the endogenous *nlp-29* transcript following knock-down of five candidate genes, all associated with the activation of other stress reporter transgenes (Additional file 7: Table S6), including *spg-7*, a well-established means of inducing the UPR$^{\text{mt}}$ [69]. Inactivation of four of them (*ant-1.1*, *atp-4*, *spg-7*, and *ucr-1*) abrogated *nlp-29* gene expression after infection to the same degree

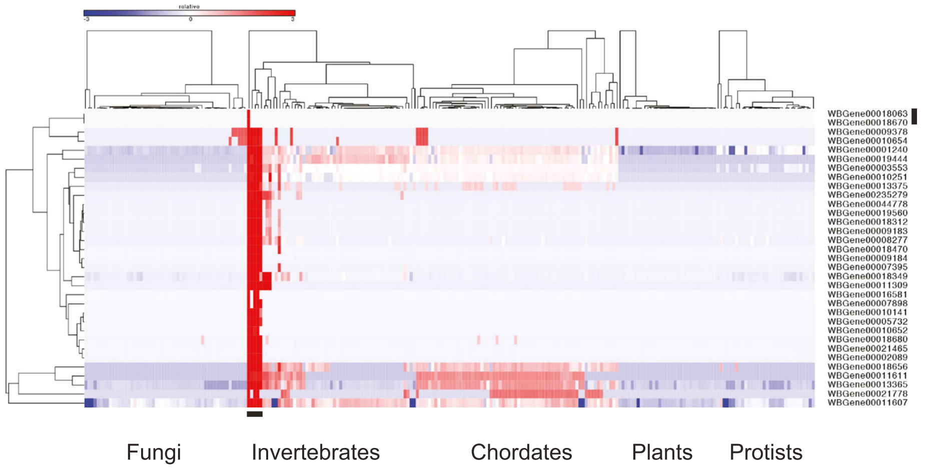


Fig. 8 Phylogenetic profiles for 33 foundling genes. Hierarchical clustering of the normalized bit scores for homologs for 33 foundling Nipi genes across the genomes of 238 species present in Ensembl (see Additional file 8: Table S7 for the list of species in the order they appear here). For each of the groups (fungi, invertebrates, etc.), the species are clustered independently. The colour code reflects the relative normalized bit score, from high (red) to low (blue) across the different species. The horizontal bar at the bottom marks the position of the five *Caenorhabditis* species, from the left to right, *C. elegans*, *C. brenneri*, *C. briggsae*, *C. remanei*, and *C. japonica*. Several distinct groups of genes can be discerned, including genes unique to *C. elegans* (i.e., orphan genes [123]), indicated by the vertical bar on the right

as the positive control, *dcar-1*, while knocking down *gas-1* did not have a statistically significant effect (Fig. 10a).

In *C. elegans*, the UPR^{mt} can involve trans-tissue signaling (reviewed in [12]). Thus, for example, provoking an UPR^{mt} just in neurons leads to an UPR^{mt} in the intestine [70]. To address the question of whether the inhibitory effect of the UPR^{mt} on *nlp-29* gene expression might also be cell non-autonomous, we assayed the effect of knocking down the same five candidate genes by RNAi specifically in the intestine in the strain MGH171 [26]. In this case, in contrast to intestinal knockdown of *dcar-1*(RNAi), which gave the same average level of expression of *nlp-29* as *sta-1*(RNAi), consistent with *dcar-1*'s cell autonomously function [20], intestinal RNAi of *ant-1.1*, *atp-4*, *spg-7*, and *ucr-1* was associated with an abrogation of *nlp-29* gene expression following *D. coniospora* infection, while *gas-1*(RNAi) again did not provoke a statistically significant effect (Fig. 10b). Overall, our results indicate that provoking the UPR^{mt} in the intestine reduces the induction of an antimicrobial peptide gene in the epidermis (Fig. 10c).

Discussion

Qualitative versus quantitative RNAi screens

Genome-wide RNAi screens have been performed in *C. elegans* for more than a decade. Their experimental basis is relatively straightforward, since RNAi by feeding is an effective technique in worms [71]. In a number of cases, the read-out has been the effect of RNAi on the expression of a reporter gene or the localization of a chimeric reporter protein, to address a specific biological question (e.g., [49, 57, 59, 66, 72–77]). Generally, these have been

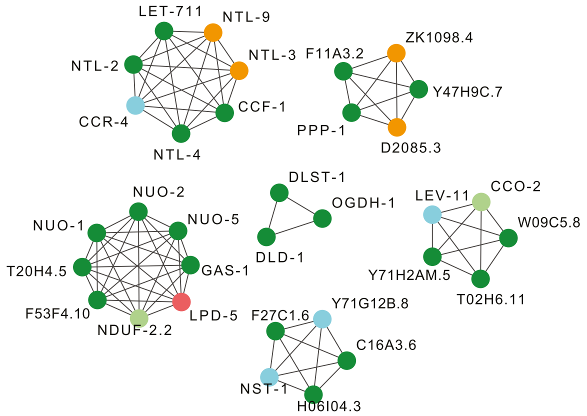


Fig. 9 Enrichment of proteins encoded by Nipi genes in selected protein complexes. The components of several known complexes are shown, namely, clockwise from top left, with the complex ID(s) in brackets (Additional file 10: Table S9), CCR4-NOT (YChrMod1004), eukaryotic translation initiation factor 2B complex (195; AMMC1231), mitochondrial ETC (218; AMMC1876), 66S pre-ribosomal particles (169; AMMC1267), mitochondrial complex I (65; AMMC1203), and in the center, mitochondrial oxoglutarate dehydrogenase complex (OGDC; YPC_CON1369). Proteins in dark green correspond to members of the final set of 338 Nipi genes, those in light green and orange to the presumed targets of clones that passed the first round of selection or that gave a single positive result in the first round, respectively. The clone that potentially knocks down *lpd-5* (corresponding protein in red) abrogated *nlp-29p::gfp* expression but had a marked effect on development and reduced *col-12p::dsRed* expression. The remaining proteins are shown in blue

visual screens. Despite certain advantages [78], visual screens include an element of subjective judgment, lack discriminatory power, and are best suited to identifying clones that provoke a marked phenotype. These will generally target genes at the central nodes of a signaling network. Full understanding of regulatory mechanisms also requires, however, the identification of genes that exert only a minor effect [23].

An alternative is to undertake automated quantitative screens. These require specialized equipment and tools for data storage and analysis [18, 19, 28, 79–82] and are thus more difficult to put in place. Further, they also suffer from the intrinsic variability of RNAi, which cannot be adequately accounted for using formal statistical analyses (Thomas Richardson, University of Washington, personal communication). Coupled with the continuous distribution of the results, this renders the definition of candidates somewhat arbitrary. In this study, as discussed in more detail elsewhere [83], we used the results for clones targeting genes known to be important for the

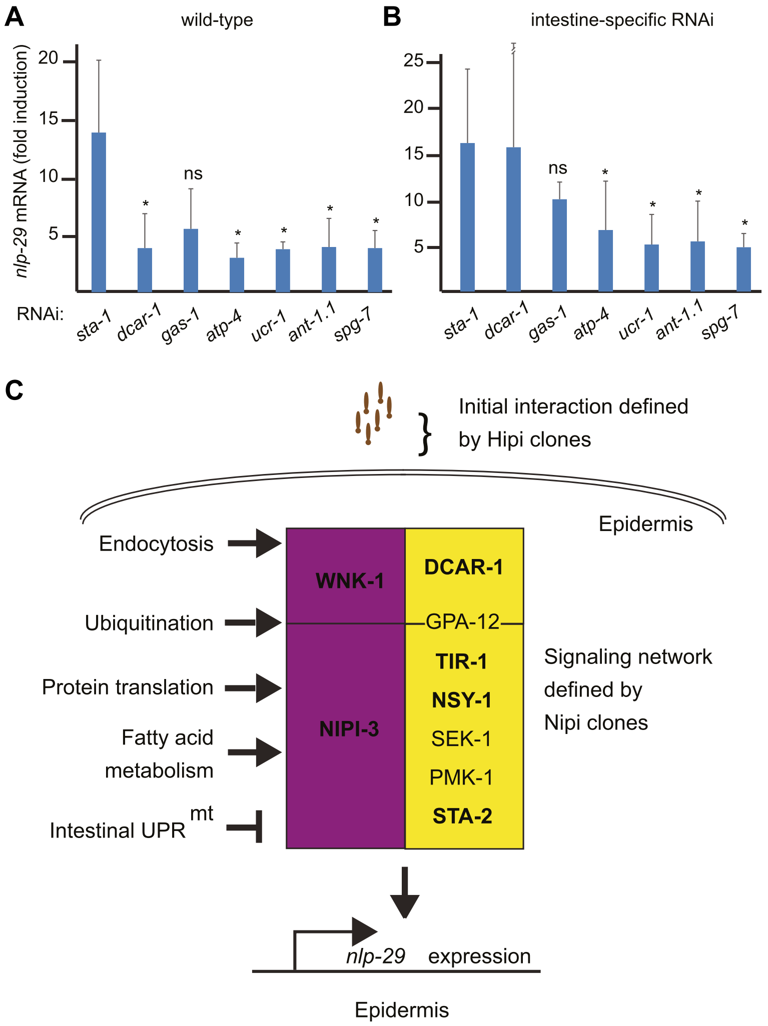


Fig. 10 Cell autonomous and non-autonomous mechanisms influence *nlp-29* expression. Quantitative RT-PCR analysis of *nlp-29* gene expression level showing the infection-associated "fold induction" (infected/non-infected values) in worms treated with positive and negative RNAi control clones (targeting *dcar-1* and *sta-1*, respectively) or clones that provoke a UPR^{mt} (targeting *ant-1.1*, *atp-4*, *spg-7*, *ucr-1*, and *gas-1*) in a wild-type (**a**) or the intestine-specific RNAi strain MGH171 (**b**) for 30 h before infection with *D. coniospora* for 18 h. Results are the average (± SD) from four and three experiments, respectively (see Additional file 11: Table S10). The difference between control and *gas-1*(RNAi) is not significant (ns) in either strain; * $P < 0.05$ (unpaired *t*-test). The SD for *dcar-1* in (b) is 14. **c** Simplified model of pathways and processes involved in the regulation of *nlp-29*. The screen identified Hipi genes that modulate the adhesion of spores to the worm cuticle, and Nipi genes (central box) required for the expression of *nlp-29* upon osmotic stress and infection (purple), or only after infection (yellow), acting downstream (below horizontal line) or upstream of or parallel to GPA-12. Only a very limited number of genes are shown; those in bold were identified in the screen. The Nipi genes fall into multiple functional categories; some are listed on the left, positioned arbitrarily; pointed and flat arrows indicate positive and negative regulation, respectively

regulation of *nlp-29* to establish cut-offs, and in the first round of screening for Nipi clones, we privileged those giving a reproducible effect. Similarly, the cut-offs we used to identify Peni and Hipi clones were based more on criteria of reproducibility rather than strict statistical criteria. The thresholds we adopted will necessarily determine the candidate genes identified, and could bias our global analyses. We have provided, however, for the first time, via a dedicated web interface, the complete set of results for the two rounds of selective screening, measuring multiple parameters for individual worms from each population. Not every clone in the library will contain the expected insert. Extrapolating from our sequencing of 388 clones, which revealed an error rate of 14 %, around 3000 clones in the complete library might be incorrect. With this caveat in mind, the complete set of results contains a substantial amount of information that we have not attempted to exploit, for example, linked to inter-individual variability of gene expression, terminal epidermal cell fate determination, or simply genes that affect the development and size of *C. elegans*. It also will be an important resource for those wishing to develop new analytical methods; these would be required to leverage the intrinsically variable quantitative data for subsequent analyses.

Genes affecting spore adhesion

One class of genes to emerge from this screen is the Hipi genes that affect the initial adhesion of *D. coniospora* spores to *C. elegans*. This requires contact between the spores' adhesive bud and the outermost layer of the nematode cuticle, the surface coat. In contrast to the underlying collagen- and cuticlin-rich cuticle, the surface coat is rich in structural glycoproteins, including mucins [84, 85]. Two of the candidate Hipi genes, *bus-2* and *bus-12*, were known to affect the surface coat. Their role in adhesion of *D. coniospora* has been investigated using the corresponding mutants [31]. These were originally isolated because, unlike wild-type worms, they were not susceptible to infection by the bacterial pathogen *Microbacterium nematophilum* that normally adheres to specific areas of the worm cuticle [86]. Both genes are important for the post-translational modification of surface-exposed proteins [29, 30]. Several other Hipi genes encode conserved enzymes. Their precise role in mediating spore adhesion will require detailed study but, as mentioned above, they too might alter surface protein maturation. Another candidate, K06A9.1, is nematode-specific. It is predicted to encode several protein isoforms, including one of more than 2200 amino acids, comprising 22 degenerate 81 amino acid repeats. Taken together with its distant similarity to mucins, this suggests that it could be a component of the surface coat.

The gene *ykt-6* encodes the worm ortholog of Ykt6p, a v-SNARE essential for endoplasmic reticulum-Golgi transport [87]. It could be required for the correct transport of surface proteins. On the other hand, *ykt-6* has been linked to insulin signaling in *C. elegans* [88] and, interestingly, *ins-6*, which encodes an insulin-like peptide, was identified as a Nipi gene in the current screen (Table 5). An *ins-6* loss-of-function mutant, however, did not display a Nipi phenotype (unpublished results). A lack of concordance between phenotypes observed using RNAi and mutant strains has previously been reported (e.g., [89]). The definitive attribution for a role in spore adhesion for the various Hipi genes must therefore await individual genetic validation.

MAPK signaling and osmotic stress responses

The results of the screen reaffirm the central place of p38 MAPK signaling in the regulation of the *nlp-29* AMP gene in the epidermis [90] and substantially expand the catalog of genes involved. Many of the same genes are also required not only for xenobiotic detoxification, the UPR, the UPR^{mt}, and the response to ROS [49], but also for the regulation of *gpdh-1*, a gene that encodes the rate-limiting enzyme in the biosynthesis of the osmoprotectant glycerol [72, 91]. The expression of *gpdh-1* does not change following infection with *D. coniospora* [2, 46], but is elevated upon exposure to high concentrations of salt, via a mechanism that involves inhibition of translation. This is mediated by the general control non-derepressible (GCN-2) kinase signaling pathway that controls eIF-2α phosphorylation and the activity of the with-no-lysine kinase and Ste20 kinases WNK-1 and GCK-3 [92]. In contrast, the results of our screen amply demonstrated that inhibition of translation does not activate *nlp-29* expression. Quite the contrary, many clones targeting genes required for translation, including those encoding aminoacyl-tRNA synthetases (*hars-1*, *lars-1*, *rars-1*, *tars-1*, *wars-1*) and eIF subunits, were required for expression of *nlp-29* after infection. Although we have not yet determined whether these effects are cell autonomous, this AMP gene therefore distinguishes itself from effectors of other stress responses, such as *irg-1* and *gpdh-1.* On the other hand, its expression upon osmotic stress does require *wnk-1* and *gck-3* [8], which is also the case for *gpdh-1* [92]. We show here that the induction of *nlp-29* expression after infection also requires *wnk-1*, acting upstream or in parallel to *gpa-12* (Additional file 5: Table S5) and to a lesser extent *gck-3* (Fig. 1b). Further, in common with two thirds of the genes that act downstream of *gpa-12* (the targets of the G-clones, including *hars-1*, *lars-1*, *rars-1* and *tars-1*), we found that *nipi-3*, which encodes a homolog of Tribbles required for the response to infection [3], is also required for the expression of *nlp-29* upon osmotic stress. These results lead to a revision of the infection/osmotic stress dichotomy [8] of our previous models for the regulation of *nlp-29* (Fig. 10c).

Cross-tissue communication

One unexpected finding regarding MAPK signaling was that *vhp-1*(RNAi) abrogates *nlp-29* expression, since VHP-1 has been described as a negative regulator of p38 PMK-1 in the nematode intestine [43, 93]. There, the p38 MAPK has a well-characterized role in defense against bacterial pathogens that colonize the gut lumen. We have previously observed that there is an overrepresentation among the genes induced by *D. coniospora* of genes repressed after infection by the bacteria *S. marcescens*, *E. faecalis*, and *P. luminescens*. This enrichment includes numerous antimicrobial peptide genes of the *nlp* and *cnc* classes [46]. In other words, bacterial infection of the gut, which switches on the p38 MAPK pathway, implicating a decrease in VHP-1 activity, abrogates epidermal antimicrobial peptide gene expression. Our current hypothesis is that *vhp-1*(RNAi) activates the p38 MAPK pathway in the intestine, and that this has the paradoxical consequence of reducing p38 MAPK activity in the epidermis despite a reduction of *vhp-1* expression in that tissue too. We currently have no plausible explanation for the observation of ectopic, intestinal expression of *nlp-29* following knockdown of *vhp-1* only in the epidermis, but note that other analogous examples of cell non-autonomous regulation have been recently reported [78].

A further example of communication between tissues was revealed in our investigation of the impact of the UPR^{mt} on AMP gene expression. Knocking-down, specifically in the intestine, one of several genes known to trigger an UPR^{mt} caused a reduction of *nlp* gene expression following *D. coniospora* infection. Recent studies have suggested that intestinal pathogens can provoke the UPR^{mt} in *C. elegans* and that this switches on defense gene expression in the intestine [67, 74]. The UPR^{mt} is negatively regulated by the Jun kinase KGB-1 [94], which in turn is negatively regulated by VHP-1 [43, 93]. While this mechanism is complemented by another pathway involving ROS-stimulated eIF2α kinase that leads to a reduction in protein translation [95], compromising overall translatory capacity can by itself cause the expression of defense genes such as *irg-1* [59, 96] and *gpdh-1* [92], but prevents ROS-induced UPR^{mt} [94]. While the precise interplay between this complex series of homeostatic and cellular defense mechanisms is far from being understood [12, 97], these different findings are compatible with a model wherein activation of antibacterial defense mechanisms in the intestine, whether directly upon infection with bacterial pathogens, by reducing VHP-1 activity, by reducing protein translation, or following a UPR^{mt}, leads to a suppression of the capacity of the epidermis to express antifungal defense genes. As such, this could constitute a mechanism to ensure an appropriate allocation of resources within the organism, with the aim of concentrating energy to defend one tissue, to the detriment of the capacity of the epidermis to express AMPs.

Fatty acid metabolism and AMP gene expression

Previous studies have suggested a possible link between fatty acid metabolism and innate immunity in *C. elegans* [8, 27, 98]. This is further reinforced by the fact that both *dld-1* and *elo-2*, respectively encoding a dihydrolipoamide dehydrogenase and a palmitic acid elongase, were identified as Nipi genes in our screen. Further, the expression of *acdh-1* and *acdh-2*, which encode mitochondrial short-chain acyl-CoA dehydrogenases that catalyze the first step of fatty acid beta-oxidation, is markedly reduced when *C. elegans* is infected either with *D. coniospora* or with a number of different bacterial intestinal pathogens [46, 99]. The *elo-2* paralog *elo-3* was previously found to be required for the expression of *acdh-1p::gfp* [73]. A total of 35 Nipi genes are also regulators of *acdh-1p::gfp* expression (Table 3). These include the mediator complex gene *mdt-15*, a major regulator of fatty acid metabolism and longevity [100, 101]. MDT-15 is also required for oxidative stress responses and the induction of specific detoxification genes in response to xenobiotics or heavy metals [102–104]. MDT-15 was recently shown to have a more direct role in innate immunity since it regulates the expression of p38 MAP kinase PMK-1-dependent immune genes and resistance to *P. aeruginosa* infection [105]. The links between fatty acid metabolism and host defense in *C. elegans* clearly merit more detailed investigation.

Further functional groups involved in AMP gene expression

Several other groups of functionally related genes were also identified among the Nipi genes. Almost 50 genes had previously been characterized as being necessary for transgene silencing [57]. This is paradoxical since, if RNAi were not efficient in our system, we would not expect an RNAi-dependent reduction in reporter gene expression. Many of the genes in this category clearly play an indirect role in transgene silencing. To give just one example, *dpy-4* encodes a cuticle collagen and is required for normal morphology. Whether they play direct roles in modulating *nlp-29p::gfp* expression remains to be established.

There was a similar overlap of Nipi genes with genes required for the correct sub-cellular localization of the RAB-11, a small GTPase involved in endocytosis, and for transport of the apical membrane protein PEPT-1 [79]. This is consistent with our observation that knocking down dynamin (encoded by*dyn-1*), which is involved in the scission of newly formed clathrin-coated endocytic vesicle from the cell membrane, or the small GTPase Rab5 (*rab-5*), which characterizes early endosomes derived from dynamin-dependent and independent endocytosis, abrogates *nlp-29* gene expression after infection [7]. Endosomal membranes may function as important platforms for innate immune signaling in *C. elegans* as in other species [106, 107].

Among the three Nipi genes (*icd-1*, *let-70*, and *let-92*) required both for transgene silencing [57] and endocytosis [79], as mentioned above, the E2 ubiquitin conjugating enzyme gene *let-70* is linked to a broad range of cellular and organismal functions. It is noteworthy that we found nine other genes involved in ubiquitination and proteasome-mediated protein catabolism (*dcaf-1*, *hecd-1*, *pas-3*, *pas-5*, *pbs-2*, *prp-19*, *rbpl-1*, *skr-1*, *usp-39*) in the present screen. It is likely therefore that, in *C. elegans*, ubiquitination plays an important role in regulating innate immune responses, as it does in many species by governing the stability of key signaling molecules [108–111].

Finally, there was enrichment for components of the CCR4-NOT complex. This complex coordinates a variety of cellular processes, acting at all levels of gene expression, including transcription and mRNA or protein stability. It is involved in cellular adaptation to external stress, including the control of the vertebrate innate immune response through the regulation of STAT1 [112]. It may act in a similar manner to influence the activity of the STAT-like transcription factor STA-2 and thereby the expression of AMP genes in *C. elegans*.

Conservation and innovation in innate immune defenses

In nature, infection represents an extremely strong selection pressure. This is reflected by the evolution of sophisticated host defense mechanisms, driven by the different pathogens that exercise a negative impact on fitness and survival in the environment. In jawed vertebrates, this has led to the emergence of the adaptive immune system, based on a specific collection of genes and mechanisms not found outside the infraphylum [113–115]. Similar specialization involving groups of TRGs involved in immunity is observed in other branches of the animal kingdom [116]. Here, we identified a number of genes required for the expression of an antimicrobial peptide in *C. elegans* that are restricted to nematodes. They are expected to be part of a lineage-specific defensive innovation. Their further study will contribute to our understanding of the evolution of immunity in *C. elegans* [117, 118]. Our results also highlighted the links that exist between antimicrobial defenses and the homeostatic mechanisms that counter abiotic stress. This supports an ancient origin for the co-adaptive evolution of stress and innate immune responses (e.g., [119]).

Conclusions

In conclusion, this genome-wide study has allowed the identification of hundreds of genes that modulate the capacity of *C. elegans* to express the AMP gene *nlp-29* following infection with *D. coniospora*. Not only has it greatly expanded the number of such Nipi genes, but it has also revealed multiple interwoven cellular regulatory mechanisms that impinge on AMP gene expression. Understanding the precise nature of the regulatory activity exercised by the Nipi genes in each of these different functional classes, as well as the many individual genes, will require focused study in the future.

Methods

Nematode strains

All strains were maintained on nematode growth media and fed with *E. coli* strain OP50. The strain for intestine-specific RNAi, MGH171 *sid-1*(*qt9*); *alxIs7*[*vha-6*p::SID-1::SL2::*gfp*] [26] was kindly provided by Justine Melo and the strain AU133 *wt*; *agIs17*[*irg-1*p::*gfp*;*myo-2*p::*mCherry*] [96] by Emily Troemel. Details about the constructions of the strains IG274 (*frIs7*[*nlp-29*p::*gfp*, *col-12*p::*DsRed*] *IV*), IG1389 (*frIs7*; *frIs30*[*col-19*p::*gpa-12**,*unc-53*pB::*gfp*] *I*), and the epidermis-specific RNAi strain IG1502 (*rde-1(ne219) V; ls[wrt-2*p*::RDE-1::unc-54 3'utr; myo-2*p*::RFP3] III; frIs7 IV*) are provided elsewhere [3, 5, 20].

A genome-wide RNAi library

In order to cover the maximum number of target genes, as specifically as possible, we combined RNAi clones from the Ahringer genomic [21] and Vidal cDNA [22] libraries. The constitution of the RNAi library was based on the data and tools for target prediction available at the time (WormMart WS220; now retired). If an Ahringer library clone was predicted by WormMart to have more than one primary target, when possible, we added Vidal library clones predicted to target individually any or all of the multiple primary targets. The Ahringer clones were directly redistributed from each 384-well library plate among four daughter 96-well plates. Of the 16,744 clones in our copy of the Ahringer library, 625 failed to grow. We equally sought to replace them with clones from the Vidal library and cherry-picked a total of 5136 clones. Of these, 32 failed to grow, leaving us with a collection of 21,223 clones (Additional file 1: Table S1). Among them, 132 Ahringer clones were present in two wells, so that the combined library included clones in 21,355 wells. This library of 21,355 wells was used in the first round of screening.

Target prediction

Because of limitations in the method used by WormBase to predict the targets of an RNAi clone, as part of this project, we developed the tool CloneMapper [28]. Out of the 21,223 clones, 20,025 were present in CloneMapper, and were predicted to target 16,565 genes (score ≥ 1). For the remaining 1198 clones, despite the known shortcomings [28], we used WormMart WS220 and WormBaseConverter [46] (WS220 to WS240) to identify a further 1304 targets. Combined, the clones are predicted to target 17,415 of the 20,540 protein coding genes (84.8 %) in WS240.

High-throughput RNAi screen

The RNAi screen was performed as previously described in detail [19, 83]. Briefly, synchronized L1 larvae were deposited in 96-well plates containing nematode growth media agar, with a different RNAi clone in each well. After 30 hours at 25 °C, when worms had reached the L3-L4 stage, a fresh solution of *D. coniospora* spores was added to each well, and worms were harvested 18 hours later for analysis using the COPAS Biosort. All data was stored in a custom-made database (Modul-Bio, Marseille, France) for subsequent analysis. Evaluation of the capacity of RNAi clones to block the increase in *nlp-29p::gfp* expression provoked by osmotic stress was performed as described [9]. Briefly, following culture on RNAi clones for 48 h, young adult worms were transferred into 96-well U-bottom plates containing 200 μL of 300 mM NaCl and gently agitated for 3 hours at 25 °C before Biosort analysis. Generally, a minimum of 80 synchronized worms were analyzed for size (TOF), extension, and green (GFP) and/or red (dsRed) fluorescence [18]. The inserts of candidate clones were sequenced to establish their identity.

Data analysis and clone selection

Data analysis was performed as previously described in detail [19, 83]. Briefly, in the first round (whole genome) screen, for each well, a mean value for the GFP/TOF ratios for each worm was calculated. From these values, for each plate, a truncated mean (discarding the 25 % lowest and the 25 % highest values) was calculated and used to normalize the average GFP/TOF values for the individual wells, to allow across-plate comparison. Normalized values for TOF (TOF/[truncated mean of TOF]) and dsRed ((dsRed/TOF)/[truncated mean of dsRed/TOF]) were similarly calculated. Details of Nipi clone selection after the second round of screening are given in Supplementary Methods.

Validation of the RNAi screening approach

A full description of the experimental validation of the screening approach can be found in a publicly available PhD thesis [83]. Of note, using the standard feeding protocol, a substantial number of RNAi clones can provoke severe developmental delays and/or larval lethality [21]. In an attempt to circumvent this, we transferred worms from their standard *E. coli* OP50 diet to RNAi bacteria at the early L3 stage and assayed the same worms when they were adults. Unfortunately, this was not a sufficiently robust method since, of the sequence-verified positive controls we tested, namely *pkc-3*, *rack-1*, and *sta-2*, only *sta-2* gave a phenotype [83]. Trying to increase the efficiency of the RNAi by using the RNAi sensitive strain *rrf-3* was also unsuccessful [83] since expression from high-copy transgenes is compromised in this background [120].

Analysis of spore adhesion

Worms treated with each of the 28 Hipi clones from the L1 stage were infected as L4s with *D. coniospora*. To directly correlate spore adhesion and reporter gene expression, worms in the population ($n \geq 30$) were visually inspected for their GFP expression, before accessing the adhesion of spores. A score was assigned, taking into account the intrinsic variability in GFP expression associated with infection (see, for example, Fig. 10b). Clones associated with a very high and homogenous induction were assigned a score of 2, clones associated with an induction similar to wild type were assigned a score of 0, and clones associated with an intermediate phenotype assigned a score of 1. Worms were then harvested in 50 mM NaCl, 0.05 % Triton, transferred to 96-well round-bottom well plates, and frozen at –80 °C. Plates were subsequently thawed and the number of spores attached to worms were counted at 230× using a Leica MZ16 stereomicroscope. Clones were assigned to three broad categories, relative to *sta-1*(RNAi)-treated control worms: 0 = 1–10 spores/worm (same as control); 1 = 10–25 spores; 2 = > 25 spores. A minimum of 30 animals were scored for each clone. Worms treated with the candidate Nipi clones from the L1 stage and infected as L4s with *D. coniospora*, as above, were analyzed slightly differently. The major part of each sample was analyzed with the Biosort, as above, and for the remainder, the number of spores attached to worms, at the head and vulva, were counted. An adhesion index was calculated: ((number of worms with $n > 1$ spores at the mouth) + (number of worms with $n > 1$ spores at the vulva))/(2 × total number worms). Clones associated with a score inferior to that of all control clones, and with a reduction of reporter gene expression greater than 50 % (i.e., loss of spore adhesion was accompanied by a reduction in the observed innate immune response) were selected. Clones selected in both duplicate tests were retained.

Bioinformatic analyses

All analyses used the WS240 WormBase release, unless otherwise stated. Programs were written in Perl and the user interface was developed using HTML, PHP, JavaScript, and MySQL. We used WormNet v3 [42], EASE 2.0 [45] with an in-house database of functional annotations [46], GOrilla [48] with the November 2015 data update, and KEGG [52] release 77.1. For clustering, we used "One minus Pearson correlation" distance matrices within GENE-E (www.broadinstitute.org/cancer/software/GENE-E/).

Data collection

In addition to the previously collected datasets for *C. elegans* functional classes used in EASE analyses and manually assembled from the literature, including differential transcriptomic and proteomic data, miRNA targets, TF targets etc., further classes were defined using data from a variety of resources.

1) We extracted all 1094 phenotypes available in WormBase WS246 from ftp://ftp.wormbase.org/pub/wormbase/.
2) A total of 1221 expression cluster datasets (WS246) were downloaded from ftp://caltech.wormbase.org/pub/wormbase/spell_download/.
3) We extracted the full list of *Drosophila* phenotypes from FlyBase release FB2014_06 via http://flybase.org/.bin/cvreport.html?cvterm=FBcv:0000347+childdepth=2, and manually collated closely related classes to give a list of 145 to which we matched the corresponding FlyBase Gene IDs. Using the DRSC Integrative Ortholog Prediction Tool (DIOPT) (http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl), we identified the *C. elegan*s orthologs for these genes.

   If more than one ortholog for a given fly gene was predicted, using an in-house perl script, we selected a best hit if the difference in the DIOPT score was greater than or equal to 2 (maximum score 10), but otherwise did not retain a worm ortholog.
4) Fly RNAi screen data was taken from http://www.flyrnai.org/RNAi_all_hits.txt (downloaded 14-11-14). RNAi clones for which no target gene was listed were excluded. Prediction of *C. elegans* orthologs was as above.
5) We downloaded all RNAi screen datasets for *Drosophila* from version 13 of GenomeRNAi.org [54] and manually collated very similar classes to give 110 datasets. Prediction of *C. elegans* orthologs was as above.
6) We extracted the genes corresponding to 136 KEGG_pathways (July 2015) and then converted identifiers to WormBase GeneIDs.

The resulting datasets are available on request.

A separate collection of predicted *C. elegans* protein-protein complexes was also assembled using experimentally determined protein complexes from other species [61–64]. Prediction of *C. elegans* orthologs was as above except that, if the difference in the DIOPT score was less than or equal to 2, the top two putative orthologs were retained.

Phylogenetic profiling

To construct phylogenetic profiles, we followed an approach somewhat similar to that of Tabach et al. [57]. We collected data for a wide range of eukaryotic species. We downloaded the complete set of predicted proteins for 66 vertebrates from Ensembl (release 78) and for 55 invertebrates, 53 fungi, 32 plants, and 32 protists from Ensembl genomes (release 25). As many genes have multiple isoforms (e.g., 30,939 for 20,493 protein-coding genes in *C. elegans*), we chose the longest transcript for each gene. We used BLASTP to compare the proteins predicted for the 33 *C. elegans* foundling genes against all 238 proteomes and we chose the best hit for each. From this, we generated a "BestHit" matrix (33 × 238), where each entry C_{ij} is the best BLAST bit score of the top hit in species "j" for *C. elegans* protein "i". As BLAST bit score depends on protein length, we normalized each bit score by calculating a self-similarity score C_{ii} (by BLASTing each *C. elegans* protein C_i against itself). We generated a normalized matrix by replacing each C_{ij} by C_{ij}/C_{ii}.

RNA preparation and quantitative RT-PCR

RNA preparation and quantitative RT-PCR were as described [3]. Results were normalized to those of *act-1* and were analyzed by the cycling threshold method. Control and experimental conditions were tested in the same 'run'. Each sample was normalized to its own *act-1* control to take into account age-specific changes in gene expression. Primers used for qRT-PCR are for:

act-1: JEP538 ccatcatgaagtgcgacattg JEP539 catggttgatg gggcaagag;
nlp-29: JEP952 tatggaagaggatatggaggatatg JEP848 tccatg tatttactttccccatcc.

Availability of data and material

The entire dataset for the first and second rounds of the RNAi screen are publically available at http://bioinformatics.lif.univ-mrs.fr/RNAiScreen/. Data from qRT-PCR experiments are provided in Additional file 11: Table S10. Custom programming scripts are available on request.

Additional files

Additional file 1: Table S1. The list of 21,223 clones used in this study. (XLSX 355 kb)

Additional file 2: Table S2. Results for Hipi clone candidates retained in the first round of screening and identification of potential target genes. (XLSX 134 kb)

Additional file 3: Table S3. Results for Hipi clone candidates in the second and third round of screening; categorization of Hipi and Peni clones; clone verification and identification of potential target genes. (XLSX 112 kb)

Additional file 4: Supplementary Methods. Details of how Nipi clones where selected after the second round of screening. (PDF 60 kb)

Additional file 5: Table S5. Clone verification and the identification of potential target genes for 360 Nipi clones; genes that are potentially targeted by more than one clone; identification of clones potentially targeting histone genes; phenotypic characterization; comparison with published data. (XLSX 211 kb)

Additional file 6: Table S4. Results for final 360 Nipi clone candidates in the first and second round of screening; identification of clones provoking severe developmental effect. (XLSX 1068 kb)

Additional file 7: Table S6. WormNet and expression analysis systematic explorer (EASE) analyses. (XLSX 131 kb)

Additional file 8: Table S7. Species used in phylogenetic clustering listed in the same order as in Fig. 8. (XLSX 47 kb)

Additional file 9: Table S8. Analysis of spore adhesion following RNAi treatment with Nipi clones; identification of genes potentially involved in spore binding. (XLSX 132 kb)

Additional file 10: Table S9. Analysis of enrichment of predicted Nipi proteins in conserved protein complexes. (XLSX 136 kb)

Additional file 11: Table S10. Data from qRT-PCR experiments. (XLSX 22 kb)

Abbreviations

AMP: Antimicrobial peptide; CCR4-NOT: Carbon catabolite repression 4-negative regulator of transcription; cnc: Caenacin (*Caenorhabditis* bacteriocin); DIOPT: DRSC integrative ortholog prediction tool; DRSC: Drosophila RNAi screening center; DsRed: Red fluorescent protein from *Discosoma* sp.; EASE: Expression analysis systematic explorer; eIF: Eukaryotic translation initiation factor; GFP: Green fluorescent protein; GO: Gene ontology; GPCR: G-protein coupled receptor; GTP: Guanosine triphosphate; Hipi: Hyper-induction of peptide expression after infection; IRG: Infection response gene; KEGG: Kyoto encyclopedia of genes and genomes; Nipi: No induction of peptide after *Drechmeria* infection; Peni: Peptide expression no infection; ROS: Reactive oxygen species; TOF: Time of flight; TRG: Taxonomically-restricted gene; UPR: Unfolded protein response; UPR^{mt}: Mitochondrial unfolded protein response.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceptualization: NP, JJE. Methodology: OZ, NT, JB, BS, JJE. Software: NT, JB, LT. Validation: OZ, NT, BS, JJE. Formal analysis: NT, JJE. Investigation: OZ, JB, BS, CLK, JS, SO, NP. Resources: LT. Data curation: LT, JJE. Writing – original draft: OZ, JJE. Writing – review & editing: OZ, NT, BS, JS, NP, JJE. Visualization: OZ, NT, NP, JJE. Supervision: NP, JJE. Project administration: JJE. Funding acquisition: NP, JJE. All authors read and approved the final manuscript.

Acknowledgments

We thank Ben Lehner, Mirko Francesconi, Nobert Perrimon, Thomas Richardson, Robin Evans, Jerome Reboul, Peter Sarkies, Yanhui Hu, and staff at the Drosophila RNAi Screening Center; Todd Harris, Kimberly van Auken, and staff at WormBase, for discussion and advice; Justine Melo and Emily Troemel for the kind gift of strains; Clement Ghigo and Sophie Cypowyj for their contributions; Pierre Golstein and Yishi Jin for critical reading of the manuscript; and the bioinformatics platforms of the CIML and the Laboratoire d'Informatique Fondamentale de Marseille for providing computing resources, database, and web servers. We are grateful to system administrators Manuel Bertrand and Kai Poutrain for their support and input, and the anonymous reviewers for their very constructive criticisms. Some nematode strains were provided by the *Caenorhabditis* Genetics Center, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440), or by the National Bioresource Project coordinated by S. Mitani.

Funding

This work was funded by institutional grants from AMU, INSERM and CNRS, program grants from the PACA Regional Council, the ANR (MIME-2007, ANR-12-BSV3-0001-01, ANR-11-LABX-0054 (Investissements d'Avenir–Labex INFORM) and ANR-11-IDEX-0001-02 (Investissements d'Avenir–A*MIDEX)). BS was supported in part by the Fondation ARC.

Author details

[1]Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université UM2, Inserm, U1104, CNRS UMR7280, 13288 Marseille, France. [2]Institut de Mathématiques de Marseille, Aix Marseille Université, I2M Centrale Marseille, CNRS UMR 7373, 13453 Marseille, France. [3]Present address: Section of Hematology/Oncology, Department of Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [4]Present address: Institut de Biologie du Développement de Marseille, CNRS, UMR6216, Case 907, Marseille, France. [5]Present address: Institut de Genomique Fonctionnelle, 141, rue de la Cardonille, 34094 Montpellier Cedex 05, France.

Received: 19 January 2016 Accepted: 18 April 2016
Published online: 29 April 2016

References

1. Couillault C, Pujol N, Reboul J, Sabatier L, Guichou JF, Kohara Y, et al. TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human SARM. Nat Immunol. 2004;5:488–94.
2. Pujol N, Zugasti O, Wong D, Couillault C, Kurz CL, Schulenburg H, et al. Anti-fungal innate immunity in *C. elegans* is enhanced by evolutionary diversification of antimicrobial peptides. PLoS Pathog. 2008;4(7):e1000105.
3. Pujol N, Cypowyj S, Ziegler K, Millet A, Astrain A, Goncharov A, et al. Distinct innate immune responses to infection and wounding in the *C. elegans* epidermis. Curr Biol. 2008;18(7):481–9.
4. Kim DH, Ewbank JJ. Signaling in the innate immune response. In: The *C. elegans* Research Community, Editor. WormBook. 2015. p. 1–51. http://www.wormbook.org. Accessed 01 May 2016.
5. Labed SA, Omi S, Gut M, Ewbank JJ, Pujol N. The pseudokinase NIPI-4 is a novel regulator of antimicrobial peptide gene expression. PLoS One. 2012; 7(3):e33887.
6. Zugasti O, Ewbank JJ. Neuroimmune regulation of antimicrobial peptide expression by a noncanonical TGF-beta signaling pathway in *Caenorhabditis elegans* epidermis. Nat Immunol. 2009;10(3):249–56.
7. Dierking K, Polanowska J, Omi S, Engelmann I, Gut M, Lembo F, et al. Unusual regulation of a STAT protein by an SLC6 family transporter in *C. elegans* epidermal innate immunity. Cell Host Microbe. 2011;9(5):425–35.
8. Lee KZ, Kniazeva M, Han M, Pujol N, Ewbank JJ. The fatty acid synthase *fasn-1* acts upstream of WNK and Ste20/GCK-VI kinases to modulate

antimicrobial peptide expression in *C. elegans* epidermis. Virulence. 2010; 1(3):113–22.

9. Ziegler K, Kurz CL, Cypowyj S, Couillault C, Pophillat M, Pujol N, et al. Antifungal innate immunity in *C. elegans*: PKCdelta links G protein signaling and a conserved p38 MAPK cascade. Cell Host Microbe. 2009;5(4):341–52.
10. van Oosten-Hawle P, Morimoto RI. Transcellular chaperone signaling: an organismal strategy for integrated cell stress responses. J Exp Biol. 2014; 217(Pt 1):129–36.
11. Taylor RC, Berendzen KM, Dillin A. Systemic stress signalling: understanding the cell non-autonomous control of proteostasis. Nat Rev Mol Cell Biol. 2014;15(3):211–7.
12. Ewbank JJ, Pujol N. Local and long-range activation of innate immunity by infection and damage in *C. elegans*. Curr Opin Immunol. 2015;38:1–7.
13. Ahringer J. Reverse genetics. In: The *C. elegans* Research Community, Editor. WormBook. 2006. doi:10.1895/wormbook.1.47.1. http://www.wormbook.org. Accesse 01 May 2016.
14. Boutros M, Ahringer J. The art and design of genetic screens: RNA interference. Nat Rev Genet. 2008;9(7):554–66.
15. Kurz CL, Ewbank JJ. Infection in a dish: high-throughput analyses of bacterial pathogenesis. Curr Opin Microbiol. 2007;10(1):10–6.
16. Pukkila-Worley R, Holson E, Wagner F, Mylonakis E. Antifungal drug discovery through the study of invertebrate model hosts. Curr Med Chem. 2009;16(13):1588–95.
17. Ewbank JJ, Zugasti O. *C. elegans*: model host and tool for antimicrobial drug discovery. Dis Model Mech. 2011;4(3):300–4.
18. Pulak R. Techniques for analysis, sorting, and dispensing of *C. elegans* on the COPAS flow-sorting system. Methods Mol Biol. 2006;351:275–86.
19. Squiban B, Belougne J, Ewbank J, Zugasti O. Quantitative and automated high-throughput genome-wide RNAi screens in *C. elegans*. J Vis Exp. 2012; 60:e3448.
20. Zugasti O, Bose N, Squiban B, Belougne J, Kurz CL, Schroeder FC, et al. Activation of a G protein-coupled receptor by its endogenous ligand triggers the innate immune response of *Caenorhabditis elegans*. Nat Immunol. 2014;15(9):833–8.
21. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature. 2003;421(6920):231–7.
22. Rual JF, Ceron J, Koreth J, Hao T, Nicot AS, Hirozane-Kishikawa T, et al. Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. Genome Res. 2004;14(10B):2162–8.
23. Friedman A, Perrimon N. Genetic screening for signal transduction in the era of network biology. Cell. 2007;128(2):225–31.
24. Murray PJ, Smale ST. Restraint of inflammatory signaling by interdependent strata of negative regulatory pathways. Nat Immunol. 2012;13(10):916–24.
25. Arthur JS, Ley SC. Mitogen-activated protein kinases in innate immunity. Nat Rev Immunol. 2013;13(9):679–92.
26. Melo JA, Ruvkun G. Inactivation of conserved *C. elegans* genes engages pathogen- and xenobiotic-associated defenses. Cell. 2012; 149(2):452–66.
27. Ward JD, Mullaney B, Schiller BJ, le He D, Petnic SE, Couillault C, et al. Defects in the *C. elegans* acyl-CoA synthase, *acs-3*, and nuclear hormone receptor, *nhr-25*, cause sensitivity to distinct, but overlapping stresses. PLoS One. 2014;9(3):e92552.
28. Thakur N, Pujol N, Tichit L, Ewbank JJ. Clone mapper: an online suite of tools for RNAi experiments in *Caenorhabditis elegans*. G3. 2014;4(11):2137–45.
29. Palaima E, Leymarie N, Stroud D, Mizanur RM, Hodgkin J, Gravato-Nobre MJ, et al. The *Caenorhabditis elegans* *bus-2* mutant reveals a new class of O-glycans affecting bacterial resistance. J Biol Chem. 2010;285(23):17662–72.
30. Gravato-Nobre MJ, Stroud D, O'Rourke D, Darby C, Hodgkin J. Glycosylation genes expressed in seam cells determine complex surface properties and bacterial adhesion to the cuticle of *Caenorhabditis elegans*. Genetics. 2011; 187(1):141–55.
31. Rouger V, Bordet G, Couillault C, Monneret S, Mailfert S, Ewbank JJ, et al. Independent synchronized control and visualization of interactions between living cells and organisms. Biophys J. 2014;106(10):2096–104.
32. Hansen M, Hsu AL, Dillin A, Kenyon C. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. PLoS Genet. 2005;1(1):119–28.
33. Lehmann S, Shephard F, Jacobson LA, Szewczyk NJ. Using multiple phenotype assays and epistasis testing to enhance the reliability of RNAi screening and identify regulators of muscle protein degradation. Genes (Basel). 2012;3(4):686–701.
34. Goto A, Matsushita K, Gesellchen V, El Chamy L, Kuttenkeuler D, Takeuchi O, et al. Akirins are highly conserved nuclear proteins required for NF-kappaB-dependent gene expression in drosophila and mice. Nat Immunol. 2008; 9(1):97–104.
35. Bonnay F, Nguyen XH, Cohen-Berros E, Troxler L, Batsche E, Camonis J, et al. Akirin specifies NF-kappaB selectivity of *Drosophila* innate immune response via chromatin remodeling. Embo J. 2014;33(20):2349–62.
36. Vanhoven MK, Bauer Huang SL, Albin SD, Bargmann CI. The claudin superfamily protein *nsy-4* biases lateral signaling to generate left-right asymmetry in *C. elegans* olfactory neurons. Neuron. 2006;51(3):291–302.
37. Andrusiak MG, Jin Y. Context specificity of stress-activated MAP Kinase signaling: the story as told by *C. elegans*. J Biol Chem. 2016;291: 7796–804.
38. Chavel CA, Caccamise LM, Li B, Cullen PJ. Global regulation of a differentiation MAPK pathway in yeast. Genetics. 2014;198(3):1309–28.
39. Friedman AA, Tucker G, Singh R, Yan D, Vinayagam A, Hu Y, et al. Proteomic and functional genomic landscape of receptor tyrosine kinase and ras to extracellular signal-regulated kinase signaling. Sci Signal. 2011; 4(196):rs10.
40. Bandyopadhyay S, Chiang CY, Srivastava J, Gersten M, White S, Bell R, et al. A human MAP kinase interactome. Nat Methods. 2010;7(10):801–5.
41. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011;12:357.
42. Cho A, Shin J, Hwang S, Kim C, Shim H, Kim H, et al. WormNet v3: a network-assisted hypothesis-generating server for *Caenorhabditis elegans*. Nucleic Acids Res. 2014;42(Web Server issue):W76–82.
43. Kim DH, Liberati NT, Mizuno T, Inoue H, Hisamoto N, Matsumoto K, et al. Integration of *Caenorhabditis elegans* MAPK pathways mediating immunity and stress resistance by MEK-1 MAPK kinase and VHP-1 MAPK phosphatase. Proc Natl Acad Sci U S A. 2004;101(30):10990–4.
44. Chauhan VM, Orsi G, Brown A, Pritchard DI, Aylott JW. Mapping the pharyngeal and intestinal pH of *Caenorhabditis elegans* and real-time luminal pH oscillations using extended dynamic range pH-sensitive nanosensors. ACS Nano. 2013;7(6):5577–87.
45. Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003;4(10): R70.
46. Engelmann I, Griffon A, Tichit L, Montanana-Sanchis F, Wang G, Reinke V. A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans*. PLoS One. 2011;6(5):e19055.
47. Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, et al. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. Nature. 2014;512(7515):400–5.
48. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
49. Shore DE, Carr CE, Ruvkun G. Induction of cytoprotective pathways is central to the extension of lifespan conferred by multiple longevity pathways. PLoS Genet. 2012;8(7):e1002792.
50. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355–60.
51. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, et al. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015;43(Database issue):D690–7.
52. Tanabe M, Kanehisa M. Using the KEGG database resource. Curr Protoc Bioinformatics. 2012; Chapter 1:Unit1 12. doi: 10.1002/0471250953.bi0112s38.
53. Flockhart IT, Booker M, Hu Y, McElvany B, Gilly Q, Mathey-Prevot B, et al. FlyRNAi.org–the database of the Drosophila RNAi screening center: 2012 update. Nucleic Acids Res. 2012;40(Database issue):D715–9.
54. Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Res. 2013;41(Database issue):D1021–6.
55. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404–13.

56. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999;96(8):4285–8.
57. Tabach Y, Billi AC, Hayes GD, Newman MA, Zuk O, Gabel H, et al. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. Nature. 2013;493(7434):694–8.
58. Rohlfing AK, Miteva Y, Hannenhalli S, Lamitina T. Genetic and physiological activation of osmosensitive gene expression mimics transcriptional signatures of pathogen infection in *C. elegans*. PLoS One. 2010;5(2):e9010.
59. Dunbar TL, Yan Z, Balla KM, Smelkinson MG, Troemel ER. *C. elegan s* detects pathogen-induced translational inhibition to activate immune signaling. Cell Host Microbe. 2012;11(4):375–86.
60. Simmer F, Moorman C, Van Der Linden AM, Kuijk E, Van Den Berghe PV, Kamath R, et al. Genome-wide RNAi of *C. elegans* using the hypersensitive *rrf-3* strain reveals novel gene functions. PLoS Biol. 2003;1(1):E12.
61. Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, et al. Panorama of ancient metazoan macromolecular complexes. Nature. 2015;525(7569):339–44.
62. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440(7084):631–6.
63. Lenstra TL, Benschop JJ, Kim T, Schulze JM, Brabers NA, Margaritis T, et al. The specificity and topology of chromatin interaction pathways in yeast. Mol Cell. 2011;42(4):536–49.
64. Benschop JJ, Brabers N, van Leenen D, Bakker LV, van Deutekom HW, van Berkum NL, et al. A consensus of core protein complex compositions for *Saccharomyces cerevisiae*. Mol Cell. 2010;38(6):916–28.
65. Inada T, Makino S. Novel roles of the multi-functional CCR4-NOT complex in post-transcriptional regulation. Front Genet. 2014;5:135.
66. Bennett CF, Vander Wende H, Simko M, Klum S, Barfield S, Choi H, et al. Activation of the mitochondrial unfolded protein response does not predict longevity in *Caenorhabditis elegans*. Nat Commun. 2014;5:3483.
67. Pellegrino MW, Nargund AM, Kirienko NV, Gillis R, Fiorese CJ, Haynes CM. Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection. Nature. 2014;516(7531):414–7.
68. Pellegrino MW, Haynes CM. Mitophagy and the mitochondrial unfolded protein response in neurodegeneration and bacterial infection. BMC Biol. 2015;13:22.
69. Nargund AM, Pellegrino MW, Fiorese CJ, Baker BM, Haynes CM. Mitochondrial import efficiency of ATFS-1 regulates mitochondrial UPR activation. Science. 2012;337(6094):587–90.
70. Durieux J, Wolff S, Dillin A. The cell-non-autonomous nature of electron transport chain-mediated longevity. Cell. 2011;144(1):79–91.
71. Kamath RS, Ahringer J. Genome-wide RNAi screening in *Caenorhabditis elegans*. Methods. 2003;30(4):313–21.
72. Rohlfing AK, Miteva Y, Moronetti L, He L, Lamitina T. The *Caenorhabditis elegans* mucin-like protein OSM-8 negatively regulates osmosensitive physiology via the transmembrane protein PTR-23. PLoS Genet. 2011;7(1): e1001267.
73. Watson E, MacNeil LT, Arda HE, Zhu LJ, Walhout AJ. Integration of metabolic and gene regulatory networks modulates the *C. elegans* dietary response. Cell. 2013;153(1):253–66.
74. Liu Y, Samuel BS, Breen PC, Ruvkun G. *Caenorhabditis elegans* pathways that surveil and defend mitochondria. Nature. 2014;508(7496):406–10.
75. Lamitina T, Huang CG, Strange K. Genome-wide RNAi screening identifies protein damage as a regulator of osmoprotective gene expression. Proc Natl Acad Sci U S A. 2006;103(32):12173–8.
76. Updike DL, Strome S. A genomewide RNAi screen for genes that affect the stability, distribution and function of P granules in *Caenorhabditis elegans*. Genetics. 2009;183(4):1397–419.
77. Nollen EA, Garcia SM, van Haaften G, Kim S, Chavez A, Morimoto RI, et al. Genome-wide RNA interference screen identifies previously undescribed regulators of polyglutamine aggregation. Proc Natl Acad Sci USA. 2004; 101(17):6403–8.
78. MacNeil LT, Pons C, Arda HE, Giese GE, Myers CL, Walhout AJ. Transcription factor activity mapping of a tissue-specific gene regulatory network. Cell Syst. 2015;1(2):152–62.
79. Winter JF, Hopfner S, Korn K, Farnung BO, Bradshaw CR, Marsico G, et al. *Caenorhabditis elegans* screen reveals role of PAR-5 in RAB-11-recycling endosome positioning and apicobasal cell polarity. Nat Cell Biol. 2012;14(7): 666–76.
80. Morton E, Lamitina T. A suite of MATLAB-based computational tools for automated analysis of COPAS Biosort data. Biotechniques. 2010;48(6):xxv–x.
81. Montanana F, Julien RA, Vaglio P, Matthews LR, Tichit L, Ewbank JJ. ICeE: an interface for *C. elegans* experiments. Worm. 2014;3(3):e959420.
82. Marza E, Taouji S, Barroso K, Raymond AA, Guignard L, Bonneu M, et al. Genome-wide screen identifies a novel p97/CDC-48-dependent pathway regulating ER-stress-induced gene transcription. EMBO Rep. 2015;16(3):332–40.
83. Squiban B. Criblage par ARN interférence du génome complet de *C. elegans* pour l' identification de nouveaux gènes impliqués dans l' immunité innée. Marseille: Aix-Marseille Université; 2012.
84. Davies KG, Curtis RH. Cuticle surface coat of plant-parasitic nematodes. Annu Rev Phytopathol. 2011;49:135–56.
85. Page AP, Johnstone IL. The cuticle. In: The *C. elegans* Research Community, Editor. WormBook. 2007. doi:10.1895/wormbook.1.138.1:1-15. Accessed 01 May 2016.
86. Gravato-Nobre MJ, Nicholas HR, Nijland R, O'Rourke D, Whittington DE, Yook KJ, et al. Multiple genes affect sensitivity of *Caenorhabditis elegans* to the bacterial pathogen *Microbacterium nematophilum*. Genetics. 2005;171(3):1033–45.
87. McNew JA, Sogaard M, Lampen NM, Machida S, Ye RR, Lacomis L, et al. Ykt6p, a prenylated SNARE essential for endoplasmic reticulum-Golgi transport. J Biol Chem. 1997;272(28):17776–83.
88. Billing O, Natarajan B, Mohammed A, Naredi P, Kao G. A directed RNAi screen based on larval growth arrest reveals new modifiers of *C. elegans* insulin signaling. PLoS One. 2012;7(4):e34507.
89. Luallen RJ, Bakowski MA, Troemel ER. Characterization of microsporidia-induced developmental arrest and a transmembrane leucine-rich repeat protein in *Caenorhabditis elegans*. PLoS One. 2015;10(4):e0124065.
90. Engelmann I, Pujol N. Innate Immunity in *C. elegans*. In: Söderhäll K, editor. Invertebrate Immunity. Austin, TX: Landes Bioscience; 2010. p. 1–17.
91. Lamitina ST, Morrison R, Moeckel GW, Strange K. Adaptation of the nematode *Caenorhabditis elegans* to extreme osmotic stress. Am J Physiol Cell Physiol. 2004;286(4):C785–91.
92. Lee EC, Strange K. GCN-2 dependent inhibition of protein synthesis activates osmosensitive gene transcription via WNK and Ste20 kinase signaling. Am J Physiol Cell Physiol. 2012;303(12):C1269–77.
93. Mizuno T, Hisamoto N, Terada T, Kondo T, Adachi M, Nishida E, et al. The *Caenorhabditis elegans* MAPK phosphatase VHP-1 mediates a novel JNK-like signaling pathway in stress response. Embo J. 2004;23(11):2226–34.
94. Runkel ED, Liu S, Baumeister R, Schulze E. Surveillance-activated defenses block the ROS-induced mitochondrial unfolded protein response. PLoS Genet. 2013;9(3):e1003346.
95. Baker BM, Nargund AM, Sun T, Haynes CM. Protective coupling of mitochondrial function and protein synthesis via the eIF2alpha kinase GCN-2. PLoS Genet. 2012;8(6):e1002760.
96. McEwan DL, Kirienko NV, Ausubel FM. Host translational inhibition by *Pseudomonas aeruginosa* Exotoxin A Triggers an immune response in *Caenorhabditis elegans*. Cell Host Microbe. 2012;11(4):364–74.
97. Cohen LB, Troemel ER. Microbial pathogenesis and host defense in the nematode *C. elegans*. Curr Opin Microbiol. 2015;23C:94–101.
98. Nandakumar M, Tan MW. Gamma-linolenic and stearidonic acids are required for basal immunity in *Caenorhabditis elegans* through their effects on p38 MAP kinase activity. PLoS Genet. 2008;4(11):e1000273.
99. Wong D, Bazopoulou D, Pujol N, Tavernarakis N, Ewbank JJ. Genome-wide investigation reveals pathogen-specific and shared signatures in the response of *Caenorhabditis elegans* to infection. Genome Biol. 2007;8(9):R194.
100. Taubert S, Van Gilst MR, Hansen M, Yamamoto KR. A Mediator subunit, MDT-15, integrates regulation of fatty acid metabolism by NHR-49-dependent and -independent pathways in *C. elegans*. Genes Dev. 2006; 20(9):1137–49.
101. Zhang P, Judy M, Lee SJ, Kenyon C. Direct and indirect gene regulation by a life-extending FOXO protein in *C. elegans*: roles for GATA factors and lipid gene regulators. Cell Metab. 2013;17(1):85–100.
102. Taubert S, Hansen M, Van Gilst MR, Cooper SB, Yamamoto KR. The Mediator subunit MDT-15 confers metabolic adaptation to ingested material. PLoS Genet. 2008;4(2):e1000021.
103. Oliveira RP, Porter Abate J, Dilks K, Landis J, Ashraf J, Murphy CT, et al. Condition-adapted stress and longevity gene regulation by *Caenorhabditis elegans* SKN-1/Nrf. Aging Cell. 2009;8(5):524–41.
104. Goh GY, Martelli KL, Parhar KS, Kwong AW, Wong MA, Mah A, et al. The conserved Mediator subunit MDT-15 is required for oxidative stress responses in *Caenorhabditis elegans*. Aging Cell. 2014;13(1):70–9.

105. Pukkila-Worley R, Feinbaum RL, McEwan DL, Conery AL, Ausubel FM. The evolutionarily conserved Mediator subunit MDT-15/MED15 links protective innate immune responses and xenobiotic detoxification. PLoS Pathog. 2014; 10(5):e1004143.
106. Blasius AL, Beutler B. Intracellular toll-like receptors. Immunity. 2010;32(3): 305–15.
107. Huang HR, Chen ZJ, Kunes S, Chang GD, Maniatis T. Endocytic pathway is required for Drosophila Toll innate immune signaling. Proc Natl Acad Sci U S A. 2010;107(18):8322–7.
108. Vandenabeele P, Bertrand MJ. The role of the IAP E3 ubiquitin ligases in regulating pattern-recognition receptor signalling. Nat Rev Immunol. 2012; 12(12):833–44.
109. Zinngrebe J, Montinaro A, Peltzer N, Walczak H. Ubiquitin in the immune system. EMBO Rep. 2014;15(1):28–45.
110. Moynagh PN. The roles of Pellino E3 ubiquitin ligases in immunity. Nat Rev Immunol. 2014;14(2):122–31.
111. Davis ME, Gack MU. Ubiquitination in the antiviral immune response. Virology. 2015;479–480:52–65.
112. Chapat C, Corbo L. Novel roles of the CCR4-NOT complex. Wiley Interdiscip Rev RNA. 2014;5(6):883–901.
113. Boehm T. Evolution of vertebrate immunity. Curr Biol. 2012;22(17):R722–32.
114. Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, Cooper MD. VLR-based adaptive immunity. Annu Rev Immunol. 2012;30:203–20.
115. Flajnik MF. Re-evaluation of the immunological Big Bang. Curr Biol. 2014; 24(21):R1060–5.
116. Du Pasquier L. Metazoa immune receptors diversification during evolution. Med Sci (Paris). 2009;25(3):273–80.
117. Schulenburg H, Kurz CL, Ewbank JJ. Evolution of the innate immune system: the worm perspective. Immunol Rev. 2004;198:36–58.
118. Irazoqui JE, Urbach JM, Ausubel FM. Evolution of host innate defence: insights from *Caenorhabditis elegans* and primitive invertebrates. Nat Rev Immunol. 2010;10(1):47–58.
119. Zhang L, Li L, Guo X, Litman GW, Dishaw LJ, Zhang G. Massive expansion and functional divergence of innate immune genes in a protostome. Sci Rep. 2015;5:8693.
120. Simmer F, Tijsterman M, Parrish S, Koushika SP, Nonet ML, Fire A, et al. Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. Curr Biol. 2002;12(15):1317–9.
121. Powell JR, Ausubel FM. Models of *Caenorhabditis elegans* infection by bacterial and fungal pathogens. In: Ewbank J, Vivier E, editors. Methods Mol Biol, vol. 415. New York, NY: Humana Press; 2008. p. 403–27.
122. Sagasti A, Hisamoto N, Hyodo J, Tanaka-Hino M, Matsumoto K, Bargmann CI. The CaMKII UNC-43 activates the MAPKKK NSY-1 to execute a lateral signaling decision required for asymmetric olfactory neuron fates. Cell. 2001; 105(2):221–32.
123. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics. 1999;15(9):759–62.
124. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, et al. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet. 2004;36(2):197–204.
125. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, et al. A gene expression map for *Caenorhabditis elegans*. Science. 2001;293(5537):2087–92.
126. Ceron J, Rual JF, Chandra A, Dupuy D, Vidal M, van den Heuvel S. Large-scale RNAi screens identify novel genes that interact with the *C. elegans* retinoblastoma pathway as well as splicing-related components with synMuv B activity. BMC Dev Biol. 2007;7:30.
127. McElwee JJ, Schuster E, Blanc E, Thomas JH, Gems D. Shared transcriptional signature in *Caenorhabditis elegans* Dauer larvae and long-lived *daf-2* mutants implicates detoxification system in longevity assurance. J Biol Chem. 2004;279(43):44533–43.
128. Mansisidor AR, Cecere G, Hoersch S, Jensen MB, Kawli T, Kennedy LM, et al. A conserved PHD finger protein and endogenous RNAi modulate insulin signaling in *Caenorhabditis elegans*. PLoS Genet. 2011;7(9):e1002299.
129. Love DC, Ghosh S, Mondoux MA, Fukushige T, Wang P, Wilson MA, et al. Dynamic O-GlcNAc cycling at promoters of *Caenorhabditis elegans* genes regulating longevity, stress, and immunity. Proc Natl Acad Sci U S A. 2010; 107(16):7413–8.
130. Guisbert E, Czyz DM, Richter K, McMullen PD, Morimoto RI. Identification of a tissue-selective heat shock response regulatory network. PLoS Genet. 2013;9(4):e1003466.
131. Parry DH, Xu J, Ruvkun G. A whole-genome RNAi Screen for *C. elegans* miRNA pathway genes. Curr Biol. 2007;17(23):2013–22.
132. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, et al. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. Nat Struct Mol Biol. 2010;17(2):173–9.
133. Curran SP, Ruvkun G. Lifespan regulation by evolutionarily conserved genes essential for viability. PLoS Genet. 2007;3(4):e56.
134. Mabon ME, Mao X, Jiao Y, Scott BA, Crowder CM. Systematic identification of gene activities promoting hypoxic death. Genetics. 2009;181(2):483–96.

2.3 Publication 3

Global biological analyses through integrated functional and phylogenetic profiling.
Nishant Thakur, Nathalie Pujol, Jacques van Helden, Laurent Tichit, Jonathan J. Ewbank.
In this publication, I developed a public *C. elegans*-specific functional enrichment tool for high throughput data analysis called YAAT. It incorporates a novel way of analysing high-throughput data using phylogenetic profiles. I constructed a large database of function annotations representing diverse data types, some not found in other tools. I established a way to keep this underlying database up-to-date automatically. I benchmarked the tool through global analyses of all 4700 datasets and showed its advantages compared to the only other existing *C. elegans* tool. I analysed various infection-related datasets to draw general conclusions about the evolution of innate immune regulators and effectors.

Global biological analyses through integrated functional and phylogenetic profiling.

Nishant Thakur[1], Nathalie Pujol[1], Laurent Tichit[2], Jonathan J. Ewbank[1]
[+other authors?]

[1]Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université, Inserm, CNRS, Marseille, France
[2]Institut de Mathématiques de Marseille, Aix Marseille Université, I2M Centrale Marseille, CNRS UMR 7373, 13453 Marseille, France

Author for correspondence: ewbank@ciml.univ-mrs.fr

NOTE: This is a draft MS. Certain figures, needs to be updated to the new Wormbase release before submission and there is no abstract.

INTRODUCTION

In recent years, there has been a boom in genomic, transcriptomic and epigenomic studies, largely fuelled by advances in sequencing technologies and the attendant reduction in costs. They often result in the production of long lists of candidate genes. Various publicly-available resources classify genes on the basis of their structure, interactions or function. A common way to interpret a new gene list is to exploit this available knowledge, looking for enrichment of genes in defined classes. This technique of gene functional enrichment has been used for more than a decade [1] and there are currently dozens of available tool, many listed at http://omictools.com/.

Despite having such a long list of tools, Wadi et al. recently raised several important issues concerning even the highly cited ones. The main problem with these tools is that many of them are not regularly updated. Among the 21 most popular tools, 12 have not been updated in last five years; 84% of citations to gene enrichment tools are for those that are outdated [2]. For example, the data for the current (June 2016) production version for DAVID (6.7) a tool cited more than 4,000 times in 2015, has not been updated since 2009. This is a problem for 2 reasons. Firstly, gene structure predictions change. For the model organism *Caenorhabditis elegans*, between the WS220 release of the database Wormbase in 2009 and WS252 (released in 2016), for instance, more than 1700 new genes have been defined, and the predicted structure of hundreds of existing genes modified. If tools do not update their source data, an increasing number of genes are either incorrectly associated with an annotation, or are simply absent. Secondly, these tools lack up-to-date information about gene function. Recently, gene ontology (GO) annotations have increased on average by 12.5% every year [2]. There is a similar trend for other annotations, like those from Reactome and KEGG. Thus, for example, fully 80% of currently available functional annotations are absent from DAVID 6.7. This is clearly a severe limitation, and can substantially bias analyses of gene lists [2].

Some tools that perform gene enrichment analyses are updated regularly. One example is g:Profiler [3]. Despite its power, in common with similar tools, g:Profiler captures annotations for multiple species from centralized databases, rather than leveraging the annotations available in species-specific databases. For example, in Wormbase WS252, 14,557 genes are associated with a GO annotation, but 7% (1,023 as of 4/7/2016) of these are not present in the GO Consortium's compendium. Similarly, Wormbase contains more-or-less detailed descriptions of the phenotypes associated with mutation or RNAi-knockdown for ~40% of genes. Although much of this information in time reaches generic databases, there is always a lag, so at a given moment, for some annotations, Wormbase is the sole source. It also harbours more than 1600 transcriptome expression datasets that are manually curated to a standard machine-readable format, allowing automated retrieval and data integration. This represents a powerful resource for functional enrichment analysis that is more readily exploited than the raw data available through the main transcriptome databases such as the Gene Expression Omnibus (GEO).

While generic databases tools have a broad appeal, there is a clear demand for model organism databases (MODs) [4] and tools that are species-specific. The recently published WormExp is one of many *C. elegans*-specific bioinformatic tools. It overcomes some of the limitations of the more generic ones, but is deliberately limited to transcriptome data [5]. Others include a database of time-resolved expression data [6, 7], catalogues of potential transcription factor binding sites, established on the basis of the DNA conservation (cisRED [8] or ChiPseq experiments (motif-disc [9]), tools such as GExplore and WormMine for large-scale data mining related to gene or protein function [10, 11] and WormNet that can generate new members for a pathway, infer functions from network neighbours, or predict "hub" genes, on the basis of annotations and expression data [12]. On the other hand, to the best of our knowledge, there are no dedicated tools for enrichment analysis that exploit the rich set of data available for *C. elegans*. We report here the creation of the web-accessible tool YAAT (for "yet another analysis tool") that performs enrichment analysis using a very extensive dataset (> 4700 classes) gathered in an automated manner and designed from the outset to be readily updated. To increase its utility, we have supplemented it with several complementary analytical methods, enriched class clustering, evaluation of conservation, and phylogenetic profiling. Global analyses with YAAT revealed hitherto unsuspected biases in the underlying datasets that have the potential to skew results. With this caveat in mind, in a proof of principle, we use YAAT and a previously unpublished RNAseq dataset to undertake a comparative analysis of the innate immune response of *C. elegans* to 2 of its natural pathogens, the Ophiocordycipitaceae family fungi *Haptocillium sphaerosporum* [13] and the well-studied *Drechmeria coniospora* [14-16]. Although infection by both pathogens starts with the attachment of non-motile spores to the nematode cuticle, as revealed by our analysis, *C. elegans* mounts very different transcriptional responses against them, characterised by marked differences in the coverage of the differentially regulated genes in the reference datasets, the enriched classes that are associated with each infection model, the overall level of conservation of the differentially expressed genes, and their patterns of conservation across species.

We extended the analysis using a selection of 56 pathogen-related datasets that we further subdivided on a structural or functional basis, to derive general conclusions about the nature of the nematode innate immune response [NB. Not in current version of MS!]. Through these examples, we demonstrate how YAAT's suite of online tools could be of broad utility to the large community of *C. elegans* researchers, and may inspire those working with other model systems to create equivalent species-specific tools.

MATERIAL AND METHODS

Data collection

To constitute the database for functional enrichment analysis, we expanded our previous collection of expression and phenotypic data from diverse resources [17]. Firstly, for expression data we extracted ~1600 Serial Pattern of Expression Levels Locator (SPELL) expression clusters, curated from 320 articles, from the Wormbase FTP site ftp://caltech.wormbase.org/pub/wormbase/spell_download/tables/. We supplemented this data with non-redundant datasets from WormExp [5]. As neither Wormbase expression clusters nor WormExp are comprehensive, we manually curated a further 188 datasets from 71 articles, building on our previously described set derived from 69 articles [15]. Most of the phenotypic data was extracted from Wormbase. Thus 1883 phenotypic classes, associated with a total of 7948 genes, were downloaded from Wormbase release WS252 (ftp://ftp.wormbase.org/pub/wormbase/releases/WS252/ONTOLOGY/phenotype_association.WS252.wb). Further, 101 phenotypic classes were extracted from DRSC FlyRNAi database "http://www.flyrnai.org/RNAi_all_hits.txt" and the correspondence between fly genes and their nematode homologues established as previously described [17]. We also automatically extracted the list of genes associated with each of the 129 *C. elegans* pathways present in KEGG. Finally, we added regulatory information into YAAT by including 202 dataset for the putative gene targets for 94 transcription factors from the last modENCODE release [9]. Collectively, these form a very extensive resource of 4723 datasets.

Enrichment analysis

For functional enrichment statistics, we used hypergeometric distribution using the R "stats" package. The p-value is calculated as

phyper(q = x-1, m = m, n = n, k = k, lower.tail = FALSE)

Where, k is the number of genes in the user's query.

N is total number of genes annotated in the catalogue of reference classes (i.e. 20173 here).

m is the number of genes annotated in the functional class of interest.

x is number of genes in the user list that are annotated in the functional class of interest.

n= N - m

For multiple comparisons, p-values are corrected with the p.adjust function in the R "stats" package, with various types of correctional statistics: ("holm" [18]), ("hochberg" [19]), ("hommel" [20]), ("BH" or its alias "fdr" and "BY" [21]).

Clustering of enriched classes.

Having defined the enriched classes for the user's list of genes, a binary matrix of enriched classes and the overlapping genes in each class is generated, where rows represents genes found in at least one enriched class and columns represents enriched classes. Genes common to the user list and the enriched class are set to be "1", otherwise genes are assigned the value "0". The distance between all the genes and all the enriched classes are calculated and based on the distances, hierarchical clustering is performed using "hclust" function in R.

Phylogenetic profiles

For phylogenetic profile analysis, we made 3 reference sets of proteomes from different ranges of species. The first dataset, NemaProf, contains 72 free-living nematode species. The second also includes 29 parasitic species (nematodes and platyhelminthes, collectively referred to as helminths; HelmProf). Both were based on the complete set of proteins predicted from genome sequences, extracted from

Wormbase. The third dataset, EukaProf, contains 56 eukaryote species (vertebrates, invertebrates, fungi, plants and protists). For this dataset, out of 66 reference proteomes defined by the Quest For Orthologs consortium (http://questfororthologs.org/; data release 2016/05/03), the sequences for the proteins of 56 species available at Ensembl/Ensemblgenomes were downloaded. Each *C. elegans* protein was compared to the complete 56 proteomes (including *C. elegans*) using BLASTP. When multiple isoforms existed for a single protein, the longest one was selected. For each *C. elegans* query protein, only the best hit in each species was considered. Based on this analysis, for each of the 3 datasets, we generated a “BinMat” matrix (n proteins × m species), where for each entry C_{ij}, if there was a BLASTP hit in species “j” for *C. elegans* protein “i” with a P-value less than 10^{-5}, then C_{ij} was set to “1”, and otherwise to “0”. Users provide lists of proteins or genes. In the latter case, for each gene, the corresponding protein (the longest isoform, if applicable) is identified. Starting with this list of “N” genes, the tool will extract a “QuBinMat” matrix (N proteins × m species) from “BinMat”. Next, the distances between the “N” proteins within the QuBinMat matrix are calculated and based on these distances, clustering is performed with the “hclust” clustering algorithm in R.

Fungal culture, infection and RNAseq analysis

Haptocillium sphaerosporum strain CBS889.85, the kind gift of Jan Dijksterhuis, CBS Fungal Biodiversity Center, Utrecht, was cultured by passaging through *C. elegans*, and spores collected essentially as described for *D. coniospora* [22]. Fresh spores were used to infect a synchronized population of L4 N2 worms cultivated 25°C. After 6 h, worms were harvested and processed for RNA extraction as previously described [15]. Age-matched, non-infected worms were used as a control. cDNA was generated from extracted polyA RNA following standard techniques [23]. RNA-seq libraries and sequencing was performed as described [24] except read lengths were 76 rather than 36 bases. Reads were aligned to the WS220 *C. elegans* genome and read counts normalized as previously described [24] to give the average depth of coverage per million reads (dcpm) [15]. Differentially-expressed genes were defined using the previously described algorithms [15].

An infection-specific dataset

A dataset was assembled containing genes identified as being differentially-regulated following infection with microsporidia (*Nematocida parisii*), fungi (*Drechmeria coniospora* and *Candida albicans*), bacteria (*Lactobacillus rhamnosus*, *Staphylococcus aureus*, *Comamonas aquatica*, *Bacillus thuringiensis, Mycobacterium nematophilus, Erwinia carotovora, Yersinia pestis, Enterococcus faecalis, Photorhabdus luminescens* and *Vibrio cholera*) and *Orsay* virus, giving in total 140 SPELL expression datasets. We then restricted the datasets to those with a maximum of 500 genes to avoid biasing the analysis. Among remaining 99 datasets, we merged time course datasets into single datasets (e.g. *Nematocida parisii* 8, 16, 30, 34, 40, 64h post-infection data merged into unique dataset). From the remaining 56 datasets containing a total of 4700 genes, we made a further refinement, and retained only genes that were found in at least 2 datasets to give a final dataset of ~1900 genes.

RESULTS

Constitution of the YAAT dataset

In order to analyse a large dataset of genes derived from a genome-wide RNAi screen, we recently developed a functional class enrichment and clustering tool [17]. Given the interest generated by this tool, we decided to improve it and adapt it to provide the web-based resource YAAT. As mentioned above, an important limitation of many available tools is the fact that they are not maintained so that the underlying data quickly becomes out-dated [2]. A prerequisite for our tool was therefore that it could be updated in a simple, ideally automated manner. Much of the source data for YAAT is derived from Wormbase, which has a regular release schedule (currently every 2 months). YAAT therefore periodically interrogates Wormbase to determine whether there has been a new release, and if this is the case, retrieves the necessary data from the Wormbase FTP site. Other data sources, on the other hand, are static. For example, the association of transcription factors and putative target genes established by the modENCODE consortium was generated using WS220 [9] and has not been updated. Other data sources, including KEGG [25], are updated regularly, but not necessarily synchronised with Wormbase releases. We previously developed a tool Wormbase Converter to deal specifically with this issue [15]. We used Wormbase Converter to ensure that we had a homogeneous set of data, using a single reference release of Wormbase genes, irrespective of the data's source.

A single gene can give rise to multiple transcripts, and hence to different protein isoforms. Some functional annotations are associated with a specific transcript, but we did not attempt to capture this level of detail in YAAT; the reference object is a gene (with unique Wormbase gene identifier). Indeed, for annotations associated with proteins (e.g. for enzyme activities or derived from proteomic studies), the corresponding gene name was used. Thus for the sake of simplicity, here we use the term "gene" in an indiscriminate way when describing classes that were defined on the gene, transcript or protein level.

Two major sources of data were the pre-computed expression clusters from the *C. elegans* implementation of Serial Pattern of Expression Levels Locator (SPELL [26]) and WormExp [5]. These sources provide a partially redundant but not comprehensive coverage of the available transcriptome data. SPELL includes only published datasets, while the underlying data in WormExp comes from the many publicly-available transcriptome databases, but not SPELL, and includes unpublished data. Both databases include expression values for genes. Our pre-existing in-house functional class database also contained data manually extracted from transcriptome studies. Since it only references gene identities, not expression values, it includes some studies for which the expression values were never made publicly-available and that are therefore absent from SPELL and WormExp (Figure 1A). Our in-house database was not restricted to transcriptome data, but additionally included lists of genes from functional screens and structural categories. The different sources were combined in the YAAT database to give 4723 functional classes (Figure 1B). This set covers essentially all (>99%) protein-coding genes in C*. elegans*. The size of classes ranged from 1 to 10,731 genes, with an average of 224 and median of 20 (Figure 1C). Classes from different sources had markedly different size distributions. Thus the 1883 classes derived from Wormbase phenotypes were biased towards small classes (median 5, average 34), while those derived from SPELL (1556 classes) were biased towards large classes (median 104, average 490; Figure 1C). Only 200 genes (<1%) were associated with a single category, in most cases (116 genes), the class "*C. elegans*-specific genes" [27], and therefore correspond to lineage-specific genes of unknown function. The median number of classes for each gene was 42 (Figure 1D), so the dataset has good coverage and the potential to provide insight into the functional relationships between the members of a gene list.

In its most basic implementation, a user inputs a list of genes of interest and sets analysis parameters. Users can choose between different methods to determine statistical significance; the default is by false discovery rate. Thresholds for *p* value and for the maximum reference class size can be defined. In the context of enrichment analysis, including classes that contain as many as 50% of all protein-coding genes makes relatively little sense; we generally, and arbitrarily set this parameter to 2000 genes (i.e. ca. 10% of all protein-coding genes). Enriched functional classes are returned in the form of a hyperlinked table that gives access to the underlying data. The results are also available as a downloadable text file.

Analysis of the YAAT dataset

To characterise the dataset and the tool, we investigated the relationship between class sizes and number of enriched gene classes, for the complete set and the classes derived from Wormbase phenotypes or SPELL. We observed a positive correlatory trend between the number of genes in a class and the number of enriched functional classes in all 3 sets (R^2=0.45 for an exponential fit for the entire set; Figure 2). There were, however, numerous classes that did not follow this trend; a number of extreme outliers are listed in Table 1. In all cases, they correspond to SPELL classes. The 4 classes that were associated with an unexpectedly low number of enriched categories are all very large classes (>5000 genes), derived from a single study that transcriptionally profiled different cell types [28]. The reasons for this are currently unclear especially since other classes that contain large numbers of genes returned a high number of enriched classes (e.g. SpellEC_1019, 6878 genes and SpellEC_1089 classes). The 5 classes that are associated with an unexpectedly high number of enriched categories were linked to the transcription response to biotic [29, 30] or abiotic (osmotic [31]) stress. Regarding the classes related to osmotic stress, while only 83 out of a total of 600 constituent genes (13.8%) were common to all 3 classes, more than half of the enriched functional categories were shared (269/528). This is a good demonstration of the power of the class-enrichment technique, and applied equally to classes that were not such outliers (results not shown).

We noticed that the vast majority of the commonly enriched classes associated with the response to osmotic stress (224/269) were derived from SPELL; only one was a phenotypic class (WBPhenotype:0000039). Since in the entire dataset, there were more WBPhenotype than SPELL classes, this suggested a bias that was investigated by comparing the enriched categories obtained with different types of input lists. To do so, and to avoid the confounding effect of class size, we took the 20 WBPhenotype classes constituted of 200-300 genes, and selected at random 20 (out of the >100) SPELL classes of the same size (Figure 3A), and ran them through YAAT. There was a striking difference in enriched gene classes for the 2 types of gene lists. While groups of genes defined by transcriptome studies (i.e. SPELL classes) were found to be enriched almost exclusively for other SPELL classes, the groups defined on the basis of a phenotype were enriched for a far more diverse range of categories (Figure 3B), in proportions that approached their overall distribution (Figure 1B). The consistent bias that was observed is a caveat for the interpretation of this type of enrichment analysis and is discussed further below. These results also support the idea that having the broadest range of types of functional classes is important to avoid biased functional enrichment analyses.

Enriched class clustering in YAAT to investigate functional connections

To investigate the utility of the tool, as a test data set, we used a list of 146 *C. elegans* protein-coding genes that are up-regulated upon infection with the fungal pathogen *D. coniospora* [15]. As expected from the global analysis of the dataset, there was a very strong bias of enriched classes towards those derived from transcriptome analyses. All of these “Dc_Up” genes were found in at least one functional class.

Consistent with prior knowledge, the analysis highlighted the enrichment of genes in several classes, including those regulated by osmotic stress [14, 31], by *nhr-25* [32] and those down-regulated by diverse bacterial intestinal pathogens, including *Serratia marcescens* and *Enterococcus faecalis* [15].

To provide a graphical representation of the relationship between the members of the enriched classes, YAAT returns the results as a hierarchical cluster plot. With the Dc_Up set of query genes, the different groups clustered independently. For example, the genes regulated upon *D. coniospora* infection and by *osm-7*, *osm-8* and *osm-11* co-clustered, as did the genes regulated by *nhr-25* and *acs-3* (Figure 4A). The results with this test dataset suggest that the tool has the potential to identify functional connections between different regulatory mechanisms.

Phylogenetic profiling in YAAT to investigate functional connections

Genes that function together are often observed to co-evolve [33]. To provide a comparative measure of gene conservation, we include in the table of results a comparison between the average level of conservation across a broad range of 56 eukaryotic species (EukaProf; vertebrates, invertebrates, fungi, plants and protists; see Materials and Methods) for the entire list of genes in an enriched class and for the genes that are shared between that list and the genes input as a query (Figure 4B). The figures for overall conservation for the members of the different classes varied widely. Very surprisingly, for an average level of conservation between 30% and 85%, there was a roughly similar number of classes (42 +/- 8). On the other hand, when we looked at the overall conservation for 2 of the main component datasets in YAAT, namely expression clusters from SPELL and phenotypic data from Wormbase, strikingly, we found a strong bias towards conservation among the classes associated with a phenotype, but not among those derived from expression studies (Figure 5). Not surprisingly, the lowest score was associated with the class of lineage-specific genes [27]. It was notable that 7 of 20 classes with the least conserved genes were lists of infection-regulated genes. At the other end of the scale, among the larger classes (>200 genes) with high scores, apart from the classes transposed from *Drosophila*, which were by their very nature conserved, those related to fundamental cellular processes such as DNA and protein synthesis were prominent. As discussed below, the observed skew in the conservation of genes found in the different types of functional classes could lead to biases when analysing datasets.

Among the functional classes that were enriched for our test set of 146 *D. coniospora* up-regulated (Dc_Up) genes, the least conserved was of those 72 genes in the category "*hda-1*(RNAi)_upregulated" (score 0.16; Figure 4B). The 14 overlapping genes included *cnc-4*, that encodes a nematode-specific antimicrobial peptide [34], and in addition the uncharacterised cluster of paralogous genes F53A9.1, F53A9.6, F53A9.7, F53A9.8 that all encode small (<90 residue) histidine and glycine-rich proteins. We speculate that these are also direct effectors of antifungal immunity. The histone deacetylase encoded by *hda-1* was found to be a Nipi gene, required for the expression of *nlp-29* upon *D. coniospora* infection ([17] and unpublished results). Our analysis suggests that it potentially has a broader role in innate immunity, co-regulating multiple defence genes.

In addition to giving a global measure of conservation, to provide insight into the pattern of conservation of the members of a gene list, YAAT offers the possibility of performing phylogenetic profiling. This can be a broad analysis against the 56 eukaryotic species mentioned above (EukaProf), or more targeted ones using data from 72 nematode species (NemaProf). As an illustration, we took the list of 278 Nipi genes required for the expression of an antimicrobial peptide gene reporter [17] as well as the list of 146 Dc_Up genes. We found exactly the same trend as mentioned above: almost all the phenotypic dataset genes were conserved across different taxa (Figure 6A), whereas most of the transcriptionally regulated genes were either found only in nematode species or just in *C. elegans* (Figure 6B). This suggests that innate immune effectors evolve faster than network genes. When we inspected the clusters of phylogenetic profiles in EukaProf of the 278 Nipi genes (Figure 6A), we found one cluster of 46 genes that is found in invertebrate and vertebrate species but not others. The list includes *sta-2* and *tir-1* (encoding a STAT-like transcription factor and a TIR domain adaptor protein, respectively) that are both central to the control of AMP gene expression [35, 36]. Submitting these 46 genes to WormNet, a phenotype-centric tool that represents known interactions between genes in a list [12], revealed a significant overall connectivity and identified a number of potential protein complexes (Figure 7). The largest includes LIN-40, DCP-66 and AKIR-1. Through a biochemical approach, we have recently confirmed the existence of a physical complex containing these 3 proteins, together with HDA-1 (Polanowska at al., in preparation). YAAT therefore can contribute to the definition of signaling modules from lists of genes on the basis of their shared function and evolutionary trajectory.

Comparison of with another fungus upregulated datasets

In order to determine whether the pattern of class enrichment and conservation observed with the set of *D. coniospora* up-regulated genes was characteristic for the epidermal anti-fungal response, we conducted transcriptional profiling with a second pathogen that infects the nematode epidermis, *H. sphaerosporum* [13]. Using the same analytical method as had previously been applied to the *D. coniospora* dataset [15], we defined a set of 202 infection upregulated genes (Hapto_Up). Functional enrichment analysis returned 31 classes (p-val < 0.0001; Table 2), compared to 170 for Dc_Up. Very surprisingly, unlike DC_Up where all the genes were found in at least one functional class, in Hapto_Up, almost 30% (57 genes) were foundlings, not present in any enriched functional class. For the remaining genes, clustering of the corresponding functional classes showed two separate groups, one largely of pathogen-related classes (#1-11; Table 2) and the other, principally development/metabolism-related classes (Figure 8). These 2 groups included classes of genes enriched for expression in the epidermis (SpellEC_807), and intestine (SpellEC_331), respectively. Only 7 functional classes were found in common for the functional enrichment analysis of the genes induced upon infection by the two fungal pathogens (Table 2). These were all in the group of pathogen-related classes, including the class of genes expressed in the epidermis (SpellEC_807) mentioned above. This latter class was not unexpected since the majority of the known defence genes are expressed in the epidermis upon fungal infection. The group also included 3 classes of genes downregulated by different intestinal bacterial pathogens. This inverse relationship has been previously noted [15] and may reflect cross-tissue coordination of gene expression. Interestingly, there was a significant overlap for both Dc_Up and Hapto_Up genes with genes induced by the bacterium *Staphylococcus aureus* in an *hlh-30(tm1978)* mutant background. There were only a small number of genes shared by all 3 datasets (Figure 8B), and 4 of these 5 corresponded to AMP genes (*cnc-4, cnc-8, fip-6, nlp-34*). The analysis also revealed a possible role for EGF signalling that merits investigation. The presence of the class "*alg-1(gk214)*_upregulated" may be an artefact since *alg-1* mutants are developmentally delayed. In such mutants, genes that have a dynamic expression level during development can appear to be differentially regulated if strains are not strictly synchronised [37, 38].

Compared to the Dc_Up genes that, as described above were poorly conserved, with an average conservation score of 0.2 for enriched classes, and for which the class with the most conserved genes had a score of 0.66, the Hapto_Up enriched classes were remarkably conserved, with an average score for all enriched classes of 0.65, and with a maximum conservation score for the enriched class "pleiotropic defects severe early emb" of 0.98. As reflected by the EukaProf profile (Figure 8C), more than 40% of the Hapto_Up genes are conserved across diverse eukaryotic species. Looking at NemaProf profile, almost 65 % of the genes are conserved across all nematodes; the rest of the genes are conserved only in *Caenorhabditis* species (Figure 8D). Thus, despite their similar level of pathogenicity and routes of infection, the two fungal species *D. coniospora* and *H. sphaerosporum* provoke the up-regulation of very distinct sets of genes, not only in terms of the exact genes involved, but also their potential physiological roles (as defined by class enrichment) and pattern of conservation (Table 3). The high number of foundling genes in the Hapto_Up set suggests that the host immune response to this pathogen involves unexplored molecular mechanisms. One exception to this concerns the genes encoding antimicrobial peptides (AMPs). In addition to the 4 shared with the Dc_Up genes, there were a further 11 AMP genes in the Hapto_Up set, once again reinforcing the notion that the production of AMPs is a fundamental means of defence against fungal infection in *C. elegans* [15].

Comparing YAAT with WormExp

To conduct a comparison between YAAT and WormExp, we first defined a group of pathogen-specific datasets (see Materials and Methods). Even though the overall number of functional classes in WormExp (1960) is less than half the number of functional classes in YAAT (4700), WormExp gave more enriched classes for all the test datasets. Unexpectedly, for the different datasets, WormExp identified between 4% (83) and 50% (988) of the 1960 functional classes as being enriched, higher by 50% in absolute terms than obtained with YAAT. To determine whether this disparity reflected a more general trend, we submitted random gene datasets of different sizes (50, 100, 500, 1000 and 2000) to both tools. WormExp consistently gave enriched classes for random datasets of at least 500 genes. The number of enriched classes then increased with increasing dataset size. With YAAT, however, no functional classes were found enriched for any of the random datasets (p-val < 0.0001; Table 4). This suggests that YATT provides enriched classes with a higher specificity.

DISCUSSION

Various stand-alone or web-based functional enrichment analysis tools have been developed over the last 15 years. Among them, WormExp [5] is the only *C. elegans*-specific tool that we are aware of. Even though WormExp harbours more information than generic tools, it is currently referenced to an out-dated release of Wormbase (WS235; data from 2012). The results of our tests indicated that functional enrichment analyses should be based on comparisons with the broadest possible collection of functional classes to avoid the biases inherent to each type of gene set. WormExp includes only transcriptome data, a subset of that publicly available for *C. elegans*. We have overcome these shortcomings with YAAT, a tool that includes a broad base of functional classes. Indeed, by collecting transcriptomic, functional and phenotypic data from the literature, Wormbase, KEGG and other resources, we amassed more than 4700 functional classes, covering essentially all protein coding genes in Wormbase, currently referenced to the 2016 release WS252. Although generic tools are valuable, just as specific model organism databases are essential for research [4], so too do species-specific analytical tools present advantages. In the case of *C. elegans*, most particularly, they allow leveraging of the most up-to-date expert-curated information.

Importantly, since data evolves, from the outset, we included a pipeline to update data from the respective resources automatically. For the static data such as modENCODE assigned transcription factor targets, the most accurate updating would involve reassigning genes to ChIPseq peaks on the basis of revised gene structure predictions. Given the practical difficulties, in these cases we opted for an intermediate strategy, tracking changes in gene identity over successive Wormbase releases using an established methodology [15].

Rigorous statistical enrichment analysis requires knowledge of the gene "universe" sampled when a class was defined. It matters whether a class was defined after sampling all genes or only a subset. For example, in our previous transcriptome studies to determine the response of *C. elegans* to infection, we have used cDNA [39], long oligonucleotide [40] and tiling microarrays [15]. These theoretically cover between 35% and close to 100% of the predicted protein coding genes. In reality, because of the sensitivity of the techniques, these figures are upper limits. Similarly, for analyses by RNAseq, in principal all genes are assayed, but in fact the number is determined by the depth of sequencing [41]. For other types of class, determining the universe is equally problematic, be it because of the high rate of false negatives in RNAi screens, or lack of relevant information for genetic screens. Indeed, it is very rare for the sampled universe to be reported, and even when it is, this information is generally not available in a machine-readable standardised format. In common with most current functional enrichment analysis tools, we have assumed that the universe for each set is the entire complement of protein coding genes. This will introduce inevitable inaccuracy in the calculation of the statistical significance of any given enrichment, especially if the universe of a class of interest is small. Different definitions of the gene universe likely contribute to the observed disparity of results obtained for random gene sets between YAAT and WormExp. The latter includes non-coding genes in its gene universe, which expands it to close to 28,000 genes. In WormExp, any query list comprised exclusively of coding genes will be inherently enriched in a dataset restricted to coding genes due to the reduced size of the real (<20,000) compared to the imposed universe. For this reason, users of WormExp and YAAT are recommended to crosscheck the data sources for classes of interest.

YAAT also goes beyond the functionality of WormExp since it provides class clustering, allowing the inference of functional links between groups of genes. Further, we have included the option to perform phylogenetic profiling, a complementary method for establishing groups of co-evolving, and hence functionally-related genes. Given the increasing interest in translating research from *C. elegans* to parasitic worms (e.g. [42]), among the options we provide searches restricted to nematodes.

In the course of benchmarking YAAT, we constituted a list of genes described as differentially-regulated in at least 2 different studies. Even though the constituent datasets were limited to those with less than 500 genes, one very striking observation was that the complete list of genes represented a quarter of all protein-coding genes (5577). There are several possible explanations for this high number. The most banal is that it reflects an inherent limitation to the definition of differential expression, and that many of the genes are in fact false-positives. On the other hand, it is also possible that different pathogens interfere with different fundamental physiological processes, hence giving rise to the observed diversity of differentially-regulated genes. One argument against this latter explanation is the fact that overall these differentially-regulated genes were poorly conserved, suggesting that they do not play fundamental roles in basic cellular physiology.

This lack of conservation contrasted with the much broader conservation of genes associated with a phenotype. This is consistent with previous observations that genes that form part of signalling networks (and that give a phenotype when mutated) are generally more conserved than their transcriptional targets, that rarely give phenotypes when mutated because of functional redundancy. Further studies are required to establish the generality of these observations in *C. elegans*.

Class	Description		Number Of Genes	Count Class
SpellEC_670	osm-7_regulated	WBPaper00035873	254	351
SpellEC_673	osmotically_regulated	WBPaper00035873	285	399
SpellEC_1309	N.parisii_64h_downregulated	WBPaper00045401	348	384
SpellEC_711	affected_by_RN6390 [*S. aureus*]	WBPaper00036464	378	365
SpellEC_672	osm-11_regulated	WBPaper00035873	386	399
SpellEC_763	AVE-neuron_expressed	WBPaper00037950	5457	14
SpellEC_749	A-class-motor-neurons_expressed	WBPaper00037950	8138	38
SpellEC_755	all-neurons_expressed	WBPaper00037950	8665	43
SpellEC_779	coelomocytes_expressed	WBPaper00037950	8673	37

Table 1. Classes with atypical numbers of enriched categories. Small classes with a high number of enriched categories (in bold) and large classes with a small number of enriched categories were defined arbitrarily (see Figure 2).

	Class	Overlap	Tot	PVal	fdr	Condition	Paper/source
1	**SpellEC_807**	**30**	**1242**	**7.33E-06**	**1.97E-05**	**hypodermis_larva_enriched**	**WBPaper00037950**
2	**SpellEC_1251**	**26**	**675**	**4.72E-09**	**2.53E-08**	**S.aureus-induced_hlh-30(tm1978)**	**WBPaper00045314**
3	**SpellEC_943**	**31**	**968**	**1.10E-08**	**4.23E-08**	**alg-1(gk214)_upregulated**	**WBPaper00040823**
4	SpellEC_918	19	498	7.37E-07	3.24E-06	H2S_24hr_upregulated	WBPaper00040285
5	**JEEC_112**	**12**	**230**	**3.92E-06**	**1.47E-05**	**Upregulated gene in let-23(sa62) animals compared with eor-1(cs28) animals (EGF/EOR-1-upregulated genes)**	**21673654**
6	SpellEC_891	4	12	4.63E-06	2.23E-06	hcf-1nc_sir-2.1down_daf-2down	WBPaper00040184
7	SpellEC_1502	5	29	9.58E-06	1.01E-05	[cgc5767]:cluster_16	WBPaper00005767
8	SpellEC_1288	11	198	5.50E-06	1.94E-05	prg-1_downregulated_L4	WBPaper00045316
9	**SpellEC_866**	**31**	**1106**	**2.20E-07**	**7.00E-07**	**S.marcescens_24hr_downregulated_RNAseq**	**WBPaper00038438**
10	**SpellEC_862**	**25**	**775**	**2.91E-07**	**1.24E-06**	**P.lumniescens_24hr_downregulated_RNAseq**	**WBPaper00038438**
11	**SpellEC_856**	**30**	**719**	**3.92E-11**	**2.85E-10**	**E.faecalis_24hr_downregulated_RNAseq**	**WBPaper00038438**
12	SpellEC_598	48	1832	4.68E-10	5.02E-10	L3_enriched	WBPaper00032528
13	SpellEC_331	44	1932	1.85E-07	1.83E-07	intestine_enriched	WBPaper00026980
14	SpellEC_67	45	1715	1.80E-09	2.32E-09	cholesterol_10-9M_regulated	WBPaper00005124
15	JEEC_251	32	353	1.98E-21	1.89E-20	Down-un stressed aak-2 vs N2	21303547
16	WBPhenotype:0000031	41	1788	4.47E-07	5.73E-07	slow growth	WBPhenotype
17	JEEC_243	14	162	1.04E-09	4.58E-09	Down unstressed aak-2 vs N2	21303547
18	JEEC_246	14	128	4.43E-11	2.68E-10	Down-stressed wild type and stressed aak-2	21303547
19	JEEC_241	15	162	9.72E-11	5.02E-10	Down paraquat-stressedN2 vs N2	21303547
20	JEEC_245	15	136	7.76E-12	5.67E-11	Down stressed aak-2 vs N2	21303547
21	SpellEC_657	13	289	7.80E-06	3.08E-05	daf-16(RNAi)_upregulated	WBPaper00035479
22	cel03010	19	121	1.35E-17	7.68E-17	Ribosome	KEGG
23	SpellEC_583	21	330	2.13E-11	2.05E-10	differentially_expressed_with_age_medoid_7	WBPaper00032165
24	SpellEC_1606	17	277	3.26E-09	1.81E-08	[cgc5767]:expression_class_E_pi(66_min)	WBPaper00005767
25	JEEC_41	24	443	2.15E-11	2.05E-10	Protein expression	11557892
26	JEEC_161	7	55	1.27E-06	2.23E-06	Genes required for paraquat triggered hsp-6::gfp induction	23516373
27	SpellEC_104	10	85	1.39E-08	3.99E-08	cluster_10	WBPaper00025032
28	WBPhenotype:0000270	13	118	2.03E-10	8.59E-10	pleiotropic defects severe early emb	WBPhenotype
29	WBPhenotype:0000055	12	199	8.64E-07	3.11E-06	early larval arrest	WBPhenotype
30	JEEC_123	19	374	8.45E-09	4.32E-08	gene inactivations that stimulate microbial aversion behavior	22500807
31	WBPhenotype:0000402	19	438	1.04E-07	5.40E-07	avoids bacterial lawn	WBPhenotype

Table 2. Raw output from YAAT obtained upon analysis of the Hapto-Up gene set. Functional classes that are enriched for both Dc_UP and Haptp_Up gene sets are highlighted in bold. The classes are listed in the same order as in Figure 7A. Classes 1-11 cluster together. Tot: total number of genes in a class.

	D. coniospora	*H. sphaerosporum*
UP regulated genes	146	203
Genes not in an enriched class	0	57
Enriched classes	170	31
Average conservation score of enriched classes	0.2	0.65
Conservation score among eukaryotes	0.17	0.43
Conservation score among nematodes	0.43	0.61

Table 3: Comparison between *D. coniospora* and *H. sphaerosporum* up-regulated gene sets.

Class (genes)	**Replicates**	**WormExp (1960)**	**YAAT (~4200)**
Haptocilium Up (203)	NA	83	56
Drechmeria Up (147)	NA	396	213
Infection Set	NA	947	603
Random (50)	5x (mean)	0	0
Random (100)	5x (mean)	0.2	0
Random (500)	5x (mean)	13.4	0
Random (1000)	5x (mean)	69.8	0
Random (2000)	5x (mean)	188.6	0

Table 4. Comparison of functional enrichment analysis by WormExp and YAAT on various real and random datasets.

REFERENCES

1. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1-13.
2. Wadi, L., Meyer, M., Weiser, J., Stein, L., and Reimand, J. (2016). Impact of knowledge accumulation on pathway enrichment analysis. bioRxiv, 049288.
3. Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res.
4. Oliver, S.G., Lock, A., Harris, M.A., Nurse, P., and Wood, V. (2016). Model organism databases: essential resources that need the support of both funders and users. BMC Biol *14*, 49.
5. Yang, W., Dierking, K., and Schulenburg, H. (2016). WormExp: a web-based application for a Caenorhabditis elegans-specific gene expression enrichment analysis. Bioinformatics *32*, 943-945.
6. Stoeckius, M., Grun, D., Kirchner, M., Ayoub, S., Torti, F., Piano, F., Herzog, M., Selbach, M., and Rajewsky, N. (2014). Global characterization of the oocyte-to-embryo transition in Caenorhabditis elegans uncovers a novel mRNA clearance mechanism. Embo J *33*, 1751-1766.
7. Grun, D., Kirchner, M., Thierfelder, N., Stoeckius, M., Selbach, M., and Rajewsky, N. (2014). Conservation of mRNA and protein expression during development of *C. elegans*. Cell reports *6*, 565-577.
8. Sleumer, M.C., Bilenky, M., He, A., Robertson, G., Thiessen, N., and Jones, S.J. (2009). Caenorhabditis elegans cisRED: a catalogue of conserved genomic elements. Nucleic Acids Res *37*, 1323-1334.
9. Araya, C.L., Kawli, T., Kundaje, A., Jiang, L., Wu, B., Vafeados, D., Terrell, R., Weissdepp, P., Gevirtzman, L., Mace, D., et al. (2014). Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. Nature *512*, 400-405.
10. Hutter, H., Ng, M.P., and Chen, N. (2009). GExplore: a web server for integrated queries of protein domains, gene expression and mutant phenotypes. Bmc Genomics *10*, 529.
11. Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K., et al. (2014). WormBase 2014: new views of curated biology. Nucleic Acids Res *42*, D789-793.
12. Cho, A., Shin, J., Hwang, S., Kim, C., Shim, H., Kim, H., Kim, H., and Lee, I. (2014). WormNet v3: a network-assisted hypothesis-generating server for *Caenorhabditis elegans*. Nucleic Acids Res *42*, W76-82.
13. Labed, S., and Pujol, N. (2011). *Caenorhabditis elegans* Antifungal Defense Mechanisms. The Journal of Invasive Fungal Infection *5*, 110-117.
14. Pujol, N., Zugasti, O., Wong, D., Couillault, C., Kurz, C.L., Schulenburg, H., and Ewbank, J.J. (2008). Anti-fungal innate immunity in *C. elegans* is enhanced by evolutionary diversification of antimicrobial peptides. PLoS Pathog *4*, e1000105.
15. Engelmann, I., Griffon, A., Tichit, L., Montanana-Sanchis, F., Wang, G., Reinke, V., Waterston, R.H., Hillier, L.W., and Ewbank, J.J. (2011). A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans*. PLoS One *6*, e19055.
16. Lebrigand, K., He, L.D., Thakur, N., Arguel, M.J., Polanowska, J., Henrissat, B., Record, E., Magdelenat, G., Barbe, V., Raffaele, S., et al. (2016). Comparative Genomic Analysis of *Drechmeria coniospora* Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes. PLoS Genet *12*, e1006017.

17. Zugasti, O., Thakur, N., Belougne, J., Squiban, B., Kurz, C.L., Soule, J., Omi, S., Tichit, L., Pujol, N., and Ewbank, J.J. (2016). A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes. BMC Biol *14*, 35.

18. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics *6*, 65-70.

19. Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika *75*, 800-802.

20. Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika *75*, 383-386.

21. Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics *29*, 1165-1188.

22. Powell, J.R., and Ausubel, F.M. (2008). Models of *Caenorhabditis elegans* Infection by Bacterial and Fungal Pathogens. In Methods Mol Biol, Volume 415, J. Ewbank and E. Vivier, eds. (Humana Press), pp. 403-427.

23. Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L.W., Sasidharan, R., Reinke, V., Waterston, R.H., and Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. Bmc Genomics *11*, 383.

24. Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. Genome Res *19*, 657-666.

25. Tanabe, M., and Kanehisa, M. (2012). Using the KEGG database resource. Curr Protoc Bioinformatics *Chapter 1*, Unit1 12.

26. Hibbs, M.A., Hess, D.C., Myers, C.L., Huttenhower, C., Li, K., and Troyanskaya, O.G. (2007). Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics *23*, 2692-2699.

27. Zhou, K., Huang, B., Zou, M., Lu, D., He, S., and Wang, G. (2015). Genome-wide identification of lineage-specific genes within *Caenorhabditis elegans*. Genomics *106*, 242-248.

28. Spencer, W.C., Zeller, G., Watson, J.D., Henz, S.R., Watkins, K.L., McWhirter, R.D., Petersen, S., Sreedharan, V.T., Widmer, C., Jo, J., et al. (2011). A spatial and temporal map of C. elegans gene expression. Genome Res *21*, 325-341.

29. Bakowski, M.A., Desjardins, C.A., Smelkinson, M.G., Dunbar, T.A., Lopez-Moyado, I.F., Rifkin, S.A., Cuomo, C.A., and Troemel, E.R. (2014). Ubiquitin-mediated response to microsporidia and virus infection in *C. elegans*. PLoS Pathog *10*, e1004200.

30. Irazoqui, J.E., Troemel, E.R., Feinbaum, R.L., Luhachack, L.G., Cezairliyan, B.O., and Ausubel, F.M. (2010). Distinct pathogenesis and host responses during infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. PLoS Pathog *6*, e1000982.

31. Rohlfing, A.K., Miteva, Y., Hannenhalli, S., and Lamitina, T. (2010). Genetic and physiological activation of osmosensitive gene expression mimics transcriptional signatures of pathogen infection in *C. elegans*. PLoS One *5*, e9010.

32. Ward, J.D., Mullaney, B., Schiller, B.J., He le, D., Petnic, S.E., Couillault, C., Pujol, N., Bernal, T.U., Van Gilst, M.R., Ashrafi, K., et al. (2014). Defects in the *C. elegans* acyl-CoA Synthase, *acs-3*, and Nuclear Hormone Receptor, *nhr-25*, Cause Sensitivity to Distinct, but Overlapping Stresses. PLoS One *9*, e92552.

33. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285-4288.

34. Pujol, N., Davis, P.A., and Ewbank, J.J. (2012). The Origin and Function of Anti-Fungal Peptides in *C. elegans*: Open Questions. Front Immunol *3*, 237.

35. Couillault, C., Pujol, N., Reboul, J., Sabatier, L., Guichou, J.F., Kohara, Y., and Ewbank, J.J. (2004). TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human SARM. Nat Immunol *5*, 488-494.
36. Dierking, K., Polanowska, J., Omi, S., Engelmann, I., Gut, M., Lembo, F., Ewbank, J.J., and Pujol, N. (2011). Unusual regulation of a STAT protein by an SLC6 family transporter in *C. elegans* epidermal innate immunity. Cell Host Microbe *9*, 425-435.
37. George-Raizen, J.B., Shockley, K.R., Trojanowski, N.F., Lamb, A.L., and Raizen, D.M. (2014). Dynamically-expressed prion-like proteins form a cuticle in the pharynx of *Caenorhabditis elegans*. Biol Open *3*, 1139-1149.
38. Francesconi, M., and Lehner, B. (2014). The effects of genetic variation on gene expression dynamics during development. Nature *505*, 208-211.
39. Mallo, G.V., Kurz, C.L., Couillault, C., Pujol, N., Granjeaud, S., Kohara, Y., and Ewbank, J.J. (2002). Inducible antibacterial defense system in *C. elegans*. Curr Biol *12*, 1209-1214.
40. Wong, D., Bazopoulou, D., Pujol, N., Tavernarakis, N., and Ewbank, J.J. (2007). Genome-wide investigation reveals pathogen-specific and shared signatures in the response of *Caenorhabditis elegans* to infection. Genome Biol *8*, R194.
41. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. Genome Res *21*, 2213-2223.
42. Keiser, J. (2015). Is *Caenorhabditis elegans* the Magic Bullet for Anthelminthic Drug Discovery? Trends Parasitol *31*, 455-456.

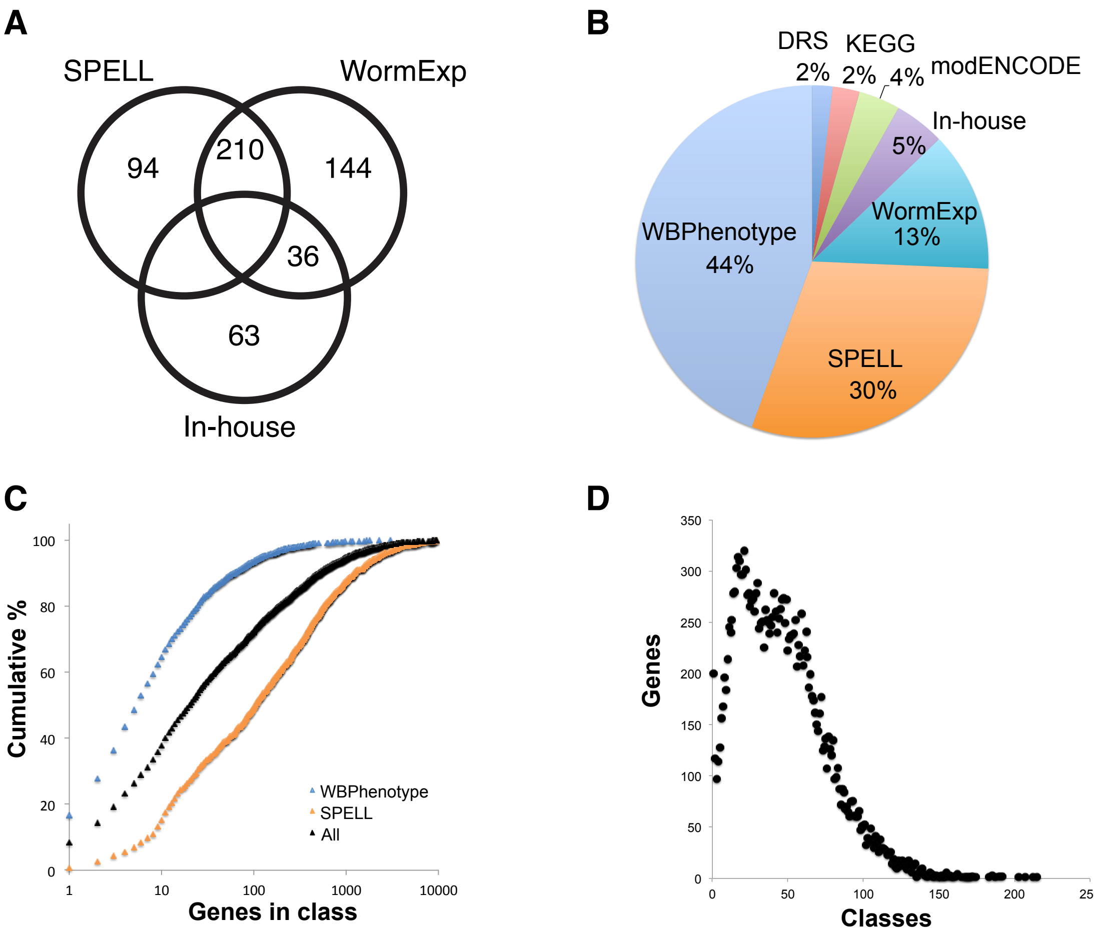


Figure 1. Sources of and types of data in YAAT. A. Venn diagram showing the data coverage by SPELL, WormExp and our original in-house collection. The figures indicate the number of articles from which transcriptome data was extracted. B. Graph showing the distribution of the major types of data in YAAT. DRS: phenotypes transposed from Drosophila RNAi screens. C. Graph showing the distribution of the class sizes, for all classes (green; a single class with 10,731 genes is omitted), and those from SPELL (orange) and Wormbase phenotypes (blue). D. Graph showing the distribution of the number of classes associated with each gene.

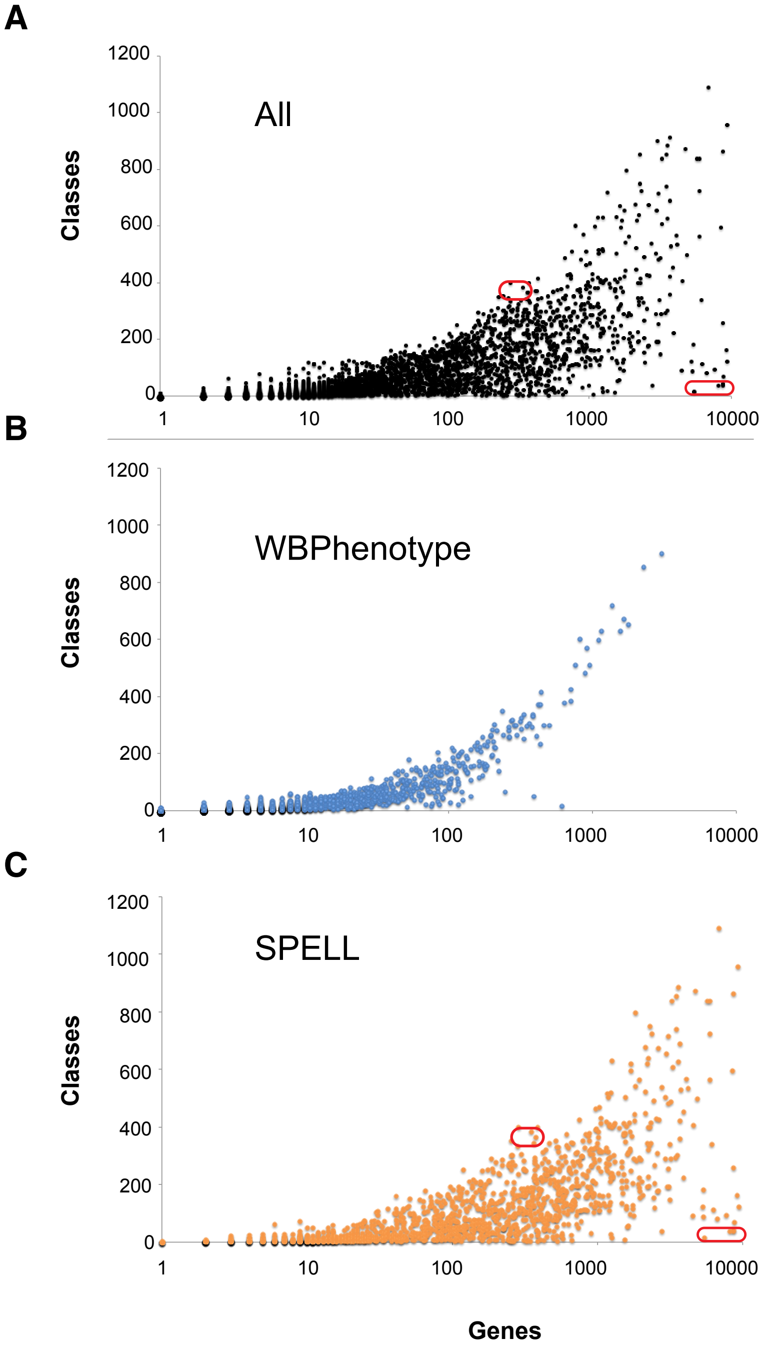


Figure 2. Graphs showing the distribution of the number of enriched classes returned (Y-axis) as a function of the size of the class used as a query (X-axis), for the entire set of classes (A), for classes derived from Wormbase phenotypes (B) and from SPELL (C).

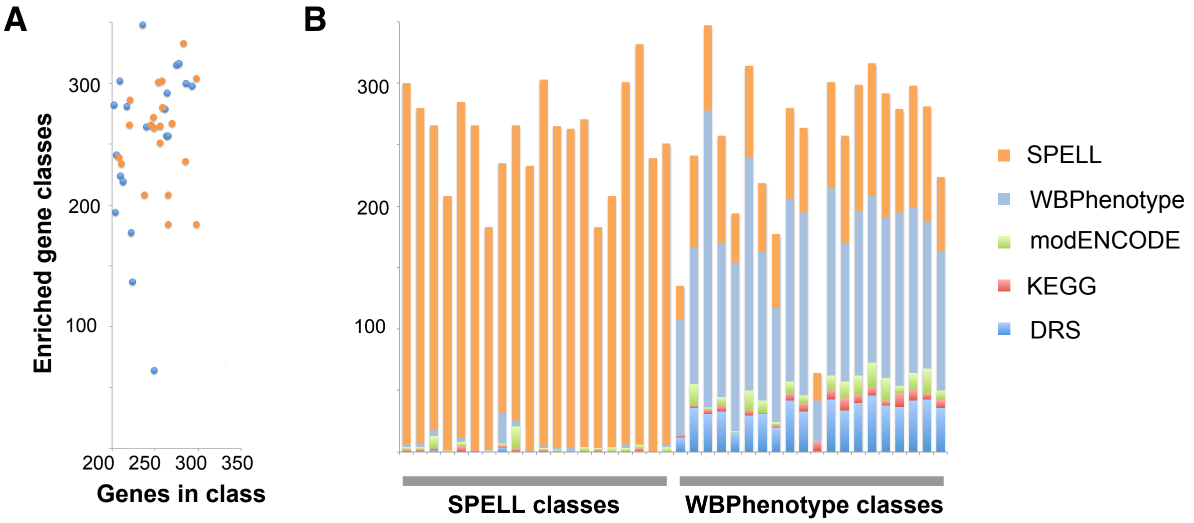
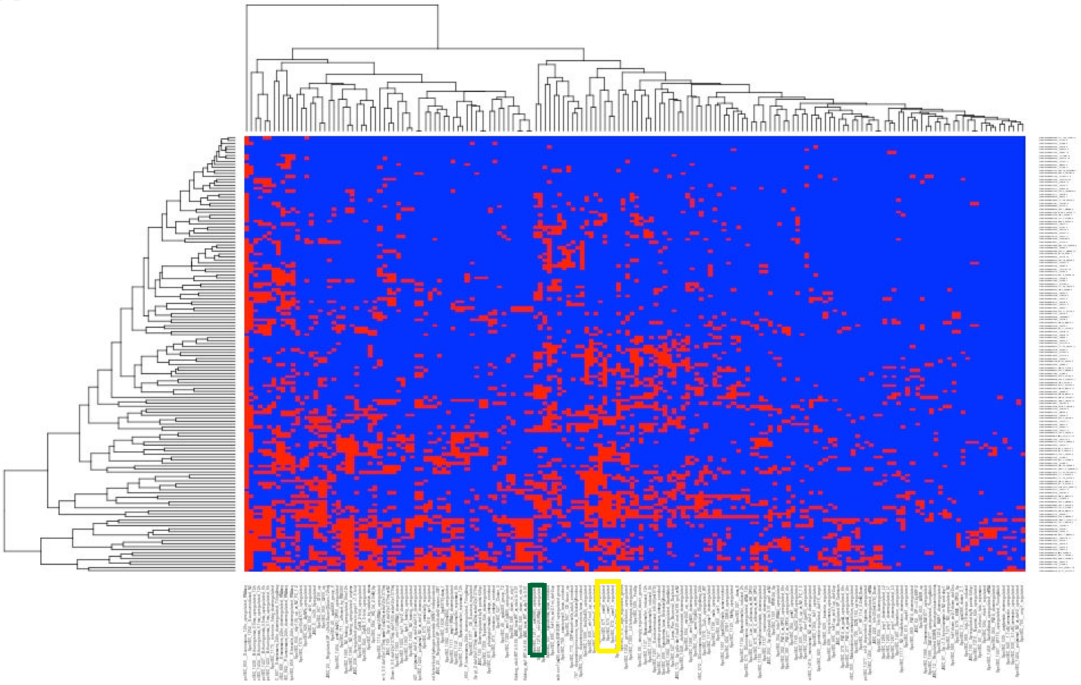


Figure 3. Different datasets are preferentially enriched for certain gene classes. A. Characteristics of the 20 sets of genes from either the SPELL or WBPhenotype categories. B. Graph showing the distribution of types of enriched gene classes for the same 20 sets of 200-300 genes from either the SPELL or WBPhenotype categories.

A



B

Class	Order	Overlap	Tot	PVal	fdr	Class	Source	Normalised score	Class conservation
SpellEC_304	117	14	72	1.30E-16	7.20E-17	hda-1(RNAi)_upregulated	WBPaper00025192	0.03	0.16
SpellEC_1117	144	14	72	1.30E-16	7.20E-17	N2_VirusInfection_down-regulated	WBPaper00042574	0.09	0.19
JEEC_112	69	38	230	4.90E-41	2.00E-40	Upregulated gene in let-23(sa62) animals compared with eor-1(cs28) animals (EGF/EOR-1-upregulated genes)	21673654	0.09	0.24
JEEC_172	31	21	749	1.20E-07	1.00E-07	Down lt_0.5 daf-2(e1370ts); swsn-1(os22ts) vs daf-2(e1370ts)	23604319	0.07	0.32
JEEC_12	151	5	38	7.90E-06	1.80E-06	Regulated DOWN_various genes,xenobiotics	16168746	0.43	0.59
SpellEC_1377	136	5	12	1.50E-08	1.10E-09	elt-2_dependent_hypoxia_up-regulated	WBPaper00045842	0.68	0.62
SpellEC_559	15	35	1695	1.20E-08	3.60E-09	slr-2_regulated	WBPaper00031832	0.32	0.63

Figure 4. A. Clustering of enriched functional classes for D. coniospora up-regulated genes. Each row is a gene, each column is a functional class. The 2 classes of genes regulated by nhr-25 and acs-3, and the 3 classes of genes regulated by osm-7, osm-8 and osm-11 are highlighted with the green and yellow boxes, respectively. B. Extract of results table, showing the top and bottom when ranked by class conservation. The columns are, Class (name in database), Order (column number in clustering graph), Overlap (number of genes shared between input gene set and class of interest), Tot (total number of genes in class of interest), PVal (p value) fdr (false discovery rate) Class (description) Source (Wormbase or Pubmed reference), Normalised score (measure of conservation within genes shared with class), Class conservation (measure of conservation for all genes in class). The thresholds were p value < 0.0001 and < 2000 genes/class.

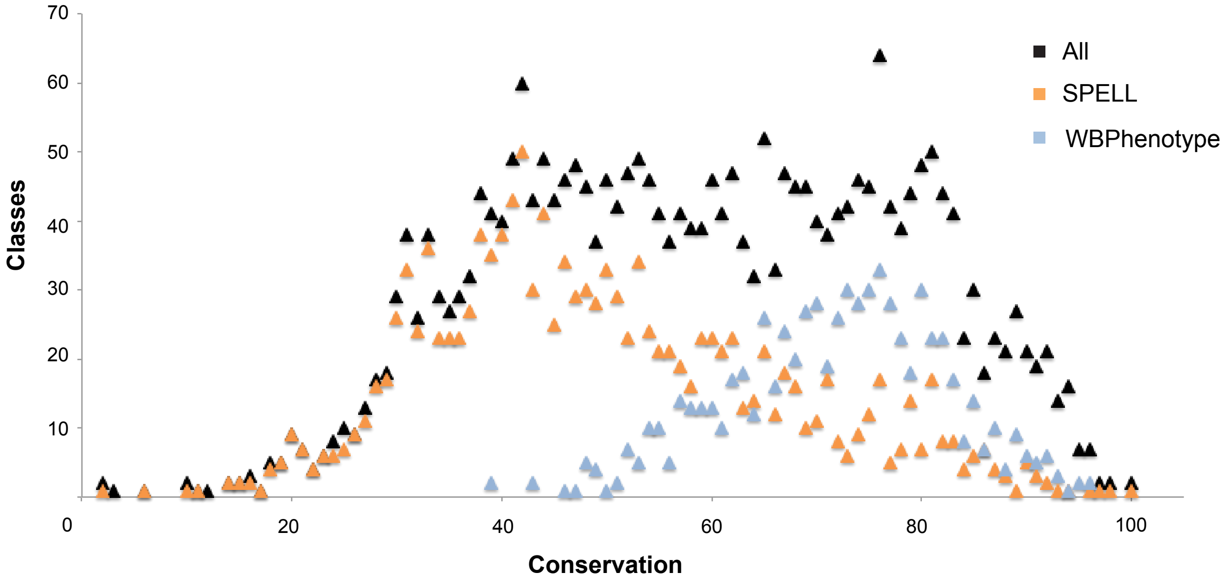


Figure 5. Average conservation score across 56 eukaryotic species for different types of gene classes, X-axis; shows the average conservation score, Y-axis; number of classes . The conservation score for all the gene classes (black) compared to gene classes from either the SPELL (orange) or WBPhenotype (blue) categories. Only classes with a minimum of 10 genes were included.

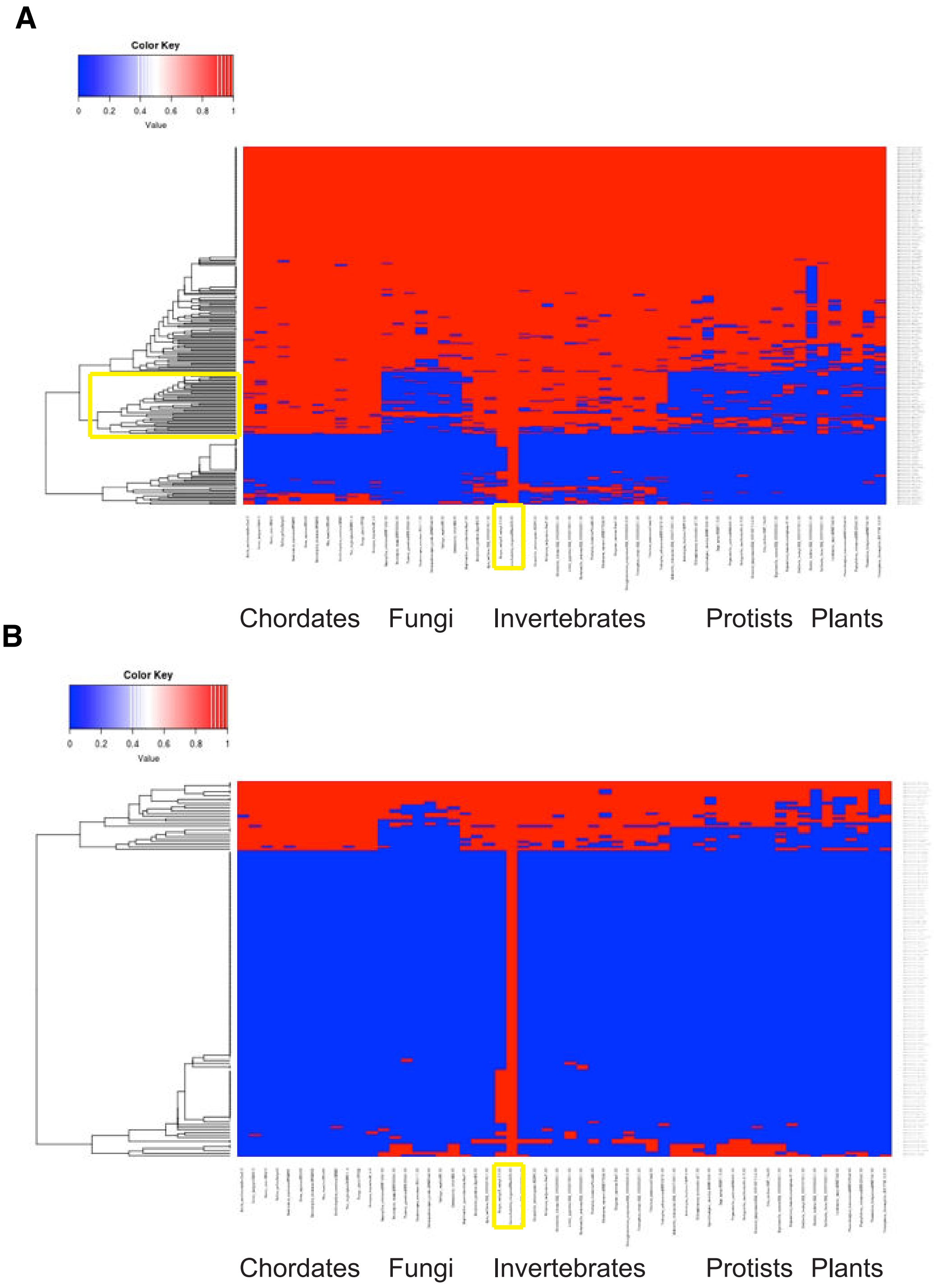


Figure 6. Distinct phylogenetic profiles for regulatory and effector genes. Profiles across 56 eukaryotic species for 278 Nipi genes identified in an RNAi screen (A) and for 146 genes upregulated upon D. coniospora infection (B). Red and blue indicate the presence and absence of an orthologues, respectively. Each row is a gene, columns represents species (ordered according to the established evolutionary relationships among each group of species). The group of 46 conserved genes mentioned in the text, and nematode species are highlighted by the yellow boxes on the left and at the bottom, respectively. In panel A and B, row represent genes and column represent species.

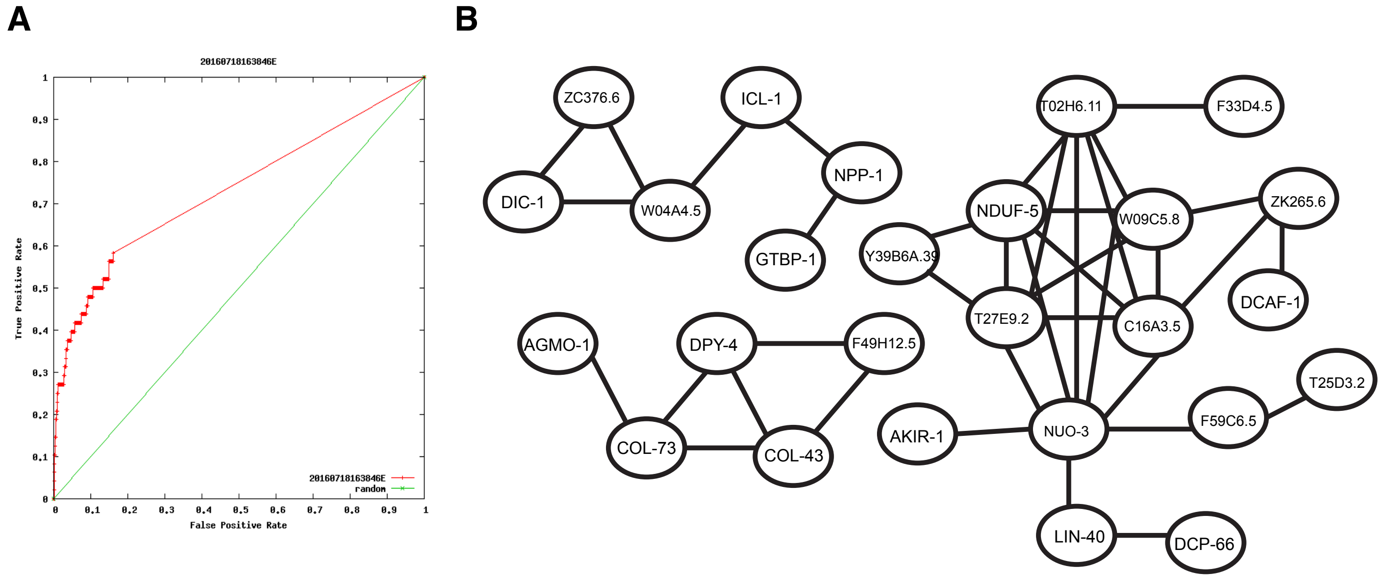


Figure 7. Conserved Nipi genes are highly connected. A. Area under the curve plot generated by WormNet for the connectivity of 46 Nipi genes conserved between vertebrates and invertebrates (red) compared to an equally sized random set (green). B. Connections between subsets of the same 46 genes as provided by WormNet. Only groups of at least 4 genes are shown.

Figure 8

Thakur et al.

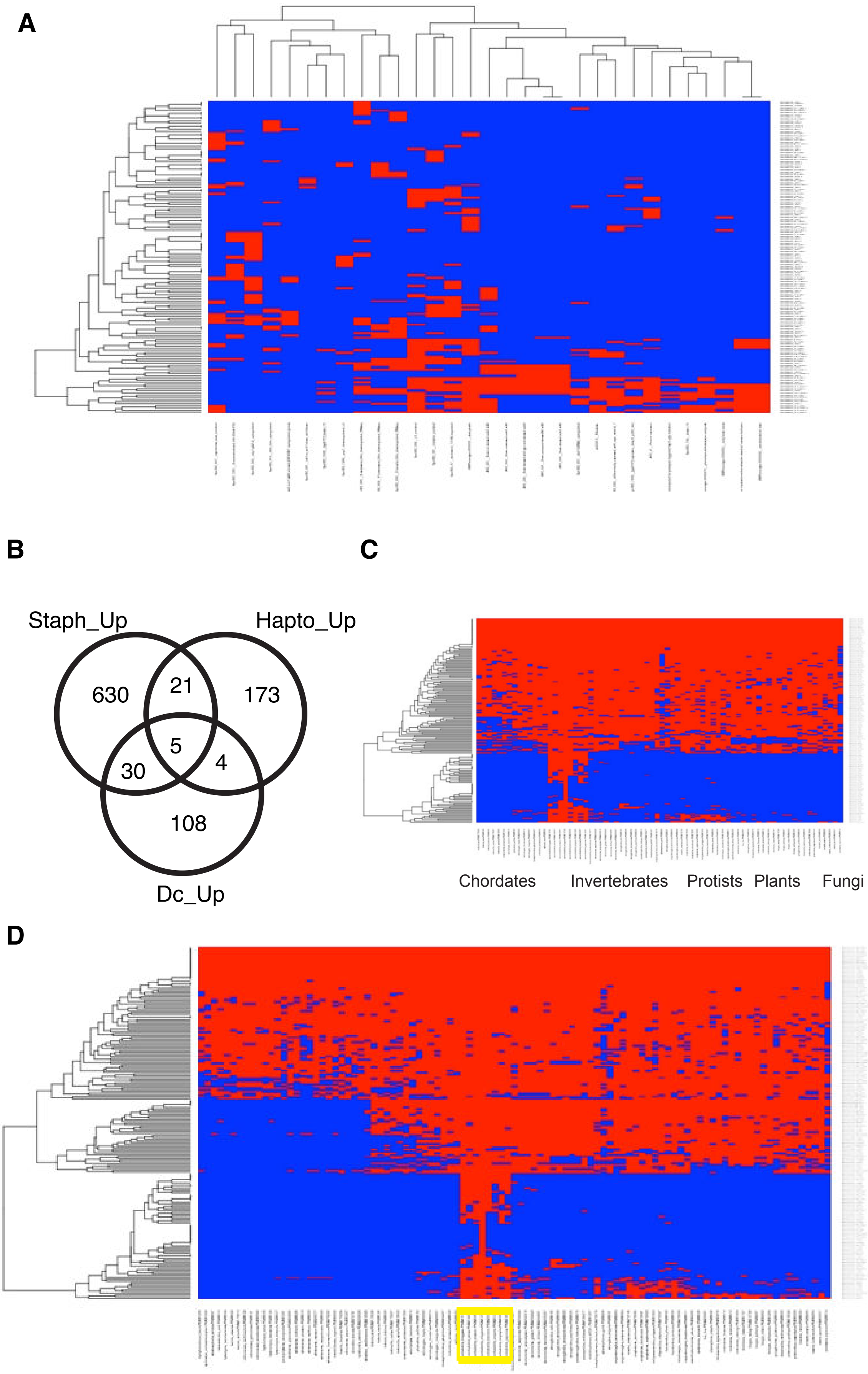


Figure 8. **A**. Clustering of enriched functional classes for *H. sphaerosporum* up-regulated genes. The identity of the classes is given in Table 2. X-axis; represent genes, Y-axis; enriched functional classes **B**. Venn diagram showing the overlap of individual genes between the classes Dc_Up, Hapto_Up and Staph (genes induced by the bacterium *Staphylococcus aureus* in an *hlh-30(tm1978)* mutant background). Phylogenetic profile of HaptoUp across 56 diverse Eukaryotic species (**C**) or specifically against nematodes (NemaProf; **D**). Each row is a gene, columns represents species (ordered according to the established evolutionary relationships among each group of species). *Caenorhabditis* species are highlighted by the yellow box at the bottom.

2.4 Publication 4

Coordinated inhibition of C/EBP by Tribbles in multiple tissues is essential for *C. elegans* development.
Kyung Won Kim , **Nishant Thakur** , Christopher A. Piggott , Shizue Om , Jolanta Polanowska , Yishi Jin and Nathalie Pujol.
I did bioinformatic analysis of the CEBP-1 TFs targets using YAAT tool (described in the publication 3). Using this analysis, I found that CEBP-1 TF′s targets cluster into two functional categories, one involved in the development and the other in stress. Further bionformatics analyses were performed to identify the cis-regulatory elements in CEBP-1 TF binding peaks. Using RSAT, I identified a conserved CEBP-1 TF binding motif. Apart from this I also performed analysis of the RNA-seq data generated for different *cebp-1* TF mutants. My analysis revealed that the underlying sequence data was of poor quality and therefore we did not include the RNA-seq analysis in the manuscript.

Coordinated inhibition of C/EBP by Tribbles in multiple tissues is essential for *C. elegans* development

Kyung Won Kim[1*], Nishant Thakur[2], Christopher A. Piggott[1], Shizue Omi[2], Jolanta Polanowska[2], Yishi Jin[1,3*] and Nathalie Pujol[2,*]

[1]Section of Neurobiology, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, USA

[2]Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université, Inserm, CNRS, Marseille, France

[3]Howard Hughes Medical Institute, University of California, San Diego, La Jolla, CA 92093, USA

* co-corresponding authors: k9kim@ucsd.edu, yijin@ucsd.edu & pujol@ciml.univ-mrs.fr

Abstract

Background

Tribbles proteins are conserved pseudokinases that function to control kinase signalling and transcription in diverse biological processes. Abnormal function in human Tribbles has been implicated in a number of diseases including leukaemia, metabolic syndromes and cardiovascular diseases. *Caenorhabditis elegans* Tribbles NIPI-3 was previously shown to activate host defense upon infection, by promoting the conserved PMK-1/p38 MAP kinase signalling pathway. Despite the prominent role of Tribbles proteins in many species, our knowledge of their mechanism of action is fragmented and the in vivo functional relevance of their interactions with other proteins remains largely unknown.

Results

Here, by characterizing *nipi-3* null mutants, we show that *nipi-3* is essential for larval development and viability. Through analyses of genetic suppressors of *nipi-3* null mutant lethality, we show that NIPI-3 negatively controls PMK-1/p38 signalling via transcriptional repression of the C/EBP transcription factor CEBP-1. We identified CEBP-1's transcriptional targets by ChIP-seq analyses and found them to be enriched in genes involved in development and stress responses. Unlike its cell-autonomous role in innate immunity, NIPI-3 is required in multiple tissues to control organismal development.

Conclusions

Together, our data uncover an unprecedented crosstalk involving multiple tissues, in which NIPI-3 acts as a master regulator to inhibit CEBP-1 and the PMK-1/p38 MAPK pathway. In doing so, it keeps innate immunity in check and ensures proper organismal development.

Keywords
Tribbles, C/EBP transcription factor, p38 MAP kinase, genetic suppression, transcriptional regulation, signal transduction, development, stress responses.

Background

The Tribbles genes encode a family of highly conserved pseudokinases, which lack key catalytic amino acids in the kinase domain [1-3]. Functional studies in multiple organisms have shown that these pseudokinases play diverse roles in innate immunity, cell signalling, energy homeostasis, and cell division [2, 4]. The Drosophila *tribbles* gene is required for cell proliferation and migration in embryogenesis and oogenesis [5-8]. The mammalian Tribbles family includes three genes, Trib1, Trib2, and Trib3, each of which plays unique roles in signalling networks regulating adipose tissue, metabolic homeostasis, and the immune system [2, 4]. Abnormal function in human Tribbles has been implicated in a number of diseases including leukaemia, metabolic syndromes and cardiovascular disease [9].

Mammalian Tribbles can interact with components of mitogen-activated protein (MAP) kinase pathway, and act as adaptor proteins to modulate the strength and output of kinase signalling cascades [10-12]. Like fly Tribbles [7, 13], mammalian Tribbles bind to the basic leucine zipper (bZIP) transcription factors. Trib1 and Trib2 can induce degradation of several C/EBP (CCAAT/enhancer-binding protein) members in a context-dependent manner [1, 14, 15]. Trib3 binds to and inhibits the transcriptional activity of both ATF4 (Activating Transcription Factor 4) and CHOP (C/EBP homologous protein) in cultured cell lines [16-19], although the in vivo functional relevance of such protein interactions remains unknown.

Given its extensive connections to different cellular processes, many questions remain to be answered regarding the role of Tribbles at the organismal level. NIPI-3 (No Induction of Peptide after *Drechmeria* Infection-3) is the single Tribbles protein in *C. elegans* [20]. A role for *nipi*-3 in the innate immune response was previously uncovered through the isolation of a partial loss-of-function mutation, which contains a missense mutation in the pseudokinase domain [20]. NIPI-3 is required for the upregulation of antimicrobial peptide (AMP) gene expression following infection by the fungus *Drechmeria coniospora* [20]. It acts upstream of a p38 MAP kinase (MAPK) pathway consisting of NSY-1/MAPKKK, SEK-1/MAPKK, and PMK-1/MAPK [20, 21]. Both NIPI-3 and all components of the MAPK cascade are required cell-autonomously in the epidermis during the immune response [20].

In this study, we generated null mutations of *nipi-3* and uncovered a novel role in animal development and viability. The lethality of *nipi-3* null animals is completely suppressed by loss of function in CEBP-1, a *C. elegans* member of the C/EBP family, previously known to be required for adult sensory axon regeneration and neuronal stress responses [22, 23]. Unexpectedly, loss of function in components of the PMK-1/p38 MAPK cascade also suppresses the lethality of *nipi-3* null animals. In *nipi-3* mutants, the levels of activated PMK-1 are increased, in a *cebp-1* dependent manner. Through ChIP-seq analyses and identification of target genes, we found that CEBP-1 binds to a conserved DNA motif. Our analyses of candidate target genes of CEBP-1 suggest a functional enrichment in development and stress responses. Importantly, in contrast to its role in innate immunity, we show that NIPI-3 acts in multiple tissues to negatively regulate transcriptional expression of *cebp-1*. This inhibition of CEBP-1 by NIPI-3 is also required across multiple tissues to enable larval development and maintain fecundity. The coordinated inhibition of CEBP-1 by NIPI-3 in multiple tissues reveals novel requirements for systemic regulation of signalling pathways in organism development.

Results

***C. elegans* Tribbles *nipi-3* is required for larval development and viability.**

To better understand the biological roles of *nipi-3*, we used CRISPR-Cas9 genome editing [24-27] to generate two deletion alleles of *nipi-3* (*fr148* and *ju1293*) and a GFP knock-in (KI) (*fr152)* (Fig. 1a; see Methods and below). The *fr148* and *ju1293* deletion alleles remove 1.6 kb and 0.6 kb of the 5' region of the gene, respectively (Fig. 1a), resulting in molecular nulls for *nipi-3*. The phenotypes of homozygous mutants of either *nipi-3* deletion allele, designated as null (0), were indistinguishable (Fig. 1c,d). Mutants arrested development at the second to third larval stages (L2-L3) (see below) and eventually died between 5-10 days after hatching. When compared to wild type larvae at the same stage (2 days post-hatching), *nipi-3(0)* arrested larvae displayed a small and dumpy body morphology. At 3 days post-hatching wild type animals reached the adult stage, as evidenced by fusion of seam cells (lateral epidermal cells), formation of adult alae and the vulva (Fig. 1b,f). By contrast, all age-matched *nipi-3(0)* animals were arrested at L2/3, as the seam cells did not fuse, adult alae were not observed, and the vulval invagination did not occur (Fig. 1c,d,f). In these mutant animals, the germline also appeared to be arrested, generally at L3 based on the size of the gonad and the number of germ cells (Fig. 1c,d,g). Occasionally, in *nipi-3(0)* animals with longer bodies, we observed some sperm or a few unfertilized oocytes. The *nipi-3(0)* animals also exhibited an abnormal pharyngeal morphology (Additional file 1: Figure S1). We rescued the larval lethality and sterility of *nipi-3(0)* by expressing the wild-type *nipi-3* genomic DNA as high-copy number extrachromosomal arrays (Fig. 1a,e,h; methods). As expression from such transgenes is silenced in the germline [28], this result indicates that the larval lethality and germline development defects of *nipi-3(0)* are both primarily due to its function in somatic tissues. Thus, the analyses of *nipi-3* null animals indicate an essential somatic role of *nipi-3* in organism development.

Knocking-in *gfp* to the *nipi-3* locus, which produced a protein tagged at its N-terminus (GFP::NIPI-3), had no adverse effect; KI animals (*fr152)* were fully viable and indistinguishable from wild type in growth and movement. We observed GFP expression in the epidermis, intestine and in neurons (Fig. 1i), consistent with the previously reported expression pattern obtained using transgenic transcriptional reporters [20]. Interestingly, although NIPI-3 does not have a clearly identifiable nuclear-localization signal, GFP::NIPI-3 expression was observed predominantly in the nuclei (Fig. 1i), with an overall intensity peaking at the L2-L3 stages. This nuclear localization suggests a role for NIPI-3 in regulation of gene expression.

Loss of *cebp-1* suppresses the lethality, but not innate immune response defect, of *nipi-3* null animals.

To dissect the molecular mechanisms involving NIPI-3, we undertook a yeast two-hybrid screen using the full-length NIPI-3 protein as bait (Cypowyj S. *et al*, MS in preparation). One prominent candidate interacting partner was CEBP-1, a member of the C/EBP family of transcription factors [22]. The CEBP-1 protein consists of 319 amino acids, with a bZIP domain at its C-terminus. In further analyses using the yeast two-hybrid assay, we found that the N-terminal region (amino acids 1-115) of CEBP-1 was sufficient for binding NIPI-3 (Additional file 2: Figure S2). This interaction is reminiscent of those observed for fly and vertebrate Tribbles proteins, which bind and degrade C/EBP family proteins [1, 7, 13-15, 29], with human Trib1 binding the N-terminus of C/EBPα [1]. Thus, the ability of Tribbles and C/EBP proteins to interact directly is likely conserved from *C. elegans* to humans.

To understand the functional significance of the observed protein interaction, we next performed genetic analysis using null mutations of *nipi-3* and *cebp-1.* Null mutants of *cebp-1* show normal development and body appearance. Remarkably, *cebp-1(0)* completely suppressed the growth and fertility defects of *nipi-3(0)* mutants (Fig. 2b,h; Methods). *cebp-1(0)* also completely suppressed the body size defect and the developmental delay observed at 25°C of the partial loss-of-function *nipi-3(fr4)* mutants (Additional file 3: Figure S3). NIPI-3 is also known to be necessary for the epidermal innate immune response upon fungal infection [20]. The expression of AMP genes after fungal infection is highly induced in wild type animals and abrogated in *nipi-3(fr4)* mutant [20]. However, *cebp-1(0) nipi-3(0)* animals did not exhibit an induction of the AMP gene *nlp-34* upon infection (Additional file 4: Figure S4a). Further, *cebp-1(0) nipi-3(0)* animals expressing wild-type *cebp-1* in a tissue-specific manner, either in the epidermis or neurons also showed no AMP gene induction (Additional file 4: Figure S4a). These results indicate that loss of *cebp-1* does not suppress the immune response defect in *nipi-3(0)*. As *cebp-1(0)* strongly impairs sensory axon regeneration after laser axotomy [22], we also tested whether *nipi-3* affected PLM axon regeneration. Neither *nipi-3(0)* nor *nipi-3(fr4)* showed significant effects in axon regeneration, and *cebp-1(0) nipi-3(0)* double mutants showed impaired axon regeneration similar to *cebp-1(0)* (Additional file 4: Figure S4b). These results show that the genetic interaction between *nipi-3* and *cebp-1* is highly specific for larval development and organism fecundity.

Loss of the PMK-1/p38 MAPK pathway also suppresses *nipi-3(0)* lethality.

To gain further insight into the mechanism underlying *nipi-3*'s role in animal development, we performed a forward genetic screen for suppressors of *nipi-3(0)* lethality. We mutagenized *nipi-3(ju1293); Tg[nipi-3(+); myo-2p::mCherry]* animals (Methods). Among their F_2 progeny, we isolated fertile animals that had lost the rescuing transgene (Fig. 2a), and established multiple suppressor lines (genotypes designated as *nipi-3(0); suppressor).* We screened ~16,000 haploid genomes and identified 7 independent suppressor alleles. All the identified suppressors in this study were associated with a full reversion of the *nipi-3(0)* lethality; fertile adults could be propagated without the *nipi-3(+)* rescuing transgene.

We characterized several suppressor mutations using candidate gene analyses in combination with whole-genome sequencing. One suppressor caused a missense mutation in the bZIP domain of *cebp-1,* and behaved like *cebp-1(0)* (Fig. 2b,h). Among the other suppressors of *nipi-3(0)*, we found missense alterations in TIR-1/SARM (Sterile alpha and TIR(Toll-interleukin receptor) motif-containing protein), NSY-1/MAPKKK, SEK-1/MAPKK, and MAK-2/MAPKAPK (MAPK-activated protein kinase) (Fig. 2c–f). These mutations are located in known functional domains, including the TIR domain for TIR-1*,* the kinase domains for NSY-1, SEK-1, and MAK-2, and the DUF4071 domain, commonly found at the N-terminus of many serine-threonine kinase-like proteins, for NSY-1 (Fig. 2c–f). We then tested known null mutants in each of these genes and found them to suppress *nipi-3(0)* to the same degree as our suppressor mutations, as shown by quantification of the body length (Fig. 2b–f,h, Additional file 5: Figure S5a). Thus, loss of function in *tir-1* and these three kinase genes causes strong suppression of *nipi-3(0).*

NIPI-3 has a specific role in the regulation of epidermal defense genes. It acts together with TIR-1, NSY-1, and SEK-1, as well as several other genes including *tpa-1*, *pmk-1*, and *sta-2* [20, 21, 30]. We therefore tested mutants for these three latter genes for their genetic interaction with *nipi-3(0).* We found that loss of *pmk-1,* but not *tpa-1* or *sta-2,* suppressed the lethality of *nipi-3(0)* (Fig. 2g,h for *pmk-1*; Additional file 5: Figure S5b for *tpa-1* and *sta-2*)*.* Loss of *pmk-1* resulted in suppression of *nipi-3(0)* phenotypes similar to the other suppressor mutants such that *nipi-3(0); pmk-1(0)* double mutants developed into fertile adults with adult alae and vulva. The rescue of body size, however, was not complete (Fig. 2g,h). As *cebp-1* is involved in two other MAPK cascades known for their roles in adult axon regeneration [22, 31, 32], we tested mutations in several other candidate genes and found that loss of function in *dlk-1, pmk-3, mlk-1* or *kgb-1* did not suppress *nipi-3(0)* defects (Additional file 5: Figure S5b). Together, our analyses from both the forward genetic screening and test of candidate mutants reveal a previously unknown role of CEBP-1 in larval development mediated by NIPI-3, and a novel genetic interaction between NIPI-3 and the PMK-1/p38 MAPK cascade.

NIPI-3 inhibits PMK-1 phosphorylation via CEBP-1.

To dissect the mechanism underlying the interaction between PMK-1/p38 MAPK and NIPI-3, we first asked how the levels of active PMK-1 might be altered in *nipi-3(0)* suppressor animals. We performed Western blot analysis using an anti-phospho-p38 MAPK antibody that specifically recognizes phosphorylated, active PMK-1 [30]. We made protein lysates from animals at 1 day post-hatching (L2) because *nipi-3(0)* animals at this stage were as healthy as wild type. We observed that levels of active PMK-1 were significantly increased in *nipi-3(0)*, but remained similar to wild type in *cebp-1(0)* and *cebp-1(0) nipi-3(0)* animals (Fig. 3a,b). Phosphorylated PMK-1 was undetectable in *nipi-3(0) sek-1(0)* animals, consistent with PMK-1 being activated by SEK-1 (Fig. 3a,b). The total PMK-1 levels were likely unchanged in *nipi-3(0)* as the mRNA levels of *pmk-1* were comparable between *nipi-3(0)* and wild type when assessed by quantitative RT-PCR (Fig. 3c)*.* We note that *mak-2(0)* did not affect phosphorylated PMK-1 (Additional file 6: Figure S6), suggesting that MAK-2 likely acts downstream of, or in parallel to, PMK-1. Together, these results suggest that the abnormally high levels of phosphorylated PMK-1 in *nipi-3(0)* are dependent on *cebp-1*.

NIPI-3 represses the transcription of *cebp-1* in multiple tissues.

To dissect how NIPI-3 inhibits CEBP-1, we examined whether *cebp-1* levels were altered in *nipi-3(0)* and in each *nipi-3(0)* suppressor. Transcriptional reporters of *cebp-1* (*Tg[cebp-1p::GFP]*) were broadly expressed in most post-embryonic tissues, including epidermis, muscles, pharynx, intestine and neurons (Fig. 3d–j). Strikingly, the expression of the *cebp-1* transcriptional reporter was highly and significantly increased in both *nipi-3(fr4)* and *nipi-3(0)* mutants, compared with wild type animals (Fig. 3d–f). Quantitative RT-PCR analysis also showed significantly increased expression of *cebp-1* mRNAs in *nipi-3* mutant animals (Fig. 3k). Consistent with the observed transcriptional regulation, we found that a translational CEBP-1::GFP reporter driven by an heterologous epidermal promoter showed no detectable differences in GFP expression in a *nipi-3(0)* background (Fig. 3l–n). The increased expression of *cebp-1p::GFP* in *nipi-3(0)* was reduced to normal levels when a *nipi-3(+)* transgene was introduced (Fig. 3g), indicating that NIPI-3 represses the transcription of *cebp-1*. Additionally, the transcriptional repression of *cebp-1* by NIPI-3 was largely independent of the PMK-1 pathway since *cebp-1p::GFP* expression in *nipi-3(0)* remained high in animals that also carried a null mutation of *nsy-1, pmk-1*, or *mak-2* (Fig. 3h–j). Together, the results show that NIPI-3 negatively regulates expression of *cebp-1* at the transcriptional level and that CEBP-1 acts upstream of the PMK-1 pathway.

The tight regulation of *cebp-1'*s expression level is critical for animal viability. We found that suppression of *nipi-3(0)* by *cebp-1(0)* was semi-dominant, as *cebp-1(0)/+* caused partial but significant suppression of the short body length of *nipi-3(0)* mutants (Additional file 7: Figure S7). Moreover, in a wild type background, the transgene *eft-3p:: CEBP-1::GFP,* which drives strong and ubiquitous expression of CEBP-1, caused dose-dependent lethality (see Methods). In addition, expression of a full-length functional translational reporter of *cebp-1* (*cebp-1p::CEBP-1::GFP*) in *nipi-3(0)* mutant animals exacerbated developmental defects and accelerated larval lethality, while the same transgene showed no such effects in a wild type background. Together, these results support the conclusion that the lethality observed in *nipi-3(0)* mutants is a direct consequence of *cebp-1* overexpression.

Overexpression of truncated forms of CEBP-1 suppresses *nipi-3(0)* lethality.

To dissect the molecular basis of CEBP-1's role in animal development, we expressed truncated forms of CEBP-1 lacking the bZIP domain (aa 1-230 or aa 1-115), or lacking the N-terminus (aa 237-319) in *nipi-3(0)* mutants. Surprisingly, in stark contrast to the strong lethality caused by overexpressing full-length CEBP-1 in *nipi-3(0)* mutants, we found that expression of either the N-terminal fragment or the bZIP domain of CEBP-1 alone resulted in significant suppression of *nipi-3(0)* defects (Fig. 4a–d). Overexpression of either N- or C-terminal truncated protein caused no defects either in the wild type or *cebp-1(0)* backgrounds. Among the three CEBP-1 fragments, the expression of C-terminal CEBP-1(aa 237-319) showed most effective rescue, judged by the fecundity of the transgenic lines and quantitative comparisons of body length (Fig. 4c,d). The expression levels of CEBP-1(aa 1-230)::GFP were markedly increased in *nipi-3(0)*, presumably reflecting the transcriptional regulation of *cebp-1* described above. We noticed the fluorescence intensity was most strongly increased in the epidermis and neurons, throughout the head region and in the ventral nerve cords (Fig. 4e–g). These observations suggest that the truncated forms of CEBP-1 act in a dominant negative manner to inhibit the activity of the endogenous CEBP-1.

CEBP-1 binds conserved DNA motifs in genes regulating development and stress response.

To gain further insight into CEBP-1's function in animal development, we next sought candidate target genes of CEBP-1 by performing chromatin immunoprecipitation and deep-sequencing (ChIP-seq) analysis on transgenic animals expressing a functional FLAG-tagged CEBP-1 protein in a *cebp-1(0)* background (Methods). We found 209 CEBP-1 ChIP-seq peaks in the genome that were associated with 212 coding genes (Additional file 11: Table S1). CEBP-1 peaks were preferentially located within the promoter regions of the target genes (169 genes, 79%), less frequently within introns (43 genes, 21%), and never within exons.

We then performed motif analysis of the genomic regions bound by CEBP-1 using motif discovery tools, MEME (Multiple Em for Motif Elicitation) [33] and RSAT (Regulatory Sequence Analysis Tools) [34]. The most over-represented motif DTTDYGAAAH was found in 139 out of 209 CEBP-1 ChIP-seq peak regions (Fig. 5a). We then compared this motif with published motifs using the motif comparison tool Tomtom [35] and found the most statistically significant similarities to vertebrate C/EBP binding motifs [36]. The conservation of CEBP-1 binding motif further reinforces the functional parallels between *C. elegans* CEBP-1 and vertebrate C/EBPs.

As CEBP-1 likely acts upstream of the PMK-1 pathway, we searched among the targets of CEBP-1 for components of the PMK-1 pathway and found CEBP-1 ChIP-seq peaks present in the promoter of *sek-1* (Additional file 8: Figure S8). When we examined the mRNA levels of *sek-1* by quantitative RT-PCR in *nipi-3(0)* animals where *cebp-1* is overexpressed, we observed increased *sek-1* mRNA levels (Fig. 5b). In contrast, in *nipi-3(0) cebp-1(0)* animals the levels of *sek-1* mRNAs were similar to wild type. These results suggest that the abnormally high levels of *cebp-1* in *nipi-3(0)* can promote the expression of *sek-1*, which in turn promotes the phosphorylation of PMK-1 [30], leading to abnormal larval development and lethality.

We then asked whether the list of potential CEBP-1 targets was enriched for genes with specific functions. To this end, we searched for enriched categories through an EASE (Expression Analysis Systematic Explorer) analysis [37, 38], using our in-house database of functional annotations as described [39]. This includes 4,600 datasets automatically updated from multiple sources including Wormbase, Flybase, KEGG, and relevant RNAi databases. The great majority (80%) of CEBP-1 target genes were associated with at least one of 33 enriched functional classes ($p < 10^{-5}$; Additional file 11: Table S1). Hierarchical clustering of the genes in each of the enriched classes identified two main groups (Fig. 5c; Additional file 9: Figure S9; Additional file 11: Table S1). One group is related to development with phenotypic classes such as "larval lethal" or "slow growth" and includes genes involved in basic cellular processes, transcription, translation or endocytosis. The second group is related to the response to biotic or abiotic stress, including response to cadmium, hygromycin or bacterial toxins. Interestingly, McEwan *et al.* have found that most of the genes upregulated in *nipi-3(fr4)* mutants are also induced by the translational inhibitory toxin ToxA. Out of the 14 stress-related classes associated with the *cebp-1* targets, 10 are shared with those found for genes upregulated in *nipi-3(fr4)* (D. L. McEwan, personal communication; $p < 10^{-10}$ Additional file 11: Table S1). On the other hand, consistent with the fact that the *nipi-3(fr4)* allele does not provoke larval lethality, only 1 out of the 10 classes in the development cluster is shared between the *cebp-1* targets and the genes up-regulated in *nipi-3(fr4)*. These analyses suggest that CEBP-1 regulates genes functioning in development and in stress responses, and might have a particularly important impact on organismal physiology when overexpressed in a *nipi-3* mutant context.

NIPI-3 is required in multiple tissues to ensure proper larval development.

We next asked in which tissue the expression of *nipi-3(+)* is required for animal viability. We expressed *nipi-3(+)* in a tissue-specific manner, using intestinal, epidermal, and pan-neuronal promoters, in the *nipi-3(0)* background (Fig. 6a–f). In contrast to the complete rescue of body size and lethality in *nipi-3(0)* mutants expressing *nipi-3(+)* under its own promoter (Fig. 1e,h), expressing *nipi-3(+)* in individual tissues failed to rescue the developmental arrest (Fig. 6a–c,f). Pan-neuronal expression of *nipi-3(+)* resulted in slightly increased body length of *nipi-3(0)* 3 days post-hatching (Fig. 6c,f). Since *nipi-3* activity strongly inhibits *cebp-1* transcription in the epidermis and neurons (Fig. 4e–g), we expressed *nipi-3(+)* in both these tissues together, and found that these transgenic animals showed an increased body size (Fig. 6e,f), compared to those expressing *nipi-3(+)* in each tissue alone, or in the intestine and epidermis simultaneously (Fig. 6a–d,f). Some of the transgenic animals with a body length closer to that of wild type animals showed improved somatic and germline development, with formation of vulva and adult alae, and produced a few viable but infertile progeny. When we expressed *nipi-3(+)* in all three tissues together, the transgenic animals showed no further improvement of body size compared to those expressing *nipi-3(+)* in both the epidermis and neurons (Fig. 6f), and did not recapitulate the rescue of lethality associated with the expression of *nipi-3* under its own promoter. Thus, we conclude that NIPI-3 is required in multiple tissues, particularly epidermis and neurons, for animal growth and development.

Suppression of *nipi-3(0)* by *cebp-1(0)* also requires a block of CEBP-1 activity in multiple tissues.

Conversely, we asked whether expression of *cebp-1(+)* in a single tissue might cause the viable *nipi-3(0) cebp-1(0)* animals to die. We expressed *cebp-1(+)* using the intestinal, epidermal, and pan-neuronal promoters in *cebp-1(0) nipi-3(0)* double mutants (Fig. 6g). Expression of *cebp-1(+)* in individual tissues was insufficient to produce larval lethal phenotypes in *cebp-1(0) nipi-3(0)* animals. Interestingly, epidermal or neuronal expression of *cebp-1(+)* in *cebp-1(0) nipi-3(0)* double mutants caused short body length, but not in *cebp-1(0)* mutants (Fig. 6g). Co-expression of *cebp-1(+)* in the epidermis and neurons resulted in further reduction in body length, although these were not as severe as those in *nipi-3(0)* animals (Fig. 6g). In addition, we noticed that the same transgenes expressing *cebp-1(+)* in both epidermis and neurons caused an abnormal pharyngeal morphology in *cebp-1(0) nipi-3(0)* animals, similar to that seen in *nipi-3(0)* mutants (Additional file 1: Figure S1). Together, our data suggest that the tight regulation of both NIPI-3 and CEBP-1 in multiple tissues is required in a systemic manner for normal animal growth and development.

Discussion

C. elegans Tribbles NIPI-3 was identified on the basis of its roles in host defense [20, 21]. Here, through generation and analyses of null alleles, we find *nipi-3* to be essential for animal development and viability. Remarkably, the larval arrest and lethality caused by complete loss of *nipi-3* is fully suppressed by loss of *cebp-1*, a C/EBP bZIP transcription factor, or by loss of function in the PMK-1/p38 MAPK cascade including *tir-1/*SARM, *nsy-1/*MAPKKK, *sek-1*/MAPKK, *pmk-1*/MAPK. Our data show that complete elimination of the function of *nipi-3* causes abnormally high expression of CEBP-1, and activation of PMK-1 MAPK. This then disrupts development and leads to death. The level of *sek-1* mRNA is increased in *nipi-3(0)* mutants but not in *cebp-1(0)* nor in *nipi-3 cebp-1* animals. The level of phosphorylated (active) PMK-1 follows the same trend. Coupled with our ChIP-seq analyses and genetic epistasis data, this suggests that CEBP-1 acts as a direct positive regulator of *sek-1*. The PMK-1 pathway is therefore activated when CEBP-1 expression is high in *nipi-3(0).* On the other hand, *cebp-1* expression levels remain high in *nipi-3(0); pmk-1(0)* animals, confirming that CEBP-1 does not act downstream of the PMK-1 pathway. Together, these results suggest that NIPI-3 negatively regulates the PMK-1 MAPK cascade, via CEBP-1, to promote animal viability and development (Fig. 6h).

In innate immunity, however, *nipi-3* cell-autonomously promotes or enhances the same p38 kinase cascade to activate host defense in the epidermis [20]. It has been shown that overexpression of *sek-1* in the epidermis rescues the block of AMP induction in *nipi-3* mutants upon fungal infection [20], and an overexpression of *nipi-3* provokes an increase in the constitutive expression of AMP which is dependent on the p38 cascade [21]. It is intriguing that NIPI-3 appears to be capable of activating or inhibiting PMK-1/p38 in the epidermis at different times or under different conditions (infection vs. development). How might NIPI-3 achieve this dual role under different stresses and in altered cellular contexts? As Tribbles proteins are well known to act as adaptors, NIPI-3 might be regulated via binding with other cofactors only present under specific circumstances. Indeed, we find that other upstream and downstream components of the epidermal immune response cascade are not involved in the developmental regulation described here. Thus, the core PMK-1/p38 MAPK cassette has evolved context-specific functions depending on different upstream regulators or cofactors [40, 41]. Members of the Tribbles family in other species have been mostly studied in the context of cell proliferation, adipocyte tissue differentiation, energy metabolism and immunity, where they function in a cell-autonomous manner. Our discovery of the opposing roles of NIPI-3 in development and in the immune response illustrates how cellular context can alter the function of highly conserved signalling molecules.

Negative regulation of C/EBP by Tribbles has been observed throughout the animal kingdom. Drosophila and mammalian Tribbles bind and degrade C/EBP proteins [1, 7, 13-15]. We find that *C. elegans* NIPI-3 represses the transcription of *cebp-1*, which has important functional consequences in vivo. This form of regulation has not been reported in other organisms. Given its nuclear localization, NIPI-3 may inhibit the transcription of *cebp-1* by interfering with other transcription factor(s). The promoter of *cebp-1* contains putative CEBP-1 binding consensus motifs, raising the possibility that NIPI-3 by binding to CEBP-1 may also alter the transcriptional activity of CEBP-1.

NIPI-3 is required to control CEBP-1 levels in multiple tissues for animal development and viability. Consistent with the inhibition of CEBP-1 expression by NIPI-3 in the epidermis and neurons, simultaneous expression of *nipi-3(+)* in both tissues makes noticeable contribution to animal development in *nipi-3(0)* mutants, compared with *nipi-3(+)* expression in single tissues. Conversely, simultaneous expression of *cebp-1(+)* in both epidermis and neurons causes noticeable defects to animal development in *nipi-3(0) cebp-1(0)* mutants, compared with *cebp-1(+)* expression in single tissues. Thus, a tightly regulated coordination of these two genes' interactions in multiple tissues is required to ensure proper development.

A key conclusion from our study is that the precise control of CEBP-1 and PMK-1/p38 MAPK pathways in multiple tissues is critical for organismal development. NIPI-3 acts as a master regulator to prevent improper activation of CEBP-1 and PMK-1, whose hyperactivation during development has deleterious consequences. Interestingly, hyperactivation of PMK-1/p38 was previously shown to block larval development when the endoplasmic reticulum unfolded protein response was altered [42]. Moreover, innate immune activation with a xenobiotic that provides protection from bacterial infection in the adult, has been shown to provoke a growth delay during development [43]. Subsequently, an elegant genetic suppressor screen revealed that mutations in the PMK-1/p38 MAPK pathway suppressed this developmental phenotype [44]. Thus the NIPI-3/CEBP-1 axis is a key mechanism by which immune effector expression is held in check during nematode development.

During normal development, both CEBP-1 and PMK-1 are maintained at a basal level by NIPI-3. The levels of inducible signalling from these pathways are, however, important for animals to protect themselves or to promote repair. For instance, following fungal infection, NIPI-3 promotes PMK-1/p38 MAPK signalling pathway in the epidermis [20]. Thus, animals can successfully defend themselves from fungal infection with activated PMK-1 locally in the epidermis, while survival is not affected as PMK-1 remains inactive in other tissues. Similarly, CEBP-1 is known to play a key role in neuronal stress responses [22, 23, 45], and we identified potential CEBP-1 target genes that are involved in different stress responses. Moreover, a concomitant study has identified NIPI-3 as a negative regulator of CEBP-1 in intestinal defense against the bacterial toxin ToxA (McEwan *et al.*). An important challenge for the future will be to understand how NIPI-3 regulates its downstream pathways and how NIPI-3 itself is regulated depending on developmental and environmental conditions. Understanding the molecular mechanism of this systemic, coordinated regulation should advance our knowledge of how animal development can be maintained in the face of environmental stresses.

Conclusions

We showed a novel essential role for the *C. elegans* Tribbles homolog NIPI-3 in animal development and viability, which requires NIPI-3's function in multiple tissues. NIPI-3 acts as a master regulator to prevent improper activation of a C/EBP transcription factor and a conserved

PMK-1/p38 MAPK signalling cascade known to control innate immunity. These findings suggest that innate immune responses are tightly controlled for proper organismal development.

Methods

Strains, transgenes and plasmids. *C. elegans* strains were maintained under standard conditions at 20˚C unless mentioned. Wild type was the N2 Bristol strain [46]. New strains were constructed using standard procedures and all genotypes were confirmed by PCR or sequencing. All strains and their genotypes used in this study are described in Additional file 12: Table S2. Extrachromosomal array transgenic lines were generated as described [47]. Expression constructs, transgenes and strain genotypes are also summarized in Additional file 12: Table S2. For all experiments, at least two independent transgenic lines were examined and quantitative data is shown for one. In our studies of *cebp-1* dosage effect, we could not generate transgenic animals when injecting 10 ng/µl of *eft-3p::CEBP-1::GFP* transgenes into wild type animals, while many transgenic lines were obtained when *cebp-1::gfp* DNA was injected with lower concentrations (*i.e.,* 1 ng/µl or 0.01 ng/µl). All plasmids used in this study are described in Additional file 13: Table S3a.

CRISPR-Cas9-mediated deletion and GFP KI. We generated the *nipi-3(ju1293)* deletion allele using the co-CRISPR method [25, 48]. We used four sgRNAs targeting the N-terminus of the *nipi-3* gene (Additional file 13: Table S3a). *U6p::nipi-3* sgRNAs were generated by Gibson assembly and injected into *cebp-1(tm2807)* worms using standard methods, in mixtures composed of 30 ng/µl of each *nipi-3* sgRNA, 50 ng/µl of *eft-3p::Cas9-SV40NLS::tbb-2 3'UTR,* 50 ng/µl of *U6p::unc-22* sgRNA and 1.5 ng/µl of *myo-2p::mCherry.* For the *nipi-3(fr148)* deletion allele, a single sgRNA targeting the N-terminus of the *nipi-3* gene (Additional file 13: Table S3a) was injected into *cebp-1(tm2807)* worms, in mixtures composed of 50 ng/µl of *nipi-3* sgRNA, 30 ng/µl of *eft-3p::Cas9-SV40NLS::tbb-2 3'UTR,* and 30 ng/µl of *col-12p::dsRed* [49]. Note that we also injected into many wild type worms with various combinations of *nipi-3* sgRNAs*,* but failed to isolate any deletion alleles.

GFP KI in the *nipi-3* locus, *nipi-3(fr152)* was generated with the same mixture as for *nipi-3(fr148)* providing 30 ng/µl of the *nipi-3* repair template pSO1. This template was generated by Gibson assembly in the SEC cassette containing vector pDD282 [50] with 716 bp and 518 bp homology arm upstream and downstream of the *nipi-3* start codon, respectively (Additional file 13: Table S3a).

Genetic screen for *nipi-3(0)* lethality suppressors. *nipi-3(ju1293)* mutant animals carrying *juEx6807[nipi-3 genomic DNA; myo-2p::mCherry]* were mutagenized using 45 mM Ethyl Methane Sulphonate (EMS) following standard procedures as described [46]. Animals were distributed onto NGM plates seeded with *E. coli* OP50 and screened in the F_2 generation for normal animal growth reaching adulthood without expressing the transgene (no pharyngeal mCherry) under a fluorescence dissecting microscope.

Mapping and cloning of *nipi-3* suppressor alleles. We first performed conventional Sanger sequencing analysis for all suppressor alleles for *cebp-1*, and determined that the *ju1367* allele affected *cebp-1.* We next sequenced *mak-2*, which was previously known to act in the same pathway as *cebp-1* in neurons [22], and found *ju1349* and *ju1352* alleles to affect *mak-2.* All other suppressors were analysed by whole genome sequencing analysis and SNP mapping following established methods [51]. Briefly, genomic DNA was prepared using Puregene Cell and Tissue Kit (Qiagen) according to the manufacturer's instruction, and 20X coverage of sequences were obtained using 90-bp paired-end Illumina Hiseq 2000 at Beijing Genomics Institute (BGI Americas). The raw sequences were mapped to the *C. elegans* reference genome (WS220/ce10) using BWA [52] in the Galaxy platform (http://usegalaxy.org) [53]. Following subtraction of the nucleotide variants in the original strains, we generated a list of candidate genes containing unique homozygous nucleotide variants that were predicated to alter the function of the gene. We then confirmed the causality of the candidate genes by testing the known null alleles on the suppression of *nipi-3(0)*.

Body Length Analysis. To examine body length of animals during development, we obtained synchronized animals. Briefly, 5-15 gravid adults were placed on a seeded NGM plate to allow egg-laying for 3 hours. Eggs laid during this period were incubated at 20 °C for 72 h, and the animals were then mounted on 2% agarose pads containing a drop of 2.5 mM levamisole and photographed with a Zeiss axioplan compound microscope, using Nomarski-DIC optics and an attached AxioCam digital camera. ImageJ software (NIH) was used to measure body length by drawing a freehand midline from the tip of the nose to the tip of the tail of each animal.

Western blot analysis. Worms of each genotype (80-100 individuals of L2 stage worms) were collected and washed with M9 buffer, and boiled in SDS sample buffer for 10 min and loaded onto SDS-PAGE gel (Bio-Rad). A 1:500 dilution of rabbit anti-phospho-p38 MAPK (Cell Signaling, #9211) and a 1:10,000 dilution of mouse anti-actin (MPbio, #08691001) were used as primary antibodies. ImageJ was used to quantify the intensity of immunoblot bands.

Quantitative RT-PCR. Quantitative real-time PCR was performed as previously described [49]. Sequences of primers are given in the Additional file 13: Table S3b. To collect synchronized *nipi-3(0)* homozygous animals, we maintained *nipi-3(0)* animals on *cebp-1* RNAi plates. Gravid adults were then treated by bleaching solution to collect *nipi-3(0)* embryos, which were placed directly on regular NGM plates for 2 days in parallel to other strains.

Fluorescence microscopy and axon regeneration by laser axotomy. Animals were mounted on 2% agarose pads and immobilized with 2.5 mM levamisole. For transcriptional *cebp-1p::GFP*, GFP expression was imaged with a Zeiss Axioplan compound microscope, using Nomarski optics and an attached AxioCam digital camera. Translational CEBP-1::GFP expression was imaged with a Zeiss LSM710 confocal microscope for quantitative analyses. For confocal images, z stacks were obtained and maximum projection images were created using Zeiss Zen 2012 software. ImageJ was used to measure the GFP intensity at nerve ring area for *cebp-1p::CEBP-1::GFP* and at each nucleus (10 per animals) for *col-154p::CEBP-1::GFP*. GFP::NIPI-3 KI expression was imaged with a spinning disk confocal microscope as described [54] to improve the signal-to-noise ratio.

We cut PLM axons and quantified the length of regrown axons as described [55].

CEBP-1 ChIP-seq analysis. We generated transgenic animals expressing a functional FLAG-tagged CEBP-1 protein in a *cebp-1(tm2807)* mutant background (*cebp-1(0); juIs418 [cebp-1p::FLAG::CEBP-1::cebp-1 3'UTR]*) (Additional file 10: Figure S10) and then immunoprecipitated FLAG-CEBP-1-associated DNA fragments using anti-FLAG antibodies (M2 anti-FLAG magnetic beads; Sigma). We collected mixed stage worms grown at 20˚C on NGM plates followed by 2% formaldehyde and sonicated the samples as described [56]. We next generated ChIP-seq DNA libraries via ligating DNA to specific adaptors and amplification with barcode primers, then sequenced them on the Illumina HiSeq-2000 platform. We performed two independent ChIP-seq experiments, with parallel genomic DNA controls prepared from the same strain. We conducted peak-calling using CLC genomics workbench 6.0 (CLCbio). To define genes associated with the peaks, we used the annotation of transcription start site (TSS) and transcription end site (TES) from WS220 and annotated the peak if it overlapped with the gene or the 3 kb upstream of the TSS. We then manually confirmed the peaks and associated genes using the UCSC browser and update to WS252. If the peak was found within the promoter for one isoform and introns for other isoforms, we categorized it as a peak within a promoter. The ChIP-seq data are available at the Gene Expression Omnibus under the accession number GSE83330.

Bioinformatic analyses. MEME [33] and RSAT [34] were used to identify over-represented motifs from 209 CEBP-1 ChIP-seq peak sequences and then Tomtom [35] was used to compare the most over-represented motif against a database of known motifs in vertebrates. All informatics tools can be found at: http://meme-suite.org and http://www.rsat.eu. Enrichment analyses were run on a newly developed database of functional annotations including 4,600 datasets [39] updated to WS252 http://www.wormbase.org. Classes considered for enrichment had a maximum size of 2,000 genes and the *P*-value for enrichment was lower than 10^{-5}.

Yeast two-hybrid assay. Full length or fragment cDNAs were cloned into the pACT2 (Gal4 activation domain) or pBTM116 (LexA DNA-binding domain) vectors (Clontech) and constructs were co-transformed into yeast strain L40. We grew transformed yeasts on agar plates with SD medium (synthetic minimal medium) lacking leucine and tryptophan; interactions were examined on plates with SD medium lacking leucine, tryptophan, and histidine.

Statistical Analysis. Statistical analysis was performed using GraphPad Prism 5. Significance was determined using unpaired *t*-tests for two samples, one-way ANOVA followed by Tukey's multiple comparison tests for multiple samples. For two nominal variables, Fisher's exact test was used to evaluate the statistical significance. *P* < 0.05 (*) was considered statistically significant. * *P* < 0.05; ** *P* < 0.01; *** *P* < 0.001.

References

1. Murphy JM, Nakatani Y, Jamieson SA, Dai W, Lucet IS, Mace PD: **Molecular Mechanism of CCAAT-Enhancer Binding Protein Recruitment by the TRIB1 Pseudokinase**. *Structure* 2015, **23**(11):2111-2121.
2. Lohan F, Keeshan K: **The functionally diverse roles of tribbles**. *Biochem Soc Trans* 2013, **41**(4):1096-1100.
3. Taylor SS, Shaw A, Hu J, Meharena HS, Kornev A: **Pseudokinases from a structural perspective**. *Biochem Soc Trans* 2013, **41**(4):981-986.
4. Dobens LL, Jr., Bouyain S: **Developmental roles of tribbles protein family members**. *Dev Dyn* 2012, **241**(8):1239-1248.
5. Seher TC, Leptin M: **Tribbles, a cell-cycle brake that coordinates proliferation and morphogenesis during *Drosophila* gastrulation**. *Curr Biol* 2000, **10**(11):623-629.
6. Mata J, Curado S, Ephrussi A, Rørth P: **Tribbles coordinates mitosis and morphogenesis in *Drosophila* by regulating string/CDC25 proteolysis**. *Cell* 2000, **101**(5):511-522.
7. Rørth P, Szabo K, Texido G: **The level of C/EBP protein is critical for cell migration during *Drosophila* oogenesis and is tightly controlled by regulated degradation**. *Mol Cell* 2000, **6**(1):23-30.
8. Grosshans J, Wieschaus E: **A genetic link between morphogenesis and cell division during formation of the ventral furrow in *Drosophila***. *Cell* 2000, **101**(5):523-531.
9. Yokoyama T, Nakamura T: **Tribbles in disease: Signaling pathways important for cellular function and neoplastic transformation**. *Cancer Sci* 2011, **102**(6):1115-1122.
10. Sung HY, Guan H, Czibula A, King AR, Eder K, Heath E, Suvarna SK, Dower SK, Wilson AG, Francis SE *et al*: **Human tribbles-1 controls proliferation and chemotaxis of smooth muscle cells via MAPK signaling pathways**. *J Biol Chem* 2007, **282**(25):18379-18387.
11. Yokoyama T, Kanno Y, Yamazaki Y, Takahara T, Miyata S, Nakamura T: **Trib1 links the MEK1/ERK pathway in myeloid leukemogenesis**. *Blood* 2010, **116**(15):2768-2775.
12. Eder K, Guan H, Sung HY, Ward J, Angyal A, Janas M, Sarmay G, Duda E, Turner M, Dower SK *et al*: **Tribbles-2 is a novel regulator of inflammatory activation of monocytes**. *Int Immunol* 2008, **20**(12):1543-1550.
13. Masoner V, Das R, Pence L, Anand G, LaFerriere H, Zars T, Bouyain S, Dobens LL: **The kinase domain of Drosophila Tribbles is required for turnover of fly C/EBP during cell migration**. *Dev Biol* 2013, **375**(1):33-44.
14. Satoh T, Kidoya H, Naito H, Yamamoto M, Takemura N, Nakagawa K, Yoshioka Y, Morii E, Takakura N, Takeuchi O *et al*: **Critical role of Trib1 in differentiation of tissue-resident M2-like macrophages**. *Nature* 2013, **495**(7442):524-528.

15. Dedhia PH, Keeshan K, Uljon S, Xu L, Vega ME, Shestova O, Zaks-Zilberman M, Romany C, Blacklow SC, Pear WS: **Differential ability of Tribbles family members to promote degradation of C/EBPalpha and induce acute myelogenous leukemia**. *Blood* 2010, **116**(8):1321-1328.

16. Ohoka N, Hattori T, Kitagawa M, Onozaki K, Hayashi H: **Critical and functional regulation of CHOP (C/EBP homologous protein) through the N-terminal portion**. *J Biol Chem* 2007, **282**(49):35687-35694.

17. Ohoka N, Yoshii S, Hattori T, Onozaki K, Hayashi H: **TRB3, a novel ER stress-inducible gene, is induced via ATF4-CHOP pathway and is involved in cell death**. *EMBO J* 2005, **24**(6):1243-1255.

18. Ord D, Meerits K, Ord T: **TRB3 protects cells against the growth inhibitory and cytotoxic effect of ATF4**. *Exp Cell Res* 2007, **313**(16):3556-3567.

19. Jousse C, Deval C, Maurin AC, Parry L, Cherasse Y, Chaveroux C, Lefloch R, Lenormand P, Bruhat A, Fafournoux P: **TRB3 inhibits the transcriptional activation of stress-regulated genes by a negative feedback on the ATF4 pathway**. *J Biol Chem* 2007, **282**(21):15851-15861.

20. Pujol N, Cypowyj S, Ziegler K, Millet A, Astrain A, Goncharov A, Jin Y, Chisholm AD, Ewbank JJ: **Distinct innate immune responses to infection and wounding in the *C. elegans* epidermis**. *Curr Biol* 2008, **18**(7):481-489.

21. Ziegler K, Kurz CL, Cypowyj S, Couillault C, Pophillat M, Pujol N, Ewbank JJ: **Antifungal innate immunity in *C. elegans*: PKCdelta links G protein signaling and a conserved p38 MAPK cascade**. *Cell Host Microbe* 2009, **5**(4):341-352.

22. Yan D, Wu Z, Chisholm AD, Jin Y: **The DLK-1 kinase promotes mRNA stability and local translation in *C. elegans* synapses and axon regeneration**. *Cell* 2009, **138**(5):1005-1018.

23. Bounoutas A, Kratz J, Emtage L, Ma C, Nguyen KC, Chalfie M: **Microtubule depolymerization in *Caenorhabditis elegans* touch receptor neurons reduces gene expression through a p38 MAPK pathway**. *Proc Natl Acad Sci U S A* 2011, **108**(10):3982-3987.

24. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA *et al*: **Multiplex genome engineering using CRISPR/Cas systems**. *Science* 2013, **339**(6121):819-823.

25. Friedland AE, Tzur YB, Esvelt KM, Colaiacovo MP, Church GM, Calarco JA: **Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system**. *Nat Methods* 2013, **10**(8):741-743.

26. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM: **RNA-guided human genome engineering via Cas9**. *Science* 2013, **339**(6121):823-826.

27. Xu S: **The application of CRISPR-Cas9 genome editing in *Caenorhabditis elegan*s**. *J Genet Genomics* 2015, **42**(8):413-421.

28. Kelly WG, Xu S, Montgomery MK, Fire A: **Distinct requirements for somatic and germline expression of a generally expressed *Caernorhabditis elegans* gene**. *Genetics* 1997, **146**(1):227-238.

29. Yamamoto M, Uematsu S, Okamoto T, Matsuura Y, Sato S, Kumar H, Satoh T, Saitoh T, Takeda K, Ishii KJ *et al*: **Enhanced TLR-mediated NF-IL6 dependent gene expression by Trib1 deficiency**. *J Exp Med* 2007, **204**(9):2233-2239.

30. Dierking K, Polanowska J, Omi S, Engelmann I, Gut M, Lembo F, Ewbank JJ, Pujol N: **Unusual regulation of a STAT protein by an SLC6 family transporter in *C. elegans* epidermal innate immunity**. *Cell Host Microbe* 2011, **9**(5):425-435.

31. Hammarlund M, Nix P, Hauth L, Jorgensen EM, Bastiani M: **Axon regeneration requires a conserved MAP kinase pathway**. *Science* 2009, **323**(5915):802-806.

32. Nix P, Hisamoto N, Matsumoto K, Bastiani M: **Axon regeneration requires coordinate activation of p38 and JNK MAPK pathways**. *Proc Natl Acad Sci U S A* 2011, **108**(26):10738-10743.

33. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202-208.

34. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C *et al*: **RSAT 2015: Regulatory Sequence Analysis Tools**. *Nucleic Acids Res* 2015, **43**(W1):W50-56.

35. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs**. *Genome Biol* 2007, **8**(2):R24.

36. Vinson CR, Sigler PB, McKnight SL: **Scissors-grip model for DNA recognition by a family of leucine zipper proteins**. *Science* 1989, **246**(4932):911-916.

37. Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE**. *Genome Biol* 2003, **4**(10):R70.

38. Engelmann I, Griffon A, Tichit L, Montanana-Sanchis F, Wang G, Reinke V, Waterston RH, Hillier LW, Ewbank JJ: **A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans***. *PLoS One* 2011, **6**(5):e19055.

39. Zugasti O, Thakur N, Belougne J, Squiban B, Kurz CL, Soule J, Omi S, Tichit L, Pujol N, Ewbank JJ: **A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes**. *BMC Biol* 2016, **14**(1):35.

40. Andrusiak MG, Jin Y: **Context Specificity of Stress-activated Mitogen-activated Protein (MAP) Kinase Signaling: The Story as Told by *Caenorhabditis elegans***. *J Biol Chem* 2016, **291**(15):7796-7804.

41. Kim DH, Ewbank JJ: **Signaling in the innate immune response**. *WormBook* 2015:1-51.

42. Richardson CE, Kinkel S, Kim DH: **Physiological IRE-1-XBP-1 and PEK-1 signaling in *Caenorhabditis elegans* larval development and immunity**. *PLoS Genet* 2011, **7**(11):e1002391.

43. Pukkila-Worley R, Feinbaum R, Kirienko NV, Larkins-Ford J, Conery AL, Ausubel FM: **Stimulation of host immune defenses by a small molecule protects C. elegans from bacterial infection**. *PLoS Genet* 2012, **8**(6):e1002733.

44. Cheesman HK, Feinbaum RL, Thekkiniath J, Dowen RH, Conery AL, Pukkila-Worley R: **Aberrant Activation of p38 MAP Kinase-Dependent Innate Immune Responses Is Toxic to *Caenorhabditis elegans***. *G3 (Bethesda)* 2016, **6**(3):541-549.

45. Kim KW, Jin Y: **Neuronal responses to stress and injury in *C. elegans***. *FEBS Lett* 2015, **589**(14):1644-1652.

46. Brenner S: **The genetics of *Caenorhabditis elegans***. *Genetics* 1974, **77**(1):71-94.

47. Mello CC, Kramer JM, Stinchcomb D, Ambros V: **Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences**. *EMBO J* 1991, **10**(12):3959-3970.

48. Arribere JA, Bell RT, Fu BX, Artiles KL, Hartman PS, Fire AZ: **Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in *Caenorhabditis elegans***. *Genetics* 2014, **198**(3):837-846.

49. Pujol N, Zugasti O, Wong D, Couillault C, Kurz CL, Schulenburg H, Ewbank JJ: **Anti-fungal innate immunity in *C. elegans* is enhanced by evolutionary diversification of antimicrobial peptides**. *PLoS Pathog* 2008, **4**(7):e1000105.

50. Dickinson DJ, Pani AM, Heppert JK, Higgins CD, Goldstein B: **Streamlined Genome Engineering with a Self-Excising Drug Selection Cassette**. *Genetics* 2015, **200**(4):1035-1049.

51. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O: ***Caenorhabditis elegans* mutant allele identification by whole-genome sequencing**. *Nat Methods* 2008, **5**(10):865-867.

52. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

53. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al*: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome Res* 2005, **15**(10):1451-1455.

54. Rouger V, Bordet G, Couillault C, Monneret S, Mailfert S, Ewbank JJ, Pujol N, Marguet D: **Independent synchronized control and visualization of interactions between living cells and organisms**. *Biophys J* 2014, **106**(10):2096-2104.

55. Wu Z, Ghosh-Roy A, Yanik MF, Zhang JZ, Jin Y, Chisholm AD: ***Caenorhabditis elegans* neuronal regeneration is influenced by life stage, ephrin signaling, and synaptic branching**. *Proc Natl Acad Sci U S A* 2007, **104**(38):15132-15137.

56. Mukhopadhyay A, Deplancke B, Walhout AJ, Tissenbaum HA: **Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in *Caenorhabditis elegans***. *Nat Protoc* 2008, **3**(4):698-709.

Figure legends

Fig. 1. *C. elegans* Tribbles *nipi-3* is required for larval development and viability.

(a) The *nipi-3* locus. Top, *nipi-3* encodes a pseudokinase of the Tribbles family. Middle, *nipi-3* deletions generated using CRISPR-Cas9 genome editing. Bottom, the extent of the *nipi-3* genomic region used to rescue the deletion mutants. (b–e) Bright-field images of worms at 3 days post-hatching; wild type (b), *nipi-3* null mutants (c,d), and a transgenic animal (Tg) expressing the wild-type *nipi-3* genomic DNA in a *nipi-3(0)* background (e). (f) Fluorescence images of worms expressing AJM-1::GFP reporter in the epithelial cells to allow visualization of seam cells. (g) Overlaid differential interference contrast (DIC) and fluorescence image of a *nipi-3(fr148)* mutant at 3 days post-hatching expressing *lag-2p::*GFP reporter. This image is corresponding to the boxed region in Fig. 1d. Two distal tip cells (DTC) are shown in green (arrow), and the germline of arrested *nipi-3* null larva is denoted by a dotted red line. (h) Body length (µm) of worms at 3 days post-hatching. Each dot represents a single animal measured as shown; each red line represents the mean value. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests). (i) Fluorescence images of endogenous NIPI-3 expression visualized in the GFP KI strain (*fr152)*. Expression is observed in the nuclei (yellow arrows) of the epidermis (left panel), the intestine (upper right panel) and the head neurons (lower right panel).

Fig. 2. *nipi-3(0)* lethality is suppressed by loss of *cebp-1* or components of a PMK-1/p38 MAPK cascade.

(a) Schematic overview of the forward genetic screen designed to identify the suppressors of *nipi-3(0)* larval arrest and lethality. (b–g) The mutations isolated from the *nipi-3(0)* suppressor screen and the deletion null mutations tested for *nipi-3(0)* suppression assay are shown in the left column and bright-field images of worms at 3 days post-hatching are shown in the right column. (h) Body length (µm) of worms at 3 days post-hatching. Each dot represents a single animal measured as shown; each red line represents the mean value; some data are replicated from Fig. 1 as shown in darker grey dots. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests).

Fig. 3. NIPI-3 represses PMK-1 phosphorylation via *cebp-1* and represses *cebp-1* transcription.

(a) Western blot analysis on total protein lysate from various animal strains using the indicated antibodies, α-phospho-p38 MAPK antibody to detect a phosphorylated form of PMK-1 proteins (p-PMK-1) or α-actin antibody as a loading control. (b) Densitometric quantifications of immunoblot signals normalized to actin. *n* = 2; Error bars represent SEM*; *P* < 0.05 (one-way ANOVA with Tukey's post hoc tests). (c) Quantitative RT-PCR (qRT-PCR) analysis of *pmk-1*. Relative abundance of *pmk-1* mRNA normalized to *actin* mRNA. *n* = 3; Error bars represent SEM; ns, not significant (one-way ANOVA with Tukey's post hoc tests). (d–j) Fluorescence images of *cebp-1p::GFP* reporter animals at L2. (k) qRT-PCR analysis of *cebp-1* in WT, *nipi-3(fr4), nipi-3(0)* animals. Relative abundance of *cebp-1* mRNA normalized to *actin* mRNA. *n =* 3; Error bars represent SEM; **P* < 0.05; ***P* < 0.01; ns, not significant (one-way ANOVA with Tukey's post hoc tests). (l,m) Confocal fluorescence images (z-stack) of *col-154p(epidermis)::CEBP-1::GFP* reporter animals at

L2. (n) Quantification of GFP intensity measured in each epidermal nucleus (10 per each animal). *n* = 6; Error bars represent SEM; ns, not significant (Student's unpaired *t*-test). Primary data for panels **b**, **c**, and **k** are provided in Additional file 14.

Fig. 4. Overexpression of truncated forms of CEBP-1 protein suppresses *nipi-3(0)* lethality.
(a–c) Bright-field images of worms at 3 days post-hatching expressing truncated forms of CEBP-1 proteins. (d) Body length (µm) of worms at 3 days post-hatching. Each dot represents a single animal measured as shown; each red line represents the mean value; some data are replicated from Fig. 1 as shown in darker grey dots. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests). (e,f) Confocal fluorescence images (z-stack) of *cebp-1p::CEBP-1(aa 1-230)::GFP* reporter animals at L2. (g) Quantification of GFP intensity measured in the region of head neurons. Error bars represent SEM; ****P* < 0.001 (Student's unpaired *t*-test).

Fig. 5. CEBP-1 binds conserved DNA motifs in genes regulating development and stress response.
(a) Motif logo of the most over-represented motif among CEBP-1 ChIP-seq peaks. (b) qRT-PCR analysis of *sek-1*. Relative abundance of *pmk-1* mRNA normalized to *actin* mRNA. *n =* 3; Error bars represent SEM. **P* < 0.05; ns, not significant (one-way ANOVA with Tukey's post hoc tests). Primary data are provided in Additional file 14. (c) Hierarchical clustering of genes and functional classes (see Additional file 9: Figure S9 and Additional file 11: Table S1 for class labels and full data); the presence of a gene in a class is represented by a red rectangle, its absence by blue.

Fig. 6. Tight regulation of both NIPI-3 and CEBP-1 is required in multiple tissues for proper organism development.
(a–e) Bright-field images and (f) the body length of worms expressing tissue-specific *nipi-3(+)* driven by the intestinal (*mtl-2*), epidermal (*col-12*), or pan-neuronal (*rgef-1*) promoters in a *nipi-3(0)* background. (g) The body length of worms expressing tissue-specific *cebp-1(+)* driven by the intestinal (*ges-1*), epidermal (*col-154*), or pan-neuronal (*rgef-1*) promoters in a *nipi-3(0) cebp-1(0)* background. (f,g) Each dot represents a single animal measured as shown; each red line represents the mean value; some data are replicated from Figs. 1 and 2 as shown in darker grey dots. ***P* < 0.01; ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests). (h) Working model for NIPI-3 function in *C. elegans* development. In wild type, presence of NIPI-3 keeps *cebp-1* expression level optimal for coordinated tissue development. In *nipi-3(0),* however, *cebp-1* and *sek-1* are overexpressed and in turn PMK-1 is hyper-activated.

List of abbreviations

AMP: Antimicrobial peptide
bZIP: basic leucine zipper
C/EBP: CCAAT/enhancer-binding protein
ChIP-seq: chromatin immunoprecipitation and deep-sequencing

CRISPR: clustered regularly interspersed short palindromic repeats
EASE: Expression analysis systematic explorer
GO: Gene ontology
KEGG: Kyoto encyclopedia of genes and genomes
MAPK: mitogen-activated protein kinase
MEME: Multiple Em for Motif Elicitation
NIPI: No induction of peptide after *Drechmeria* infection
RSAT: Regulatory Sequence Analysis Tools

Declarations

Acknowledgements

We thank J. Ewbank for support on the project, A. D. Chisholm and members of the Jin and Chisholm laboratories for valuable discussion and critical comments on the manuscript, and D. L. McEwan and F. M. Ausubel for discussion and sharing unpublished data. We thank S. Cypowyj for the initial findings on *cebp-1*, J. Ruiz for assistance in strain construction, Z. Wu for help in axotomy analyses, M. Bulle and A. Boned for technical assistance, D. J. Dickinson and Z. Wang for CRISPR plasmids, H-R. Li and Y. Zhou in X-D. Fu's lab at UC San Diego for help in Illumina sequencing and advice on mapping ChIP-seq reads, and WormBase as an information resource. Y.J. gratefully acknowledges the support from IMéRA, Aix-Marseille University. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440), and by National Bioresource Project of Japan (NBRP).

Funding

This work was supported by institutional grants from AMU, INSERM and CNRS, the ANR (ANR-12-BSV3-0001-01, ANR-11-LABX-0054 (Investissements d'Avenir–Labex INFORM) and ANR-11-IDEX-0001-02 (Investissements d'Avenir–A*MIDEX) and ANR-10-INBS-04-01 (France Bio Imaging), NIH grant R01 NS035546 to Y.J., and American Heart Association postdoctoral fellowship to K.W.K.. Y.J. is an investigator of the Howard Hughes medical Institute.

Authors' contributions

K.W.K. and Y.J. designed the genetic suppressor screen, K.W.K and C.A.P. performed the screen. K.W.K. performed all experiments on *nipi-3* suppressors and CEBP-1 ChIP-seq analyses. S.O. and J.P. performed experiments and N.T. the bioinformatic analyses. Y.J. and N.P. supervised the experiments. K.W.K., Y.J. and N.P. analysed the data and wrote the manuscript.

Availability of data and material

All data generated or analyzed during this study are included in this published article (and its supplementary information files). The ChIP-seq data have been deposited in the Gene Expression Omnibus database under the accession number GSE83330 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=avylwwgwvtmrlmz&acc=GSE83330). Requests for material should be made to the corresponding authors.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Additional files

Additional file 1: Figure S1. CEBP-1 expression in multiple tissues causes an abnormal pharyngeal morphology in *nipi-3(0) cebp-1(0)* animals. (a) DIC images of worms at 3 days post-hatching. Pharynx and intestine are denoted by a dotted red and yellow line, respectively. (b) Co-expression of *cebp-1(+)* in the epidermis and neurons in *nipi-3(0) cebp-1(0)* animals caused pharyngeal morphology defect. **P* < 0.05; ***P* < 0.01; ns, not significant (two-tailed Fisher's exact test).

Additional file 2: Figure S2. NIPI-3 interacts with CEBP-1 in yeast two-hybrid assay. CEBP-1 variants were fused to the Gal4 activation domain and tested for their interaction with full-length NIPI-3 fused to the LexA DNA-binding domain. A large CEBP-1 fragment (Δ2, amino acids 1-115) was sufficient for the NIPI-3 interaction. "–" refers to no growth and "+" refers to growth in the absence of histidine.

Additional file 3: Figure S3. Loss of *cebp-1* rescues the *nipi-3(fr4)* phenotype. (a) Bright-field images and (b) body length of worms at 3 days post-hatching grown at 25˚C. (b) Each dot represents a single animal measured as shown; each red line represents the mean value. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests).

Additional file 4: Figure S4. *cebp-1* and *nipi-3* are dispensable in immune response and axon regeneration, respectively. (a) qRT-PCR analysis of AMP gene, *nlp-34.* Relative abundance of *nlp-34* mRNA normalized to *actin* mRNA. *n* = 3; Error bars represent SEM. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests). Primary data are provided in Additional file 14. (b) Axotomy analysis. Normalized PLM axon regrowth is shown in the bar graph. Error bars represent SEM. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests).

Additional file 5: Figure S5. Quantification of the body length of suppressor alleles identified from the screen and loss-of-function mutants of *tpa-1, pmk-1, sta-2, dlk-1, pmk-3, mlk-1* and *kgb-1*. (a,b) Body length of worms at 3 days post-hatching. Each dot represents a single animal measured as shown; each red line represents the mean value; some data are replicated from Fig. 1 as shown in darker grey dots. ****P* < 0.001 (one-way ANOVA with Tukey's post hoc tests).

Additional file 6: Figure S6. Phosphorylated PMK-1 levels are unchanged in *nipi-3(0); mak-2(0)* animals. (a) Western blot analysis on total protein lysate from various animal strains using the indicated antibodies, α-phospho-p38 MAPK antibody to detect a phosphorylated form of PMK-1 proteins (p-PMK-1) or α-actin antibody as a loading control. (b) Densitometric quantifications of immunoblot signals normalized to actin. $n = 4$; Error bars represent SEM*;* ns, not significant (Student's paired *t*-test). Primary data are provided in Additional file 14.

Additional file 7: Figure S7. *cebp-1* shows a dosage sensitive effect in *nipi-3(0)* mutants*.* (a) Bright-field images and (b) body length of worms at 3 days post-hatching. (b) Each dot represents a single animal measured as shown; each red line represents the mean value. ****P* < 0.001 (one-way ANOVA with Tukey's post hoc tests).

Additional file 8: Figure S8. The promoter of *sek-1* contains a ChIP-seq peak of CEBP-1*.* The promoter region of *sek-1* contains two consensus DNA-binding motifs for CEBP-1 (black triangles). Top, the *sek-1* locus. Middle, sequencing reads from CEBP-1-IP. Bottom, sequencing reads from genomic DNA input.

Additional file 9: Figure S9. Hierarchical clustering of genes and functional classes. The presence of a gene in a class is represented by a red rectangle, its absence in blue. See Additional file : Table S1 for class labels and full data.

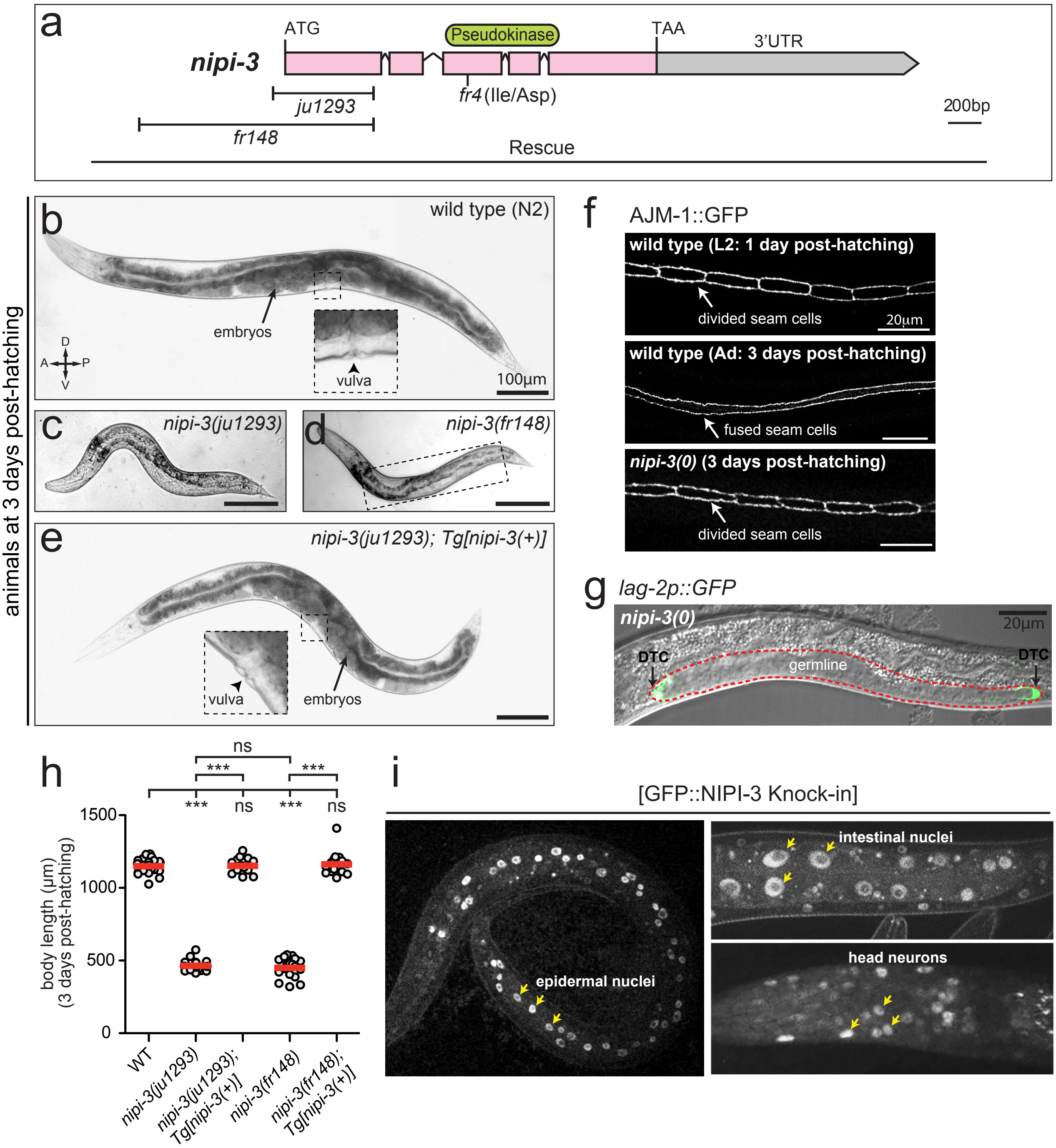
Additional file 10: Figure S10. FLAG-tagged CEBP-1 protein rescues PLM axon regeneration defects of *cebp-1(0).* Axotomy analysis. Normalized PLM axon regrowth is shown in the bar graph. Error bars represent SEM. ****P* < 0.001; ns, not significant (one-way ANOVA with Tukey's post hoc tests).

Additional file 11: Table S1. A list of CEBP-1 target genes and hierarchical clustering of genes and functional classes.

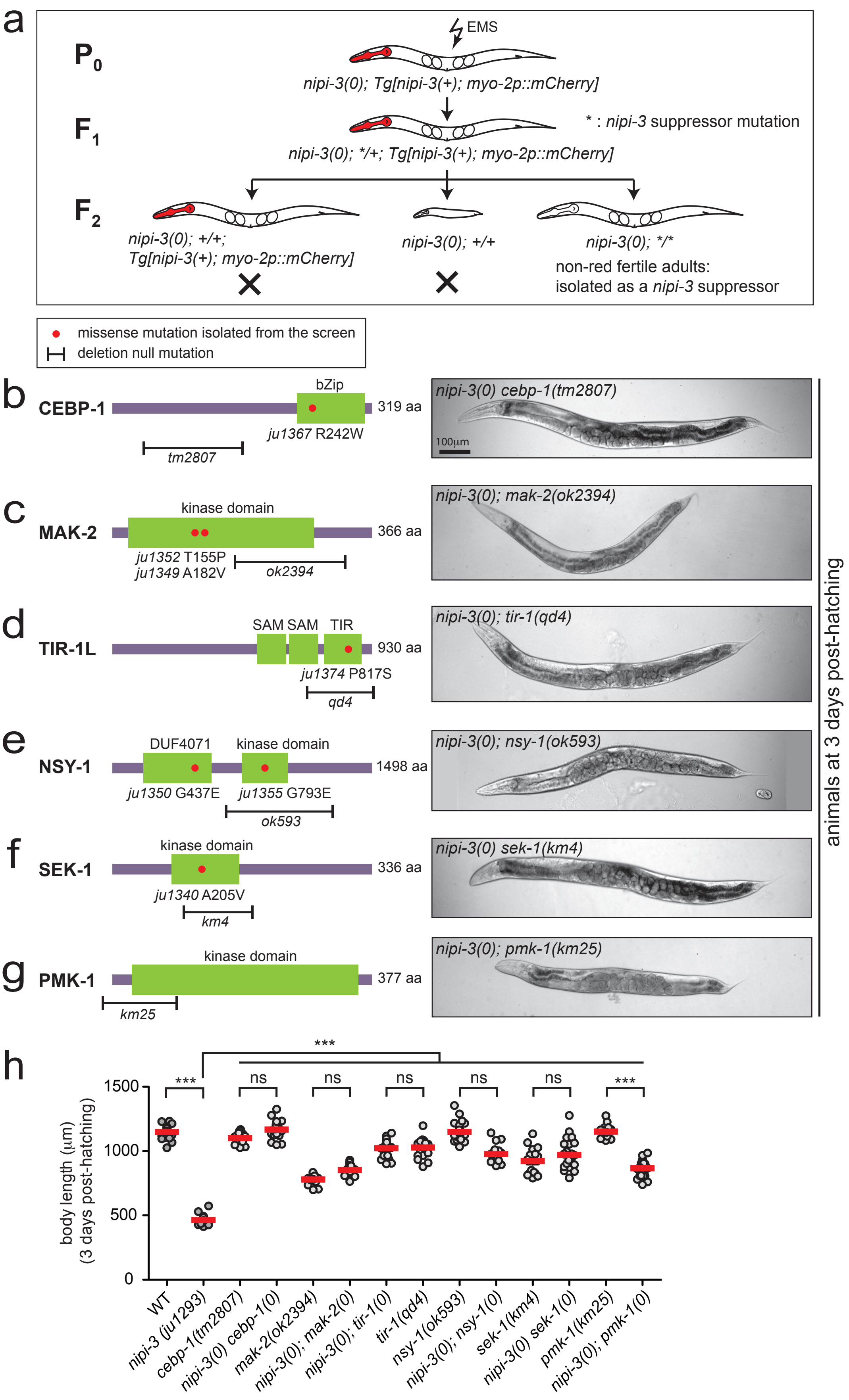
Additional file 12: Table S2. A list of strains and alleles.

Additional file 13: Table S3. A list of plasmids and primers used for qRT-PCR.

Additional file 14: Primary data for figures 3b, 3c, 3k, 5b, S4a, and S6b.

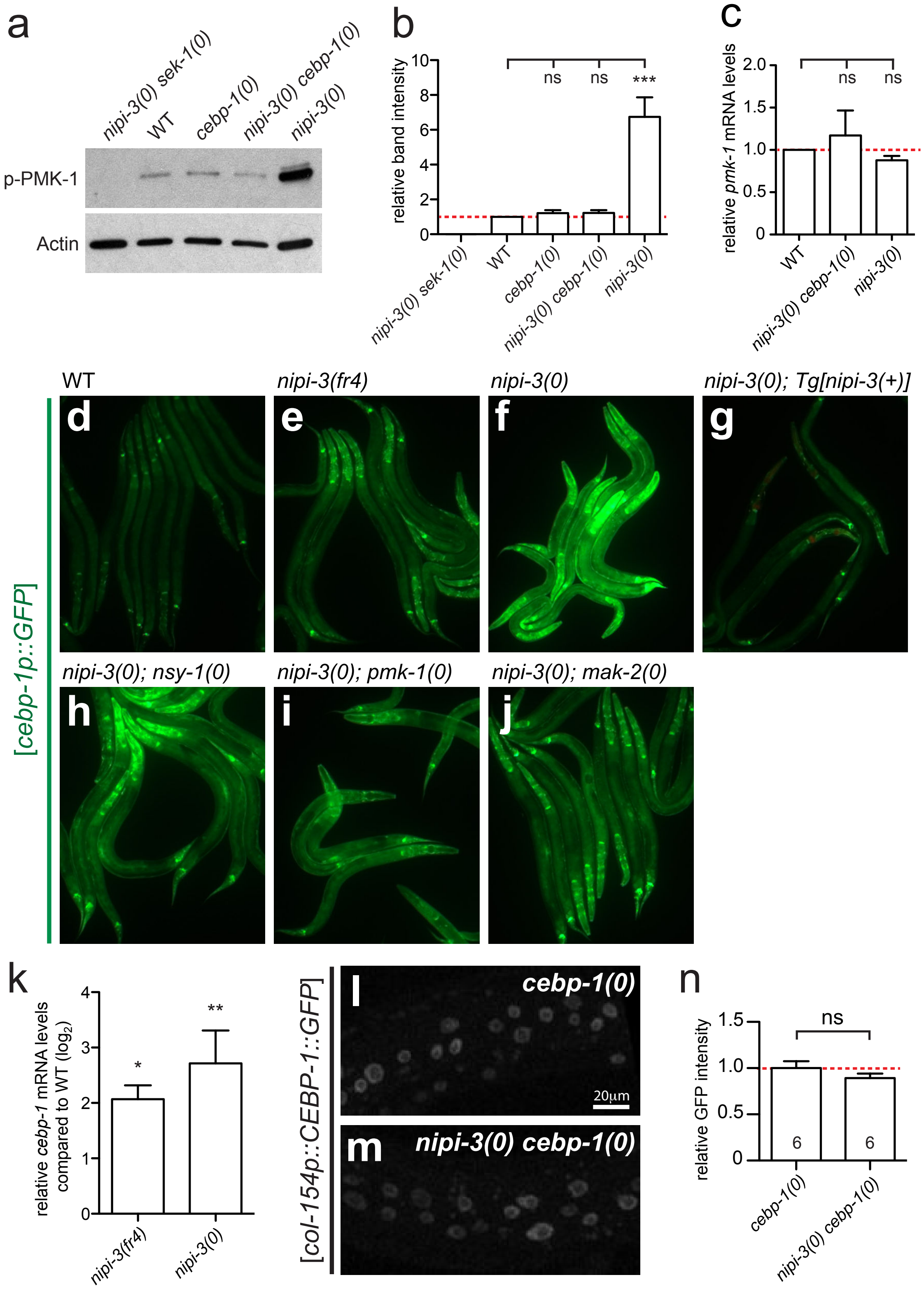

a
nipi-3
ATG
Pseudokinase
TAA
3'UTR
ju1293
fr4 (Ile/Asp)
200bp
fr148
Rescue
animals at 3 days post-hatching
b
wild type (N2)
embryos
vulva
D
A
P
V
100µm
c
nipi-3(ju1293)
d
nipi-3(fr148)
e
nipi-3(ju1293); Tg[nipi-3(+)]
vulva
embryos
f
AJM-1::GFP
wild type (L2: 1 day post-hatching)
divided seam cells
20µm
wild type (Ad: 3 days post-hatching)
fused seam cells
nipi-3(0) (3 days post-hatching)
divided seam cells
g
lag-2p::GFP
nipi-3(0)
20µm
DTC
germline
DTC
h
ns
***
***
***
ns
***
ns
body length (µm)
(3 days post-hatching)
1500
1000
500
0
WT
nipi-3(ju1293)
nipi-3(ju1293);
Tg[nipi-3(+)]
nipi-3(fr148)
nipi-3(fr148);
Tg[nipi-3(+)]
i
[GFP::NIPI-3 Knock-in]
epidermal nuclei
intestinal nuclei
head neurons

Kim_Figure 2



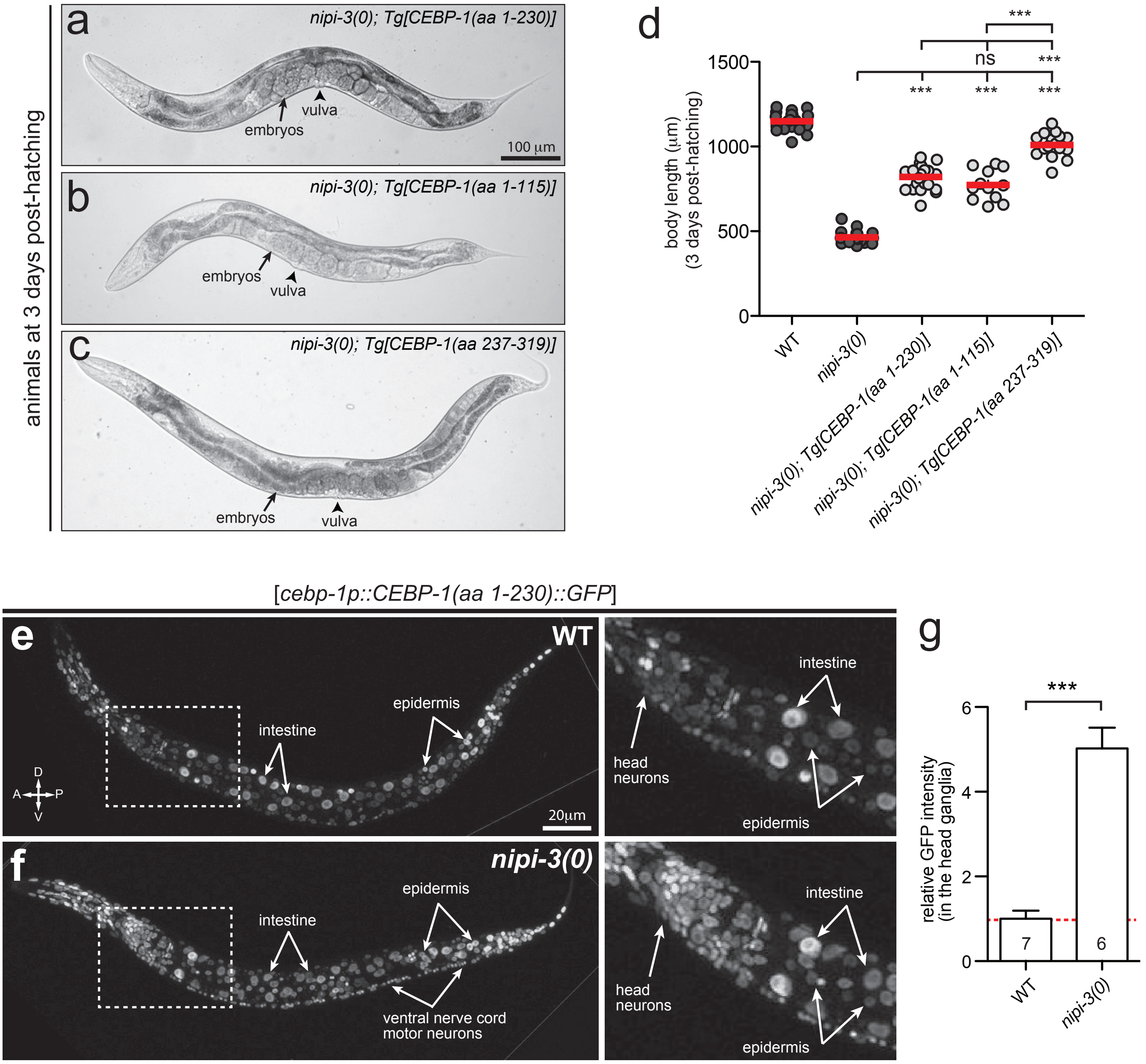
a
P0
EMS
nipi-3(0); Tg[nipi-3(+); myo-2p::mCherry]
F1
* : nipi-3 suppressor mutation
nipi-3(0); */+; Tg[nipi-3(+); myo-2p::mCherry]
F2
nipi-3(0); +/+;
Tg[nipi-3(+); myo-2p::mCherry]
nipi-3(0); +/+
nipi-3(0); */*
non-red fertile adults:
isolated as a nipi-3 suppressor
● missense mutation isolated from the screen
deletion null mutation
b
CEBP-1
bZip
319 aa
ju1367 R242W
tm2807
nipi-3(0) cebp-1(tm2807)
100µm
c
MAK-2
kinase domain
366 aa
ju1352 T155P
ju1349 A182V
ok2394
nipi-3(0); mak-2(ok2394)
d
TIR-1L
SAM SAM TIR
930 aa
ju1374 P817S
qd4
nipi-3(0); tir-1(qd4)
e
NSY-1
DUF4071
kinase domain
1498 aa
ju1350 G437E
ju1355 G793E
ok593
nipi-3(0); nsy-1(ok593)
f
SEK-1
kinase domain
336 aa
ju1340 A205V
km4
nipi-3(0) sek-1(km4)
g
PMK-1
kinase domain
377 aa
km25
nipi-3(0); pmk-1(km25)
animals at 3 days post-hatching
h
***
***
ns
ns
ns
ns
ns
***
1500
1000
500
0
body length (µm)
(3 days post-hatching)
WT
nipi-3 (ju1293)
cebp-1(tm2807)
nipi-3(0) cebp-1(0)
mak-2(ok2394)
nipi-3(0); mak-2(0)
nipi-3(0); tir-1(0)
tir-1(qd4)
nsy-1(ok593)
nipi-3(0); nsy-1(0)
sek-1(km4)
nipi-3(0) sek-1(0)
pmk-1(km25)
nipi-3(0); pmk-1(0)

Kim_Figure 3

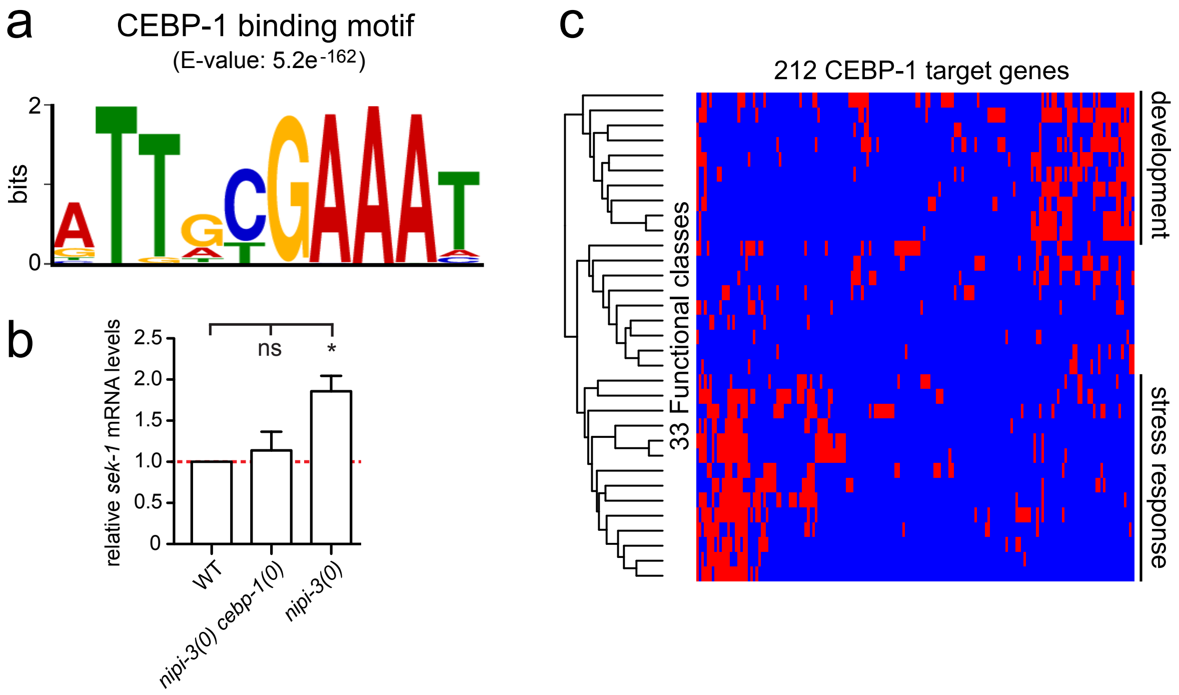


a
nipi-3(0) sek-1(0)
WT
cebp-1(0)
nipi-3(0) cebp-1(0)
nipi-3(0)
p-PMK-1
Actin
b
relative band intensity
0
2
4
6
8
10
ns
ns
***
nipi-3(0) sek-1(0)
WT
cebp-1(0)
nipi-3(0) cebp-1(0)
nipi-3(0)
c
relative pmk-1 mRNA levels
0
0.5
1.0
1.5
2.0
ns
ns
WT
nipi-3(0) cebp-1(0)
nipi-3(0)
[cebp-1p::GFP]
WT
d
nipi-3(fr4)
e
nipi-3(0)
f
nipi-3(0); Tg[nipi-3(+)]
g
nipi-3(0); nsy-1(0)
h
nipi-3(0); pmk-1(0)
i
nipi-3(0); mak-2(0)
j
k
relative cebp-1 mRNA levels compared to WT (log_2)
0
1
2
3
4
*
**
nipi-3(fr4)
nipi-3(0)
[col-154p::CEBP-1::GFP]
l
cebp-1(0)
20μm
m
nipi-3(0) cebp-1(0)
n
relative GFP intensity
0
0.5
1.0
1.5
ns
6
6
cebp-1(0)
nipi-3(0) cebp-1(0)

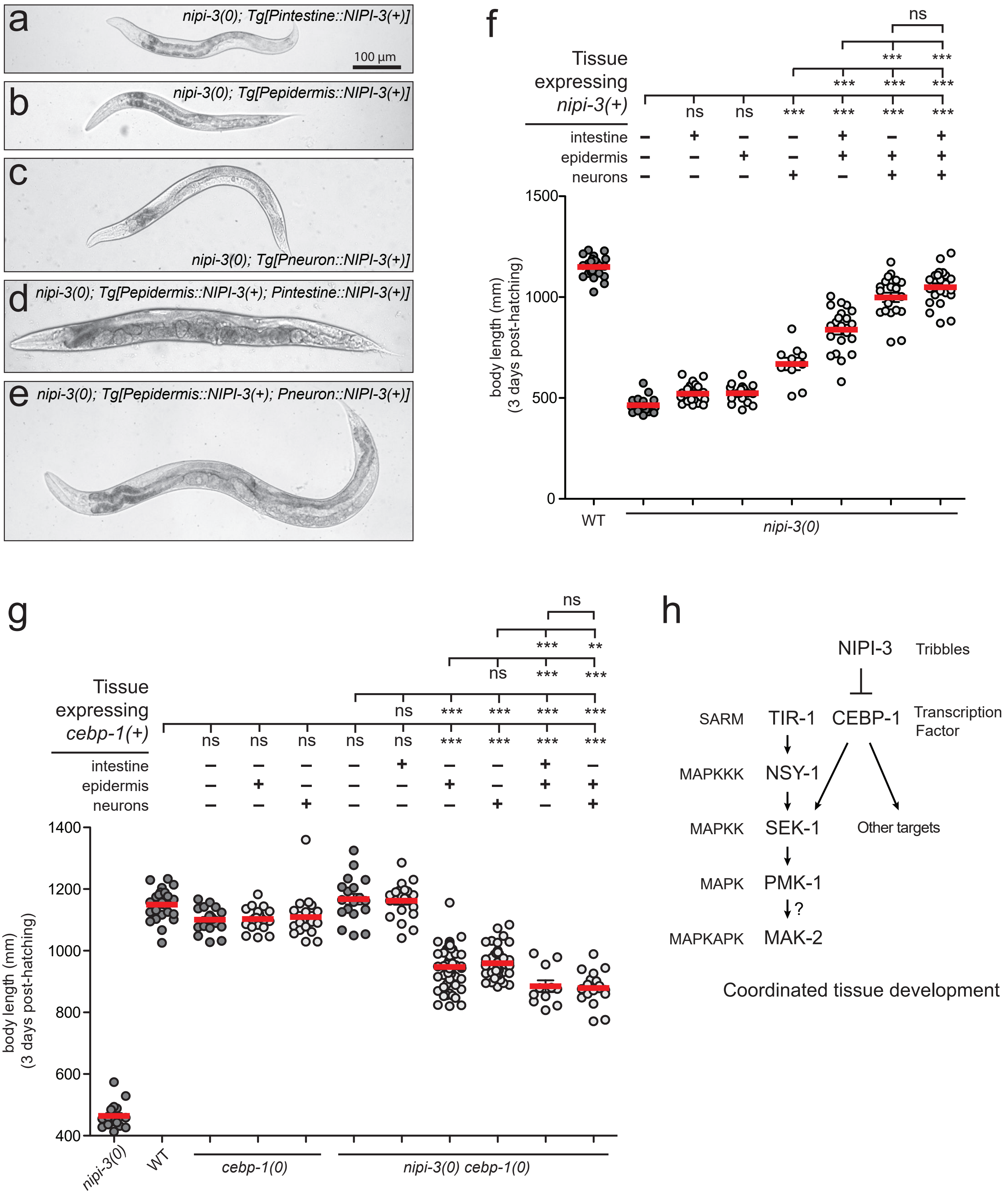
Kim_Figure 4


animals at 3 days post-hatching
a
nipi-3(0); Tg[CEBP-1(aa 1-230)]
embryos
vulva
100 µm
b
nipi-3(0); Tg[CEBP-1(aa 1-115)]
embryos
vulva
c
nipi-3(0); Tg[CEBP-1(aa 237-319)]
embryos
vulva
d
body length (µm)
(3 days post-hatching)
1500
1000
500
0
***
ns
***
***
***
***
WT
nipi-3(0)
nipi-3(0); Tg[CEBP-1(aa 1-230)]
nipi-3(0); Tg[CEBP-1(aa 1-115)]
nipi-3(0); Tg[CEBP-1(aa 237-319)]
[cebp-1p::CEBP-1(aa 1-230)::GFP]
e
WT
intestine
epidermis
D
A
P
V
20µm
head neurons
intestine
epidermis
f
nipi-3(0)
epidermis
intestine
ventral nerve cord motor neurons
head neurons
intestine
epidermis
g
relative GFP intensity
(in the head ganglia)
6
4
2
0
***
7
6
WT
nipi-3(0)

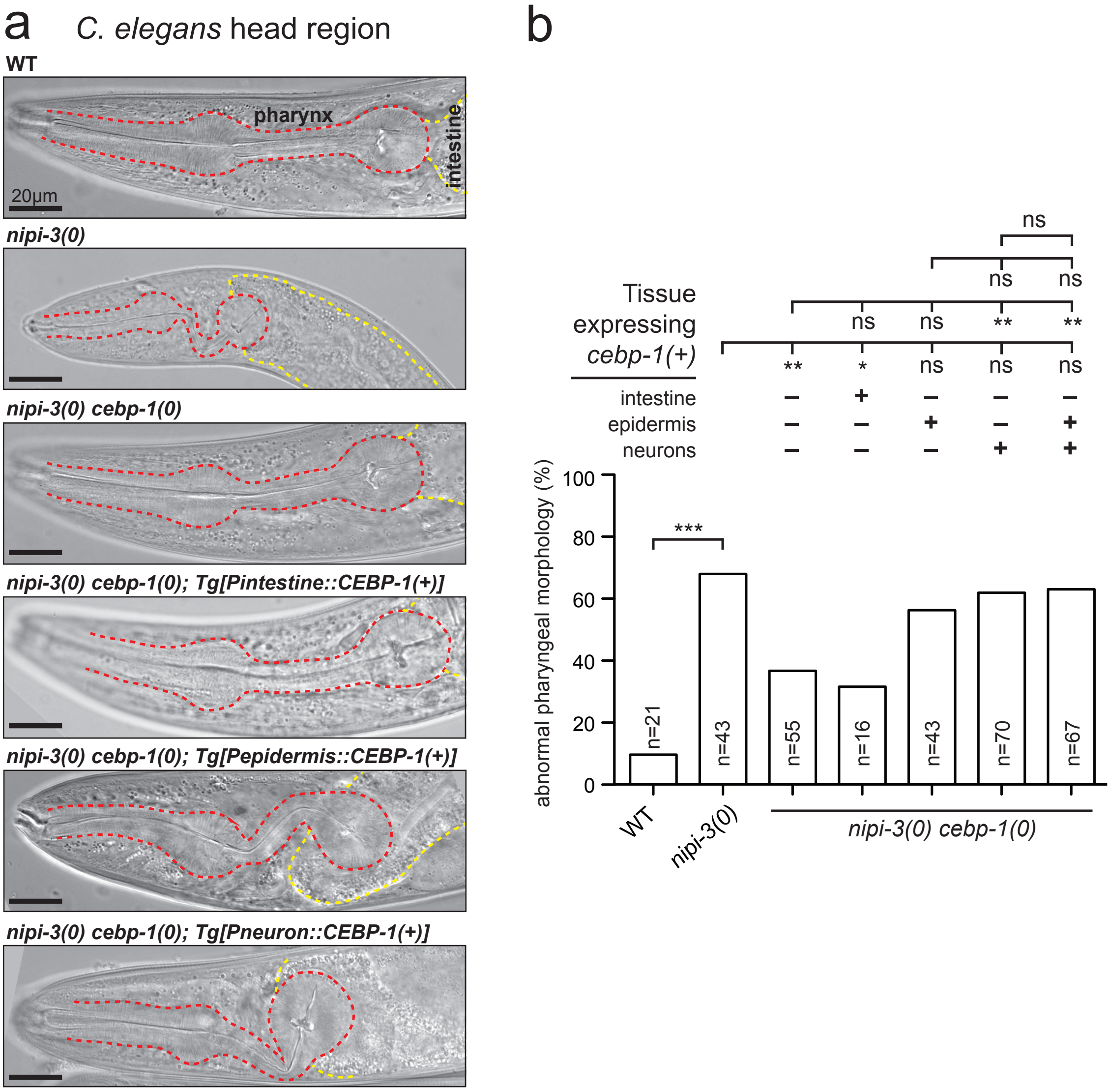
Kim_Figure 5


a
CEBP-1 binding motif
(E-value: 5.2e^{-162})
bits
2
1
0
b
relative *sek-1* mRNA levels
2.5
2.0
1.5
1.0
0.5
0
ns
*
WT
nipi-3(0) cebp-1(0)
nipi-3(0)
c
212 CEBP-1 target genes
33 Functional classes
development
stress response

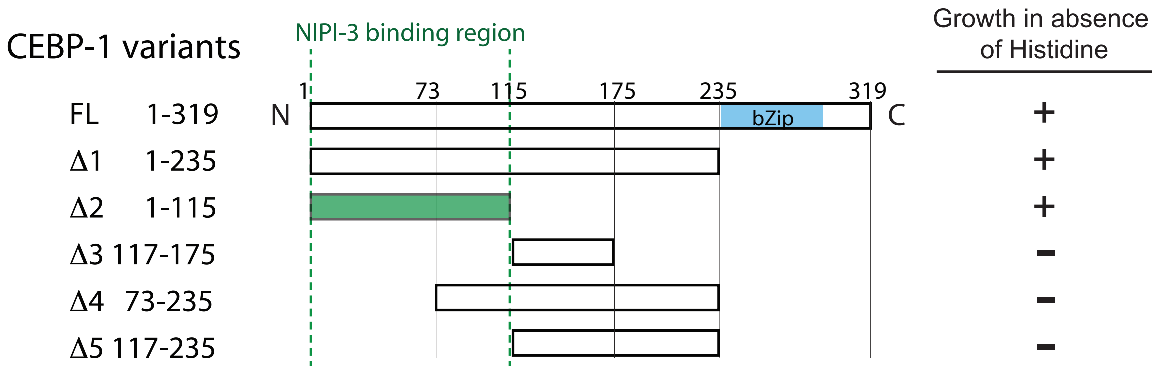
Kim_Figure 6


a
nipi-3(0); Tg[Pintestine::NIPI-3(+)]
100 µm
b
nipi-3(0); Tg[Pepidermis::NIPI-3(+)]
c
nipi-3(0); Tg[Pneuron::NIPI-3(+)]
d
nipi-3(0); Tg[Pepidermis::NIPI-3(+); Pintestine::NIPI-3(+)]
e
nipi-3(0); Tg[Pepidermis::NIPI-3(+); Pneuron::NIPI-3(+)]
f
Tissue expressing nipi-3(+)
intestine – + – – + – +
epidermis – – + – + + +
neurons – – – + – + +
ns ns *** *** *** ***
*** *** ***
*** ***
ns
body length (mm) (3 days post-hatching)
0
500
1000
1500
WT
nipi-3(0)
g
Tissue expressing cebp-1(+)
intestine – – – – + – – + –
epidermis – + – – – + – + +
neurons – – + – – – + – +
ns ns ns ns ns *** *** *** ***
ns *** *** *** ***
ns *** ***
*** **
ns
body length (mm) (3 days post-hatching)
400
600
800
1000
1200
1400
nipi-3(0)
WT
cebp-1(0)
nipi-3(0) cebp-1(0)
h
NIPI-3
Tribbles
SARM
TIR-1
CEBP-1
Transcription Factor
MAPKKK
NSY-1
MAPKK
SEK-1
Other targets
MAPK
PMK-1
?
MAPKAPK
MAK-2
Coordinated tissue development

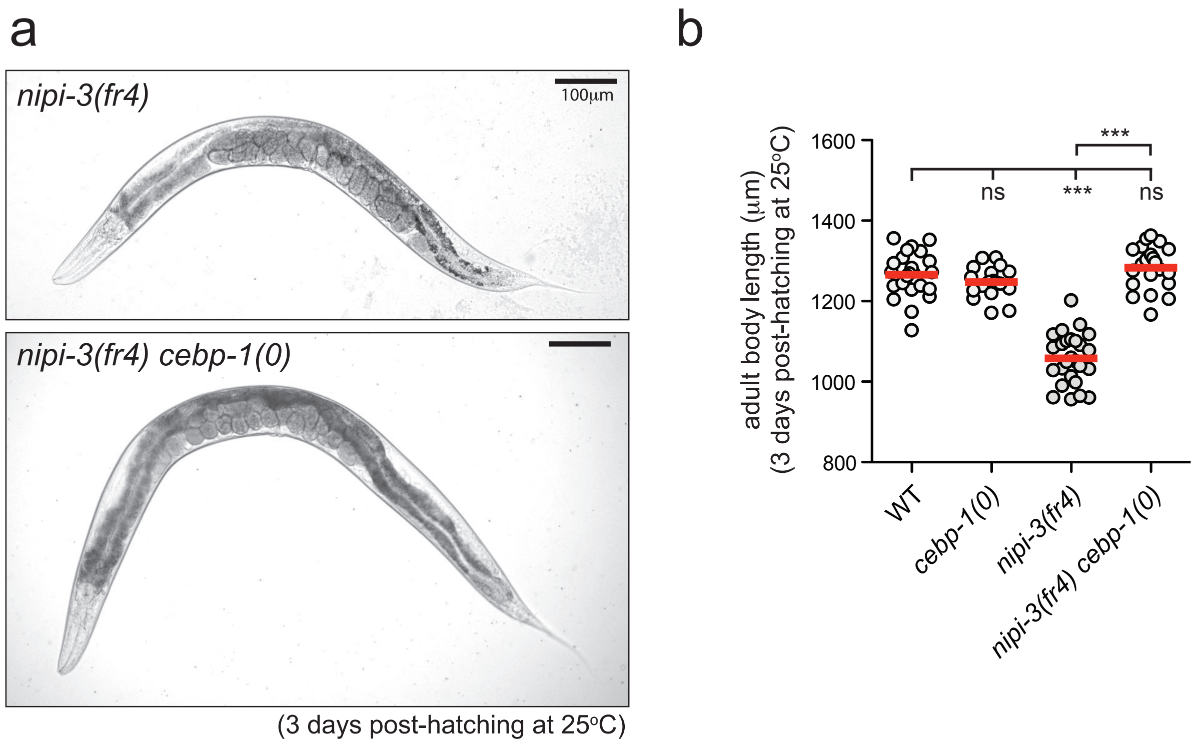
Kim_Supplementary Figure 1



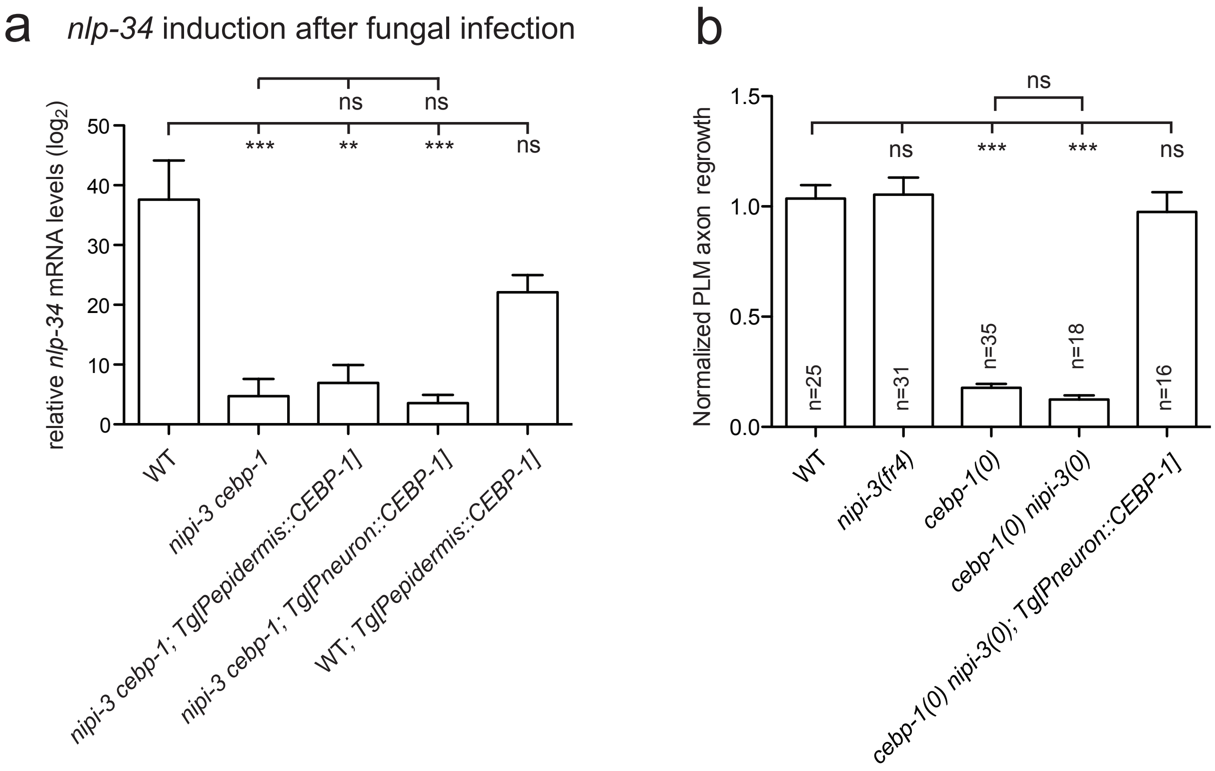
Kim_Supplementary Figure 2



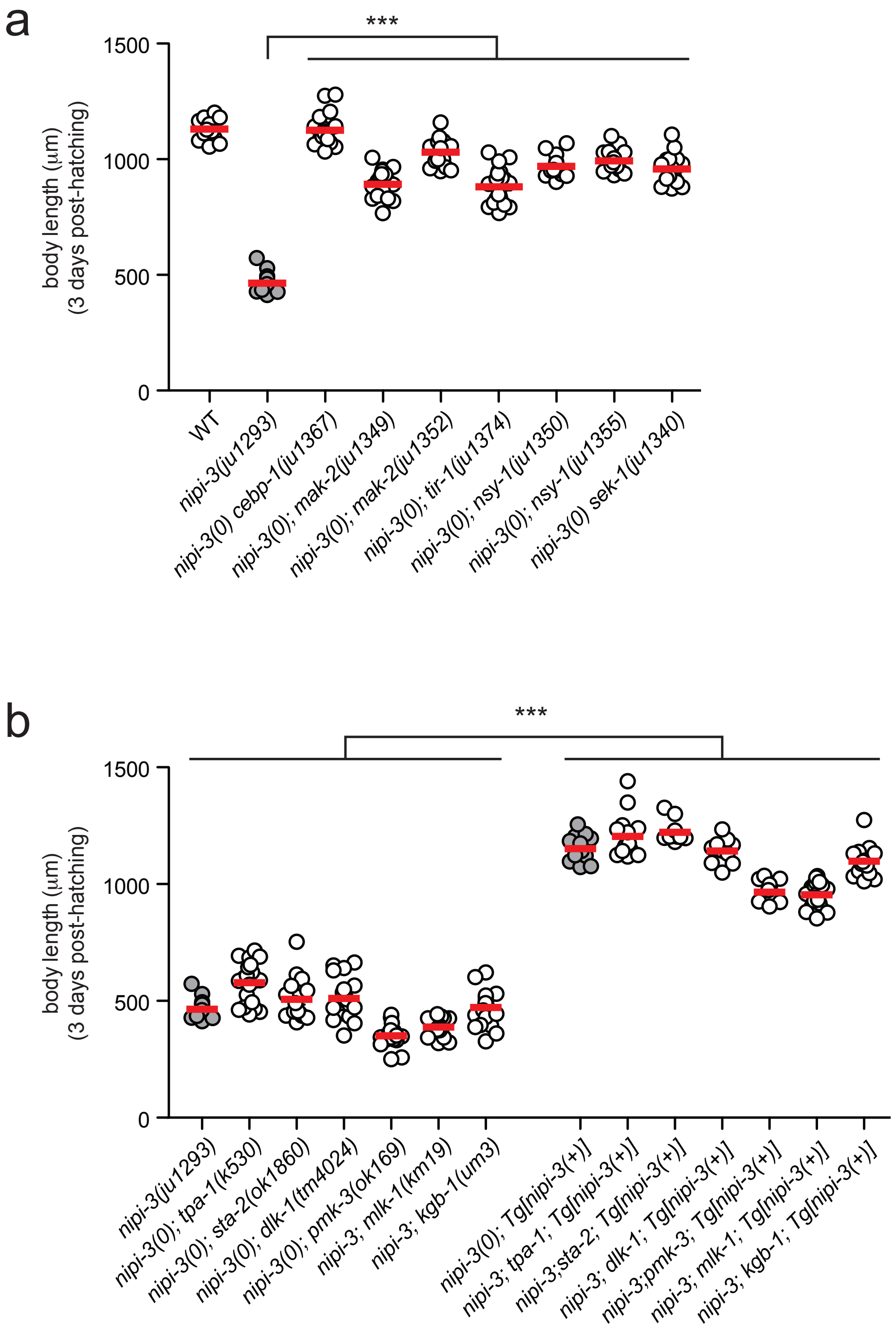
Kim_Supplementary Figure 3



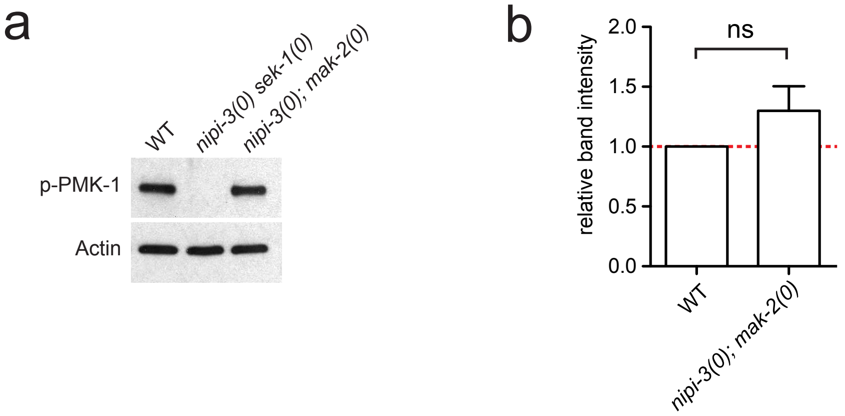
Kim_Supplementary Figure 4



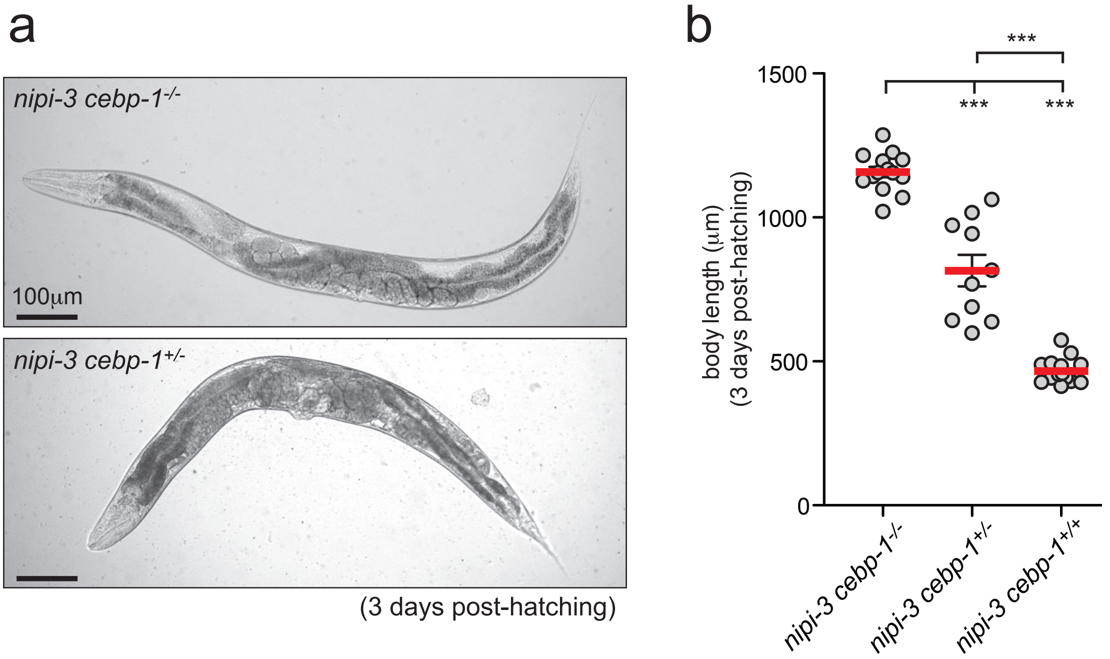
Kim_Supplementary Figure 5



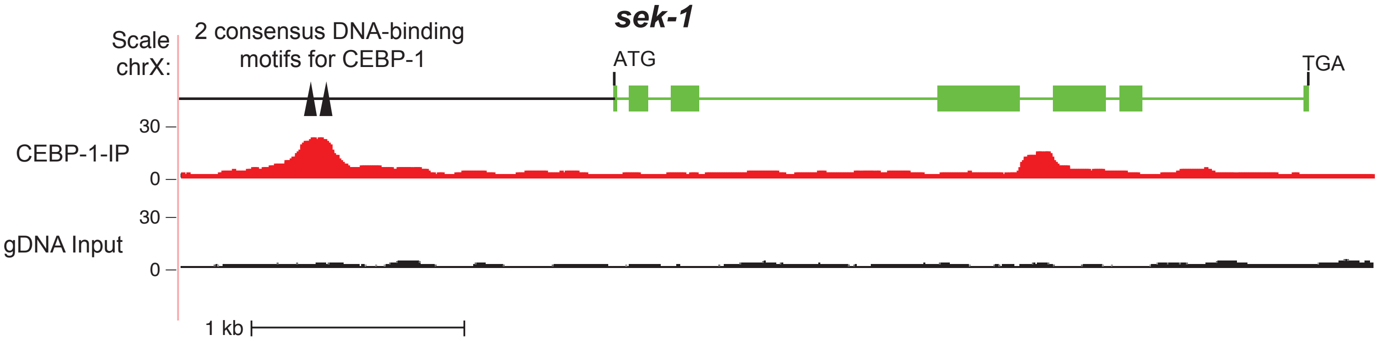
Kim_Supplementary Figure 6



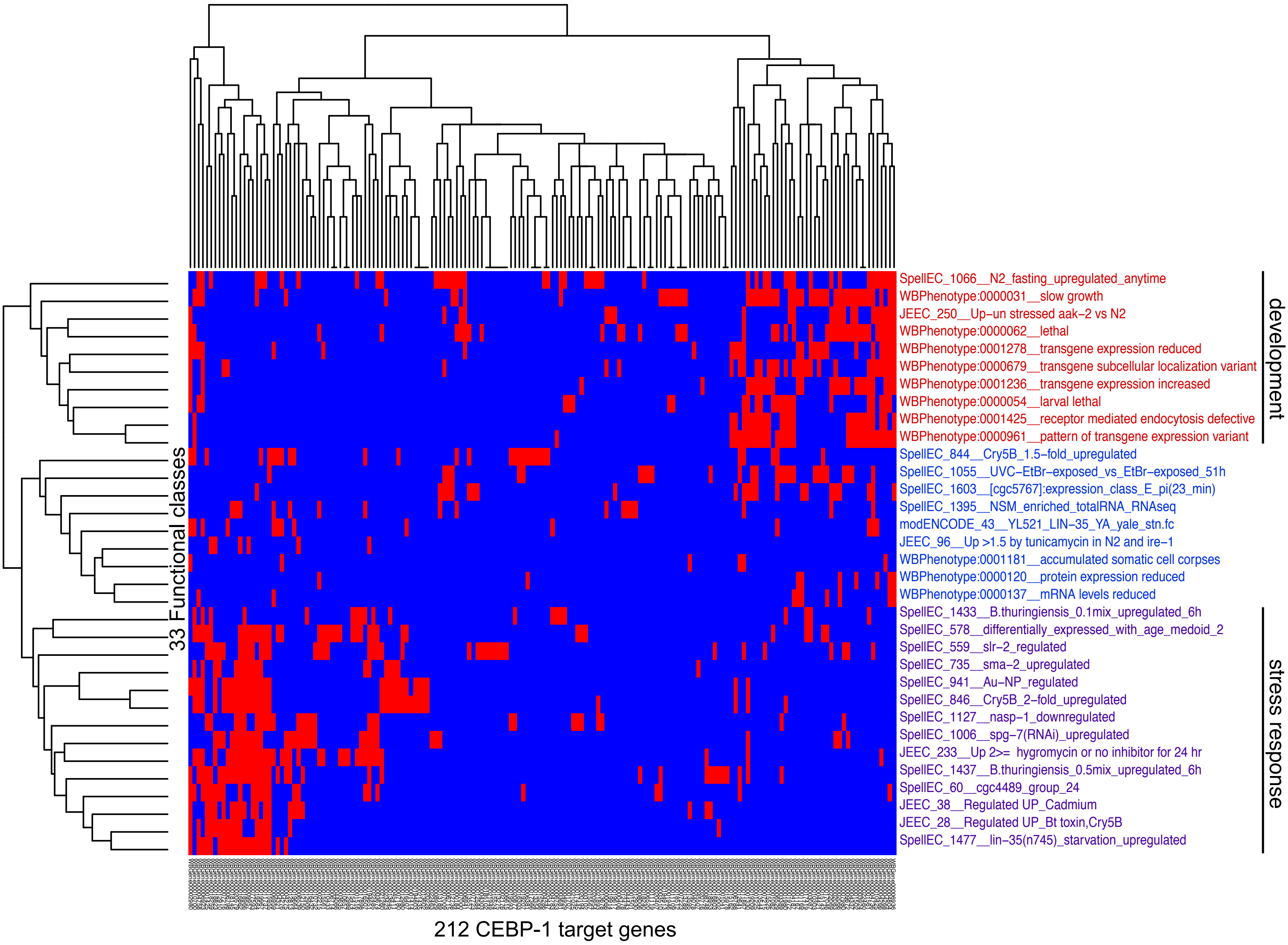
Kim_Supplementary Figure 7



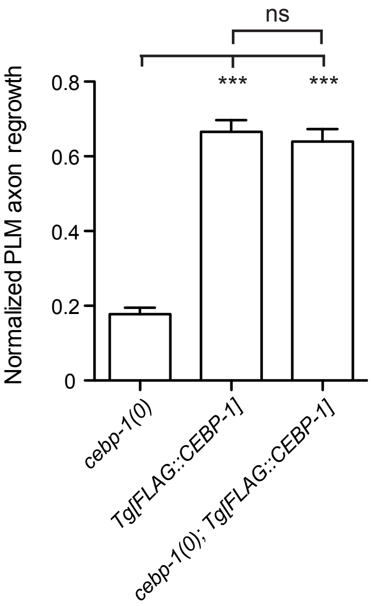
Kim_Supplementary Figure 8



Kim_Supplementary Figure 9



Kim_Supplementary Figure 10



2.5 Publication 5

Comparative Genomic Analysis of *Drechmeria coniospora* Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes.

Lebrigand K, He LD, **Thakur N**, Arguel MJ, Polanowska J, Henrissat B, Record E, Magdelenat G, Barbe V, Raffaele S, Barbry P, Ewbank JJ, PLoS Genet, 2016

In this study, my main contribution was the analysis of stage-specific and infection-related gene expression (RNA-seq data analysis). Apart from this I contributed to the comparative analysis of protein domains and carbohydrate-active enzymes cluster analysis. In addition, I set up in-house analysis tools (local BLAST server, etc) to facilitate genome analysis.

RESEARCH ARTICLE

Comparative Genomic Analysis of *Drechmeria coniospora* Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes

Kevin Lebrigand[1]☯, Le D. He[2]☯, Nishant Thakur[2], Marie-Jeanne Arguel[1], Jolanta Polanowska[2], Bernard Henrissat[3,4,5], Eric Record[6,7], Ghislaine Magdelenat[8], Valérie Barbe[8], Sylvain Raffaele[9,10], Pascal Barbry[1]*, Jonathan J. Ewbank[2]*

1 CNRS and University Nice Sophia Antipolis, Institute of Molecular and Cellular Pharmacology, Sophia Antipolis, France, **2** Centre d'Immunologie de Marseille-Luminy, Aix Marseille Université UM2, Inserm, U1104, CNRS UMR7280, Marseille, France, **3** CNRS UMR 7257, Aix-Marseille University, Marseille, France, **4** INRA, USC 1408 AFMB, Marseille, France, **5** Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, **6** INRA, UMR1163 Biodiversité et Biotechnologie Fongiques, Aix-Marseille Université, Polytech Marseille, CP 925, Marseille, France, **7** Aix-Marseille Université, UMR1163 Biodiversité et Biotechnologie Fongiques, Faculté des Sciences de Luminy-Polytech, CP 925, Marseille, France, **8** Commissariat à l'Energie Atomique, Institut de Génomique, Génoscope, Laboratoire de Biologie Moleculaire pour l'Etude des Génomes (LBioMEG), Evry, France, **9** INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, Castanet Tolosan, France, **10** CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, Castanet Tolosan, France

☯ These authors contributed equally to this work.
* barbry@ipmc.cnrs.fr (PB); ewbank@ciml.univ-mrs.fr (JJE)

OPEN ACCESS

Citation: Lebrigand K, He LD, Thakur N, Arguel M-J, Polanowska J, Henrissat B, et al. (2016) Comparative Genomic Analysis of *Drechmeria coniospora* Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes. PLoS Genet 12(5): e1006017. doi:10.1371/journal.pgen.1006017

Editor: Paul M. Richardson, MicroTrek Incorporated, UNITED STATES

Received: June 19, 2015

Accepted: April 8, 2016

Published: May 6, 2016

Data Availability Statement: All relevant data are either within the paper and its Supporting Information files or available from the NCBI; MIseq reads used for de novo genome assembly and Hiseq reads used for RNAseq de novo have been deposited at SRA under accession numbers SRX883538 (http://www.ncbi.nlm.nih.gov/sra/?term=SRX883538) and SRX969055 (http://www.ncbi.nlm.nih.gov/sra/?term=SRX969055), respectively. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JYHR00000000 (http://www.ncbi.nlm.

Abstract

Drechmeria coniospora is an obligate fungal pathogen that infects nematodes via the adhesion of specialized spores to the host cuticle. *D*. *coniospora* is frequently found associated with *Caenorhabditis elegans* in environmental samples. It is used in the study of the nematode's response to fungal infection. Full understanding of this bi-partite interaction requires knowledge of the pathogen's genome, analysis of its gene expression program and a capacity for genetic engineering. The acquisition of all three is reported here. A phylogenetic analysis placed *D*. *coniospora* close to the truffle parasite *Tolypocladium ophioglossoides*, and *Hirsutella minnesotensis*, another nematophagous fungus. Ascomycete nematopathogenicity is polyphyletic; *D*. *coniospora* represents a branch that has not been molecularly characterized. A detailed *in silico* functional analysis, comparing *D*. *coniospora* to 11 fungal species, revealed genes and gene families potentially involved in virulence and showed it to be a highly specialized pathogen. A targeted comparison with nematophagous fungi highlighted *D*. *coniospora*-specific genes and a core set of genes associated with nematode parasitism. A comparative gene expression analysis of samples from fungal spores and mycelia, and infected *C*. *elegans*, gave a molecular view of the different stages of the *D*. *coniospora* lifecycle. Transformation of *D*. *coniospora* allowed targeted gene knock-out and the production of fungus that expresses fluorescent reporter genes. It also permitted the initial characterisation of a potential fungal counter-defensive strategy, involving interference

nih.gov/nuccore/JYHR00000000). The version described in this paper is version JYHR01000000.

Funding: This work was supported by a program grant from the ANR (ANR-12-BSV3-0001-01) and institutional funding from INSERM, CNRS and AMU. The Functional Genomics Platforms of Nice Sophia Antipolis and Toulouse are supported by the Program "Investissements d'Avenir" managed by the National Research Agency (ANR-10-09-INBS) with support from the National Infrastructure France Génomique (ANR-10-INBS-09-03 and ANR-10-INBS-09-02), the Cancéropôle PACA, the Fondation pour la Recherche Médicale (DEQ20130326464 to PB) and the Labex Signalife (ANR-11-LABX-0028-01); microscopy used the France-BioImaging infrastructure supported by ANR-10-INSB-04-01; worm sorting infrastructure supported by ANR-11-LABX-0054 (Investissements d'Avenir–Labex INFORM) and ANR-11-IDEX-0001-02 (Investissements d'Avenir–A*MIDEX). SR is supported by a Marie Curie grant (MC-CIG 334036 project SEPAraTE), a starting grant of the European Research Council (ERC-StG 336808 project VariWhim) and the French Laboratory of Excellence project TULIP (ANR-10-LABX-41; ANR-11-IDEX-0002-02). LDH received a China Scholarship Council fellowship, and NT supported by an AMU doctoral fellowship and Labex INFORM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

with a host antimicrobial mechanism. This high-quality annotated genome for *D*. *coniospora* gives insights into the evolution and virulence of nematode-destroying fungi. Coupled with genetic transformation, it opens the way for molecular dissection of *D*. *coniospora* physiology, and will allow both sides of the interaction between *D*. *coniospora* and *C*. *elegans*, as well as the evolutionary arms race that exists between pathogen and host, to be studied.

Author Summary

Some soil-living fungi can kill nematodes and are used as biocontrol agents against plant parasitic nematodes. Certain species trap their prey using adhesive knobs or nets. For others, like *Drechmeria coniospora*, infection starts with the adhesion of specialized non-motile spores to the nematode cuticle. We have sequenced and annotated the *D. coniospora* genome. Comparative and functional genomic analyses provide insights into how its nematode-destroying lifestyle has evolved. We identified genes that were found only in *D. coniospora*, others found only in nematophagous species; many were highly expressed and differentially regulated during the different stages of fungal growth or during nematode infection. We have also developed methods for the genetic modification of *D. coniospora* that can be used to probe the function of its genes, allowing the dissection of this mode of nematode killing. We used them to probe a specific interaction between *D. coniospora* and *C. elegans*, involving the potential interference by the pathogen of a host antimicrobial mechanism.

Introduction

Species of nematophagous fungi have evolved a variety of strategies to invade and kill their nematode hosts in order to use them as a source of nutrients. Some species, exemplified by *Arthrobotrys oligospora*, form specialized and elaborate hyphal structures that trap nematodes, while others, like *Monacrosporium haptotylum*, use adhesive branches [1]. In addition to the fundamental interest in understanding these remarkable adaptations, these fungi are of economic importance because of their long-recognized potential as biocontrol agents of plant parasitic nematodes [2]. Insights into the molecular mechanisms that underlie the virulence of nematophagous species, and into their evolution, have been obtained from a series of genomic analyses (e.g. [3–9]).

Drechmeria coniospora produces non-motile spores (conidia) that stick to the nematode cuticle, via a specialized adhesive bud [10, 11]. Shortly after, the spores germinate, producing an appresorium that allows the fungus to pierce the nematode cuticle and send hyphae into its epidermis [10]. Until now, there has been essentially no molecular characterisation of *D. coniospora*. Thus, at the start of this project, nothing was known about its genetic makeup, apart from 1.05 kb of rRNA sequence in Genbank (GI:16763389; AF106012) that had been used to assign *D. coniospora* to the hypocrealean family, Clavicipitaceae, which includes many fungal pathogens of arthropods, such as *Beauveria bassiana* [12]. The same single sequence was used in a subsequent analysis that removed *Drechmeria* from the Clavicipitaceae and recognized it as one of six genera within the Ophiocordycipitaceae [13].

D. coniospora was adopted as a model fungal pathogen of *C. elegans* 30 years ago. Since the first studies in this domain [14, 15], *C. elegans* has emerged as a powerful model system for the investigation of host-pathogen interactions [16–22], and *D. coniospora* shown to be a natural pathogen of *C. elegans* [23]. We have put considerable effort into understanding the host defences that are triggered by *D coniospora* infection (e.g. [24–28]). Great strides in dissecting host defences in other organisms have been gained by investigating how pathogens evade or subvert these mechanisms (e.g. [29–32]). Understanding what is happening on the pathogen side during infection in the *D. coniospora*-*C. elegans* model could therefore be key to unravelling completely the host defence network, especially as the two protagonists are likely to have co-evolved [23, 33].

Completing a high-quality draft genome of *D. coniospora* is a very useful first step for understanding its virulence mechanisms. Combined with RNAseq transcriptomic and *in silico* analyses, it allowed us to predict a first complete gene set for *D. coniospora*. A comparison with other fungi, including nematode-destroying species, has revealed genes potentially involved in virulence and given insights into the evolution of the infectious capability of *D. coniospora*. To be able to exploit this knowledge, we established a method for fungal transformation, and, in a proof-of-principle, used it to generate recombinant knock-out and knock-in strains. These different approaches allowed us to initiate the investigation of a potentially novel fungal counter-defensive strategy.

Results

Sequencing output processing, *de novo* genome assembly and scaffolding

D. coniospora genomic DNA was sequenced on an Illumina MIseq sequencer as 2 x 150 bp paired-end reads. After filtering, we obtained 11.3 million reads, for a 100X coverage of a genome originally estimated to be 30 Mb. To determine which frequently used *de novo* genome assembly program performed best with this set of data, we tested four, Velvet [34], SPAdes [35], SOAPdenovo2 [36] and ABySS [37]. Each assembly was scaffolded with SSPACE [38], using two libraries of mate-paired 2 x 60 bp SOLiD reads. We applied a stringent filter, keeping only very high-quality reads, to limit errors during scaffolding. The libraries finally contained 23.2 and 23.6 million mate-paired reads, with insert sizes of 1.5 kb and 3 kb, respectively. Two contigs were scaffolded, using SSPACE, only when supported by at least 5 shared mate-paired reads. The overall characteristics of the *de novo* genome assemblies are shown in Table 1, before and after the SSPACE scaffolding step. ABySS (with kmer 96) and SPAdes performed well, giving low numbers of both contigs and unknown nucleotides, with ABySS maximizing the N50 value (2.09 Mb; length for which the collection of contigs of that length or longer contains half the total length of all contigs).

Optical mapping data integration

We then used optical mapping to test further the quality of the different assemblies generated before and after SSPACE scaffolding. Individual chromosomes were stretched on a glass slide and cut *in situ* with a restriction enzyme. The resulting fragments were visualized using a fluorescent microscope and their lengths measured. These lengths were compared to the predicted lengths of fragments from the longer scaffolds (i.e. with a length of at least 20 kb). The same approach was applied for each assembly. The nine distinct maps that were identified by optical mapping are indicative of a genome organization into 9 distinct chromosomes, with sizes ranging from 0.58 to 11.3 Mb (S1 Table).

The assembly obtained with ABySS (kmer = 96) was selected since it maximized the remapping of the scaffolds on the optical map whilst at the same time minimizing the number of mis-assemblies observed in the optical map analysis (4 versus 20 for Spades and 9 for ABySS at a

Table 1. Descriptive statistics of the different assemblies before and after SSPACE scaffolding.

	Before SSPACE					After SSPACE				
	Vel	**So**	**Sp**	**A64**	**A96**	**Vel**	**So**	**Sp**	**A64**	**A96**
Number contigs	455	12451	1075	7880	1356	190	11700	470	7157	1023
Total length (Mb)	31.6	33.1	31.9	32.3	32.0	31.7	34.0	31.9	32.8	32.1
Max contig length (Mb)	1.27	1.94	0.51	1.36	1.20	3.00	4.54	4.50	2.83	4.80
N50 (Mb)	0.32	0.58	0.12	0.35	0.24	1.40	1.76	1.21	1.70	2.09
% of Ns	0.77	1.85	0.00	0.24	0.12	1.13	4.58	0.16	1.85	0.33

Total number of contigs; total size of the assembly; length of the longest contig; N50; percentage of unknown bases. Values are given for each assembler (Vel: Velvet with kmer = 63, So: SOAPdenovo with kmer = 63, Sp: Spades with kmer = 127; A64 and A96: ABySS with kmer = 64 and 96, respectively) before and after scaffolding with SSPACE.

doi:10.1371/journal.pgen.1006017.t001

kmer value of 64; S1 Table). This low occurrence of scaffolding errors following SSPACE indicates that the first scaffolding step based on SOLiD reads would probably already have been of sufficient quality for an accurate assembly. During optical map analysis integration, we decided not to join scaffolds located on the same optical map when separated by a gap larger than 150 kb. Rather, for omap6937 and omap49267, the resulting map was divided into sub-chromosomal sequences. For instance, omap49267, the longest map, has a length of 11.3 Mb. It is represented in the final assembly by 2 sub-chromosomal sequences: omap49267a, with a length of 1.6 Mb and omap49267b, with a length of 10 Mb. The discrepancy in size between the sum of the lengths of the 2 sub-chromosomal sequences and that of omap49267 (11.6 Mb instead of 11.3 Mb) is explained by the fact that we did not break the scaffolds' distal extremities during scaffold concatenation. It is noteworthy that the sequences both at the 3' end of omap49267a and the 5' end of omap49267b (i.e. within the corresponding chromosome) are of low complexity (Fig 1), which can explain the difficulty in mapping properly these extremities on the optical map. We then put aside all contigs smaller than 0.5 kb (a collection totalling 128 kb). The final genome assembly therefore includes 75 sequences for a total size of 31.9 Mb. It contains less than 0.2% of unknown nucleotides, and its N50 value is equal to 3.86 Mb. The combination of 2 types of short-read sequencing (SOLiD and MISeq) with optical mapping therefore provided the basis for a high-quality assembly of the *D. coniospora* genome.

Transposable elements, repeat sequences, tRNAs, rRNAs and mtDNA

Amplification of transposable elements (TE) can contribute to chromosomal rearrangements, altering genomic structure and gene expression. It is believed to be an important route to speciation in some fungi. We therefore characterized the set of TEs predicted in the *D. coniospora* genome using TransposonPSI (http://transposonpsi.sourceforge.net/). A total of 600 elements were detected, falling into 12 classes (Table 2), and covering 1.6% of the genome (516 kb).

We compared these results with the number of TEs found in the genomes of 11 other fungi, chosen on the basis of their phylogenetic position and/or lifestyle, using the name assigned by the NCBI Taxonomy Database: (1) *Arthrobotrys oligospora*, *Hirsutella minnesotensis*, *Monacrosporium haptotylum* and *Pochonia chlamydospora*, 4 nematophagous fungi that feed on different species and stages of nematode worms; (2) the entomopathogenic fungi *Metarhizium acridum*, *Metarhizium anisopliae* and *Ophiocordyceps sinensis*; (3) the plant pathogens *Fusarium graminearum* and *Fusarium oxysporum*; (4) the mycoparasite *Tolypocladium ophioglossoides*; (5) *Trichoderma reesei*, a model mesophilic and filamentous fungus. Although being numerous in *D. coniospora*, the number of TEs was in no way exceptional. For example, there

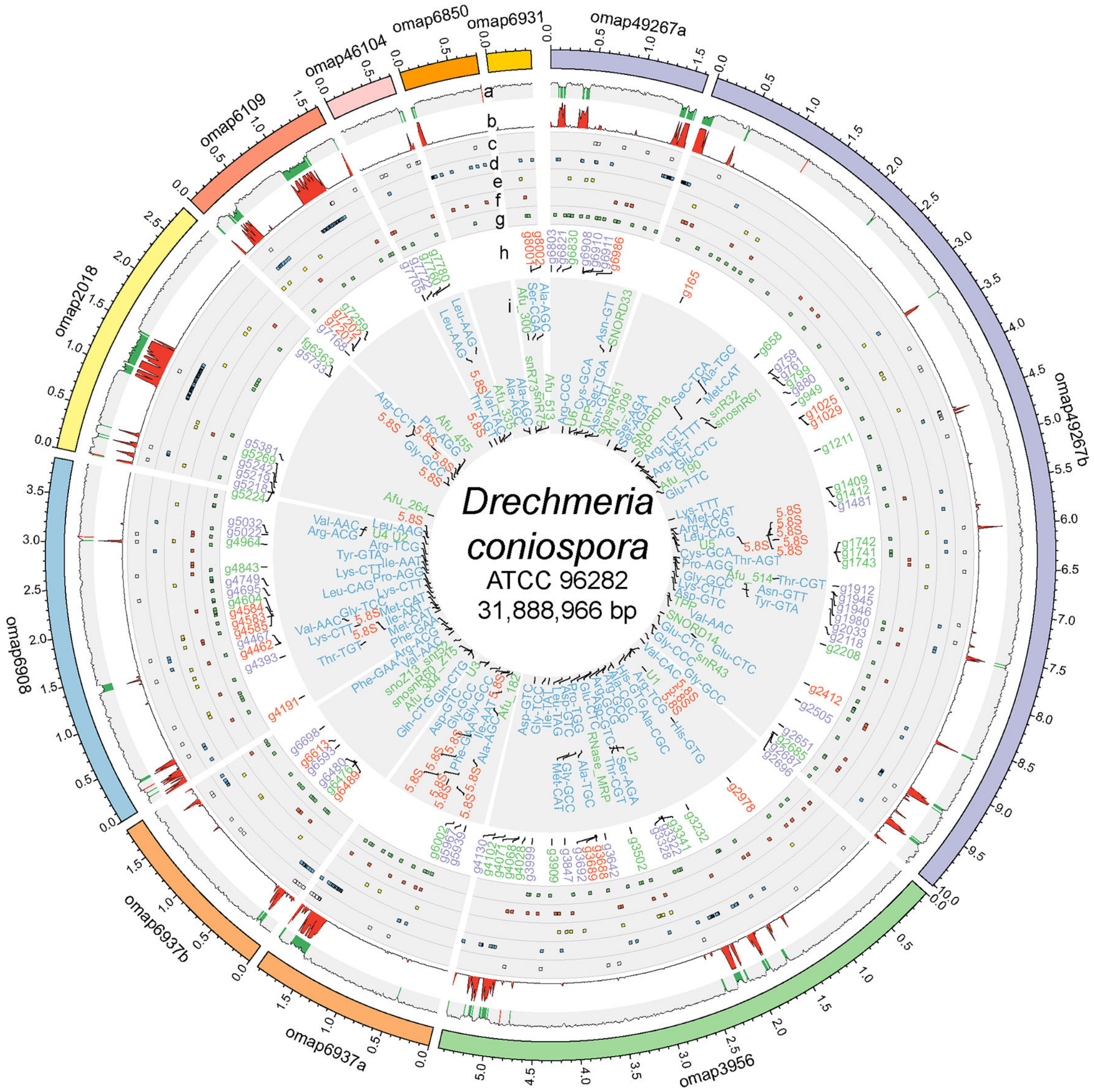


Fig 1. The *Drechmeria coniospora* genome. The optical maps (omap), potentially corresponding to distinct chromosomes, are depicted using Circos [39] with coloured sectors on the outer layer. As explained in the text, the optical maps 6937 and 49267 are split into two pieces. Scaffold43, corresponding to 23.8kb of mitochondrial DNA, and a further 63 other unanchored scaffolds, including 41 containing at least one predicted gene, and totalling 755,493 bp of genomic sequence, are not printed on the Circos plot. Each layer depicts, from the outside to the inside: **(a)** Percentage of G+C (red > 0.65, green < 0.45); **(b)** Percentage of repeat elements (red > 10%); **(c and d)** CLASS II and CLASS I transposable elements (white and blue blocks, respectively); **(e)** Members of three superfamilies encoding glutathione-S-transferases (GSTs), cytochrome P450 monooxygenases (P450s) and carboxyl/cholinesterases (CCEs) important for xenobiotic detoxification and oxidative stress resistance in entomopathogenic species [40] are depicted in yellow; **(f)** TM7 transmembrane proteins (red); **(g)** ABC proteins (green); **(h)** groups of genes discussed in the text that encode: putative nonribosomal peptide synthetases (red), diverse proteases (purple) and enterotoxin-like proteins in green (their names without the.t1 suffix are shown); **(i)** non-coding RNA genes: tRNAs (blue), rRNAs (red), others (green).

doi:10.1371/journal.pgen.1006017.g001

has been a remarkable proliferation of retrotransposons in *Ophiocordyceps sinensis* [41]; its genome is predicted to contain 12862 TEs (S2 Table).

As part of its overall characterisation, we mined the *D. coniospora* genome for repetitive elements using RepeatScout [42]. The resulting specific repeat library was used for masking the genome with RepeatMasker (www.repeatmasker.org) and finally covered 2.73 Mb (8.6%) of

Table 2. Transposable elements in the *D. coniospora* genome.

Class 1 DNA transposons, LTR retroelements		Class 2 DNA transposons			
Name	**Number**	**Name**	**Number**	**Name**	**Number**
TY1_Copia	193	hAT	38	MuDR_A_B	15
gypsy	193	cacta	41	piggybac	2
DDE_1	46	mariner	23	helitronORF	2
LINE	26	mariner_ant1	20	ISC1316	1

doi:10.1371/journal.pgen.1006017.t002

the full genome. Genes for ribosomal 28S and 18S RNAs were identified on scaffold58 (6.3 kb), near a 5.8S rRNA subunit gene. In addition, 26 copies of 5S RNAs were detected scattered throughout the entire genome. A further 33 non-coding, non-tRNAs/rRNAs were detected. We also searched for predicted tRNAs and found a total of 123 nuclear tRNA loci (Fig 1). Eighteen mitochondrial-specific tRNAs were located on Scaffold43 (23.8 kb, G+C 25.7%) after a tRNAscan search [43] with the option "organelle". The 141 predicted tRNAs correspond to 49 out of 64 codons. This degree of coverage compares very favourably with other sequenced fungal genomes (S3 Table). A more comprehensive investigation of Scaffold43 with TBLASTX [44] against a database of mitochondrial genes led to the identification of homologues for multiple mitochondrial proteins and established it as the mitochondrial DNA (S1 Text).

Gene prediction and genome annotation

From the genomic DNA, Augustus [45] was used to predict the coding DNA sequence (CDS) for a total of 8733 genes, with a mean length of 1425 bp and a GC content of 61.3% (higher than for the overall genome sequence, as is generally the case). The predicted set represents a total of 12.4 Mb of coding nucleotides. To assess the quality and completeness of the prediction, we performed a BUSCO analysis [46], which is based on expectations of gene content from near-universal single-copy orthologs (USCOs). Using a set of more than 1400 fungal USCOs, again the annotation of the *D. coniospora* genome appeared to be of high quality, at least as good as that of the other fungi used in this study (Table 3).

With these results in mind, it is therefore interesting to compare this first comprehensive prediction for the *D. coniospora* genome with those of the other 11 fungi (Table 4).

D. coniospora was recently reassigned to the family Ophiocordycipitaceae [13]. Overall, compared to the 3 other family members included in our analysis, the *D. coniospora* genome much more closely resembles that of *Tolypocladium ophioglossoides* than either *Hirsutella minnesotensis* or *Ophiocordyceps sinensis*, in terms of size, GC content and the number of predicted genes. Both species have comparatively small genomes but a relatively high complement of predicted protein-coding genes.

Phylogeny

A previous study placed *D. coniospora* in Ophiocordycipitaceae on the basis of a single DNA sequence [13]. In order to carry out a more thorough phylogenetic analysis, using BUSCO, we identified a set of 97 high-confidence orthologous proteins present in all 12 fungal species. Concatenated sequences (S4 Table) were aligned using MAFFT [47] and phylogenetic distances calculated using PhyML [48]. The overall phylogeny was in line with recent phylogenetic studies of *H. minnesotensis* and *P. chlamydosporia* [9, 49], and confirmed *D. coniospora*'s place in the Ophiocordycipitaceae family. Consistent with the general features of their respective genomes, the analysis placed *D. coniospora* closest to *T. ophioglossoides* (Fig 2). The results

Table 3. Universal single-copy ortholog prediction in *D. coniospora* and 11 other species.

	Complete single-copy	Complete duplicated	Fragmented	Missing	Source[a]
Drechmeria coniospora	98% (1412)	10% (153)	1.5% (22)	0.2% (4)	This study; PRJNA269584
Trichoderma reesei	97% (1407)	11% (167)	1.6% (24)	0.4% (7)	Trire2/Trire2.home.html
Fusarium graminearum	97% (1406)	11% (167)	2% (29)	0.2% (3)	Fusgr1/Fusgr1.home.html
Arthrobotrys oligospora	97% (1404)	11% (159)	2.1% (31)	0.2% (3)	Artol1/Artol.home.html
Monacrosporium haptotylum	96% (1390)	10% (156)	2.7% (39)	0.6% (9)	Monha1/Monha1.home.html
Hirsutella minnesotensis	96% (1388)	12% (183)	2.3% (34)	1.1% (16)	PRJNA67943
Fusarium oxysporum	96% (1386)	35% (511)	3.2% (47)	0.3% (5)	Fusox1/Fusox1.home.html
Metarhizium anisopliae	95% (1376)	11% (165)	3.5% (51)	0.7% (11)	Metan1/Metan1.home.html
Tolypocladiumophioglossoides	95% (1374)	10% (157)	1.6% (24)	2.7% (40)	PRJNA91059
Metarhizium acridum	95% (1370)	10% (152)	4% (58)	0.6% (10)	Metac1/Metac1.home.html
Pochonia chlamydosporia	88% (1270)	11% (168)	9% (130)	2.6% (38)	www.fungalinteractions.org/index.php/en/genome
Ophiocordyceps sinensis	71% (1026)	7.3% (105)	10% (152)	18% (260)	PRJNA59569

The table shows the percentage of the different categories of USCOs, with the corresponding number of proteins in brackets, as calculated by BUSCO. Orthologs are classified as 'complete' when their lengths are within two standard deviations of the BUSCO group mean length, otherwise they are classified as 'fragmented' (length not within the threshold) or 'missing'. 'Complete' orthologs found with more than one copy are classified as 'duplicated'.

[a]NCBI bioproject number, full URL, or end of URL at genome.jgi.doe.gov

doi:10.1371/journal.pgen.1006017.t003

support the conjecture that in Hypocreales invertebrate-pathogenic fungi form a monophyletic group, distinct from cellulolytic, plant pathogenic filamentous fungi [49], while also providing a further illustration of the multiple independent origins of nematode pathogenic fungi and the distinct evolutionary trajectories of the trapping fungi such as *Arthrobotrys oligospora*, as opposed to the conidial species including *D. coniospora*.

Functional annotation of the predicted proteome

In order to obtain a first overview of the set of proteins predicted from the *D. coniospora* genome, we functionally annotated the protein sequences with InterproScan [50]. This assigned at least one annotation to more than three quarters of them (6734/8733; 77.1%). There was a bias in the distribution of annotations; longer proteins were more likely to have an annotation. Thus while the vast majority (90.7%) of proteins as long or longer than the median (393 amino acids) had an annotation, only 63.4% of proteins shorter than the median had one. For smaller proteins, the effect was even more marked. Indeed, only half (50.6%) of proteins shorter than 250 amino acids long had an annotation (S5 Table). In line with previous observations [51], this bias was mirrored in the pattern of conservation. While overall, 60% (5248/8733) of the predicted proteins have a homologue in the curated UniprotKB/Swissprot database, 74.8% of the proteins as long or longer than the median were assigned a homologue, while only 34.6% of the proteins less than 250 amino acids were. It should be noted that these are conservative estimates of homology since although of high quality, the UniprotKB/Swissprot database is not exhaustive. Indeed, for example, on the basis of BLASTP analyses using the current NCBI non-redundant database, many (19/31) of the predicted proteins longer than 1000 amino acids but without any annotation have homologues in other fungi. The closest homologue was most often found in *Tolypocladium ophioglossoides* (S5 Table), consistent with the phylogenetic analysis.

Table 4. General characteristics of the *D. coniospora* genome compared to other species.

	Drechmeria coniospora	*Metarhizium acridum*	*Metarhizium anisopliae*	*Pochonia chlamydosporia*	*Arthrobotrys oligospora*	*Monacrosporium haptotylum*	*Fusarium graminearum*	*Fusarium oxysporum*	*Trichoderma reesei*	*Tolypocladium ophioglossoides*	*Hirsutella minnesotensis*	*Ophiocordyceps sinensis*
Sequencing depth (fold coverage)	100	107	100	136	37	28	85	-	48	76	128	241
Genome size (Mb)	31.9	39.4	39.2	42.4	40.1	39.5	36.5	61.4	33.5	31.2	51.1	78.5
Chromosomes[a]	9 + MT	-	-	-	-	-	4 + MT	15 + MT	7 + MT	-	-	-
GC content (%)	55.2	49.9	51.5	49.9	44.5	45.3	48.3	48.4	52.8	57.3	52.1	43.5
# of scaffolds	75	241	176	956	215	1279	31	114	87	172	736	10603
Scaffold N50 (Mb)	3.86	0.33	1.96	0.23	2.04	0.19	5.35	1.98	1.22	0.67	0.38	0.01
% of unknown (N's)	0.2	3.49	0.27	2.91	0.27	0.09	0.61	2.32	0.14	3.15	2.44	5.77
No. predicted protein-coding genes	8733	9849	10583	11079	11479	10959	13322	17708	9849	9317	12700	6972
Average exons per gene	2.83	2.7	2.68	-	3.17	3.31	2.82	2.7	2.88	-	2.5	-
Average exon length (bp)	504	549	568	-	473	470	508	498	512	-	-	-
Median protein length (aa)	393	406	416	393	407	416	366	366	408	410	403	362
Mitochondrial DNA (kb) [a]	23.8	145	24.7	25.6	170.4	140.2	107.7	84.8	42.1	-	40.6	42.2
Repetitive sequence in Mb (%)	2.73 (8.6)	1.16 (2.9)	0.59 (1.5)	0.06 (0.1)	0.54 (1.4)	0.73 (1.9)	0.35 (1.0)	11.85 (19.3)	0.17 (0.5)	0.37 (1.2)	13.22 (25.9)	51.64 (65.8)
Transposable elements in Mb (%)	0.52 (1.6)	0.15 (0.4)	0.33 (0.8)	0.14 (0.3)	0.17 (0.4)	0.32 (0.8)	0.04 (0.1)	3.176 (5.2)	0.11 (0.3)	0.07 (0.2)	3.64 (7.1)	8.51 (10.8)
NCBI accession	PRJNA269584	ADNI00000000.1	PRJNA156697	AOSW00000000	ADOT00000000.1	AQGS00000000.1	AACM00000000.2	AAXH01000000	PRJNA118357	PRJNA91059	PRJNA67943	PRJNA59569

[a] See S1 Text

doi:10.1371/journal.pgen.1006017.t004

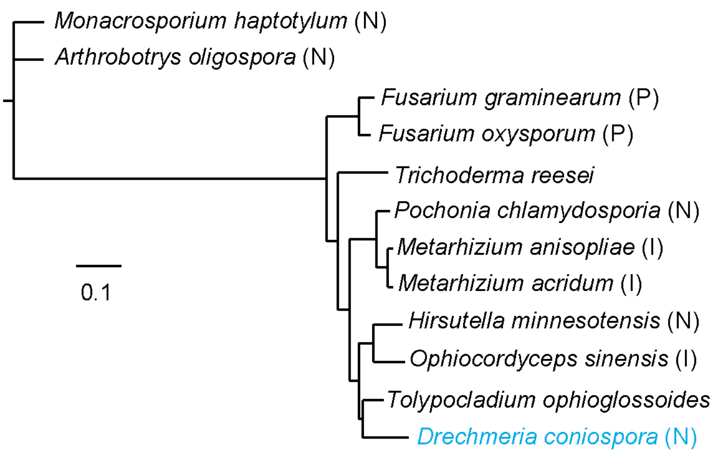

Monacrosporium haptotylum (N)
Arthrobotrys oligospora (N)
Fusarium graminearum (P)
Fusarium oxysporum (P)
Trichoderma reesei
Pochonia chlamydosporia (N)
Metarhizium anisopliae (I)
Metarhizium acridum (I)
Hirsutella minnesotensis (N)
Ophiocordyceps sinensis (I)
Tolypocladium ophioglossoides
Drechmeria coniospora (N)
0.1

Fig 2. Phylogenetic tree for 12 Pezizomycotina fungi. Phylogenetic tree for 12 species based on alignments for a concatenation of 97 conserved protein sequences. Branch-lengths are drawn in proportion to the estimated number of substitutions per site. Species known to infect insects (I), nematodes (N) and plants (P) are indicated. All branches are fully supported (100/100 bootstraps).

doi:10.1371/journal.pgen.1006017.g002

To define gene families and assign proteins to orthologous and paralogous groups, we performed an OrthoMCL analysis [52]. Most proteins (6851) were assigned to an OrthoMCL orthologous group, together with 241 for which an orthologue but no group was assigned (in total 81.2%; S6 Table). For the great majority (82%) of proteins allocated to an OrthoMCL orthologous group, the closest homolog was in *Fusarium graminearum* (the anamorph name of *Gibberella zeae*), the only Hypocrealean fungal species represented in OrthoMCL (S6 Table). The most populated groups, with 20 members (OrthoMCL group OG5_126718) correspond to predicted nonribosomal peptide synthases (see below), followed by ABC transporters and subtilisin-like serine proteases (OG5_134254 and OG5_137388, respectively, with 15 members each). When no homolog is found in any of the 150 species in OrthoMCL, proteins are placed into paralogous groups on the basis of their sequence [52]. The largest group of paralogs, with 15 members, corresponds to the heat-labile enterotoxin alpha chain, present in several insect pathogenic fungal species including *B. bassiana* and *Metarhizium robertsii* (e.g. EXU96489) [53]. A further 3 groups (containing a total of 18 predicted proteins) exhibited a more-or-less strong similarity to the heat-labile enterotoxin alpha chain, and one group of 5 to subtilisin-like serine proteases, also mentioned above (Tables 5 and S6). As discussed below, these all potentially play a role in fungal virulence or in the interaction of *D. coniospora* with other microbes. This analysis also revealed many groups of proteins currently unique to *D.*

Table 5. Families of predicted *D. coniospora* paralogs.

OrthoMCL group(s)*	Number of members	Identity
1, 4, 9, 17	15, 8, 6, 4	Heat-labile enterotoxin alpha chain
2, 5, 6, 7, 8, 10, 14, 15, 16, 19	10, 8, 7, 7, 6, 6, 4, 4, 4, 4	*Drechmeria*-specific; unknown function
3, 12, 13, 18, 20	9, 5, 4, 4, 4	Hypothetical protein conserved in certain fungal species; unknown function
11	5	Subtilisin-like serine protease

*See S6 Table

doi:10.1371/journal.pgen.1006017.t005

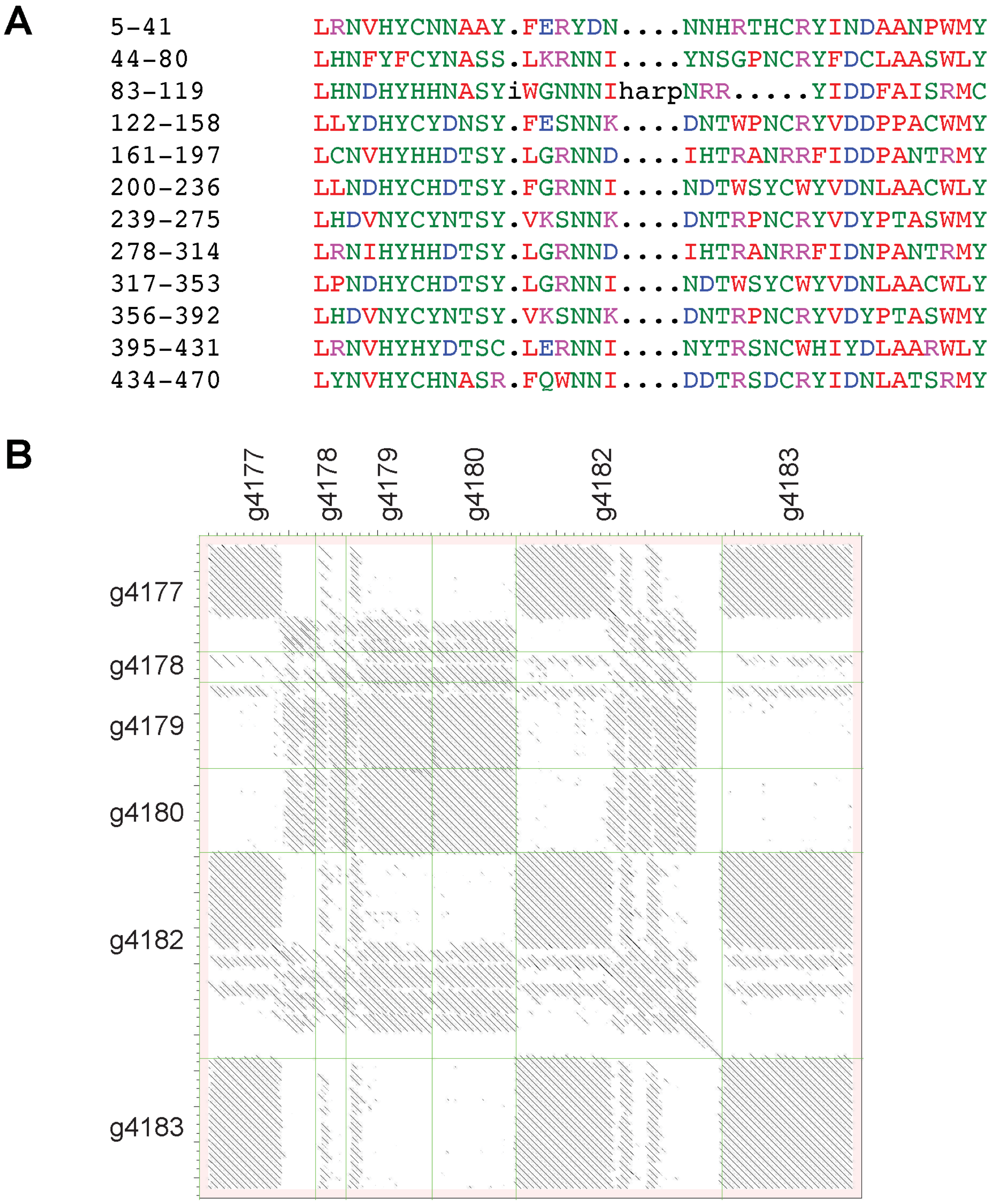
coniospora. In the most dramatic example, a group of proteins (OrthoMCL paralogous group 2) with no recognisable domain, each entirely composed of highly repeated sequences, has expanded to 10 members, with the majority found in a cluster of less than 50 kb on scaffold omap6908, with a very complex pattern of conservation (S6 Table and Figs 3 and S1). A further 9 out of the 20 groups with at least 4 members currently correspond to proteins unique to *D. coniospora* (Table 5).

Comparative analysis of protein domains

To look in more detail at protein domains, we next compared the PFAM annotations [56] for the entire set of predicted *D. coniospora* proteins with those of the 11 other fungi obtained with InterproScan using the same set of parameters (S7 Table). We completed the analyses of specific proteins and protein families by manual inspection. A total of 4287 PFAM domains were identified in at least one protein from one or more of the species, with 3510 (81.9%) represented in predicted *D. coniospora* proteins and more than half (55.7%) in all 12 predicted proteomes. These presumably reflect core eukaryotic and/or fungal biological processes. In the expectation of revealing domains that were potentially functionally related, we hierarchically clustered protein domain families present in *D. coniospora* and not more than 4 of the other fungi, but no obvious associations were found (Fig 4). A number of PFAM domains were predicted for *D. coniospora* proteins but not for proteins of any of the other 11 fungi (Fig 4 and S7 Table and S1 Text). Within this group, each PFAM domain is present in a single protein, with 3 exceptions. Three predicted *D. coniospora* proteins (OrthoMCL paralogous group 33, S6 Table and S1 Text) contain PF12810, the "glycine rich protein" domain, characterised by several glycine rich motifs interspersed through the sequence. Currently, no orthologues have been described in any other species, and no hint of a function can be garnered from the sequence. Two lipid-binding MORN (Membrane Occupation and Recognition Nexus; PF02493) domains [57] are predicted, towards the C-terminus, in proteins from each of 2 adjacent *D. coniospora* genes (g5037.t1 and g5038.t1; OrthoMCL orthologous group OG5_154358). MORN domains are relatively uncommon in fungi, but tandemly arranged orthologues for these 2 proteins do exist in one species, *Trichoderma gamsii*, and orthologs are currently also found in various other fungi, such as the brown-rot Basidiomycota *Hydnomerulius pinastri*. The conserved N-terminal portion of these proteins is shared with a number of related toxins, including the hemolytic factor neoverrucotoxin from stonefish venom (S5 Table). Whether these proteins may play a role in fungal virulence and interactions with other microorganisms is a matter for speculation. The presence of the Saposin A domain (PF02199) in 2 predicted proteins (g3895.t1 and g1982.t1, with 3 and 2 domains respectively) is equally atypical (S5 and S7 Tables). We address its possible role below.

On the basis of their constituent domains, several other atypical or highly represented protein families (Fig 4 and S7 Table) are also potentially linked to virulence. These include the single iron-sequestrating lipocalin (PF13924) and the deuterolysin M35 metalloprotease (PF02102; called here M35) domains. The M35 domain is unusually highly represented compared to other fungi [58, 59], being present in 10 predicted *D. coniospora* proteins.

Of a total of 777 domains absent from *D. coniospora*, 32 were present in all 11 other species (S7 Table). As a most striking example, *D. coniospora* lacks proteins containing the NACHT domain (PF05729), which is present between 8 and 117 times in the other species, suggesting that this is unlikely to simply be a problem of sequence coverage or gene prediction. In ascomycete fungi, the NACHT domain can be found together with the HET domain (PF06985) in heterokaryon incompatibility proteins. It acts as a death effector domain, preventing viable heterokaryotic cells from being formed by the fusion of filaments from different wild-type strains



```
A   5-41      LRNVHYCNNAAY.FERYDN....NNHRTHCRYINDAANPWMY
    44-80     LHNFYFCYNASS.LKRNNI....YNSGPNCRYFDCLAASWLY
    83-119    LHNDHYHHNASYiWGNNNIharpNRR.....YIDDFAISRMC
    122-158   LLYDHYCYDNSY.FESNNK....DNTWPNCRYVDDPPACWMY
    161-197   LCNVHYHHDTSY.LGRNND....IHTRANRRFIDDPANTRMY
    200-236   LLNDHYCHDTSY.FGRNNI....NDTWSYCWYVDNLAACWLY
    239-275   LHDVNYCYNTSY.VKSNNK....DNTRPNCRYVDYPTASWMY
    278-314   LRNIHYHHDTSY.LGRNND....IHTRANRRFIDNPANTRMY
    317-353   LPNDHYCHDTSY.LGRNNI....NDTWSYCWYVDNLAACWLY
    356-392   LHDVNYCYNTSY.VKSNNK....DNTRPNCRYVDYPTASWMY
    395-431   LRNVHYHYDTSC.LERNNI....NYTRSNCWHIYDLAARWLY
    434-470   LYNVHYCHNASR.FQWNNI....DDTRSDCRYIDNLATSRMY
```

Fig 3. Unusual structure of *Drechmeria*-specific proteins and complex structural relationship between neighbouring proteins. (A) RADAR analysis [54] reveals the repeated structure in the sequence of g4180.t1, a 471 a.a. protein from OrthoMCL-defined paralogous group 2 (S6 Table). (B) All-against-all dot-plot representation [55] of the alignment of the predicted protein sequences from g4180.t1 and from 5 neighbouring genes on scaffold omap6908, all from the OrthoMCL paralogous group 2. Dots represent regions of sequence similarity (within a 100 a.a. sliding window). The intensity of each dot is proportional to the corresponding alignment score. The ".t1" suffix has been removed from all sequence names.

doi:10.1371/journal.pgen.1006017.g003

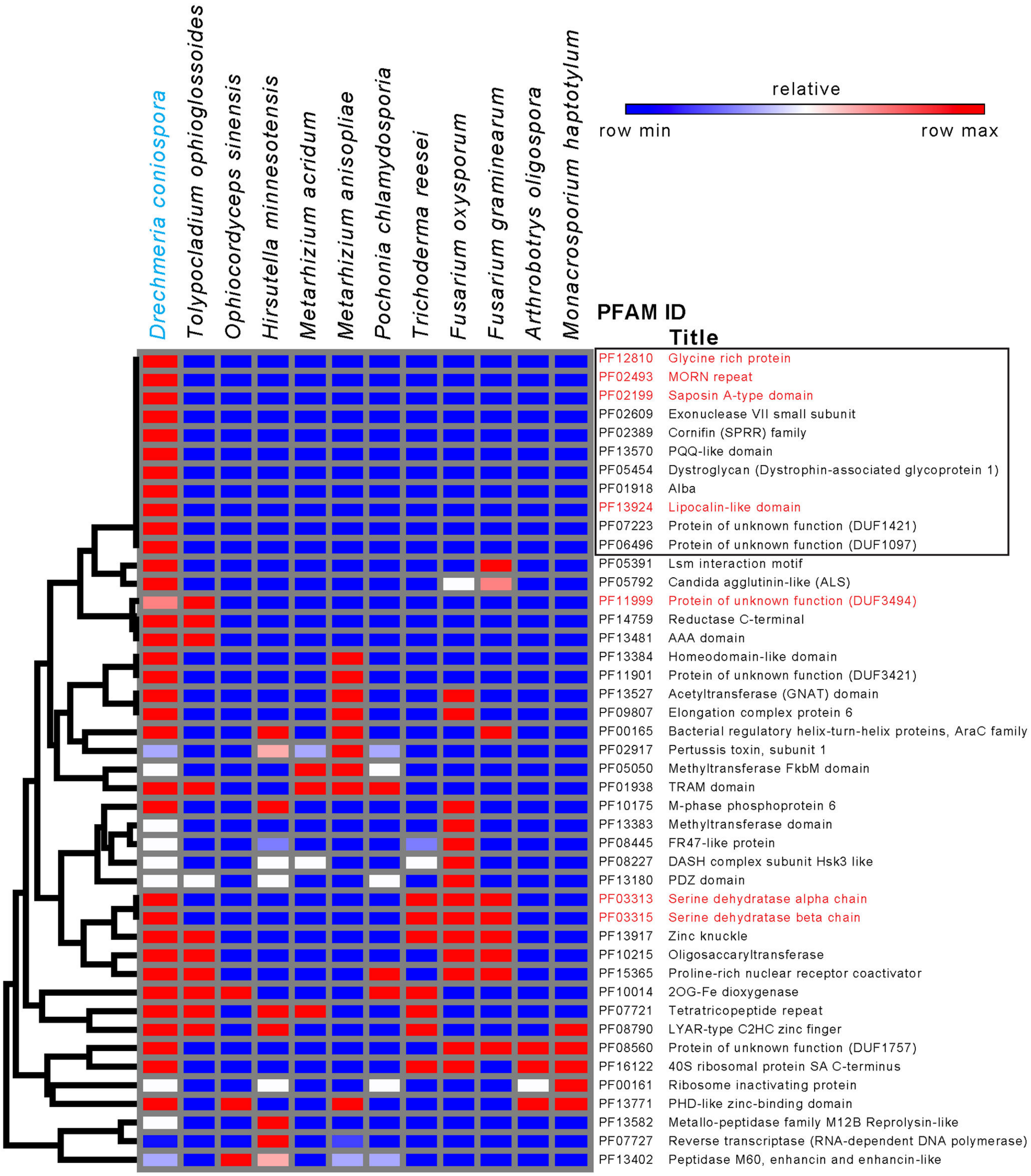


Fig 4. Species-specific or atypical protein domain families in the predicted *D. coniospora* proteome. Hierarchical clustering of protein domain families present in *D. coniospora* and not more than 4 of the other fungi, on the basis of the corresponding number of proteins. PFAM domains discussed in the text are highlighted in red. The box highlights 11 families specific to *D. coniospora* (see S1 Text). The colour code reflects the relative abundance of proteins with each domain, from high (red) to low (blue) across the different species. The serine dehydratase alpha and beta domains (PF03313 and PF03315, respectively), cluster since they occur in a single highly conserved protein (g4699.t1 in *D. coniospora*).

doi:10.1371/journal.pgen.1006017.g004

[60]. In *D. coniospora*, of the 6 proteins predicted to contain a HET domain, 5 (g3856.t1, g5845.t1, g5969.t1, g6809.t1, g7038.t1) do not contain any other domains, while in common with many other fungal proteins of this class, one (g215.t1) is coupled to an ankyrin repeat (PF12796) domain. No sexual stage has been described for *D. coniospora*, and in common with 3 of the 4 other nematophagous fungi, no mating type protein (MATα, PF04769) domain was predicted (S7 Table), making it unlikely that these different proteins play any role in heterokaryon incompatibility.

To see whether there was any general pattern for the 32 domains absent from *D. coniospora* suggestive of a coordinate evolutionary process, we used dcGO analysis to look for ontology term enrichment at the domain level [61]. This revealed a potential involvement of the NACHT domain together with 4 others (PF00931; PF01966; PF02178; PF09273) in the biological process, "regulation of response to stress" and of 4 of them (PF00931; PF01966; PF05729; PF09273) in "regulation of defense response". This is discussed further below. Submitting all 777 domains absent from *D. coniospora* to dcGO analysis highlighted the absence of 16 related domains, all annotated as being involved in hydrolase activity, acting on glycosyl bonds, including PF00331 that corresponds to the Carbohydrate-Active Enzyme (CAZyme) glycosyl hydrolase (GH) family 10. There was also a significant ($p = 2.5 \times 10^{-4}$) enrichment for the more specific ontology term "alpha-N-arabinofuranosidase activity" (PF05270, PF06964, PF09206), suggesting an alteration in the capacity of *D. coniospora* to metabolise different carbohydrates compared to other fungi analysed.

Carbohydrate-active enzymes

A species' set of CAZymes can often give insights into its biology, in particular into nutrient sensing and acquisition. Given the differences revealed by the dcGO analysis, we conducted a targeted examination of CAZymes in *D. coniospora* and 10 of the 11 fungi chosen for the other comparative analyses (S8 Table). *P. chlamydiosporia* was not included as it will be the subject of a dedicated study. Overall, CAZyme profiling recapitulated the phylogenetic analysis, except that the two *Metarhizium* species clustered together with *H. minnesotensis* and *O. sinensis* (Fig 5). This grouping of the various fungi reflects their respective requirements for carbon acquisition. The two nematode-trapping fungi *A. oligospora* and *M. haptotylum* have a large repertoire of enzymes for feeding on plant cell wall polysaccharides. They make a separate group and are neighbours of the saprophytes, reflecting their dual parasitic and saprophytic lifestyles. *T. reesei* has an intermediate position consistent with its evolving from an ancestral saprophyte lifestyle to become a mycoparasite. The remaining fungi, which are the most specialized and have evolved by gene loss, group together by virtue of their common loss of an arsenal of plant polysaccharide degradative enzymes. Thus the nematophagous and insectivorous fungi in this group are not separate from the mycoparasite *T. ophioglossoides*; the same range of CAZymes is probably needed for the three types of substrate and this is accompanied by a similar loss of the plant-targeting CAZymes. Regarding *D. coniospora* in detail, it has lost virtually all enzymes, from multiple families, that participate in cellulose binding (Carbohydrate Binding Module, CBM1), the breakdown of cellulose/hemicellulose and pectin-rich plant cell walls (e.g. GH7, GH45, PL8, and CE8 family proteins; S8 Table). The few GH5 proteins that remain in *D. coniospora* are predicted to be involved in the metabolism of fungal cell wall ß-glucans, not the digestion of plant cellulose or mannan. The GH13 family, involved in both starch and glycogen breakdown, has also shrunk to just two members. The two remaining proteins show strong similarities to glycogen branching and debranching enzymes and are thus most likely involved in the fungal glycogen cycle. Collectively, and coupled with the expansion of protease families

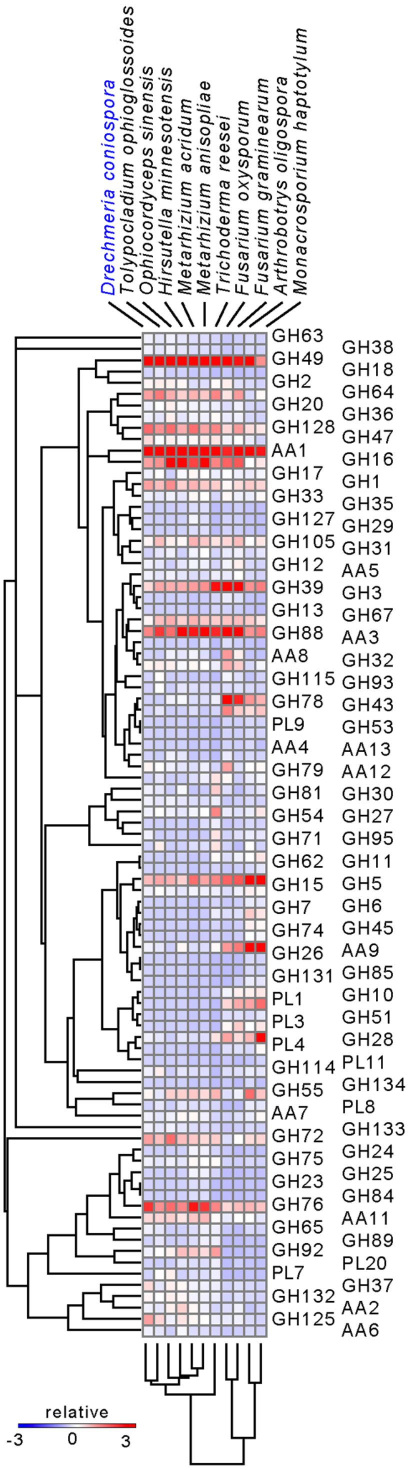


Fig 5. Supervised clustering of selected CAZy families in 11 fungal species. The distribution of the CAZy families involved in complex carbohydrate breakdown (AA, GH and PL classes) across the given species is shown. Clustering of families is based on the number of genes in each family. The colour code reflects the relative abundance of proteins within each family, from high (red) to low (blue) across an individual species.

doi:10.1371/journal.pgen.1006017.g005

and the acquisition of virulence factors, these changes appear to reflect the shift of *D. coniospora* to an obligate nematophagous lifestyle.

Secondary metabolite biosynthesis

Another hallmark of each fungal species is its complement of genes involved in the production of secondary metabolites. The "backbone" genes for these biosynthetic pathways include those encoding nonribosomal peptide synthases (NRPSs), polyketide synthases (PKSs), and prenyltransferases (DMATSs), responsible for the production of bioactive peptides, polyketides, and indole alkaloids, respectively. SMURF analysis [62] revealed that the number and type of backbone genes predicted for *D. coniospora* was comparable to those of other Hypocreales fungi (Table 6). The majority of the NRPS and NRPS-like genes belong to OrthoMCL group OG5_126718. Several of the genes (g1025.t1 and g1029.t1; g3636.t1 and g3638.t1; g7201.t1, g7202.t1 and g7204.t1; g8001.t1 and g8002.t1) are close together in the genome suggesting that they may be functionally related (S9 Table). The molecules synthesized by NRPSs, PKSs and DMATS are frequently modified by "decorating" enzymes before secretion. The genes encoding the proteins necessary for these different steps are often found in genomic clusters, and are co-ordinately regulated by specific Zn_2Cys_6 transcription factors and/or by the global secondary metabolism regulator LaeA [62]. Of the 29 backbone genes, 24 had an associated gene cluster, and of these, 4 included a Zn_2Cys_6 gene. These are therefore candidate regulators of their respective clusters. Two clusters (14 and 15) included genes in the proximity of the *D. coniospora* LaeA-encoding gene g6733.t1 (S9 Table). In *D. coniospora*, LaeA could play a conserved role in secondary metabolism. It is important to note that in many fungal pathogens, secondary metabolites are essential for virulence [63].

PHI-base analysis

To gain a general view of proteins potentially involved in virulence, we made use of the pathogen-host interactions database, PHI-base (phi-base.org; [65]). Of the 2104 proteins matching a PHI-base entry, 990 had the annotation "reduced virulence" or "loss of pathogenicity", indicating that a homologous protein in at least one other species plays a demonstrated role as a virulence factor in a particular model of infection (S5 Table).

Among the highly represented (≥5) hits in PHI-base, several were characterised by the presence of ABC transporter domains, and so predicted to be involved in ATP-dependent export of organic anions or drugs from the cytoplasm (S10 Table and Fig 1). For predicted NRPS proteins, 9/20 members of the orthoMCL group OG5_126718 were assigned a PHI-base annotation (PHI:2511). The analysis also highlighted the potential role of multiple degradative enzymes, 11 chitinases (GH18; PHI:144; 6/6 OG5_126929, 2/2 OG5_142806, 2/2 OG5_210539, 1/1 OG5_152762), known to be important for the virulence of nematophagous fungi [7, 66], and

Table 6. Number of secondary metabolite backbone genes predicted from the *D. coniospora* genome compared to other species.

	DMAT	NRPS	NRPS-Like	PKS	PKS-Like	Reference
Drechmeria coniospora	1	10	9	7	2	This study
Tolypocladium ophioglossoides	0	14	6	16	2	[64]
Trichoderma reesei	0	8	5	11	1	[62]
Hirsutella minnesotensis	6	21	21	27	4	This study
Fusarium oxysporum	2	7	12	9	2	[62]
Fusarium graminearum	0	10	11	14	1	[62]

doi:10.1371/journal.pgen.1006017.t006

subtilisin-like and extracellular metalloproteases (PHI:2117 and 479, with 9 and 5 members, respectively). These are often found in expanded gene families in pathogenic fungi (e.g. [67]). There were also 11 Pth11-like receptors (PHI:404), which can be involved in host sensing and have established roles in virulence in other species [68]. Three of them contain a CFEM domain (PF05730), found in proteins with proposed roles in fungal pathogenesis [69]. As expected from the OrthoMCL analysis, there were also multiple hits to enterotoxin A proteins (PHI:698, with the PF01375 domain), scattered throughout the genome (g496.t1, g964.t1, g2819.t1, g5058.t1, g6833.t1, g7169.t1, g7949.t1). The *D. coniospora* genome is therefore predicted to encode a large range of virulence factors, some of which have expanded markedly in number compared to other fungal species.

D. coniospora secretome

To be able to act as virulence factors, many proteins, for example chitinases and proteases, need to be secreted. Some virulence factors are secreted into host cells, and can be targeted to specific organelles. We therefore complemented the InterproScan and PHI-base analysis with a focused and more thorough *in silico* investigation of the *D. coniospora* secretome. A total of 608 proteins (7%) were predicted with high confidence to be secreted (Fig 6 and S11 Table). They included the Saposin A domain protein g3895.t1, as well as 6/10 M35 domain proteins. More than a third of the putative secreted proteins (242/608), including 5 of the 6 secreted M35 domain proteins, and multiple proteins containing several different glycosyl hydrolase domains (GH2, 3, 16, 18, 20, 31, 35 47 and 65; see above), were also predicted to target a host cell organelle (e.g. nucleus or mitochondria), and of these 27 were homologous to proteins present in PHI-base with a demonstrated role in virulence (Fig 6 and Tables 7 and S11). In addition to chitinases, among the 27 predicted proteins, there were alkaline, aspartic, metallo-, subtilisin-like and cuticle-degrading proteases, all of which potentially contribute to the destruction of host tissue. There were also heat-labile enterotoxin homologues that would

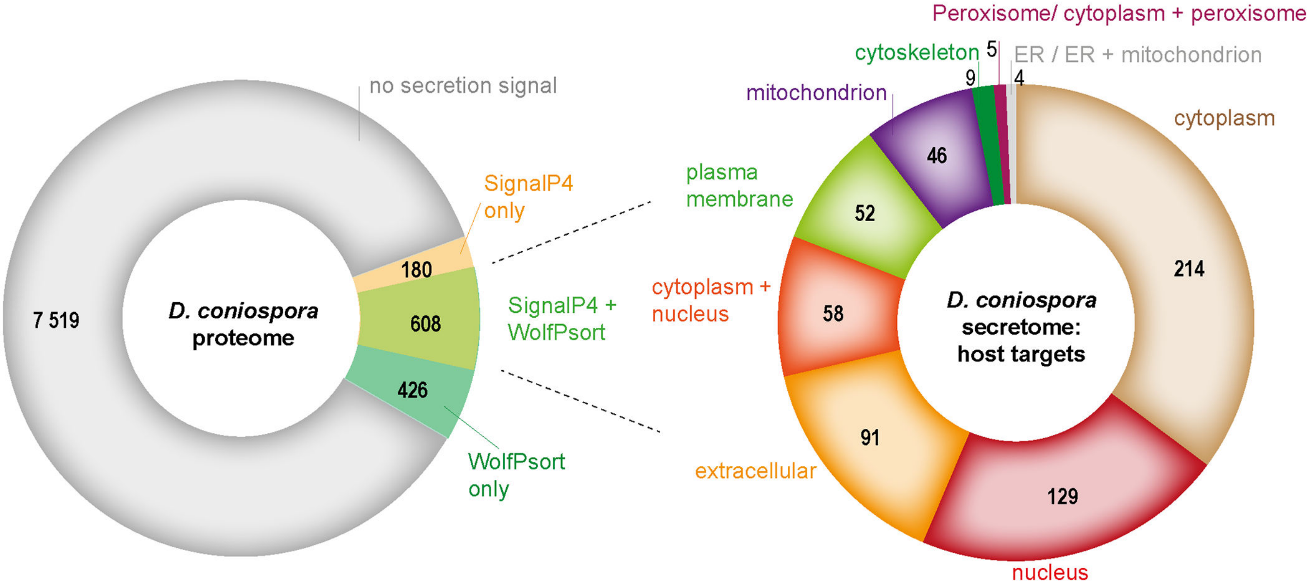


Fig 6. Predicted secreted proteins in *D. coniospora*. The left hand chart shows the distribution of protein predicted to be secreted by 2 different computational methods. For the proteins predicted to be secreted by both, the right hand chart indicates the predicted sub-cellular localisation.

doi:10.1371/journal.pgen.1006017.g006

Table 7. Potential virulence factors among *D. coniospora* proteins predicted to be secreted and targeted to a host organelle.

Gene ID(s)	PHIBASE ID	BLASTP hit	Brief description
g3389.t1	PHI:434	GEL1_ASPFC	1,3-beta-glucanosyltransferase
g6911.t1	PHI:2117	ORYZ_ASPCL	Alkaline protease
g1715.t1	PHI:73	YPS3_YEAST	Aspartic proteinase
g2067.t1	PHI:1046	SOL5_ALTSO	Bifunctional solanapyrone synthase
g2448.t1	PHI:383	SOD5_ARTBC	Cell surface Cu-only superoxide dismutase
g375.t1, g3187.t1	PHI:144	CHI1_APHAL	Chitinase 1
g2033.t1	PHI:2117	CUDP_METAN	Cuticle-degrading protease
g3550.t1	PHI:2570	CYB2_WICAO	Cytochrome b2, mitochondria
g858.t1	PHI:3078	E9DYG9_METAQ	Endonuclease/exonuclease/phosphatase family protein
g656.t1, g6818.t1	PHI:479	MEP1_COCP7	Extracellular metalloprotease
g960.t1	PHI:1071	GLU2A_SCHPO	Glucosidase 2 subunit alpha
g964.t1	PHI:698	E2AA_ECOLX	Heat-labile enterotoxin IIA
g496.t1	PHI:698	E2BA_ECOLX	Heat-labile enterotoxin IIB
g1983.t1	PHI:2920	ASO_CUCPM	L-ascorbate oxidase
g2355.t1, g5020.t1	PHI:785	MU157_SCHPO	Meiotically up-regulated gene 157 protein
g6255.t1	PHI:184	PRY2_YEAST	Pathogenesis-related protein 2
g4517.t1	PHI:1071	AGDC_ASPFU	Probable alpha/beta-glucosidase agdC
g1115.t1	PHI:68	OPSB_ASPOR	Probable aspartic-type endopeptidase
g2153.t1	PHI:184	PRY1_ARTBC	Probable pathogenesis-related protein
g5968.t1, g6831.t1	PHI:2654	A2965_ARTBC	Putative amidase
g1191.t1	PHI:1166	ATG15_CHAGB	Putative lipase (Autophagy-related protein 15)
g4757.t1	PHI:2117	SUB2_PSED2	Subtilisin-like protease 2
g7479.t1	PHI:891	G4MVB6_MAGO7	Uncharacterized protein

doi:10.1371/journal.pgen.1006017.t007

similarly be predicted to be direct effectors of virulence or play a role in *D. coniospora*'s interactions with other microbes (see below).

Comparative analysis of nematode-destroying fungi

The above results illustrate how secreted proteins can be key to virulence. To investigate commonalities and differences in the molecular basis of nematode infection, we therefore conducted a comparative analysis of predicted secretomes between the 12 fungal species, focusing on the 5 that are nematopathogenic. Using reciprocal BLASTP analyses, we first determined high-confidence clusters of orthologous proteins among the 12 species. We then concentrated on the 1548 clusters containing at least one protein from a nematopathogenic species (S11 Table), and calculated their distribution across the 5 species (Fig 7A). While there was a substantial overlap between *A. oligospora* and *M. haptotylum*, with 395/700 shared clusters unique to these 2 species, very few of the clusters uniquely shared between *D. coniospora* and *H. minnesotensis* or *P. chlamydosporia* were restricted to nematopathogenic species (3/38 and 4/44, respectively). Indeed, there were only 9 clusters present in *D. coniospora* and another nematopathogenic fungus but not any of the 7 non-nematopathogenic species. Only one of these clusters corresponded to proteins with a conserved domain, namely fungal hydrophobin (PF01185), also found in rodlet proteins, a major component the hydrophobic sheath, or rodlet

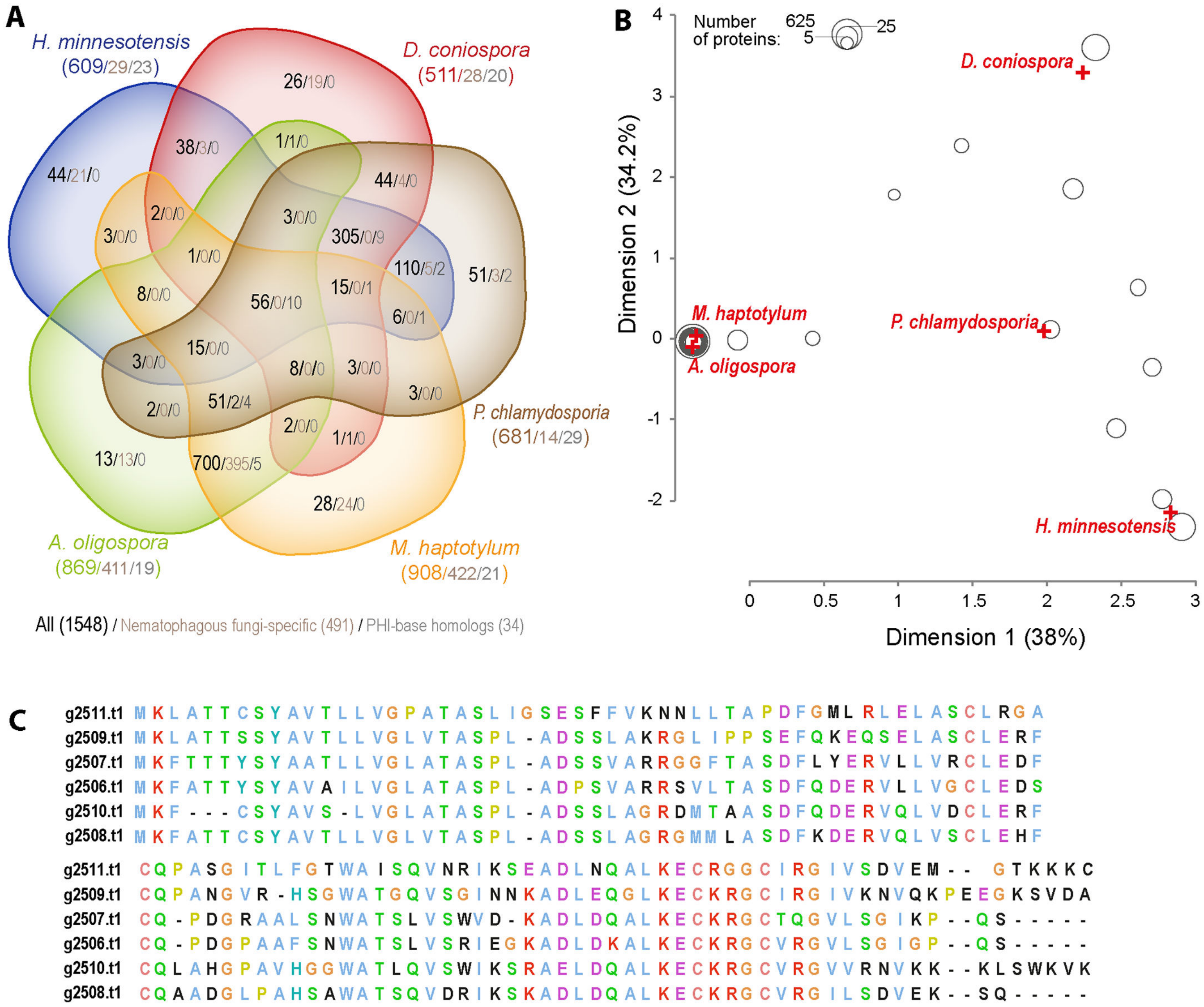


Fig 7. Comparative analysis of the predicted secretome of *D. coniospora*. (A) Distribution of sequence-based protein clusters across five nematopathogenic fungal genomes. Only clusters containing at least one secreted protein are shown. Except for empty sectors, in each sector, there are 3 numbers: total number of clusters/number of clusters with members only in nematophagous fungi/number of clusters with at least one member that matches a PHI base entry with an annotation of reduced virulence or loss of pathogenicity. (B) A correspondence analysis of sequence-based clusters of secreted proteins from five nematopathogenic fungi. The first two dimensions are shown. Crosses represent the position of the individual fungal species and circles represent protein clusters. Circles are sized according the number of constituent proteins as indicated. When clusters have identical coordinates, the size of the circle represents the sum of the number of proteins in each cluster. For example, the circle at (2.33, 3.6) corresponds to 19 clusters of proteins, in this case unique to *D. coniospora*, including Cluster01087. The proximity of each circle to the species' apices is a measure of the contribution of the species to that cluster's content. The distance between the circles is a measure of the similarity of their content (number of proteins from each species). (C) Multiple sequence alignment of proteins from the *D. coniospora*-specific cluster Cluster01087. Only 3 of the 6 proteins are predicted to be secreted (g2506.t1, g2508.t1, g2511.t1; S11 Table).

doi:10.1371/journal.pgen.1006017.g007

layer, that covers the surface of fungal spores, required in *Aspergillus nidulans* for efficient spore dispersal [70].

To gain a more synthetic and rigorous overview of these results, we used correspondence analysis of the clusters of the secreted proteins from the 5 nematopathogenic fungi. Correspondence analysis is conceptually similar to principal component analysis, but can be used with categorical rather than continuous data. The first 2 dimensions explained more than 70% of the observed distribution and separated the species. As expected from the numerical overlaps of protein clusters, *A. oligospora* and *M. haptotylum* were much closer together, and furthest from *D. coniospora* and *H. minnesotensis* (Fig 7B). The separation of the latter 2 species reflects their unique clusters. For *D. coniospora*, the most populated cluster of proteins (Cluster01087) contains 6 members (Fig 7C). The corresponding genes (g2506.t1-g2511.t1) are together in the genome. While the 6 proteins have no annotation or predicted function (S5 Table), 5 other clusters include at least one member with a PFAM domain. Consistent with our previous domain-centric analyses, among them we found the M35 (PF02102) and heat-labile enterotoxin alpha chain (PF01375) domains, reinforcing the notion that these domains potentially characterize the pathogenic capacity of *D. coniospora*. The remaining 3 were a domain of unknown function (PF11999), and 2 expected to be involved in fungal adhesion, the GLEYA (PF10528) domain, [71], and the hydrophobic surface binding protein A (PF12296) domain discussed in the next section. It is noteworthy that GLEYA domain proteins are highly expressed by nematode-trapping fungi during infection of *C. briggsae* [5].

Stage-specific gene expression

While analysis of the genome sequence reveals predicted proteins putatively involved in virulence or other aspects of fungal physiology, it is necessary to establish when the corresponding genes are in fact expressed. To gain a first insight into the genes potentially important at different stages of the life cycle of *D. coniospora*, we chose to compare gene expression between mycelia that had been grown in liquid for several generations in the absence of nematodes, and spores harvested from infected *C. elegans* and starting to germinate *in vitro* (S2 Fig). These two morphological forms were expected to provide a broad though not necessarily exhaustive repertoire of expressed transcripts. The combined set of 48.6 million paired-end reads were mapped to the set of predicted genes. Inspection of the most highly expressed genes (arbitrarily the top 50) in mycelia and spores, revealed a substantial number of genes encoding basic metabolic enzymes, such as glyceraldehyde 3-phosphate dehydrogenase, as well as multiple proteins involved in translation (ribosomal proteins and elongation factors) and protein folding (chaperones of the heat shock protein family), presumably reflecting the need for protein synthesis during growth. We then used stringent criteria to define a small set of genes that were differentially regulated between germinating spores and mycelia, (61 and 86 genes, respectively; S12 Table). The list of spore-specific genes was particularly interesting. It included 16 encoding predicted proteins with no identified PFAM domain or homolog in UniProtKB. Several are found in the secretome-associated and/or orthoMCL-defined clusters described above including g2508.t1 and g2509.t1 (group 6 and Cluster01087), g4179.t1 (group 2), and g2083.t1 (group 34). The functional role in spores of the members these different paralogous groups merits further investigation in the future.

Two genes g3607.t1 and g6474.t1 encode proteins that contain the hydrophobic surface binding protein A domain (PF12296) mentioned above. By analogy with the eponymous protein from *Aspergillus oryzae* [72], we hypothesize that this protein is important for spore attachment to the cuticle of *C. elegans* and that further, they may act together with g5675.t1 that contains a CFEM domain (PF05730 [69]), that is structurally related to fungal adhesins and is highly preferentially expressed in spores. The differential expression of such genes will depend on stage-specific expression of transcription factors, such as g5153.t1 that is predicted to be a transcription factor with Zn_2Cys_6 (PF00172) and fungal-specific (PF11951) domains, and is also preferentially expressed in spores (S12 Table).

Genes that are more highly expressed in mycelia than spores would be predicted to be important for vegetative growth but also potentially for virulence. The list of these genes included candidates in both categories (S12 Table). Thus, on the basis of the domains found in the corresponding predicted proteins, 5 genes are associated with carbohydrate metabolism (GO:0005975; g1835.t1, g1896.t1, g3555.t1, g7200.t1, g7950.t1), while g7950.t1 corresponds to a pyruvate/2-oxoglutarate dehydrogenase, a key metabolic enzyme, and fg7260.t1 to a sulfide: quinone oxidoreductase that catalyzes the first step in the mitochondrial metabolism of H_2S. These are expected to fulfil metabolic needs that are not present during the early growth of spores.

Otherwise, there is g3889.t1 that encodes a highly conserved protein, annotated as a putative NRPS-like enzyme in multiple fungal species; it was not identified as such by SMURF analysis. g3999.t1 encodes a subtilisin and was also preferentially expressed in mycelia. Subtilisin-type proteases are associated with virulence, including nematicidal activity [73], in many species [6]. As in other nematophagous fungi [8], as detailed above, *D. coniospora* has a large family of subtilisin-type proteases (PF00082; S7 Table and Table 5). These different genes are all potentially linked to growth in the nematode host. Strikingly, 46 of the 86 (53.5%) proteins corresponding to genes preferentially expressed in mycelia currently have no identifiable domains nor homologs in the UniProtKB database, compared to 30.8% for the full set of 8733 predicted proteins (S5 and S12 Tables), and among them 41/46 are shorter than the median length. Manual searches suggest that many may have homologues in other fungal species. As a single example, fg6211.t1 is expressed at almost 10-fold higher levels in mycelia than spores, encodes a predicted 70 amino acid protein and matches a predicted 64 amino acid protein of unknown function from *Trichoderma reesei* (Genbank XP_006962079). Determining the role of these different genes will require extensive functional analyses in the future.

Gene expression during infection of *C. elegans*

While this analysis revealed genes potentially involved in virulence, an important question is what fungal genes are actually expressed during infection. Having the annotated *D. coniospora* genome in hand allowed a re-examination of RNAseq data obtained from samples of *C. elegans* infected by *D. coniospora* (NCBI SRA SRX036882 and [74]). From sequencing of samples taken 5 and 12 h post-infection (p.i.), a small number of reads among those that did not align to the *C. elegans* genome could be aligned to predicted *D. coniospora* genes. Together, 537 gene models were covered by at least one read, with 339 only at 5 h (p.i), 142 only at 12 h (p.i.) and 56 at both time-points (S13 Table).

Focusing on genes for which there were at least 3 matching reads (S13 Table), as might be expected, many corresponded to genes that were highly expressed in mycelia and/or spores (30/47 within the top 15 percentile for expression; i.e. >1192 and >1298 reads, for mycelia and spores respectively; S12 Table). In addition to 14 ribosomal genes, they included the CFEM domain protein, g5675.t1, mentioned above. Six encode proteins homologous to ones present in PHI-base with a demonstrated role in virulence. For example, g6659.t1 corresponds to a component of the mitochondrial membrane ATP synthase complex (S13 Table). As with the other such genes, its role in virulence probably reflects a general function in fungal physiology and growth.

For the remaining 17 genes that were not highly expressed in mycelia and/or spores, BLASTP searches at NCBI lead to the identification of potential homologs for 15 of the corresponding predicted proteins (e-value $<10^{-10}$), across a range of species (Table 8). Among them, 3 encoded homologues of PHI-base listed virulence factors (S5 Table and Table 7). The first, g2153.t1 (PHI:184), potentially encodes a cysteine-rich secretory protein family

Table 8. Selection of *D. coniospora* genes expressed during the infection of *C. elegans*.

	Read counts					Best Genbank BLASTP hit			PFAM domains	
Gene ID	**5 hour p.i.**	**12 hour p.i.**	**Mycelia**	**Spores**	**Length (a.a.)**	**.Genbank**	**Length (a.a)**	**Species**	**ID**	**Name**
g2153.t1	8	0	81	189	307	GI:302895657	324	*Nectria haematococca*	PF00188	Cysteine-rich secretory protein family
g2389.t1	3	0	40	29	118	GI:799240667	128	*Hirsutella minnesotensis*	-	
g2390.t1	4	0	57	34	303	GI:667661323	308	*Beauveria bassiana*	PF11327	Protein of unknown function (DUF3129)
g261.t1	0	4	411	167	69	GI:952549466	69	*Aspergillus lentulus*	PF11034	Protein of unknown function (DUF2823)
g3607.t1	7	2	147	593	428	GI:908391879	373	*Tolypocladium ophioglossoides*	PF12296	Hydrophobic surface binding protein A
g4094.t1	10	11	26	73	111	-	-	-	-	
g4805.t1	3	0	123	82	323	GI:666406534	327	*Stachybotrys chartarum*	-	
g4815.t1	3	1	3	35	354	GI:908393692	353	*Tolypocladium ophioglossoides*	PF11790	Glycosyl hydrolase catalytic core
g4816.t1	3	2	0	54	343	GI:949387848	548	*Rosellinia necatrix*	PF12430	Abscisic acid G-protein coupled receptor
g4817.t1	4	0	3	262	446	-	-	-	-	
g5054.t1	3	0	23	111	363	GI:969886810	403	*Trichoderma gamsii*	-	
g5671.t1	1	5	422	198	163	GI:799246947	108	*Hirsutella minnesotensis*	-	
g6474.t1	10	0	40	426	331	GI:389629282	257	*Magnaporthe oryzae*	PF12296	Hydrophobic surface binding protein A
g6475.t1	3	0	89	166	600	GI:743663295	600	*Metarhizium guizhouense*	PF01532	Glycosyl hydrolase family 47
g6844.t1	2	1	77	159	160	GI:531863496	231	*Ophiocordyceps sinensis*	-	
g6908.t1	2	1	5	4	563	GI:667643613	395	*Beauveria bassiana*	PF00082/ PF05922	Subtilase family/Peptidase inhibitor I9
g7182.t1	3	0	107	161	416	GI:672821344	241	*Mortierella verticillata*		

The Gene IDs in bold indicate genes that are neighbours in the genome. Those that are underlined are homologous to proteins in PHI-base annotated as being important for virulence.

doi:10.1371/journal.pgen.1006017.t008

(PF00188) member, similar to the pathogen-related protein Pry1. Members of this family have different roles in various pathogenic fungal species, including the neutralization of host defenses, and antimicrobial activity to inhibit the growth of competing microorganisms [75]. The second, g2390.t1 (PHI:257), encodes a widely conserved cell surface protein of unknown function, and the third, g6908.t1 (PHI:2117), an alkaline serine protease, homologous to the peptidase S8 of *Beauveria bassiana*. Interestingly, among the 17 genes, there were 3 groups of neighbours, suggesting a coordination of gene expression at the genome level. These included a group of 3 genes, g4815.t1–g4817.t1, that encode proteins that are completely unrelated in sequence: g4815.t1, one of 3 *D. coniospora* GH128 proteins, from a recently described glycoside hydrolase family ([76]; S8 Table), g4816.t1, a G-protein coupled receptor, and g4817.t1 that does not currently have homologues in any other species. For the other predicted proteins, it is notable that 2 encode proteins with a hydrophobic surface binding protein A (PF12296) domain, also found in chitinases (e.g. KID89971). One of these, g6474.t1, was mentioned above since it is preferentially expressed in spores, and may be important for the initial adhesion to and penetration of the nematode cuticle. This is consistent with the fact that the corresponding RNAseq reads were only found at the early time-point of infection.

Transformation and genome modification

The different *in silico* analyses reported above led to the identification of a very large number of candidate virulence genes. Addressing their functional importance would be greatly facilitated by the availability of techniques for the genetic transformation of *D. coniospora* and for targeted editing of its genome. By screening a number of different liquid media (C. Couillault, personal communication), we found that *D. coniospora* grew well in a rich, cholesterol-supplemented medium. We used fresh liquid cultures of *D. coniospora* mycelia to generate protoplasts, which were then transformed using a standard technique of polyethylene glycol (PEG)/ $CaCl_2$-mediated DNA uptake [77]. As a proof of principle, we transformed protoplasts with a plasmid (pLH4237) in which expression of a gene encoding a chimeric hygromycin B phosphotransferase::GFP protein [78] was driven by the *D. coniospora* ß-tubulin promoter (*ß-tub*p::HPH::GFP). The resultant recombinant fungus exhibited hygromycin resistance and strong GFP fluorescence in both spores and mycelia (Fig 8A).

To test the possibility of specifically knocking out a gene's function, we chose to target the *D. coniospora* homolog of the *so* (*soft*) gene (NCBI Gene ID: 3880225), required in other fungi for anastomosis (mycelial fusion) since this was expected to give a clear viable and visible phenotype [79], while at the same time not greatly altering virulence [80]. We therefore flanked our *ß-tub*p::HPH::GFP construct with arms homologous to the 5' and 3' regions of the *D. coniospora so* gene (*Dso*; g1469.t1) and used this construct (in pLH4256) to transform protoplasts. We obtained hygromycin-resistant GFP-expressing transformants in which, as demonstrated by PCR (S3 Fig) and sequencing, the *Dso* gene was replaced by the *ß-tub*p:: HPH::GFP cassette. The mutant exhibited the expected anastomosis defect. In contrast to the wild-type strain (Fig 8A and 8B, S2B and S2C Fig), neighbouring mycelia were never observed to fuse *in vitro* (Fig 8C) or during infection of *C. elegans* (Fig 8D and 8E). We have therefore the capacity to make targeted modifications of the *D. coniospora* genome.

Having fungal strains that express a fluorescent protein opens many new possibilities for future research. One immediate consequence is that we were able to follow the infection *in vivo* directly using fluorescence microscopy (Fig 8A, 8B, 8D and 8E). We also wondered whether this would offer a new way to quantify the progression of the infection. Fluorescence in *C. elegans* can be measured *in vivo* using the Complex Object Parametric Analyzer and Sorter (COPAS) Biosort. When the Biosort's Profiler is used, as well as a single measurement for each

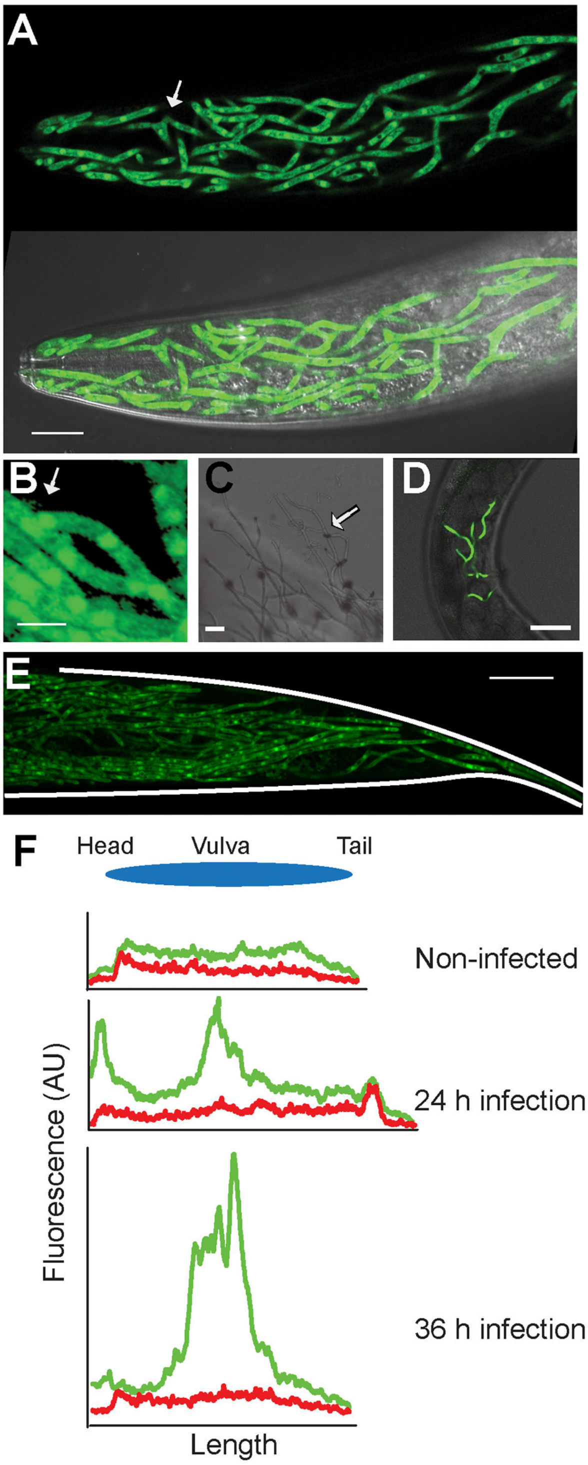


Fig 8. Transformation of *D. coniospora*. (A) A recombinant strain of *D. coniospora*, expressing GFP under the control of a ß-tubulin promoter, viewed by fluorescence microscopy (upper panel) combined with differential interference contrast microscopy (lower panel; scale bar, 20 μm). The fusion of 2 mycelia is highlighted by an arrow. (B) Higher magnification view of fused mycelia (scale bar, 5 μm). (C-E) The *Dso* mutant has a defect in anastomosis. (C) As highlighted with the arrow, in culture, mycelia are seen to grow across one another but never fuse. The mutant strain was engineered to express GFP constitutively. Fungal

mycelia growing in living animals, visualized using a stereo fluorescence dissecting microscope (D), or confocal fluorescence microscope (E; the shape of the worm is traced by white lines), were not observed to fuse. Worms had been infected overnight (D) or for 60 h (A, B, E) before images were taken. Scale bars in C, D, E: 20, 50, 40 μm. (F) The growth of the fungus can also be followed using the Profiler of the COPAS Biosort. The graphs show fluorescence profiles for green and red channels for an uninfected worm (top); a worm infected at the head and vulva (peak in green signal on the left and in the middle, respectively in the middle graph) and analysed after 24 h; another worm infected at the vulva (peak in green signal in the middle, bottom graph) and analysed after 36 h. Fluorescence and length are measured in arbitrary but constant units.

doi:10.1371/journal.pgen.1006017.g008

worm, one obtains a readout of fluorescence intensity along the length of the worm [81]. We found that the COPAS Biosort was sufficiently sensitive to allow us to follow the progression of the infection in a qualitative (Fig 8F) and potentially quantitative manner.

Initial characterisation of a potential fungal counter-defensive strategy

Our *in silico* analysis highlighted the atypical presence of 2 proteins with saposin A (PF02199) domains, one of which (g3895.t1) was predicted to be secreted (S11 Table). Mammalian saposins are synthesized as precursor proteins (prosaposin) that contain four Saposin-B domains (PF05184; PF03489) and two Saposin-A domains that are removed during the process of activation. Saposin-B domains also occur in proteins without Saposin-A domains across many species including nematodes. In *C. elegans*, they are represented by a large family of 23 proteins, also called caenopores [82, 83]. They are structurally similar to the innate defense proteins of the SAPLIP family, including vertebrate NK-lysin and granulysin [84]. Several of them are upregulated upon infection, some by multiple pathogens ([85] reviewed in [22]), including *spp-2*, *spp-6*, *spp-13*, *spp-14* and *spp-15* that are induced upon infection by *D. coniospora* [74]. A number of the SPP caenopores/saposins have been demonstrated to play a role in host innate immunity [86, 87] suggesting that certain SPP proteins could be also be direct effectors of anti-fungal defense.

The *D. coniospora* protein g3895.t1 is characterized by the presence of 3 Saposin-A domains, but no Saposin-B domain. There are currently no clear orthologs in any species in publically available databases. Given its unusual structure, we hypothesised that this protein, which we call here SapA, might act as an inhibitor of one or more nematode Saposin-B domain-containing caenopores/saposins and thereby interfere with host defense.

As a first step in the analysis of *sapA*, we chose to assay directly its expression during the infection of *C. elegans*. Using the *D. coniospora* actin gene as a control, by RT-PCR we observed a clear increase in the relative level of expression of *sapA* across the time-course of infection (S4A Fig). To define *in vivo* the spatio-temporal expression pattern of the corresponding protein, we made use of our capacity for transformation to produce recombinant fungus expressing the SapA protein tagged with dsRed at its C-terminus (SapA::dsRed), under the control of its own promoter. Strong dsRed expression was observed at the surface of spores, but not on mycelia early in the infection. Expression was then seen at the tips of growing hyphae at the moment when they approached the apical surface of the epidermis, before penetrating the cuticle from the inside (Fig 9A).

To test the hypothesis that the *D. coniospora* SapA might interact physically with one or more of the *C. elegans* caenopores/saposins, we incubated an extract of proteins purified from the *D. coniospora* strain expressing SapA::dsRed with 3 different purified recombinant *C. elegans* SPP proteins [83, 87], each possessing a C-terminal His-tag. The SapA::dsRed was then immunoprecipitated, together with any bound SPP protein, using an anti-dsRed antibody. To probe for a possible interaction with the SPP proteins, the immunoprecipitated material was analysed by Western blotting, using an anti-His-tag antibody. Although the 3 samples had

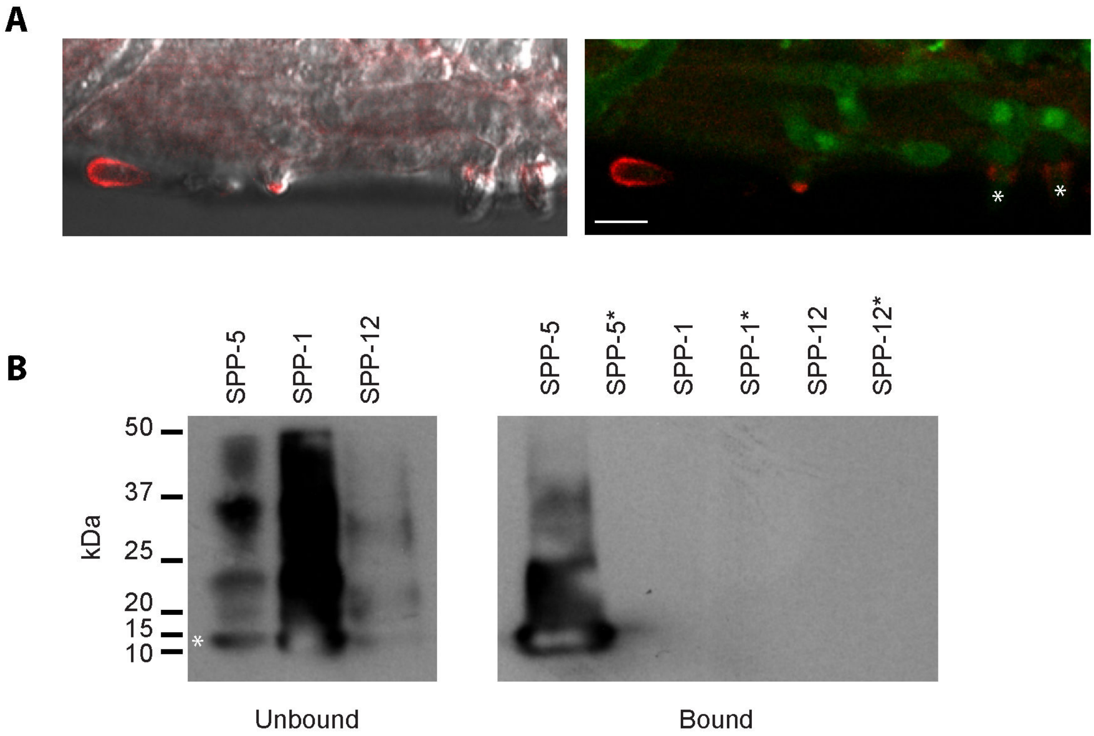


Fig 9. Saposin A-domain protein expression during infection of *C. elegans* and its *in vitro* interaction with SPP-5. (A) A strain of *D. coniospora* engineered to express GFP constitutively and a SapA::dsRed chimeric protein under the control of the *sapA* promoter, visualized using confocal fluorescence microscope (right panel) combined with differential interference contrast microscopy (left panel), during the infection of *C. elegans*. In both panels, a bright red fluorescent spore can be seen on the left. In the centre, a mycelium that is starting to exit the worm shows bright red fluorescence at its tip. On the right, 2 adjacent mycelia that have emerged can be seen. At the point where they leave the epidermis, a ring of SapA::dsRed can be seen, marked by asterisks in the right-hand panel. A general, less concentrated, signal can also be seen in the infected tissue. Worms had been infected for 60 h before images were taken. Scale bar = 5 μm. (B) Physical interaction between SPP-5 and SapA::dsRed. Recombinant His-tagged SPP-1, SPP-5 or SPP-12 was mixed with a protein extract from fungi expressing SapA::dsRed. The mix was analysed by Western blot probed with an anti-His-tag antibody before (left hand panel) or after (right hand panel) immunoprecipitation with anti-dsRed antibody-coated beads. In the sample before immunoprecipitation, in addition to the band at the expected size (11.9 kDa) marked by an asterisk, higher molecule weight species were detected, corresponding to the previously described oligomerization [86]. SPP-5 was co-immunoprecipitated with SapA::dsRed, principally in its monomeric form, but not if incubated with blocked beads (lanes marked with an asterisk; a control for non-specific binding). Neither SPP-1 nor SPP-12 gave any indication of co-immunoprecipitating with SapA::dsRed, even if SPP-1 was more abundant in the sample before immunoprecipitation.

doi:10.1371/journal.pgen.1006017.g009

been incubated with equal quantities of fungal protein, while there was no indication of any interaction between SapA::dsRed and SPP-1 or SPP-12, we observed a clear co-immunoprecipition of SapA::dsRed and SPP-5 (Fig 9B). Interestingly, SPP-5 is markedly more closely related to SPP-2 (79% similar) and the other infection-induced SPPs than are either SPP-1 or SPP-12 (S4B Fig).

Discussion

D. coniospora was first described 75 years ago as a parasite of *Rhabditis* nematodes in leaf mold, and then named *Meria coniospora* [88]. Its spores adhere to many different nematode species, and it is capable of infecting a relatively broad range of hosts including *C. elegans*, soybean cyst and root knot nematodes [14, 89–93]. While the different steps in the infectious process have been documented at the ultrastructural level [10, 94], nothing is known at the molecular level. Here, we have provided a first annotated genome sequence of *D. coniospora* that will serve as a starting point for future functional studies, as well as for more refined predictions of gene structure (S1 Text).

To make a comparative study of the nematophagous lifestyle, we chose 4 fungal species that use different strategies to infect nematodes. Like *D. coniospora*, *H. minnesotensis* is an endoparasitic fungus of the family Ophiocordycipitacea. It naturally infects soybean cyst nematodes using non-motile spores. Although it has never been reported to be one of its natural pathogens, it can infect *C. elegans* in the laboratory [9, 95]. *M. haptotylum* and *A. oligospora* are phylogenetically distant species that infect diverse nematodes, including *C. elegans*, after trapping them with their adhesive knobs and nets, respectively. *P. chlamydosporia* lies between these 2 pairs phylogenetically, infects nematode eggs, and in common with the nematode trapping fungi can grow as a saprophyte [3, 4, 6, 96–102].

Our phylogenetic analysis confirms the assignment of *D. coniospora* to the family Ophiocordycipitacea. It is interesting to note that other members of the Ophiocordycipitaceae family, *Haptocillium sphaerosporum* and *Harposporium spp*. are also pathogens of *C. elegans* [20, 74]. Among sequenced fungi, *D. coniospora*'s closest relative is currently the truffle-parasite *Tolypocladium ophioglossoides*. Our analysis thus provides a further illustration of the polyphyletic nature of nematode parasitism and will contribute to the on-going debate regarding the acquisition of host specificity in pathogenic fungi [6, 67].

One factor that can contribute to the emergence of virulence traits is the amplification of transposable elements that facilitate genome rearrangements. The *H. minnesotensis* genome represents an extreme example since 35% of its genome was reported to be composed of transposable elements [9]; in *D. coniospora* they make up less than 2%. *D. coniospora* also has a substantially reduced range of glycoside hydrolase (GH) enzymes compared to the nematode-trapping fungi, in direct contrast to the extensive repertoire of *P. chlamydiosporia* [98]. This presumably reflects a decreased capacity to adapt to diverse environments [103]. Unlike *P. chlamydiosporia* that can infect plants as well as nematodes, *D. coniospora* is an obligate nematode-specific parasite [104]. This slimming down of the genome is reflected in diverse other protein families. Thus, in addition to the different families described above, for example, there are few multi copper oxidases and lytic polysaccharide mono-oxygenases, so-called "auxiliary activities" linked to lignocellulose conversion [105]. Its genome also contains fewer genes related to sugar/inositol transport, which are involved in the establishment of plant-fungus relationships in *M. anisopliae* [106]. Thus *D. coniospora* does not show the pattern of gene family expansions observed in characterized nematode-trapping fungi, which are more similar to those seen in plant pathogens than to insect and animal pathogens [5].

An atypical loss from the *D. coniospora* genome is of genes encoding NACHT-domain proteins, which as mentioned above are present in multiple copies in all the 11 other fungal species analysed. The NACHT domain is a constituent domain in one of the two main classes of NOD domain proteins. The second class has the NB-ARC (PF00931) domain, a signalling motif shared by plant resistance gene products and regulators of cell death in animals [107]. These too are absent from the predicted set of *D. coniospora* proteins. Again this is highly unusual, since they are found often in all the other species, (8, 9, 10 and 15 times in *M. haptotylum*, *A. oligospora*, *P. chlamydosporia* and *H. minnesotensis*, respectively). Indeed, NOD domain proteins are present broadly among fungi and have important roles in fungal non-self recognition and in defence systems [60]. Determining the reasons and consequences of this intriguing loss of NOD domain proteins from *D. coniospora* remains a challenge for the future.

On the other hand, both insect and nematode parasitism requires a broad set of genes involved in detoxification and resistance to oxidative stress, such as glutathione S-transferases, cytochrome P450 genes, and carboxylesterases [40]. These are present in the *D. coniospora* genome in similar numbers to its entomopathogenic neighbors. More than 100 protein kinases were detected. Many are likely to regulate the infection process, as previously described, for example, for a *M. anisopliae* protein kinase A [108] for which there is a direct *D. coniospora* ortholog (g2806.t1). They may in turn be regulated by different families of transcription factors, including the Zn_2Cys_6 fungal-type, bZIP and bromodomain-containing families, which are all well-represented in the *D. coniospora* genome, and are often involved in the control of gene expression linked to virulence [109].

As previously observed for other nematophagous species, the specialization of *D. coniospora* is also reflected in the high degree of structural innovation in its predicted proteome. Thus, for example, in *H. minnesotensis*, 11% of proteins were described as species-specific and lacking any recognisable domain [9]. In *D. coniospora* the corresponding figure is currently 16.3%. With increased sampling of closely related species, these figures will drop in the future. Nevertheless, in addition to the many readily predictable virulence factors, there is a wealth of novel biology to be explored using *D. coniospora*.

Regarding the broad spectrum of potential virulence effectors, certain merit further discussion. The *D. coniospora* genome is predicted to encode multiple heat-labile enterotoxins. The presence of bacterially-derived enterotoxins in entomopathogenic fungi such as *B. bassiana* already presented a conundrum as these species are assumed to lack *per os* infectivity [53]; the same is true for *D. coniospora*. Among them, at least 10 are predicted to be secreted (Tables 4 and S6 and S9), so may conceivably be delivered into the host cytoplasm where they would be expected to perturb cellular homeostasis. An alternative explanation is that they may play a role in *D. coniospora*'s interactions with other microbes. Competition has already been reported between *D. coniospora* and *A. oligospora* during nematode infection [110]; antagonistic effects are likely to exist with other species and could rely on enterotoxin production.

Subtilisin-like serine proteases were mentioned several times above. This family has been linked to fungal virulence in nematophagous fungi in some studies (e.g. [8, 73, 111]). Consistent with such a link, subtilisin genes are highly expressed by *A. oligospora* and *M. haptotylum* during infection of the nematode *C. briggsae* [5]. There are 29 genes predicted to encode subtilisins in *D. coniospora*. This family is, however, well represented in all the fungal species analysed. For the 11 other species, the number ranges from 12 to 56, with a median value of 31. Since related pathogenic and non-pathogenic species can show the same type of gene expansion, and thus gene number is not correlated with pathogenicity, it has been suggested that the number of serine proteases in a species is related primarily to their role in digestion, whether or not the food source is dead or alive [112].

The genome also contains multiple cysteine-rich secretory protein family genes. These are frequently associated with fungal host adaptation or specialization [67]. As mentioned above, one was found to be preferentially expressed during infection of *C. elegans*, despite the poor coverage of the fungal transcriptome. This coverage was in fact remarkably low, despite deep sequencing, with only 0.0011% and 0.0012% of 77 million and 123 million RNAseq reads corresponding to *D. coniospora* transcripts from samples of worms infected for 5 and 12 hours, respectively. Clearly, reliable profiling of the *D. coniospora* transcriptome at early time points will require the development of methods to enrich fungal mRNA from samples of infected worms. Fortunately, several methods already exist (e.g. [113]); adapting them to this model will require further work.

Such an analysis is likely to be necessary to prioritize the overwhelming number of candidate virulence factors for in-depth functional study. This is particularly true in cases where gene families have expanded and where there is therefore the possibility of functional redundancy. Characterizing expression profiles can allow genes with non-overlapping patterns to be identified; they are less likely to be redundant. While we have shown that *D. coniospora* can be readily genetically transformed, this remains a relatively time-consuming process that cannot be implemented on a large scale. Indeed, further work is also needed to develop functional assays to assess the role of the many genes potentially involved in pathogenesis or currently lacking any predictable role. The COPAS Biosort allows the analysis of hundreds of worms each minute. Having the possibility to transform *D. coniospora* with a fluorescence protein encoding reporter gene thus provides a non-invasive, non-destructive means to following infection at the level of individual worms at a large scale. It represents a first method of the type that will need to be applied to probe *D. coniospora* gene function during its infection of *C. elegans*.

We have also given one example of how tagging a protein fluorescently can constitute a first step in its functional characterization. We were able to determine the *in vivo* expression pattern for SapA::dsRed and demonstrate an *in vitro* interaction between it and the host antimicrobial effector SPP-5. We do not yet know whether this interaction mirrors an *in vivo* interaction. We currently favour the idea that SapA::dsRed will be capable of interacting with the host SPP caenopores/saposins that are up-regulated upon infection and thereby inhibit their antimicrobial activity. SPP-5 is very similar in sequence to SPP-2, SPP-6 and SPP-14 that are differentially regulated in *D. coniospora* infected worms. Clearly much work remains to be done to clarify the significance of these observations, including establishing which if any of the host SPP caenopores/saposins bind SapA *in vivo*, and determining whether this has any consequence for the progression of the fungal infection. Given the multiplicity of SPP proteins in *C. elegans*, this remains a substantial challenge for the future. Clearly, our analysis of the *D. coniospora* genome, and the tools we have developed, have opened many avenues for future investigation of this fungus's antagonistic interaction with *C. elegans*.

Materials and Methods

D. coniospora culture and nucleic acid purification

The *D. coniospora* strain ATCC 96282 (the kind gift of Hans-Börge Jansson; S1 Text) was either cultured at 25°C on solid Nematode Growth Medium (NGM) in the presence of *C. elegans* as previously described [114], or in liquid NGM with Yeast extract (NGMY see S1 Methods) in the absence of nematodes. For the extraction of fungal genomic DNA, roughly 10^9 fresh spores were inoculated in 100 mL NGMY medium. After 7 days culture, mycelia were harvested by vacuum filtering through a sterile 10 μm nylon membrane. The filter was placed in a 30°C incubator for around 3 h, until the mycelia were dry. The mycelia were then manually ground in a liquid nitrogen-cooled mortar. A 20 mg aliquot was transferred to a 1.5 mL tube and the DNA extracted using the DNeasy Plant Mini Kit (Qiagen), following the manufacturer's instructions.

To prepare samples for RNA extraction, fresh spores were first collected from 10 cm NGM plates of infected worms as previously described [114] and inoculated in 100 mL of NGMY liquid medium, and either cultured for 4 days before harvesting, or serially cultured (5 x 8 d cultures). For this, a 1 mL aliquot from an 8 d culture was strongly agitated for 5 min to disrupt the tight balls of mycelia and inoculated into fresh 100 mL NGMY liquid medium then cultured for a further 8 days. In both cases, samples were collected by filtering cultures through 0.22 μm Steritop units (Milipore), flash frozen in liquid nitrogen and stored at -80°C. Aliquots of 200 mg of spores or mycelia were dissolved in 1 mL Trizol (Invitrogen) in lysing matrix tubes (MP Biomedicals), homogenized for 20 s at 6 m/s in a FastPrep-24 (MP Biomedicals), incubated on ice for 2 min and homogenized a second time. RNA was purified using a standard protocol [115] and cleaned with an RNeasy mini Kit (Qiagen).

Library construction and sequencing

MIseq paired-end libraries were prepared from genomic DNA and sequenced on an Illumina MIseq sequencer as 2 x 150 bp paired-end reads following the manufacturer's standard procedures. For SOLiD sequencing, two mate-paired libraries were prepared from *D. conisopora* genomic DNA following the manufacturer's instructions (Mate-Paired Library Preparation user guide 4460958 Rev.B; Life Technologies, Carlsbad, USA). DNA was sheared to 1.5 kb or 3 kb fragments with Covaris System and Covaris Blue miniTUBES. Following nick-translation for 9.5 min, to generate fragments of < 400 bp, the library was amplified through 14 rounds of PCR. Fragments between 250 and 300 bp were size-selected using 4% acrylamide gel and purified with SOLiD Library Micro Column Purification Kit prior to conversion for analysis with the 5500 WildFire system (Life Technologies, Carlsbad, USA).

cDNA library construction and Illumina sequencing of mRNA from spores and mycelia using an Illumina HiSeq 2000 platform was performed at the Beijing Genomics Institute (Shenzhen, China; http://www.genomics.cn/index.php) using their standard pipeline. More than 48 million 90 bp paired-end reads were obtained from each 200 bp insert library.

De novo genome and transcriptome assembly

MIseq reads were processed with BBDuk software, part of BBMap suite (http://sourceforge.net/projects/bbmap/) to filter out contaminants and low quality reads. The remaining reads were used for *de novo* genome assembly with Velvet [34], SPAdes [35], SOAPdenovo2 [36] and ABySS [37] using standard input parameters except for ABySS for which k values of 64 and 96 were used. The resulting assemblies were then scaffolded using SOLiD mate-paired reads. Only very high quality reads were used for this scaffolding step using SSPACE v2.0 [38] with the k parameter set to 5 (default value).

For use subsequent in gene prediction (see below) a *de novo* transcriptome assembly was performed on the combined sets of reads from the two sequenced libraries (from spores and mycelia) with Trinity [116] using default parameters.

Optical mapping

A whole genome map of *D. coniospora* was generated using the Argus Whole-Genome Mapping System (www.opgen.com). To obtain size-optimized restriction fragments (6–12 kb on average and no fragment larger than 80 kb across the genome) we used Enzyme Chooser (OpGen Inc., Gaithersburg, MD) that led to the selection of *Xba* I. We sorted out 42,622 molecules longer than 200 kb (average 310 kb) used for the assembly, performed with MapSolver software (www.opgen.com). The resulting map contigs were manually validated leading to 9 maps (ranging from 0.58 to 11.3 Mb) with a cumulative size of 31.8 Mb. These were used for comparisons with the *in silico Xba* I digestion profile of the scaffolds obtained after sequencing.

Functional annotation

Genome annotation was performed using standard open source software. Repetitive elements were mined using RepeatScout version 1.0.5 [42]. A specific repeat library was generated in order to mask the genome with RepeatMasker version 4-0-5 (Smit, Hubley, & Green, RepeatMasker Open-4.0; www.repeatmasker.org). TransposonPSI v08222010 (transposonpsi.sourceforge.net) was used to characterize different types of transposable elements. Non-coding RNA were identified using Rfam scan perl script v1.0 [117] and tRNAscan-SE v1.3.1 [43] for transfer RNAs.

Gene prediction and annotation

We performed a first round of gene prediction using Augustus [45], trained using the Trinity-derived *D. coniospora* transcripts to predict 8111 protein-coding genes. In a second round, we used Augustus trained using *Fusarium graminearum* (available from http://bioinf.uni-greifswald.de/augustus/) and retained 631 additional predicted proteins that were either (i) conserved, (ii) supported by RNAseq data or (iii) contained a PFAM domain. Inspection of the combined set led to the removal of 9 aberrant proteins (S1 Text), giving a final set of 8733 predicted proteins. Subsequent inspection of the alignment of the unassembled RNAseq reads and the Trinity-derived transcripts to the predicted gene models revealed occasional inconsistencies. Thus as with any first round genome-wide gene prediction and sequence annotation, some errors will need to be resolved in future versions of the genome (see S1 Text).

The set of Augustus-predicted protein-coding genes were annotated using Interproscan [118], release 5.16–55 (data package 55), with Hamap (201511.02), ProDom (2006.1), PIRSF (3.01), PANTHER (10.0), Pfam (28.0), SMART (6.2), Gene3D (3.5.0), Coils (2.2.1), ProSiteProfiles (20.113), TIGRFAM (15.0), PRINTS (42.0), SUPERFAMILY (1.75), and ProSitePatterns (20.113).

The same protocol of InterProScan annotation was performed to annotate the predicted set of proteins from eleven additional fungi. The names indicated are from the NCBI Taxonomy Database; commonly used synonyms and/or NCBI genome assembly accession numbers are in brackets: *Arthrobotrys oligospora* (ADOT00000000.1) [8], *Fusarium graminearum* (*Gibberella zeae*; AACM00000000.2) [119], *Fusarium oxysporum* (AAXH00000000.1) [120], *Hirsutella minnesotensis* (JPUM00000000.1) [9], *Metarhizium acridum* (ADNI00000000.1) [40], *Metarhizium anisopliae* (AZNF00000000.1) [40], *Monacrosporium haptotylum* (*Dactylellina haptotyla*; AQGS00000000.1) [5], *Ophiocordyceps sinensis* (ANOV00000000.1) [41], *Pochonia chlamydosporia* (*Metacordyceps chlamydospora*; AOSW00000000.1) [98], *Tolypocladium ophioglossoides* (*Elaphocordyceps ophioglossoides*; LFRF00000000.1) [64] and *Trichoderma reesei* (*Hypocrea jecorina*; AAIL00000000.2) [121].

SignalP v4.1 [122], TargetP v1.1 [123] and Tmhmm v2.0 [124] were used to predict respectively signal peptide, target peptide and transmembrane domains. A Blast analysis (BLASTP with e value $< 10^{-5}$) versus PHI-base proteins [65] was performed to associate *D. coniospora* genes to experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens.

Phylogenetic analysis

We refined the set of BUSCO-defined orthologues by restricting it to proteins that did not differ by more than 10% in total length across all 12 fungal species. This left us with a set of 97 high-confidence orthologous proteins present in all species. The respective sequences were concatenated were aligned using MAFFT [47] and phylogenetic distances calculated using the maximum likelihood-based method implemented within PhyML [48]. Altering the order of concatenation had no influence on the calculated phylogenetic distances. These analyses were performed within the Mobyle Web environment [125] at http://mobyle.pasteur.fr and the output plotted using the tree drawing engine implemented in the ETE toolkit [126].

CAZy analysis

Each *D. coniospora* protein model was compared using BLASTP [127] to proteins listed in the CAZy database (www.cazy.org; [128]). Because the e-value depends on the length of the aligned segment (for instance a 30% sequence identity results in widely different e-values, from non-significant to highly significant, if the two aligned proteins are 40, 100, 250 or 500 residues in length), CAZy family assignments rather included examination of sequence conservation (percentage identity over CAZy domain length). Proteins that gave more than 50% identity over the entire domain length of an entry in CAZy were directly assigned to the same family. Proteins with less than 50% identify to a protein in CAZy were all manually inspected and conserved features such as catalytic residues were searched. The variable modular structure of CAZymes was integrated by performing alignments with isolated functional domains [129]. The same methods were used for all fungi that were compared to *D.* *coniospora*. For clustering of protein families and domains, we used “One minus Pearson correlation” distance matrices within GENE-E (www.broadinstitute.org/cancer/software/GENE-E/).

Secretome and comparative PFAM analysis

Secretome analysis was carried out as previously described [130] by combining predictions from SignalP v4.1 [122], WolfPSORT [131,132] and NucPred [133]. A WolfPSORT search using mature secreted proteins and model ‘ANIMAL’ was used to determine probable target protein localization in host. For sequence-based clustering, a database consisting of predicted proteomes of *D. coniospora*, *H. minnesotensis*, *P. chlamidosporia*, *M. haptotylum*, *A. oligospora* and PHIbase v3.6 [65] entries was built. The result of a BLASTP search of this database against itself with an e-value cutoff of 10^{-30} was used as input for clustering with the MCL program in Biolayout Express 3D [134]. Some of these clusters (S11 Table) were projected onto the Circos plot (Fig 1, in red, purple and green, respectively): OG5_126718 as putative nonribosomal peptide synthetases; OG5_127207 (serine carboxypeptidase S28), OG5_138644 (deuterolysin metalloprotease (M35) family), OG5_137388 (PA domain; subtilase family), OG5_128249 (subtilase family), OG5_149879 (serine carboxypeptidase), together with orthoMCL-defined paralogGroup 11 (subtilisin-like serine protease; S6 Table) as diverse proteases; paralogGroups 1, 4, 9 and 17 as enterotoxin-like proteins. Clusters that did not contain predicted secreted proteins from any of the 5 fungal species were discarded. Correspondence analysis was performed using the FactomineR package in R.

TargetP v1.1 [123] was used to determine probable protein localization, using mature proteins for secreted proteins or full-length sequences otherwise. A consensus predicted localization was derived using the following rules: TargetP predictions with reliability < = 3 only were considered; NLS were considered if predicted by NLStradamus and with a NucPred score <0.6; PredGPI predictions were considered if probability > = 90. PFAM domains were identified through a search against PFAM 28.0 database using gathering thresholds.

Gene expression analysis

For samples from mycelia and spores, Illumina paired-end (2 x 90 bp) RNAseq reads were aligned to the genome assembly using STAR [135]. Reads were assigned to the 8733 gene models using the htseq-count script within HTSeq [136]. To establish lists of differentially regulated genes, we used previous described methods [74] and retained the genes that were commonly defined by both. For samples from *C. elegans* infected with *D. coniospora*, the unaligned reads (640069 and 237794 reads from 5 and 12 h samples, respectively; kindly provided by LaDeana Hillier) from a previous RNAseq analysis [74] were aligned and assigned to gene models as above.

Protoplast preparation and fungal transformatn

N2 worms at the L4 stage were infected with fungal spores as described [114] on NGMY plates spread with the *E. coli* strain OP50 and incubated for 24 h, then transferred into NGMY liquid medium and cultured for another 30 h. Mycelia were collected by filtration as above and protoplasts prepared and transformed using polyethylene glycol (PEG)/$CaCl_2$-mediated DNA uptake as described [137], using expression vectors (see S1 Methods) containing a hygromycin selection marker, derived from pPK2*hphgfp* [78], a kind gift from Martijn Rep (Swammerdam Institute for Life Sciences, Amsterdam). Transformants were selected for antibiotic resistance on medium containing 15 µg/ml hygromycin and screened for fluorescence.

Spore protein lysates and pull-down assays

Worms were infected on NGM plates with spores from SapA::DsRed expressing fungus. After 15 to 30 day culture at 25°C, spores were harvested in 50 mM NaCl as previously described [114]. They were extensively washed in cold 50 mM NaCl, pelleted and resuspended in an equal volume of lysis buffer (50mM Tris-Cl, pH 7.5, 100mM NaCl, 3 mM MgCl2, 0.5% Triton X-100, protease inhibitors (Complete, Roche), 5% glycerol) and flash-frozen. They were then sonicated on a high setting for 7 minutes (30 sec on and 30 sec off; Bioruptor, Diagenode,) and vortexed for 5 min with acid-washed glass beads (Sigma). The supernatant from a high-speed centrifugation was used as a whole protein extract for pull-down assays. For this, 800 µg of protein extracts were incubated with 20 µg of purified His-tagged SPP-1, SPP-5, or SPP-12 (the generous gift of M. Leippe, Kiel university), for 2 h. Preformed complexes were then immunoprecipitated with anti-dsRed/RFP agarose beads (Chromotek, RFP-Trap), at 4°C, overnight. As a binding control, preformed complexes were incubated under identical conditions with blocked matrix beads. After incubation, beads were washed three times in lysis buffer and three times in wash buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% NP-40, 1 mM $MgCl_2$). After a final wash (1 mM Tris-Cl, 150 mM NaCl and 1 mM $MgCl_2$) beads were eluted in SDS-PAGE loading buffer. Samples were then resolved on a 4–12% gradient denaturating PAGE gel and transferred onto a membrane. The blot was probed with anti-His antibody (Upstate, H8 clone) or anti-dsRed/RFP antibody (Rockland, Inc). Blots were visualized using enhanced chemiluminescence (SuperSignal West Pico, Pierce).

Nucleotide sequence accession numbers

MIseq reads used for *de novo* genome assembly and Hiseq reads used for RNAseq *de novo* have been deposited at SRA under accession numbers SRX883538 and SRX969055, respectively. The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JYHR00000000. The version described in this paper is version JYHR01000000.

Supporting Information

S1 Table. Descriptive statistics of the different assemblies before and after SSPACE scaffolding. Correspondence of the 9 optical maps (named as chromosomes) to the scaffolds from Velvet (kmer = 63), SOAPdenovo (kmer = 63), Spades (kmer = 127), and ABySS (kmer = 64 and 96), before and after scaffolding with SSPACE. Two values are provided, which correspond to the number of scaffolds localized on each optical map and the total coverage in Mb. The total number of mis-assemblies observed within each analysis is also indicated. ND: not determined.
(XLSX)

S2 Table. Comparative analysis of transposable elements in the genome of *D. coniospora* and of 11 other fungal species. The sources for the genome sequences used in this analysis are given in the Materials and Methods.
(XLSX)

S3 Table. Comparative analysis of tRNAs in the genome of *D. coniospora* and of 11 other fungal species.
(XLSX)

S4 Table. Species and sequences used for phylogenetic analysis.
(XLSX)

S5 Table. InterproScan analysis of the 8733 proteins predicted from the *D. coniospora* genome. The first sheet presents the results of an InterproScan analysis of the predicted proteome, complemented with dedicated BLASTP searches against UniprotKB and PHI-base databases. The second gives the sequences of 31 predicted proteins longer than 1000 amino acids, but without any InterproScan annotation. The third sheet gives a summary of the BLASTP analysis against the NCBI non-redundant set of proteins for the 31 sequences on the previous sheet. The results for the predicted proteins are split into 3 groups: with a homologue (only top 10 hits shown), with only remote similarity to other proteins (all hits shown), and with no similarity to any protein in the current NCBI non-redundant protein sequence database.
(XLSX)

S6 Table. OrthoMCL analysis of the 8733 proteins predicted from the *D. coniospora* genome. The first sheet lists the species present in OrthoMCL, with the frequency of top hits per species. The second sheet indicates the top OrthoMCL hit for the 7092 *D. coniospora* proteins assigned to an OrthoMCL cluster. The third sheet gives the number of predicted proteins in each OrthoMCL cluster, with corresponding PFAM domain annotation (from S5 Table) for those with more than 3 members. The fourth sheet lists the members of the clusters determined by OrthoMCL to be specific to *D. coniospora* (i.e. not represented in the 150 species in the first sheet). The fifth sheet gives the CLUSTAL Omega alignments for the 15 predicted proteins in the OrthoMCL paralogous group 1 (from fourth sheet).
(XLSX)

S7 Table. Comparative analysis of the occurrence of PFAM domains in the 8733 proteins predicted from the *D. coniospora* genome and in the predicted proteins from 11 other fungal species. Only PFAM domains present in at least one species are shown. The first sheet shows all the data. The second and third, extracted from the first, shows PFAM domains absent and present in *D. coniospora*, respectively.
(XLSX)

S8 Table. Detailed analysis of *D. coniospora* CAZy proteins and comparative analysis with the predicted CAZy proteins from 10 other fungal species. The first sheet indicates the classification of the predicted 240 CAZy proteins in *D. coniospora*, with notes from manual annotation. The following sheets show the occurrence of different CAZy family proteins in *D. coniospora* and 10 other fungi. Only CAZy families present in at least one species are shown.
(XLSX)

S9 Table. Predicted *D. coniospora* proteins involved in the production of secondary metabolites. The first sheet lists the proteins identified by SMURF analysis as 'backbone'; the second lists the proteins that putatively form part of a functional cluster.
(XLSX)

S10 Table. Analysis of hits in the PHI-base database for the 8,733 proteins predicted from the *D. coniospora* genome. The first sheet gives the number of times a given PHI-base entry was returned as the top hit for a *D. coniospora* protein (indicated in the column 'Occurrence'; data derived from the column 'PHIBASE' in S5 Table). The second sheet lists the PHI-base

entries for those hit by 5 or more predicted *D. coniospora* proteins. Certain indicated functional categories are highlighted in colour. The third sheet is an extract of the InterproScan data in S5 Table for all the *D. coniospora* proteins listed in the second sheet, and the following 7 sheets are extracts for the proteins corresponding to the PHI-base entries hit at least 10 times.
(XLSX)

S11 Table. Details of the secretome analysis for the 670 *D. coniospora* predicted with high confidence to be secreted. The first sheet lists the *D. coniospora* proteins predicted to be secreted, together with the scores from the analysis programs, and annotations taken from S5 Table. The second sheet gives the data sources for the comparative analysis. The last sheet shows the results of sequence based clustering of proteins from five nematophagous fungal species. Only clusters including at least one predicted secreted protein are considered.
(XLSX)

S12 Table. RNAseq analysis of transcripts from spore and mycelial samples. The first sheet gives the read counts for each gene from the 2 different samples. The 50 most highly expressed genes for each sample are highlighted. In the next 2 sheets they are listed together with annotations taken from S5 Table. The fourth and the fifth sheets list the genes assigned to the over-represented category in spores and in mycelia, respectively, together with annotations taken from S5 Table. On the fourth sheet, neighbouring genes are highlighted in yellow. The sixth and seventh list their respective constituent PFAM domains, with a score that reflects the confidence of the assignment. Seven domains are found in both lists (PF00005, PF00083, PF00172, PF00501, PF05730, PF07690, PF13193).
(XLSX)

S13 Table. RNAseq analysis of *D. coniospora* transcripts from samples extracted from infected *C. elegans*. The first sheet gives the raw read counts for genes covered by at least one RNAseq read from the 2 samples, listed by name and total read counts. The second sheet lists the genes covered by at least 3 reads, gives the read counts for the samples from spores and mycelia (data from S12 Table; genes in the top 15th percentile highlighted) and includes different functional annotations (from S5 Table).
(XLSX)

S1 Fig. RADAR analysis [54] reveals the repeated structure in the sequence of g8068.t1, a 1045 a.a. protein from OrthoMCL-defined paralogous group 2 (see S6 Table).
(PDF)

S2 Fig. Photomicrographs of mycelial (A, B, C) or spore (D) preparations of *D. coniospora*, taken shortly before processing for RNA extraction. (A) Under the conditions of liquid culture used, *D. coniospora* forms compact balls of up to several mm in diameter. (B, C) At a higher magnification, it can be seen that the mycelia are devoid of spores and the fusion of hyphae can be clearly observed (white arrows). (D) While the majority of spores have started to germinate (red arrows), some have not (white arrows). A smaller proportion is not mature, lacking the adhesive bud (yellow arrows). Scale bars (white) in C and D, 10 μm.
(PDF)

S3 Fig. PCR-based verification of the insertion of a hygromycin-resistance expression cassette into the *Dso* locus. The top part of the figure shows the position of PCR primers relative to the genomic and recombinant DNA sequences. The 2 tables indicate the expected sizes and occurrences of PCR amplicons. The bottom part of the figure shows that the expected bands are obtained from the wild-type (WT) and knocked-in strain (Dso).
(PDF)

S4 Fig. (A) PCR products from reverse-transcribed mRNA corresponding to the *D. coniospora* saposin A-domain protein-encoding gene g3895.t1 (SapA) and the actin gene g2551.t1 (Actin) from mycelia and at the indicated times post-infection (p.i.). The size markers in the outside lanes are, from top to bottom, 300, 200 and 100 bp. (B) Clustal multiple alignment of infection-induced saposin proteins (in bold) and those used for to assay for a possible interaction between a host saposin and the fungal SapA protein.
(PDF)

S1 Text. Contains comments on the quality of the genome sequence and gene annotation, as well as a description of the isolation history of ATCC 96282 and its derivatives.
(PDF)

S1 Methods. Contains details of media, plasmid constructions and primer sequences.
(DOCX)

Acknowledgments

We thank Jérôme Gouzy for his input on an initial round of genome annotation and Aude Barani, Jérôme Belougne, Carole Couillault, Stéphane Coulon, Jan Dijksterhuis, Laurent Hannouche, LaDeana Hillier, Matthias Leippe, Luis Lopez-Llorca, Hans-Börge Jansson, Martijn Rep and particularly Nathalie Pujol for their contributions, advice and/or the generous gift of reagents. We also thank the anonymous reviewers for their constructive criticisms.

Author Contributions

Conceived and designed the experiments: KL JP ER VB SR PB JJE. Performed the experiments: KL LDH NT MJA JP GM SR JJE. Analyzed the data: KL BH NT ER SR PB JJE. Contributed reagents/materials/analysis tools: KL NT SR. Wrote the paper: KL NT BH ER PB JJE.

References

1. Barron GL. Nematophagous destroying fungi. Guelph: Lancester press; 1977. 1–140 p.
2. Linford MB. Stimulated activity of natural enemies of nematodes. Science. 1937; 85:123–4. PMID: 17754209
3. Ahren D, Tholander M, Fekete C, Rajashekar B, Friman E, Johansson T, et al. Comparison of gene expression in trap cells and vegetative hyphae of the nematophagous fungus *Monacrosporium haptotylum*. Microbiology. 2005; 151(Pt 3):789–803. PMID: 15758225
4. Fekete C, Tholander M, Rajashekar B, Ahren D, Friman E, Johansson T, et al. Paralysis of nematodes: shifts in the transcriptome of the nematode-trapping fungus *Monacrosporium haptotylum* during infection of *Caenorhabditis elegans*. Environ Microbiol. 2008; 10(2):364–75. PMID: 18028414
5. Meerupati T, Andersson KM, Friman E, Kumar D, Tunlid A, Ahren D. Genomic mechanisms accounting for the adaptation to parasitism in nematode-trapping fungi. PLoS Genet. 2013; 9(11):e1003909. Epub 2013/11/19. doi: 10.1371/journal.pgen.1003909 PMID: 24244185
6. Andersson KM, Kumar D, Bentzer J, Friman E, Ahren D, Tunlid A. Interspecific and host-related gene expression patterns in nematode-trapping fungi. Bmc Genomics. 2014; 15:968. Epub 2014/11/12. doi: 10.1186/1471-2164-15-968 PMID: 25384908
7. Liu K, Zhang W, Lai Y, Xiang M, Wang X, Zhang X, et al. *Drechslerella stenobrocha* genome illustrates the mechanism of constricting rings and the origin of nematode predation in fungi. Bmc Genomics. 2014; 15:114. Epub 2014/02/11. doi: 10.1186/1471-2164-15-114 PMID: 24507587
8. Yang J, Wang L, Ji X, Feng Y, Li X, Zou C, et al. Genomic and Proteomic Analyses of the Fungus *Arthrobotrys oligospora* Provide Insights into Nematode-Trap Formation. PLoS Pathog. 2011; 7(9): e1002179. Epub 2011/09/13. doi: 10.1371/journal.ppat.1002179 PMID: 21909256
9. Lai Y, Liu K, Zhang X, Li K, Wang N, Shu C, et al. Comparative genomics and transcriptomics analyses reveal divergent lifestyle features of nematode endoparasitic fungus *Hirsutella minnesotensis*. Genome Biol Evol. 2014; 6(11):3077–93. Epub 2014/11/02. doi: 10.1093/gbe/evu241 PMID: 25359922

10. Dijksterhuis J, Veenhuis M, Harder W. Ultrastructural study of adhesion and initial stages of infection of the nematode by conidia of *Drechmeria coniospora*. Mycological research. 1990; 94(1):1–8.
11. Rouger V, Bordet G, Couillault C, Monneret S, Mailfert S, Ewbank JJ, et al. Independent Synchronized Control and Visualization of Interactions between Living Cells and Organisms. Biophysical journal. 2014; 106(10):2096–104. doi: 10.1016/j.bpj.2014.03.044 PMID: 24853738
12. Gernandt DS, Stone JK. Phylogenetic analysis of nuclear ribosomal DNA place the nematode parasite, *Drechmeria coniospora*, in Clavicipitaceae. Mycologia. 1999; 91(6):993–1000.
13. Quandt CA, Kepler RM, Gams W, Araujo JP, Ban S, Evans HC, et al. Phylogenetic-based nomenclatural proposals for *Ophiocordycipitaceae* (*Hypocreales*) with new combinations in *Tolypocladium*. IMA Fungus. 2014; 5(1):121–34. doi: 10.5598/imafungus.2014.05.01.12 PMID: 25083412
14. Jansson HB, Jeyaprakash A, Zuckerman BM. Differential adhesion and infection of nematodes by the endoparasitic fungus *Meria coniospora* (*Deuteromycetes*). Appl Envir Microbiol. 1985; 49:552–5.
15. Coles GC, Dicklow MB, Zuckerman BM. Protein changes associated with the infection of the nematode *Caenorhabditis elegans* by the nematophagous fungus *Drechmeria coniospora*. Int J Parasitol. 1989; 19:733–6.
16. Fuchs BB, Mylonakis E. Using non-mammalian hosts to study fungal virulence and host defense. Curr Opin Microbiol. 2006; 9(4):346–51. PMID: 16814595
17. Ewbank JJ, Zugasti O. *C*. *elegans*: model host and tool for antimicrobial drug discovery. Dis Model Mech. 2011; 4(3):300–4. Epub 2011/04/21. doi: 10.1242/dmm.006684 PMID: 21504910
18. Clark LC, Hodgkin J. Commensals, probiotics and pathogens in the *Caenorhabditis elegans* model. Cellular microbiology. 2014; 16(1):27–38. Epub 2013/10/31. doi: 10.1111/cmi.12234 PMID: 24168639
19. Cohen LB, Troemel ER. Microbial pathogenesis and host defense in the nematode *C*. *elegans*. Curr Opin Microbiol. 2015; 23C:94–101. Epub 2014/12/03.
20. Labed S, Pujol N. *Caenorhabditis elegans* Antifungal Defense Mechanisms. The Journal of Invasive Fungal Infection. 2011; 5(4):110–7.
21. Ewbank JJ, Pujol N. Local and long-range activation of innate immunity by infection and damage in *C*. *elegans*. Curr Opin Immunol. 2016; 38:1–7. doi: 10.1016/j.coi.2015.09.005 PMID: 26517153
22. Kim DH, Ewbank JJ. Signaling in the Immune Response. 2015 Dec 22. In: WormBook [Internet]. http://www.wormbook.org; [1–51]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26694508.
23. Felix MA, Duveau F. Population dynamics and habitat sharing of natural populations of *Caenorhabditis elegans* and *C*. *briggsae*. BMC Biol. 2012; 10(1):59. Epub 2012/06/27.
24. Zugasti O, Bose N, Squiban B, Belougne J, Kurz CL, Schroeder FC, et al. Activation of a G protein-coupled receptor by its endogenous ligand triggers the innate immune response of *Caenorhabditis elegans*. Nat Immunol. 2014; 15(9):833–8. Epub 2014/08/05. doi: 10.1038/ni.2957 PMID: 25086774
25. Squiban B, Belougne J, Ewbank J, Zugasti O. Quantitative and automated high-throughput genome-wide RNAi screens in *C*. *elegans*. J Vis Exp. 2012; 60:e3448. Epub 2012/03/08.
26. Labed SA, Omi S, Gut M, Ewbank JJ, Pujol N. The pseudokinase NIPI-4 is a novel regulator of antimicrobial peptide gene expression. PLoS One. 2012; 7(3):e33887. Epub 2012/04/04. doi: 10.1371/journal.pone.0033887 PMID: 22470487
27. Dierking K, Polanowska J, Omi S, Engelmann I, Gut M, Lembo F, et al. Unusual regulation of a STAT protein by an SLC6 family transporter in *C*. *elegans* epidermal innate immunity. Cell Host Microbe. 2011; 9(5):425–35. Epub 2011/05/18. doi: 10.1016/j.chom.2011.04.011 PMID: 21575913
28. Zugasti O, Ewbank JJ. Neuroimmune regulation of antimicrobial peptide expression by a noncanonical TGF-beta signaling pathway in *Caenorhabditis elegans* epidermis. Nat Immunol. 2009; 10(3):249–56. doi: 10.1038/ni.1700 PMID: 19198592
29. Nomura K, Debroy S, Lee YH, Pumplin N, Jones J, He SY. A bacterial virulence protein suppresses host innate immunity to cause plant disease. Science. 2006; 313(5784):220–3. Epub 2006/07/15. PMID: 16840699
30. Elde NC, Malik HS. The evolutionary conundrum of pathogen mimicry. Nat Rev Microbiol. 2009; 7 (11):787–97. Epub 2009/10/07. doi: 10.1038/nrmicro2222 PMID: 19806153
31. Kepp O, Senovilla L, Galluzzi L, Panaretakis T, Tesniere A, Schlemmer F, et al. Viral subversion of immunogenic cell death. Cell Cycle. 2009; 8(6):860–9. Epub 2009/02/18. PMID: 19221507
32. Shames SR, Auweter SD, Finlay BB. Co-evolution and exploitation of host cell signaling pathways by bacterial pathogens. Int J Biochem Cell Biol. 2009; 41(2):380–9. Epub 2008/09/09. doi: 10.1016/j.biocel.2008.08.013 PMID: 18775503

33. Pujol N, Zugasti O, Wong D, Couillault C, Kurz CL, Schulenburg H, et al. Anti-fungal innate immunity in *C*. *elegans* is enhanced by evolutionary diversification of antimicrobial peptides. PLoS Pathog. 2008; 4(7):e1000105. Epub 2008/07/19. doi: 10.1371/journal.ppat.1000105 PMID: 18636113

34. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–9. doi: 10.1101/gr.074492.107 PMID: 18349386

35. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology: a journal of computational molecular cell biology. 2012; 19(5):455–77.

36. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012; 1(1):18-. doi: 10.1186/2047-217X-1-18 PMID: 23587118

37. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol Iß. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19(6):1117–23. doi: 10.1101/gr.089532.108 PMID: 19251739

38. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27(4):578–9. Epub 2010/12/15. doi: 10.1093/bioinformatics/btq683 PMID: 21149342

39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19(9):1639–45. Epub 2009/06/23. doi: 10.1101/gr.092759.109 PMID: 19541911

40. Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, et al. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M*. *acridum*. PLoS Genet. 2011; 7(1):e1001264. Epub 2011/01/22. doi: 10.1371/journal.pgen.1001264 PMID: 21253567

41. Hu X, Zhang Y, Xiao G, Zheng P, Xia Y, Zhang X, et al. Genome survey uncovers the secrets of sex and lifestyle in caterpillar fungus Chinese Science Bulletin. 2013; 58(23):2846–54.

42. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics (Oxford, England). 2005; 21 Suppl 1:i351–8.

43. Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res. 1997; 25(5):0955–964.

44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. Epub 1990/10/05. PMID: 2231712

45. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008; 24(5):637–44. doi: 10.1093/bioinformatics/btn013 PMID: 18218656

46. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31 (19):3210–2. doi: 10.1093/bioinformatics/btv351 PMID: 26059717

47. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. doi: 10.1093/molbev/mst010 PMID: 23329690

48. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52(5):696–704. PMID: 14530136

49. Lin R, Liu C, Shen B, Bai M, Ling J, Chen G, et al. Analysis of the complete mitochondrial genome of *Pochonia chlamydosporia* suggests a close relationship to the invertebrate-pathogenic fungi in Hypocreales. BMC Microbiol. 2015; 15(1):5. Epub 2015/02/01.

50. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001; 17(9):847–8. PMID: 11590104

51. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. The relationship of protein conservation and sequence length. BMC Evol Biol. 2002; 2:20. PMID: 12410938

52. Li L, Stoeckert CJ, Roos D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9):2178–89. PMID: 12952885

53. Xiao G, Ying SH, Zheng P, Wang ZL, Zhang S, Xie XQ, et al. Genomic perspectives on the evolution of fungal entomopathogenicity in *Beauveria bassiana*. Sci Rep. 2012; 2:483. Epub 2012/07/05. doi: 10.1038/srep00483 PMID: 22761991

54. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 2015. Epub 2015/04/08.

55. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 1995; 167(1–2):GC1–10. PMID: 8566757

56. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res. 2004; 32(Database issue):D138–41. PMID: 14681378

57. Habicht J, Woehle C, Gould SB. *Tetrahymena* Expresses More than a Hundred Proteins with Lipid-binding MORN Motifs that can Differ in their Subcellular Localisations. J Eukaryot Microbiol. 2015. Epub 2015/04/08.

58. Li J, Zhang KQ. Independent expansion of zincin metalloproteinases in Onygenales fungi may be associated with their pathogenicity. PLoS One. 2014; 9(2):e90225. Epub 2014/03/04. doi: 10.1371/journal.pone.0090225 PMID: 24587291

59. Li J, Yu L, Tian Y, Zhang KQ. Molecular evolution of the deuterolysin (M35) family genes in Coccidioides. PLoS One. 2012; 7(2):e31536. Epub 2012/03/01. doi: 10.1371/journal.pone.0031536 PMID: 22363666

60. Dyrka W, Lamacchia M, Durrens P, Kobe B, Daskalov A, Paoletti M, et al. Diversity and variability of NOD-like receptors in fungi. Genome Biol Evol. 2014; 6(12):3137–58. doi: 10.1093/gbe/evu251 PMID: 25398782

61. Fang H, Gough J. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. 2013; 41(Database issue):D536–44. Epub 2012/11/20. doi: 10.1093/nar/gks1080 PMID: 23161684

62. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010; 47(9):736–41. Epub 2010/06/18. doi: 10.1016/j.fgb.2010.06.003 PMID: 20554054

63. Scharf DH, Heinekamp T, Brakhage AA. Human and plant fungal pathogens: the role of secondary metabolites. PLoS Pathog. 2014; 10(1):e1003859. Epub 2014/02/06. doi: 10.1371/journal.ppat.1003859 PMID: 24497825

64. Quandt CA, Bushley KE, Spatafora JW. The genome of the truffle-parasite *Tolypocladium ophioglossoides* and the evolution of antifungal peptaibiotics. Bmc Genomics. 2015; 16(1):553.

65. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M, et al. PHI-base update: additions to the pathogen host interaction database. Nucleic Acids Res. 2008; 36(Database issue): D572–6. Epub 2007/10/19. PMID: 17942425

66. Shen B, Xiao J, Dai L, Huang Y, Mao Z, Lin R, et al. Development of a high-efficiency gene knockout system for *Pochonia chlamydosporia*. Microbiol Res. 2015; 170:18–26. doi: 10.1016/j.micres.2014.10.001 PMID: 25458554

67. Hu X, Xiao G, Zheng P, Shang Y, Su Y, Zhang X, et al. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. P Natl Acad Sci USA. 2014; 111(47):16796–801. Epub 2014/11/05.

68. Xue C, Hsueh YP, Heitman J. Magnificent seven: roles of G protein-coupled receptors in extracellular sensing in fungi. FEMS Microbiol Rev. 2008; 32(6):1010–32. Epub 2008/09/25. doi: 10.1111/j.1574-6976.2008.00131.x PMID: 18811658

69. Kulkarni RD, Kelkar HS, Dean RA. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. Trends Biochem Sci. 2003; 28(3):118–21. Epub 2003/03/14. PMID: 12633989

70. Stringer MA, Dean RA, Sewall TC, Timberlake WE. *Rodletless*, a new *Aspergillus* developmental mutant induced by directed gene inactivation. Genes Dev. 1991; 5(7):1161–71. PMID: 2065971

71. Linder T, Gustafsson CM. Molecular phylogenetics of ascomycotal adhesins—a novel family of putative cell-surface adhesive proteins in fission yeasts. Fungal Genet Biol. 2008; 45(4):485–97. Epub 2007/09/18. PMID: 17870620

72. Ohtaki S, Maeda H, Takahashi T, Yamagata Y, Hasegawa F, Gomi K, et al. Novel hydrophobic surface binding protein, HsbA, produced by *Aspergillus oryzae*. Applied and environmental microbiology. 2006; 72(4):2407–13. Epub 2006/04/07. PMID: 16597938

73. Yang J, Zhao X, Liang L, Xia Z, Lei L, Niu X, et al. Overexpression of a cuticle-degrading protease Ver112 increases the nematicidal activity of *Paecilomyces lilacinus*. Appl Microbiol Biotechnol. 2011; 89(6):1895–903. Epub 2010/11/27. doi: 10.1007/s00253-010-3012-6 PMID: 21110018

74. Engelmann I, Griffon A, Tichit L, Montanana-Sanchis F, Wang G, Reinke V, et al. A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C*. *elegans*. PLoS One. 2011; 6(5):e19055. Epub 2011/05/24. doi: 10.1371/journal.pone.0019055 PMID: 21602919

75. Teixeira PJ, Thomazella DP, Vidal RO, do Prado PF, Reis O, Baroni RM, et al. The fungal pathogen *Moniliophthora perniciosa* has genes similar to plant PR-1 that are highly expressed during its interaction with cacao. PLoS One. 2012; 7(9):e45929. doi: 10.1371/journal.pone.0045929 PMID: 23029323

76. Sakamoto Y, Nakade K, Konno N. Endo-beta-1,3-glucanase GLU1, from the fruiting body of *Lentinula edodes*, belongs to a new glycoside hydrolase family. Appl Environ Microbiol. 2011; 77(23):8350–4. doi: 10.1128/AEM.05581-11 PMID: 21965406

77. Punt PJ, van den Hondel CA. Transformation of filamentous fungi based on hygromycin B and phleomycin resistance markers. Methods Enzymol. 1992; 216:447–57. Epub 1992/01/01. PMID: 1479914

78. Michielse CB, van Wijk R, Reijnen L, Cornelissen BJ, Rep M. Insight into the molecular requirements for pathogenicity of *Fusarium oxysporum* f. sp. *lycopersici* through large-scale insertional mutagenesis. Genome biology. 2009; 10(1):R4. Epub 2009/01/13. doi: 10.1186/gb-2009-10-1-r4 PMID: 19134172

79. Fleissner A, Sarkar S, Jacobson DJ, Roca MG, Read ND, Glass NL. The *so* locus is required for vegetative cell fusion and postfertilization events in *Neurospora crassa*. Eukaryot Cell. 2005; 4(5):920–30. Epub 2005/05/10. PMID: 15879526

80. Prados Rosales RC, Di Pietro A. Vegetative hyphal fusion is not essential for plant infection by *Fusarium oxysporum*. Eukaryot Cell. 2008; 7(1):162–71. Epub 2007/11/28. PMID: 18039941

81. Duverger Y, Belougne J, Scaglione S, Brandli D, Beclin C, Ewbank JJ. A semi-automated high-throughput approach to the generation of transposon insertion mutants in the nematode *Caenorhabditis elegans*. Nucleic Acids Res. 2007; 35(2):e11. PMID: 17164286

82. Banyai L, Patthy L. Amoebapore homologs of *Caenorhabditis elegans*. Biochim Biophys Acta. 1998; 1429(1):259–64. PMID: 9920402

83. Roeder T, Stanisak M, Gelhaus C, Bruchhaus I, Grotzinger J, Leippe M. Caenopores are antimicrobial peptides in the nematode *Caenorhabditis elegans* instrumental in nutrition and immunity. Dev Comp Immunol. 2010; 34(2):203–9. Epub 2009/10/13. doi: 10.1016/j.dci.2009.09.010 PMID: 19818806

84. Mysliwy J, Dingley AJ, Stanisak M, Jung S, Lorenzen I, Roeder T, et al. Caenopore-5: the three-dimensional structure of an antimicrobial protein from *Caenorhabditis elegans*. Dev Comp Immunol. 2010; 34(3):323–30. doi: 10.1016/j.dci.2009.11.003 PMID: 19917307

85. Wong D, Bazopoulou D, Pujol N, Tavernarakis N, Ewbank JJ. Genome-wide investigation reveals pathogen-specific and shared signatures in the response of *Caenorhabditis elegans* to infection. Genome Biol. 2007; 8(9):R194. PMID: 17875205

86. Hoeckendorf A, Leippe M. SPP-3, a saposin-like protein of *Caenorhabditis elegans*, displays antimicrobial and pore-forming activity and is located in the intestine and in one head neuron. Dev Comp Immunol. 2012; 38(1):181–6. doi: 10.1016/j.dci.2012.05.007 PMID: 22677064

87. Hoeckendorf A, Stanisak M, Leippe M. The saposin-like protein SPP-12 is an antimicrobial polypeptide in the pharyngeal neurons of *Caenorhabditis elegans* and participates in defence against a natural bacterial pathogen. Biochem J. 2012; 445(2):205–12. Epub 2012/04/24. doi: 10.1042/BJ20112102 PMID: 22519640

88. Drechsler C. Some hyphomycetes parasitic on free-living terricolous nematodes. Phytopathology. 1941; 31:773–802.

89. Jansson HB, Jeyaprakash A, Zuckerman BM. Control of Root-Knot Nematodes on Tomato by the Endoparasitic Fungus *Meria coniospora*. J Nematol. 1985; 17(3):327–9. PMID: 19294101

90. Meyer SL, Huettel RN, Sayre RM. Isolation of Fungi from *Heterodera glycines* and *in vitro* Bioassays for Their Antagonism to Eggs. J Nematol. 1990; 22(4):532–7. PMID: 19287754

91. Poinar GO, Jansson HB. Susceptibility of *Neoaplectana* spp. and *Heterorhabditis heliothidis* to the Endoparasitic Fungus *Drechmeria coniospora*. J Nematol. 1986; 18(2):225–9. Epub 1986/04/01. PMID: 19294171

92. Dijksterhuis J, Veenhuis M, Harder W. Conidia of the nematophagous fungus *Drechmeria coniospora* adhere to but barely infect *Acrobeloides buetschilii*. FEMS Microbiology Letters. 1993; 113(2):183–8.

93. Jansson HB. Adhesion to Nematodes of Conidia from the Nematophagous Fungus *Drechmeria coniospora*. Journal of General Microbiology. 1993; 139:1899–906.

94. van den Boogert PH, Dijksterhuis J, Velvis H, Veenhuis M. Adhesive knob formation by conidia of the nematophagous fungus *Drechmeria coniospora*. Antonie Van Leeuwenhoek. 1992; 61(3):221–9. PMID: 1519917

95. Sun J, Park SY, Kang S, Liu X, Qiu J, Xiang M. Development of a transformation system for *Hirsutella spp*. and visualization of the mode of nematode infection by GFP-labeled *H*. *minnesotensis*. Sci Rep. 2015; 5:10477. doi: 10.1038/srep10477 PMID: 26190283

96. Rosso LC, Finetti-Sialer MM, Hirsch PR, Ciancio A, Kerry BR, Clark IM. Transcriptome analysis shows differential gene expression in the saprotrophic to parasitic transition of *Pochonia chlamydosporia*. Appl Microbiol Biotechnol. 2011; 90(6):1981–94. doi: 10.1007/s00253-011-3282-7 PMID: 21541788

97. Olivares CM, Lopez-Llorca LV. Fungal egg-parasites of plant-parasitic nematodes from Spanish soils. Rev Iberoam Micol. 2002; 19(2):104–10. PMID: 12828513

98. Larriba E, Jaime MD, Carbonell-Caballero J, Conesa A, Dopazo J, Nislow C, et al. Sequencing and functional analysis of the genome of a nematode egg-parasitic fungus, *Pochonia chlamydosporia*. Fungal Genet Biol. 2014; 65:69–80. Epub 2014/02/18. doi: 10.1016/j.fgb.2014.02.002 PMID: 24530791

99. Tahseen Q, Clark IM, Atkins SD, Hirsch PR, Kerry BR. Impact of the nematophagous fungus *Pochonia chlamydosporia* on nematode and microbial populations. Commun Agric Appl Biol Sci. 2005; 70 (1):81–6. PMID: 16363363

100. Mendoza De Gives PM, Davies KG, Clark SJ, Behnke JM. Predatory behaviour of trapping fungi against *srf* mutants of *Caenorhabditis elegans* and different plant and animal parasitic nematodes. Parasitology. 1999; 119(1):95–104.

101. Migunova VD, Byzov BA. Determinants of trophic modes of the nematophagous fungus *Arthrobotrys oligospora* interacting with bacterivorous nematode *Caenorhabditis elegans*. Pedobiologia. 2005; 49 (2):101–8.

102. Niu X-M, Zhang K-Q. *Arthrobotrys oligospora*: a model organism for understanding the interaction between fungi and nematodes. Mycology. 2011; 2(2):59–78.

103. van den Brink J, de Vries RP. Fungal enzyme sets for plant polysaccharide degradation. Appl Microbiol Biotechnol. 2011; 91(6):1477–92. Epub 2011/07/26. doi: 10.1007/s00253-011-3473-2 PMID: 21785931

104. Dijksterhuis J, Harder W, Wyss U, Veenhuis M. Colonization and digestion of nematodes by the endoparasitic nematophagous fungus *Drechmeria coniospora*. Mycological Research. 1991; 95:873–8.

105. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels. 2013; 6(1):41. doi: 10.1186/1754-6834-6-41 PMID: 23514094

106. Fang W, Leger RJ St. *Mrt*, a gene unique to fungi, encodes an oligosaccharide transporter and facilitates rhizosphere competency in *Metarhizium robertsii*. Plant Physiol. 2010; 154(3):1549–57. Epub 2010/09/15. doi: 10.1104/pp.110.163014 PMID: 20837701

107. van der Biezen EA, Jones JD. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol. 1998; 8(7):R226–7. PMID: 9545207

108. Fang W, Pava-ripoll M, Wang S, Leger R St. Protein kinase A regulates production of virulence determinants by the entomopathogenic fungus, *Metarhizium anisopliae*. Fungal Genet Biol. 2009; 46 (3):277–85. Epub 2009/01/07. doi: 10.1016/j.fgb.2008.12.001 PMID: 19124083

109. Lu J, Cao H, Zhang L, Huang P, Lin F. Systematic analysis of Zn_2Cys_6 transcription factors required for development and pathogenicity by high-throughput gene knockout in the rice blast fungus. PLoS Pathog. 2014; 10(10):e1004432. Epub 2014/10/10. doi: 10.1371/journal.ppat.1004432 PMID: 25299517

110. Dijksterhuis J, Sjollema KA, Veenhuis M, Harder W. Competitive interactions between two nematophagous fungi during infection and digestion of the nematode *Panagrellus redivivus*. Mycological Research. 1994; 98(12):1458–62.

111. Zou CG, Tao N, Liu WJ, Yang JK, Huang XW, Liu XY, et al. Regulation of subtilisin-like protease prC expression by nematode cuticle in the nematophagous fungus *Clonostachys rosea*. Environ Microbiol. 2010; 12(12):3243–52. doi: 10.1111/j.1462-2920.2010.02296.x PMID: 20636375

112. Muszewska A, Taylor JW, Szczesny P, Grynberg M. Independent subtilases expansions in fungi associated with animals. Mol Biol Evol. 2011; 28(12):3395–404. doi: 10.1093/molbev/msr176 PMID: 21727238

113. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, Kemen E, et al. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. Science. 2010; 330(6010):1549–51. Epub 2010/12/15. doi: 10.1126/science.1195203 PMID: 21148394

114. Powell JR, Ausubel FM. Models of *Caenorhabditis elegans* Infection by Bacterial and Fungal Pathogens. In: Ewbank J, Vivier E, editors. Methods Mol Biol. 415: Humana Press; 2008. p. 403–27.

115. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987; 162(1):156–9. Epub 1987/04/01. PMID: 2440339

116. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29(7):644–52. doi: 10.1038/nbt.1883 PMID: 21572440

117. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res. 2009; 37(Database issue):D136–40. doi: 10.1093/nar/gkn766 PMID: 18953034

118. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014; 30:1236–40. doi: 10.1093/bioinformatics/btu031 PMID: 24451626

119. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, et al. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science. 2007; 317(5843):1400–2. Epub 2007/09/08. PMID: 17823352

120. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 2010; 464(7287):367–73. Epub 2010/03/20. doi: 10.1038/nature08850 PMID: 20237561

121. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol. 2008; 26(5):553–60. Epub 2008/05/06. doi: 10.1038/nbt1403 PMID: 18454138

122. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods. 2011; 8(10):785–6. Epub 2011/10/01. doi: 10.1038/nmeth.1701 PMID: 21959131

123. Hawkins J, Boden M. Detecting and sorting targeting peptides with neural networks and support vector machines. J Bioinform Comput Biol. 2006; 4(1):1–18. Epub 2006/03/29. PMID: 16568539

124. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998; 6:175–82. Epub 1998/10/23. PMID: 9783223

125. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, et al. Mobyle: a new full web bioinformatics framework. Bioinformatics. 2009; 25(22):3005–11. doi: 10.1093/bioinformatics/btp493 PMID: 19689959

126. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. Mol Biol Evol. 2016.

127. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. PMID: 9254694

128. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014; 42(Database issue):D490–5. Epub 2013/11/26. doi: 10.1093/nar/gkt1178 PMID: 24270786

129. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009; 37 (Database issue):D233–8. Epub 2008/10/08. doi: 10.1093/nar/gkn663 PMID: 18838391

130. Badet T, Peyraud R, Raffaele S. Common protein sequence signatures associate with *Sclerotinia borealis* lifestyle and secretion in fungal pathogens of the Sclerotiniaceae. Front Plant Sci. 2015; 6:776. doi: 10.3389/fpls.2015.00776 PMID: 26442085

131. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007; 35(Web Server issue):W585–7. PMID: 17517783

132. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics. 2009; 10:202. Epub 2009/07/01. doi: 10.1186/1471-2105-10-202 PMID: 19563654

133. Brameier M, Krings A, MacCallum RM. NucPred—predicting nuclear localization of proteins. Bioinformatics. 2007; 23(9):1159–60. Epub 2007/03/03. PMID: 17332022

134. Theocharidis A, van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat Protoc. 2009; 4(10):1535–50. doi: 10.1038/nprot.2009.177 PMID: 19798086

135. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. doi: 10.1093/bioinformatics/bts635 PMID: 23104886

136. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31(2):166–9. doi: 10.1093/bioinformatics/btu638 PMID: 25260700

137. Turgeon BG, Condon B, Liu J, Zhang N. Protoplast transformation of filamentous fungi. Methods Mol Biol. 2010; 638:3–19. Epub 2010/03/20. doi: 10.1007/978-1-60761-611-5_1 PMID: 20238257

2.6 Internship at OICR Toronto

Duration: **July 15 to April 15**.
Supervisor: **Lincoln Stein.**

Project: 1) Bayesian inference of innate immunity network in *C. elegans*.
Project: 2) R package for converting Reactome pathways to logical pathways.

2.6.1 Bayesian inference of innate immunity network in *C. elegans*.

In the lab using genetics, genomics, proteomics studies we have identified at least four different (p38 MAPK, TGF-beta, Insulin signaling, WNK signaling) pathways that regulates the expression of defense AMPs upon *D. Coniospora infection*, for details on pathways please see section 1.2.3. Separately, we have identified various components of each pathway, but, how these pathways works together is largely unknown. The Reactome group at OICR is trying to apply the Bayesian network learning and inference on different Reactome pathways and more specifically on cancer pathways. In more than a decade, our lab at CIML, have identified various genes of these pathways, Figure 2.1. Based on the information that we have and new data we can infer the regulation of immune network using Bayesian networks. Our aim was to do the same kind of analysis for our immunity pathways in *C. elegans*. So during my stay at OICR, I learned and implemented Bayesian networks on our curated immunity pathway in *C. elegans*. This work is ongoing and we are trying different aspects on our immunity network.

I have already talked about basics of Bayesian networks in Introduction chapter, section 1.7. Here i will describe more about the Bayes theorem and Bayesian networks.

2.6.1.1 Bayes theorem

Bayes Theorem describes the relationship between the new (or posterior) probability of a Hypothesis (H), after having learned a piece of evidence (E). Using the theorem

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

we can derive a formula known as Bayes' rule.

P(H) is the prior probability of hypothesis.

P(E|H) is the likelihood of evidence given the hypothesis.

P(E) is the expectedness/marginal probability of the evidence or probability of E over all possibilities. P(H|E) is the posterior probability of the hypothesis, the new probability given the Evidence.

2.6.1.2 Bayesian networks

Bayesian networks (BNs) or belief networks or Bayes nets in short are probabilistic graphical models that represent a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, it can represent the relationships between diseases and symptoms probabilistically. So that if we are given

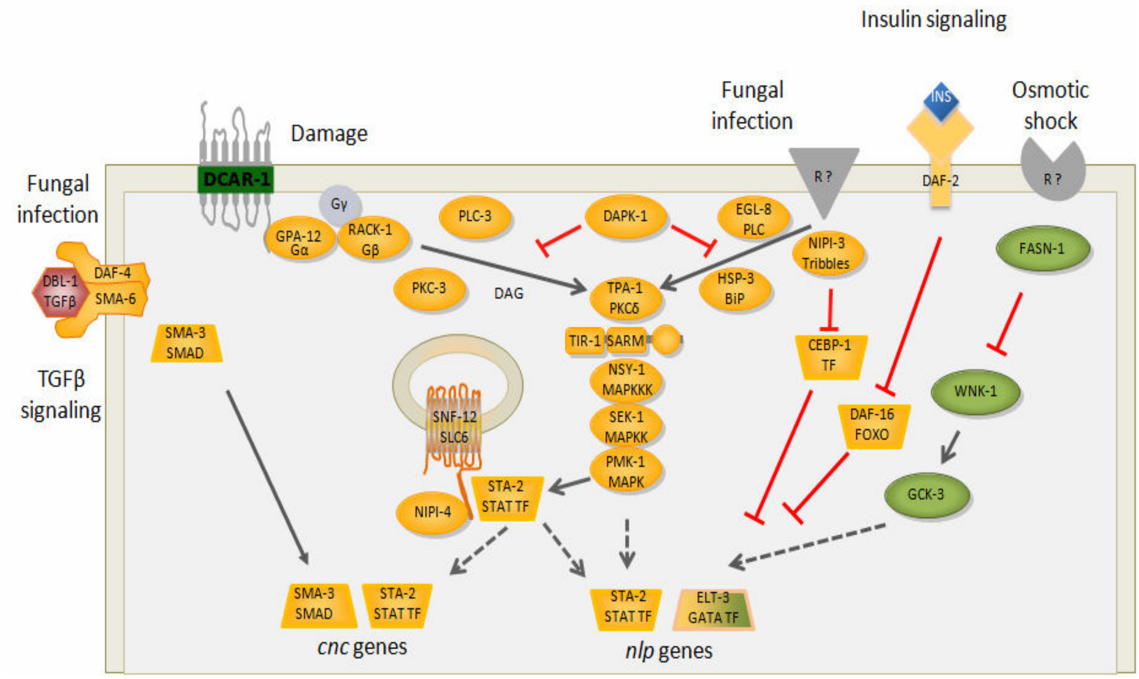


Figure 2.1: Simplified representation of *C. elegans* immune signaling pathways involved in AMP regulation. Although *wnk-1* and *gck-3* are required for the up-regulation of *nlp-29* expression upon infection (Zugasti et al., 2016), the osmotic and infection pathways are separated here for the sake of clarity.

symptoms, a Bayesian network can calculate the probabilities of the occurrence of diseases. In BNs, each node in the graph represents a random variable and the edges between the nodes represent probabilistic dependencies among the corresponding random variables. Computational and statistical methods are used to calculate these conditional dependencies of the random variables in the graph. Given a set of random variables, BNs can be used for effective representation and computation of the joint probability distribution (JPD) of these sets of random variables. In BNs, we have parent child relationships; a child state can only be influenced by its direct parents, meaning that the child output/state is independent of all non-parent nodes: P(Xi|Parents(Xi)). This conditional distribution is represented as conditional probability tables (CPT), here Xi is the child. So if there are 5 nodes [A, B, E, J, M] in the network Figure 2.2, JPD for network will be P(A,B,E,J,M) = P(B).P(E).P(A|B,E).P(J|A).P(M|A) [chain rule or general product rule]. In Figure 2.2, B and E are parents of A, A is a parent of J and M, so M and J are independent of B and E as they are not the direct parents. Like this we calculate CPT for all the nodes, which will be used later to calculate the posterior probability of the nodes as described in the Bayes theorem above.

2.6.1.3 Bayesian networks for immunity pathway in *C. elegans*

We tried to use Bayesian networks for our *C. elegans* immunity pathway. At OICR, Hossein Radfar has developed a Bayesian network tool called Libnet, which can learn and infer the effect on the pathways based on prior knowledge and high throughput data. The workflow of network analysis of *C. elegans* immunity network is shown in Figure 2.3. This is subdivided into various steps as follows.
Step 1: Converting biological pathways to logical model/pathway.
Step 2: Converting logical model/pathway into factor graphs for Bayesian inference.

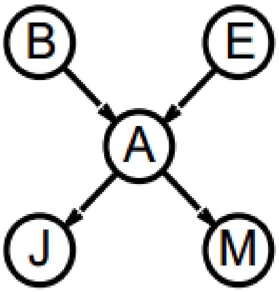


Figure 2.2: Dummy network of five random variables/nodes.

Step 3: Doing Bayesian inference on the logical pathway by doing some in-silico mutations.
Step 4: Looking for non-obvious conditions and conducting experimental validation.

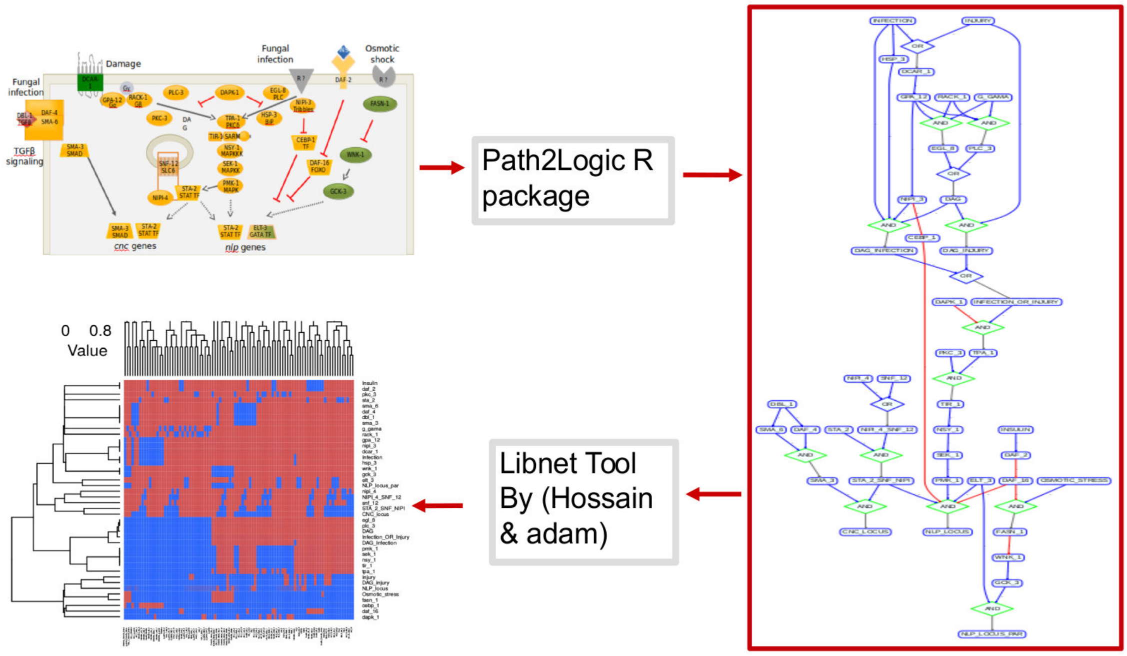


Figure 2.3: Workflow of Bayesian network inference using LibNet and Path2Logic packages. In this figure, top left; cartoon of the network pathway genes identified in the lab, top right; cartoon pathways are converted into logical (”AND” or ”OR” logic between various pathway components) pathways. left bottom; the Bayesian inference of the network genes(in rows) activity (red: active and blue: inactive) for a various in-silico (in columns) mutants.

In the first step we generate the pathway (Figure 2.1) based on the published research from our lab. There are 57 nodes/genes in the network so far. Each gene has parent child relation with other genes in the network. This pathway was converted to the logical pathway as shown in Figure 2.4, with AND or OR logic. Each child/gene have some kind of relation/logic (AND, OR) with its parent or vice-versa. This logical tables shows whether parents work together (AND) or they can work alone (OR) to produce or activate/deactivate a child node. This is done with the help of the R package Path2Logic that I developed in OICR. This package first generates a truth table for each child node and then generates the conditional probability table, which is converted into LibNet readable factor graph format by the same Path2Logic package. Finally, these logical pathway-based factor graphs are used to for Bayesian inference, setting different parameter of the LibNet which I will not detail here. Briefly, one of the parameters for this tool is the observation file in which we provide observation data. In this case it will be *in-silico* mutation of combination of node/genes. So finally, after *in-silico* mutation of all the genes (one at a time) and combination of genes in the network, the LibNet tool calculates the JPD for all the genes in the network as shown in the bottom heatmap in Figure 2.3. Here each row represents all the genes in the networks and each column represents the condition/mutants. Red (high probability of being active) in the heatmap shows that a particular gene is active in the respective mutant condition and blue means inactive (low probability of being active). Color gradient is from blue (0/low probability) to red (1/high probability).

2.6.1.4 Web server for Bayesian network inference

I developed web-server interface for generating logical pathways from user provided data in tab delimited file format. In the same web application, I have implemented the Bayesian network inference functionality as where using Path2Logic user first generate logical pathway and then do the in-silico mutation of the nodes of the generated logical pathway, in results this webinterface returns Bayesian network probabilities for all the nodes/genes of the networks, Figure 2.5 shows an Bayesian network inference for *in-silico sma-3* mutant gene.

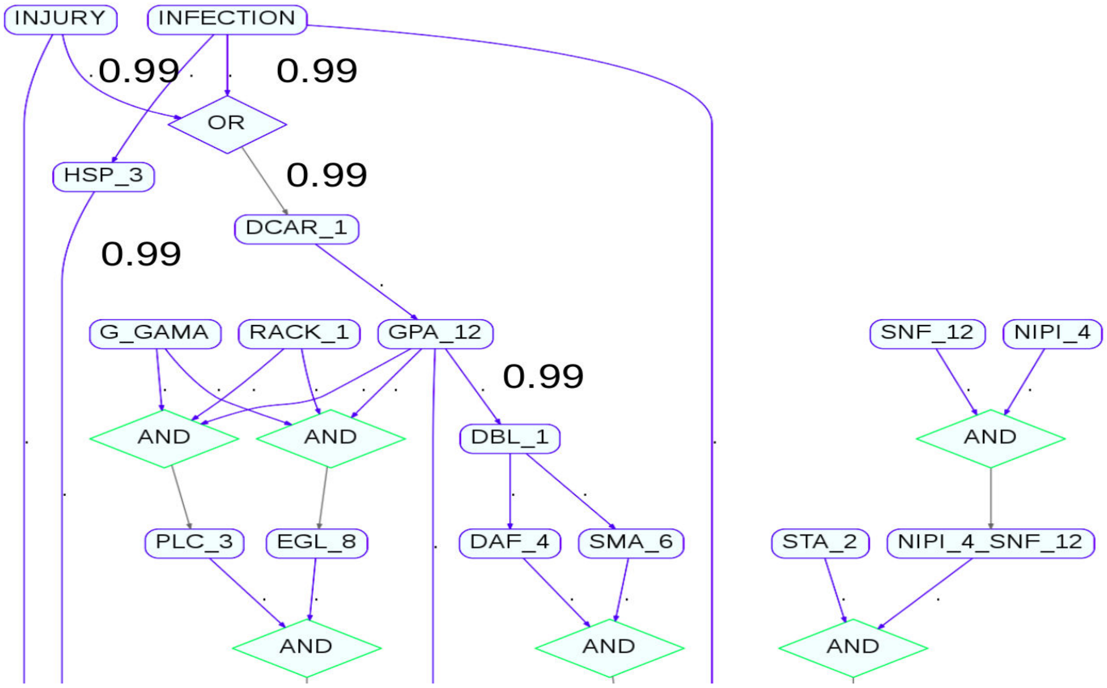


Figure 2.4: Logical representation of the *C. elegans* immunity pathway.In this figure, OR represent either of the genes are required for downstream regulation, whereas in case of AND all the genes are required for the downstream regulation.

2.6.2 R package for converting Reactome pathways to logical pathways.

The Reactome database is one the few manually curated databases of reactions and pathways. It is used heavily by researchers all around the world. It contains more than

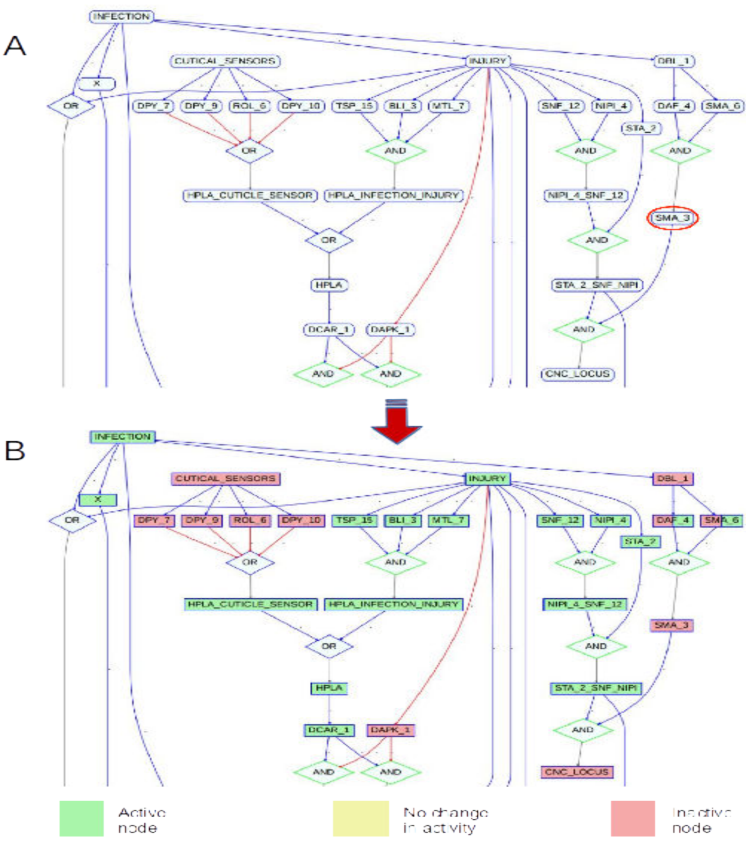


Figure 2.5: Bayesian network inference for in-silico mutant of sma-3 gene. A) Logical immunity pathway of *C. elegans* for *D. coniospora* infection, injury and osmotic stress. B) Bayesian network inference for all the pathway genes in sma-3 in-silico mutants (rounded in red lines), here each node can be either in active (green), inactive(red) or no-change(yellow) state. Each node shows the probability of node/genes to be in each state. Red line shows repressive relation between parent and child node, blue lines shows activating link between parent and child node.

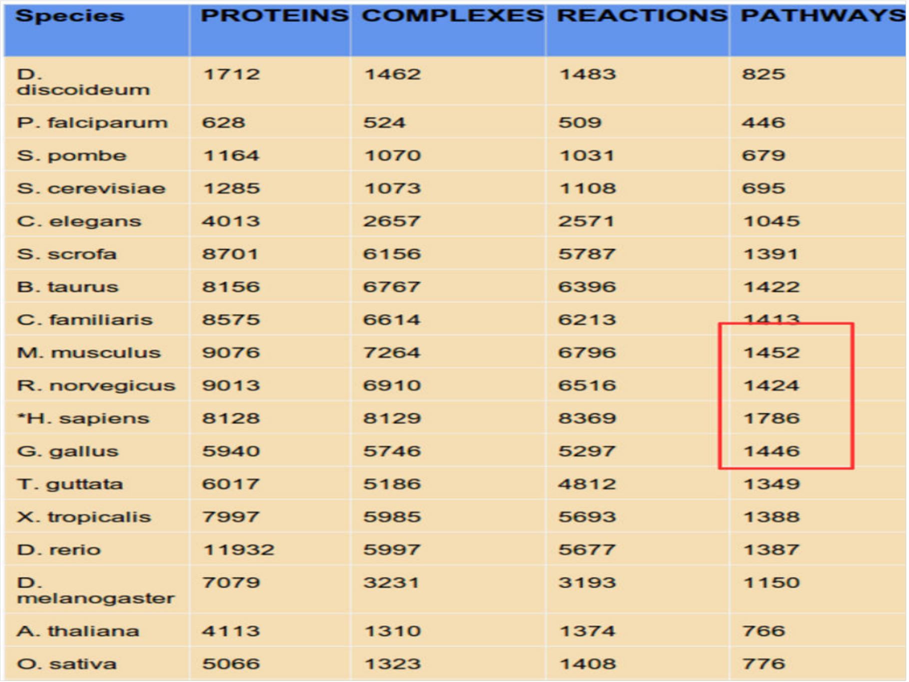
10,000 pathways for 19 important species as shown in Figure 2.6. Even though this database is a very useful resource for researchers, these pathways are not presented as logical pathways (Figure 2.7). This means that if there is a complex/set of interactions, we cant say whether we need all the reactions (AND) or we need just one (OR) of the reaction/interaction to produce the output/product. Only AND and OR logic is considered in the current version of this package. In simple words, logic is missing from these pathways, which stops users from doing network or Bayesian analysis. There is a great need of conversion of these useful Reactome pathways to logical pathways. During my stay at OICR, I started developing a R programming language package that can automatically convert any Reactome pathway into a logical pathway, which can further be used for various kind of analysis, for example, Bayesian inference etc. Most likely these logical pathways would appear in next release of Reactome.

Main features of this package are as follows.

1) Converting any Reactome pathway tab separated file into R pathway object. Objects can later be used for other kinds of analysis on the pathway as described below.

2) Converting pathway to a truth table or a conditional probability table (CPT; Figure 2.8B).

3) Generating the logical pathway graphs as shown in Figure 2.8C and Figure 2.4. In the logical pathway graphs, a red edge represents repression of the child gene by the



Species	PROTEINS	COMPLEXES	REACTIONS	PATHWAYS
D. discoideum	1712	1462	1483	825
P. falciparum	628	524	509	446
S. pombe	1164	1070	1031	679
S. cerevisiae	1285	1073	1108	695
C. elegans	4013	2657	2571	1045
S. scrofa	8701	6156	5787	1391
B. taurus	8156	6767	6396	1422
C. familiaris	8575	6614	6213	1413
M. musculus	9076	7264	6796	1452
R. norvegicus	9013	6910	6516	1424
*H. sapiens	8128	8129	8369	1786
G. gallus	5940	5746	5297	1446
T. guttata	6017	5186	4812	1349
X. tropicalis	7997	5985	5693	1388
D. rerio	11932	5997	5677	1387
D. melanogaster	7079	3231	3193	1150
A. thaliana	4113	1310	1374	766
O. sativa	5066	1323	1408	776

Figure 2.6: Reactome pathway statistics for 19 organisms.

parent gene and blue lines means the parent gene is an activator of the downstream child gene. Numbers on the edge or link is the probability of dependence of the child on the parent node. Blue diamonds show OR logic, that means any parent can activate/deactivate the downstream child node/gene, whereas green diamonds show AND logic, which means that to activate/deactivate the downstream child node we need all the parent genes active/inactive.

4) These truth tables or CPT are then converted to LibNet readable factor graph files. Factor graphs are bipartite graph representing the factorization of the function, each factor represent each node in the pathway. The overall workflow of the Path2Logic tool is shown in Figure 2.8.

In conclusion, my internship exposed me to an entirely new work environment. It allowed me to learn a novel set of computing skills, and was validated by the creation of Path2Logic. This will be a great resource for low and high-throughput pathway analyses.

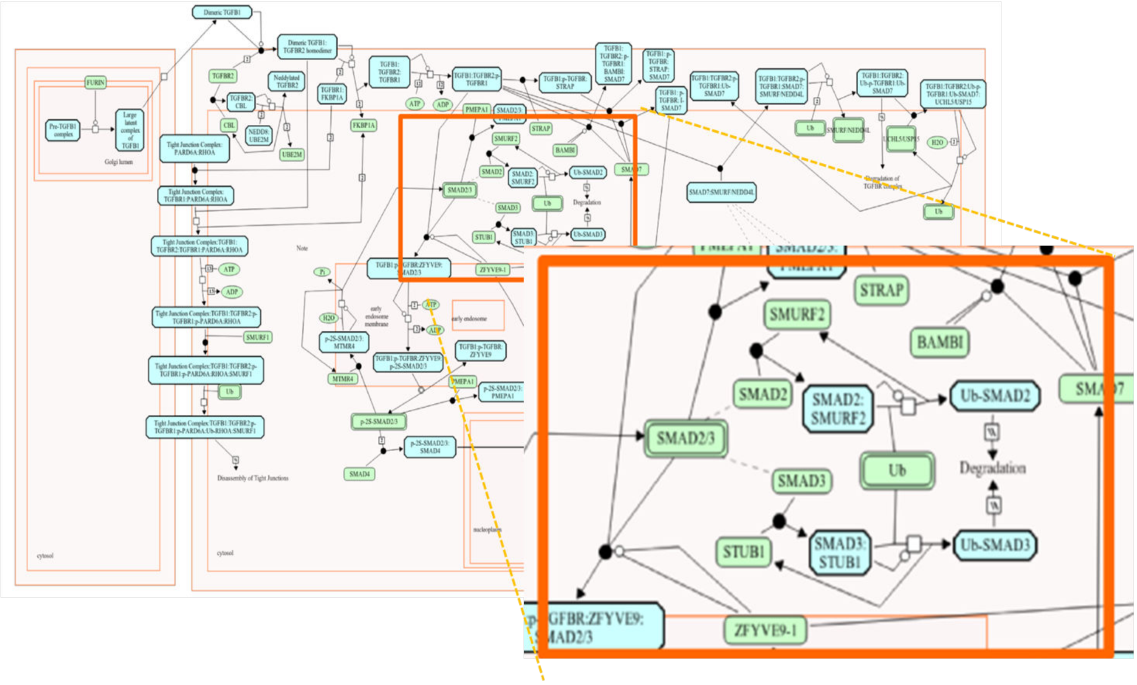
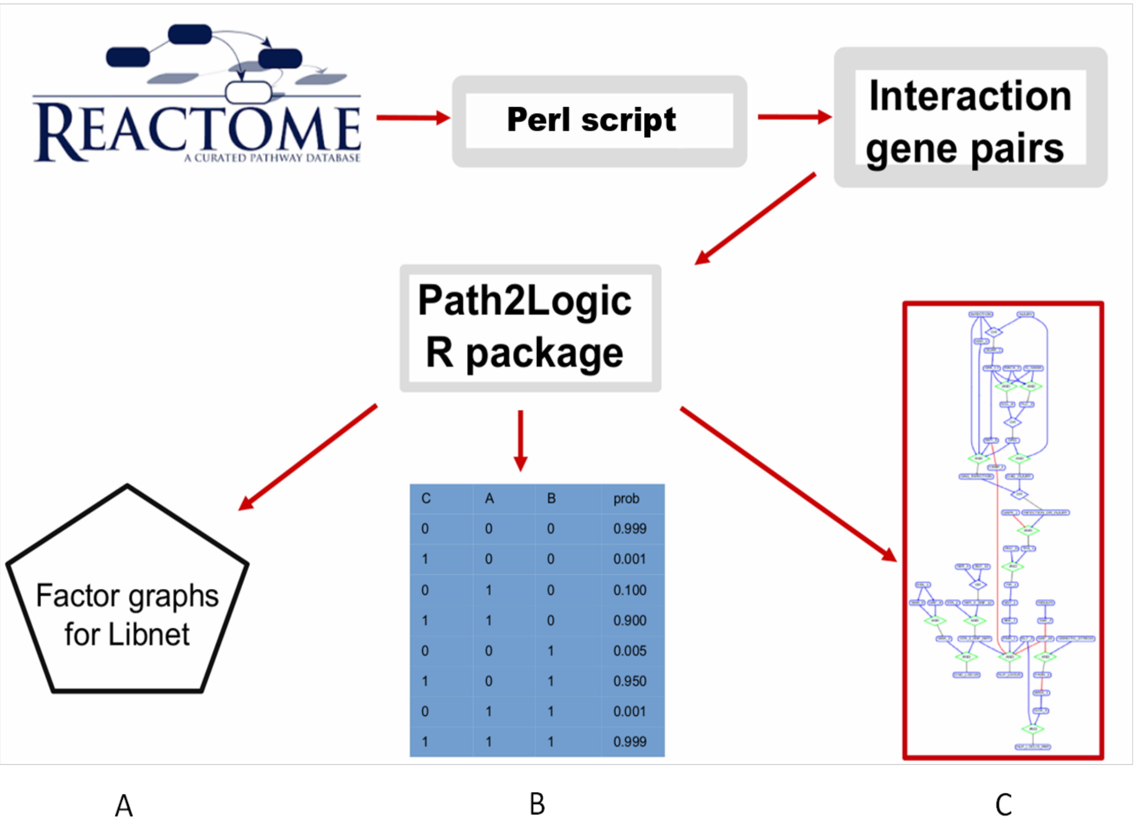


Figure 2.7: Reactome TGF-beta Pathway, box shows that there is no logical representation of the complexes in the pathway.



C	A	B	prob
0	0	0	0.999
1	0	0	0.001
0	1	0	0.100
1	1	0	0.900
0	0	1	0.005
1	0	1	0.950
0	1	1	0.001
1	1	1	0.999

Figure 2.8: Path2Logic R package workflow, from Reactome to the A) Factorgraph file, B) Conditional probability table (CPT) . C) Logical pathway.

Chapter 3

Discussion

C. elegans has been established as a powerful model to study host-pathogen interactions for more than a decade. Like other invertebrates, to defend itself against pathogens *C. elegans* relies entirely on its innate immune system. Studies have revealed that the innate immune system of *C. elegans* involves evolutionarily conserved signaling pathways and many lineage-specific innovations (Kim and Ewbank, 2015). During evolution, immune systems became more complex. In jawed vertebrates, the adaptive immune system came into existence. One could ask why higher organisms had to invent this complex adaptive immune system if the innate immunity is sufficient to protect other species? One possible explanation that has been proposed by immunologists is that with advancing evolution, the host-pathogen interactions becomes more complicated (Travis, 2009), which necessitated the host organisms to evolve and to fine-tune their immune systems accordingly. If this is correct, one approach to understand better the complexity of adaptive systems, we need first to understand ”simple” innate immune systems. In our lab, we mainly focus on the innate immune response of *C. elegans* to its natural pathogen *D. coniospora*. After infection many defense AMPs are induced by *C. elegans* that protect host from the pathogen. The aim of my thesis was to build a gene regulatory network to describe the regulation of these AMPs. In the following sections, I will discuss how these 3 manuscripts contributes to a better understanding of the regulatory gene networks that control the innate immune response of *C. elegans* to *D. coniospora*. To understand the host immunity, understanding pathogen virulence is also very important. During my PhD thesis, I also looked and analyzed the newly sequenced *D. coniospora* genome. Using the comparative and the functional genomic bioinformatics analysis we gained insight into the evolution and virulence of the nematode-destroying fungi. Those studies will not be discussed further here, but are the subject of ongoing research.

3.1 Discussion of Publication 1 (Section II.1)

Clone mapper: an online suite of tools for RNAi experiments in *Caenorhabditis elegans.*
Thakur N, Pujol N, Tichit L, Ewbank JJ, G3, 2014

3.1.1 Summary of results

Many genome-wide RNAi screens in *C. elegans* have been published in last 15 years. Most used bacterial clones that produce the specific double-stranded RNAs. Surprisingly, there was no proper and easy way to analyze the results from such large RNAi screens. In this article, we introduced an online suite of tools called CloneMapper which contains two separate algorithms for RNAi screen analysis. One is for verification of the sequenced bacterial clones and the second algorithm for identification of target genes. We compared our algorithm with existing clone target identification resources like UP-TORR (Hu et al., 2013) and Wormbase (Harris et al., 2014) and showed that our tool outperformed the existing resources due to better algorithm design. This suite of tools is freely accessible at http://bioinformatics.lif.univ-mrs.fr/RNAiMap/index.html.

3.1.2 Clone verification and tool updates

After a genome-wide screen, the most common method to verify a clones identity was by blasting the relevant sequence against the *C. elegans* genome. Manually verifying clone sequences for genome-wide studies is tedious and potentially introduces errors. No automated and easy pipeline was available for such analysis. We developed the CloneMapper tool that allows automated clone verification on the fly, in a simple and easy way. For this analysis, the predicted clone sequences are generated *in-silico* from the genomic features provided by Wormbase. The Wormbase database is updated every two months and sometimes there are significant changes made to the genome sequence and annotation. This tool was developed using genomic data available at Wormbase release WS240 and in an ideal world, this should be re-done for every release. Due to time and resource constraints, this is not possible yet. This problem will be solved in coming future as this tool will be a part of Wormbase and it will be updated on every release.

3.1.3 Off-target identification

The clone target identification is complicated by an off-target effects of RNAi clones. Sometimes, one RNAi clone generates siRNAs that targets two or more genes. The existing tools like UP-TORR and Wormbase many times either miss these off-target genes or they predict the wrong targets, which is obviously problematic for downstream analysis for high-throughput RNAi screen results. In this tool, we tried to overcome the shortcomings of existing tools. After a comparative analysis of whole RNAi clone library targets, we found that due to under-prediction of clone off-targets, Wormbase missed more than 20% of the high-confidence targets identified by Clone Mapper. This shows the scale at which downstream analysis can be directly affected by target prediction. We further demonstrated the potential of Clone Mapper tool in identifying novel targets for some RNAi clones taken from published studies.

3.1.4 Secondary siRNA target identification

Many studies have shown the importance of spreading of target silencing (e.g. (Pak and Fire, 2007)). Once the primary siRNAs are processed by dicer, these primary siRNAs chop the target mRNA and this targeted RNA becomes the source of production of secondary siRNAs. These secondary siRNAs target any complementary mRNA sequences. In most cases, these sequences correspond to the gene that was targeted by the primary siRNAs, from which they were produced. In some cases, however, they can target mRNAs from other genes, if there is sufficient sequence similarity. In turn, these mRNAs can give rise to new siRNAs that potentially will target a still more diverse set of mRNAs. In principle, this process can keep repeating itself indefinitely, limited only by the common occurrence of fragments of sufficient sequence identity across multiple mRNAs. To the best of our knowledge, there has been no systematic study of the effective range of this RNAi spreading. In the current version of Clone Mapper, we have therefore not provided a tool for the identification of potential secondary siRNA targets. In future, if there is sufficient demand this functionality of secondary target identification could be added.

3.1.5 Data integration for better siRNA target identification

It is a frequent observation in RNAi-based experiments that knocking down a gene does not provoke the expected phenotype. This can often be explained by the spatial and temporal expression pattern of the target gene. Several tissues, such as neurons, are refractory to RNAi in *C. elegans*. This barrier can be circumvented through different experimental approaches (Firnhaber and Hammarlund, 2013). Alternatively, a given gene might be expressed only at a particular developmental stage, at a very low or very high level, or in a cyclic manner. It could be useful to integrate spatial and temporal expression information for potential target genes when drawing up the lists of candidate RNAi targets. This, in turn could lead to a better understanding of regulatory networks since integrating such information would allow the identification of genes expected to be recalcitrant to RNAi. This would enhance network construction from genome-wide RNAi screens.

3.1.6 Conclusion

In this article, we presented a novel suite of tools for analyzing genome-wide RNAi screens. This tool provides an automatic batch mode RNAi clone sequence verification functionality. We also introduced a novel algorithm for RNAi clone target identification, based on what is known about the molecular mechanisms involved. We showed that our tool is better than Wormbase in target identification and with this tool we found 20 % more targets compared to Wormbase. In the tool we also described how our algorithm can identify potential off-target genes that other algorithm can not identify. With this tool, we re-analyzed 3 genome-wide RNAi screen results and identified many new targets that were not found in the original studies. This tool provides an easy and accurate way to analyze high-throughput RNAi screen data. Since its publication, the tool has been accessed from more than 150 different IP addresses.

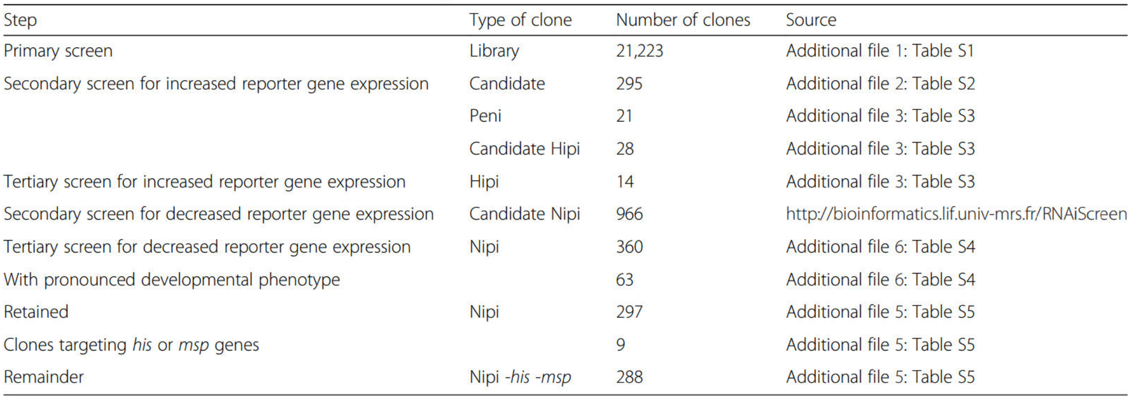
3.2 Discussion of Publication 2 (Section II.2)

A quantitative genome-wide RNAi screen in *C. elegans* for antifungal innate immunity genes.
Zugasti O#, **Thakur N#**, Belougne J, Squiban B, Kurz CL, Soul J, Omi S, Tichit L, Pujol N, Ewbank JJ, BMC Biol, 2016
#=equal contribution

3.2.1 Summary of results

The first step in building a regulatory network is to elucidate its backbone, comprised of the genes that are its main nodes. Removal of any one of these main nodes would be expected to have a major phenotypic effect. These genes can often be identified by forward genetic screens. In direct mutagenesis screens for Nipi (for no induction of antimicrobial peptides after infection) genes, we identified a number of candidates. Although relatively large-scale screens have been conducted, (e.g. 8 alleles have been isolated for snf-12, (Dierking et al., 2011; Labed et al., 2012)), they are certainly not saturating. With the aim of identifying further backbone nodes, as well as more peripheral players in the regulatory network, using an automated method (Squiban et al., 2012), we conducted a genome-wide RNAi screen with a collection of 21,223 RNAi clones from the Ahringer (Kamath et al., 2003) and Vidal (Rual et al., 2004a) RNAi libraries. We screened these clones twice to quantify the infection-induced expression of the nlp-29p::gfp reporter gene, followed by successive rounds of refining screens. The overall results of this RNAi screen are summed up in Table 1 taken from the paper (Zugasti et al., 2016).

Table 1. Overview of screen results (Zugasti et al. 2016)



Step	Type of clone	Number of clones	Source
Primary screen	Library	21,223	Additional file 1: Table S1
Secondary screen for increased reporter gene expression	Candidate	295	Additional file 2: Table S2
	Peni	21	Additional file 3: Table S3
	Candidate Hipi	28	Additional file 3: Table S3
Tertiary screen for increased reporter gene expression	Hipi	14	Additional file 3: Table S3
Secondary screen for decreased reporter gene expression	Candidate Nipi	966	http://bioinformatics.lif.univ-mrs.fr/RNAiScreen
Tertiary screen for decreased reporter gene expression	Nipi	360	Additional file 6: Table S4
With pronounced developmental phenotype		63	Additional file 6: Table S4
Retained	Nipi	297	Additional file 5: Table S5
Clones targeting *his* or *msp* genes		9	Additional file 5: Table S5
Remainder	Nipi -*his* -*msp*	288	Additional file 5: Table S5

The distribution of phenotypes provoked by the RNAi clones was similar to that seen for many other quantitative screens Figure 3.1, with the clones giving the less-pronounced phenotypes predicted to target genes encoding proteins acting outside the centre of the network Figure 3.2.

In the RNAi screen, having eliminated clones that provoked a strong developmental phenotype, we identified 297 Nipi clones that abrogate AMP gene induction. Using Clone Mapper (Thakur et al., 2014) [discussed in results section 2.1] we identified 338 target genes for these 297 Nipi clones. Only 6 had previously been associated with innate immune gene regulation. Both *pkc-3* and *tir-1* had been assigned an immune regulatory function through a candidate gene approach; no alleles for these genes were ever recovered in our genetic screens. Consistent with this, they gave moderately mild phenotypes. The remaining 4 genes were associated with relatively strong phenotypes. Among them, the clones targeting *sta-2* and *nipi-3* gave the most robust phenotypes

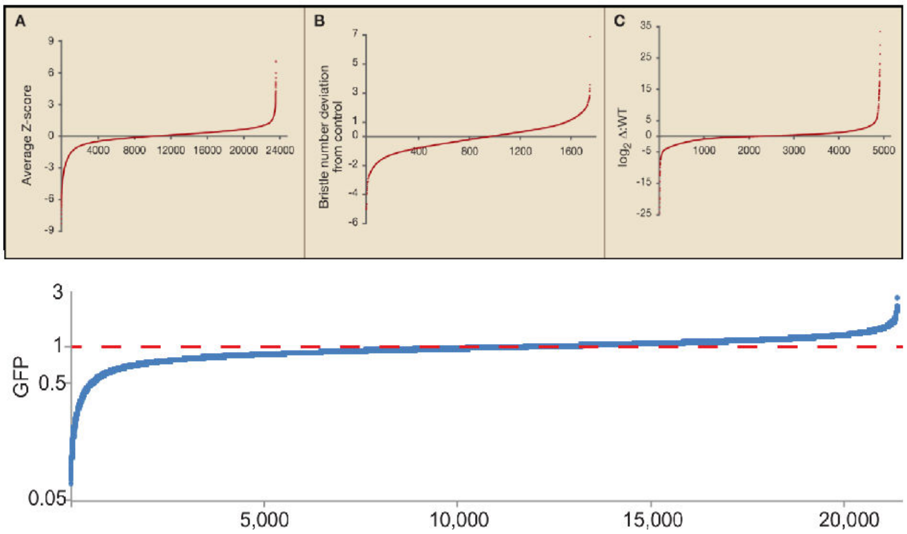


Figure 3.1: Continuous Distributions of Quantitative Signaling Readouts Upper panel, figure and legend taken from (Friedman and Perrimon, 2007) (A) Quantitative assay for ERK activation in a cell-based *Drosophila* RNAi screen (Friedman and Perrimon, 2007). Effect of each double-stranded RNA is represented as a Z score relative to control ERK activation. (B) In vivo quantitative P element collection screen in *Drosophila* assaying deviation of bristle number (Norga et al., 2003). (C) Genome-wide quantitative deletion screen for G protein/MAPK signaling in *S. cerevisiae* (Chasse et al., 2006). Effect of each deletion on mating factor stimulation relative to an internal wild-type control. Unpublished dataset is courtesy of H. Dohlman. Lower panel, quantitative signaling readout distributions for Nipi RNAi screen (Zugasti et al., 2016).

and were both in the top 20 of all clones (top 0.1%), established on the basis of the average of the six trials in the 1st and 2nd rounds of screening Figure 3.3. This is consistent with the expectation that genes central to signaling are those that give the strongest phenotypes and can be isolated in forward genetic screens.

Despite having filtered out clones provoking developmental defects, many of the Nipi clones targeted genes that at first sight are unlikely to play a specific role in innate immunity. Looking only for the targets of the top 20 clones, there were genes involved in transcription and translation (ZC376.6 that encodes a subunit of the integrator complex that associates with RNA polymerase II and mediates 3′end processing of small nuclear RNAs, and F11A3.2 that encodes a subunit of eukaryotic translation initiation factor 2B), ubiquitination (F52C6.12 and *skr-1*) or metabolism (R08D7.7, a xylulokinase). The list also contained *spg-7* that encodes a metalloprotease important for the proteolytic degradation of mitochondrial proteins. We showed that it exerts its effect indirectly, via an induction of the UPRmt in the intestine.

These results indicate the degree to which one must be cautious in assigning a specific immune function to the genes that are associated with the strongest phenotypes. Two other mitochondrial genes were also in the list, *atad-1* and *nuo-3*. Whether these play a direct or indirect role in AMP gene regulation has not yet been established. In the case of *nuo-3*, this will be particularly difficult to assay due to the genes very particular gene structure; the locus encodes two distinct, non-overlapping proteins, a mitochondrial NADH-ubiquinone oxidoreductase alpha subunit, but also, from the 5′end, an orthologous of BRICK1, a component of the WAVE1 complex involved in actin-remodeling Figure 3.4.

There is a further caveat to the interpretation of these results since the phenotypes

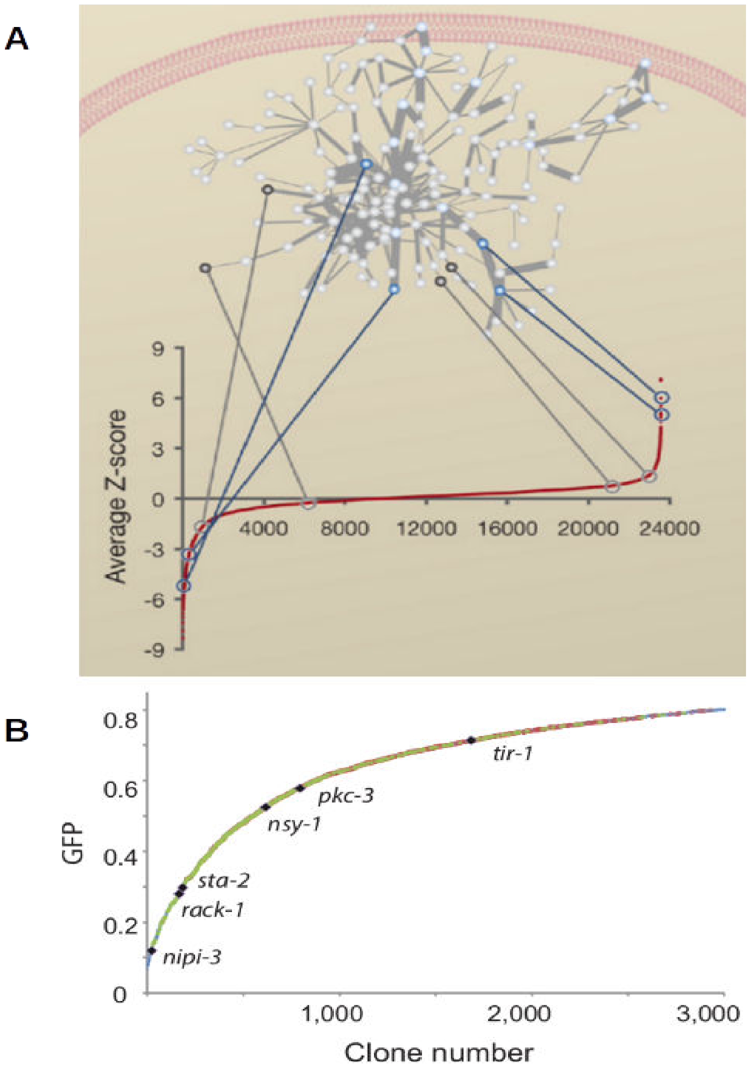


Figure 3.2: A) Relationship between cell-based quantitative screen output and a network signaling model. In this screen, canonical components have the greatest impact on signaling output and thus are represented at the extreme ends of the distribution. More distant and quantitatively less important proteins appear internally along the distribution. Measuring signaling output at one of these proteins may result in canonical components having weak effects in the assay, Figure and legend are taken from (Friedman and Perrimon, 2007). B) The averages of the two values for normalized GFP expression from the first round of screening for the first 3000 RNAi clones. The 966 and 360 clones that passed first and second round screening are indicated in red and green, respectively. The results for selected known signaling components are indicated in black. Figure and legend are taken from (Zugasti et al., 2016).

associated with each positive RNAi clone were extremely variable Figure 3.5, which makes a robust ranking of the clones almost impossible.

Through a bioinformatic network analysis of these Nipi genes, we confirmed the central role of MAPK pathways in the induction of AMPs. Functional enrichment analysis revealed an enrichment of genes involved in the mitochondrial unfolded response (UPRmt) and we subsequently proved experimentally the role of these UPRmt genes in AMP induction. Indeed, we found evidence for multiple types of cross-tissue communication being involved in the regulation of defense AMPs. We also discovered the involvement of the mRNA deadenylase CCR4-NOT protein complex in AMP regulation. Overall in this study, we were able to identify 338 genes potentially involved

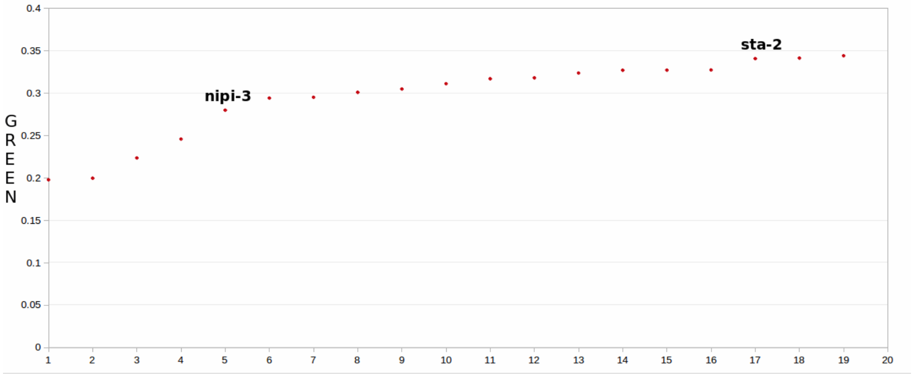


Figure 3.3: Top 20 Nipi clones with strongest phenotypes.

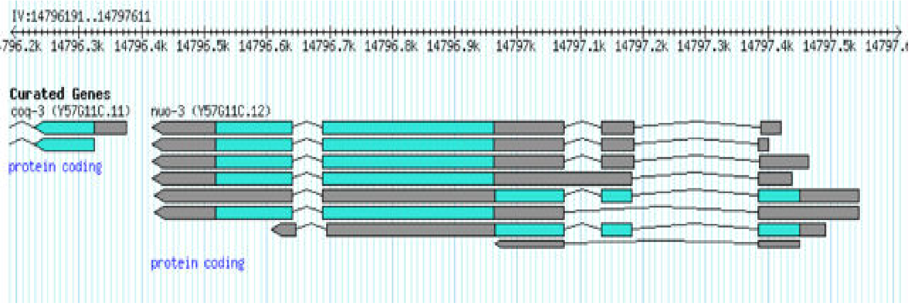


Figure 3.4: Genomic locus containing *nuo-3* gene. Screen grab from Wormbase WS253.

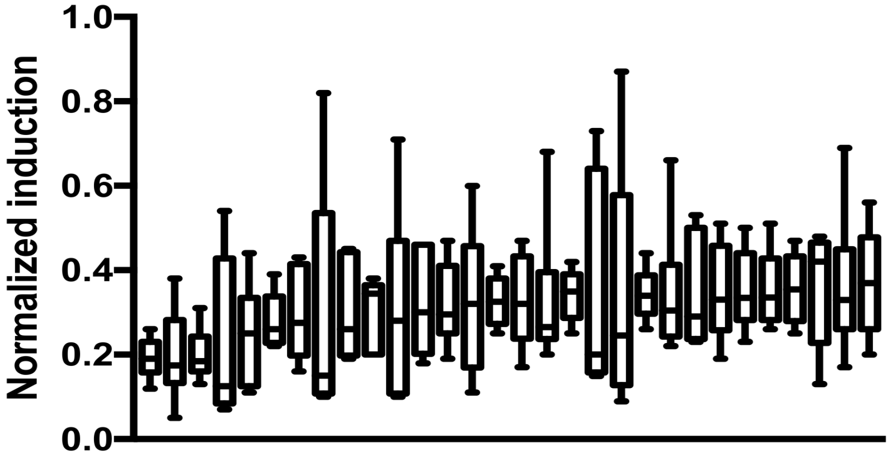


Figure 3.5: Box and whisker plots of the 30 clones associated with the strongest Nipi phenotypes, ranked on the basis of the average of their normalised effect on reporter gene expression from 6 independent experiments (2 from round 1 and 4 from round 2; see (Zugasti et al., 2016)). The bars represent the median values and the limits are defined by the highest and lowest individual values.

in the immune regulatory network, acting in various signaling complexes.

3.2.2 KEGG Pathways

One of the resources we used for our analysis was the KEGG pathway database (Kanehisa et al., 2016). This is a manually curated molecular interaction/reaction pathway database that harbours information for 484 reference pathways in multiple species. For *C. elegans*, there are 129 pathways in KEGG. When we did functional enrichment analysis using the set of Nipi genes, we found a total of 120 Nipi genes are linked to 41 KEGG pathways and more than 7 pathways are highly enriched with Nipi genes. I will focus my discussion on particular pathways. One of the most enriched KEGG pathway is oxidative phosphorylation. Cells releases energy by oxidizing nutrients with the help of enzymes of the oxidative phosphorylation pathway. Apart from its vital role in metabolism, this pathway also produces harmful reactive oxygen species (ROS) such as superoxide and hydrogen peroxide, which leads to disease, aging, etc. Recently ROS/nitric oxide (NO) has been shown to be important for innate immunity in various organism (Phillip West et al., 2011; Kohchi et al., 2009; Yang et al., 2013; Nathan and Cunningham-Bussel, 2013). Mitochondrialy-derived ROS promotes wound repair in *C. elegans* following physical injury (Xu and Chisholm, 2014). There have been claims that *D. coniospora* infection provokes the production of ROS in *C. elegans* (Zou et al., 2013). Dual oxidase (Ce-Duox1/BLI-3)-dependent ROS production has been demonstrated to have a protective effect against intestinal bacterial infections (Chávez et al., 2009). Thus lower expression of Ce-Duox1/BLI-3 leads to higher susceptibility to *E. faecalis* infection. This infection-induced ROS production activates SKN-1, (mammalian ortholog of Nrf transcription factors) in a p38 MAPK dependent manner (Hoeven et al., 2011). Thus, ROS generation has been suggested to be important for *C. elegans* innate immune signaling (van der Hoeven et al., 2012). Strikingly, in this study, we found 23 genes out of a total of 110 known oxidative phosphorylation genes to be Nipi. In the context of our validation of candidate mitochondrial genes, as mentioned above and detailed in (Zugasti et al., 2016), we attributed the observed abrogation of AMP expression to an indirect effect involving induction of the UPRmt. Further studies will be needed to determine whether some of the clones targeting genes required for oxidative phosphorylation exert their effect via affecting ROS production.

Several enriched pathways were related to transcription and translation, including RNA transport, spliceosome and ribosome assembly pathways. Another pathway found to be enriched among the Nipi genes is that of mRNA surveillance. This is a mechanism to ensure messenger RNA (mRNA) quality by detecting and degrading abnormal mRNAs. There are three kinds of mRNA surveillance pathways in an organism, the nonsense-mediated mRNA decay pathway (NMD), the Nonstop Mediated mRNA decay pathways (NSD) and the No-go Mediated mRNA decay pathway (NGD). The NMD pathway is the best-studied RNA surveillance pathway and its role in physiology, stress, immune response and viral replication has been reviewed extensively (Hug et al., 2016; Gardner, 2010). Under stress conditions, the NMD pathway is suppressed due to inhibition of mRNA translation. This leads to upregulation of certain genes/proteins important for coping with stress (Karam et al., 2013; Gardner, 2010). Suppression of this pathway also leads to an up-regulation of the unfolded protein response (UPR), generating a further stress signal (Karam et al., 2013). Interestingly, (Sun et al., 2011) identified 224 genes whose inactivation leads to the production of P bodies, cytoplasmic processing bodies involved in post-transcriptional processes like mRNA decay, nonsense-mediated RNA decay (NMD), storage of silenced mRNA, etc. Strikingly, 83 out of these 224 genes are associated to the NMD pathway, and 33 of these were also Nipi genes. Several CCR4/NOT1 complex genes were also found to regulate P-body production. This was again a category enriched in the set of Nipi genes (Zugasti et al., 2016). This suggests that immune signaling pathways in *C. elegans* are regulated at the level of mRNA stability. A host has to protect itself from pathogens, so immune signaling pathways have to be activated quickly. One possibility is to keep expressing and degrading defense gene mRNAs under normal conditions, but as soon as a pathogen is encountered to switch off mRNA degradation. This would lead to the rapid production of defense effector proteins. The potential role of mRNA stability in immunity in *C. elegans* remains open for the future studies.

3.2.3 MAPK signaling at a center of *C.elegans* fungal immunity pathways

With this screen, we recovered the *tir-1* and *nsy-1* genes, known components of the *pmk-1* p38 MAPK pathway. MAPK signaling is central to innate immune defense in many species, including *C.elegans* and vertebrates (Andrusiak and Jin, 2016; Arthur and Ley, 2013). More than 1/5 (22 %) of the Nipi gene were found to be associated with MAPK signaling in at least one of the three species (Human, Yeast, Drosophila). Some of these genes have already been shown to be part of a *C.elegans* MAPK pathway, for example, *hda-1*, Histone deacetylase 1, which is linked to the JNK PMK pathway upon metal stress (Hattori et al., 2013). Similarly, *rack-1*, related to Gβ proteins, was shown to act upstream of the p38 MAPK cascade (Ziegler et al., 2009b). Most of these MAPK signaling genes were found to be connected in a WormNet analysis [Results section 2.2,(Zugasti et al., 2016)]. Another 15 genes from the screen were found linked to this MAPK network ((Zugasti et al., 2016), Additional file 7: Table S6). In total, 50 genes have been found associated with the MAPK signaling pathway, extending the known network considerably. Further experimental studies are required to explore the role of these genes in MAPK signaling. Recently, (Pukkila-Worley et al., 2014) identified 29 genes that regulated the expression of the antibacterial immune effector F08G5.6 in the intestinal immune response in a p38 MAPK dependent manner. Among them, we found 6 genes (*nipi-3, rpn-9, pabp-2, mdt-15, tir-1, ftt-2*) in our final list of 338 Nipi genes. When we looked at the screen data for the remaining 23 genes, we found that 3 other clones, targeting *acs-19, sdhb-1* and *fat-7* passed the first round of Nipi screening and that 4 clones blocked reporter gene induction in at least one replicate of the first round. The fact that 13 genes among 29 validated regulators of the intestinal p38 MAPK pathway are associated with a Nipi phenotype shows the degree of similarity between MAPK pathways in the epidermis and intestine.

3.2.4 Mediator complex

One of the genes implicated in innate immunity in both tissues is *mdt-15*. This encodes a subunit of the Mediator complex, a transcriptional coactivator of 21-26 proteins, common to all eukaryotes. The Mediator complex is involved in the transcriptional initiation, transcriptional elongation, and organization of genomic DNA. *mdt-15* was shown to play an important role in the response to infection and abiotic stress in *C. elegans* (Pukkila-Worley et al., 2014), whereas *mdt-23* and *mdt-24* are required for the response to *Microbacterium nematophilum* (Nicholas and Hodgkin, 2004). This complex has also been found to be involved in other stress responses as reviewed in (Grants et al., 2015) and to be important for plant innate immunity (An and Mou, 2013). MED14-16, 18, 19, 25, CDK8 are known to be involved in the defense response to various kind of stress (Samanta and Thakur, 2015). Figure 3.6

3.2.5 Transcription factors (TFs)

The Mediator complex interacts with a broad range of TFs. Using an established list of worm TFs (Reece-Hoyes et al., 2005), in the genome-wide Nipi screen, we identified 11 TF genes that potentially regulate the expression of AMP genes. Among

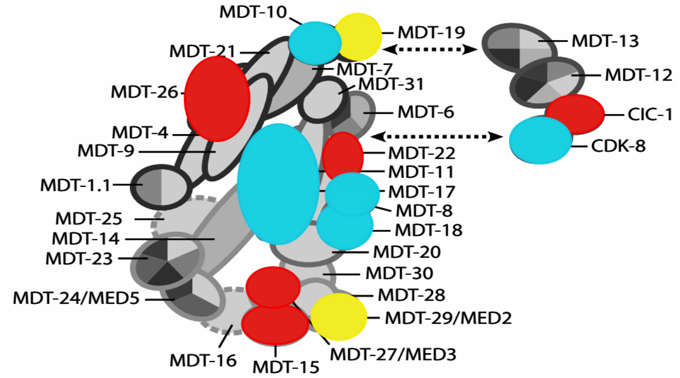


Figure 3.6: Protein components of the *C. elegans* mediator complex. Components shown in red were in the final Nipi gene list. The genes corresponding to the components in yellow are those that passed the first round of the Nipi screen only, while those in blue correspond to genes that were classified as Nipi in at least one replicate of first Nipi screen. Figure adapted from (Grants et al., 2015).

them, DCP-66 and LIN-40 play roles in chromatin remodeling. Surprisingly, the one previously identified TF gene, *sta-2*, is not among those catalogued by Reece et al. Together, this gives a total of 12 genes potentially directly involved in AMP gene expression. As outlined below, their role in regulating the innate immune response is being investigated.

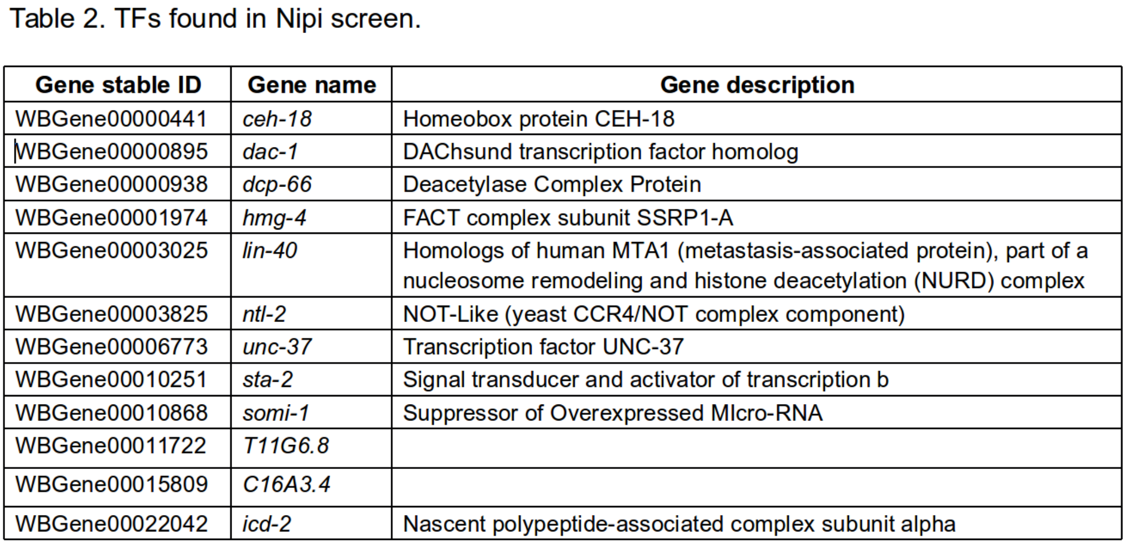


Table 2. TFs found in Nipi screen.

Gene stable ID	Gene name	Gene description
WBGene00000441	*ceh-18*	Homeobox protein CEH-18
WBGene00000895	*dac-1*	DAChsund transcription factor homolog
WBGene00000938	*dcp-66*	Deacetylase Complex Protein
WBGene00001974	*hmg-4*	FACT complex subunit SSRP1-A
WBGene00003025	*lin-40*	Homologs of human MTA1 (metastasis-associated protein), part of a nucleosome remodeling and histone deacetylation (NURD) complex
WBGene00003825	*ntl-2*	NOT-Like (yeast CCR4/NOT complex component)
WBGene00006773	*unc-37*	Transcription factor UNC-37
WBGene00010251	*sta-2*	Signal transducer and activator of transcription b
WBGene00010868	*somi-1*	Suppressor of Overexpressed MIcro-RNA
WBGene00011722	*T11G6.8*	
WBGene00015809	*C16A3.4*	
WBGene00022042	*icd-2*	Nascent polypeptide-associated complex subunit alpha

3.2.6 Phylogenetic profile analysis

In the published article, we reported the phylogenetic profile of a small subset of Nipi genes. When we extended this analysis to the complete set of 278 Nipi genes, strikingly, we found that about 60 % of the corresponding proteins are evolutionary conserved across all the diverse eukaryotic species (Cluster C1 in Figure 3.7). In cluster C3, we found taxon-restricted genes (TRG) including the GPCR DCAR-1. Given DCAR-1s role in antifungal immunity (Zugasti et al., 2014), these TRG might be *C. elegans* specific immune signaling genes. In cluster C2, on the other hand, we found genes that are specific to invertebrates and chordates species. They include *sta-2, ceh-18, lin-40, akir-1* (Polanowska et al., in preparation), and *tir-1* all of which have been the subject of detailed study and shown to be important for AMP gene regulation (Dierking et al., 2011; Couillault et al., 2004), acting via the p38 MAPK pathway. It will be of great interest to characterize the function of the remaining genes as they may as well as conserved immune regulators.

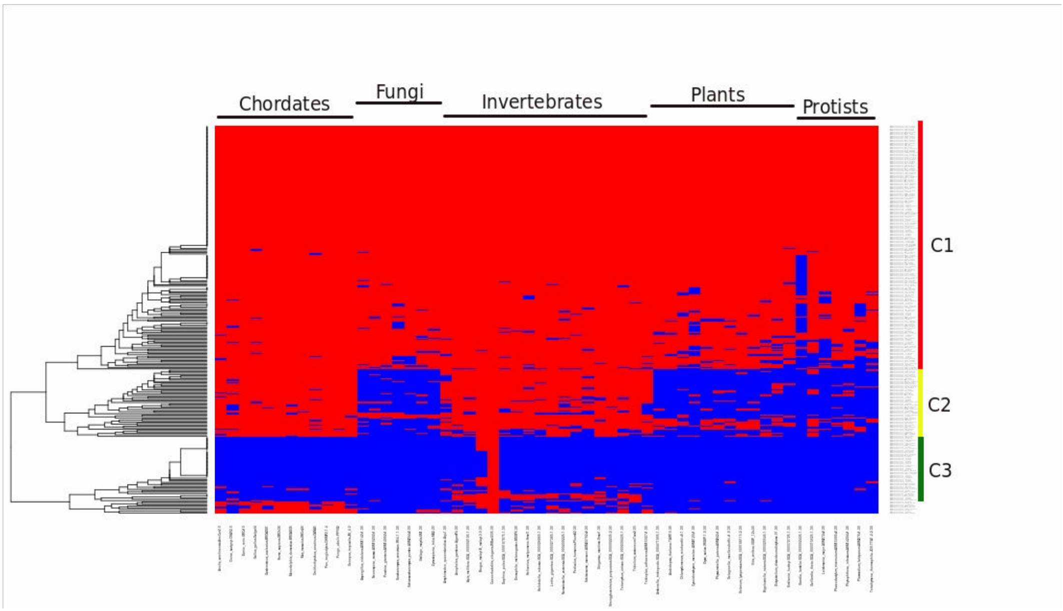


Figure 3.7: Phylogenetic conservation of 278 Nipi genes across 56 diverse eukaryotic species. Here rows represent genes and columns represent species. Three different clusters have been labelled on the right side of the figure. Red and blue mean presence or absence of the orthologs of given *C. elegans* protein in particular species, respectively.

3.2.7 Extensive crosstalk between different pathways

Other interesting functional classes found enriched through our EASE analysis were p38 MAPK-independent inducers of antibacterial effector gene *irg-1* and four classes related to osmotic stress. Osmotic stress was previously shown to regulate antimicrobial peptide gene expression independent of p38 MAPK pathway (Pujol et al., 2008b; Lee et al., 2010; Rohlfing et al., 2010). We tried to explore this link by directly testing the capacity of the 297 RNAi clones to induce expression of an irg-1p::gfp reporter gene (called I-clones) or block the increase in nlp-29p::gfp expression provoked by osmotic stress (called O-clones). Among 297 Nipi RNAi clones, 100 were found to be I-clones and 131 were O-clones. Epistasis analysis, testing the potential of these 297 Nipi clones to abrogate the elevated nlp-29p::gfp expression associated with a constitutively active form of GPA-12 (GPA-12*; (Ziegler et al., 2009a)), the alpha subunit of a heterotrimeric G protein that acts between DCAR-1 and TIR-1 (Zugasti et al., 2014), were also performed. Clones that abrogated this constitutive expression were called G-clones. These G-clones are expected to be p38 MAPK dependent (Ziegler et al., 2009a). Interestingly, 104 of these G-clones were also found to be O-clones. This epistatic analysis shows the complex crosstalk between different pathways.

3.2.8 Cross-tissue communication

In the screen we unexpectedly found that upon epidermis-specific knock-down of *vhp-1*, a known negative regulator of the p38 MAPK pathway (Kim et al., 2004), nlp-29p::gfp was induced in the intestine of the worm. This clearly shows that there is some kind of immune signal between two tissues. It will be extremely interesting to do an epidermis specific RNAi screen to identify more vhp-1 like genes which will help to build a cross-tissue network.

3.2.9 Conclusion

In this study, we identified more than 300 new immune signaling pathway genes, some of which, like *dcar-1* (Zugasti et al., 2014) will represent important but previously uncharacterised key regulators of the network. Through this quantitative screen, despite the limitations described above, we managed to capture even the network genes associated with a weak phenotype. Positioning them in a coherent network remains a challenge for the future.

3.3 Discussion of Publication 3 (Section II.3)

Global biological analyses through integrated functional and phylogenetic profiling.
Nishant Thakur, Nathalie Pujol, Jacques van Helden, Laurent Tichit, Jonathan J. Ewbank.

3.3.1 Summary

Among the techniques most frequently used to interpret high-throughput data, functional class enrichment analysis offers the promise of providing insights into overall biological mechanisms. Dozens of functional enrichment tools are already available in the public domain but almost all of them are generic. They lack the up-to-date and complete information available in species-specific databases. Regarding *C. elegans*, there is only one species-specific functional enrichment analysis tool, WormExp, which contains only transcriptome data; phenotypic data is completely missing. Here we present a *C. elegans* specific functional enrichment tool that we named YAAT, which is based on 4700 functional classes. Apart from phenotypic and transcriptomic data from the *C. elegans* database, Wormbase, it includes datasets manually extracted from literature, as well as data from other resources like DRSC Drosophila RNAi screens, C. elegans TF targets from the modENCODE consortium and assignments to specific pathways from the KEGG database. In this tool, we provide novel phylogenetic profiling and functional class clustering techniques that can be applied to any list of genes provided by the user. In this study, we analyzed multiple datasets and compared the results with WormExp tool and showed that YAAT gives more meaningful and complete results. We also analyzed different sets of random genes and found that WormExp tool gives enriched classes even for these random sets. This is clearly a significant limitation and could lead researchers to faulty conclusions. Finally, using global analysis of the YAAT tool datasets, we made a striking observation that if a query list is derived from a transcriptomic study, almost all the enriched classes will be too, whereas for gene lists constituted, for example, from a functional study one gets mixed enriched classes, transcriptomic and phenotypic. Unlike other tools, this tool is automatically updated and will be made freely available shortly.

3.3.2 Species-specific tools

There are many species-specific resources that are freely available and are manually updated very frequently compared to the general resources which contain partial and outdated data for different species. These latter resources are used for functional enrichment analysis. Either the data for these general tools has to be updated simultaneously with the organism-specific resources or organism-specific functional enrichment tools need to be made. Failure to use up-to-date data can lead to wrong conclusions (Wadi et al.). On the other hand, keeping resources up-to-date is not always simple. Our prior experience with Wormbase has shown that unanticipated changes to database structure can render programmes that interrogate Wormbase inoperable. Even the most established databases undergo significant changes, as witnessed by the imminent retirement of Sequence Gene Identifiers (GI) from Genbank (http://bit.ly/29JsibU). As the NCBI succinctly explained, Any code that parses GI numbers from NCBI FASTA records (from any NCBI source) will break. A tool that relies on data from multiple databases will be more susceptible to this type of problem.

Nevertheless, hundreds of general enrichment tool are available. Some tools use only Gene Ontology (GO) data, others use diverse kinds of functional data. Only a few tools are organism-specific and most of these tools are not updated frequently. In this article, we developed YAAT for *C. elegans* specific functional enrichment analysis.

3.3.3 Novel analysis and visualization techniques

With YAAT, we provided novel ways of data interpretation and visualization. In a result table of functional enrichment analysis, along with the P-value of enriched classes, the tool also provides conservation information of the enriched genes or all the genes in the reference class. Users can directly see whether the genes in enriched classes are evolutionary conserved or not. YAAT also provided clustering graphs of enriched functional classes, which can help in establishing and interpreting the links between different enriched classes. If two or more enriched classes are clustered together, this might signify some biological linkage between them. For example, in the enrichment analysis of 278 Nipi genes, the class JEEC_32 (Energy generation) is clustered with JEEC_16 (Mitochondrial protein), which makes biological sense, because mitochondrial proteins are necessary for energy generation. Other examples have been discussed in the YAAT tool manuscript. This novel clustering visualization technique provides a global picture of the enriched classes and help users to interpret the enrichment results. Another novel application of this tool is the phylogenetic profile clustering. As mentioned in the results section of the YAAT tool manuscript, genes that function together are often observed to co-evolve (Pellegrini et al., 1999). In this tool user can generate two different kinds of phylogenetic profiles, one with 56 diverse eukaryotic species and the second restricted to 72 nematodes species. The tool′s interface provides global conservation information for the user′s gene list. Unexpectedly, we found a bias in overall conservation of different kind of datasets. Phenotypic classes are strikingly more conserved than the transcriptomic classes from SPELL. Clustering of these phylogenetic profiles brings genes with the same conservation pattern together. This provides users with another level of functional coupling since the genes that cluster together might play a role in a similar pathway or biological process.

3.3.4 Improvements

In the coming release of this tool, we would like to add additional functionalities like the generation and visualization of a network of enriched classes. Firstly, the Hamming distance between the enriched classes will be calculated and based on a predetermined threshold, a distance network will be constructed as shown in Figure 3.8. This functionality will provide more meaningful interpretation of the enriched classes and help in understanding the biological links between the connected functional classes. Secondly, in the current release we just have eukaryotic species for phylogenetic profile analysis. We are contemplating incorporating of bacterial species because their integration might give some deep evolutionary insight of eukaryotic molecular functions as explained by the endosymbiotic theory (Martin et al., 2015; Archibald, 2015).

3.3.5 Conclusion

In this study we have developed a novel tool for high-throughput data analysis with the novel functionality of functional class clustering and phylogenetic profiling. We showed the various applications of this tool and analyzed many infection related datasets like Nipi genes (Zugasti et al., 2016), fungal up-regulated *C. elegans* effectors, etc. This

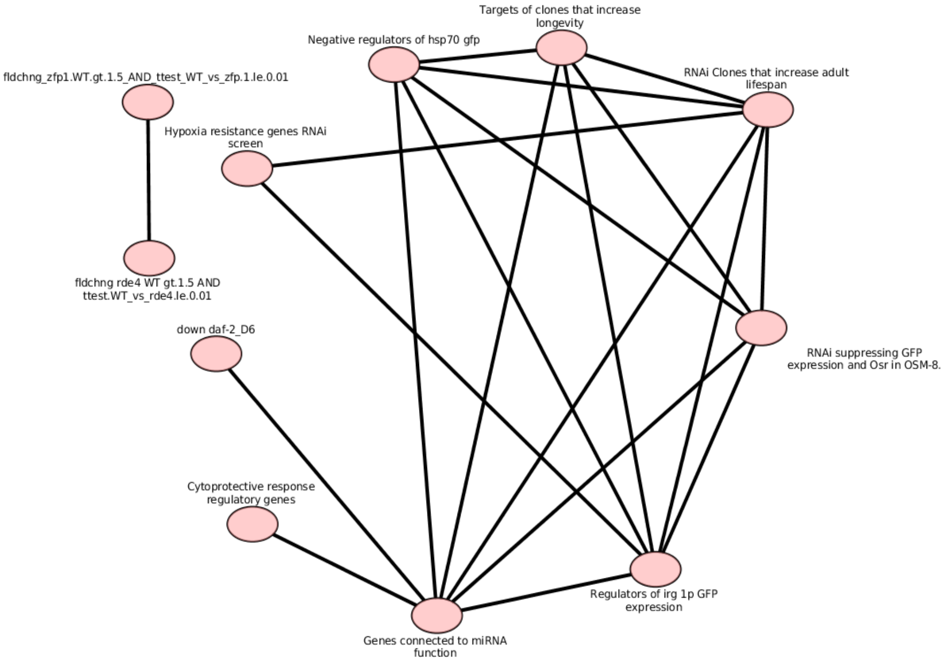

Targets of clones that increase longevity
Negative regulators of hsp70 gfp
fldchng_zfp1.WT.gt.1.5_AND_ttest_WT_vs_zfp.1.le.0.01
RNAi Clones that increase adult lifespan
Hypoxia resistance genes RNAi screen
fldchng rde4 WT gt.1.5 AND ttest.WT_vs_rde4.le.0.01
down daf-2_D6
RNAi suppressing GFP expression and Osr in OSM-8.
Cytoprotective response regulatory genes
Regulators of irg 1p GFP expression
Genes connected to miRNA function

Figure 3.8: Network representation of enriched classes (data used for Figure 5 in (Zugasti et al., 2016). A threshold of 20 for Hamming distances was used to filter out less significant links/interactions. Certain classes like RNAi Clones that increase adult lifespan and Targets of clones that increase longevity are very similar but not identical. Due to their different gene composition, only the former class is linked to Hypoxia resistance genes RNAi screen. Its is a big challenge to define parameters to automatically identify and remove redundant classes, or to consolidate classes that should include the same set of genes when these classes have been generated in independent studies.

tool will be freely available and will be automatically updated with every release of the Wormbase database. This tool will allow the *C. elegans* community to analyze high-throughput data from a totally different perspective.

3.4 On-going studies, perspectives and concluding remarks

During my thesis, I tried to build a regulatory network to describe the regulation of AMP genes.During my PhD, I followed the workflow shown in Figure 3.9. I analyzed various kind of data to achieve this goal. The first task, to identify the genes of the network was accomplished by genome-wide RNAi screen we identified 362 activator and repressor genes that are involved in AMP regulation. After multiple rounds of validation, we identified 28 genes with repressor activity. We showed that inactivation of these repressors increases spore adhesion and that this leads to a greater infectious burden, leading to higher induction of AMPs. These, therefore, do not represent elements of a regulatory network *in sensu stricto*. Given the widespread occurrence of the negative-regulatory circuits involved in immune responses across diverse species (e.g. (Aggarwal and Silverman, 2008; Liu et al., 2016; Forster et al., 2015; Pedraza-Alva et al., 2015)), it would be surprising if genuine negative regulators did not also exist in *C. elegans*. Identifying such genes will require targeted genetic screens. After multiple rounds of experimental validation, the genome-wide Nipi screen gave us 338 genes that are required for the activation of AMP gene expression. These potential regulatory genes were analyzed using various bioinformatic techniques. In order to perform gene enrichment analysis, we were obliged to remove the multiple histones and major sperm protein genes from the list. While eliminating one bias linked to inaccurate annotation, this clearly introduces another since these genes are not included in the query set. There is currently no obvious means to overcome this problem. The 278 non-histone and non-msp Nipi genes were enriched for many stress-related functional classes, like cytoprotective genes, genes involved in the mitochondrial UPR, mitochondrial surveillance, hypoxia-resistance, osmotic stress, etc.. We experimentally validated the role of UPRmt in AMP regulation, but showed that the effect of the corresponding Nipi genes was indirect, and indeed the consequence of cross-tissue signaling. This affects the cell-autonomous p38 MAPK pathway that is central to AMP gene regulation in the epidermis. We identified 50 genes potentially involved in this MAPK signaling pathway. We initiated epistasis analyses to position these different genes in the pathway; this work is still on-going. With the bioinformatics analysis, we also identified Ccr4-Not1 and NMD mRNA surveillance regulatory modules that post-transcriptionally regulate AMP genes expression. How these exert a specific function during the innate immune response will require further investigation. The next objective of my thesis was to obtain a time series of the transcriptional changes in *C. elegans* that follow *D. coniospora* infection.

To do so, we did RNA-sequencing at 2, 4, 6 and 8 hours post infection Figure 3.9B. Through a bioinformatics analysis, I identified more than 300 up-regulated genes. A brief overview of the RNA-seq analysis pipeline and are results is shown in the Figure 3.10. This list contains at least 20 AMPs, including 16 nlp and cnc AMPs. I analysed the ChIP-seq data for 94 TFs available through modENCODE consortium, Figure 3.9C. I did not find any TF enriched for these differentially expressed genes. So we chose to take an experimental approach. We selected 48 genes from among the 300 up-regulated genes and validated their differential expression using the Fluidigm high-throughput RT-PCR method, Figure 3.10B & C. We then took a total of 24 TFs for experimental validation (the 12 TFs found in Nipi RNAi screen together with 12 stress-related TFs selected from literature), Figure 3.10D and assayed whether knocking down a specific TF gene affected the infection-induced expression of any or all of the candidate effector genes. These experiments are still ongoing. Our hope is that the results will contribute to building the gene regulatory network Figure Figure 3.10E.

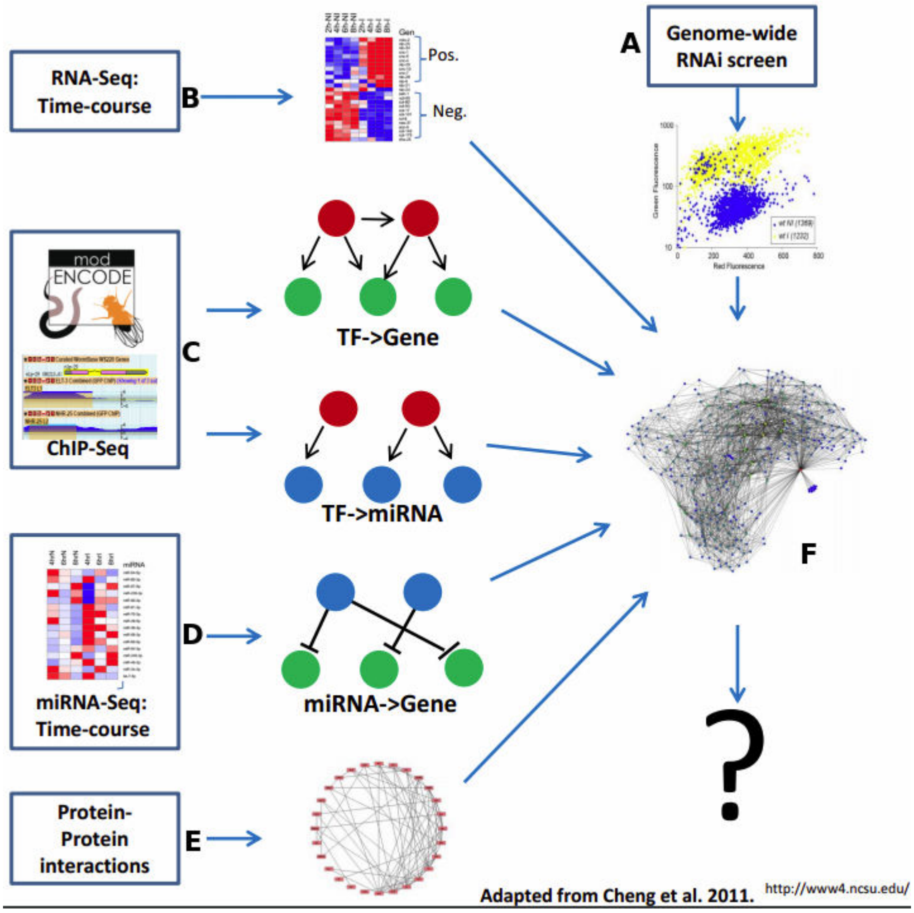


Figure 3.9: Workflow that was followed for my PhD thesis. A) Genome-wide RNAi screen data used for inferring the network nodes. B) Time course transcriptome profiling of infected C. elegans. C) ChIP-seq data analysis for identifying TFs that regulates AMPs. D) Small RNA-sequencing for identifying small non-coding RNAs that regulates AMP gene expression. E) Protein-protein interaction (PPI) data generated in-house and from publically available PPI resources. F) Integrated network construction from A-E data.

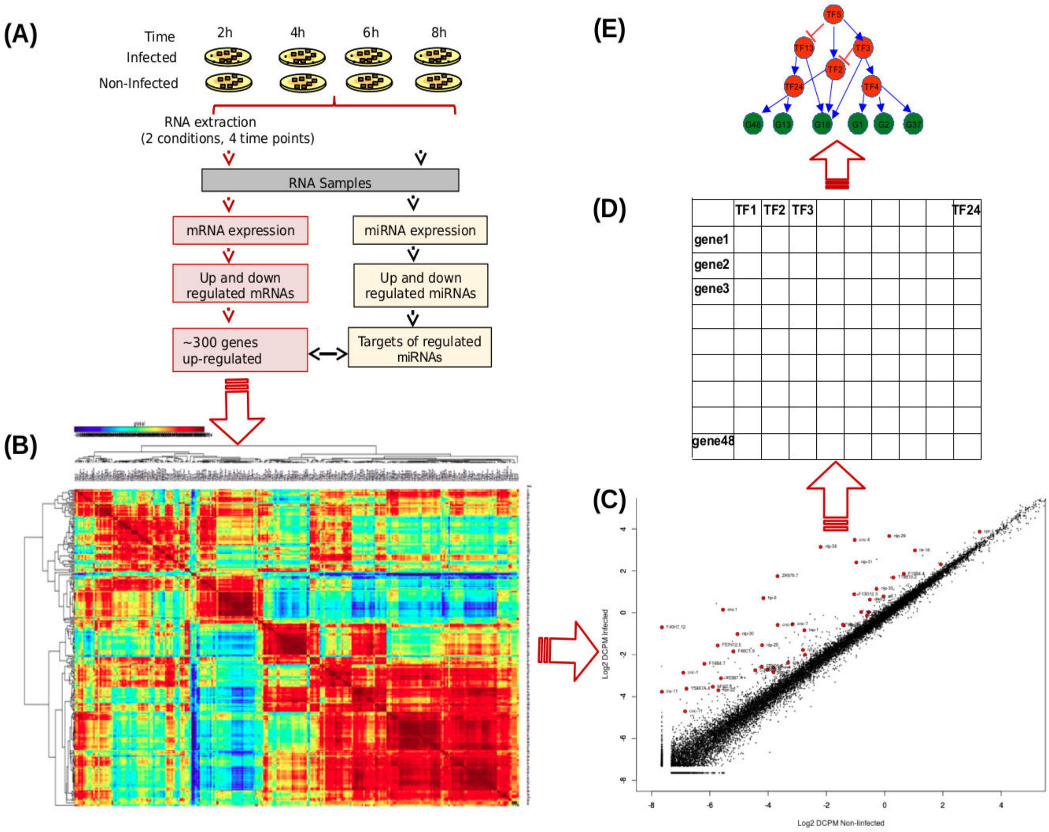


Figure 3.10: Workflow from RNA-seq to gene regulatory network construction. A) Time-course RNA sample preparation and analysis pipeline. B) Gene co-expression heatmap. We selected 48 genes from different clusters for Fludigm-based validation. C) Scatter plot showing expression level of genes in infected for 4 hours (x-axis) against non-infected worms (y-axis). The red dots are the 48 validated targets. DCPM: depth of coverage per million mapped reads (an expression metric). D) Experimental design for testing 48 effectors genes (rows) in 24 TFs (columns). E) Gene regulatory network construction from experimental data generated in step D.

Small RNAs like microRNAs (miRNA) plays an important role in post-transcriptional gene regulation Figure 3.9D. We have also conducted an RNA-sequencing of the same time-course samples that were used for RNA-seq. This led to the identification of a few differentially-regulated miRNAs. Further analysis and experimental validation is still pending. Therefore, although I have explored different aspects of AMP gene regulation, the main integrated network has yet to be constructed. The necessary data will all be available soon. Combining it to form a coherent representation of the innate immune network will be a challenge for the future.

Appendix A

Code

Although I developed various programming codes and packages for the analysis of only my data, but, I believe it can be useful for the *C. elegans* research community in general. I made all codes available to the public through https://github.com/nishantthakur/YAAT GitHub repository.

Similarly, I also developed a pipeline for identification of RNAi clone targets. Anyone can use this code and modify as needed. For both the tools, I tried my best, but, I hereby do not take any kind of responsibility for the authenticity of the results generated by these tools.

Perl code for RNAi clone target identification is on next page, to run this code, at first, you need to install MPSCAN program on your system.

```
##################################
## Developed by Nishant Thakur  ##
## Jonathan Ewbank's Lab        ##
## CIML, Marseille, France      ##
## October, 2014                ##
##################################
use Getopt::Std;
getopts('i:w:o:d:f:p:');
$in=$opt_i;#input clone sequence file in fasta format
$db=$opt_d;#input transcript file in fasta format,as follows
#>3R5.1a:WBGene00007065
#atgttttcaccgctcgagtgtcgtcttgctgttgctt

$win=$opt_w; #window size or mer size

$out=$opt_o; #output file name
$fold=$opt_f; # folder to process files , please provide the full path (no absolute)
$parallel=$opt_p; # number of parallel runs you want to execute

#subroutine to convert fasta file to single line fasta (idea from GPCR package)
sub fasta2sfasta (@fasta){
    @fasta_seq=@_;
    undef @ret_sfasta;
    $counter=-1;
    system "rm $fold/SFASTA";
    open(SF,">>$fold/SFASTA");
    foreach $fasta_seq_line (@fasta_seq){
        chomp($fasta_seq_line);
        if($fasta_seq_line=~/>/){if($counter  eq -1){print SF $fasta_seq_line."##"; $counter++}else{print SF "\n",$fasta_seq_line."##";}}
        else{print SF $fasta_seq_line;}
        }
            chmod(0777, "$fold/SFASTA");
            close SF;
    }

#subroutine to make fixed window (mers) in-silico siRNAs
sub make_mers{
        undef $line;
        $line=$_[0];
        chomp($line);
        $j=0;
        undef $file_name;
        $file_name=$_[1];
        open(OUT_SFASTA,">>$file_name");
        $win=$_[2];
        $psudoname=$j++;
        undef @split;
        @split=split("##",$line);
        for($i=0;$i<(length($split[1])-$win);$i++){
            $sub=lc(substr($split[1],$i,$win));
            print OUT_SFASTA "$split[0]\@$psudoname:$i:F\n$sub\n";
            undef $revcomp;
            $revcomp = reverse($sub);
            $revcomp =~ tr/ACGTacgt/TGCAtgca/;
            print OUT_SFASTA "$split[0]\@$psudoname:$i:RC\n$revcomp\n";
        }
        close OUT_SFASTA;
    }

```

```

if($in eq '' || $win eq '' || $out eq '' || $db eq ""||$fold eq "" ){print "Usage: perl run_mpscan.pl -i<input fasta file> -d <database(fasta sequence of the target)>-o <output> -w <mer size> -f<folder name>  -p <number of parallel run>\n";}
else{

        print "Making $fold/split folder...\n";
        system "rm -rf  $fold/split";
        system "mkdir $fold/split";

        #read each clone into an array
        open(IN,$in);
        @in=<IN>;
        close IN;

        #Fasta to single line fasta conversion
        print "Converting clone fasta to sfasta file...\n";
        fasta2sfasta(@in);
        print "Converted clone fasta to sfasta file\n";

        #Read single line fasta (SFASTA) to array
        open(SFASTA,"$fold/SFASTA");
        @clone_sf=<SFASTA>;
        close SFASTA;

        #making hash %c_l to store the length of each clone, will be used later
        print "Making $win mers from clones hash...\n";
        undef %mer_hash;
        undef %c_l;

        #foreach loop through each sfasta line and store the length into %c_l
        foreach $clone_sf(@clone_sf){
            chomp($clone_sf);
            undef @sp_clone;
            @sp_clone=split("##",$clone_sf);
            undef $cname;
            $cname=$sp_clone[0];
            $cname=~s/>//g;
            $c_l{$cname}=length($sp_clone[1]);
        }

        print "Making $win mers from clones file...\n";
        open(FH,"$fold/SFASTA");#read clones file(sfasta)

        #Define the number of parallel processing, this step is needed because sometimes it takes a long time to process the whole genome and also because sometimes the MPSCAN algorithm stuck in the loop if there are high repetitive 22 mers. I divided this it into multiple parallel runs (to avoid the recursive looping), finally I combine results together.

        while($clone=<FH>){
                   chomp($clone);
                   $pseudo_fname="SPLIT_0";
                   make_mers($clone,"$fold/split/$pseudo_fname",$win);

            for ($runi=1;$runi<=$parallel;$runi++){
                if($clone=<FH>){
                   chomp($clone);
                   $pseudo_fname="SPLIT_".$runi;
                   make_mers($clone,"$fold/split/$pseudo_fname",$win);
                }
            }
        }
        close FH;
        print "Making $win mers from clones file completed\n";



        ##change permission if needed
        chmod(0777, $fold);
        chmod(0777, "$fold/split");


        ##change directory to split folder


        $path=`pwd`;chdir("$fold/split");

        # get the list of the split files
        @list=`ls $fold/split/SPLIT_*`;
        chomp($path);

        ## read split files and then run mpscan on each file.
        $pids="";

        ### running parallel MPSCAN runs, this step is very very crucial if there are too many repetitive sequences, background processes (MPSCAN) will take a long time to finish and program will skip this loop, start calculating the targets (it could be incomplete sometimes)
        print "running parallel MPSCAN runs\n";
        foreach $job(@list){
            chomp($job);
            $job=~s/\s+//g;
            $job=~/SPLIT_(.+)/;
            $name=$1;
            $n++;
            system ("nohup $fold/mpscan.linux32 -p $fold/split/SPLIT_$name -t $db -r $fold/split/mpscan_out_$name &");
            $pids.=system($!);$pids.=" ";

        }

system("wait $pids");
        # wait for parallel MPSCAN run to finish total_clones/number_of_split
        sleep((scalar(@clone_sf)/scalar(@list))+(0.15*scalar(@clone_sf)));
        print "MPSCAN runs done...\n";

        $path=`pwd`;
        chdir("$fold/split");

        #cating all the mpscan result togather into one file MPSCAN_OUT
        system "cat mpscan_out_* |sed 's/;/\\t/'|sort -u>$fold/split/MPCSAN_OUT";

        #reading MPSCAN_OUT into "@targets" array
        open(FH,"$fold/split/MPCSAN_OUT");
        open(OUT1,">../$out");
        @targets=<FH>;
        close FH;

        ### mer target list
        undef %mertarget;$increase=0;
        undef @mapped_mer;

        ## MPSCAN_OUT file should look like this sjj_B0035.90@0:102:F B0035.9:WBGene00001920;68

        undef %mer22;
        undef %target;
        undef %results; undef $clone;undef $target;undef %results_clone;undef %clone_length;

#######################################################################################################################################################
        print "going into target mer matched.....\n";
        foreach $line(@targets){
                chomp $line;
                undef @split_line;
                @split_line=split("\t",$line);
                $clone=$split_line[0];

                $clone=~s/>//g;
                undef $clone_temp;
                $clone_temp=$clone;#sjj_B0035.90@0:102:F
                $clone=~s/:[0-9]+:[A-Z]+$//g;
                undef @seq;

                ################################### processing clone name ####################################
                undef @col_clone;

                ##split clone name in MPSCAN_OUT file #sjj_B0035.90@0:102:F
                @col_clone=split(":",$clone_temp);

                ##length of the query sequence
                $length_clone=$c_l[0];

                ##store this information in clone_length array
                $clone_length{$clone}=$length_clone;

                ##define $clone part
                undef $clone_part;
                $clone_part=sprintf("%d",($col_clone[1]/$win)); #102/$win

                ################################# processing target name########################################
                ###F54E12.3:WBGene00001930;54
                undef @col;
                @col=split(":",$split_line[1]);
                $target=$col[0];###$target=F54E12.3

                ##storing the total non-overlaping mer covered
                $results_clone{$clone}{$target}{$clone_part}++;

                ##storing the total overlaping mer matched in target
                $results{$clone}{$target}++;
        }
##################################################################################################################

```

```
        print "Out of the results_clone_loop!!!\n";

        #sjj_B0035.90@0:102:F
        print "going into results_clone loop.....\n";
        foreach $clone(keys %results_clone){
            chomp $clone;
            undef @new_array;
            @new_array=split(":",$clone);
            $f_c_s=$new_array[0];
            $f_c_s=~s/@.+//g;

            $length_clone=$c_l{$f_c_s};
            undef $pno;
            undef $pos;
            ##posible non-overlaping segements
            $pno=sprintf("%d",(($length_clone-1)/$win));

            ##posible overlaping segments
            $pos=$length_clone-$win;

            ##looping for each target of each clone
            while ( $clone, $tar = each(%{$results_clone{$clone}}) ) { # Problem is here   217
                chomp $tar;
                undef $mos;
                undef $mno;
                $count_seg=scalar(keys %{$results_clone{$clone}{$tar}});
                $count_seg1=sprintf("%.5f",$count_seg);

                #print "\n$clone\t$tar\t$pno\t$count_seg\t$pos\t$results{$clone}{$tar}\n";

                ##matched non-overlaping segments
                $mno=scalar(keys %{$results_clone{$clone}{$tar}});

                ##matched overlaping segments
                $mos=$results{$clone}{$tar};

                ## nos ratio
                undef $nos_ratio;
                $nos_ratio=($mno/$pno);
                $nos_ratio1=sprintf("%.5f",$nos_ratio);

                ## os ratio
                undef $os_ratio;
                $os_ratio=($mos/$pos);
                $os_ratio1=sprintf("%.5f",$os_ratio);


                ##final score
                undef $final_score;
                $final_score=(($nos_ratio1)*($nos_ratio1))*$os_ratio;
                $additive=($final_score)*($mos*0.1)*100;
                $additive1=sprintf("%.5f",$additive);

                if($additive1>100){$final_NT=100;}
                else{$final_NT=$additive;}
                print OUT1  "$clone\t$tar\t$pno\t$mno\t$pos\t$mos\t$nos_ratio1\t$os_ratio1\t$final_NT\n";
#               print OUT1
"$clone\t$tar\t$pno\t$mno\t$pos\t$mos\t$nos_ratio1\t$os_ratio1\t$final_NT\t$additive1\t$additive\t$final_score\n";
           }

       }
       #remove unwanted files
       system "rm -rf ../split ../SFASTA";
}

################################################################
### Please contact Nishant thakur at thakur@ciml.univ-mrs.fr ##
### if there is some problem running this program        ##
################################################################

```

Bibliography

Aggarwal, K. and N. Silverman (2008). "Positive and negative regulation of the Drosophila immune response." BMB Rep. **41**(4): 267-277.

Akira, S. and K. Takeda (2004). "Toll-like receptor signalling." Nat. Rev. Immunol. **4**(7): 499-511.

Akutsu, T., S. Miyano and S. Kuhara (1999). "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model." Pac. Symp. Biocomput.: 17-28.

Akutsu, T., S. Miyano and S. Kuhara (2000). "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function." J. Comput. Biol. **7**(3-4): 331-343.

Alegado, R. A., M. C. Campbell, W. C. Chen, S. S. Slutz and M.-W. Tan (2003). "Characterization of mediators of microbial virulence and innate immunity using the Caenorhabditis elegans host–pathogen model." Cell. Microbiol. **5**(7): 435-444.

Alegado, R. A. and M.-W. Tan (2008). "Resistance to antimicrobial peptides contributes to persistence of Salmonella typhimurium in the C. elegans intestine." Cell. Microbiol. **10**(6): 1259-1273.

Alper, S., S. J. McBride, B. Lackford, J. H. Freedman and D. A. Schwartz (2007). "Specificity and complexity of the Caenorhabditis elegans innate immune response." Mol. Cell. Biol. **27**(15): 5544-5553.

An, C. and Z. Mou (2013). "The function of the Mediator complex in plant immunity." Plant Signal. Behav. **8**(3): e23182.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol. **11**(10): R106.

Anders, S., P. T. Pyl and W. Huber (2014). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics: btu638.

Andrusiak, M. G. and Y. Jin (2016). "Context specificity of stress-activated MAP Kinase signaling: the story as told by C. elegans." J. Biol. Chem.

Annunziato, A. T. "DNA Packaging: Nucleosomes and Chromatin." Nature Education **1(1):26**.

Arala-Chaves, M. and T. Sequeira (2000). "Is there any kind of adaptive immunity in invertebrates?" Aquaculture **191**(1–3): 247-258.

Arber, W. and S. Linn (1969). "DNA modification and restriction." Annu. Rev. Biochem. **38**: 467-500.

Archibald, J. M. (2015). "Endosymbiosis and Eukaryotic Cell Evolution." Curr. Biol. **25**(19): R911-921.

Arthur, J. S. C. and S. C. Ley (2013). "Mitogen-activated protein kinases in innate immunity." Nat. Rev. Immunol. **13**(9): 679-692.

Ashrafi, K., F. Y. Chang, J. L. Watts, A. G. Fraser, R. S. Kamath, J. Ahringer and G. Ruvkun (2003). "Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes." Nature **421**(6920): 268-272.

Ashwell, J. D. (2006). "The many paths to p38 mitogen-activated protein kinase activation in the immune system." Nat. Rev. Immunol. **6**(7): 532-540.

Ausubel, F. M. (2005). "Are innate immune signaling pathways in plants and animals conserved?" Nat. Immunol. **6**(10): 973-979.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein and S. A. Teichmann (2004). "Structure and evolution of transcriptional regulatory networks." Curr. Opin. Struct. Biol. **14**(3): 283-291.

Bader, G. D., D. Betel and C. W. V. Hogue (2003). "BIND: the Biomolecular Interaction Network Database." Nucleic Acids Res. **31**(1): 248-250.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes and M. L. Bulyk (2009). "Diversity and complexity in DNA recognition by transcription factors." Science **324**(5935): 1720-1723.

Bahar, A. A. and D. Ren (2013). "Antimicrobial peptides." Pharmaceuticals **6**(12): 1543-1575.

Bailey, T., P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim and J. Zhang (2013). "Practical guidelines for the comprehensive analysis of ChIP-seq data." PLoS Comput. Biol. **9**(11): e1003326.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li and W. S. Noble (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Res. **37**(Web Server issue): W202-208.

Bandyopadhyay, S., C.-Y. Chiang, J. Srivastava, M. Gersten, S. White, R. Bell, C. Kurschner, C. H. Martin, M. Smoot, S. Sahasrabudhe, D. L. Barber, S. K. Chanda and T. Ideker (2010). "A human MAP kinase interactome." Nat. Methods **7**(10): 801-805.

Bányai, L. and L. Patthy (1998). "Amoebapore homologs of Caenorhabditis elegans." Biochim. Biophys. Acta **1429**(1): 259-264.

Baolin, L. and H. Bo (2007). HPRD: A High Performance RDF Database. Network and Parallel Computing. K. Li, C. Jesshope, H. Jin and J.-L. Gaudiot, Springer Berlin Heidelberg**:** 364-374.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-1712.

Barron, G. L. (1977). "Nematophagous destroying fungi." Guelph: Lancester press **1–140**.

Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano (2005). "Reverse engineering of regulatory networks in human B cells." Nat. Genet. **37**(4): 382-390.

Berger, M. F., G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Peña-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk and T. R. Hughes (2008). "Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences." Cell **133**(7): 1266-1276.

Berggård, T., S. Linse and P. James (2007). "Methods for the detection and analysis of protein–protein interactions." Proteomics **7**(16): 2833-2842.

Bigger, C. H., P. L. Jokiel and W. H. Hildemann (1983). "Cytotoxic transplantation immunity in the sponge Toxadocia violacea." Transplantation **35**(3): 239-243.

Bischof, L. J., C.-Y. Kao, F. C. O. Los, M. R. Gonzalez, Z. Shen, S. P. Briggs, F. G. van der Goot and R. V. Aroian (2008). "Activation of the unfolded protein response is required for defenses against bacterial pore-forming toxin in vivo." PLoS Pathog. **4**(10): e1000176.

Boehnisch, C., D. Wong, M. Habig, K. Isermann, N. K. Michiels, T. Roeder, R. C. May and H. Schulenburg (2011). "Protist-type lysozymes of the nematode Caenorhabditis elegans contribute to resistance against pathogenic Bacillus thuringiensis." PLoS One **6**(9): e24619.

Boutros, M. and J. Ahringer (2008). "The art and design of genetic screens: RNA interference." Nat. Rev. Genet. **9**(7): 554-566.

Brandt, J. P. and N. Ringstad (2015). "Toll-like Receptor Signaling Promotes Development and Function of Sensory Neurons Required for a C. elegans Pathogen-Avoidance Behavior." Curr. Biol. **25**(17): 2228-2237.

Brewster, J. L., T. de Valoir, N. D. Dwyer, E. Winter and M. C. Gustin (1993). "An osmosensing signal transduction pathway in yeast." Science **259**(5102): 1760-1763.

Butte, A. J. and I. S. Kohane (1999). "Unsupervised knowledge discovery in medical databases using relevance networks." Proc. AMIA Symp.: 711-715.

Campeau, E. and S. Gobeil (2011). "RNA interference in mammals: behind the screen." Brief. Funct. Genomics **10**(4): 215-226.

Chasse, S. A., P. Flanary, S. C. Parnell, N. Hao, J. Y. Cha, D. P. Siderovski and H. G. Dohlman (2006). "Genome-scale analysis reveals Sst2 as the principal regulator of mating pheromone signaling in the yeast Saccharomyces cerevisiae." Eukaryot. Cell **5**(2): 330-346.

Chávez, V., A. Mohri-Shiomi and D. A. Garsin (2009). "Ce-Duox1/BLI-3 generates reactive oxygen species as a protective innate immune mechanism in Caenorhabditis elegans." Infect. Immun. **77**(11): 4983-4989.

Cheesman, H. K., R. L. Feinbaum, J. Thekkiniath, R. H. Dowen, A. L. Conery and R. Pukkila-Worley (2016). "Aberrant Activation of p38 MAP Kinase-Dependent Innate Immune Responses Is Toxic to Caenorhabditis elegans." G3: Genes|Genomes|Genetics **6**(3): 541-549.

Cheng, Y. and F. Perocchi (2015). "ProtPhylo: identification of protein-phenotype and protein-protein functional associations via phylogenetic profiling." Nucleic Acids Res. **43**(W1): W160-168.

Choe, K. P. and K. Strange (2008). "Systemic osmotic signaling pathways function upstream of WNK and GCK-VI kinases to regulate hypertonic stress resistance in C. elegans." The FASEB Journal **22**(1 Supplement): 933.939-933.939.

Couillault, C., N. Pujol, J. Reboul, L. Sabatier, J.-F. Guichou, Y. Kohara and J. J. Ewbank (2004). "TLR-independent control of innate immunity in Caenorhabditis elegans by the TIR domain adaptor protein TIR-1, an ortholog of human SARM." Nat. Immunol. **5**(5): 488-494.

Cromar, G. L., A. Zhao, X. Xiong, L. S. Swapna, N. Loughran, H. Song and J. Parkinson (2016). "PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya." Database **2016**.

Darzacq, X., Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair and R. H. Singer (2007). "In vivo dynamics of RNA polymerase II transcription." Nat. Struct. Mol. Biol. **14**(9): 796-806.

Date, S. V. and E. M. Marcotte (2003). "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages." Nat. Biotechnol. **21**(9): 1055-1062.

Dey, G., A. Jaimovich, S. R. Collins, A. Seki and T. Meyer (2015). "Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling." Cell Rep.

Dey, G. and T. Meyer (2015). "Phylogenetic Profiling for Probing the Modular Architecture of the Human Genome." cels **1**(2): 106-115.

Dierking, K., J. Polanowska, S. Omi, I. Engelmann, M. Gut, F. Lembo, J. J. Ewbank and N. Pujol (2011). "Unusual regulation of a STAT protein by an SLC6 family transporter in C. elegans epidermal innate immunity." Cell Host Microbe **9**(5): 425-435.

Dierking, K., W. Yang and H. Schulenburg (2016). "Antimicrobial effectors in the nematode Caenorhabditis elegans: an outgroup to the Arthropoda." Philos. Trans. R. Soc. Lond. B Biol. Sci. **371**(1695).

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic and C. French StatOmique (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." Brief. Bioinform. **14**(6): 671-683.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras "STAR: ultrafast universal RNA-seq aligner."

Dowling, D. K. and L. W. Simmons (2009). "Reactive oxygen species as universal constraints in life-history evolution." Proc. Biol. Sci. **276**(1663): 1737-1745.

Ecker, J. R. and R. W. Davis (1986). "Inhibition of gene expression in plant cells by expression of antisense RNA." Proc. Natl. Acad. Sci. U. S. A. **83**(15): 5372-5376.

Engelmann, I. and N. Pujol (2010). "Innate immunity in C. elegans." Adv. Exp. Med. Biol. **708**: 105-121.

Estes, K. A., T. L. Dunbar, J. R. Powell, F. M. Ausubel and E. R. Troemel (2010). "bZIP transcription factor zip-2 mediates an early response to Pseudomonas aeruginosa infection in Caenorhabditis elegans." Proc. Natl. Acad. Sci. U. S. A. **107**(5): 2153-2158.

Evans, E. A., W. C. Chen and M.-W. Tan (2008). "The DAF-2 insulin-like signaling pathway independently regulates aging and immunity in C. elegans." Aging Cell **7**(6): 879-893.

Evans, E. A., T. Kawli and M.-W. Tan (2008). "Pseudomonas aeruginosa Suppresses Host Immunity by Activating the DAF-2 Insulin-Like Signaling Pathway in Caenorhabditis elegans." PLoS Pathog. **4**(10): e1000175.

Ewbank, J. J. (2002). "Tackling both sides of the host-pathogen equation with Caenorhabditis elegans." Microbes Infect. **4**(2): 247-256.

Ewbank, J. J. (2006). "Signaling in the immune response." WormBook: 1-12.

Ewbank, J. J. and N. Pujol (2015). "Local and long-range activation of innate immunity by infection and damage in C. elegans." Curr. Opin. Immunol. **38**: 1-7.

Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res. **8**(3): 175-185.

Félix, M.-A., A. Ashe, J. Piffaretti, G. Wu, I. Nuez, T. Bélicard, Y. Jiang, G. Zhao, C. J. Franz, L. D. Goldstein, M. Sanroman, E. A. Miska and D. Wang (2011). "Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses." PLoS Biol. **9**(1): e1000586.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver and C. C. Mello (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-811.

Firnhaber, C. and M. Hammarlund (2013). "Neuron-Specific Feeding RNAi in C. elegans and Its Use in a Screen for Essential Genes Required for GABA Neuron Function." PLoS Genet. **9**(11): e1003921.

Fjell, C. D., J. A. Hiss, R. E. W. Hancock and G. Schneider (2012). "Designing antimicrobial peptides: form follows function." Nat. Rev. Drug Discov. **11**(1): 37-51.

Forster, S. C., M. D. Tate and P. J. Hertzog (2015). "MicroRNA as Type I Interferon-Regulated Transcripts and Modulators of the Innate Immune Response." Front. Immunol. **6**: 334.

Franceschini, A., J. Lin, C. von Mering and L. J. Jensen (2016). "SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles." Bioinformatics **32**(7): 1085-1087.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering and Others (2013). "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Res. **41**(D1): D808-D815.

Friedman, A. and N. Perrimon (2006). "A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling." Nature **444**(7116): 230-234.

Friedman, A. and N. Perrimon (2007). "Genetic screening for signal transduction in the era of network biology." Cell **128**(2): 225-231.

Friedman, R. C., K. K.-H. Farh, C. B. Burge and D. P. Bartel (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Res. **19**(1): 92-105.

Galcheva-Gargova, Z., B. Dérijard, I. H. Wu and R. J. Davis (1994). "An osmosensing signal transduction pathway in mammalian cells." Science **265**(5173): 806-808.

Gardner, L. B. (2010). "Nonsense-mediated RNA decay regulation by cellular stress: implications for tumorigenesis." Mol. Cancer Res. **8**(3): 295-308.

Garsin, D. A., J. M. Villanueva, J. Begun, D. H. Kim, C. D. Sifri, S. B. Calderwood, G. Ruvkun and F. M. Ausubel (2003). "Long-lived C. elegans daf-2 mutants are resistant to bacterial pathogens." Science **300**(5627): 1921.

Gewirtz, A. T., T. A. Navas, S. Lyons, P. J. Godowski and J. L. Madara (2001). "Cutting edge: bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression." J. Immunol. **167**(4): 1882-1885.

Glass, L. and S. A. Kauffman (1973). "The logical analysis of continuous, non-linear biochemical control networks." J. Theor. Biol. **39**(1): 103-129.

Gordy, M. A., E. A. Pila and P. C. Hanington (2015). "The role of fibrinogen-related proteins in the gastropod immune response." Fish Shellfish Immunol. **46**(1): 39-49.

Grants, J. M., G. Y. S. Goh and S. Taubert (2015). "The Mediator complex of Caenorhabditis elegans: insights into the developmental and physiological roles of a conserved transcriptional coregulator." Nucleic Acids Res. **43**(4): 2442-2453.

Gravato-Nobre, M. J., H. R. Nicholas, R. Nijland, D. O'Rourke, D. E. Whittington, K. J. Yook and J. Hodgkin (2005). "Multiple genes affect sensitivity of Caenorhabditis elegans to the bacterial pathogen Microbacterium nematophilum." Genetics **171**(3): 1033-1045.

Graveley, B. R. (2001). "Alternative splicing: increasing diversity in the proteomic world." Trends Genet. **17**(2): 100-107.

Grishok, A. (2005). "RNAi mechanisms in Caenorhabditis elegans." FEBS Lett. **579**(26): 5932-5939.

Grove, C. A., F. De Masi, M. I. Barrasa, D. E. Newburger, M. J. Alkema, M. L. Bulyk and A. J. M. Walhout (2009). "A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors." Cell **138**(2): 314-327.

Gumienny, T. L. and C. Savage-Dunn (2013). "TGF-β signaling in C. elegans." WormBook: 1-34.

Guo, Y., S. Mahony and D. K. Gifford (2012). "High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints." PLoS Comput. Biol. **8**(8): e1002638.

Haas, B. J. and M. C. Zody (2010). "Advancing RNA-Seq analysis." Nat. Biotechnol. **28**(5): 421-423.

Hamilton, B., Y. Dong, M. Shindo, W. Liu, I. Odell, G. Ruvkun and S. S. Lee (2005). "A systematic RNAi screen for longevity genes in C. elegans." Genes Dev. **19**(13): 1544-1555.

Hammami, R., J. Ben Hamida, G. Vergoten and I. Fliss (2009). "PhytAMP: a database dedicated to antimicrobial plant peptides." Nucleic Acids Res. **37**(Database issue): D963-968.

Han, J., J. D. Lee, L. Bibbs and R. J. Ulevitch (1994). "A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells." Science **265**(5173): 808-811.

Hardison, R. C. and J. Taylor (2012). "Genomic approaches towards finding cis-regulatory modules in animals." Nat. Rev. Genet. **13**(7): 469-483.

Harford, J. B. and D. R. Morris (1997). MRNA Metabolism & Post-Transcriptional Gene Regulation, Wiley.

Harris, T. W., J. Baran, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, J. Done, C. Grove, K. Howe, R. Kishore, R. Lee, Y. Li, H.-M. Muller, C. Nakamura, P. Ozersky, M. Paulini, D. Raciti, G. Schindelman, M. A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, J. D. Wong, K. Yook, T. Schedl, J. Hodgkin, M. Berriman, P. Kersey, J. Spieth, L. Stein and P. W. Sternberg (2014). "WormBase 2014: new views of curated biology." Nucleic Acids Res. **42**(Database issue): D789-793.

Hartman, R. S. and R. D. Karp (1989). "Short-term immunologic memory in the allograft response of the American cockroach, Periplaneta americana." Transplantation **47**(5): 920-922.

Haskins, K. A., J. F. Russell, N. Gaddis, H. K. Dressman and A. Aballay (2008). "Unfolded protein response genes regulated by CED-1 are required for Caenorhabditis elegans innate immunity." Dev. Cell **15**(1): 87-97.

Hattori, A., T. Mizuno, M. Akamatsu, N. Hisamoto and K. Matsumoto (2013). "The Caenorhabditis elegans JNK signaling pathway activates expression of stress response genes by derepressing the Fos/HDAC repressor complex." PLoS Genet. **9**(2): e1003315.

Hayakawa, Y. and M. J. Smyth (2006). "CD27 dissects mature NK cells into two subsets with distinct responsiveness and migratory capacity." J. Immunol. **176**(3): 1517-1524.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman and R. Apweiler (2004). "IntAct: an open source molecular interaction database." Nucleic Acids Res. **32**(Database issue): D452-455.

Hildemann, W. H., C. H. Bigger, I. S. Johnston and P. L. Jokiel (1980). "Characteristics of transplantation immunity in the sponge, Callyspongia diffusa." Transplantation **30**(5): 362-367.

Hildemann, W. H., P. L. Jokiel, C. H. Bigger and I. S. Johnston (1980). "Allogeneic polymorphism and alloimmune memory in the coral, Montipora verrucosa." Transplantation **30**(4): 297-301.

Hillier, L. W., V. Reinke, P. Green, M. Hirst, M. A. Marra and R. H. Waterston (2009). "Massively parallel sequencing of the polyadenylated transcriptome of C. elegans." Genome Res. **19**(4): 657-666.

Hodgkin, J., P. E. Kuwabara and B. Corneliussen (2000). "A novel bacterial pathogen, Microbacterium nematophilum, induces morphological change in the nematode C. elegans." Curr. Biol. **10**(24): 1615-1618.

Hoeckendorf, A., M. Stanisak and M. Leippe (2012). "The saposin-like protein SPP-12 is an antimicrobial polypeptide in the pharyngeal neurons of Caenorhabditis elegans and participates in defence against a natural bacterial pathogen." Biochem. J **445**(2): 205-212.

Hoeven, R. v. d., K. C. McCallum, M. R. Cruz and D. A. Garsin (2011). "Ce-Duox1/BLI-3 generated reactive oxygen species trigger protective SKN-1 activity via p38 MAPK signaling during infection in C. elegans." PLoS Pathog. **7**(12): e1002453.

Hoffmann, J. A., F. C. Kafatos, C. A. Janeway and R. A. Ezekowitz (1999). "Phylogenetic perspectives in innate immunity." Science **284**(5418): 1313-1318.

Hu, Y., C. Roesel, I. Flockhart, L. Perkins, N. Perrimon and S. E. Mohr (2013). "UP-TORR: online tool for accurate and Up-to-Date annotation of RNAi Reagents." Genetics **195**(1): 37-45.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res. **37**(1): 1-13.

Huang, G., L. Z. Shi and H. Chi (2009). "Regulation of JNK and p38 MAPK in the immune system: signal integration, propagation and termination." Cytokine **48**(3): 161-169.

Huang, S., X. Tao, S. Yuan, Y. Zhang, P. Li, H. A. Beilinson, Y. Zhang, W. Yu, P. Pontarotti, H. Escriva, Y. Le Petillon, X. Liu, S. Chen, D. G. Schatz and A. Xu (2016). "Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination." Cell **166**(1): 102-114.

Hug, N., D. Longman and J. F. Cáceres (2016). "Mechanism and regulation of the nonsense-mediated decay pathway." Nucleic Acids Res. **44**(4): 1483-1495.

Hunter, C. P., W. M. Winston, C. Molodowitch, E. H. Feinberg, J. Shih, M. Sutherlin, A. J. Wright and M. C. Fitzgerald (2006). "Systemic RNAi in Caenorhabditis elegans." Cold Spring Harb. Symp. Quant. Biol. **71**: 95-100.

Ideo, H., K. Fukushima, K. Gengyo-Ando, S. Mitani, K. Dejima, K. Nomura and K. Yamashita (2009). "A Caenorhabditis elegans glycolipid-binding galectin functions in host defense against bacterial infection." J. Biol. Chem. **284**(39): 26493-26501.

Irazoqui, J. E., A. Ng, R. J. Xavier and F. M. Ausubel (2008). "Role for beta-catenin and HOX transcription factors in Caenorhabditis elegans and mammalian host epithelial-pathogen interactions." Proc. Natl. Acad. Sci. U. S. A. **105**(45): 17469-17474.

Irazoqui, J. E., E. R. Troemel, R. L. Feinbaum, L. G. Luhachack, B. O. Cezairliyan and F. M. Ausubel (2010). "Distinct pathogenesis and host responses during infection of C. elegans by P. aeruginosa and S. aureus." PLoS Pathog. **6**: e1000982.

Irazoqui, J. E., J. M. Urbach and F. M. Ausubel (2010). "Evolution of host innate defence: insights from Caenorhabditis elegans and primitive invertebrates." Nat. Rev. Immunol. **10**(1): 47-58.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). "Data Clustering: A Review." ACM Comput. Surv. **31**(3): 264-323.

Janeway, C. A., Jr. (1989). "Approaching the asymptote? Evolution and revolution in immunology." Cold Spring Harb. Symp. Quant. Biol. **54 Pt 1**: 1-13.

Janssens, S., B. Pulendran and B. N. Lambrecht (2014). "Emerging functions of the unfolded protein response in immunity." Nat. Immunol. **15**(10): 910-919.

Jansson, H. B. (1994). "Adhesion of Conidia of Drechmeria coniospora to Caenorhabditis elegans Wild Type and Mutants." J. Nematol. **26**(4): 430-435.

Ji, X., W. Li, J. Song, L. Wei and X. S. Liu (2006). "CEAS: cis-regulatory element annotation system." Nucleic Acids Res. **34**(Web Server issue): W551-554.

Jia, K., C. Thomas, M. Akbar, Q. Sun, B. Adams-Huet, C. Gilpin and B. Levine (2009). "Autophagy genes protect against Salmonella typhimurium infection and mediate insulin signaling-regulated pathogen resistance." Proc. Natl. Acad. Sci. U. S. A. **106**(34): 14564-14569.

Jim, K., K. Parmar, M. Singh and S. Tavazoie (2004). "A cross-genomic approach for systematic mapping of phenotypic traits to genes." Genome Res. **14**(1): 109-115.

Jollès, P. (1996). "Lysozymes: model enzymes in biochemistry and biology." EXS.

Jones, J. M. and M. Gellert (2004). "The taming of a transposon: V(D)J recombination and the immune system." Immunol. Rev. **200**: 233-248.

Jonkers, I. and J. T. Lis (2015). "Getting up to speed with transcription elongation by RNA polymerase II." Nat. Rev. Mol. Cell Biol. **16**(3): 167-177.

Jothi, R., S. Cuddapah, A. Barski, K. Cui and K. Zhao (2008). "Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data." Nucleic Acids Res. **36**(16): 5221-5231.

Kaderali, L. and N. Radde (2008). Inferring Gene Regulatory Networks from Expression Data. Computational Intelligence in Bioinformatics. A. Kelemen, A. Abraham and Y. Chen, Springer Berlin Heidelberg**:** 33-74.

Kamaladevi, A. and K. Balamurugan (2015). "Role of PMK-1/p38 MAPK defense in Caenorhabditis elegans against Klebsiella pneumoniae infection during host-pathogen interaction." Pathog. Dis. **73**(5).

Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen and J. Ahringer (2003). "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi." Nature **421**(6920): 231-237.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Res. **44**(D1): D457-462.

Karam, R., C.-H. Lou, H. Kroeger, L. Huang, J. H. Lin and M. F. Wilkinson (2015). "The unfolded protein response is shaped by the NMD pathway." EMBO Rep. **16**(5): 599-609.

Karam, R., J. Wengrod, L. B. Gardner and M. F. Wilkinson (2013). "Regulation of nonsense-mediated mRNA decay: implications for physiology and disease." Biochim. Biophys. Acta **1829**(6-7): 624-633.

Karlebach, G. and R. Shamir (2008). "Modelling and analysis of gene regulatory networks." Nat. Rev. Mol. Cell Biol. **9**(10): 770-780.

Kato, Y., T. Aizawa, H. Hoshino, K. Kawano, K. Nitta and H. Zhang (2002). "abf-1 and abf-2, ASABF-type antimicrobial peptide genes in Caenorhabditis elegans." Biochem. J **361**(Pt 2): 221-230.

Kauffman, S., C. Peterson, B. Samuelsson and C. Troein (2003). "Random Boolean network models and the yeast transcriptional network." Proc. Natl. Acad. Sci. U. S. A. **100**(25): 14796-14799.

Kauffman, S. A. The Origins of Order: Self-Organization and Selection in Evolution. Spin Glasses and Biology. year = 2012 edition =, WORLD SCIENTIFIC address = year = 2012 edition =**:** 61-100.

Kawli, T. and M.-W. Tan (2008). "Neuroendocrine signals modulate the innate immunity of Caenorhabditis elegans through insulin signaling." Nat. Immunol. **9**(12): 1415-1424.

Keshet, Y. and R. Seger (2010). "The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions." Methods Mol. Biol. **661**: 3-38.

Kharchenko, P. V., M. Y. Tolstorukov and P. J. Park (2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nat. Biotechnol. **26**(12): 1351-1359.

Kikugawa, S., K. Nishikata, K. Murakami, Y. Sato, M. Suzuki, M. Altaf-Ul-Amin, S. Kanaya and T. Imanishi (2012). "PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset." BMC Syst. Biol. **6 Suppl 2**: S7.

Kim, D. H. and J. J. Ewbank (2015). "Signaling in the innate immune response." WormBook: 1-51.

Kim, D. H., R. Feinbaum, G. Alloing, F. E. Emerson, D. A. Garsin, H. Inoue, M. Tanaka-Hino, N. Hisamoto, K. Matsumoto, M.-W. Tan and F. M. Ausubel (2002). "A conserved p38 MAP kinase pathway in Caenorhabditis elegans innate immunity." Science **297**(5581): 623-626.

Kim, D. H., N. T. Liberati, T. Mizuno, H. Inoue, N. Hisamoto, K. Matsumoto and F. M. Ausubel (2004). "Integration of Caenorhabditis elegans MAPK pathways mediating immunity and stress resistance by MEK-1 MAPK kinase and VHP-1 MAPK phosphatase." Proc. Natl. Acad. Sci. U. S. A. **101**(30): 10990-10994.

Kohchi, C., H. Inagawa, T. Nishizawa and G.-I. Soma (2009). "ROS and innate immunity." Anticancer Res. **29**(3): 817-821.

Koropatnick, T. A., J. T. Engle, M. A. Apicella, E. V. Stabb, W. E. Goldman and M. J. McFall-Ngai (2004). "Microbial factor-mediated development in a host-bacterial mutualism." Science **306**(5699): 1186-1188.

Kurz, C. L. and J. J. Ewbank (2007). "Infection in a dish: high-throughput analyses of bacterial pathogenesis." Curr. Opin. Microbiol. **10**(1): 10-16.

Kurz, C. L. and M.-W. Tan (2004). "Regulation of aging and innate immunity in C. elegans." Aging Cell **3**(4): 185-193.

Labed, S. A., S. Omi, M. Gut, J. J. Ewbank and N. Pujol (2012). "The pseudokinase NIPI-4 is a novel regulator of antimicrobial peptide gene expression." PLoS One **7**(3): e33887.

Labrousse, A., S. Chauvet, C. Couillault, C. L. Kurz and J. J. Ewbank (2000). "Caenorhabditis elegans is a model host for Salmonella typhimurium." Curr. Biol. **10**(23): 1543-1545.

Lähdesmäki, H., I. Shmulevich and O. Yli-Harja (2003). "On Learning Gene Regulatory Networks Under the Boolean Network Model." Mach. Learn. **52**(1-2): 147-167.

Lamitina, T., C. G. Huang and K. Strange (2006). "Genome-wide RNAi screening identifies protein damage as a regulator of osmoprotective gene expression." Proc. Natl. Acad. Sci. U. S. A. **103**(32): 12173-12178.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol. **10**(3): 1-10.

Latchman, D. S. (2002). Gene Regulation: A Eukaryotic Perspective, Garland Science.

Lebrigand, K., L. D. He, N. Thakur, M.-J. Arguel, J. Polanowska, B. Henrissat, E. Record, G. Magdelenat, V. Barbe, S. Raffaele, P. Barbry and J. J. Ewbank (2016). "Comparative Genomic Analysis of Drechmeria coniospora Reveals Core and Specific Genetic Requirements for Fungal Endoparasitism of Nematodes." PLoS Genet. **12**(5): e1006017.

Ledergerber, C. and C. Dessimoz (2011). "Base-calling for next-generation sequencing platforms." Brief. Bioinform. **12**(5): 489-497.

Lee, K.-Z., M. Kniazeva, M. Han, N. Pujol and J. J. Ewbank (2010). "The fatty acid synthase fasn-1 acts upstream of WNK and Ste20/GCK-VI kinases to modulate antimicrobial peptide expression in C. elegans epidermis." Virulence **1**(3): 113-122.

Lee, S. S., R. Y. N. Lee, A. G. Fraser, R. S. Kamath, J. Ahringer and G. Ruvkun (2003). "A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity." Nat. Genet. **33**(1): 40-48.

Lehner, B., J. Tischler and A. G. Fraser (2006). "RNAi screens in Caenorhabditis elegans in a 96-well liquid format and their application to the systematic identification of genetic interactions." Nat. Protoc. **1**(3): 1617-1620.

Lemaitre, B., E. Nicolas, L. Michaut, J. M. Reichhart and J. A. Hoffmann (1996). "The dorsoventral regulatory gene cassette spätzle/Toll/cactus controls the potent antifungal response in Drosophila adults." Cell **86**(6): 973-983.

Levasseur, A., M. Bekliz, E. Chabrière, P. Pontarotti, B. La Scola and D. Raoult (2016). "MIMIVIRE is a defence system in mimivirus that confers resistance to virophage." Nature **531**(7593): 249-252.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, H., J. Ruan and R. Durbin (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res. **18**(11): 1851-1858.

Li, R., Y. Li, K. Kristiansen and J. Wang (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-714.

Li, Y., S. E. Calvo, R. Gutman, J. S. Liu and V. K. Mootha (2014). "Expansion of biological pathways based on evolutionary inference." Cell **158**(1): 213-225.

Li, Y. and Z. Chen (2008). "RAPD: a database of recombinantly-produced antimicrobial peptides." FEMS Microbiol. Lett. **289**(2): 126-129.

Liang, S., S. Fuhrman and R. Somogyi (1998). "Reveal, a general reverse engineering algorithm for inference of genetic network architectures." Pac. Symp. Biocomput.: 18-29.

Liao, Y., G. K. Smyth and W. Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." Bioinformatics **30**(7): 923-930.

Liberati, N. T., K. A. Fitzgerald, D. H. Kim, R. Feinbaum, D. T. Golenbock and F. M. Ausubel (2004). "Requirement for a conserved Toll/interleukin-1 resistance domain protein in the Caenorhabditis elegans immune response." Proc. Natl. Acad. Sci. U. S. A. **101**(17): 6593-6598.

Liu, J., C. Qian and X. Cao (2016). "Post-Translational Modification Control of Innate Immunity." Immunity **45**(1): 15-30.

Liu, X., W. L. Kraus and X. Bai (2015). "Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways." Trends Biochem. Sci. **40**(9): 516-525.

Lu, Z. and S. Xu (2006). "ERK1/2 MAP kinases in cell survival and apoptosis." IUBMB Life **58**(11): 621-631.

Maas, S. (2010). "Gene regulation through RNA editing." Discov. Med. **10**(54): 379-386.

Machanick, P. and T. L. Bailey (2011). "MEME-ChIP: motif analysis of large DNA datasets." Bioinformatics **27**(12): 1696-1697.

Maia, A. F., M. E. Tanenbaum, M. Galli, D. Lelieveld, D. A. Egan, R. Gassmann, C. E. Sunkel, S. van den Heuvel and R. H. Medema (2015). "Genome-wide RNAi screen for synthetic lethal interactions with the C. elegans kinesin-5 homolog BMK-1." Sci Data **2**: 150020.

Mallo, G. V., C. L. Kurz, C. Couillault, N. Pujol, S. Granjeaud, Y. Kohara and J. J. Ewbank (2002). "Inducible antibacterial defense system in C. elegans." Curr. Biol. **12**(14): 1209-1214.

Marcotte, E. M., I. Xenarios, A. M. van Der Bliek and D. Eisenberg (2000). "Localizing proteins in the cell from their phylogenetic profiles." Proc. Natl. Acad. Sci. U. S. A. **97**(22): 12115-12120.

Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics **7 Suppl 1**: S7.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res. **18**(9): 1509-1517.

Marsh, E. K., M. C. W. van den Berg and R. C. May (2011). "A two-gene balance regulates Salmonella typhimurium tolerance in the nematode Caenorhabditis elegans." PLoS One **6**(3): e16839.

Martin, J. A. and Z. Wang (2011). "Next-generation transcriptome assembly." Nat. Rev. Genet. **12**(10): 671-682.

Martin, W. F., S. Garg and V. Zimorski (2015). "Endosymbiotic theories for eukaryote origin." Philos. Trans. R. Soc. Lond. B Biol. Sci. **370**(1678): 20140330.

Maxwell, C. S., W. S. Kruesi, L. J. Core, N. Kurhanewicz, C. T. Waters, C. L. Lewarch, I. Antoshechkin, J. T. Lis, B. J. Meyer and L. R. Baugh (2014). "Pol II Docking and Pausing at Growth and Stress Genes in C. elegans." Cell Rep. **6**(3): 455-466.

May, R. C. and R. H. A. Plasterk (2005). "RNA interference spreading in C. elegans." Methods Enzymol. **392**: 308-315.

McElwee, J. J., E. Schuster, E. Blanc, J. H. Thomas and D. Gems (2004). "Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance." J. Biol. Chem. **279**(43): 44533-44543.

McLean, C. Y., D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger and G. Bejerano (2010). "GREAT improves functional interpretation of cis-regulatory regions." Nat. Biotechnol. **28**(5): 495-501.

Medina-Rivera, A., M. Defrance, O. Sand, C. Herrmann, J. A. Castro-Mondragon, J. Delerce, S. Jaeger, C. Blanchet, P. Vincens, C. Caron, D. M. Staines, B. Contreras-Moreira, M. Artufel, L. Charbonnier-Khamvongsa, C. Hernandez, D. Thieffry, M. Thomas-Chollier and J. van Helden (2015). "RSAT 2015: Regulatory Sequence Analysis Tools." Nucleic Acids Res.

Medzhitov, R. and C. A. Janeway, Jr. (1997). "Innate immunity: the virtues of a nonclonal system of recognition." Cell **91**(3): 295-298.

Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander and B. E. Bernstein (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Miller, E. V., L. N. Grandi, J. A. Giannini, J. D. Robinson and J. R. Powell (2015). "The Conserved G-Protein Coupled Receptor FSHR-1 Regulates Protective Host Responses to Infection and Oxidative Stress." PLoS One **10**(9): e0137403.

Miltsch, S. M., P. H. Seeberger and B. Lepenies (2014). "The C-type lectin-like domain containing proteins Clec-39 and Clec-49 are crucial for Caenorhabditis elegans immunity against Serratia marcescens infection." Dev. Comp. Immunol. **45**(1): 67-73.

Mlotshwa, S., O. Voinnet, M. F. Mette, M. Matzke, H. Vaucheret, S. W. Ding, G. Pruss and V. B. Vance (2002). "RNA silencing and the mobile silencing signal." Plant Cell **14 Suppl**: S289-301.

Muir, R. E. and M. W. Tan (2008). "Virulence of Leucobacter chromiireducens subsp. solipictus to Caenorhabditis elegans: characterization of a novel host-pathogen interaction." Appl. Environ. Microbiol.

Murphy, C. T., S. A. McCarroll, C. I. Bargmann, A. Fraser, R. S. Kamath, J. Ahringer, H. Li and C. Kenyon (2003). "Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans." Nature **424**(6946): 277-283.

Nathan, C. and A. Cunningham-Bussel (2013). "Beyond oxidative stress: an immunologist's guide to reactive oxygen species." Nat. Rev. Immunol. **13**(5): 349-361.

Nathoo, A. N., R. A. Moeller, B. A. Westlund and A. C. Hart (2001). "Identification of neuropeptide-like protein gene families in Caenorhabditis elegans and other species." Proceedings of the National Academy of Sciences **98**(24): 14000-14005.

Neves, G., J. Zucker, M. Daly and A. Chess (2004). "Stochastic yet biased expression of multiple Dscam splice variants by individual cells." Nat. Genet. **36**(3): 240-246.

Nicholas, H. R. and J. Hodgkin (2004). "The ERK MAP kinase cascade mediates tail swelling and a protective response to rectal infection in C. elegans." Curr. Biol. **14**(14): 1256-1261.

Norga, K. K., M. C. Gurganus, C. L. Dilda, A. Yamamoto, R. F. Lyman, P. H. Patel, G. M. Rubin, R. A. Hoskins, T. F. Mackay and H. J. Bellen (2003). "Quantitative analysis of bristle number in Drosophila mutants identifies genes involved in neural development." Curr. Biol. **13**(16): 1388-1396.

O'Neill, L. A. J. and A. G. Bowie (2007). "The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling." Nat. Rev. Immunol. **7**(5): 353-364.

O'Neill, L. P. and B. M. Turner (1996). "Immunoprecipitation of chromatin." Methods Enzymol. **274**: 189-197.

O'Rourke, D., D. Baban, M. Demidova, R. Mott and J. Hodgkin (2006). "Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of C. elegans with M. nematophilum." Genome Res. **16**(8): 1005-1016.

Oshlack, A., M. D. Robinson and M. D. Young (2010). "From RNA-seq reads to differential expression results." Genome Biol. **11**(12): 220.

Pak, J. and A. Fire (2007). "Distinct populations of primary and secondary effectors during RNAi in C. elegans." Science **315**(5809): 241-244.

Palm, N. W. and R. Medzhitov (2009). "Pattern recognition receptors and control of adaptive immunity." Immunol. Rev. **227**(1): 221-233.

Pasupuleti, M., A. Schmidtchen and M. Malmsten (2012). "Antimicrobial peptides: key components of the innate immune system." Crit. Rev. Biotechnol. **32**(2): 143-171.

Pedraza-Alva, G., L. Pérez-Martínez, L. Valdez-Hernández, K. F. Meza-Sosa and M. Ando-Kuri (2015). "Negative regulation of the inflammasome: keeping inflammation under control." Immunol. Rev. **265**(1): 231-257.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc. Natl. Acad. Sci. U. S. A. **96**(8): 4285-4288.

Pellegrino, M. W., A. M. Nargund, N. V. Kirienko, R. Gillis, C. J. Fiorese and C. M. Haynes (2014). "Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection." Nature.

Pepke, S., B. Wold and A. Mortazavi (2009). "Computation for ChIP-seq and RNA-seq studies." Nat. Methods **6**: S22-S32.

Phillip West, A., G. S. Shadel and S. Ghosh (2011). "Mitochondria in innate immune responses." Nat. Rev. Immunol. **11**(6): 389-402.

Piotto, S. P., L. Sessa, S. Concilio and P. Iannelli (2012). "YADAMP: yet another database of antimicrobial peptides." Int. J. Antimicrob. Agents **39**(4): 346-351.

Poulin, G., R. Nandakumar and J. Ahringer (2004). "Genome-wide RNAi screens in Caenorhabditis elegans: impact on cancer research." Oncogene **23**(51): 8340-8345.

Powell, J. R. and F. M. Ausubel (2008). "Models of Caenorhabditis elegans infection by bacterial and fungal pathogens." Methods Mol. Biol. **415**: 403-427.

Powell, J. R., D. H. Kim and F. M. Ausubel (2009). "The G protein-coupled receptor FSHR-1 is required for the Caenorhabditis elegans innate immune response." Proc. Natl. Acad. Sci. U. S. A. **106**(8): 2782-2787.

Pujol, N., S. Cypowyj, K. Ziegler, A. Millet, A. Astrain, A. Goncharov, Y. Jin, A. D. Chisholm and J. J. Ewbank (2008). "Distinct innate immune responses to infection and wounding in the C. elegans epidermis." Curr. Biol. **18**(7): 481-489.

Pujol, N., O. Zugasti, D. Wong, C. Couillault, C. L. Kurz, H. Schulenburg and J. J. Ewbank (2008). "Anti-fungal innate immunity in C. elegans is enhanced by evolutionary diversification of antimicrobial peptides." PLoS Pathog. **4**(7): e1000105.

Pukkila-Worley, R., F. M. Ausubel and E. Mylonakis (2011). "Candida albicans infection of Caenorhabditis elegans induces antifungal immune defenses." PLoS Pathog. **7**(6): e1002074.

Pukkila-Worley, R., R. L. Feinbaum, D. L. McEwan, A. L. Conery and F. M. Ausubel (2014). "The Evolutionarily Conserved Mediator Subunit MDT-15/MED15 Links Protective Innate Immune Responses and Xenobiotic Detoxification." PLoS Pathog. **10**(5): e1004143.

Qu, W., C. Ren, Y. Li, J. Shi, J. Zhang, X. Wang, X. Hang, Y. Lu, D. Zhao and C. Zhang (2011). "Reliability analysis of the Ahringer Caenorhabditis elegans RNAi feeding library: a guide for genome-wide screens." BMC Genomics **12**: 170.

Radonjic, M., J.-C. Andrau, P. Lijnzaad, P. Kemmeren, T. T. J. P. Kockelkorn, D. van Leenen, N. L. van Berkum and F. C. P. Holstege (2005). "Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit." Mol. Cell **18**(2): 171-183.

Rapaport, F., R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci and D. Betel (2013). "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data." Genome Biol. **14**(9): R95.

Razick, S., G. Magklaras and I. M. Donaldson (2008). "iRefIndex: a consolidated protein interaction database with provenance." BMC Bioinformatics **9**: 405.

Reboul, J., P. Vaglio, N. Tzellas, N. Thierry-Mieg, T. Moore, C. Jackson, T. Shin-i, Y. Kohara, D. Thierry-Mieg, J. Thierry-Mieg, H. Lee, J. Hitti, L. Doucette-Stamm, J. L. Hartley, G. F. Temple, M. A. Brasch, J. Vandenhaute, P. E. Lamesch, D. E. Hill and M. Vidal (2001). "Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans." Nat. Genet. **27**(3): 332-336.

Reece-Hoyes, J. S., B. Deplancke, J. Shingles, C. A. Grove, I. A. Hope and A. J. M. Walhout (2005). "A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks." Genome Biol. **6**(13): R110.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-2309.

Ren, M., H. Feng, Y. Fu, M. Land and C. S. Rubin (2009). "Protein kinase D is an essential regulator of C. elegans innate immunity." Immunity **30**(4): 521-532.

Richardson, C. E., T. Kooistra and D. H. Kim (2010). "An essential role for XBP-1 in host protection against immune activation in C. elegans." Nature **463**(7284): 1092-1095.

Robinson, J. D. and J. R. Powell (2016). "Long-term recovery from acute cold shock in Caenorhabditis elegans." BMC Cell Biol. **17**: 2.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Roeder, R. G. (2005). "Transcriptional regulation and the role of diverse coactivators in animal cells." FEBS Lett. **579**(4): 909-915.

Roeder, T., M. Stanisak, C. Gelhaus, I. Bruchhaus, J. Grötzinger and M. Leippe (2010). "Caenopores are antimicrobial peptides in the nematode Caenorhabditis elegans instrumental in nutrition and immunity." Dev. Comp. Immunol. **34**(2): 203-209.

Rohlfing, A.-K., Y. Miteva, S. Hannenhalli and T. Lamitina (2010). "Genetic and physiological activation of osmosensitive gene expression mimics transcriptional signatures of pathogen infection in C. elegans." PLoS One **5**(2): e9010.

Roy, B., L. M. Haupt and L. R. Griffiths (2013). "Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity." Curr. Genomics **14**(3): 182-194.

Rual, J.-F., J. Ceron, J. Koreth, T. Hao, A.-S. Nicot, T. Hirozane-Kishikawa, J. Vandenhaute, S. H. Orkin, D. E. Hill, S. van den Heuvel and M. Vidal (2004). "Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library." Genome Res. **14**(10B): 2162-2168.

Ruepp, A., B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone and H. W. Mewes (2010). "CORUM: the comprehensive resource of mammalian protein complexes--2009." Nucleic Acids Res. **38**(Database issue): D497-501.

Sadreyev, I. R., F. Ji, E. Cohen, G. Ruvkun and Y. Tabach (2015). "PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles." Nucleic Acids Res.

Salter-Cid, L. and C. H. Bigger (1991). "Alloimmunity in the Gorgonian Coral Swiftia exserta." Biol. Bull. **181**(1): 127-134.

Samanta, S. and J. K. Thakur (2015). "Importance of Mediator complex in the regulation and integration of diverse signaling pathways in plants." Front. Plant Sci. **6**: 757.

Sarkies, P., A. Ashe, J. Le Pen, M. A. McKie and E. A. Miska (2013). "Competition between virus-derived and endogenous small RNAs regulates gene expression in Caenorhabditis elegans." Genome Res. **23**(8): 1258-1270.

Schlitt, T. and A. Brazma (2006). "Modelling in molecular biology: describing transcription regulatory networks at different scales." Philos. Trans. R. Soc. Lond. B Biol. Sci. **361**(1467): 483-494.

Schmitt, T., C. Ogris and E. L. L. Sonnhammer (2014). "FunCoup 3.0: database of genome-wide functional coupling networks." Nucleic Acids Res. **42**(Database issue): D380-388.

Schulenburg, H. and J. J. Ewbank (2007). "The genetics of pathogen avoidance in Caenorhabditis elegans." Mol. Microbiol. **66**(3): 563-570.

Schulenburg, H., M. P. Hoeppner, J. Weiner, 3rd and E. Bornberg-Bauer (2008). "Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode Caenorhabditis elegans." Immunobiology **213**(3-4): 237-250.

Shabalina, S. A. and E. V. Koonin (2008). "Origins and evolution of eukaryotic RNA interference." Trends Ecol. Evol. **23**(10): 578-587.

Sifri, C. D., J. Begun, F. M. Ausubel and S. B. Calderwood (2003). "Caenorhabditis elegans as a model host for Staphylococcus aureus pathogenesis." Infect. Immun. **71**(4): 2208-2217.

Simmer, F., C. Moorman, A. M. van der Linden, E. Kuijk, P. V. E. van den Berghe, R. S. Kamath, A. G. Fraser, J. Ahringer and R. H. A. Plasterk (2003). "Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions." PLoS Biol. **1**(1): E12.

Simonsen, M., S. R. Maetschke and M. A. Ragan (2012). "Automatic selection of reference taxa for protein–protein interaction prediction with phylogenetic profiling." Bioinformatics **28**(6): 851-857.

Singh, V. and A. Aballay (2009). "Regulation of DAF-16-mediated Innate Immunity in Caenorhabditis elegans." J. Biol. Chem. **284**(51): 35580-35587.

Squiban, B., J. Belougne, J. Ewbank and O. Zugasti (2012). "Quantitative and automated high-throughput genome-wide RNAi screens in C. elegans." J. Vis. Exp.(60).

Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res. **34**(Database issue): D535-539.

Stuart, A. (1993). "Kaufmann. The Origins of Order: Self-Organization and Selection in Evolution." Oxford: Oxford University Press **353**: 354.

Stuart, L. M. and R. A. Ezekowitz (2008). "Phagocytosis and comparative innate immunity: learning on the fly." Nat. Rev. Immunol. **8**(2): 131-141.

Su, A. A. H. and L. Randau (2011). "A-to-I and C-to-U editing within transfer RNAs." Biochemistry **76**(8): 932-937.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc. Natl. Acad. Sci. U. S. A. **102**(43): 15545-15550.

Sun, Y., P. Yang, Y. Zhang, X. Bao, J. Li, W. Hou, X. Yao, J. Han and H. Zhang (2011). "A genome-wide RNAi screen identifies genes regulating the formation of P bodies in C. elegans and their functions in NMD and RNAi." Protein Cell **2**(11): 918-939.

Sundararajan, V. S., M. N. Gabere, A. Pretorius, S. Adam, A. Christoffels, M. Lehväslaiho, J. A. C. Archer and V. B. Bajic (2012). "DAMPD: a manually curated antimicrobial peptide database." Nucleic Acids Res. **40**(D1): D1108-D1112.

Tabach, Y., A. C. Billi, G. D. Hayes, M. A. Newman, O. Zuk, H. Gabel, R. Kamath, K. Yacoby, B. Chapman, S. M. Garcia, M. Borowsky, J. K. Kim and G. Ruvkun (2013). "Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence." Nature **493**(7434): 694-698.

Tabach, Y., T. Golan, A. Hernández-Hernández, A. R. Messer, T. Fukuda, A. Kouznetsova, J.-G. Liu, I. Lilienthal, C. Levy and G. Ruvkun (2013). "Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling." Mol. Syst. Biol. **9**: 692.

Tan, M.-W. (2002). "Cross-species infections and their analysis." Annu. Rev. Microbiol. **56**: 539-565.

Tan, M. W. (2001). "Genetic and genomic dissection of host-pathogen interactions using a P. aeruginosa-C. elegans pathogenesis model." Pediatr. Pulmonol. **32**: 96-97.

Tenor, J. L. and A. Aballay (2008). "A conserved Toll-like receptor is required for Caenorhabditis elegans innate immunity." EMBO Rep. **9**(1): 103-109.

Thakur, N., N. Pujol, L. Tichit and J. J. Ewbank (2014). "Clone mapper: an online suite of tools for RNAi experiments in Caenorhabditis elegans." G3 **4**(11): 2137-2145.

Thomas-Chollier, M., C. Herrmann, M. Defrance, O. Sand, D. Thieffry and J. van Helden (2012). "RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets." Nucleic Acids Res. **40**(4): e31.

Timmons, L., D. L. Court and A. Fire (2001). "Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans." Gene **263**(1-2): 103-112.

Timmons, L. and A. Fire (1998). "Specific interference by ingested dsRNA." Nature **395**(6705): 854.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nat. Biotechnol. **31**(1): 46-53.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat. Protoc. **7**(3): 562-578.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat. Biotechnol. **28**(5): 511-515.

Travis, J. (2009). "Origins. On the origin of the immune system." Science **324**(5927): 580-582.

Troemel, E. R., M.-A. Félix, N. K. Whiteman, A. Barrière and F. M. Ausubel (2008). "Microsporidia are natural intracellular parasites of the nematode Caenorhabditis elegans." PLoS Biol. **6**(12): 2736-2752.

van den Berg, L. M., S. I. Gringhuis and T. B. H. Geijtenbeek (2012). "An evolutionary perspective on C-type lectins in infection and immunity." Ann. N. Y. Acad. Sci. **1253**: 149-158.

van der Hoeven, R., K. C. McCallum and D. A. Garsin (2012). "Speculations on the activation of ROS generation in C. elegans innate immune signaling." Worm **1**(3): 160-163.

van Nimwegen, E. (2003). "Scaling laws in the functional content of genomes." Trends Genet. **19**(9): 479-484.

Vance, R. E., R. R. Isberg and D. A. Portnoy (2009). "Patterns of pathogenesis: discrimination of pathogenic and nonpathogenic microbes by the innate immune system." Cell Host Microbe **6**(1): 10-21.

Vinayagam, A., J. Zirin, C. Roesel, Y. Hu, B. Yilmazel, A. A. Samsonova, R. A. Neumüller, S. E. Mohr and N. Perrimon (2014). "Integrating protein-protein interaction networks with phenotypes reveals signs of interactions." Nat. Methods **11**(1): 94-99.

Voinnet, O. (2005). "Non-cell autonomous RNA silencing." FEBS Lett. **579**(26): 5858-5871.

Wadi, L., M. Meyer, J. Weiser, L. D. Stein and J. Reimand "Impact of knowledge accumulation on pathway enrichment analysis."

Waghu, F. H., L. Gopi, R. S. Barai, P. Ramteke, B. Nizami and S. Idicula-Thomas (2014). "CAMP: Collection of sequences and structures of antimicrobial peptides." Nucleic Acids Res. **42**(Database issue): D1154-1158.

Walhout, M., M. Vidal and J. Dekker (2012). Handbook of Systems Biology: Concepts and Insights, Elsevier Science.

Wan, C., B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S. Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte and A. Emili (2015). "Panorama of ancient metazoan macromolecular complexes." Nature **525**(7569): 339-344.

Wang, J., W. A. Mohler and C. Savage-Dunn (2005). "C-terminal mutants of C. elegans Smads reveal tissue-specific requirements for protein activation by TGF-beta signaling." Development **132**(15): 3505-3513.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat. Rev. Genet. **10**(1): 57-63.

Watson, F. L., R. Püttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel and D. Schmucker (2005). "Extensive diversity of Ig-superfamily proteins in the immune system of insects." Science **309**(5742): 1874-1878.

Wilson, R. C. and J. A. Doudna (2013). "Molecular mechanisms of RNA interference." Annu. Rev. Biophys. **42**: 217-239.

Wittkopp, P. J. and G. Kalay (2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence." Nat. Rev. Genet. **13**(1): 59-69.

Wong, D., D. Bazopoulou, N. Pujol, N. Tavernarakis and J. J. Ewbank (2007). "Genome-wide investigation reveals pathogen-specific and shared signatures in the response of Caenorhabditis elegans to infection." Genome Biol. **8**(9): R194.

Wu, T. D. and S. Nacu (2010). "Fast and SNP-tolerant detection of complex variants and splicing in short reads." Bioinformatics **26**(7): 873-881.

Xie, Q. and H.-S. Guo (2006). "Systemic antiviral silencing in plants." Virus Res. **118**(1-2): 1-6.

Xu, S. and A. D. Chisholm (2014). "C. elegans epidermal wounding induces a mitochondrial ROS burst that promotes wound repair." Dev. Cell **31**(1): 48-60.

Yang, Y., A. V. Bazhin, J. Werner and S. Karakhanova (2013). "Reactive oxygen species in the immune system." Int. Rev. Immunol. **32**(3): 249-270.

Yook, K. and J. Hodgkin (2007). "Mos1 mutagenesis reveals a diversity of mechanisms affecting response of Caenorhabditis elegans to the bacterial pathogen Microbacterium nematophilum." Genetics **175**(2): 681-697.

Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich and G. Cesareni (2002). "MINT: a Molecular INTeraction database." FEBS Lett. **513**(1): 135-140.

Zeitlinger, J., A. Stark, M. Kellis, J.-W. Hong, S. Nechaev, K. Adelman, M. Levine and R. A. Young (2007). "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." Nat. Genet. **39**(12): 1512-1516.

Zhang, S.-M., C. M. Adema, T. B. Kepler and E. S. Loker (2004). "Diversification of Ig superfamily genes in an invertebrate." Science **305**(5681): 251-254.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based Analysis of ChIP-Seq (MACS)." Genome Biol. **9**(9): 1-9.

Zhu, C., K. J. R. P. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer and M. L. Bulyk (2009). "High-resolution DNA-binding specificity analysis of yeast transcription factors." Genome Res. **19**(4): 556-566.

Zhu, L. J., C. Gazin, N. D. Lawson, H. Pagès, S. M. Lin, D. S. Lapointe and M. R. Green (2010). "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data." BMC Bioinformatics **11**: 237.

Ziegler, K., C. L. Kurz, S. Cypowyj, C. Couillault, M. Pophillat, N. Pujol and J. J. Ewbank (2009). "Antifungal innate immunity in C. elegans: PKCdelta links G protein signaling and a conserved p38 MAPK cascade." Cell Host Microbe **5**(4): 341-352.

Zou, C.-G., Q. Tu, J. Niu, X.-L. Ji and K.-Q. Zhang (2013). "The DAF-16/FOXO transcription factor functions as a regulator of epidermal innate immunity." PLoS Pathog. **9**(10): e1003660.

Zugasti, O., N. Bose, B. Squiban, J. Belougne, C. L. Kurz, F. C. Schroeder, N. Pujol and J. J. Ewbank (2014). "Activation of a G protein-coupled receptor by its endogenous ligand triggers the innate immune response of Caenorhabditis elegans." Nat. Immunol.

Zugasti, O. and J. J. Ewbank (2009). "Neuroimmune regulation of antimicrobial peptide expression by a noncanonical TGF-beta signaling pathway in Caenorhabditis elegans epidermis." Nat. Immunol. **10**(3): 249-256.

Zugasti, O., N. Thakur, J. Belougne, B. Squiban, C. L. Kurz, J. Soulé, S. Omi, L. Tichit, N. Pujol and J. J. Ewbank (2016). "A quantitative genome-wide RNAi screen in C. elegans for antifungal innate immunity genes." BMC Biol. **14**(1): 35.