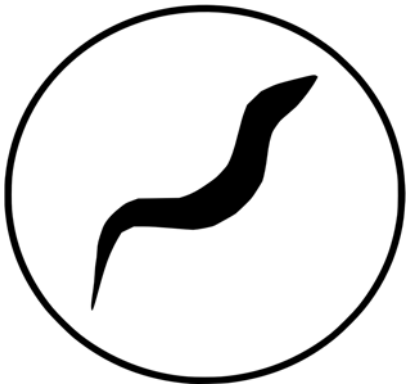
Omics analysis in *Caenorhabditis elegans*: pattern inference and interpretation



Dissertation

in fulfillment of the requirements for the degree

Doctor rerum naturalium

of the Faculty of Mathematics and Natural Sciences

at the University of Kiel

Submitted by Wentao Yang

Department of Evolutionary Ecology and Genetics

Zoological Institute, Kiel University

Kiel, 2017

First referee: Prof. Dr. Hinrich Schulenburg

Second referee: Prof. Dr. Philip Rosenstiel

Date of oral examination: 08.03.2017

Approved for publication: 08.03.2017

Signature:_____

Contents

Declaration

I, **Wentao Yang**, declare that:

Apart from my supervisor's guidance the content and design of the thesis is all my own work;

Specific aspects of my thesis were supported by colleagues; their contribution is specified in detail in the following section "Authors' Contributions";

The thesis has not already been submitted neither partially nor wholly as part of a doctoral degree to another examining body. Apart from the included unpublished paper (Chapter II, aFold: a new method to infer fold change and differential gene expression from RNA-Seq data) other parts of the thesis have been published;

The thesis has been prepared subject to the Rules of Good Scientific Practice of the German Research Foundation (DFG).

Signature:_____

Authors' Contributions

This PhD thesis consists of eight chapters, each represented by a publication, or unpublished manuscript. Wentao Yang developed original ideas and wrote the manuscripts with major contribution for Chapter I, II, III, IV and VI, and on collaborative basis for Chapter V and VII.

Chapter I

ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences. *BMC Genomics* (2016) 17:541

Wentao Yang, Philip Rosenstiel, Hinrich Schulenburg

WY had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; PR contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript.

Chapter II

aFold: a new method to infer fold change and differential gene expression from RNA-Seq data. *Unpublished manuscript.*

Wentao Yang, Philip Rosenstiel, Hinrich Schulenburg

WY had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; PR contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript.

Chapter III

WormExp: a web-based application for Gene Set Enrichment Analysis on *Caenorhabditis elegans.* *Bioinformatics* (2016) 32:943-945

Wentao Yang, Katja Dierking and Hinrich Schulenburg

WY had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; KD contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript

---

Chapter IV

Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* towards pathogenic *Bacillus thuringiensis*. *Dev. Comp. Immunol*. (2015) 51:1-9

Wentao Yang, Katja Dierking, Daniela Esser, Andreas Tholey, Matthias Leippe, Philip Rosenstiel and Hinrich Schulenburg

WY and KD had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; DE, AT, ML and PR contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript

---

Chapter V

Contrasting invertebrate immune defense behaviors caused by a single gene, the *Caenorhabditis elegans* neuropeptide receptor gene *npr-1*. *BMC Genomics* (2016) 17:280

Rania Nakad, Basten Snoek, Wentao Yang, Sunna Ellendt, Franziska Schneider, Timm Mohr, Lone Rösingh, Anna Masche, Philip Rosenstiel, Katja Dierking, Jan Kammenga and Hinrich Schulenburg

RN conceived the study, performed and participated in all experiments, analyzed the data, and drafted the manuscript. LBS conceived and performed QTL analysis and drafted the manuscript. WY performed transcriptomic analysis and drafted the manuscript. SE, FS, TGM, LR, ACM, KD helped

performing phenotypic and functional genetic experiments. PCR conceived and participated in the transcriptomic analysis. JEK provided the RIL and IL libraries, conceived the QTL analysis, and drafted the manuscript. HS conceived the study, analyzed the data, and drafted the manuscript.

Chapter VI

GATA transcription factor as a likely key regulator of the *Caenorhabditis elegans* innate immune response against gut pathogens. *Zoology* (2016) 119:244-253

Wentao Yang, Katja Dierking, Philip Rosenstiel and Hinrich Schulenburg

WY and KD had the initial idea to the approach, designed the study and wrote the manuscript. WY performed analyses. PR contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript

---

Chapter VII

Antimicrobial effectors in the nematode *C. elegans* – an outgroup to the Arthropoda. *Phil Trans R Soc Lond B.* (2016) 371: 20150299

Katja Dierking, Wentao Yang and Hinrich Schulenburg

KD had the initial idea to the approach, designed the study and wrote the manuscript. WY performed analyses and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript

---

Hiermit bestätige ich als Betreuer die oben stehenden Angaben.

Summary

High-throughput molecular technologies have greatly enhanced our understanding of biological processes by characterizing expression changes of genes (microarray and RNA-Seq data) and proteins (proteomics data), or transcription factor targets and epigenetics states (ChIP-chip and ChIP-Seq data). Among them, transcriptome studies based on microarrays or RNA-Seq have the ability to identify genes involved in the response to environmental change or specific stressors, thereby helping us to infer the underlying biological processes.

During my PhD, I mainly focused on transcriptomic data analysis, using in most cases the nematode *Caenorhabditis elegans* as a model taxon. In particular, I have addressed seven specific projects: i) development of ABSSeq, an improved detection approach of differential gene expression for RNA-Seq data; ii) development of aFold, a method to fully moderate fold-change of RNA-Seq data and to improve gene ranking and visualization; iii) development of WormExp, a knowledge-based approach for interpreting gene sets in *C. elegans*; iv) exploration of the regulation of the *C. elegans* immune system using curated data sets from WormExp; v) characterization of putative major effectors (GATA transcription factors) in the *C. elegans* innate immune system; vi) comparison of the immune response of *C. elegans* at protein and transcript level.

In general, our work facilitates high-throughput data analysis via improving pattern inference and interpretation, which in practice provides new insights into the immune system of *C. elegans*.

Zusammenfassung

Die Charakterisierung der Expressionsänderungen von Genen (Microarray- und RNA-Seq-Daten) und Proteinen (Proteomikdaten) oder Transkriptionsfaktor-Targets und epigenetischen Zuständen (ChIP-Chip und ChIP-Seq-Daten) mittels molekularer Hochdurchsatztechnologien hat unser Verständnis von biologischen Prozessen maßgeblich verbessert. Unter diesen haben Transkriptomstudien, die auf Microarrays oder RNA-Seq basieren, die Fähigkeit, Gene zu identifizieren, die an der Reaktion auf Umweltveränderungen oder spezifische Stressoren beteiligt sind. Dies hilft, auf die zugrundeliegenden biologischen Prozesse schließen zu können.

Während meiner Promotion konzentrierte ich mich vor allem auf die transkriptomische Datenanalyse, wobei in den meisten Fällen der Nematode *Caenorhabditis elegans* als Modell verwendet wurde. Im Speziellen behandelte ich dabei sieben spezifische Projekte: i) Entwicklung von ABSSeq, einem verbesserten Erkennungsansatz der differentiellen Genexpression für RNA-Seq-Daten; ii) Entwicklung von aFold, einer Methode zur vollständigen Normierung des Fold Change von RNA-Seq-Daten und zur Verbesserung des Gen-Rankings und der Visualisierung; iii) Entwicklung von WormExp, einem wissensbasierten Ansatz zur Interpretation von *C. elegans*-Gensätzen; iv) Erforschung der Regulation des *C. elegans*-Immunsystems unter Verwendung von kuratierten Datensätzen von WormExp; v) Charakterisierung von potentiellen Haupteffektoren (GATA-Transkriptionsfaktoren) im angeborenen Immunsystem von *C. elegans*; vi) Vergleich der Immunantwort von *C. elegans* auf Protein- und Transkriptebene.

Zusammenfassend erleichtert unsere Arbeit die Hochdurchsatzdatenanalyse durch die Verbesserung der Musterinferenz und -interpretation, die in der Praxis neue Einblicke in das Immunsystem in *C. elegans* liefert.

Introduction

Pattern inference and interpretation are the major interests for omics studies, which have greatly enhanced our understanding of biological processes. My PhD project focused on transcriptome data analysis in the well-established model organism *C. elegans*, which relies on identifying differentially expressed genes (pattern inference) between different conditions, as well as inferring their potential biological functions (pattern interpretation). In the present work I developed new approaches to detect differential expression (DE) in RNA-Seq datasets and specifically infer underlying biological functions in *C. elegans*. With the help of these new approaches, we addressed the regulation effectors of *C. elegans* innate immune responses. This introduction provides information necessary to understand the present work and includes four major topics: *C. elegans* as a model organism and its immunity system, omics studies in *C. elegans*, DE detection of RNA-Seq data and biological function inference on DE.

1 *C. elegans* – an invertebrate model organism for innate immunity research

The innate immune system serves as the first line to defend animals and plants against pathogenic infections. It shares common features in molecular pathways across vertebrates and invertebrates, such as flies, nematodes and mammals [1]. Because of these evolutionary conserved characteristics, the study of invertebrate host defenses can provide a better understanding of vertebrate innate immunity, including those of relevance for humans. The nematode *C. elegans,* with its completely sequenced genome, genetic tractability, and susceptibility to a number of human and other animal pathogens; is widely used as a powerful invertebrate model organism to study innate immunity [2, 3].

1.1 Known effectors in *C. elegans* immune system

Previous studies revealed that this nematode's immune system relies on

several signaling pathways conserved across invertebrates and vertebrates including: the p38 mitogen-activated protein kinase (MAPK), c-Jun N-terminal kinase (JNK) MAPK, extracellular signal-regulated kinase (ERK) MAPK, transforming growth factor-β (TGF-β), and also the insulin-like receptor (ILR) pathways [4-6]. Several transcription factors have been identified to contribute to *C. elegans* immune defense, such as: the GATA transcription factor ELT-2 [7], the basic-region leucine zipper (bZIP) transcription factors ATF-7 [8], ATFS-1 [9], ZIP-2 [10], and SKN-1 [11], the basic helix-loop-helix (bHLH) transcription factor HLH-30 [12], the signal transducer and activator of transcription (STAT)-like transcription factor STA-2 [13], and the activator protein 1 (AP-1) transcription factor dimer JUN-1/FOS-1 [14]. Moreover, pathogen elimination involves certain antimicrobial effectors [15], including for example the caenacins and related peptides [16], the caenopores [17, 18], and additionally the generation of reactive oxygen species (ROS) [19, 20]. While it remains unclear if and how pathogens are directly recognized by *C. elegans*, nematode defense can also be activated indirectly through a cellular surveillance system and/or damage signals, allowing the worms to respond to the cellular disturbance caused by an infection [21-23].

2 Omics and their application on *C. elegans* research

High-throughput molecular technologies have greatly enhanced our understanding of biological processes by characterizing expression changes of genes (microarray and RNA-Seq data) and proteins (proteomics data) or transcription factor targets and epigenetic states (ChIP-chip and ChIP-Seq data).

Although most studies on *C. elegans* rely on functional genetic approaches, numerous transcriptomic and proteomic analyses have additionally been performed to explore the sets of genes, which are activated or repressed upon specific conditions or life stages. Over the last decade, more than 350 high-throughput expression studies have been published, covering a large variety of research themes, such as immunity, aging, development, and stress responses.

The resulting lists of differentially expressed genes are publicly available and can be related to a specific experimental design, environmental condition, and/or gene defect. Because they capture a variety of inducible expression responses of this particular organism, they might be highly useful in interpreting new *C. elegans* gene lists [24, 25] or predicting candidates for downstream analysis [26].

However, omics studies usually yield hundreds or thousands of differentially regulated genes or proteins that are not always easy to interpret. Validation of the numerous differentially expressed genes is usually not possible. Uncovering the underlying organizational principles from such large gene lists requires computational and statistical approaches as well as precise biological reference information.

2.1 Omics on *C. elegans* immunity

Currently, gene or protein expression change in *C. elegans* upon exposure to 21 pathogens (including gram-positive and negative bacteria, fungi and virus) and six non-pathogenic bacterial strains have been quantified. Several of these studies involved more than one pathogen and showed an overlapping signature in the response to the various pathogens [6, 27], indicating the presence of a common regulatory mechanism in the worm's immune system.

3 RNA-Seq and differential expression inference

3.1 RNA-Seq

RNA Sequencing or RNA-Seq is a recent and popular technology for transcriptome studies, which is based on next generation sequencing. In contrast to array-based approaches, RNA-Seq could quantify genome-wide gene expression without genome annotation and thus is widely used to study both model and non-model organisms [28]. The underlying aim of transcriptome or RNA-Seq studies is to understand inducible biological functions through an analysis of differential gene expression (DE), which is usually inferred from comparison of two different treatments, life stages or tissues, among other

conditions that can be compared.

Read count of RNA-Seq data requires normalization before DE inference in order to reduce possible biases from variation in sequencing depth, library preparation, sequencing in different lanes or other random factors [29, 30].

3.2 Differential expression analysis on RNA-Seq

Current statistical approaches for DE analysis in RNA-Seq rely on fitting the distribution of read counts with probabilistic models [31-35]. These methods usually detect DE via false discovery rate (FDR) adjusted p-value, which highly depends on mean-variance relationship [31, 36, 37]. However, variance could be arbitrarily small or even zero (under-estimated) even after borrowing information from other genes, which often results in highly statistically significant DE [38, 39] as well as high type I error rate and FDR, but with extremely small fold-change [40-42]. There is therefore often a clear-cut inconsistency between the statistical result and the preferred cut-off value in practice. Efficiently reducing such potential artifacts requires additional cutoffs in fold change [40-42], which somehow agrees with the interests of biologists in term of worthwhile change, for instance a fold change of at least 1.5 or 2.0 [43-45]. Moreover, false positives of DE are also commonly present in genes with high coefficient of variation (usually at low expression level), therefore a third cutoff of expression value seems also necessary [40-42]. Even though these sources of problems in DE estimation are clear, there is still the need to develop strategies to avoid introducing the new problem of an arbitrary choice of the fold-change cut-off and expression value cut-off.

4. Gene set enrichment analysis – inferring biological function

Gene set enrichment analysis represents a powerful tool to link the identified differentially expressed gene lists to biological processes and functions. They are based on the statistical evaluation of the overlap between the generated gene set and a specified reference list of genes. These enrichment analyses are usually based on public databases such as those defined by Gene Ontology

[46] and KEGG pathways [47]. However, these existing databases have important drawbacks. First, the annotations are incomplete and only a subset of known genes are functionally annotated [48]. For example, functional information is only available for approximately 60% of the gene repertoire of the nematode *C. elegans* [49]. Second, the included functional information is often imprecise, as it usually represents an extrapolation from experimental data of a different taxon and thus assumes a high level of functional conservation across evolution, which may not always be the case [50]. Third, functional information is predicted for most organisms from protein domains. Taxon-specific genes or protein domains may thus be missed. Taxon-specific gene sets, which explicitly consider taxon-restricted genes and also taxon-specific expression responses, are thus required for improved functional genomic analyses. Several applications such as GSEA [51] and EASE [52] have been developed to permit performance of enrichment analyses with curated gene sets, derived for example from published expression studies in the same organism. Yet, a systematic assessment of the value of taxon-specific enrichment analyses is still missing.

In general, omics studies produce numerous data that retains noise and thus requires efficient statistical tools to explore the underlying biological signatures and functions. During the PhD period, I developed approaches to improve DE detection of RNA-Seq and biological function inference upon DE, which has been applied on the study of *C. elegans* immunity system. This thesis summarizes these works in seven Chapters.

PhD Thesis Content

The aim of the PhD project was to improve pattern inference and interpretation for omics study. The current PhD thesis embraces three aspects: i) pattern inference (DE detection); ii) pattern interpretation (gene set enrichment analysis); iii) their application on *C. elegans* immune response. I use seven chapters to describe and discuss these three aspects.

Chapter I represents a published approach for DE detection on RNA-Seq data with title “ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences”. Here we introduce a new analysis approach, ABSSeq, which uses a negative binomal distribution to model absolute expression differences between conditions, taking into account variations across genes and samples as well as magnitude of differences. In comparison to alternative methods, ABSSeq shows higher performance on controlling type I error rate and at least a similar ability to correctly identify differentially expressed genes. This chapter refers to pattern inference.

Chapter II is a manuscript that is ready for submission. Here, we introduce a new approach, aFold (i.e., accurately estimation of fold change from RNA-Seq data), which provides a statistical framework to solve the problem of an arbitrary choice of cut-off values by integrating all sources of variation into fold change calculation. This approach models the uncertainty of read count via a polynomial function of sample mean and standard deviation. aFold also provide an efficient strategy for determining cutoff of fold change across all significant levels. Instead of modelling read count distribution, aFold employs a zero-centered normal distribution on testing shifted log fold changes against a global standard deviation, which therefore avoid the influence of extremely small variance. This chapter refers to pattern inference.

Chapter III is a published approach for gene set enrichment analysis specifically on *C. elegan* with title “WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis”. We here present a web-based application for a taxon-specific gene set exploration and enrichment analysis, which is expected to yield novel functional insights into newly determined gene sets. The approach is based on the complete collection of curated high-throughput gene expression data sets for the model nematode *C. elegans*, including 1980 gene sets from more than 350 studies. This chapter refers to pattern interpretation.

Chapter IV is a published study on *C. elegans* immune response to *Bacillus*

thuringiensis with title “Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* towards pathogenic *Bacillus thuringiensis*”. Here, we compare transcriptome and proteome data generated after infection of the nematode and model organism *C. elegans* with the Gram-positive pathogen *B. thuringiensis*. Our analysis revealed a high overlap between abundance changes of corresponding transcripts and gene products, especially for genes encoding C-type lectin domain-containing proteins, indicating their particular role in worm immunity. We additionally identified a unique signature at the proteome level, suggesting that the *C. elegans* response to infection is shaped by changes beyond transcription. Such effects appear to be influenced by AMP-activated protein kinases (AMPKs), which may thus represent previously unknown regulators of *C. elegans* immune defense. This chapter refers to application of pattern inference and interpretation.

Chapter V is a published study on *C. elegans* immune response to *B. thuringiensis* and *Pseudomonas aeruginosa*. with title “Contrasting invertebrate immune defense behaviors caused by a single gene, the *Caenorhabditis elegans* neuropeptide receptor gene *npr-1*”. Here, we demonstrate in the model invertebrate *C. elegans* that a single gene, a homolog of the mammalian neuropeptide Y receptor gene, *npr-1*, mediates contrasting defense phenotypes towards two distinct pathogens, the Gram-positive *B. thuringiensis* and the Gram-negative *P. aeruginosa*. Subsequent transcriptional profiling of *C. elegans* wildtype and *npr-1* mutant suggested that *npr-1* mediates defense against both pathogens through *p38* MAPK signaling, insulin-like signaling, and C-type lectins. Importantly, increased defense towards *P. aeruginosa* seems to be additionally influenced through the induction of oxidative stress genes and activation of GATA transcription factors, while the repression of oxidative stress genes combined with activation of Ebox transcription factors appears to enhance susceptibility to *B. thuringiensis*. This chapter refers to application of pattern inference and interpretation.

Chapter VI is a published study of meta-analysis on *C. elegans* immunity system with title “GATA transcription factor as a likely key regulator of the *C. elegans* innate immune response against gut pathogens”. In this study, we take advantage of WormExp in order to explore commonalities and differences in the regulation of nematode immune defense against a large variety of pathogens versus food microbes. We identified significant overlaps in the transcriptional response towards microbes, especially pathogenic bacteria. We also found that the GATA motif is overrepresented in many microbe-induced gene sets and in targets of other previously identified regulators of worm immunity. This chapter refers to application of pattern inference and interpretation.

Chapter VII is a review on *C. elegans* immunity system with title “Antimicrobial effectors in the nematode *C. elegans* – an outgroup to the Arthropoda”. In this review, we discuss putative *C. elegans* antimicrobial effector proteins, such as lysozymes, caenopores (or saposin-like proteins), defensin-like peptides, caenacins and neuropeptide-like proteins, in addition to the production of reactive oxygen species and autophagy. We provide an overview of *C. elegans* immune effector proteins and mechanisms. We summarize the experimental evidence of their antimicrobial function and involvement in the response to pathogen infection. We further evaluate the microbe-induced expression of effector genes using WormExp (Chapter III). We emphasize the need for further analysis at the protein level to demonstrate an antimicrobial activity of these molecules both in vitro and in vivo. This chapter refers to application of pattern inference and interpretation.

References - Introduction

1. Ausubel, F.M., *Are innate immune signaling pathways in plants and animals conserved?* Nature immunology, 2005. **6**(10): p. 973-979.

2. Kurz, C.L. and J.J. Ewbank, *Caenorhabditis elegans: an emerging genetic model for the study of innate immunity.* Nature Reviews Genetics, 2003. **4**(5): p. 380-390.

3. Marsh, E.K. and R.C. May, *Caenorhabditis elegans, a model organism for investigating immunity.* Applied and environmental microbiology, 2012. **78**(7): p. 2075-2081.

4. Pukkila-Worley, R. and F.M. Ausubel, *Immune defense mechanisms in the Caenorhabditis elegans intestinal epithelium.* Current opinion in immunology, 2012. **24**(1): p. 3-9.

5. Engelmann, I. and N. Pujol, *Innate immunity in C. elegans*, in *Invertebrate Immunity*. 2010, Springer. p. 105-121.

6. Irazoqui, J.E., et al., *Distinct pathogenesis and host responses during infection of C. elegans by P. aeruginosa and S. aureus.* PLoS Pathog, 2010. **6**(7): p. e1000982.

7. Shapira, M., et al., *A conserved role for a GATA transcription factor in regulating epithelial innate immune responses.* Proceedings of the National Academy of Sciences, 2006. **103**(38): p. 14086-14091.

8. Shivers, R.P., et al., *Phosphorylation of the conserved transcription factor ATF-7 by PMK-1 p38 MAPK regulates innate immunity in Caenorhabditis elegans.* PLoS Genet, 2010. **6**(4): p. e1000892.

9. Pellegrino, M.W., et al., *Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection.* Nature, 2014.

10. Estes, K.A., et al., *bZIP transcription factor zip-2 mediates an early response to Pseudomonas aeruginosa infection in Caenorhabditis elegans.* Proceedings of the National Academy of Sciences, 2010. **107**(5): p. 2153-2158.

11. Papp, D., P. Csermely, and C. Soti, *A role for SKN-1/Nrf in pathogen resistance and immunosenescence in Caenorhabditis elegans.* PLoS Pathog, 2012. **8**(4): p. e1002673.

12. Visvikis, O., et al., *Innate host defense requires TFEB-mediated transcription of cytoprotective and antimicrobial genes.* Immunity, 2014. **40**(6): p. 896-909.

13. Dierking, K., 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): p. 425-435.

14. Kao, C.-Y., et al., *Global functional analyses of cellular responses to pore-forming toxins.* PLoS Pathog, 2011. **7**(3): p. e1001314.

15. Dierking, K., W. Yang, and H. Schulenburg, *Antimicrobial effectors in the nematode C. elegans–an outgroup to the Arthropoda.* Phil Trans R Soc Lond B., 2016. **in press**.

16. Couillault, C., et al., *TLR-independent control of innate immunity in Caenorhabditis elegans by the TIR domain adaptor protein TIR-1, an ortholog of human SARM.* Nature immunology, 2004. **5**(5): p. 488-494.

17. Roeder, T., et al., *Caenopores are antimicrobial peptides in the nematode Caenorhabditis elegans instrumental in nutrition and immunity.* Developmental & Comparative Immunology, 2010. **34**(2): p. 203-209.

18. Mysliwy, J., et al., *Caenopore-5: The three-dimensional structure of an antimicrobial protein from Caenorhabditis elegans.* Developmental & Comparative Immunology, 2010. **34**(3): p. 323-330.

19. Chávez, V., A. Mohri-Shiomi, and D.A. Garsin, *Ce-Duox1/BLI-3 generates reactive oxygen species as a protective innate immune mechanism in Caenorhabditis elegans.* Infection and immunity, 2009. **77**(11): p. 4983-4989.

20. Van Der Hoeven, R., et al., *Ce-Duox1/BLI-3 generated reactive oxygen species trigger protective SKN-1 activity via p38 MAPK signaling during infection in C. elegans.* PLoS Pathog, 2011. **7**(12): p. e1002453-e1002453.

21. Melo, J.A. and G. Ruvkun, *Inactivation of conserved C. elegans genes engages pathogen-and xenobiotic-associated defenses.* Cell, 2012. **149**(2): p. 452-466.

22. Ewbank, J.J. and N. Pujol, *Local and long-range activation of innate immunity by infection and damage in C. elegans.* Current opinion in immunology, 2016. **38**: p. 1-7.

23. Zugasti, O., et al., *Activation of a G protein-coupled receptor by its endogenous ligand triggers the innate immune response of Caenorhabditis elegans.* Nature immunology, 2014. **15**(9): p. 833-838.

24. Yang, W., et al., *Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode Caenorhabditis elegans toward pathogenic Bacillus thuringiensis.* Developmental & Comparative Immunology, 2015. **51**(1): p. 1-9.

25. Engelmann, I., et al., *A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans.* PloS one, 2011. **6**(5): p. e19055.

26. Block, D.H., et al., *The Developmental Intestinal Regulator ELT-2 Controls p38-Dependent Immune Responses in Adult C. elegans.* 2015.

27. Wong, D., et al., *Genome-wide investigation reveals pathogen-specific and shared signatures in the response of Caenorhabditis elegans to infection.* Genome Biol, 2007. **8**(9): p. R194.

28. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews genetics, 2009. **10**(1): p. 57-63.

29. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome biology, 2010. **11**(3): p. 1.

30. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome biology, 2010. **11**(12): p. 1.

31. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

32. Hardcastle, T.J. and K.A. Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 2010. **11**(1): p. 422.

33. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

34. Li, J., et al., *Normalization, testing, and false discovery rate estimation for RNA-sequencing data.* Biostatistics, 2012. **13**(3): p. 523-538.

35. Srivastava, S. and L. Chen, *A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.* Nucleic acids research, 2010. **38**(17): p. e170-e170.

36. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 1.

37. Law, C.W., et al., *Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome biology, 2014. **15**(2): p. 1.

38. Feng, J., et al., *GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data.* Bioinformatics, 2012. **28**(21): p. 2782-2788.

39. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics, 2013. **14**(1): p. 91.

40. Li, S., et al., *Detecting and correcting systematic variation in large-scale RNA sequencing data.* Nature biotechnology, 2014. **32**(9): p. 888-895.

41. Li, S., et al., *Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study.* Nature biotechnology, 2014. **32**(9): p. 915-925.

42. Yang, W., P.C. Rosenstiel, and H. Schulenburg, *ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences.* BMC Genomics, 2016. **17**: p. 541.

43. McCarthy, D.J. and G.K. Smyth, *Testing significance relative to a fold-change threshold is a TREAT.* Bioinformatics, 2009. **25**(6): p. 765-771.

44. Patterson, T.A., et al., *Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project.* Nature biotechnology, 2006. **24**(9): p. 1140-1150.

45. DeRisi, J., et al., *Use of a cDNA microarray to analyse gene expression patterns in human cancer.* Nature genetics, 1996. **14**(4): p. 457-460.

46. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.

47. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

48. King, O.D., et al., *Predicting gene function from patterns of annotation.* Genome research, 2003. **13**(5): p. 896-904.

49. Petersen, C., P. Dirksen, and H. Schulenburg, *Why we need more ecology for genetic models such as C. elegans.* Trends in Genetics, 2015. **31**(3): p. 120-127.

50. Khatri, P. and S. Drăghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. **21**(18): p. 3587-3595.

51. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

52. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome Biol, 2003. **4**(10): p. R70.

Chapter I

ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences

Wentao Yang[a,*], Philip C. Rosenstiel[b], Hinrich Schulenburg[a,*]

Affiliations

[a] Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany

[b] Centre for Molecular Biology, Institute for Clinical Molecular Biology, CAU Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany

* Correspondence: wyang@zoologie.uni-kiel.de, hschulenburg@zoologie.uni-kiel.de

Yang et al. BMC Genomics (2016) 17:541
DOI 10.1186/s12864-016-2848-2

BMC Genomics

SOFTWARE Open Access

ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences

Wentao Yang[1*], Philip C. Rosenstiel[2] and Hinrich Schulenburg[1*]

Abstract

Background: The recent advances in next generation sequencing technology have made the sequencing of RNA (i.e., RNA-Seq) an extemely popular approach for gene expression analysis. Identification of significant differential expression represents a crucial initial step in these analyses, on which most subsequent inferences of biological functions are built. Yet, for identification of these subsequently analysed genes, most studies use an additional minimal threshold of differential expression that is not captured by the applied statistical procedures.

Results: Here we introduce a new analysis approach, ABSSeq, which uses a negative binomal distribution to model absolute expression differences between conditions, taking into account variations across genes and samples as well as magnitude of differences. In comparison to alternative methods, ABSSeq shows higher performance on controling type I error rate and at least a similar ability to correctly identify differentially expressed genes.

Conclusions: ABSSeq specifically considers the overall magnitude of expression differences, which enhances the power in detecting truly differentially expressed genes by reducing false positives at both very low and high expression level. In addition, ABSSeq offers to calculate shrinkage of fold change to facilitate gene ranking and effective outlier detection.

Keywords: RNA-Seq, Transcriptome analysis, Differential gene expression, ABSSeq, Negative binomial distribution

Background

Transcriptome studies usually aim at understanding inducible biological functions through an analysis of differential gene expression (DE). Since relatively recently, the variation in gene expression is commonly studied through RNA sequencing or RNA-Seq, based on next generation sequencing (NGS) technologies. In these study approaches, DE is usually inferred from comparison of two different treatments, developmental stages, or different tissues. A key step in these analyses is the reliable identification of significant DE. Most current statistical approaches employ a probabilistic model, such as the Negative Binomial (NB) [1–3], Poisson [4], the Generalized Poisson (GP) model [5], and use information on gene expression variation in the data to account for ambiguity caused by sample size, biological and technical biases, overall levels of expression and the presence of outliers. DE inference is usually based on the null hypothesis that the means of read counts among conditions are the same or follow the same distribution. These tests neglect the magnitude of encountered differences and might report statistically highly significant DE with arbitrarily small fold change, at least if the number of sequencing counts is large enough [6, 7]. However, small fold changes may represent artifacts and often cannot be validated experimentally (e.g., through Realtime PCR approaches or functional genetic analysis). Thus, they might not be worth further investigation. A currently common solution is sought by combining the statistical indication (i.e., an FDR-adjusted p-value) with a specified minimum fold change [8, 9]. This approach has the possible problem of a high number of identified candidate genes with low count numbers (which may produce high fold change by chance) and its dependence on an arbitrarily chosen fold-change cut-off value.

* Correspondence: wyang@zoologie.uni-kiel.de; hschulenburg@zoologie.uni-kiel.de
[1]Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany
Full list of author information is available at the end of the article

An alternative approach has so far only been established for ChIP-Seq data and relies on an analysis of count differences between test and reference conditions [10, 11]. In this case, the statistics are based on a measure that considers the magnitude of count differences and the level of expression variation across replicates with the effect that genes with only minor expression levels and only small fold change are selected against. In consideration of such potential advantages, such an approach may prove useful for reliable DE identification in RNA-Seq data.

Here, we introduce ABSSeq (i.e., differential expression analysis of ABSolute differences of RNA-Seq data), which employs an NB distribution to model count differences between conditions. It permits testing the magnitude of observed count differences taking into consideration background expression level variation. In particular, ABSSeq accounts for heterogeneous dispersions in expression level across genes by adding expected values (pseudocounts) to reads count according to the smoothed mean-variance relationship [1], which thus adjusts parameters in the NB distribution (mean and size). In addition, ABSSeq imposes a penalty on the dispersion estimation, it uses a new outlier detection strategy, and it also inroduces a procedure for shrinkage of fold change to disfavor identification of candidate genes with abnormal high dispersions and extremely low expression. Using real and simulated datasets, we demonstrate that our method is highly efficient in reducing the false discovery rate (FDR) and thus in identifying truly differentially expressed genes in RNA-Seq data. It therefore shows an at least similar performance than several frequently used, alternative approaches like those implemented in the software packages DESeq [1], DESeq2 [1, 12], edgeR [3, 13] (referred as edgeR-robust when applied on data set with outliers), limma [14, 15] (referred to as Voom), baySeq [2], and EBSeq [16].

Implementation

ABSSeq has been implemented in the software package ABSSeq for the cross-platform environment R [17]. ABSSeq is released under the GPL-3 license as part of the Bioconductor project [18] at URL: http://bioconductor.org/packages/devel/bioc/html/ABSSeq.html.

Results and discussion

We firstly introduce our approach with the help of the modencodefly and ABRF datasets (see Datasets). Thereafter, performance of our method is compared with that of several previously developed and currently popular methods (always used under default settings, for example limma under eBayes settings and the TMM normalization for limma and edgeR; see Additional file 1), including one, EBSeq, which allows to evaluate DE at both transcript and also gene level [16]. We exlcude Cuffdiff2 [19] from our assessment because it was previously compared with the other available approaches and generally found to produce higher rates of false positives without an increase in sensitivity [20]. Method evaluation is based on two types of data sets. On the one hand, we use simulated data, for which data structure can be efficiently controlled and which have been widely used to evalute methods of differential expression analysis [2, 7, 21–24]. We use the same strategy and identical simulated data sets as Soneson et al. [7] and compare method performance according to two criteria: (i) the ability to control type I error rates; and (ii) the ability to rank truly DE genes ahead of non-DE ones. On the other hand, we also evaluate our approach with the help of real data sets, as described in more detail below.

Control of type I error rate

Minimizing the type I error rate (i.e., the null hypothesis is falsely rejected) or false positive rate is a primary goal of differential expression analysis [20, 25]. Type I error is often introduced by under-estimation of disperson in RNA-Seq data and occurs at genes with very low or high counts [20]. We thus compare the ability of the alternative approaches to control type I error rates, using two real data sets and also the simulated data sets from Soneson et al. [7]. DE genes are defined by a p-value cutoff of 0.05 for each method except baySeq and EBSeq, which are excluded from this comparison since they report DE by posterior probalilities instead of a p-value. The simulated datasets are assumed to lack DE genes, facilitating computation of the type I error rate by dividing the number of DE genes identified by each method with the total number of genes. Figure 1 summarizes the results from the modencodefly data set (Fig. 1a and 1b) and two different simulation settings (Fig. 1c), including data sets of various replicate sample sizes and, in each case, ten independent repetitions (see also Additional file 2). Additional file 3 shows the results for the ABRF data set.

The first comparison is based on a real data set for the fruitfly *Drosophila melanogaster*, the modencodefly data set [26], which characterizes the developmental transcriptome across 30 distinct stages (conditions) with technical replicates ranging from 4 to 6. We randomly select 4 replicates for each condition and separate them into two groups, which should thus only be characterized by stochastic variations but not true DE. The results of our analysis is summarized in Fig. 1a. At the p-value cutoff of 0.05, ABSSeq identifies an average of 17 DE genes and thus significantly fewer DE genes than all alternative methods (Wilcoxon rank test, $p < 0.01$). DESeq2 also performs well on this real data set, while

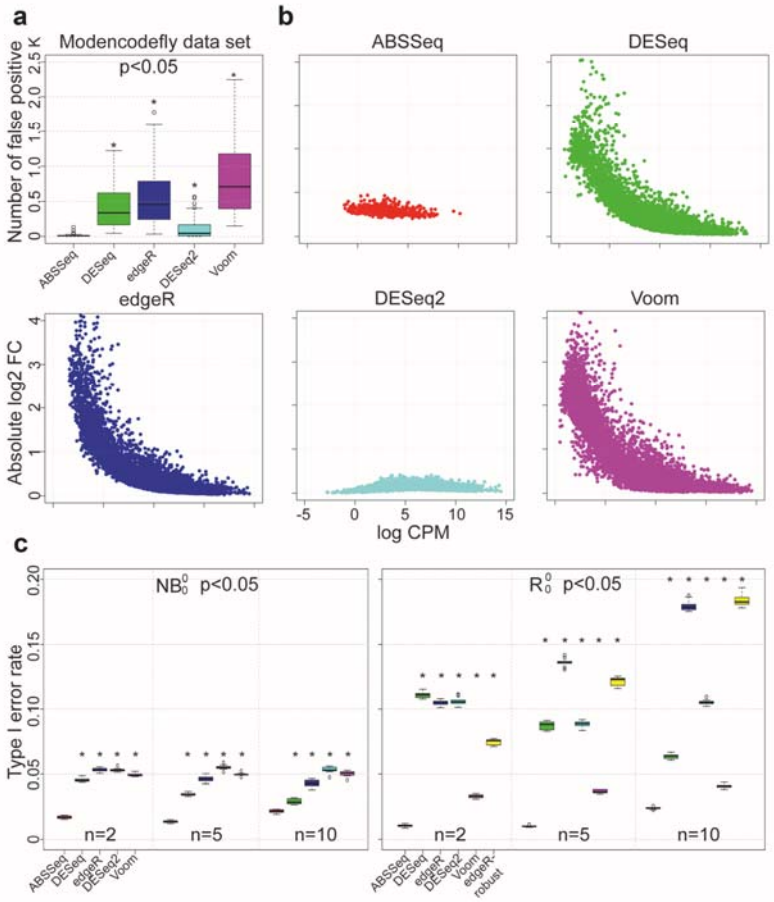


Fig. 1 Method-dependent variation in type I error. Type I error rates for ABSSeq and five alternative methods using the modencodefly real data set (**a**) and two simulation settings: Negative Binomial (NB, c left panel), and NB with random outliers (R, c right panel). **b** Points show the absolute log fold change (FC, y-axis) distribution of false positives against the expression level (logCPM, x-axis). **c** Each boxplot summarizes the type I error rates across 10 independent simulated data sets. Asterisk indicates a statistically significant difference in type I error between ABSSeq and any of the other methods. n indicates the number of RNA replicates considered in each case (ranging from 2 to 10). Under all conditions, ABSSeq reduced the type I error rate

the highest type I error rate is obtained for limma (873 identified cases of DE).

Next, we examined the distribution of false positives along absolute log2 fold change and expression level (log of average counts per million, logCPM, calculated by edgeR) using the data from Fig. 1a. As shown in Fig. 1b, false positives with low logCPM (x-axis) tend to have a high fold change (DESeq, edgeR and Voom), and vice versa. This skewed distribution is very similar to the quadratic mean-variance relationship [1, 5, 15], suggesting there might be a general under-estimation of variance or dispersion for these methods. In contrast, ABSSeq and DESeq2 both shrink the fold change according to variance. As a consequnce, they both exhibit a pronnounced reduction of false positives at low expression (logCPM < 0). Morever, ABSSeq also reduces false positives at high expression level (logCPM > 10), which likely have a very low smoothed dispersion [1] and are often inferred to be highly statistically significant but show only very small fold change.

As the modencodefly data set only allows us to consider two replicates per group and condition, the resulting statistical power may be limited. Therefore, we repeated this assessment using another real data set, the ABRF data set [27] (see Datasets in Methods section), which is based on an RNA-Seq analysis of the same two samples across three independent laboraties and thus comprises for each sample three replicates that should

only show variation caused by differences among the laboratories such as library preparation methods or sample processing procedures (6 comparisons in total, Additional file 3). The analysis of the ABFR data set confirms the previous results. It demonstrates that ABSSeq produces the smallest number of false positives and especially reduces fold change for the genes with generally low expression.

Overall, the results from the two real data sets suggest that ABSSeq has the ability to handle very small expression changes by considering the magnitude of absolute differences and penalizing the estimated dispersion (See Methods). Our results also suggest that the alternative methods should allow enhanced reduction of the type I error rate if combined with additional filtering approaches, such as usage of a fold-change cut-off as discussed in [28, 29] and also further below.

In addition to the two real data sets, we also compare the ability of the alternative approaches to control type I error rates on simulated data (Fig. 1c). Generally, all methods are able to control type I error rate under 0.05 when applied on the NB distributed data (Fig. 1c left panel, denoted NB_0^0, 0 indicates the number of up or down-regulated genes) but exhibit high diversity on the NB distributed data with randomly introduced outliers (abnormally high counts, multiplying a randomly generated factor between 5 and 10 with counts of genes randomly selected with a probability of 0.05, denoted by S_0^0, Fig. 1c right panel). As already highlighted in [7], DESeq has excellent power to control type I error rates on NB_0^0. The performance of Voom is relatively unaffected by sample size and outliers, implying advantages of log-transformation on dealing with high value outliers. In contrast, edgeR does not control type I error rates efficiently when applied on data with outliers. Since both DESeq2 and edgeR-robust integrate strategies to handle outliers, they expectedly reduce the type I error rate on S0 0, especially when compared to the earlier program versions (e.g., DESeq at $n = 10$ or edgeR at $n = 2$ or 5). ABSSeq performs best in both cases (Tukey's, $p < 1.0e\text{-}3$), but slightly decreases its performance with increasing sample size ($n = 10$).

Taken together, ABSSeq is able to efficiently control type I error rates for the real and simulated data sets (Fig. 1a-c) and it also reduces type I error rate at both low and high expression levels. In addition, outliers impact the ability of controlling type I error rate for most methods except ABSSeq and Voom, which might be caused by shrinkage of the observed dispersion (edgeR, edgeR-robust and DESeq2) or replacing the observed with a smoothed dispersion (DESeq). In contrast, ABSSeq uses the observed dispersion directly, apparently enhancing control of type I errors to a rate of below 0.05.

Discrimination of DE versus non-DE genes in simulation studies

An ideal DE inference method should be more sensitive to DE than non-DE genes, that is, it should be able to discriminate true DE genes against non-DE ones. Here, we evaluate the discriminative power of ABSSeq and other selected methods in terms of the true and false positive rates and also the area under Receiver Operating Characteristic (ROC) curve (AUC), using again the simulated data and general approach of Soneson et al. [7]. The AUC was shown repeatedly to be informative as a measure of the overall discriminative performance of a method [30–32]. In particular, for our comparison, we extract a set of genes from the simulated data set using a given p-value or posterior probability (baySeq) threshold. Thereafter, the obtained genes are divided into a truly positive group and a truly negative group according to pre-defined DE genes in the simulated data. This information then allows us to calculate the true positive and the false positive rate for all possible thresholds, construct ROC curves and compute AUCs using the ROC package in Bioconductor [18]. For all simulations, we choose 10 % of the 12,500 genes as DE and symmetrically divide them into up- and down-regulated genes (e.g., 625 up- and 625 down-regulated genes, indicated below by super- and subscripts, respectively). We summarize the results using boxplots for four different simulation settings, including data sets with various replicate sample sizes and, in each case, ten independent repetitions (Fig. 2, Additional file 2).

When applied on the data set simulated using the NB distribution (denoted by NB_{625}^{625}, where the super- and subscripts indicate the number of up- and down-regulated genes, respectively; Fig. 2 left), ABSSeq always performs at least as good as the alternative methods at the considered replicate sample sizes (denoted by n). EBSeq performs worse than the other approaches when applied on data with small sample size ($n = 2$). The performance of ABSSeq and the other methods are generally improved as the sample size increases, revealing a positive power of sample size on identifying true DEs. Overall, these results suggest that our NB model fits the over-dispersion data at least as well as the NB model implemented in other methods.

We next test the influence of outliers, which we introduce into the NB distributed data using a similar approach as above (denoted by R_{625}^{625}, Fig. 2 right) and which may show abnormally high counts, resulting in high fold changes and also false postives. For these simulated data sets, ABSSeq shows an advantage (Tukey's, $p < 0.01$) at all replicate sizes, especially for the R_{625}^{625} data set (Tukey's, $p < 1.0e\text{-}6$) whose AUC is even greater than 0.9 at $n = 10$ (Fig. 2c, 2d). This result indicates that ABSSeq outlier detection is efficient. Interestingly, performance of the

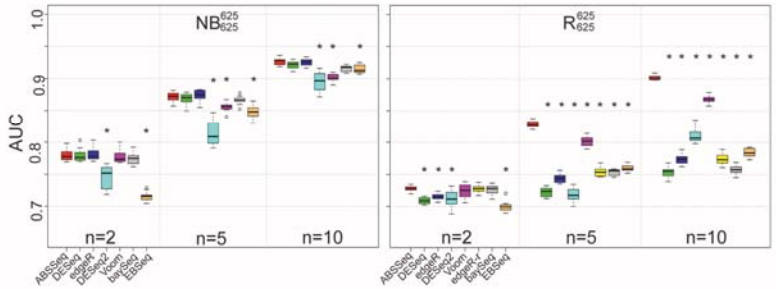


Fig. 2 AUC comparison on simulated data. Area under the curve (AUC) for ABSSeq and five alternative methods under two simulation settings: Negative Binomial (NB, *left panel*) and NB with random outliers (R, *right panel*). Each boxplot summarizes the AUCs across 10 independently simulated data sets. Asterisk indicates a statistically significant difference in AUC between ABSSeq and any of the other methods. n indicates the number of considered RNA-Seq replicates, from 2 to 10. Under all conditions, ABSSeq is highly effective in correctly identifying differentially expressed genes

alternative methods also shows substantial variability. For example, Voom generally performs better at large sample size (i.e. higher AUC in R_{625}^{625} except ABSSeq), but similar at small sample size with other methods; DESeq2 performs better at $n = 10$ due to outlier detection but worse at $n = 2$ and $n = 5$ ($n \geq 7$ required for outlier detection); baySeq shows little improvement in performance as the sample size increases for the R_{625}^{625} data set; EBSeq shows lowest AUC at $n = 2$ and improves performance at large sample sizes ($n = 5$ or 10); edgeR-robust (denoted by edgeR-r) shows an improved ability to handle outliers at small sample size ($n = 2$ or 5) compared to edgeR.

Overall, ABSSeq is at least as good as alternative methods in discriminating between DE and non-DE genes, it is highly robust towards outliers at all sample sizes, while increasing the sample size improves the discriminative performance for all methods. The high performance of ABSSeq on the outlier data sets supports the efficiency of the implemented approach based on moderated median absolute deviation (MAD) in outlier detection even at small sample size (see Methods). Together with the results in Fig. 1, our model on count differences seems to perform at least as good as other models using NB distributed data.

Differential expression analysis on qRT-PCR validated real data

As simulated data are by nature artificial, we further evaluate method performance on real data sets. The first of these relate to the MAQC study, for which RNA-Seq-identified DE genes were validated by quantitative reverse transcription PCR (qRT-PCR) [33] based on the commercially available TaqMan and PrimePCR methodologies. Although there is no single "gold standard" for assessment of RNA-Seq data reliability [29], qRT-PCR based methods have widely been proposed and applied as a validation tool for DE results from both microarray [34] and RNA-Seq studies [35]. Here, we analysed two qRT-PCR validated data sets from the MAQC study: the TaqMan data set from MAQC-I, which included an assessment of a very small fraction of the total genes (1044 out of more than 50,000 genes from hg19 annotation) and may thus be subject to biases, and additionally the PrimePCR data set from SEQC (equivalent to MAQC-III), which covers more than 20,000 genes [29]. These two data sets were used to derive ROC curves and AUC measures for the compared analysis methods. We consider this approach to provide at least an indication of the reliability and sensitivity of the analysis approach. We here follow the general strategy from [36] and [20] and divide the TaqMan and PrimePCR gene sets into a DE (true positive) group and a non-DE (false positive) group based on whether their absolute log fold change (logFC) is larger or smaller than a defined threshold. We use a logFC threshold of 0.5 (1.4 fold change) to derive ROC curves.

The results for both data sets are essentially identical (Figs. 3a and 3b). While the alternative methods can detect approximately half of the TaqMan validated DE genes without false positives, ABSSeq is even able to identify more than 75 % of the true DE genes with a false positive rate of less than 0.25 (Fig. 3a). ABSSeq reaches the highest AUC of 0.853 among six methods (baySeq: 0.840, Voom: 0.817, edgeR: 0.802, DESeq: 0.795, EBSeq: 0.783 and DESeq2: 0.777). For the PrimePCR data set, the AUC for each method decreases as the number of validated genes increases (Fig. 3b). Again, ABSSeq performs best among all seven methods, supporting its ability to discriminate efficiently between DE and non-DE genes.

Analysis approaches, which do not consider the magnitude of expression differences, might yield

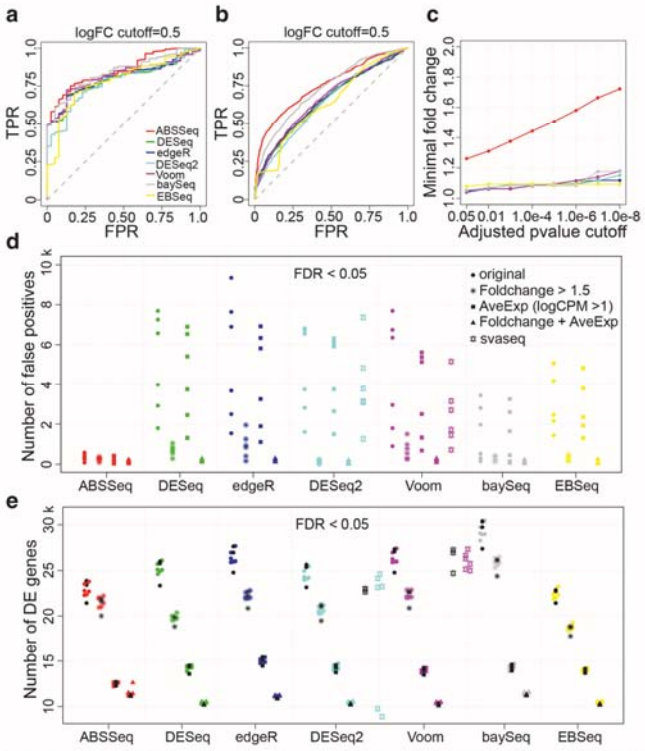


Fig. 3 Comparison of methods using validated real data sets. **a**-**c** based on data from the MAQC study; **d**-**e** based on the ABRF data set. ROC analysis for (**a**) TaqMan and (**b**) PrimePCR data sets at a qRT-PCR absolute log-ratio (logFC) threshold of 0.5. TPR, true positive rate; FPR, false positive rate. ABSSeq performs better than other methods in detecting true differential expression. A gene was considered to be not differentially regulated if its logFC was less than 0.2. **c** Minimal fold changes under various ajusted *p*-value cutoffs for the MAQC II data set. **d** Number of false postives in comparisons of samples from same condition but different lab sites and (**e**) number of DE genes in comparison of samples from two conditons under additional filtering and confounding factor assessment approaches. Symbols in black show results from comparison of conditions from same laboratory and colored symbols those from comparison of conditions across laboratories. Genes are counted under 5 situations: orginal, without filtering (circle symbols); Foldchange, with a value greater than 1.5 (star symbols); AveExp, with average logCPM greater than 1 (square symbols); combination of Foldchange and AveExp (triangle symbols); and svaseq tested only for DESeq2 and Voom (pentacle symbols)

highly statistically significant DE for genes with only small fold change (as shown in Fig. 3c), which may however often be the result of chance. The number of these type of DE genes is usually not reduced by using an adjusted *p*-value cutoff in the alternative approaches, even if the cutoff is below 1.0e-8. Therefore, other cutoff criteria are required such as fold change, which has the problem that the biologically relevant cutoff point is not clear. The ABSSeq-based analysis instead produces high correlation between the minimal fold change and the inferred adjusted *p*-value, indicating that the *p*-value alone will select against DE genes with small fold change. Additional cutoff criteria therefore do not seem to be necessary for reliable DE gene identification.

Influence of cut-off criteria and confounding factor analysis procedures

We next investigate the influence of additional cut-off criteria on DE detection with the help of the ABRF data set, which is based on RNA-Seq data generated for the same sample in three different laboratories. We apply the considered methods on this data set, which only contains variation caused by differences among the conisdered laboratories, such as biases during library preparation [28], but not true DE, thus allowing us to assess the efficacy of the methods to reduce the number of false positives ([28]; see also above). In spite of varying numbers of detected DE genes, ABSSeq reports lowest number of false positives among all methods, irrespective of any additional filtering approach (Fig. 3d).

baySeq and EBSeq also produce small numbers of false positives then compared to the remaining methods excluding ABSSeq. For all methods, the number of false positives reduces dramatically when filtered by fold-change (>1.5; star symbols in Fig. 3c) but less so when filtered by expression level (AveExp, logCPM > 1; square symbols in Fig. 3d). This finding strongly suggests that a high foldchange cut-off increases power to control the false positive rate, yet with the problem that the choice of cut-off value will usually be arbitrary.

In addition, high specificity (i.e., efficient control of the false positive rate) might lead to low senstitivity (i.e., reduced efficiency to detect true positives). To evaluate the ability of ABSSeq to detect true positives, we apply ABSSeq and alternative methods on the ABRF data set whereby in this case we focus on the comparison of the two considered conditions (i.e., tissues) either within the considered laboratories (i.e., condition A and B from the same laboratory are compared) or across the laboratories (i.e., condition A from laboratory 1 is compared with condition B from laboratory 2, and so on for all possible combinations between the two conditions). The results are shown in Fig. 3e (black for comparison of conditions from same laboratory and other colors for comparison of conditions across laboratories). All seven methods report similar numbers of DE genes, especially after fold-change filtering. This result indicates that ABSSeq retains similar sensitivity than that shown by the alternative approaches.

Confounding variation can originate from library preparation or other kinds of batch effects. To remove its influence on DE detection, it can be modeled and thus integrated into the statistical analysis [28], as implemented in svaseq [37]. To illustrate the possible influence of such variation, we applied svaseq togethor with DESeq2 and Voom. Svaseq together with Voom is able to remove more than 50 % false positves for the ABRF data set (Fig. 3d, indicated by the pink pentacle symbols), in consistency with the previous application of the svaseq approach on data from the SEQC study [28]. However, when svaseq is combined with DESeq2 it leads to only a small decrease in the number of false positives (Fig. 3d, indicated by light blue pentacle symbols). This result may suggest that the performance of svaseq depends on the DE detection method itself and/or the linear model used in such methods. Moreover, the application of svaseq does not decrease sensitivity when combined with Voom and only to a small extent when combined with DESeq2 (Fig. 3e), suggesting that svaseq mainly improves removal of false positives but does not bias detection of true DE. In general, the usage of such confounding factor assessment procedures, including svaseq and also PEER [38] can help improve DE detection. Yet, at the moment, its combination with the various DE analysis methods is not straightforward, because both svaseq and PEER produce non-integer values, whereas several of the current DE analysis methods (including ABSSeq) rely on integer count data. It thus represents a promising challenge to further develop these procedures as integrated modules of the common DE detection methods.

Assessment of statistical power via signal to noise ratio

To evaluate the statistical power of each method in measuring the magnitude of DE in dependence of its variance, we repeated above comparison using genes that are exclusively expressed in only one condition of the MAQC-II data set following the approach from [20]. The magnitude of DE of genes expressed in only one condition is ideally shown as a signal-to-noise (SNR) ratio (mean over standard deviation), which should be monotonically correlated with the p-value [20]. A poor correlation between SNR ratio and p-value might lead to reduced sensitivity (type II error) by assigning a large p-value to small SNR ratio (i.e. high variance). The monotonic dependency between predictor (SNR ratio) and response (adjusted p-value) is inferred through an isotonic regression on 1514 paired variables (genes). Results are shown in Fig. 4. All methods but DESeq and edgeR exhibit the desired monotonic behavior between SNR ratio and adjusted p-value, in consistency with previous results from [20]. Two empirical Bayes based approaches: baySeq and EBSeq yield quite similar correlations between SNR ratio and posterior probabilities. In addition, Voom assigns a more significant adjusted p-value for one specific gene with high SNR ratio but low expression (marked by green elipse in Fig. 4) whereas alternative methods produce adjusted p-values of around 0.05 (gray dashed line), suggesting Voom is more sensitive to DE at low expression level. Since DESeq2 and Voom test DE on log fold change, we postulate that the closer correlation between SNR ratio and adjusted p-value of ABSSeq is due to modelling directly the magnitude of DE difference. Overall, these results suggest that ABSSeq seems to model the magnitude of count difference with higher accuracy, which might help DE inference by reducing false positives.

Differential expression analysis of real data with unbalanced designs

Another real data set (HapMap-CEU) is taken from [39], consisting of 41 highly dispersed cDNA samples from 17 females and 24 males. DE genes are inferred from male–female comparisons. Following [23], a sensitivity analysis is predicted to find an over-representation of inferred DE genes from the sex chromosomes. Indeed, the top ten DE genes always include genes from sex chromosomes (Table 1). All methods except ABSSeq and Voom

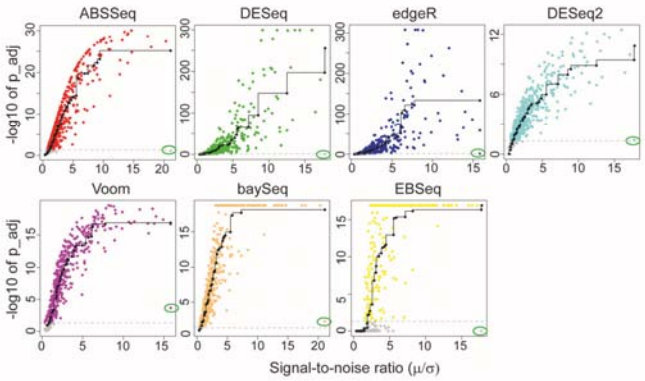


Fig. 4 Correlation between signal-to-noise ratio and *p*-value with true DE present in only one condition. Evaluation is based on a total of 1514 genes that are exclusively expressed in one condition in the MAQC-II data set. Gray points indicate genes with adjusted *p*-value value ≥ 0.05. The data point highlighted by the green elipse refers to the gene with high signal-to-noise ratio but low expression. The correlation is inferred using isotonic regression (black line)

identify DE genes beyond sex chromosomes. This may indicate that ABSSeq and Voom retain higher specificity than the remaining methods and that alternative methods may not well model variance introduced by unequal smaple sizes. EBSeq produces the lowest number of DE genes from sex chromosomes but the highest number from autosomes, confirming the previously observed lack of power of this method for the analysis of data with such high dispersion and uneven sample sizes [13]. Given the unequal sample size in this data set, the similar performance of ABSSeq to that of alternative methods also suggests that our model is able to handle unequal sample sizes and high dispersion. In particular, in ABSSeq, we attempted to compensate for unequal sample size by adding expected reads counts to the smaller group until sample sizes are equal. We always take the mean reads count of the small group as expected count, in order to minimize possible biases in subsequent variance estimations (see also Methods). This compensation step is likely crucial for unbalanced data designs, especially in case of even larger differences than in our test data set. In the future, it may be worth exploring in more detail alternative compensation procedures.

Table 1 Number of DE genes from sex chromosomes detected by each method in the HapMap-CEU data set at FDR-ajusted *p*-value of 0.05

Method	Sex/Total	Sex in Top 10
ABSSeq	7/7	7
DESeq	7/25	5
edgeR	7/20	7
DESeq2	7/10	7
Voom	7/7	7
baySeq	9/27	7
EBSeq	2/38	2

Moderating fold change

Fold change often serves as a more informative indicator for biologists to identify DEs. It is also utilized in gene ranking to select candidates for further investigation and visualization (e,g, heatmap of several comparisons). However, the fold change neglects variance across samples and might not necessarily be informative, especially for genes with low counts (see also discussion above). To overcome this problem, DESeq2 introduces an empirical Bayes shrinkage for fold change estimation, which moderates the log fold change according to gene-specific dispersion [12]. Fold change can also be represented as a function of absolute count differences (see Methods), suggesting a potential moderation of fold change via counts difference (e.g., expected counts difference). Therefore, we introduce a fold change shrinkage procedure according to count differences and dispersion. Figure 5 shows how it works using the Bottomly data set [40]. Genes with small counts tend to have high raw fold changes (Fig. 5a), which constrains reliable gene ranking by fold change at dynamic expression level. Shrinking fold change by adding pseudocounts according to expression level (see Methods) removes this trend (Fig. 5b).

However, this shrinkage approach neglects the gene-specific dispersion and thus shows no effects on non-DE genes with high dispersion as well as high expression (Fig. 5a and b, marked by green elipses). After taking account of gene-specific dispersion (Fig. 5c), the fold

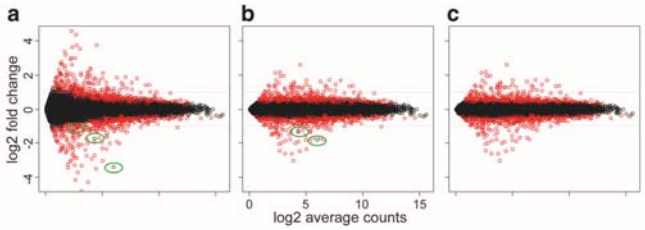


Fig. 5 Moderation of log2 fold change. **a** Raw data (without shrinkage) of the Bottomfly study. **b** The same data corrected by expression level. **c** The same data corrected by expression level and gene-specific dispersion. DE genes (adjusted pvalue <0.05) are shown in red. Non-DE genes with high log2 fold change are marked by green elipses

changes approximately reflect DE genes (in red under adjusted *p*value <0.05) and produce nearly evenly distributed fold change values, apparently improving gene ranking and visualization. Notably, shrinkage with gene-specific dispersion only influences a small part of genes with high dispersion. Unlike the approach in DESeq2, our shrinkage method on gene-specific dispersion is based on *p*-values and therefore does not change the number of inferred significant DE genes. In practice, users can obtain all three types of fold change values (raw, shrinked by smoothed dispersion according to mean, and shrinked by both smoothed and gene-specific dispersions) in ABSSeq.

Conclusions

Here we introduce a new method for differiential expression analysis of RNA-Seq count data, ABSSeq. Distinct from other current methods, ABSSeq infers DE genes through the absolute differences in gene expression and assumes the differences to be influenced by two sources of variation: that found for average gene expression levels and that found for the magnitude of differential expression. Our approach employs a NB distribution to model these two parts and, as a consequence, it is able to detect DE genes more effectively than existing methods, as demonstrated by our analysis of both real and simulated data. In particular, ABSSeq shows an advantage in discriminating DE genes against non-DE ones, it applies an efficient outlier detection approach and is thus robust against outliers. Morever, ABSSeq inferred *p*-values correlate with the magnitude of count differences, thus producing a linear relationship between SNR ratio and *p*-value. As a result, it reduces type I error rates at both very low and high expression level and it also leads to a smaller number of highly significant DE genes with small fold change.

In addition, ABSSeq introduces a procedure to shrink fold change according to the smoothed dispersion across expression level and observed dispersion (gene-specific), which permits fold change comparisons across genes and thus might favor downstream analyses, such as gene set enrichment analysis by ranking [41], clustering and visualization (heatmap) or candidate selection. A potential improvement of our approach in the future may be to adapt it to allow usage of more complex models which consider multiple conditions, its combination with additional normalization procedures, such as those implemented in PEER and svaseq [37, 38], which can further help to filter out unwanted variation, and also adjustment of our approach to allow for analysis of DE at the transcript level (in addition to the gene level, currently implemented). In summary, based on our analysis, we conclude that ABSSeq represents a highly efficient approach for identification of significant DEs across a wide range of conditions and may help efficient downstram analysis of DEs.

Methods

Datasets

In this study, the performance of methods is assessed with the help of two types of data sets: simulated and real. The simulated data sets are derived from the study of Soneson et al. [7]. Following the approach in [21], Soneson et al. used the mean and variances from Pickrell's RNA-Seq dataset ([42]; 69 lymphoblastoid human cell lines derived from unrelated Nigerian individuals) as parameters to generate read counts for each gene from a Poisson or NB distribution. The simulated data sets were generated to follow either a NB distribution (denoted by NB), half NB and half Poisson (denoted by P), NB with single sample outliers (denoted by S) and NB with random outliers (denoted by R). Each set includes 10 independently repeated simulations of two treatment groups and different replicate sample sizes of 2, 5 or 10 for each group. A total of 12,500 genes with high expression (reads count) is considered, for which expression variation is simulated with or without DE genes according to the tests performed.

Five real data sets were considered. Four of these (all except of ABRF) were downloaded from http://bowtie-

bio.sourceforge.net/recount/ [43]. The MicroArray Quality Control (MAQC) study has been used to evaluate the performance of different gene expression analysis methods [36]. It is based on replicated RNA samples of the human whole body (UHR) and brain (BHR) [44, 45]. We use the MAQC II data set for analysis of performance of DE detection methods. For each group (body or brain), seven technical replicates are produced. We filtered out genes with zero read counts across samples before analysis. The raw data of MAQC-II are available from the NCBI SRA database under SRA010153. Moreover, we also use two qRT-PCR validated data sets, either based on the TaqMan methodology, comprising more than 1000 genes from MAQC I, available at NCBI Gene expression Omnibus database under GSE5350, and that based on PrimePCR including more than 20,000 genes from SEQC (MAQC III), available under GSE56457.

The modencodefly data set served to study gene expression during the development of *Drosophila melanogaster* [26], covering 30 distinct developmental stages. Each of the stages consists of 4 up to 6 technical replicates. We subsample from each stage 4 replicates to construct a 2:2 pairwise study.

The HapMap-CEU data set [39] includes 41 samples based on immortalized B-cells from 41 unrelated CEPH grandparents. It contains 17 female samples and 24 male samples.

The ABRF data set is the Association of Biomolecular Resource Facilities next-generation sequencing (ABRF-NGS) study on RNA-seq, which aims to assess RNA-Seq data across laboratory sites and platforms [27] and relies on the the same samples used in the Sequencing Quality Control (SEQC) study [29]. Here we use data from two samples generated via a ribo-depleted protocol, namely RNA from cancer cell lines and also RNA from pooled normal human brain tissues. We thus exclude data from mixtures of these samples and that based on other protocols. The raw data and counts tables are available at the Gene Expression Omnibus database under accession number GSE48035. This study compared RNA-Seq data for the same samples assessed in different laboratories.

The Bottomly data set is from a study that characterized transcriptomic differences between two inbred mouse strains (C57BL/6J and DBA/2J) with 10 and 11 replicates each, respectively [40]. We filtered out genes with zero read counts across samples before analysis.

Data structure and normalization

RNA-Seq data is represented as count of reads (c_{ij}) for genes (i) and samples (j) at different conditions (A, B or more), which are discrete. Due to technical and other reasons, the total number of reads varies between samples or even sequencing lanes. The read count must thus be normalized before comparison across samples. The most common practice is to scale the counts according to the total number of reads of each sample [46, 47]. However, this approach was shown to introduce a bias in DE inference since DE genes can be responsive for large variations in total read number [36]. Here, for ABSSeq, we chose the quantile-based procedure, which yielded much better concordance with the qRT-PCR data [36]. In addition, we also offer geometric mean based normalization procedure in ABSSeq, which we borrowed from DESeq.

Outlier detection and replacement

Outliers mask the statistical significance by influencing the estimation of mean and variance. Given extreme high read counts outliers are often present in one or more RNA-Seq samples and thus it is essential for DE inference to reduce the impact of outliers [12, 13, 48]. Since RNA-Seq data could be treated to be log-normally distributed, ABSSeq utilizes the median absolute deviation (MAD) to detect the outliers in log-transformed read counts. However, due to typically limited sample size in RNA-Seq data, MAD could be extremely small or even zero possibly resulting in over-detection. To solve this problem, we adjusted the MAD of each gene using the highest population standard deviation (SD) σ_0, that is

$$\hat{M}_{iA||B} = \sqrt{\frac{n_{A||B}\hat{M}^2_{iA||B} + n_0\sigma_0^2}{n_{A||B} + n_0}} \tag{1}$$

where $n_{A||B}$ is the sample size for each condition and n_0 is the weight for σ_0. It is similar to empirical Bayes in limma and shrinks the observed MAD toward the highest population SD, thus avoiding small MADs in further analyses. In practice, we set $n_0 = 2$ and $\sigma_0 = \sigma_{\mu=1}$ due to the quadratic mean-variance relationship in RNA-Seq data (highest dispersion at lowest expression level), and also provide an interface for the user to change these two values. Thus, the outliers are defined as

$$\log\left(c_{i,j\in A||B} + 1\right) - median\left(log\left(c_{i,j\in A||B} + 1\right)\right) - 2\hat{M}_{iA||B} > 0 \tag{2}$$

and replaced by $median\left(log\left(c_{i,j\in A||B} + 1\right)\right) + \hat{M}_{iA||B}$. The natural exponent of the read counts after outlier replacement is then used as input for DE testing in ABSSeq.

Inferring DE genes based on absolute expression differences between conditions

DE inference relies on an assessment of the difference of expression levels between two conditions (or more) as well as the variance across replicate samples. The popular null hypothesis for testing DE is that the mean read count for a particular gene is identical between

conditions. However, the standard analysis of such a hypothesis neglects the magnitude of encountered differences. Here we use a distinct test statistic: the absolute difference of read counts between conditions (specifically, A and B), which was firstly applied to detect differential expression, epigenetics changes and transcription factors binding sites in the program EpiCenter [10], that is

$$D_i = \left| \sum_{j \in A} c_{ij} - \sum_{j \in B} c_{ij} \right| \quad (3)$$

When the sample sizes between groups are not equal, D_i introduces a bias by favoring the larger group, which has a higher likelihood to reach higher sum counts by chance, thus more likely resulting in non-zero D_i). For this reason, ABSSeq compensates the smaller group with the most likely read counts: the mean. In these cases, D_i might not be an integer and needs to be rounded to the nearest integer.

D_i is always discrete and apparently overdispersed as D_i inherits variance from $\sum_j c_{ij}$ and is less than $\sum_j c_{ij}$ ($\sum_j c_{ij}$ is overdispersed [1, 3]), which suggests that it follows a NB distribution. Based on this idea, ABSSeq employs a NB distribution to model D_i, which has two parameters, the mean m_i and size factor r_i, that is

$$D_i \sim \mathrm{NB}(m_i, r_i) \quad (4)$$

m_i can be treated as the expected value or baseline of D_i which is proportional to average expression level (larger expected value of D_i at higher expression level) or determined using the coefficient of variation (CV) in the tested data. Therefore, m_i is

$$m_i = \alpha c_i \quad (5)$$

where c_i and α are larger value of sum counts, general CV. r_i as the size factor is dependent on the mean-variance relationship and determines the scale of information contained byD_i. We assumeD_i to inherit dispersion from c_i (i.e., the shape of its distribution is similar to that of c_i). As c_i could be written as $c_i = n\mu_i$ $n = \max(n_A, n_B)$, $\mu_i = \max(\mu_{iA}, \mu_{iB})$ (the $\mu_{A||B}$ denotes mean of each condition), we assume c_i has the same dispersion under μ_i. Therefore, the dispersion of c_i becomes

$$v_i = \frac{\left(s_{iA}^2 + s_{iB}^2\right) - \mu_i}{\mu_i^2} \quad (6)$$

whereby v_i and $s_{iA||B}^2$ denote the pooled dispersion factor, the mean and variance of each condition, respectively. r_i is then given as

$$r_i = 1/v_i \quad (7)$$

As a result, DE detection is based on the magnitude of D_i against its expected value m_i and dispersion r_i. ABSSeq also allows DE detection on paired samples by replacing $s_{iA}^2 + s_{iB}^2$ with variance drived directly from paired differences.

Moderating m_i, r_i

It is well-known that the mean-variance relationship of RNA-Seq data is basically quadratic [5], which suggests a relative higher uncertainty of c_i (higher m_i) for genes with low expression levels. To account for the dynamic uncertainty, we moderated m_i by adding pseudocounts to c_{ij} according to the mean-variance relationship, which has no influence onD_i and $s_{iA||B}^2$ but μ_i and r_i, that is

$$\hat{\mu}_i = \mu_i + \mu_{0i} \; \hat{c}_i = c_i + n\mu_{0i} \quad (8)$$

$\hat{\mu}_i$ indeedly represents the upper bound of μ_i. To estimate μ_{0i}, we firstly construct the mean-variance relationship by applying local regression [49] on the graph $\left(\sqrt{v_i}, \mu_i\right)$ with *locfit* package from R, which has been introduced by DESeq. That is

$$\hat{v}_i = f(\sqrt{v_i})^2 \quad (9)$$

Then the smoothed or expected variance for each gene is given by

$$\hat{s}_i^2 = \mu_i + \hat{v}_i \mu_i^2 \quad (10)$$

Since the uncertainty of μ_i always decreases as the expression level or sample size increases, we assume that $\hat{\mu}_{0i}$ could be written as

$$\mu_{0i} = \sqrt{\frac{\theta}{\mu_i} \frac{\hat{s}_i^2}{n-1}} \quad \theta = \sqrt{mean\left(\frac{s_{iA}^2 + s_{iB}^2}{2}\right)} \quad (11)$$

where θ serves as background of uncertainty across all genes.

When the observed variance is 0 (i.e., c_{ij} is the same in all samples), the dispersion of sum counts c_i simply becomes $\hat{v}_i/n$ (combined NB distributed variables with sum size factor $n/\hat{v}_i$), which suggests $\hat{v}_i/n$ serves as the background of v_i. However, $\hat{v}_i$ is usually obained from part of the tested data (on $v_i > 0$), indicating underestimation of $\hat{v}_i$. To penalize this, we add a basic dispersion factor v_0 to v_i, which becomes

$$\hat{r}_i = \frac{1}{\overline{v}_i} \quad \overline{v}_i = v_0 + v_i + \hat{v}_i/n \quad (12)$$

v_0 is determined by quantile estimation on v_i with $v_i > 0$, that is

$$v_0 = quantile\left(v_i | v_i > 0, \sqrt{\beta}\right) \quad (13)$$

where β is the percentage of v_i on $v_i < 0$. Generally, it permits a smaller v_0 for lower β.

Notably, the small variance of μ_i ($s_{iA}^2 + s_{iB}^2 \leq \mu_i$) is not caused by r_i. However, neglecting this variance will introduce false positives at low expression level since the small variance has a higher impact on μ_i when μ_i is small. On the other hand, this small variance could be treated as noise for μ_i or c_i. In light of this, we add a small value to c_i. m_i becomes

$$\hat{m}_i = \alpha(\hat{c}_i + \varepsilon) \;\; \varepsilon = \sqrt{n_i \max\left(s_{iA}^2 + s_{iB}^2, \mu_i\right)} \quad (14)$$

Estimating α

After shifting read counts according to the mean-variance relationship, we simply assume that CVs of all genes are identical. While the SD of log-transformed data stands for the CV at original scale, we get α by

$$\alpha = mean(\sigma_i) \quad (15)$$

where σ_i is obtained by fitting a linear model to log-transformed counts from limma. In practice, the estimated α usually ranges from 0.1 to 0,3. α could also be provided by the user (i.e., testing DEs on prior threshold).

P-value calling

Following (2), we can calculate the p-value for each gene by the cumulative distribution function of $\mathrm{NB}(\hat{m}_i, , \hat{r}_i)$. The false discovery rate (FDR) by Benjamini-Hochberg is used to account for multiple testing as a default.

Moderating log fold change

The log fold change can be described as

$$FC_i = \log\left(\frac{c_i}{c_i - D_i}\right) \quad (16)$$

Thus, we can moderate it busing c_i or D_i. Indeed, in (7), we moderate c_i by adding pseudocounts, which mainly shrinks fold change in response to uncertainity across expression level but not gene-specific dispersion (observed). The gene-specific dispersion $\hat{r}_i$ also determines the scale of information contained by fold change, i.e. a high dispersion indicates low information of fold change, and vice versa [12]. On the other hand, $\hat{r}_i$ also controls the information contained by D_i, indicating a possible moderation of D_i as well as fold change by shrinkage of $\hat{r}_i$. Under certain pvalue from $\mathrm{NB}(\hat{m}_i, , \hat{r}_i)$, increasing $\hat{r}_i$ (decreasing dispersion) will reduce expectation of D_i and thus fold change. Based on this idea, we obain a new D_i by replacing $\hat{r}_i$ with the dispersion obtained through the probability quantile function from the NB distribution, that is

$$\hat{D}_i = q\,\mathrm{NB}(p_i, \hat{m}_i, r_0) \;\; r_0 = \max(\bar{r}_i, 1/mean(\bar{v}_i)) \quad (17)$$

where p_i is the $p$value for gene i. Notably, the moderation is only applied on genes with $\hat{r}_i$ less than r_0. Using this approach the log fold change is then calculated by (16) with $\hat{c}_i$ and $\hat{D}_i$, which approximately normalizes fold change toward the common dispersion (mean). In addition, we also provide an interface for user to change r_0.

Software tools

The figures in this study have been plotted using R.

Additional files

Additional file 1: Overview and command lines for differential expression analysis in R. (PDF 15 kb)

Additional file 2: Tables S1 and S2. On the results of the statistical comparison of differential expression analysis methods. (PDF 322 kb)

Additional file 3: Method-dependent variation in type I error on ABRF data set. (PDF 314 kb)

Abbreviations

AUC, area under curve; DE, differential expression; FC, fold change; FDR, false discovery rate; FPR, false positve rate; logFC, log2 of fold change; NB, negative binomial; RNA-Seq, (high-throughput) sequencing of RNA; ROC, receiver operating characteristic; SEQC, sequencing quality control; TPR, true positive rate

Acknowledgements

We thank Charlotte Soneson and Mauro Delerenzi for kindly providing the simulated data sets. We are grateful to the Rechenzentrum of the University of Kiel for access to the Linux cluster and technical support. WY is a member of the International Max-Planck Research School (IMPRS) for Evolutionary Biology at the University of Kiel.

Funding

The study was funded by the German Science Foundation within the priority program SPP1399 on host-parasite coevolution, individual grants SCHU 1415/8, SCHU1415/9 and RO 2994/3.

Availability of data and material

The datasets, supporting the conclusions of this article, are available in case of the reads count tables for MAQC-II, modencodefly, Bottomly and Hapmap-CEU data sets from Recount (http://bowtie-bio.sourceforge.net/recount/), in the case of the ABRF data set from the Gene Expression Omnibus database under accession number GSE48035, and in case of the simulated data sets via compcodeR from Bioconductor. The raw data of MAQC-II is available at NCBI SRA database under SRA010153, the PrimePCR data set from SEQC under GSE56457; and the TaqMan data set from MAQC (MAQC-I) under GSE5350.

Availability and requirements

Project name: ABSSeq

Project home page: https://bioconductor.org/packages/release/bioc/html/ABSSeq.html

Operating system: Platform-independent

Programming language: R

Other requirements: R 2.10 or Higher

License: GPL 3.0

Authors' contributions

WY had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; PR contributed to study design and wrote the manuscript. HS supervised the study, contributed to study design, and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany. [2]Centre for Molecular Biology, Institute for Clinical Molecular Biology, CAU Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany.

Received: 16 December 2015 Accepted: 20 June 2016
Published online: 04 August 2016

References

1. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
2. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11(1):422.
3. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
4. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523–38.
5. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. 2010;38(17):e170.
6. Feng J, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics. 2012;28(21):2782–8.
7. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14(1):91.
8. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu T-M, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006;24(9):1140–50.
9. Liu S, Lin L, Jiang P, Wang D, Xing Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res. 2011;39(2):578–88.
10. Huang W, Umbach DM, Jordan NV, Abell AN, Johnson GL, Li L. Efficiently identifying genome-wide changes with next-generation sequencing data. Nucleic Acids Res. 2011;39(19):e130.
11. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
13. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42(11):e91.
14. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):3.
15. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
16. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8): 1035–43. doi:10.1093/bioinformatics/btt087.
17. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R foundation for Statistical Computing; 2013. http://www.R-project.org/.
18. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
19. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53.
20. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14(9):R95.
21. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics. 2012;13(1):484.
22. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012;28(13):1721–8.
23. Zhou Y-H, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011;27(19):2672–8.
24. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23(21):2881–7.
25. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210.
26. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW. The developmental transcriptome of Drosophila melanogaster. Nature. 2011;471(7339):473–9.
27. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014;32(9):915–25.
28. Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014;32(9):888–95.
29. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014;32(9):903–14.
30. Van Rooij I, Broekmans F, Te Velde E, Fauser B, Bancsi L, De Jong F, Themmen A. Serum anti-Müllerian hormone levels: a novel measure of ovarian reserve. Hum Reprod. 2002;17(12):3065–71.
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845.
32. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59.
33. Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol. 2006; 24(9):1115–22.
34. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat Methods. 2005;2(5): 337–44.
35. Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011;473(7347):398–402.
36. Bullard J, Purdom E, Hansen K, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11(1):94.
37. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research. 2014;doi: 10.1093/nar/gku864.
38. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500–7.
39. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS. Polymorphic cis-and trans-regulation of human gene expression. PLoS Biol. 2010;8(9):e1000480.
40. Bottomly D, Walter N, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One. 2011;6(3):e17820.
41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545–50.

42. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.
43. Frazee AC, Langmead B, Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011;12(1):449.
44. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu T-M, Goodsaid FM, Pusztai L. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28(8):827–38.
45. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006;24(9):1151–61.
46. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
47. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
48. George NI, Bowyer JF, Crabtree NM, Chang C-W. An Iterative Leave-One-Out Approach to Outlier Detection in RNA-Seq Data. PLoS One. 2015;10(6):e0125224.
49. Loader C. Local Regression and Likelihood. New York: Springer; 1999

Additional file 1: Overview and command lines for differential expression analysis in R.

R commands

Here, we list the R commands for each method that were used to analyze the data in this study. As explained in the main text, we run each method in default settings. All analyses were performed with R version 3.2.2. The data matrix and conditions are denoted as cdat and cgroup, respectively.

ABSSeq

The ABSSeq package, version 1.6.1, is available in Bioconductor.

```
> library(ABSSeq)
> obj <- ABSDataSet( cdat, cgroup )
> obj <- ABSSeq(obj)
> res <- results(obj, c("Amean","Bmean","foldChange","pvalue","adj.pvalue"))
```

DESeq

The DESeq package, version 1.20.0, can be obtained from Bioconductor.

```
> library(DESeq)
> cds <- newCountDataSet(cdat, cgroup)
> cds <- estimateSizeFactors(cds)
> if(length(cgroup) == 2){
+ cds <- estimateDispersions(cds, method= "blind", fitType='local', sharingMode='fit-only')
+   }else{
+     cds <- estimateDispersions(cds, method= "per-condition", fitType='local')
+   }
> res <- nbinomTest(cds, "condA", "condB")
```

edgeR

The edgeR package, version 3.10.2, can be obtained from Bioconductor.

```
> library(edgeR)
> y <- DGEList(counts=cdat, group=cgroup)
> y <- calcNormFactors(y)
> if(2 == length(cgroup)){
+       y$common.dispersion = 0.4
+   }else{
+     y <- estimateCommonDisp(y)
+     y <- estimateTagwiseDisp(y)
+   }
> et <- exactTest(y)
> res <- topTags(et,n=nrow(et))
```

DESeq2

The DESeq2 package, version 1.8.1, can be obained from Bioconductor.

```
> library(DESeq2)
> ds <- DESeqDataSetFromMatrix(countData = cdat, colData = data.frame(cgroup), design =
+                                                ~ cgroup)
> ds <- DESeq(ds)
> res <- results(ds)
```

limma

The limma package, version 3.24.15, can be obtained from Bioconductor. It has been evaluated with two different settings: Voom and QN.

Here is Voom.

```
> library(limma)
> library(edgeR)
> design <- model.matrix(~0+cgroup)
> colnames(design) <- levels(cgroup)
> nf <- calcNormFactors(cdat)
> dat <- voom(cdat, design, plot=FALSE, lib.size=colSums(cdat) * nf)
> fit <- lmFit(dat,design)
> contrast.matrix <- makeContrasts("condB - condA", levels=design)
> fit2 <- contrasts.fit(fit, contrast.matrix)
> fit2 <- eBayes(fit2)
> res=decideTests(fit2,p.value=q.cut,lfc=lfc)
> tab<-topTable(fit2, adjust = "BH", number=nrow(fit2), sort.by='logFC')
```

Here is QN.

```
> library(limma)
> design <- model.matrix(~0+cgroup)
> colnames(design) <- levels(cgroup)
> dat <- log2(cdat+1)
> dat <- normalizeBetweenArrays(dat,method='quantile')
> fit <- lmFit(dat,design)
> contrast.matrix <- makeContrasts("condB - condA", levels=design)
> fit2 <- contrasts.fit(fit, contrast.matrix)
> fit2 <- eBayes(fit2)
> res=decideTests(fit2,p.value=q.cut,lfc=lfc)
> tab<-topTable(fit2, adjust = "BH", number=nrow(fit2), sort.by='logFC')
```

baySeq

The baySeq package, version 2.2.0, can be obtained from Bioconductor.

```
> library(baySeq)
> cl=NULL
```

```
> bcd <- new("countData", data = cdat, replicates = cgroup,
+                groups = list(NDE = rep(1, length(cgroup)), DE = cgroup))
> bcd@libsizes <- getLibsizes(bcd)
> bcd <- getPriors.NB(bcd,smaplesize=5000, cl = cl)
> bcd <- getLikelihoods.NB(bcd, cl = cl)
> baySeq.posteriors.DE <- exp(bcd@posteriors)[, 2]
> baySeq.FDR <- topCounts(bcd, group = 'DE', FDR = 1)$FDR[match(rownames(cdat),
+                                  rownames(topCounts(bcd, group = 'DE', FDR = 1)))]
```

EBSeq

The EBSeq package, version 1.10.0, can be obtained from Bioconductor.

```
> library(EBSeq)
> Sizes <- MedianNorm(cdat)
> bcd <- EBTest(Data=cdat, Conditions=as.factor(cgroup),
+         sizeFactors=Sizes,maxround=5)
> res <- GetDEResults(bcd)
```

Additional file 2: Tables S1 and S2. On the results of the statistical comparison of differential expression analysis methods.

Table S1. Multiple comparisons of type I error performance of all methods

	Tukey multiple comparisons of Means						Wilcoxon tests
	NB			R			Modencodefly data set
	N=2	N=5	N=10	N=2	N=5	N=10	
Paired comparison	**p adj**	**p adj**	**p adj**	**p adj**	**p adj**	**p adj**	**p adj**
DESeq-ABSSeq	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.21E-10
DESeq2-ABSSeq	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.34E-03
edgeR-ABSSeq	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.21E-10
Voom-ABSSeq	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.21E-10
DESeq2-DESeq	0.00E+00	0.00E+00	0.00E+00	5.11E-05	9.81E-01	0.00E+00	2.08E-05
edgeR-DESeq	0.00E+00	0.00E+00	0.00E+00	8.51E-01	0.00E+00	0.00E+00	9.19E-02
Voom-DESeq	1.70E-06	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.32E-03
edgeR-DESeq2	1.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.56E-07
Voom-DESeq2	3.80E-06	0.00E+00	6.93E-02	0.00E+00	0.00E+00	0.00E+00	2.21E-08
Voom-edgeR	4.20E-06	1.81E-04	0.00E+00	0.00E+00	0.00E+00	0.00E+00	9.19E-02
edgeR-robust-edgeR	/	/	/	0.00E+00	0.00E+00	1.97E-02	/
edgeR-robust-ABSSeq	/	/	/	0.00E+00	0.00E+00	0.00E+00	/
edgeR-robust-Voom	/	/	/	0.00E+00	0.00E+00	0.00E+00	/
edgeR-robust-DESeq	/	/	/	0.00E+00	0.00E+00	0.00E+00	/
edgeR-robust-DESeq2	/	/	/	0.00E+00	0.00E+00	0.00E+00	/

* edgeR-robust was only applied on data sets that contain outliers (indicated by R)

Table S2. Tukey's multiple comparisons of means for AUCs

	Tukey multiple comparisons of Means					
	NB			R		
	N=2	N=5	N=10	N=2	N=5	N=10
Paired comparison	**p adj**	**p adj**	**p adj**	**p adj**	**p adj**	**p adj**
EBSeq-ABBSeq	0.00E+00	0.00E+00	2.62E-02	0.00E+00	0.00E+00	0.00E+00
baySeq-ABSSeq	8.73E-01	9.08E-01	5.58E-02	9.99E-01	0.00E+00	0.00E+00
DESeq-ABSSeq	1.00E+00	9.98E-01	7.86E-01	1.30E-04	0.00E+00	0.00E+00
DESeq2-ABSSeq	0.00E+00	0.00E+00	0.00E+00	1.00E-03	0.00E+00	0.00E+00
edgeR-ABSSeq	1.00E+00	9.98E-01	1.00E+00	2.65E-02	0.00E+00	0.00E+00
edgeR-robust-ABSSeq	/	/	/	1.00E+00	0.00E+00	0.00E+00
Voom-ABSSeq	9.96E-01	3.05E-02	0.00E+00	9.68E-01	0.00E+00	0.00E+00
EBSeq-baySeq	0.00E+00	6.11E-03	1.00E+00	0.00E+00	4.71E-01	0.00E+00
DESeq-baySeq	9.15E-01	9.97E-01	6.97E-01	7.05E-04	0.00E+00	9.81E-01
DESeq2-baySeq	3.90E-06	0.00E+00	4.10E-06	4.77E-03	0.00E+00	0.00E+00
edgeR-baySeq	7.76E-01	6.09E-01	1.08E-01	8.94E-02	9.50E-02	8.78E-04
edgeR-robust-baySeq	/	/	/	9.99E-01	1.00E+00	9.70E-04
Voom-baySeq	9.95E-01	3.79E-01	3.08E-03	9.99E-01	0.00E+00	0.00E+00
EBSeq-DESeq	0.00E+00	8.58E-04	5.13E-01	2.25E-01	0.00E+00	0.00E+00
DESeq2-DESeq	0.00E+00	0.00E+00	0.00E+00	9.99E-01	9.12E-01	0.00E+00
edgeR-DESeq	1.00E+00	9.21E-01	9.09E-01	7.58E-01	1.00E-07	2.60E-05
edgeR-robust-DESeq	/	/	/	1.05E-04	0.00E+00	2.90E-05
Voom-DESeq	9.99E-01	1.21E-01	1.12E-05	4.93E-03	0.00E+00	0.00E+00
EBSeq-DESeq2	1.40E-06	1.00E-07	1.22E-05	6.34E-02	0.00E+00	0.00E+00
edgeR-DESeq2	0.00E+00	0.00E+00	0.00E+00	9.69E-01	0.00E+00	0.00E+00
edgeR-robust-DESeq2	/	/	/	8.19E-04	0.00E+00	0.00E+00
Voom-DESeq2	3.00E-07	0.00E+00	5.28E-01	2.72E-02	0.00E+00	0.00E+00
EBSeq-edgeR	0.00E+00	1.50E-05	5.42E-02	3.02E-03	1.53E-04	9.49E-02
edgeR-robust-edgeR	/	/	/	2.25E-02	3.01E-02	1.00E+00
EBSeq-edgeR-robust	/	/	/	0.00E+00	7.57E-01	8.86E-02
EBSeq-Voom	0.00E+00	6.15E-01	7.57E-03	4.00E-07	0.00E+00	0.00E+00
Voom-edgeR	9.84E-01	5.94E-03	1.00E-07	2.97E-01	0.00E+00	0.00E+00
Voom-edgeR-robust	/	/	/	9.57E-01	0.00E+00	0.00E+00

* edgeR-robust was only applied on data sets that contain outliers (indicated by R)

Additional file 3: Method-dependent variation in type I error on ABRF data set.

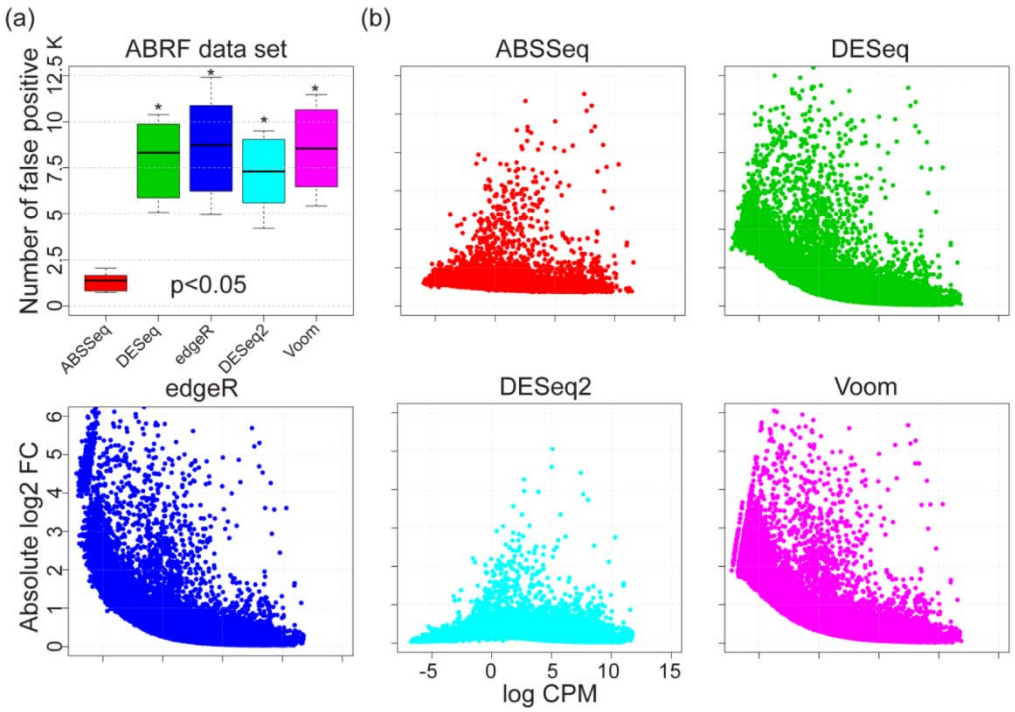


Figure S1. Method-dependent variation in type I error using the ABRF data set. (a) Number of DE genes reported by each method; (b) Points show the absolute log fold change (FC, y-axis) distribution of false positives against the expression level (logCPM, x-axis). DE calling uses data for the same sample generated by different laboratories. Only ABSSeq reports less than 5% DE genes (35766 in total) under $p < 0.05$. DESeq2 also reduces type I error at low expression level but not low fold-change.

Chapter II

aFold: a new method to infer fold change and differential gene expression from RNA-Seq data

Wentao Yang[a,*], Philip Rosenstiel[b], Hinrich Schulenburg[a,*]

[a] Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany

[b] Centre for Molecular Biology, Institute for Clinical Molecular Biology, CAU Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany

* Correspondence: hschulenburg@zoologie.uni-kiel.de

Abstract

Identification of significant differential expression represents a crucial initial step in RNA-Seq analysis. Here, we introduce two procedures to enhance assessment of differential gene expression: a normalization method, qtotal, based on the overall distribution of read count, and an analysis approach, aFold, to calculate fold change and significance of differential expression. aFold uses a polynomial function to model the uncertainty (or variance) of read count, and thus takes into consideration the variance of expression levels across treatments and genes. In comparison to alternative methods, aFold shows at least a similar ability to correctly identify differentially expressed genes. The inferred fold change values are comparable across experiments and might thus facilitate data clustering, visualization, and other downstream applications. We conclude that aFold represents a highly efficient new approach for fold change estimation and identification of significant differential expression across distinct data distributions..

Keywords

RNA-Seq, Transcriptome analysis, Differential gene expression, ABSSeq, aFold

Background

RNA Sequencing or RNA-Seq has become a popular approach for the analysis of gene expression variation and uses the enormous recent advances in next generation sequencing technology. In contrast to array-based methods, RNA-Seq permits the quantification of gene expression without detailed prior genome information, such as gene annotations. Thus it is widely used for both classical model organisms and also non-model taxa [1]. The most common aim of such RNA-Seq studies is to understand inducible biological functions, usually through the analysis of differential gene expression (DE), based on comparison of gene expression levels between two different biological states, as defined by exprimental treatments, developmental stages, or different tissues.

Current statistical approaches for DE analysis in RNA-Seq rely on fitting the distribution of read counts with probabilistic models [2-6]. These methods usually detect significant DE via an inferred probability value, usually adjusted for multiple testing with the false discovery rate (FDR), a procedure, which highly depends on mean-variance relationships [2, 7, 8]. However, the available variance can be arbitrarily small or even zero (indicating under-estimation), even after employing models that adjust individual variance levels according to mean-variance relationships across genes. Such small variance may result in highly statistically significant DE [9, 10] yet also high type I error and FDR at extremely small fold-change [11-13]. To reduce the number of resulting artifacts, additional cut-offs in fold change are commonly used [11-13] and often explicitly warranted, in order to be able to focus on only large changes for subsequent downstream analysis. Commonly used thresholds are a fold change of at least 1.5 or 2.0 [14-16]. Moreover, as false positives of DE are also frequently found for genes with a high coefficient of variation, usually at low expression levels, another cut-off for a minimum expression value or read count is also widely applied [11-13]. Both strategies are not ideal, because they rely on an arbitrary choice of the applied threshold for either minimum fold-change and/or minimum

expression value.

Alternative solutions are based on the idea of merging these cut-offs into a single statistical model or by reducing the effect of high coefficients of variation. For example, TREAT for *t*-test in microarray data partially addresses this problem via testing the significance of DE on a given fold-change threshold [14]. DESeq2 utilizes empirical Bayes method to shrink log fold changes toward zero in consideration of read count dispersion [7]. GFOLD generalizes fold changes based on the posterior distribution of log fold change for RNA-Seq data without replicates [9]. However, these methods only provide a partial solution to the problem. The approach in TREAT still requires that the user provides a cut-off value for fold change. The DESeq2 approach identifies significant DE via a Wald-test comparison of the standard error of log fold change estimates with a normal distribution, which might still result in false positives with extremely small fold-changes [11-13]. The GFOLD method can only be used for data without replication.

Here, we introduce a new approach, aFold (i.e., accurate estimation of fold change from RNA-Seq data), which provides a statistical framework to address the problem of an arbitrary choice of cut-off values by integrating different sources of variation into the calculation of fold-change values. The observed read counts of RNA-Seq data are characterized by several levels of uncertainty (resulting in observed variance) as a consequence of biological variation, but also due to systematic or non-systematic biases during libarary preparation and sequencing [12, 17]. Our approach tries to avoid the implicit assumption of a specific distribution of the read count data (e.g., Poisson or negative binomial, NB [8]). Instead, we explicitly model the uncertainty in the read count data via a polynomial function of the sample mean and standard deviation. aFold takes into account two sources of variance for fold change calculations: 1) the observed variance in gene expression (read count variation across replicates); and 2) the hidden or unknown variance due to limited sample size, which is accommodated via fitting the mean-variance relationship (borrowing

information from genes). aFold additionally penalizes high uncertainty of variance estimates, thus ensuring comparability of fold changes across genes and treatments.

In addition to estimating fold change itself, aFold also provides an efficient strategy for determining fold change cut-off values for different significance levels, thus yielding a statistical test of DE. To achieve this, aFold does not directly model read count distributions. Instead, it employs a zero-centered normal distribution on estimated log fold changes and compares them with the global standard deviation, which avoids the influence of extremely small variances on significance inference. Moreover, we also introduce a new procedure to improve the normalization of RNA-Seq samples under DE, which represents a key problem in transcriptomic studies [10, 18]. Using real and simulated datasets, we demonstrate that aFold is more efficient in DE ranking, DE visualization, and FDR reduction than the two currently most popular RNA-Seq analysis approaches, DESeq2 [7] and limma [19, 20]. For our analysis, we specifically focused on these two alternatives, but did not consider other methods such edgeR, DESeq and baySeq, beause these were previously shown by colleagues or us to perform worse than DESeq2 and Voom when tested with the same data sets [7, 20, 21].

RESULTS and DISCUSSION

We firstly introduce a new normalization procedure, qtotal, which we implemented in the aFold package and which aims at standardizing reads count variation by accommodating the influence of DE on the total number of reads count. Thereafter, we illustrate the aFold approach to model fold change and assess its statistical significance with the help of the HapMap-CEU data set. Thereafter, performance of aFold is compared with that of DESeq2 and Voom, always used under default settings (see Additional file 1). These two methods also consider log fold change for DE inference and report moderated (DESeq2)

or raw (Voom) fold changes as output. Method performance is evaluated based on three complementary criteria: 1) correct gene ranking, that is the ability to rank truly DE genes ahead of non-DE genes; 2) minimization of errors, in particular FDR and type I error rate; and 3) visualization of reported fold changes. We use different well-studied real data sets to assess the performance of each method (Table 1). In addition, we also use simulated data in method evaluation, for which data structure can be efficiently controlled and which have been widely used to evaluate similar DE analysis methods [3, 10, 13, 22-25].

qtotal normalization of reads count data

Reads count of RNA-Seq data requires normalization before DE inference in order to reduce possible biases from variation in sequencing depth, library preparation, sequencing in different lanes, or other random factors [18, 26]. A variety of different normalization procedures have been developed, which adjust individual reads count values across replicates and treatments to achieve a standardization of: 1) total number of reads count (a procedure termed total) as an indicator of sequencing depth; this procedure is however easily influenced by outliers of read counts at high expression level and DE [18]; 2) number of reads count in the lower quartiles (a procedure termed quartile), which is introduced by baySeq to avoid a possible bias due to outliers [3]; this procedure highly depends on sequencing depth that largely impacts quartile function; 3) geometric mean of all reads count (called geometric), which is used by DESeq [2] and DESeq2 [7] to reduce the influence of outliers on approximating the total number of sequence reads; this approach is also sensitive to sequencing depth and DE which might alter the total number of expressed genes as well as the geometric mean of reads count from all genes (see also below data analysis); 4) Trimmed mean of M values (called TMM), which is implemented in edgeR and Voom and is based on the assumption that the majority of genes with high expression are not DE [18]. In general, all above

listed methods rely on the assumption that the majority of genes are equally regulated (up and down) or show no change in expression level, while only very few genes show true DE. However, this assumption may not apply in many situations, for example when gene expression is compared between certain tissues or development stages, for which expression of most or at least a large number of genes can show dramatic changes.

Here we introduce a new normalization procedure, termed qtotal, to address this problem. It is based on the idea that true DE alters the overall reads count distribution (either more or less dispersed), which is reflected by a change in the coefficient of variation (CV) between samples, while variation in sequencing depth does not affect the CV. qtotal quantifies differences in CV between samples and then uses this information to adjust sequence library size, thus explicitly taking into account that there is variation in overall DE between samples (see Methods for details). We used data sets from SEQC, ABRF, and MAQC-II to illustrate the potential problems of different normalization procedures (See Datasets for details). These data sets are based on replicated RNA samples of the human whole body (UHR) and brain (BHR) [27, 28] and show different sequencing depths (ABRF>SEQC>MAQC-II, Table 1). They include validated DE genes by quantitative real-time PCR (qRT-PCR) [29] using commercially available PrimePCR methodologies from SEQC, which covers more than 20,000 genes [17]. We used the PrimePCR results to define upregulated DE genes (log2 fold change >0.5, true positives) and downregulated DE genes (log2 fold change <0.2, false positives).The three data sets show large differences in the number of DE genes of more than 70% (Table 1). Moreover, the BHR data set has a larger number of down-regulated genes than the UHR data set (60% of DE belongs to down-regulation according to the PrimePCR data set under log2 fold change cutoff of 0.5) [13, 17, 21].

The normalization procedures affect the discriminative power of subsequent DE inference. This influence can be assessed with the help of the true and false positive rates (TPR and FPR, respectively) and the area under Receiver

Operating Characteristic (ROC) curve (AUC). The AUCs were inferred with the ROC package in Bioconductor [30], whereby the ROCs were generated based on ordinary fold change under each normalization procedure. We use these three approaches to evaluate the performance of the normalization procedures on the three above listed data sets (Figure 1). The performance of the compared methods varies across the three data sets. The discriminative power of the quartile method decreases as the sequence depth decreases (Figure 1, from left to right). Normalization with total reads number is generally good, indicating that it truly reflects the sequence depth in these three data sets. The TMM and geometric methods perform worse than the other three methods, which might be due to the fact that the majority of genes in the data sets are DE, in apparent contradiction to the methods' underlying assumption. The qtotal method produces the highest AUCs on all three data sets (i.e., 0.836, 0.862 and 0.806 for the ABRF, SEQC and MAQC-II data sets, respectively). qtotal performance is not influenced by sequencing depth. These results suggest that the qtotal approach is able to normalize RNA-Seq data according to its true sequencing depth, facilitating subsequent DE detection. The below application of aFold (e.g., for fold change and DE inference) is thus almost always based on qtotal normalization unless specified otherwise.

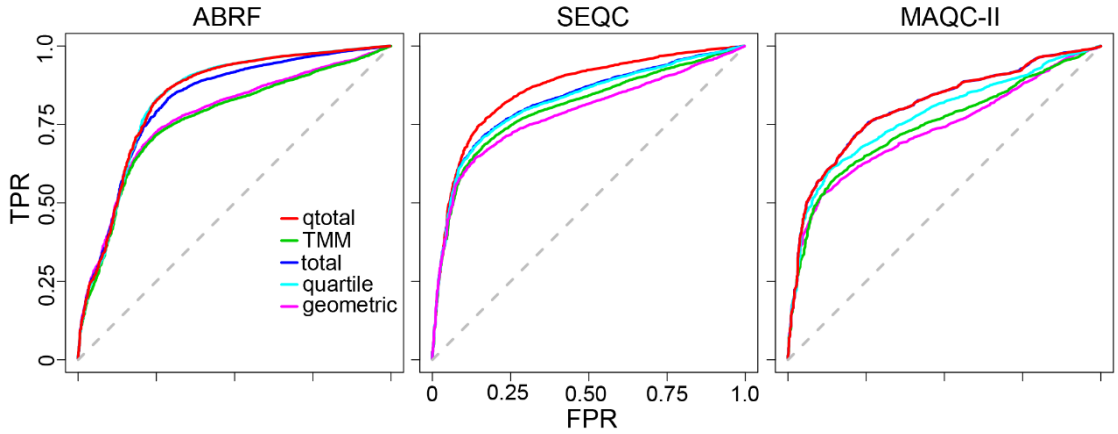


Figure 1. Normalization of RNA-Seq data. ROC analysis using the qRT-PCR validated data sets: ABRF, SEQC and MAQC-II. ROC analysis for PrimePCR data sets at a qRT-PCR absolute log-ratio (logFC) threshold of 0.5. TPR, true positive rate; FPR,

false positive rate. A gene was considered to be not differentially regulated if its logFC in the PrimePCR data was less than 0.2. Five normalization procedures are analyzed: qtotal, TMM, total, quantile and geometric. ROCs are based on ordinary log fold changes.

Ilustration of aFold with the SEQC and HapMap-CEU data sets

Ordinary fold change indicates the extent of DE for a specific gene, although it is usually not comparable across genes or data sets because of differences in variance. To address this problem, the common idea is to shrink fold changes according to dispersion of reads count so that the shrinkage is strong if dispersion for a certain gene is high. DESeq2 employs an empirical Bayes approach to shrink the log fold change according to the mean and dispersion of a gene. The Bayes approach relies on two rounds of fitting a generalized linear model (GLM) to the data: 1) GLM is fitted on reads count to obtain maximum-likelihood estimates (MLEs) for the log fold changes and a zero-centered normal distribution of MLEs from all genes; 2) a second GLM is fitted again on the reads count data using the zero-centered normal distribution as a prior. Interestingly, the second GLM, which relies on the zero-centered normal distribution of MLEs from all genes, might be influenced by the number of genes with significant DE. If the number of DE genes is high, then the inferred normal distribution shows a flat structure and thus little moderation of fold-change (see below). This could potentially introduce a bias in the obtained fold change values. aFold estimates fold change through modelling uncertainty of reads count (see Methods section). In contrast to DESeq2, aFold modelling is not influenced by differences between treatments and thus variation in the number of DE genes. Instead, fold change from aFold is a function of the expression level and dispersion of a specific gene.

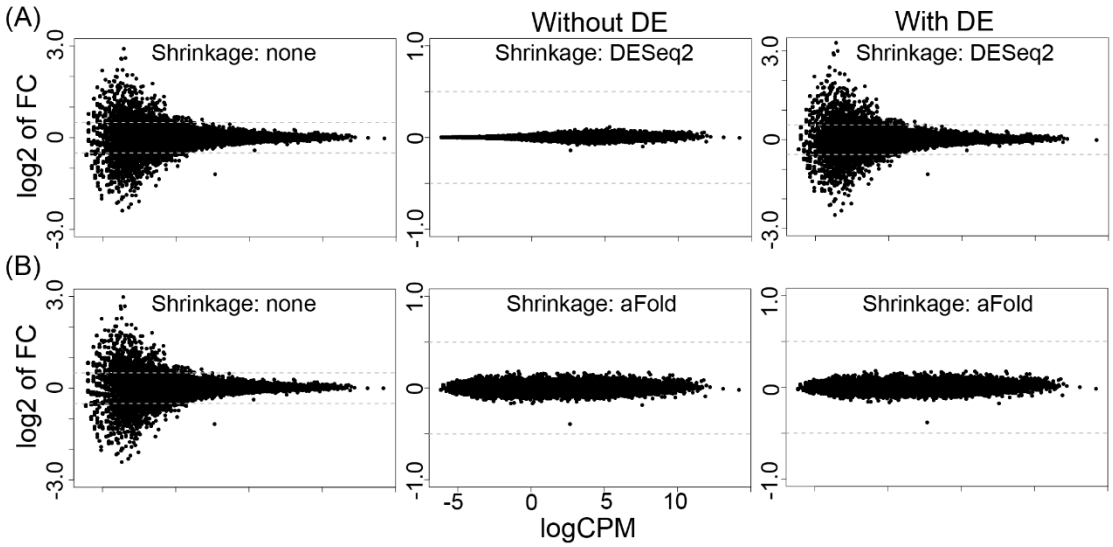


Figure 2. Fold change shrinkage of the aFold and DESeq2 methods. Results are based on the SEQC data set. Fold change is studied for two test comparisons, generated by randomly combining samples from the SEQC data set (four UHR and two BHR). For the first test comparison, samples from identical conditions are combined (all UHR), resulting in the absence of true DE (labeled “Without DE; the left and middle panels). The second test comparison additionally includes samples from a different condition (next to UHR also BHR), yielding a data set with true DE (labeled with DE; right panels). Results for DESeq2 and aFold are based on geometric and qtotal normalization.

The differences between DESeq2 and aFold are demonstrated in Figure 2 based on the SEQC data set and calculation of logCPM with the function from the edgeR package [4]. Six samples in total are randomly selected from this data set (four from UHR and two from BHR) to define two test comparisons. The first of these was set up to contain no significant DE by randomly comparing two UHR with two other UHR samples (thus, all data sets coming from identical conditions, labeled “Without DE”). This test comparison shows a skewed fold change distribution across different expression levels before application of any fold shrinkage procedure (left panels of Figure 2). In this case, both DESeq2 and aFold shrink fold change towards zero according to expression level

(dispersion) but the shrinkage is stronger in DESeq2 (Figure2, middle panels). For the second test comparison, we introduced significant DE into the first test comparison. For this, we combined two of the above used UHR samples and compared them with two randomly chosen BHR samples, yielding a large number of significant DE (because of the differences in tissues). This UHR-BHR combination was added to the above used data set without any DE, resulting in a data set with about 40 % of DE (See Additional File 1 for details). For this test comparison, the DESeq2-based shrinkage procedure leads to almost no change in the fold-change distribution, while that by aFold still results in similar shrinkage as seen for the first test comparison (Figure 2, right panels). This result suggests that fold change moderation by DESeq2 strongly depends on the number of truly DE genes in the data set, which influences shape of the inferred zero-centered normal distribution. In contrast, moderation of aFold appears to be less affected by DE gene numbers but mainly depends on expression level and dispersion (gene specific and overall dispersion).

We next illustrate the aFold approach with the help of HapMap-CEU data set, which consists of 41 highly dispersed samples from 17 females and 24 males. The results are shown in Figure 3. logCPM is again calculated with the function implemented in the edgeR package [4]. Following [24], a sensitivity analysis is predicted to find an over-representation of inferred DE genes from the sex chromosomes.

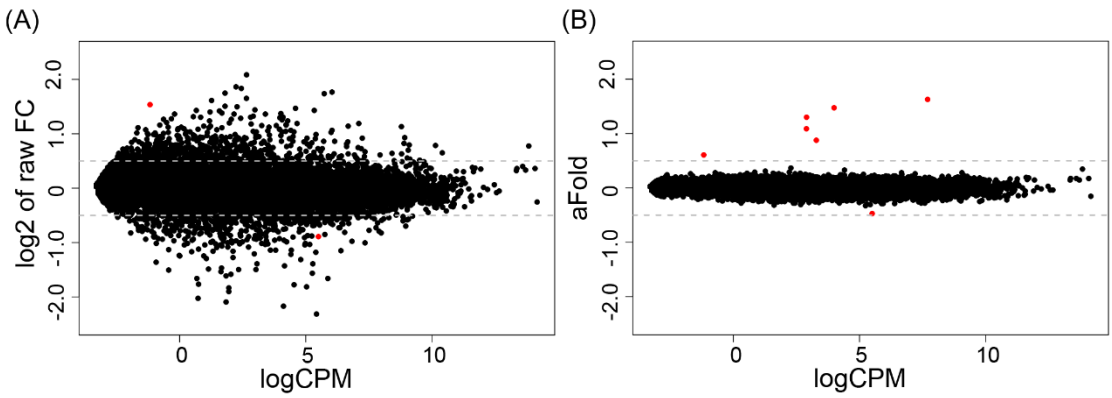


Figure 3. Ilustration of the aFold approach with the HapMap-CEU data set.

Seven genes on sex chromosomes are marked by red color. (A) Raw fold change (without shrinkage). Five genes on sex chromosome are out of y-axis range. (B) Fold change values calculated through the aFold approach. All seven genes from sex chromosomes show largest fold changes.

The ordinary fold changes between female and male samples exhibit high variability (Figure 3A) due to high dispersion of the HapMap-CEU data. In accordance with our previous study [13], seven genes on sex chromosomes are truly DE (shown in red points). For these genes, the ordinary fold change values are very large, thus five of the calculated values fall outside of the y-axis range. Such high ordinary fold changes are often produced by genes with low expression level in at least one of the conditions, which often display a high degree of variance. In these cases, the high ordinary fold change does not necessarily reflect the true DE, but represent an artefact resulting from chance effects at very low expression levels. aFold takes read count variation and expression levels into fold change calculation and thus reports comparable estimates across expression levels. After shrinkage of variance using the approach of aFold, fold change values are much smaller and the truly DE genes appear more distinct from the remaining genes (Figure 3B). These observations may suggest that aFold is able to rank the truly DE before the non-DE genes and produce fold change estimations that directly imply DE.

Table 2. Number of DE genes from sex chromosomes detected by three method in the HapMap-CEU data set at a FDR-adjusted p-value of 0.05.

Method	Sex[1]/Total[2]	Sex in Top 10 (Rank)	
		p-value	Fold-change
aFold	7/7	7	7
DESeq2	7/10	7	6
Voom	7/7	7	5

Number of genes identified by each methods in sex chromosomes (1) and total (2).

Statistical assessment of genes with significant DE confirms the above results

(Table 2). Under an adjusted p-value cut-off of 0.05, all three considered methods (aFold, DESeq2, Voom) identify seven genes on sex chromosomes with significant DE. Both Voom and aFold only report genes located on the sex chromosomes, while DESeq2 also finds three additional genes from other chromosomes. If ranking by p-value, all seven sex chromosome genes are within the top 10 DE genes. However, if genes are ranked by fold changes, then only aFold is able to find these seven gene within the top 10. These results may suggest that fold changes calculated with aFold are more robust than conventional moderated fold change calculations (Voom and DESeq2, respectively) in ranking truly DE genes. Consideration of statistical significance of DE highlights that aFold has similar power than Voom, yet higher specificity in comparison to DESeq2.

Discrimination of DE versus non-DE genes on qRT-PCR validated real data

We next evaluate the discriminative power of the three considered methods with the help of three additional data sets. In Table 2, we show that aFold is more efficient in ranking true DE before non-DEs. However, the few DE genes of the HapMap-CEU data set might lack resolution to reliably assess method performance. Therefore, we additionally consider data from the ABRF, SEQC and MAQC-II studies.

The considered ABRF data set consists of RNA-Seq data from the same mRNA sample generated by three different laboratories [12]. The ABRF data set consists of two conditions (mRNA samples from human whole body and brain), which are sequenced with three replicates at three labs. Therefore, the ARBF data set contains true DE (two conditions) as well as noise (e.g., from library preparation and sequencing), which could be used to assess the accuracy of DE detection approaches, especially their ability to discriminate between signal and noise. Here, we pooled samples for the same condition from three labs into one group (i.e. a comparison of 9:9, nine samples for body and nine for brain).

Similarly, the SEQC and MAQC-II data sets contain samples from body and brain but with different sequence depths and number of replicates (See table 1 and methods for further details). Thereafter, method performance is assessed with TPR and FPR, using ROC curves and resulting AUCs. The AUC has been shown repeatedly to be an informative measure of the overall discriminative power of a method [31-33]. The results are shown in Figure 4. aFold appears to outperform the other two methods, irrespective of ranking criterions (p-value or fold change) and sequencing depth of data sets (ABRF>SEQC>MAQC-II). Notably, p-value and fold-change for aFold are monotonically correlated (see Methods for details), so its ranking is only based on fold change. aFold reaches the highest AUC value of 0.860, 0.824 and 0.774 on the ABRF, SEQC and MAQC-II, respectively.

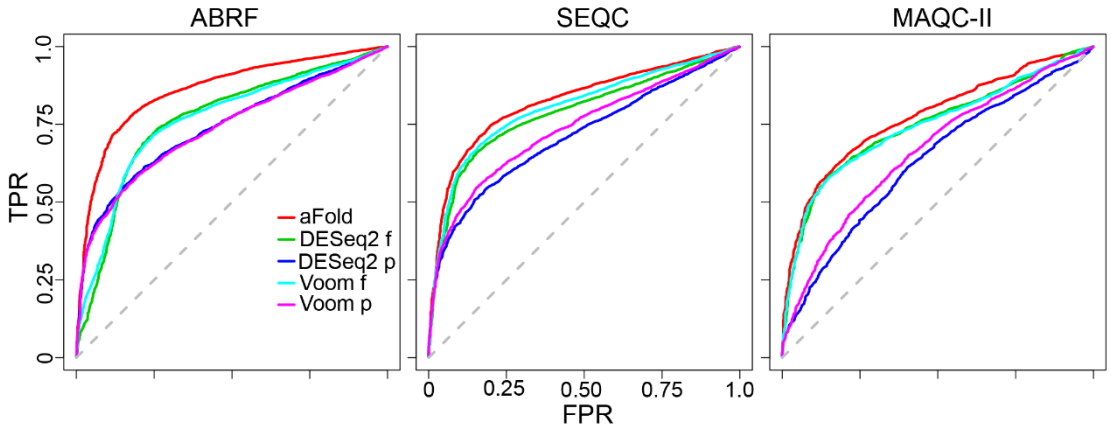


Figure 4. ROC analysis using the qRT-PCR validated data sets. TPR and FPR are defined as in Figure 1. aFold outperforms DESeq2 and Voom in ranking true differential expression at either model-based p-value (denoted by DESeq2 p and Voom p), raw fold change (Voom f) or moderated fold change (DESeq2 f). ROC analysis is performed on three data sets ABRF, SEQC and MAQC-II.

Interestingly, ROC analysis suggests that ranking by fold change is more powerful than p-values to detect true DE [11-13]. However, fold changes may fail to indicate DE in highly dispersed data (usually genes with low expression) (i.e., low FPR, at the beginning of the curve on the ABRF data set). Our model moderates fold changes with information from expression level and dispersion

and might be more powerful to detect and rank DE genes than ordinary fold changes and p-values. Noticeably, aFold performs slightly worse than the ordinary fold change approach on the SEQC and MAQC-II data set in terms of AUCs (0.824 and 0.774 compare to 0.862 and 0.806) but better on the ABRF data set (0.860 to 0.836) (Figure 1 and 4). This might be caused by true DE genes with high dispersion (result in strong moderation of fold change), which could be improved by an increased sequencing depth. In summary, these results suggest that fold change from aFold represents a better way to rank truly DE genes than the methods based on ordinary fold change, pvalue or a combination of both.

Discrimination of DE versus non-DE genes on simulated real data

The negative binomial (NB) distribution is most commonly used to increase reliability of DE detection as RNA-Seq data shows over-dispersed variance [2, 4, 7, 13]. Here, we evaluate the ability of aFold through ROC analysis on data, which was simulated based on the NB distribution, using mean and variances from Pickrell's RNA-Seq dataset [34]. For all simulations, we choose 10% of the 12,500 genes to be DE and symmetrically divide them into up- and down-regulated genes (e.g., 625 up- and 625 down-regulated genes, indicated below by super- and subscripts, respectively). We summarize the results using boxplots for two different simulation settings, including data sets with various replicate sample sizes and, in each case, ten independent repetitions (Figure 5).

When applied on the data that is fully overdispersed according to the NB distribution (denoted by NB_{625}^{625}, Figure 5A), aFold generally yields higher AUCs than alternative methods at large sample size and shows a significant advantage over DESeq2 (n=5) and Voom (n=10) (Tukey's test, $p<0.01$). While DESeq2 directly employs a NB model to identify DE, its performance improves as the sample size increases (Figure 5A). At all three considered sample sizes, aFold produces higher AUC values than DESeq2 and Voom, suggesting that

aFold fits the NB data at least as well as the models used in the other two methods.

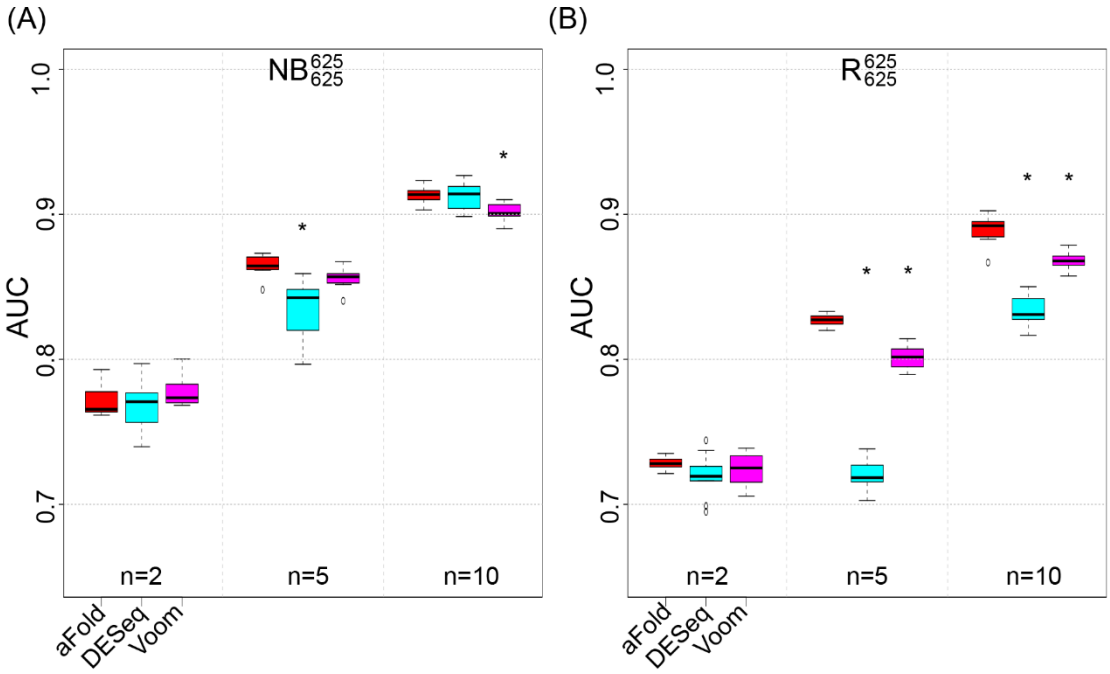


Figure 5. AUC comparison on simulated data. Area under the curve (AUC) for aFold and two alternative methods under two simulation settings: (A) Negative Binomial (NB) distribution and (B) NB distribution with random outliers (R). Each boxplot summarizes the AUCs across 10 independently simulated data sets. Asterisk indicates a statistically significant difference in AUC between aFold and any of the other methods. n indicates the number of considered RNA-Seq replicates, from (2, 5, 10). Under all conditions, aFold is highly effective in correctly identifying differentially expressed genes.

Since aFold uses sample variance to calculate fold change and identify DEs, we next test the influence of outliers that highly impact the sample variance. The outliers are introduced into the NB distributed data by multiplying a randomly generated factor between 5 and 10 with the read count of all genes in all groups obtained through random sampling with a probability of 0.05. The resulting data set (denoted R^{625}_{625}) still has 625 up- and 625 down-regulated genes, in addition to random outliers. For these simulated data sets, aFold demonstrates a significant advantage (Tukey's, $p<0.01$) at large sample sizes

(n=5 or 10) and even reaches an AUC of 0.9 at n=10 (Figure 5B). This result suggests that aFold together with the outlier detection procedure, which we already introduced in ABSSeq, is comparatively mildly affected by outliers. Interestingly, performance of the alternative methods also shows variation. For example, Cook's distance from DESeq2 requires a high number of replicates to improve its performance in presence of outliers. DE detection of Voom, as implemented in limma, is based on log-transformation, which is more robust against outliers and thus results in higher AUC values than DESeq2.

Overall, aFold is at least as good as alternative methods in discriminating between DE and non-DE genes in the presence of outliers, irrespective of data sample size.

Control of false discovery rate and type I error rate

Another important aim of reliable DE detection is to control the false discovery rate (FDR) and minimize the type I error rate (i.e., the null hypothesis is falsely rejected) while identifying a large number of DE genes [21, 35]. To assess these two aspects, we compare the ability of the alternative approaches to control

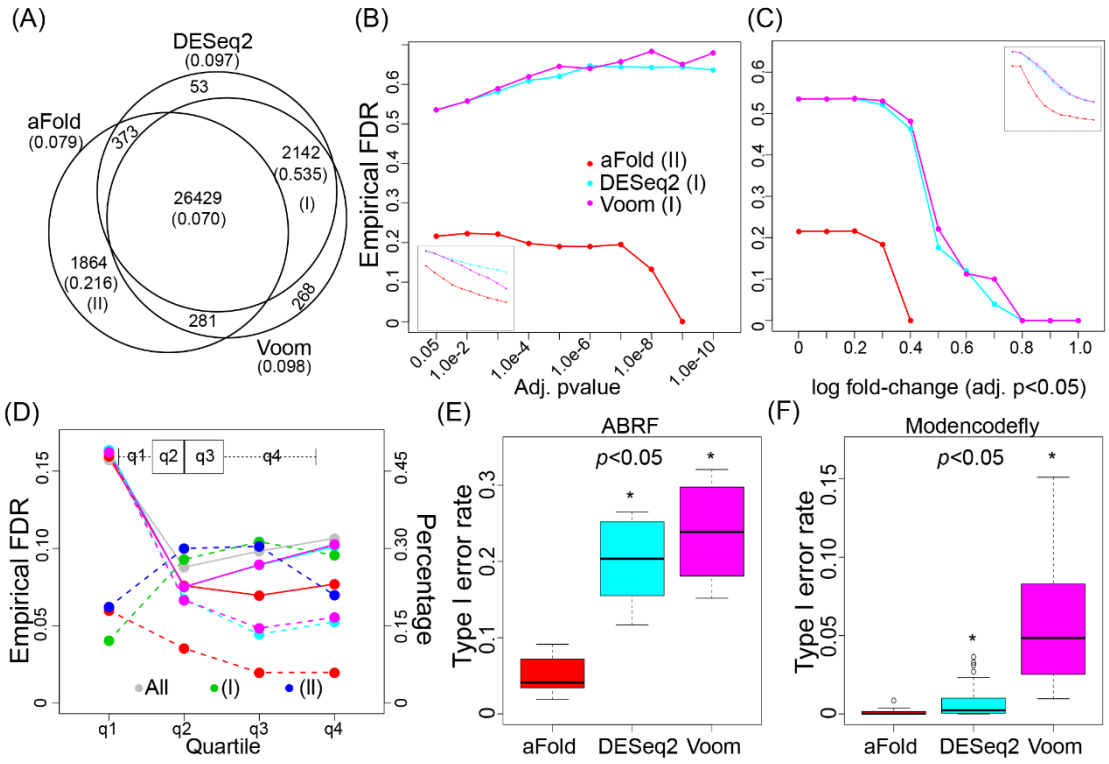


Figure 6. Comparison of methods using real data sets. (A-E) Analysis results based on the ABRF data set. (F) Analysis with the modencodefly data set. True and false positives (TPs, FPs) are defined as described for Figure 2 and used to calculate the empirical false discovery rate (eFDR) as FPs/(TPs+FPs). (A) Venn diagram of the number of DE genes identified by the three methods. Numbers in brackets indicate the eFDR. The specific gene sets of aFold alone or DESeq2 and Voom combined are indicated by roman numbers I and II, respectively. (B-C) eFDR as a function of different cut-offs of either adjusted p-value (B) or fold change (C) for gene sets I (aFold) and II (DESeq2 and Voom). The two inlets show the results based on all DE genes (rather than the subset of genes). (D) eFDR (left Y axis) and percentage of detected DE genes (right Y axis) for different quartiles of the data (X axis). Solid lines indicate eFDR under adjusted p-values of 0.05, dashed lines under adjusted p-values of 0.05 and a log fold change of at least 0.5. Red, turquois, and magenta are as in B and C. Grey line and points show eFDR for all genes (including both DEs and non-DEs). Genes were grouped according to expression (q1, q2, q3 and q4 in boxplot). Lines in blue and green show percentages of detected true DE genes across quartiles for gene set I and II, respectively. (E-F) Type I error rates for the ABRF (E) and modencodefly (F) data sets. Type I error rates are calculated via the number of DEs under p-value < 0.05 divided by the total number of genes.

FDR and type I error rates, using again the ABRF data set and, additionally, the modencodefly data set. Results are summarized in Figure 6.

We firstly evaluate the three methods using the ABRF data set, based on the same structure as above (e.g. results shown in Figure 4). Method performance is assessed with the help of empirical FDR (eFDR), which is the ratio between the number of true false positives and the sum of true and false positives (total number of detected DE genes) (Figure 6A-D). We also investigated the influence of expression levels (Figure 6D) and additional cut-offs (Figure 6B-D) on eFDR. The three methods identified similar number of DE genes under the adjusted p-value of 0.05, whereby Voom reports the largest number (29120), followed by DESeq2 (28997) and aFold (28947, Figure 6A). Moreover, when

cut-offs for fold change, expression level and adjusted p-value are combined, then these three methods report nearly the same number of DE genes, namely 12970, 14251 and 14250 for aFold, DESeq2 and Voom, respectively (84% overall overlap). This result suggests that the above observed differences between aFold and the other two methods result from genes with low expression level and/or fold change, which is consistent with findings from previous studies [11, 13]. As aFold identifies the smallest number of DE genes of the three methods, it also produces a lower overall eFDR (0.079) than both DESeq2 (0.097) and Voom (0.098). This may indicate that aFold is able to control FDR without reducing sensitivity (total number of DE genes).

Interestingly, the genes commonly identified by the three methods retain an eFDR of 0.070 which is closed to the used adjusted p-value cut-off. The additional difference in found DE genes may thus be due to model-dependent biases, either as a consequence of the normalization or the statistical approach implemented. In fact, the eFDRs for the method-specific genes are much higher than those for the commonly identified genes. In particular, the genes only revealed by aFold (denoted as the gene subset II) have an eFDR of 0.216, while those jointly identified by DESeq2 and Voom (denoted as the gene subset I) produce an eFDR of 0.535 (Note that other subsets were not considered because they included only a small number of genes, which does not permit reliable eFDR calculation). The higher eFDR for gene subset I relative to gene subset II may suggest a larger bias caused by DESeq2 and Voom. Similar results are also observed in the SEQC and MAQC-II data sets (Supp. Figure 1C and D). Interestingly, when data is normalized by TMM (Voom) or the geometric mean approach (DESeq2), both subsets are reduced (Additional file 2, Supp. Figure 1A and B). At this situation, only few genes are detected uniquely by aFold, suggesting that aFold retains higher specificity than alternative methods. The subset I is a result of the normalization procedure in aFold (qtotal), which retains low eFDR of 0.216 and supports the efficiency of qtotal normalization. However, it also suggests that genes in subset I actually

have comparatively low fold changes. These three DE and normalization methods yield similar results when applied on a data set that contains a small percentage of DE genes (Bottomly, Additional file 2, Supp. Figure 1E, F and G). Next, we try to reduce eFDR for these two gene subsets by applying more stringent adjusted p-value cut-offs (Figure 6B) or additional fold change cut-offs (Figure 6C). Both alternatives can improve the overall eFDR (for the entire set of DE genes, inlet figure in Figure 6B and C). However, eFDR for gene set I is not reduced through adjusted p-value cut-offs but rather increases with higher cut-off values. Fold change together with adjusted p-value can efficiently decrease eFDR for subset I to a level of 0.05. On the other hand, both cut-offs consistently reduce eFDR of subset II to 0.05 (adjusted p-value =1.0e-9 or 0.05 with log fold-change = 0.4). These results suggest that high eFDR of subset I and II is due to low fold changes (low dispersion). Since false positives often result from under-estimation of variances (with low fold change but high expression or high fold change but low expression) [11-13], we compared eFDRs across different categories of expression level (four quartiles, Figure 6D). Indeed, many genes from subset I and II come from the 1st (low expression) and 4th (high expression) quartile (given in light blue and green in Figure 6D, Y axis on the right side of the panel). Generally, eFDRs at 1st and 4th quartile are higher than 2nd and 3rd for total (grey line). aFold (red line) shows generally lower eFDRs in all quartiles but the 1st one than those obtained for all genes (both DE and non-DE genes, grey line in Figure 6D, Y axis on left side), whereas DESeq2 (blue line) and Voom (pink line) show a similar pattern than that found for all genes. This observation may suggest that aFold is able to improve eFDR at most of expression levels.

Then an additional fold change cut-off of 0.5 is used (under log2-tranformation), then eFDR reduces to around 0.05 in all quartiles for aFold but only the upper ones (3rd and 4th) for DESeq2 and Voom, which produce no change at 1st quartile (0.164 to 0.164) and only a slight improvement at 2nd quartile (0.075 to 0.067). In fact, reducing eFDR for DESeq2 and Voom in 1st quartile to a similar

value of 0.05 requires an extremely high log fold change cut-off of 4.0. Such a cut-off additionally decreases the total number of DE genes to 3564 and 3224 for DESeq2 and Voom, respectively. At the same time, applying a log fold change cut-off of 0.5 for aFold still yields a total of 15325 DE genes. A more efficient way to reduce eFDR at 1st quartile for DESeq2 and Voom is to use a combination of cut-offs for expression level and also p-value (eFDR=0 under logCPM>0 & adjusted p-value <0.05). These results suggest that aFold is able to control FDR by reducing false positives at all expression level while retaining sensitivity, even when more stringent cut-offs are used.

We next compare the methods in their ability to control type I error rates (i.e., the null hypothesis is falsely rejected and thus result in false positives). We use two gene expression data sets: 1) the ABRF data set as above, including data from the same RNA sample but generated by three different laboratories sites, each with 3 replicates; and 2) the modencodefly data set, which contains data for development processes of the fruitfly *Drosophila melanogaster* [36], with technical replicates ranging from 4 to 6. For the modencodefly data set, we randomly selected 4 replicates for each condition and separated them into two groups, which should thus only be characterized by stochastic variations but not true DE. The results of our analysis is summarized in Figure 6E (ABRF) and 6F (modencodefly). At the p-value cut-off of 0.05, only aFold is able to control the type I error rate around 0.05 for the ABRF data set while the other two methods produce a rate above 0.2. For the modencodefly data set, all three methods are able to control type I error rate around 0.05, but aFold reports the smallest number of false positives (average of 14), followed by DESeq2 (119) and Voom (867). Thus, for both data sets, aFold reduces the type I error rate to a larger extent than the alternative methods (Wilcoxon rank test, $p<0.1$), consistent with above results for eFDR.

Taken together, the approach implemented in aFold is able to control FDR and type I error rates more efficiently than the two tested alternative approaches. Moreover, p-values inferred from aFold are directly deduced from and thus

monotonically correlated with fold-change, which allows to apply single cut-off values to select candidate DE genes for further analysis. More importantly, aFold also takes into consideration uneven dispersion across expression levels, which avoids possible biases in inferred DE genes due to large fold change at low expression level [13] and thus permits comparable analysis of DE across different types of data distributions (and thus gene expression characteristics).

Improved visualization of RNA-Seq data

The results of transcriptomic studies are often visualized using a heatmap, which usually takes log fold change as input [37, 38]. However, ordinary fold change ignores sample variance and might result in contradictory patterns in the obtained graphical visualization (e.g., a gene may be upregulated in one

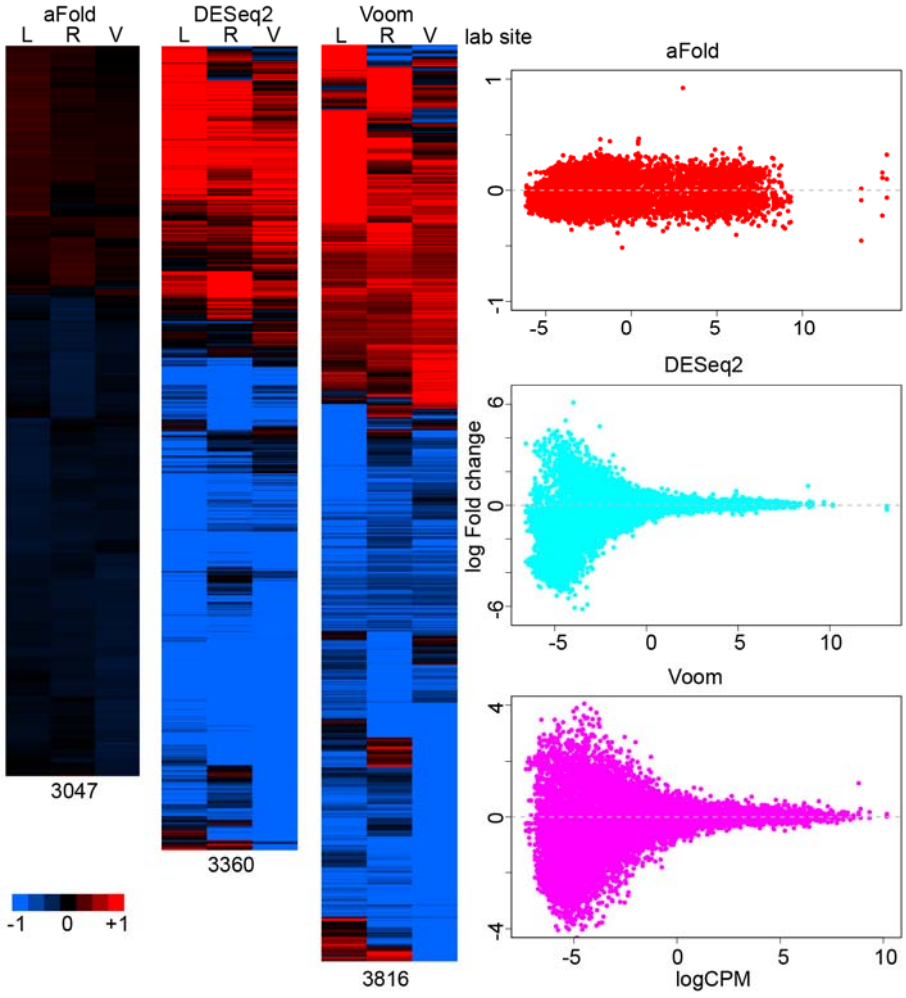


Figure 7. aFold improves visualization of RNA-Seq data. Heatmap of DEs from the same condition but different lab sites that only show significance in one of the

lab sites (based on the ABRF data set). Numbers below heatmaps indicate the total number of genes included. Scatter plots show log fold change distribution across expression levels in each heat map. L, R and V stand for lab sites.

but downregulated in another replicate of the same experiment). aFold takes the observed and mean related variance into account during fold change calculation and, thus, it produces more consistent fold change measures across groups. Here we used the ABRF data set to demonstrate how aFold improves the visualization of RNA-Seq data. The ABRF data sets consist of RNA-Seq data from the same RNA samples, measured under two conditions, but processed and analyzed by three different laboratories. The inferred DE variation among lab sites should only result from random or batch effects (unwanted) environmental or procedural variations, for example due to some differences during library preparation and/or sequencing error. The analysis results indeed identifies a high overlap across lab sites of more than 80% for DE genes detected by the three methods.

However, there are still unique genes identified at each lab site for each method, which are most likely caused by variance under-estimation due to limited sample size (n=3). We take genes that show significance at one lab site from each method as unique genes for each method (adjusted p-value < 0.05), as illustrated in Figure 7. aFold identifies the smallest number of genes with unique DE (most of them retain log fold change <0.5) at only one site but similar pattern across three lab sites (all are up or down-regulation). In contrast, DESeq2 and Voom report many genes that show opposite regulation patterns with high fold change (log fold change >1), which are likely caused by high dispersion across samples and lab sites.

Fold change measures reported by DESeq2 and Voom are unable to capture the magnitude of expression differences and therefore might result in unreal opposite regulation pattern. Indeed, over 75% of genes in each unique set show very low expression (logCPM < 0). These genes also often exhibit high variance combined with high fold change (Figure 7, scatter plot) [2, 8, 13], thus requiring

shrinkage of fold change or additional filtering of expression level to reduce false positives [9, 11-13]. The difficulty here is that there is no universal cut-off value for expression levels because reliability additionally depends on sample size (e.g., large sample size can enhance reliable variance estimation in highly dispersed data and thus also reduces error rates). Here, we demonstrate that aFold is able to accurately estimate fold change by taking into account variance. Thus, aFold improves the visualization of expression data by reducing DE variation, which in turn will facilitate pattern discovery (clustering) and gene set enrichment analyses [38].

Conclusions

Here we first introduce a new normalization procedure, qtotal, which adjusts for the influence of the number of DE genes on the overall reads count distribution and accurately approximates the true sequence depth. qtotal is highly compatible with sequence library size and DE when applied on real data set and thus might help standardization of RNA-Seq data and downstream analyses. In addition to qtotal, we also present a new method, aFold, for fold change estimation and differential expression analysis of RNA-Seq data. Distinct from other current methods, aFold produces fold changes which take into account observed variance and mean-related variance. It thus permits reliable fold-change comparisons across genes, which will help ranking of genes or isoforms for selection of candidates for subsequent analysis and gene set enrichment analysis [38]. Using real and simulated data sets, we demonstrate that aFold is capable of discriminating DE and non-DE. We also introduce a statistical framework based on aFold to infer statistically significant DE genes. This approach shows high power to control FDR and type I error rate across expression levels. In consideration of variance from all sources, aFold produces consistent fold change measures across experiments and might facilitate data clustering and visualization. Based on our analysis, we

conclude that aFold represents a highly efficient novel approach for fold change estimation and identification of significant DE across a wide range of conditions. It may help the experimentalist to avoid an arbitrary choice of cut-off thresholds and may enhance subsequent downstream functional analyses.

MATERIALS AND METHODS

Datasets

Results of this study are based on two types of data sets: simulated and real data. The simulated data sets are derived from the study of Soneson et al. [10], which simulated read count for 12,500 genes from a NB distribution with mean and variances from Pickrell's RNA-Seq dataset [34]. Pickrell's data set consists 69 lymphoblastoid human cell lines derived from unrelated Nigerian individuals. The simulated data is generated under two conditions: NB and NB with random outliers (denoted by R). Each set includes 10 independently repeated simulations of two treatment groups and different replicate sample sizes of 2, 5 or 10 for each group.

In addition, seven real data sets were used to assess the performance of DE inference methods (Table 1). The first four data sets are based on replicated RNA samples of the human whole body (UHR) and brain (BHR) [27, 28]: ABRF, MAQC-II, SEQC and PrimePCR (qRT-PCR validated data set to define true DE). The ABRF data set refers to the Association of Biomolecular Resource Facilities next-generation sequencing (ABRF-NGS) study, which assessed RNA-Seq data variation across laboratory sites and platforms [12] . Here we use data from two samples generated via a ribo-depleted protocol, namely RNA from cancer cell lines and also RNA from pooled normal human brain tissues. We thus exclude data from mixtures of these samples and that based on other protocols. The raw data and counts tables are available at the Gene Expression Omnibus database under accession number GSE48035. The considered RNA-Seq data compares two conditions (UHR and BHR), whereby the same RNA

samples are analyzed in three different laboratories. Any variation between these laboratories should not be due relevant biological differences, but result from variations across sites in environmental and also procedural factors. The MAQC-II data set consists of seven replicates for each condition and is generated by the MicroArray Quality Control (MAQC) study to evaluate the performance of different gene expression analysis methods [39]. The raw data of MAQC-II are available from the NCBI SRA database under SRA010153 and counts table is downloaded from http://bowtie-bio.sourceforge.net/recount/ [40]. The SEQC data set consists of five replicates and is generated by Sequencing Quality Control (SEQC) study [17] available under GSE49712. The PrimePCR data set is based on the PrimePCR approach of qRT-PCR and includes more than 20,000 validated DE genes from SEQC (MAQC III), available under GSE56457.

Three additional real data sets were downloaded from http://bowtie-bio.sourceforge.net/recount/ [40]. The first of these is the modencodefly data set from the modENCODE project [41], which assesses gene expression during the development of *Drosophila melanogaster* [36], covering 30 distinct developmental stages. Each of the stages consists of 4 up to 6 technical replicates, which provides an opportunity to construct subgroups per developmental stage to study stochastic variations but not true DE. We accordingly subsample from each stage to construct a 2:2 pairwise study.

The next real data set is the HapMap-CEU data set [42], which includes 41 samples based on immortalized B-cells from 41 unrelated CEPH grandparents. It contains a relatively large sample size (17 female samples and 24 male samples) and high variations in read count due to genetic diversity. It is well-studied and useful for measuring the ability of DE detection models on large samples and variations [7, 8, 13].

We also considered the Bottomly data set, which is from a study that characterized transcriptomic differences between two inbred mouse strains (C57BL/6J and DBA/2J) with 10 and 11 replicates each, respectively [43]. We

filtered out genes with zero read counts across samples before analysis.

The basic statistics for all real data sets are summarized in Table 1, including the average total number of reads count, the number of present genes, sample size and the average number of DE genes.

Table 1. Overiew of the used real data sets.

Set name	Average library size	#Present Genes	Sample size	#DEs[1]	Used for
ABRF	82297717	35647	18	28996	DE & Type I error
SEQC	57635606	20821	10	17054	DE
MAQC-II	1421992	11907	14	8386	DE
Modencodefly	13709954	13244	147	-	Type I error
HapMap-CEU	5187226	12410	41	8	DE
Bottomly	4904164	13932	21	1119	DE
PrimePCR	-	20801	-	16603	True DE

(-) indicates that the statistics are not applied. (1) the number of DEs represents average DEs reported by aFold, DESeq2 and Voom.

Normalization and outlier detection

Reads count of RNA-Seq data requires normalization before DE inference in order to reduce possible biases from sequence depth, library preparation or even analysis in sequencing lanes [18, 26]. Current approaches for normalization rely on the assumption that the majority of genes (or at least those with high expression) are not DE. This assumption might not be valid under certain biological processes, such as development or aging, when gene expression shows dramatic biological variation. In contrast to current normalization approaches, we assess the overall data dispersion between samples and then use it to adjust normalization. The ratio of library size between two samples can be represented as

$$r = \frac{n_A \phi_A}{n_B \phi_B} \tag{1}$$

where $n_{A||B}$ and $\phi_{A||B}$ stands for the number of present genes (i.e., genes that show expression in samples A or B) and expected reads count per gene in the sample A and B, respectively. n_A and n_B are usually assumed to be identical in current approaches. However, the total number of present genes can vary due to DE and thus the ratio between n_A and n_B needs to be estimated from the data. According to G.4.2 in [44], the effective degree of freedom (effective sample size) is inversely proportional to the CV. As a result, the ration of n_A and n_B can be approximated as

$$\frac{n_A}{n_B} \approx \frac{1/cv_A^2}{1/cv_B^2} = \frac{cv_B^2}{cv_A^2} \tag{2}$$

where $cv_{A||B}$ are the CVs for reads count in the sample A and B, respectively. The ratio between ϕ_A and ϕ_B can be estimated from the observed average reads count in each sample as

$$r_0 = \frac{\phi_A}{\phi_B} = \frac{\mu_A}{\mu_B} \tag{3}$$

While the genes with high expression show low dispersion [7, 8], an estimation of r_0 on genes with high expression should be more stable. However, due to DE, the order of genes according to expression is not identical across samples and ranking via A or B will report different group of genes with high expression as well as r_0. Here we choose the sample that shows small overall CV to rank the genes because the sample with smaller overall CV is less likely influenced by extreme values of reads count (outliers). In addition, since the larger overall CV in a sample indicates stronger up-regulation of genes (increased dispersion) in this sample, ranking via this sample will move the genes with stronger up-regulation towards the right tail (high expression), which in turn enlarges the

estimation of r_o. After ranking, the r_o is then obtained via sliding windows, that is

$$r_{j,o} = \frac{\mu_{A,k|window}}{\mu_{B,k|window}} \quad \text{rank by} \begin{cases} A \mid r_{cv} \le 1 \\ B \mid r_{cv} > 1 \end{cases} \quad r_{cv} = \frac{cv_A}{cv_B} \tag{4}$$

where *j* is the index of the sliding window. $r_{j,o}$ might be influenced by outliers which could result in abnormal $r_{j,o} / r_{j-1,o}$. We then trim $r_{j,o}$ based on $r_{j,o} / r_{j-1,o}$ via median absolute deviation (MAD) and keep the $r_{j,o}$ with the largest *k* as observed ratio r_o. The trimming is also applied on (2). The actual ratio is then calculated from (1). In practice, we select one sample (default is the one with the largest number of reads count) as control and then apply this procedure on all samples to obtain the size factor for normalization. This procedure is implemented as *qtotal* in ABSSeq and set as default normalization procedure for aFold.

For the other two methods, we used the default normalization procedures (voom and TMM for Voom, geometry mean for DESeq2). Outliers influence DE detection through shifting both mean and variance [7, 13], which thus needs to be corrected for. Here we integrate the procedure from our previous ABSSeq approach into aFold, which utilizes the median absolute deviation (MAD) to detect the outliers in log-transformed read count and shrink the read count of outliers toward median of read count from one condition.

Moderating uncertainty of read count

Due to biological and/or other sources of variance, the observed expression value for the ith gene g_i is given as the mean μ_i with uncertainty ε_i.

$$c_i = \mu_i + \varepsilon_i \tag{5}$$

In practice, the uncertainty is represented as the standard deviation (SD) of samples if the SD is independent of the mean. However, In RNA-Seq data or

microarray data, the SD is not independent of μ_i and could be generally written as

$$\sigma_i = a_i \mu_i \quad a_i > 0 \tag{6}$$

This implies that there is propagation of error (uncertainty) in measurement of SD based on μ_i. Therefore, an accurate reads uncertainty measurement should also include the propagation of error from (6). In theory, the propagation uncertainty of SD could be written as

$$\varepsilon_{i,s} = a_i SD(g_i) = a_i s_i \tag{7}$$

where s_i is the sample SD of g_i. Thus, the uncertainty of reads counts for each gene becomes

$$\varepsilon_i = s_i + \varepsilon_{i,s} = s_i + a_i s_i \tag{8}$$

a_i in (2) actually serves as the CV as

$$a_i = \frac{\sigma_i}{\mu_i} \approx \frac{s_i}{\mu_i} \tag{9}$$

Simply, the uncertainty of g_i becomes a polynomial function of sample SD s_i

$$\varepsilon_i = s_i + \frac{s_i^2}{\mu_i} \tag{10}$$

In addition to the observed variance, there are still hidden variances upon expression levels, which are usually described as

$$\omega_i^2 = \mu_i \tag{11}$$

which is dominated by mean read counts of each gene [7]. As a result, the uncertainty from expression level becomes

$$\hat{\varepsilon}_i = \omega_i = \sqrt{\mu}_i \tag{12}$$

We leave out the second term of the polynomial function of (10) in (12) because ω_i is the expected SD for each gene and contains no propagation error. $\hat{\varepsilon}_i$ actually sums up uncertainty across samples and thus requires moderation of

the sample size as

$$\hat{\varepsilon}_i = \frac{\omega_i}{m_i} = \frac{\sqrt{\mu}_i}{m_i} \tag{13}$$

where m_i is defined as the effective sample size for each gene. We use the effective sample size instead of the real sample size in (13) to capture the data structure (i.e, overal dispersion of CVs). According to G.4.2 in [44], a global effective sample size (effective degrees of freedom) can be obtained via

$$m = \frac{mean(\overline{v}_i)^2}{\mathrm{var}(\overline{v}_i)} \quad \overline{v}_i = \frac{s_i}{\mu_i + \varepsilon_i} \tag{14}$$

Instead of using original CVs, *m* is calculated from moderating CVs, which retains information of uncertainty of μ_i and is more stable (Figure 8A), thus avoiding under-estimation of *m*. The effective sample size m_i actually varies across expression levels since the biological variation is more difficult to capture at low than high expression levels. We thus assume that the genes with highest expression retain m as m_i and the rest of genes have a decreasing m_i as

$$m_i = \frac{k\overline{v}_0^2}{\tilde{v}_i^2} \quad \tilde{v}_i = f(\overline{v}_i)\ \overline{v}_0 = \max(v_0, \min(\frac{\omega_i}{\mu_i})) \tag{15}$$

where $\tilde{v}_i$ is the smoothed CV by the *locfit* package from R [45] and v_0 is prior value that could be provided by users to avoid over-estimation of m_i (default is 0.05). We use $\overline{v}_i^2 / \overline{v}_0^2$ instead of $\overline{v}_i / \overline{v}_0$ because *k* is proportional to $\overline{v}_i^2$ as in (15). The final uncertainty is then called as

$$\overline{\varepsilon}_i = \varepsilon_i + \hat{\varepsilon}_i = \varepsilon_i + \frac{\sqrt{\mu}_i}{m_i} \tag{16}$$

Moderating fold change by uncertainty of read count

In our previous study [13], we show that the log fold change can be described as

$$lfc_i = \log(\frac{c_i}{c_i - \Delta_i}) \quad c_i = \max(\mu_{iA}, \mu_{iB}) \; \Delta_i = \mid \mu_{iA} - \mu_{iB} \mid \tag{17}$$

This relationship specifies that log fold change depends on the expression level and the mean read count difference between two conditions (denoted as A and B). Under (17), same mean difference refers to larger log fold change at lower expression level. However, a large uncertainty for μ_i may imply that the actual expression level of i[th] gene is higher than μ_i and, thus, the uncertainty could be treated as unobserved read count. The actual or robust fold change can then be written as

$$lfc_i = \log(\frac{\hat{c}_i}{\hat{c}_i - \Delta_i}) \quad \hat{c}_i = c_i + \overline{\varepsilon}_i \tag{18}$$

The fold change is thus shrunk toward 0 according to uncertainty or variance.

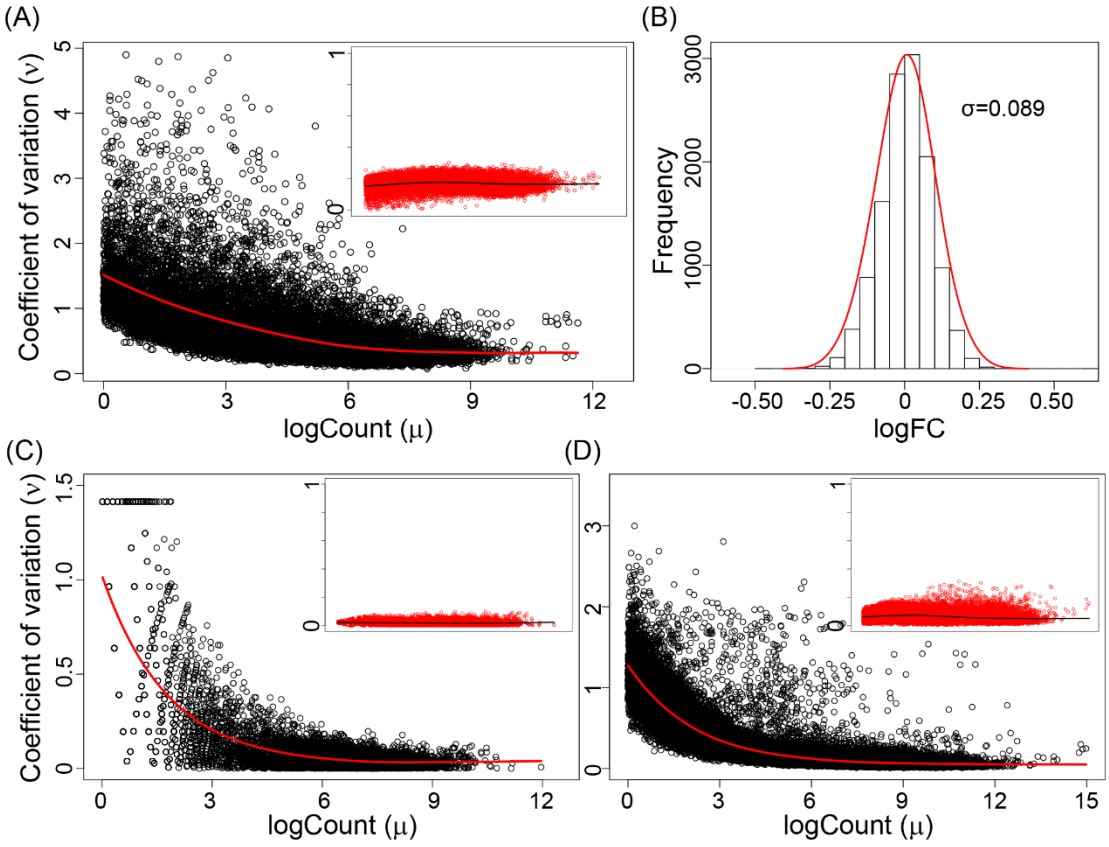


Figure 8. aFold modeling. Illustration based on the HapMap-CEU (A-B, large sample size n=24), Modencodefly (small sample size n=2, C) and ABRF (middle sample size n=9, D) data sets. (A,C-D) Mean-variance modeling and coefficient of

variation (CV) normalization. Grey horizontal line indicates the baseline of CV. Red points in the inlet show CVs after uncertainty transformation. Red line (main panel) and black line (inlet) represent the fitted value of CV via *locfit*. (B) Distribution of aFold. Red line indicates a zero-centered normal distribution with an estimated standard deviation (SD) of 0.089.

As a result, the fold change from (18) presents a robust way of measuring differential expression since it fully accounts for the mean and variance of expression values. We thus term this procedure accurate fold change (aFold) approach.

Determination of the cut-off of aFold

While the ordinary log fold changes usually follows a normal distribution with zero mean [7, 9, 46], aFold also has a zero-centered normal distribution (Figure 8B, HapMap-CEU). Therefore, the cut-off (significance threshold) of aFold can be determined by estimating the SD of the zero-centered normal distribution. Notably, the aFold calculation approach is equivalent to adding the pseudocounts ($\overline{\varepsilon_i}$) to read count. This has no influence on read count variance, but stabilizes the CV (variance stabilization, inlets in Figure 8A, 8C and 8D). Based on this procedure (i.e., adding $\overline{\varepsilon_i}$), we obtain for each data set a general CV for the count level or SD for log transformation of counts (also as the SD for aFold). We can next calculate the general SD under log transformation via moments estimation as

$$\sigma = mean(s_{\log}) / \sqrt{n-1} \tag{19}$$

σ from (19) well fits the distribution of aFold (Figure 8B, red line). Then the p-value of each aFold is generated via the normal distribution as

$$p = pnorm(lfc_i, 0, \sigma) \tag{20}$$

After an adjustment of multiple testing (i.e, Benjamini-Hochberg in default), a data-specific aFold cut-off is obtained in consideration of the significance level.

Implementation

aFold has been implemented and integrated in the software package ABSSeq for the cross-platform environment R [47]. aFold is released under the GPL-3 license as part of the Bioconductor project [30] at URL: http://bioconductor.org/packages/devel/bioc/html/ABSSeq.html.

Software tools

The figures in this study have been plotted using R.

Abbreviations

AUC: area under curve; DE: differential expression; FC: fold change; FDR: false discovery rate; FPR: false positive rate; logFC: log2 of fold change; NB: negative binomial; ROC: receiver operating characteristic; RNA-Seq: (high-throughput) sequencing of RNA; SEQC: Sequencing Quality Control; TPR: true positive rate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WY had the initial idea to the approach, designed the study, performed analyses and wrote the manuscript; PR contributed to study design and wrote the manuscript; HS supervised the study, contributed to study design, and wrote the manuscript.

Acknowledgements

We are grateful to the Rechenzentrum of the University of Kiel for access to the Linux cluster and technical support. The study was funded by the German Science Foundation within the priority program SPP1399 on host-parasite

coevolution, individual grants SCHU 1415/8 and SCHU1415/9. WY is a member of the International Max-Planck Research School (IMPRS) for Evolutionary Biology at the University of Kiel. PR is supported by BMBF DEEP TP 2.3. and EU H2020 SYSCID under the contract number 733100.

References

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature reviews genetics* 2009, **10:**57-63.

2. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.

3. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinformatics* 2010, **11:**422.

4. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

5. Li J, Witten DM, Johnstone IM, Tibshirani R: **Normalization, testing, and false discovery rate estimation for RNA-sequencing data.** *Biostatistics* 2012, **13:**523-538.

6. Srivastava S, Chen L: **A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.** *Nucleic acids research* 2010, **38:**e170-e170.

7. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**1.

8. Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome biology* 2014, **15:**1.

9. Feng J, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y: **GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data.** *Bioinformatics* 2012, **28:**2782-2788.

10. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14:**91.

11. Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C: **Detecting and correcting systematic variation in large-scale RNA sequencing data.** *Nature biotechnology* 2014, **32:**888-895.

12. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y: **Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study.** *Nature biotechnology* 2014, **32:**915-925.

13. Yang W, Rosenstiel PC, Schulenburg H: **ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences.** *BMC Genomics* 2016, **17:**541.

14. McCarthy DJ, Smyth GK: **Testing significance relative to a fold-change threshold is a TREAT.** *Bioinformatics* 2009, **25:**765-771.

15. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu T-M, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR: **Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project.** *Nature biotechnology* 2006, **24:**1140-1150.

16. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer.** *Nature genetics* 1996, **14:**457-460.

17. Consortium SM-I: **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.** *Nature biotechnology* 2014, **32:**903-914.

18. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome biology* 2010, **11:**1.

19. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**3.

20. Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome biology* 2013, **15:**1.

21. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.** *Genome biology* 2013, **14:**R95.

22. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM: **Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing.** *BMC Genomics* 2012, **13:**484.

23. Glaus P, Honkela A, Rattray M: **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics* 2012, **28:**1721-1728.

24. Zhou Y-H, Xia K, Wright FA: **A powerful and flexible approach to the analysis of RNA sequence count data.** *Bioinformatics* 2011, **27:**2672-2678.

25. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23:**2881-2887.

26. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results.** *Genome biology* 2010, **11:**1.

27. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu T-M, Goodsaid FM, Pusztai L: **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nature biotechnology* 2010, **28:**827-838.

28. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nature biotechnology* 2006, **24:**1151-1161.

29. Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY: **Evaluation of DNA microarray results with quantitative gene expression platforms.** *Nature biotechnology* 2006, **24:**1115-1122.

30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome biology* 2004, **5:**R80.

31. Van Rooij I, Broekmans F, Te Velde E, Fauser B, Bancsi L, De Jong F, Themmen A: **Serum anti-Müllerian hormone levels: a novel measure of ovarian reserve.** *Human Reproduction* 2002, **17:**3065-3071.

32. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988**:**837-845.

33. Bradley AP: **The use of the area under the ROC curve in the evaluation of machine learning algorithms.** *Pattern recognition* 1997, **30:**1145-1159.

34. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.

35. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4:**210.

36. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW: **The developmental transcriptome of Drosophila melanogaster.** *Nature* 2011, **471:**473-479.

37. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B: **Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.** *Journal of the National Cancer Institute* 2006, **98:**262-272.

38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences* 2005, **102:**15545-15550.

39. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC bioinformatics* 2010, **11:**94.

40. Frazee AC, Langmead B, Leek JT: **ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets.** *BMC Bioinformatics* 2011, **12:**449.

41. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM: **Unlocking the secrets of the genome.** *Nature* 2009, **459:**927-930.

42. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS: **Polymorphic cis-and trans-regulation of human gene expression.** *PLoS biology* 2010, **8:**e1000480.

43. Bottomly D, Walter N, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R: **Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays.** *PloS one* 2011, **6:**e17820.

44. Mesures BIdPe, internationale Cé, normalisation Oid: *Guide to the expression of uncertainty in measurement.* International Organization for Standardization; 1995.

45. Loader C: **Local Regression and Likelihood. 1999.** *NY Springer-Verlag*.

46. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays.** *Journal of the American statistical Association* 2004, **99:**909-917.

47. Team RC: **R: A language and environment for statistical computing.** *R foundation for Statistical Computing* 2005.

Additional file 1: Supplementary Methods. Overview and command lines for differential expression analysis in R.

R commands

Here, we list the R commands for each method that were used to analyze the data in this study. As explained in the main text, we run each method in default settings. All analyses were performed with R version 3.4.0. The data matrix and conditions are denoted as cdat and cgroup, respectively. In addition, we list the function for data mixture in Figure 2.

ABSSeq-aFold

aFold is integrated into the ABSSeq package, version 1.22.2, which is available in Bioconductor.

```
> library(ABSSeq)
> obj <- ABSDataSet( cdat, cgroup )
> obj <- ABSSeq(obj, useaFold=TRUE)
> res <- results(obj, c("Amean","Bmean","foldChange","pvalue","adj.pvalue"))
```

DESeq2

The DESeq2 package, version 1.16.0, can be obtained from Bioconductor.

```
> library(DESeq2)
> ds <- DESeqDataSetFromMatrix(countData = cdat, colData = data.frame(cgroup), design =
+                                                                    ~ cgroup)
> ds <- DESeq(ds)
> res <- results(ds)
```

limma-Voom

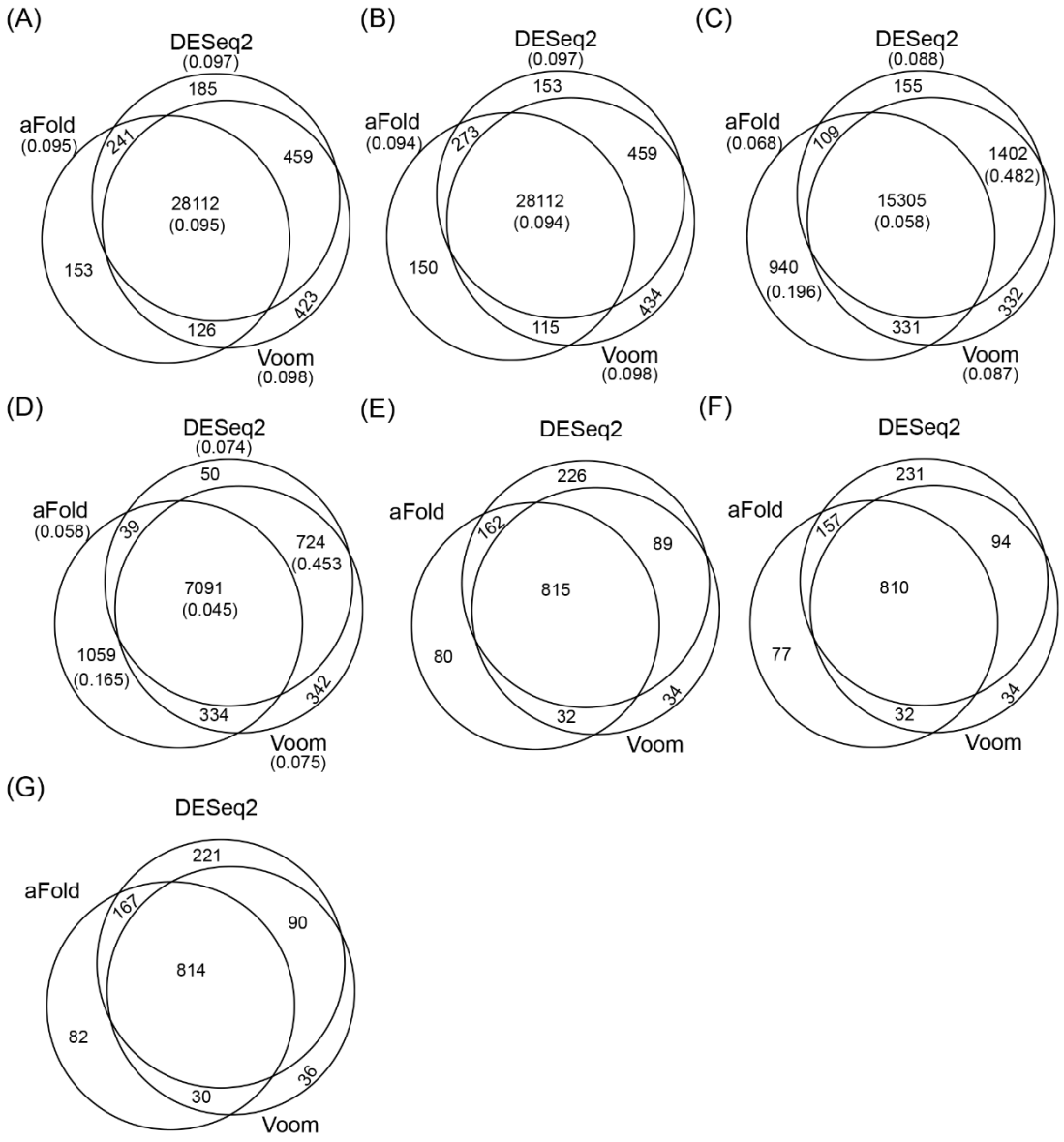
The limma package, version 3.32.0, can be obtained from Bioconductor.

```
> library(limma)
> library(edgeR)
> design <- model.matrix(~0+cgroup)
> colnames(design) <- levels(cgroup)
> nf <- calcNormFactors(cdat)
> dat <- voom(cdat, design, plot=FALSE, lib.size=colSums(cdat) * nf)
> fit <- lmFit(dat,design)
> contrast.matrix <- makeContrasts("condB - condA", levels=design)
> fit2 <- contrasts.fit(fit, contrast.matrix)
> fit2 <- eBayes(fit2)
> res=decideTests(fit2,p.value=q.cut,lfc=lfc)
> tab<-topTable(fit2, adjust = "BH", number=nrow(fit2), sort.by='logFC')
```

Mix of data

```
> mixData=function(UHR,BHR,normmethod)
> {
> sizefactor<-normmethod(cbind(UHR[,3:4],BHR[,3:4]))
> dsize <-sizefactor[1:2]/sizefactor[3:4]
> dout <- t(t(BHR[,3:4])*dsize)
> rout <- round(dout)
> dif <- dout -rout
> rout[dif>0.5] <- rout[dif>0.5]+1
> out <-rbind(as.matrix(UHR),as.matrix(rout))
> rownames(out)=1:nrow(out)
> return(out)
> }
```

Additional file 2: Supplementary Figure. Venn diagrams show number of DE identified by aFold, DESeq2 and Voom on four data sets.



Supp. Figure 1. Venn diagram. Venn diagrams show number of DE identified by aFold, DESeq2 and Voom on four data sets: (A-B) the ABRF data set; (C) the SEQC data set; (D) the MAQC-II data set; (E-G) the Bottomly data set. aFold detects DE under three normalization procedures: qtotal (C-E), TMM (A and F) and geometric mean (B and G). Numbers in brackets indicate the eFDR.

Chapter III

WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis

Wentao Yang[1], Katja Dierking[1] and Hinrich Schulenburg[1,*]

[1]Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany.

* Correspondence: hschulenburg@zoologie.uni-kiel.de

Application notes

WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis

Wentao Yang[1], Katja Dierking[1] and Hinrich Schulenburg[1,*]

[1]Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany.

ABSTRACT

Motivation: A particular challenge of the current omics age is to make sense of the inferred differential expression of genes and proteins. The most common approach is to perform a gene ontology (GO) enrichment analysis, thereby relying on a database that has been extracted from a variety of different organisms and that can therefore only yield reliable information on evolutionary conserved functions.

Results: We here present a web-based application for a taxon-specific gene set exploration and enrichment analysis, which is expected to yield novel functional insights into newly determined gene sets. The approach is based on the complete collection of curated high-throughput gene expression data sets for the model nematode *Caenorhabditis elegans*, including 1786 gene sets from more than 350 studies.

Availability and implementation: WormExp is available at http://wormexp.zoologie.uni-kiel.de.

Contacts: wyang@zoologie.uni-kiel.de, kdierking@zoologie.uni-kiel.de, or hschulenburg@zoologie.uni-kiel.de

Supplementary information: available at *Bioinformatics* online.

1 INTRODUCTION

High-throughput molecular technologies have greatly enhanced our understanding of biological processes by characterizing expression changes of genes (microarray and RNA-Seq data) and proteins (proteomics data) or transcription factor targets and epigenetics states (ChIP-chip and ChIP-Seq data). These technologies usually yield hundreds or thousands of differentially regulated genes or proteins that are not always easy to interpret. Validation of the numerous differentially expressed genes is usually not possible. Uncovering the underlying organizational principles from such large gene lists requires computational and statistical approaches as well as precise biological reference information.

Gene set enrichment analysis represents a powerful tool to link the identified differentially expressed gene lists to biological processes and functions. They are based on the statistical evaluation of the overlap between the generated gene set and a specified reference list of genes. These enrichment analyses are usually based on public databases such as those defined by Gene ontology (Ashburner, et al., 2000) and KEGG pathways (Kanehisa and Goto, 2000). However, these existing databases have important drawbacks. First, the annotations are incomplete and only a subset of known genes are functionally annotated (King, et al., 2003). For example, functional information is only available for approximately 60% of the gene repertoire of the nematode *Caenorhabditis elegans*, one of the most intensively studied model organisms in biological research (Petersen, et al., 2015). Second, the included functional information is often imprecise, as it usually represents an extrapolation from experimental data of a different taxon and thus assumes a high level of functional conservation across evolution, which may not always be the case (Khatri and Drăghici, 2005). Third, functional information is predicted for most organisms from protein domains. Taxon-specific genes or protein domains may thus be missed. Taxon-specific gene sets, which explicitly consider taxon-restricted genes and also taxon-specific expression responses, are thus required for improved functional genomic analyses. Several applications such as GSEA (Subramanian, et al., 2005) and EASE (Hosack, et al., 2003) have been developed to permit performance of enrichment analyses with curated gene sets, derived for example from published expression studies in the same organism. Yet, a systematic assessment of the value of taxon-specific enrichment analyses is still missing. WormExp provides a web-resource to explore such an approach for the nematode *Caenorhabditis elegans*.

This nematode has a well annotated genome sequence and is widely used as a powerful model organism in biological research. Over the last decade, more than 350 high-throughput expression studies have been published, covering a large variety of research themes, such as immunity, aging, development, and stress responses. The resulting lists of differentially expressed genes are publicly available and can be related to a specific experimental design, environmental condition, and/or gene defect. Because they capture a variety of inducible expression responses of this particular organism, they might be highly useful in interpreting new *C. elegans* gene lists (Engelmann, et al., 2011; Yang, et al., 2015) or predicting candidates for downstream analysis (Block, et al., 2015). In this manuscript, we present WormExp, a web-based application for gene set enrichment analysis in *C. elegans*. We collated nearly all published high-throughput expression data sets of *C. elegans* from public databases and also the available literature. We classified these gene sets in nine categories according to the experimental designs or specific condition used. WormExp accepts Wormbase (Harris, et al., 2010) identifiers (IDs), sequence names, gene names or a mixture of these as input. It offers tools for performing an enrichment analysis based on the taxon-specific 1786 gene sets, searching specific gene lists, and downloading complete data sets.

*To whom correspondence should be addressed.

2 METHODS AND FEATURES

WormExp utilizes a curated database built from published high-throughput expression studies in the nematode *C. elegans*. The database can be downloaded and will be updated continuously. Users start their analysis by uploading a gene list, for example a set of genes, whose expression is induced upon *C. elegans* exposure to a certain condition. The user then has two options (Fig. 1 and workflow in manual): (i) perform an enrichment analysis using either the entire database as reference or selected categories of gene lists (e.g. "mutants"), or (ii) search for overlaps between the uploaded gene set and specific gene lists (e.g. "up *pmk-1* mutant"), selected from the database with the help of keywords (e.g. "pmk-1"). For enrichment analysis, WormExp employs the adjusted Fisher exact test from the program EASE, which penalizes or removes one gene within a given gene set from the test list and calculates the p-value (for details on the tail distribution of jackknife Fisher exact probabilities, see manual in supplementary file). This modulation makes the Fisher exact test more robust when applied on gene sets supported by few genes, thus reducing false positives (for details, please see (Hosack, et al., 2003)). WormExp is available as a webserver with an interface from InterMine (Smith, et al., 2012), developed by Java 2 Enterprise System (J2EE) and Java Remote Method Invocations (RMI). RMI ensures fast responses due to memory-oriented query. See supplementary files 1 and 2 for manual and detailed information.

Fig.1 Homepage of WormExp

2.1 Data Structure

Currently, the WormExp database includes 1786 gene sets derived from 361 studies and collated from the NCBI GEO database (Barrett, et al., 2009), ArrayExpress database (Brazma, et al., 2003), Stanford microarray database (Sherlock, et al., 2001), Princeton University MicroArray database, and directly from the literature (Supplementary Data 1). According to experimental design and used conditions, we classified these 1786 gene sets into nine categories: Kim Mountains ((Kim, et al., 2001)); Mutants (differentially expressed genes in mutants or upon RNA interference-silencing of a particular gene); Microbes (exposure to various microorganisms), TF Targets (transcription factor targets inferred by knock-down/knock-out of the respective transcription factors), Tissue (tissue specific expression), Development/Dauer/Aging (differential expression in the various developmental stages and during aging), DAF/Insulin/food (differential expression in response to food, starvation, or insulin-like receptor activation/de-activation), Chemicals/stress (exposure to chemical compounds or other stressors), and Other (all gene sets not included above). The database will be updated regularly to integrate new *C. elegans* expression studies.

ACKNOWLEDGEMENTS

We thank the Schulenburg lab for advice and the Kiel University computer center, especially S. Lorenz and U. Schwarz, for support.

Funding: The work was funded by grants of the German Science foundation to HS (DFG grants SCHU 1415/8 and SCHU 1415/9). WY is additionally supported by the International Max-Planck Research School for Evolutionary Biology.

REFERENCES

Ashburner, M., et al. (2000) Gene Ontology: tool for the unification of biology, Nature genetics, 25, 25-29.

Barrett, T., et al. (2009) NCBI GEO: archive for high-throughput functional genomic data, Nucleic acids research, 37, D885-D890.

Block, D.H., et al. (2015) The Developmental Intestinal Regulator ELT-2 Controls p38-Dependent Immune Responses in Adult C. elegans.

Brazma, A., et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI, Nucleic acids research, 31, 68-71.

Engelmann, I., et al. (2011) A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans, PloS one, 6, e19055.

Harris, T.W., et al. (2010) WormBase: a comprehensive resource for nematode research, Nucleic acids research, 38, D463-D467.

Hosack, D.A., et al. (2003) Identifying biological themes within lists of genes with EASE, Genome Biol, 4, R70.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, Nucleic acids research, 28, 27-30.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics, 21, 3587-3595.

Kim, S.K., et al. (2001) A gene expression map for Caenorhabditis elegans, Science, 293, 2087-2092.

King, O.D., et al. (2003) Predicting gene function from patterns of annotation, Genome research, 13, 896-904.

Petersen, C., Dirksen, P. and Schulenburg, H. (2015) Why we need more ecology for genetic models such as C. elegans, Trends in Genetics, 31, 120-127.

Sherlock, G., et al. (2001) The stanford microarray database, Nucleic Acids Research, 29, 152-155.

Smith, R.N., et al. (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data, Bioinformatics, 28, 3163-3165.

Subramanian, A., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

Yang, W., et al. (2015) Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode Caenorhabditis elegans toward pathogenic Bacillus thuringiensis, Developmental & Comparative Immunology, 51, 1-9.

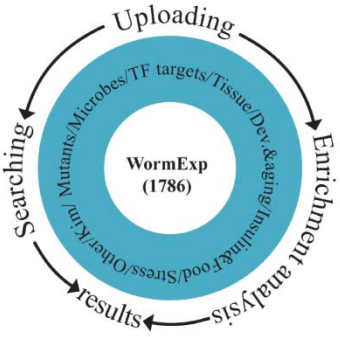
Supplementary files S2: Manual for WormExp

Manual for WormExp

Aims + Workflow

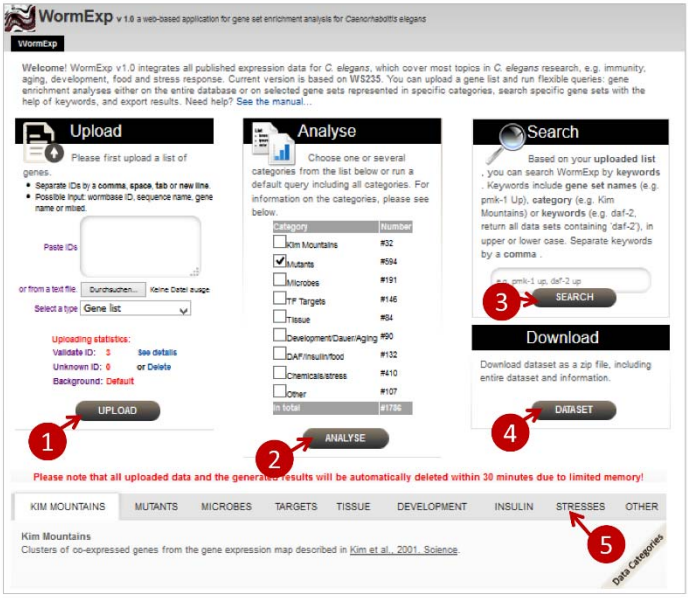
Aims:
WormExp integrates all published expression data for *C. elegans*. It allows performance of taxon-specific enrichment analyses and searches, which may help to identify functions that are taxon-specific and can thus not be captured with GO term or KEGG pathway analyses.

Workflow:
Users start their analysis by uploading a gene list, for example a set of genes, whose expression is induced upon *C. elegans* exposure to a certain condition. The user then has two options: (i) perform an enrichment analysis using either the entire database as reference or selected categories of gene lists (e.g. "mutants"), or (ii) search for overlaps between the uploaded gene set and specific gene lists (e.g. "up *pmk-1* mutant"), selected from the database with the help of keywords (e.g. "pmk-1"). The results of either approach is given in a new window and can be downloaded in table format, including list of the supporting genes.



Text in circle indicates categories of WormExp.

Homepage and Overview



Main sections:
(see red numbers and arrows)

1. **Upload**: uploading list of *C. elegans* gene identifiers (e.g., *clec-4*, Y38E10A.5, WBGene00012583)
2. **Analysis**: Use the uploaded list and select categories to run the enrichment analysis.
3. **Search**: search the database by uploading list of genes or keywords
4. **Download**: download the entire dataset and information
5. **Data categories**: information about the data categories.

Please note that any uploaded data set or any generated result is deleted after 30 min because of memory limitation.

Methods

For the **statistical evaluation during enrichment analysis**, we implemented the method developed for the **program EASE** (Hosack et al. 2003).

Assume a test list with 192 differentially regulated genes, a curated gene set with 2551 genes (e.g., from a previous study), the population size (background of all considered genes) is 27770, and then 34 out of the 192 genes from the test list are present in the curated gene set . Then the 2x2 contigency table for the Fisher Exact test is:

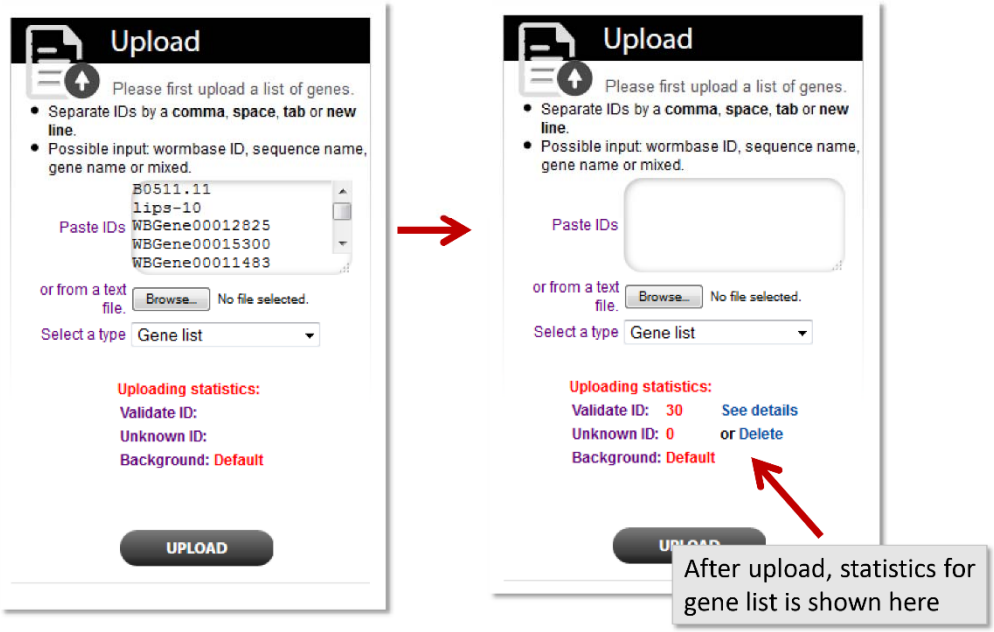
User case	Background: present	Background: absent	Background: total
present	33(34-1)	192-33	192
absent	2551-33	27770-2551-192+33	27770-192
total	2551	27770-2551	27770

User	Background	
	33	159
	2518	25060

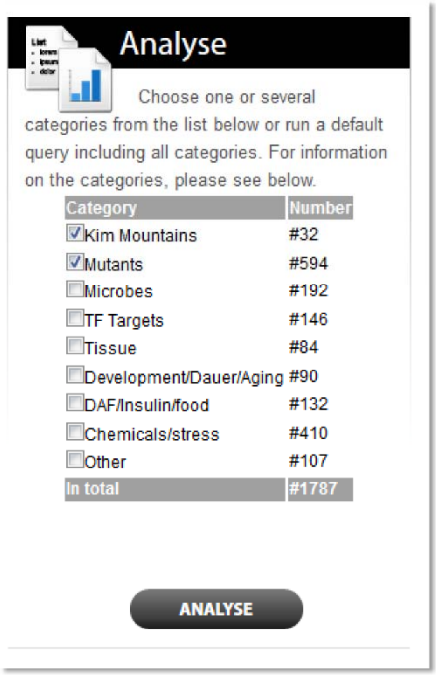
The one-tailed (greater than 33) p-value is 0.0003292. Detailed information on the statistical approach can be found in the description of the program EASE: Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA: Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003, 4(10):R70.

First step: uploading list of genes

A list of genes (e.g., a list of upregulated genes after cadmium exposure) can be uploaded either as a file (see option “Browse”) or by pasting it into the open field. It is possible to upload a query list or a new reference/background gene list.

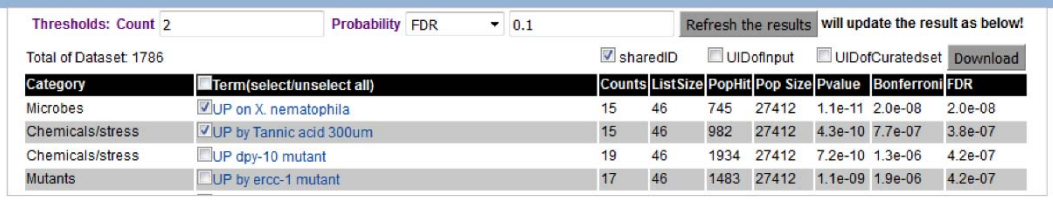


Enrichment analysis



An enrichment analysis can be performed with uploaded test list of genes. To start the enrichment analysis, select categories of the curated gene lists or use all of them (none selected – by default). Then click on the "Analyse" button. The results are shown in a separate window or tab.

Results for enrichment analysis



Category	Term(select/unselect all)	Counts	ListSize	PopHit	Pop Size	Pvalue	Bonferroni	FDR
Microbes	UP on X. nematophila	15	46	745	27412	1.1e-11	2.0e-08	2.0e-08
Chemicals/stress	UP by Tannic acid 300um	15	46	982	27412	4.3e-10	7.7e-07	3.8e-07
Chemicals/stress	UP dpy-10 mutant	19	46	1934	27412	7.2e-10	1.3e-06	4.2e-07
Mutants	UP by ercc-1 mutant	17	46	1483	27412	1.1e-09	1.9e-06	4.2e-07

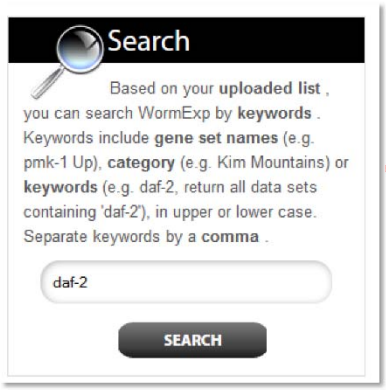
The results for the enrichment analysis are shown as a table in a new tab or window. The table includes categories (Category), name of the curated reference gene set (Term), the number of overlapping genes between the test list and the curated gene set (Counts), the total number of genes in the uploaded test list (ListSize), the number of genes in the respective curated gene set (PopHit), background size, that is the total number of genes in the entire considered data base (Pop Size), the inferred p-value before adjustment (Pvalue) and two adjusted p-values (Bonferroni and FDR).

The link underlying the name of the curated gene set leads to more detailed information on the genes contained in this gene set. The user can set thresholds on the minimum number of overlapping genes (Thresholds: Count) or on the probability (Probability; either p-value, bonferroni or FDR).

The results can be downloaded as a table with additional options, including the list of overlapping genes between input and curated list (shared ID; default; options on top of table), or the genes unique for input or for curated gene set (ID of input or ID of curated set) as well as selected curated gene sets (options in second column). Please note that file size may become very large if all three options are chosen.

Search the database

Based on the uploaded list of genes, the user can search the database for an overlap between the gene list and particular keywords for the curated gene sets. An example for such a keyword is *daf-2*, which will assess all curated gene sets related to this genes, for example transcriptome studies on *daf-2* mutants.



Use *daf-2* as an example and click "Search"

Total of Dataset: 59

GeneID	Down xbp-1 mutant on daf-2(e1368)	Down xbp-1 mutant on daf-2(e1370)	UP by daf-2;rsks-1 mutant	down by daf-2;rsks-1 mutant
B0511.11	0	0	1	0
lips-10	0	0	0	0
WBGene00012825	0	0	0	0
WBGene00015300	0	0	0	0
WBGene00011483	0	0	0	0
WBGene00023069	0	0	0	0
WBGene00044611	0	0	0	0
WBGene00020066	0	0	1	0
WBGene00044333	0	0	0	0
WBGene00022763	0	0	0	0
WBGene00019436	0	0	1	0
C04E6.4	0	0	0	0
dct-8	0	0	0	1
B0507.10	0	0	1	0
fbxa-224	0	0	0	0

The returned results table shows all curated gene sets related to *daf-2*. The rows refer to the genes from the test gene list, whereas columns show the curated gene sets related to *daf-2*. 1 and 0 indicate presence or absence of a gene from the list in the curated gene sets. The results table can be downloaded.

Download datasets

The entire database and further information are available for user download.

Note: the supplementary tables are too large to be attached here and they are available at https://doi.org/10.1093/bioinformatics/btv667

Chapter IV

Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* towards pathogenic *Bacillus thuringiensis*

Wentao Yang[1]*, Katja Dierking[1]*, Daniela Esser[2], Andreas Tholey[3], Matthias Leippe[4], Philip Rosenstiel[2] and Hinrich Schulenburg[1]

[1] Evolutionary Ecology and Genetics, Zoological Institute, Christian-Albrechts University of Kiel, Germany

[2] Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Germany

[3] Systematic Proteome Research and Bioanalytics, Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Germany

[4] Comparative Immunology, Zoological Institute, Christian-Albrechts University of Kiel, Germany

* These authors contributed equally to this work

Corresponding author: Hinrich Schulenburg, Evolutionary Ecology and Genetics, Zoological Institute, Christian-Albrechts University of Kiel, 24098 Kiel, Germany; Tel: +49-431-8804143; Fax: +49-431-8802403; Email: hschulenburg@zoologie.uni-kiel.de

Developmental and Comparative Immunology 51 (2015) 1–9

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* toward pathogenic *Bacillus thuringiensis*

Wentao Yang [a,1], Katja Dierking [a,1], Daniela Esser [b], Andreas Tholey [c], Matthias Leippe [d], Philip Rosenstiel [b], Hinrich Schulenburg [a,*]

[a] Evolutionary Ecology and Genetics, Zoological Institute, Christian-Albrechts University of Kiel, Germany
[b] Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Germany
[c] Systematic Proteome Research and Bioanalytics, Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Germany
[d] Comparative Immunology, Zoological Institute, Christian-Albrechts University of Kiel, Germany

ARTICLE INFO

Article history:
Received 28 January 2015
Revised 10 February 2015
Accepted 11 February 2015
Available online 23 February 2015

Keywords:
Innate immunity
C-type lectins
Proteomics
RNA-Seq
Caenorhabditis elegans
Bacillus thuringiensis

ABSTRACT

Pathogen infection can activate multiple signaling cascades that ultimately alter the abundance of molecules in cells. This change can be measured both at the transcript and protein level. Studies analyzing the immune response at both levels are, however, rare. Here, we compare transcriptome and proteome data generated after infection of the nematode and model organism *Caenorhabditis elegans* with the Gram-positive pathogen *Bacillus thuringiensis*. Our analysis revealed a high overlap between abundance changes of corresponding transcripts and gene products, especially for genes encoding C-type lectin domain-containing proteins, indicating their particular role in worm immunity. We additionally identified a unique signature at the proteome level, suggesting that the *C. elegans* response to infection is shaped by changes beyond transcription. Such effects appear to be influenced by AMP-activated protein kinases (AMPKs), which may thus represent previously unknown regulators of *C. elegans* immune defense.

1. Introduction

The innate immune system acts as an early line of defense against microbial infection in all animals. It relies on recognition of microbe- or damage-associated molecules by so called pattern recognition receptors (PRRs), followed by activation of multiple signaling cascades that ultimately alter the abundance of molecules in cells. The study of these changes by "omics" technologies has greatly enhanced our understanding of the molecular mechanisms that determine defense responses against potential pathogens. The classical examples include gene expression analysis using microarrays. Already 13 years ago, the study of transcriptional changes caused by colonization of germ-free mice with *Bacteroides thetaiotaomicron* revealed a wide impact of the bacteria on fundamental intestinal functions (Hooper et al., 2001) and led to the discovery of angiogenins as a family of endogenous antimicrobial proteins (Hooper et al., 2003). Similarly, the early microarray studies of the *Drosophila* immune response demonstrated the central regulatory role of the Toll and Imd signaling cascades (Boutros et al., 2002; De Gregorio et al., 2001, 2002; Irving et al., 2001). In addition to transcriptional responses, pathogen infection also causes changes at the protein level. These can be a direct consequence of transcription and translation. However, many proteins additionally undergo post-translational modification and/or are at some time point degraded (Chen et al., 1995; Izzi and Attisano, 2004; Mukhopadhyay and Riezman, 2007). Ten years ago, the first proteome analysis of the *D. melanogaster* immune response allowed identification of previously unrecognized putative immune response factors (Vierstraete et al., 2004). Likewise, early proteome analysis of human immunity newly identified cellular targets of viral immune modulators (Bartee et al., 2006). Surprisingly, transcriptome and proteome analyses undertaken in the same study systems often report only little

Abbreviations: AMP, adenosine monophosphate; AMPK, AMP-activated protein kinase; BT, *Bacillus thuringiensis*; CTLD, C-type lectin domain-containing proteins; CREB, cyclic AMP-response element binding protein; CRD, carbohydrate recognition domain; FDR, false discovery rate; GO, gene ontology; ILR, insulin-like receptor; iTRAQ, isobaric tags for relative and absolute quantitation; LC-MS/MS, liquid chromatography- tandem mass spectrometry; MAPK, mitogen-activated protein kinase; NGM, nematode growth medium; PBS, phosphate buffered saline; PFM, peptone-free NGM; PRR, pattern recognition receptors; RNAi, RNA interference; RNA-Seq, RNA sequencing; UPR, unfolded protein response.

* Corresponding author. Evolutionary Ecology and Genetics, Zoological Institute, Christian-Albrechts University of Kiel, 24098 Kiel, Germany. Tel: +49 431 8804143; fax: +49 431 8802403.
E-mail address: hschulenburg@zoologie.uni-kiel.de (H. Schulenburg).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.dci.2015.02.010

congruence among the differentially regulated genes and proteins (Ghazalpour et al., 2011; Gygi et al., 1999; Ragno et al., 2001; Scherl et al., 2006; Vogel and Marcotte, 2012), suggesting either technological biases and/or a biologically relevant difference. Hence, further work is essential to characterize the relationship between the transcriptomic and proteomic signatures of immune defense.

The nematode *Caenorhabditis elegans* is widely used as a powerful model for studying innate immunity, especially because of its amenability to genetic manipulation, its susceptibility to a number of human and animal pathogens, and its conserved immune signaling pathways (Kurz and Ewbank, 2003; Marsh and May, 2012). Similarly to other model systems, transcriptome analyses have become a central first step for elucidating pathogen-induced responses in *C. elegans*. For example, the importance of the DBL-1/TGFβ pathway in *C. elegans* immunity was uncovered through a microarray gene expression analysis upon *Serratia marcescens* infection (Mallo et al., 2002). Moreover, NLP-29 and NLP-31 were identified as members of a new family of antimicrobial peptides via microarray analysis after infection of *C. elegans* with the fungus *Drechmeria coniospora* (Couillault et al., 2004). Transcriptomic analysis of the *C. elegans* response to *Pseudomonas aeruginosa* infection revealed a conserved role of the GATA transcription factor ELT-2 in regulating epithelial innate immune responses (Shapira et al., 2006). In contrast to the numerous transcriptome analyses (more than 30), only a few studies characterized the *C. elegans* immune response at the protein level (Bogaerts et al., 2010; Couillault et al., 2012; Simonsen et al., 2011; Treitz et al., 2015; Ziegler et al., 2009). None of the current proteome studies included a comparison with the corresponding transcriptomic response to pathogens.

In this study, we compared newly generated transcriptome results with recently published quantitative proteome data (Treitz et al., 2015) generated after infection of *C.elegans* with the Gram-positive pathogen *Bacillus thuringiensis* (BT). Based on this comparison, our aim was to characterize the functions that associate with changes of transcript and protein abundances upon pathogen infection. For our approach, we used the same worm and pathogen strains, including the standard *C. elegans* laboratory strain N2 and the nematocidal BT strain MYBT18247. Previous functional genetic characterizations and microarray-based gene expression analyses revealed that the nematode's defense against this particular BT strain is influenced by the insulin-like signaling cascade and lysozymes (Boehnisch et al., 2011; Hasshoff et al., 2007; Wang et al., 2012). Moreover, defense against the Cry5B toxin expressed by another nematocidal BT strain relies centrally on glycolipid receptors in the intestine and signaling by the JNK MAPK pathway (Griffitts et al., 2005; Kao et al., 2011). We here re-assessed the nematode's transcriptomic response to MYBT18247 using RNA-Seq. The results were compared to recently published quantitative proteomics data that were acquired by isobaric labeling (iTRAQ) in combination with LC-MS/MS analysis (Treitz et al., 2015). Using these data sets, we asked which functions are enriched in the overlap between proteomic and corresponding transcriptomic response, and which ones in the unique signature of the proteome data.

2. Materials and methods

2.1. C. elegans *and bacteria strains*

The *C. elegans* Bristol N2 strain was used in all experiments. Worms were generally maintained at 20 °C on nematode growth medium (NGM) inoculated with the *Escherichia coli* strain OP50, following standard procedures (Stiernagle, 2006). For the infection experiments, worms were exposed to the nematocidal BT strain MYBT18247. As controls, we used the non-nematocidal BT strains DSM 350 (proteome study) and 407 cry- (transcriptome study) with equivalent avirulent effects in *C. elegans*.

2.2. Proteomics

Proteins were extracted 12 h after initial exposure for three replicates per treatment, followed by tryptic digestion, isobaric labeling and analysis by 2D-LC ESI MS/MS (Treitz et al., 2015). Quantitative results of $\log_2$ transformed iTRAQ-protein ratios describing the condition pathogenic/non-pathogenic BT were tested against all $\log_2$ transformed control-ratios of biological replicates in the same iTRAQ experiment using the two sided Welch's t-test. To adjust for multiple testing, p-values were corrected by permutation based FDR approach with 1000 randomizations. Protein groups with $\log_2$ ratios of at least ±0.485 (which corresponds to iTRAQ-ratios ≥ 1.4 or ≤ 0.71) with FDR-corrected p-value = q ≤ 0.05 were considered to be differentially expressed. See Treitz et al. (2015) for more detailed information.

2.3. Transcriptomic analysis by RNA-Seq

Spore-toxin mixtures of MYBT18247 (1.6×10^{10} particles/ml) and spore cultures of 407 cry- (1.8×10^{10} particles/ml) were mixed at a ratio of 1:2 and 1:10 with *E. coli* OP50, which was grown in LB at 37 °C overnight and suspended in PBS with a final OD of 5. 250 µl of the *B. thuringiensis-E. coli* mixture, or *E. coli* alone was then pipetted onto the center of a 9 cm peptone-free nematode growth medium (PFM) plate. The inoculated PFM plates were left overnight at 20 °C to dry. On the day of the infection, worms synchronized at the fourth instar larval (L4) stage were washed off NGM plates with phosphate buffered saline (PBS), added to the assay plates by pipetting, and incubated at 20 °C. Three replicates were used per treatment. At the respective time points (6 h and 12 h after initial exposure), worms were washed off the assay plates with PBS containing 0.3% Tween20®, and subsequently centrifuged. The worm pellet was resuspended in 800 µl TRIzol® (Life Technologies) reagent and worms were broken up prior to RNA extraction by treating the worm suspension five times with a freeze-and-thaw cycle using liquid nitrogen and a thermo block at 45 °C. RNA was extracted using a NucleoSpin® miRNA extraction kit (Macherey-Nagel), treated with DNAse, and stored at −80 °C. RNA libraries were prepared for sequencing using standard Illumina protocols. Libraries were sequenced on an Illumina HiSeq™ 2000 sequencing machine with paired-end strategy at read length of 100 nucleotides. The raw data are available from the GEO database (Barrett et al., 2013; Edgar et al., 2002) under the GSE number GSE64401.

RNA-Seq reads were mapped to the *C. elegans* genome (Wormbase version WS235; www.wormbase.org) by Tophat2 (Kim et al., 2013) using option *–b2-very-sensitive*, other default settings, and without a transcriptome reference. Tophat2 aligns RNA-Seq reads to a genome based on the ultra-fast short read mapping program Bowtie (Langmead et al., 2009). It can find splice junctions without a reference annotation. Transcript abundance was estimated by Cufflinks (Trapnell et al., 2013) guided by WBCell235 from EMBL and using the following options: *–library-norm-method*, *–multi-read-correct* and *–frag-bias-correct*, and significantly differentially expressed genes were identified by Cuffdiff (Trapnell et al., 2013) using an additional parameter: *–library-norm-method quartile* (Robinson and Oshlack, 2010). Cuffdiff is a program from the Cufflinks package and aims at finding significant changes in transcript expression in consideration of all possible annotated isoforms for a particular gene. Transcripts with a significant change between different conditions (adjusted p-value < 0.01) were treated as a signature for each comparison (pathogenic/non-pathogenic BT, 6 h and 12 h). The $\log_2$ transformed fold-changes (pathogenic/non-pathogenic BT) were taken as input for k-means cluster analysis using cluster 3.0 (de Hoon et al., 2004) with 4 initial clusters. A heat map was generated by TreeView version 1.1.4r3 (Saldanha, 2004).

2.4. Gene ontology and gene set enrichment analysis

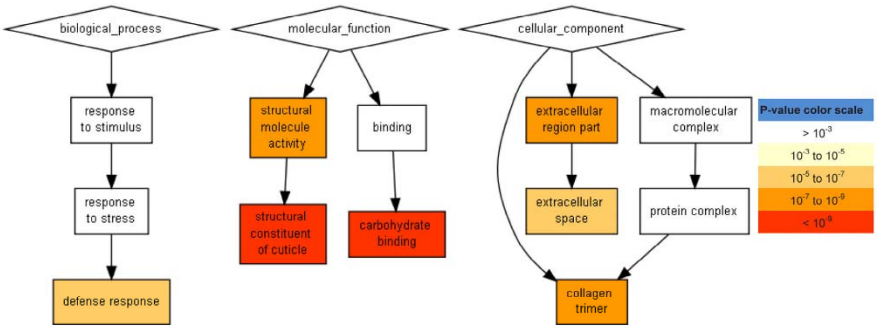
Gene ontology (GO) analysis was performed using GOrilla (Eden et al., 2009) with a cutoff of FDR < 0.01. We additionally carried out a gene set enrichment analysis, a widely used approach for assessing the significant overlap among gene lists, here the list of significantly differentially expressed genes from our study and those from comparable published transcriptome data sets (Sherman and Lempicki, 2009; Subramanian et al., 2005). This analysis was performed with the approach implemented in the program EASE (Hosack et al., 2003), a free, stand-alone software package from DAVID bioinformatics resources (http://david.abcc.ncifcrf.gov/). EASE measures the significance of an overlap between two gene lists by the one-tailed Fisher exact probability, which is calculated using Gaussian hypergeometric probability distribution. As a reference, we constructed a gene expression database from previously published *C. elegans* transcriptome studies, which is thus more specific to our model taxon than the conventionally used GO term reference base and should thus enhance the identification of significantly enriched biological functions of direct relevance for this nematode. For this database, we used the compilation of Engelmann et al. (2011) as a starting point and added a large number of other published transcriptomic data sets for *C. elegans* (see references in Table S4). The resulting database included a total of 590 data sets that cover a variety of conditions such as exposure to pathogens and other stressors, starvation, different developmental stages, and also analyses of mutants and RNAi-treated worms. Based on this database, we performed the EASE analysis using an adjustment of the critical significance threshold through the Bonferroni approach in order to correct for multiple statistical testing (Hosack et al., 2003).

3. Results

3.1. Concordances and differences between protein and mRNA abundances in the C. elegans response to BT infection

The quantitative proteome analysis yielded a total of 288 out of more than 3,600 quantified proteins with a significant change in abundance. Of these, 171 had higher and 117 had lower abundance. GOrilla analysis identified several enriched GO terms for these proteins (Fig. 1, Table S1). The most significantly enriched GO term is "structural constituent of cuticle" (FDR = 1.26E-17), which suggests that the structure of the cuticle is modified in worms challenged with BT. The second most significant GO term is "carbohydrate binding" (FDR = 4.48E-7), which is caused by an enrichment of differentially abundant galectins and C-type lectins (see also below). We also observed an enrichment of the GO terms "defense response" (FDR = 2.30E-3), which is based on differential yields of the MAP kinase kinase MEK-1, the lysozymes LYS-1 and LYS-2, the galectin LEC-8, and the C-type lectins CLEC-41 and CLEC-62. Interestingly, the GO terms "extracellular region" and "extracellular space" were also enriched, indicating that a set of secreted proteins may contribute to the defense against BT infection.

The complementary enrichment analysis with EASE (Table 1, see supplementary Table S4 for reference information) revealed



GO term	Description	P-value	FDR q-value	Count	Enrichment
GO:0042302	structural constituent of cuticle	9.97E-21	1.26E-17	26	8.71
GO:0030246	carbohydrate binding	7.10E-10	4.48E-07	18	5.36
GO:0005198	structural molecule activity	2.54E-08	1.07E-05	37	2.57
GO:0044421	extracellular region part	3.93E-08	2.27E-05	17	4.55
GO:0005581	collagen trimer	5.32E-08	1.54E-05	10	7.88
GO:0005615	extracellular space	5.69E-07	1.10E-04	12	5.36
GO:0006952	defense response	7.87E-07	2.30E-03	17	3.8

Fig. 1. GO term enrichment analysis of proteins with significant abundance changes. A total number of 288 unranked proteins with significant abundance changes in *C. elegans* after infection with MYBT18247, were taken as input for the GOrilla web-server. Top panel: Functional enrichment map; Bottom panel: a table of specific GO terms including additional information about number of over-represented genes, enrichment score as well as adjusted pvalue (FDR). GO terms with a p-value < 1.0e-5 (FDR < 1.0e-2) are shown.

significant over-representation of gene sets, which were previously identified to be differentially transcribed upon *C. elegans* exposure to the BT Cry5B toxin, the heavy metal cadmium (Huffman et al., 2004), another nematocidal BT strain DB27 (Sinha et al., 2012), and the BT strain NRRL B-18247. The latter is the ancestor of the here used MYBT18247 strain, toward which the transcriptomic response was previously assessed in the three different worm strains, N2, MY15, and MY18, using microarrays (Boehnisch et al., 2011). Moreover, for the more abundant proteins, we observed significant over-representation of genes influenced by the DAF-2 insulin-like receptor (ILR) signaling pathway (Halaschek-Wiener et al., 2005; McElwee et al., 2003; Murphy et al., 2003). This is consistent with previous results on the role of the DAF-2/ILR pathway in defense against BT (Hasshoff et al., 2007; Wang et al., 2012). Taken together, these results demonstrate high concordance between the proteomic data set and the results of previous transcriptomic analyses on the *C. elegans* immune response to BT infection.

However, at a closer look we also found discrepancies between proteomic and previous transcriptomic data sets. For example, only 21 out of the 171 and 5 out of the 117 proteins with higher and lower amounts, respectively, after BT-challenge were also found in the up-regulated or down-regulated gene groups of our previous microarray-based gene transcription analysis of N2 worms exposed to NRRL B-18247 (Boehnisch et al., 2011) (Table 1). These differences between the two studies might be due to different experimental procedures. To improve comparability of the transcriptome and proteome data sets, we here performed a BT infection experiment with similar experimental conditions as in the study of Treitz et al. (2015), using the same worm and pathogenic BT strains and also the same analysis time point (12 h after initial exposure) for RNA extraction. Two different dilutions of BT with *E. coli* (1:2 and 1:10) were used, and another time point (6 h after initial exposure) was added to evaluate the effect of pathogenicity and progress of infection. Moreover, transcriptomic variation was assessed by RNA-Seq, a method which is usually assumed to provide a more precise measurement of transcript levels than the previously used microarray approach (Wang et al., 2009).

In contrast to the comparison with the previous microarray study (Boehnisch et al., 2011), our new transcriptomic data yielded more than 35% of differentially transcribed genes that were also observed with different abundance at the protein level for all BT concentrations and time points (Fig. 2A, Table S2), suggesting that RNA-Seq indeed improves sensitivity. Fifty-eight of the genes are up-regulated and 38 are down-regulated. More than 40% of the proteins with consistent abundance changes fall into the top 400 highest differentially transcribed genes (Fig. 2B), indicating that the most pronounced changes in mRNA abundance are likely to translate into differential protein amounts. The heatmap in Fig. 2C (see supplementary Table S2 for further details) highlights the partial correlation between protein and mRNA abundances (r = 0.51, 0.51, 0.40, and 0.42 for 6 h and 12 h and the 1:2 and 1:10 dilutions, respectively, by Pearson's correlation in R (Team, 2012)). Interestingly, the correlations between protein abundances at 12 h after exposure and mRNA abundances at 6 h and 12 h after exposure in our study are nearly the same. Enrichment analysis using GOrilla on these overlapping gene sets showed that the GO terms "carbohydrate binding", "defense response", and "extracellular region" were enriched only for the genes and proteins with higher abundance, while the GO term "metabolic process" was enriched for those with lower abundance (see Table S3 for details). In contrast to the GO analysis results for all proteins with abundance changes (Fig. 1), the GO term "structural constituent of cuticle" was not significantly enriched, indicating that the change in cuticle structure is subject to post-transcriptional regulation.

In addition to concordant expression patterns (either higher or lower abundances at both protein and mRNA levels), the heatmap also showed opposite patterns of mRNA and protein abundance levels, and revealed also abundance changes restricted exclusively to the protein level. The latter group includes more than half of

Table 1
Gene set enrichment analysis of proteins with significant abundance changes[a].

Set	Gene sets	Count	Adjusted p-value
UP	UP_Bt toxin,Cry5B	56	1.04E-45
	UP_Cadmium	45	4.26E-28
	UP_*B. thuringiensis* DB27	61	1.07E-18
	Down_*daf-16* mut	24	3.94E-10
	UP_BT247 on MY15	18	5.53E-10
	UP_BT247 on N2	21	1.11E-09
	Down_*daf-2* mut&RNAi	15	9.90E-05
	UP_BT247 on MY18	12	7.79E-04
	DOWN_*daf-2* mut (Day 6)	13	7.05E-03
DOWN	DOWN_BT247 on MY15	24	5.72E-17
	DOWN_BT247 on MY18	21	6.25E-08
	DOWN_Cadmium	17	1.58E-05

[a] The enrichment analysis was based on the 288 proteins with significant abundance changes upon pathogen exposure, using the program EASE and a *C. elegans*-specific reference database, which contains custom gene lists taken from previously published transcriptome studies generated under a variety of environmental conditions, defined nematode mutants, or developmental stages (see Table S4 for references). A significant overlap between the here inferred differentially expressed gene sets and the custom gene sets was assessed with a Fisher's exact test and Bonferroni-adjusted significance levels to take account of multiple statistical testing, as implemented in EASE. Abbreviations: UP, increased abundance; DOWN, decreased abundance; mut, mutant.

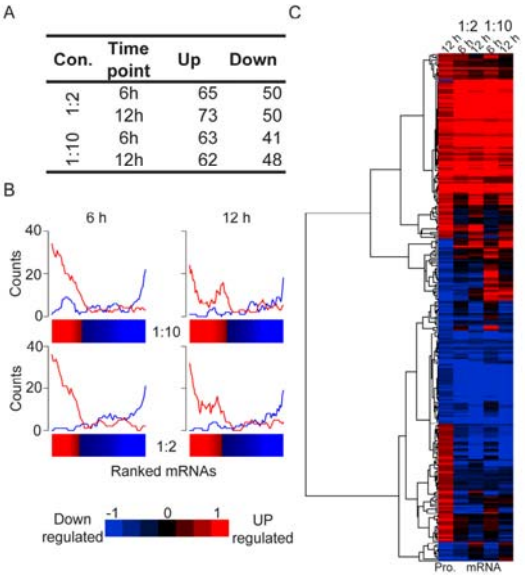


Fig. 2. Overview of different patterns of transcript and protein level abundance changes. This analysis is based on proteins and mRNAs with significant abundance changes. The latter were identified under adjusted p-value < 0.01 including only protein encoding mRNAs. Two concentrations (Con., 1:2 and 1:10) and two time points (6 h and 12 h) were analyzed in the transcriptional study. (A) Numbers of overlapping signatures. (B) Distribution of proteins with abundance changes in ranked differentially transcribed genes (ranked mRNAs). The number of proteins presented in the ranked mRNAs is calculated by a sliding window of 400 with a step of 40. (C) Heatmap of abundance changes at mRNA and protein levels. Signatures are organized by hierarchical clustering under uncentered correlation by Cluster 3.0. "Pro." indicates proteomics data in the left column.

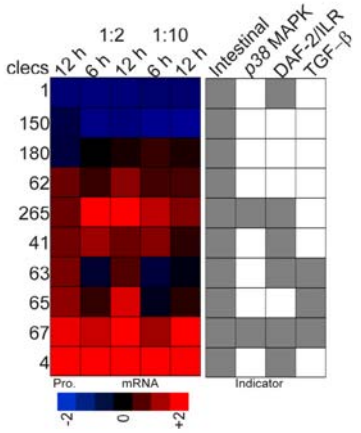


Fig. 3. Abundance changes of C-type lectin domain-containing genes and proteins (CTLDs). Eight out of ten CTLDs show similar regulation at mRNA and protein level (left panel). Seven of them are regulated by three common immune pathways (right panel). "Pro." denotes the proteomic data in the left column.

the proteins with significant abundance changes, indicating that the nematode's immune response to BT is influenced by post-transcriptional mechanisms.

3.2. C-type lectin domain-containing proteins (CTLDs) show consistent abundance changes at both protein and mRNA levels

The quantitative proteome data highlighted CTLDs as one of the highly responsive protein groups. In particular, the abundance of ten CTLDs was found to be altered upon bacterial challenge (Treitz et al., 2015). Eight of these also show a similar expression pattern at mRNA level (Fig. 3, left panel). The high concordance of CTLD protein and mRNA abundances strongly suggests their direct transcriptional regulation. Consistent with this idea, seven of the ten CTLDs are jointly regulated by central signaling cascades of the *C. elegans* immune system, such as the TGF-β (Mochii et al., 1999), p38 MAPK (Troemel et al., 2006), or DAF-2/ILR pathways (Halaschek-Wiener et al., 2005; McElwee et al., 2004) (Fig. 3 right panel). Moreover, all of them are expressed in the intestine (Stein et al., 2001) (Fig. 3 right panel), where most bacterial infections take place in *C. elegans*. These findings support a direct involvement of CTLDs in the inducible immune response against BT.

3.3. Unique proteome signatures reveal new candidate regulators of the nematode immune response

As mentioned above, approximately half of the proteins with different abundances after BT challenge do not show significant variations at the corresponding transcript level, indicating that they undergo some kind of post-transcriptional modification. To further assess the latter effects, we defined a set of 112 proteins (80 and 32 with higher or lower amounts, respectively), which all either produced a negative correlation with gene expression at the transcript level or an mRNA fold-change of less than 1.2 in at least three out of four conditions (from the two time points and two concentrations; see supplementary Table S2 for details). This protein set was subsequently subjected to enrichment analysis. GOrilla revealed an enrichment of the GO term "cytoskeleton" for the proteins with increased abundance (Fig. 4), indicating that BT infection has an impact

Table 2
Gene set enrichment analysis of unique proteomic signatures[a].

Set	Gene sets	Count	Adjusted p-value
UP	UP_*dcr-1* mut	19	1.64E-05
	UP_*lin-35* mut	20	1.84E-05
	UP_*osm-11* mut	9	7.99E-03
DOWN	Down_ *xbp-1* mut	13	6.93E-15
	UP_ *hpl-1* mut	15	5.33E-12
	DOWN_*nhr-23* RNAi	11	3.85E-10
	DOWN_*crh-1* mut	13	9.49E-10
	UP_*alg-1* mut	13	3.04E-09
	DOWN_*tom-1* mut	9	1.71E-07
	DOWN_*lin-14* mut	9	3.18E-06
	DOWN_*aak-2* oe	8	1.11E-04
	DOWN_*tax-6* mut	11	1.27E-04
	Down_*fer-1* mut	7	1.23E-03
	UP_*unc-62* RNAi	5	9.30E-03

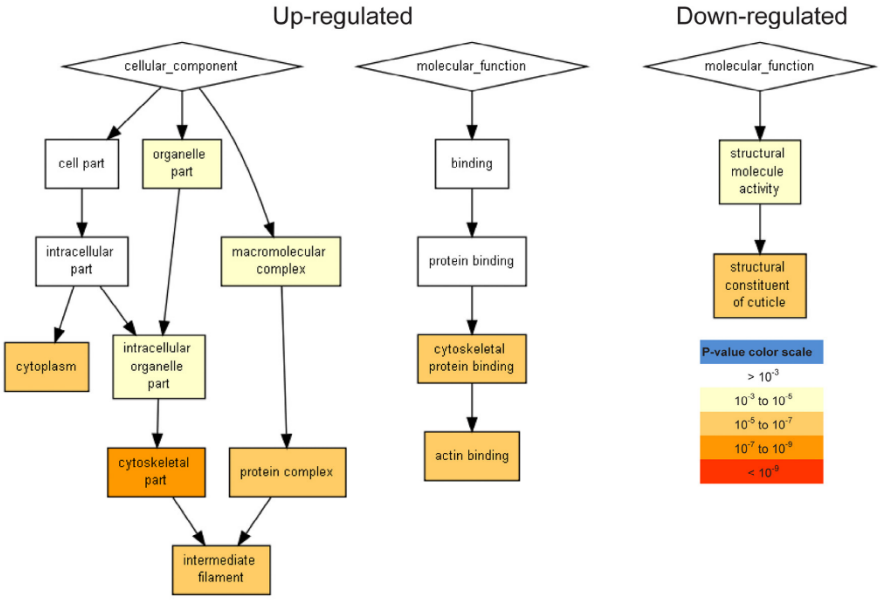
[a] The enrichment analysis was focused on the set of 112 proteins with significant differential expression at the protein level only, and thus without a corresponding differential expression at mRNA level. The analysis was based on the program EASE and a *C. elegans*-specific reference database (see legend of Table 1 and Table S4 for references). Abbreviations: UP, increased abundance; DOWN, decreased abundance; mut, mutant; oe, overexpressed.

on the cytoskeleton structure of cells, which is only realized after transcription. Moreover, the GO term "structural constituent of cuticle" was enriched for proteins with decreased abundance (Fig. 4), consistent with our GOrilla analysis of the whole proteomic data set.

EASE analysis identified 3 and 11 (p-value <0.01) gene sets enriched for proteins with higher or lower amounts, respectively. Surprisingly, none of the enriched groups were also among the over-represented gene sets identified by the EASE analysis on all of the proteins with significant abundance changes (Table 1), indicating that these gene sets are indeed distinct. Most of the enriched gene categories were inferred from previous transcriptomic analysis of worms, in which a specific gene was mutated or its expression down-regulated by RNAi (Table 2). Although some of these genes have been implicated in *C. elegans* pathogen-defense before (e.g., *dcr-1* (Iatsenko et al., 2013) *xbp-1* (Richardson et al., 2010), *osm-11* (Pujol et al., 2008) or *nhr-23* (Sahu et al., 2012)), others have not yet been linked to worm immunity and may thus represent novel regulatory genes. These sets include genes that function in multiple cellular processes, such as metabolic processes (*crh-1* and *aak-2*), development (*lin-35*, *tom-1*, and *lin-14*), RNA interference (*alg-1*), membrane fusion (*fer-1*), and cell migration (*unc-62*), while others function in sensory neurons (*tax-6*), and encode a *C. elegans* heterochromatin protein (*hpl-1*) (please see the Table S4 for references information).

4. Discussion

We here provide one of the very few studies that compare a transcriptomic with a proteomic response to pathogen infection (Encinas et al., 2010; Ragno et al., 2001). Using the nematode *C. elegans* and nematocidal BT as a model system, we show both concordances and differences between these two levels. Our analysis reveals that the abundance of several CTLD genes changes consistently across both levels, supporting a role of CTLDs as directly activated components of the *C. elegans* immune response. We further identified a unique set of proteins, whose abundances only changed at the proteome level, suggesting post-transcriptional modifications. This set is enriched in proteins, which are involved in cytoskeleton structuring and which are regulated by a number of genes, some of which are possibly novel immune regulators and/or targets of pathogenic BT manipulation.



	GO term	Description	P-value	FDR q-value	Count	Enrichment
UP	GO:0044430	cytoskeletal part	1.44E-08	1.06E-05	12	8.27
	GO:0005737	cytoplasm	2.38E-07	8.72E-05	19	3.72
	GO:0043234	protein complex	3.03E-06	5.54E-04	17	3.5
	GO:0005882	intermediate filament	2.52E-06	6.15E-04	4	39.58
	GO:0003779	actin binding	2.54E-06	4.53E-03	6	15.06
	GO:0008092	cytoskeletal protein binding	8.33E-06	7.42E-03	7	9.42
Down	GO:0042302	structural constituent of cuticle	1.04E-06	1.85E-03	6	16.6

Fig. 4. Functional annotation of unique proteomic signatures. Eighty proteins with increased and 32 proteins with decreased abundance were taken as input for the GOrilla web-server. Top panel: Functional enrichment map; the left two maps refer to the proteins with higher abundance, the right map to those with lower abundance. Bottom panel: table with specific GO terms and additional information. GO terms with a p-value < 1.0e-5 (FDR < 1.0e-2) are shown.

In detail, our study revealed four patterns for the relationship between transcriptomic and proteomic changes: (1) positively correlated abundances of mRNAs and proteins, (2) negatively correlated abundances of mRNAs and proteins, (3) unique abundance changes at the mRNA level, and (4) unique abundance changes in protein amounts. Thus, only some of the differentially expressed genes are also found among the proteins with significant abundance changes, which is consistent with previous studies (Encinas et al., 2010; Ragno et al., 2001). We observed that differentially expressed genes with a higher fold-change were more likely to be represented at the protein level (more than 40% of the proteins with abundance changes are among the top 400 highest differentially transcribed genes). There are three possible non-exclusive explanations for this observation: (i) Transcripts with higher fold-change are more reliable indicators for a change in gene expression, and therefore the corresponding proteins are especially those that can be recovered at the proteome level. Transcript abundance variations at lower fold-change are in this case likely dominated by artifacts. This would also mean that candidate genes, which are identified in transcriptome studies and then chosen for further functional analyses, should come from the top regulated genes. (ii) Only transcripts at high fold-change translate directly into proteins at the same or at least close time points, whereas transcripts at lower fold-change may then be translated less immediately or not at all. (iii) Technical resolution at the proteome level is much more constrained, resulting in identification of only those proteins with high abundance differences, which is then also reflected at high fold-change at the transcript level. We indeed noticed a much lower proteome coverage than transcriptome coverage. The raw data generated by the proteomic analysis included only about 3,600 proteins, possibly indicating that only a small fraction of the total proteome could be analyzed. However, further distinction between the above three alternatives is currently not possible and clearly warrants further investigation. For this reason, our more detailed enrichment analysis also

ignored the pattern (3), for which only transcripts varied in abundance, but not proteins, especially as the higher responsiveness at the transcript level could in this case be explained by a high degree of noise in the data.

For pattern (1), we find a convincing overlap among previous studies on the transcriptomic response of *C. elegans* to BT, our new RNA-Seq data, and the results of the proteome study. This overlap suggests that there is a robust worm response to BT exposure that can be consistently identified across mRNA and protein levels. A strong signal for this overlap comes from carbohydrate binding, mainly due to differential expression of CTLD genes. More precisely, we observed a highly consistent change in abundance of *C. elegans* genes encoding CTLDs at the mRNA and protein level and identified eight CTLDs, which are all localized in the *C. elegans* intestine, and of which seven are potentially regulated by known nematode immunity pathways. With the exception of *clec-1*, these CTLD genes have previously been shown to be up-regulated at the mRNA level after exposure to at least one other pathogen (Schulenburg et al., 2008). However, only *clec-65* has previously been examined in a functional genetic analysis, demonstrating that the gene's knock-down by RNAi increased susceptibility to the pathogenic *E. coli* strain LF82 (Simonsen et al., 2011). CTLDs contain carbohydrate-recognition domains (CRDs) that are homologous to the CRDs in the animal C-type (Ca^{2+}-dependent) lectin family. CTLDs mediate a variety of protein–carbohydrate interactions in the vertebrate immune system, functioning as PRRs or antimicrobial effector proteins (Janeway and Medzhitov, 2002; Mukherjee et al., 2013; Weis et al., 1998). For example, dectin-1 and dectin-2 mediate the response against fungi by recognizing and binding β-glucans (Brown and Gordon, 2001) and α-mannans (Saijo et al., 2010), respectively. DC-SIGN is a receptor for *Schistosoma mansoni* egg antigens, binding to the glycan antigen Lewis x (van Die et al., 2003). In *C. elegans*, genes encoding CTLDs are differentially expressed at the mRNA level after infection with various pathogens (Schulenburg et al., 2008), and the function of several of these has been tested by using the respective mutants, RNAi knock-down or over-expression, including *clec-17*, *clec-60*, *clec-61*, *clec-70*, and *clec-71* (Irazoqui et al., 2010; O'Rourke et al., 2006). Increased protein amounts of CLEC-63 have been found after *C. elegans* exposure to the Gram-negative bacterium *Aeromonas hydrophila* (Bogaerts et al., 2010) and the *E. coli* strain LF82, in this case together with CLEC-65 and CLEC-85 (Simonsen et al., 2011). Interestingly, the *clec-63* and *clec-85* genes showed no-change at the mRNA level after infection with *E. coli* LF82. Overall, these observations suggest that CTLDs play an important role in host defense to BT infection and might also be regulated by yet unknown post-transcriptional mechanisms.

Post-transcriptional modification, post-translational changes, and other processes that act after transcription produce changes at the protein level that are not directly correlated with mRNA abundances and are thus indicated by negative correlation between protein and mRNA changes (pattern 2), as well as abundance differences at the protein level only (pattern 4). To investigate the role of such mechanisms in regulating the immune response, we combined the proteins from patterns (2) and (4) and used them as unique signatures at the protein level. Our enrichment analysis on these unique signatures identified the GO term "cytoskeleton" to be over-represented among these proteins, but not among the genes that showed a consistent regulation at mRNA and protein levels. It is well known that pathogen infection induces the reorganization of the cytoskeleton, which is critical for the immune defense (Bhavsar et al., 2007; Dramsi and Cossart, 1998). A similar enrichment of cytoskeleton proteins was also found in a proteomic study in Zebrafish infected with viral hemorrhagic septicemia virus (VHSV) (Encinas et al., 2010), possibly indicating a general cytoskeleton response toward pathogens (Dustin and Cooper, 2000), and/or that BT and VHSV pathogens manipulate the skeleton structure. Among the unique protein set with decreased abundance, we observed an enrichment for the GO term "cuticle structure". As BT infects *C. elegans* orally and the infection takes place in the intestine, where BT crystal toxins lead to damage of the intestinal epithelium, this result is surprising. It might, however, indicate that upon BT exposure, the cuticle is modulated either by the host or by the pathogen irrespective of the site of infection.

In addition, for this group of proteins with unique abundance changes, the worm-specific enrichment analysis with the EASE approach and our custom database identified several genes that might regulate the immune response to BT, including previously known regulators of *C. elegans* immunity as well as new candidate immune regulators. Among the proteins with decreased abundance, we observed a significant enrichment of *xbp-1* and *nhr-23* regulated gene sets. *xbp-1*, encoding a bZIP transcription factor that is required for the unfolded protein response (UPR), has previously been shown to be important for the protection of *C. elegans* against exposure to the BT Cry5B toxin (Bischof et al., 2008), and also for the maintenance of endoplasmic reticulum homeostasis during infection with *Pseudomonas aeruginosa* (Richardson et al., 2010). *nhr-23*, encoding a nuclear hormone receptor, is known to contribute to *C. elegans* resistance to *Vibrio cholerae* cytolysin (Sahu et al., 2012). Among the proteins with increased abundance, we furthermore found an enrichment of *dcr-1* and *osm-11* regulated genes. *dcr-1*, an endoribonuclease that is required for RNA interference, has been identified as a negative regulator of the *C. elegans* immune response to the *B. thuringiensis* strain DB27 (Iatsenko et al., 2013). *osm-11* encodes a Notch ligand functioning among others in the *C. elegans* response to osmotic stress and has been shown to negatively regulate the expression of the epidermal antimicrobial peptide gene *nlp-29* (Pujol et al., 2008). The identification of these known *C. elegans* immune regulators confirms that the unique proteomic signature set includes proteins that exhibit essential functions in *C. elegans* immunity.

Moreover, we identified new candidate immune regulators within the unique proteome signature set. Perhaps the most interesting candidates are *crh-1* and *aak-2*, which are both enriched among the down-regulated proteins and which are both linked to adenosine monophosphate (AMP). While *crh-1* encodes a homolog of the cyclic AMP-response element binding protein (CREB), *aak-2* encodes one of two *C. elegans* homologs of the catalytic alpha subunit of AMP-activated protein kinases (AMPKs), which is responsive to the AMP/ATP ratio within cells and, interestingly, is known to link energy levels to DAF-2/IlR signaling in worms (Apfeld et al., 2004). As DAF-2/IlR signaling regulates the *C. elegans* response to BT (Hasshoff et al., 2007; Wang et al., 2012). *aak-2* might play a role in integrating information from DAF-2 signaling on the one hand and from endogenous stress signals (AMP/ATP ratio) generated during BT infection on the other hand. Indeed, food intake in BT infected worms is heavily impeded in the intestine due to the damage of the epithelium by BT toxins and in general by a reduction of ingestion as part of the behavioral defense response of the worm (Hasshoff et al., 2007; Schulenburg and Müller, 2004). The decline in food intake may result in reduced energy availability and increasing AMP levels, which in turn could activate AMPKs like AAK-2. In summary, the unique proteomic signatures provide new insights into the *C. elegans* immune response to BT infection and reveal *aak-2* as a new candidate mediator of host defense.

To conclude, the combination of transcriptomic with proteomic analyses of the host response to infection has the power to uncover new candidate mediators and regulators of the immune response. Finally, we propose that AMPKs are new regulators of the immune response to infection with BT and possibly other pathogens. However, the exact role of AMPKs in immunity needs to be confirmed by functional analyses in the future.

Acknowledgements

We thank Daniela Haase and the Kiel ICMB sequencing team (especially Markus Schilhabel, Melanie Friskovec, Melanie Schlapkohl) for technical assistance. We are grateful for financial support from the German Science Foundation to HS and PR within the German priority program SPP 1399 on host–parasite coevolution (DFG grants SCHU 1415/8; SCHU 1415/9; RO 2994/3). WY was additionally supported by the International Max Planck Research School for Evolutionary Biology; KD by funds from the Christian-Albrechts-University Kiel; and AT, ML, PR, and HS by the Cluster of Excellence 'Inflammation at Interfaces' of the German Science Foundation, including Cluster laboratory CL-X.

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.dci.2015.02.010.

References

Apfeld, J., O'Connor, G., McDonagh, T., DiStefano, P.S., Curtis, R., 2004. The AMP-activated protein kinase AAK-2 links energy levels and insulin-like signals to lifespan in *C. elegans*. Genes Dev. 18, 3004–3009.
Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., et al., 2013. NCBI GEO: archive for functional genomics data sets – update. Nucleic Acids Res. 41, D991–D995.
Bartee, E., McCormack, A., Früh, K., 2006. Quantitative membrane proteomics reveals new cellular targets of viral immune modulators. PLoS Pathog. 2, e107.
Bhavsar, A.P., Guttman, J.A., Finlay, B.B., 2007. Manipulation of host-cell pathways by bacterial pathogens. Nature 449, 827–834.
Bischof, L.J., Kao, C.-Y., Los, F.C., Gonzalez, M.R., Shen, Z., Briggs, S.P., et al., 2008. Activation of the unfolded protein response is required for defenses against bacterial pore-forming toxin in vivo. PLoS Pathog. 4, e1000176.
Boehnisch, C., Wong, D., Habig, M., Isermann, K., Michiels, N.K., Roeder, T., et al., 2011. Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. PLoS ONE 6, e24619.
Bogaerts, A., Temmerman, L., Boerjan, B., Husson, S.J., Schoofs, L., Verleyen, P., 2010. A differential proteomics study of *Caenorhabditis elegans* infected with *Aeromonas hydrophila*. Dev. Comp. Immunol. 34, 690–698.
Boutros, M., Agaisse, H., Perrimon, N., 2002. Sequential activation of signaling pathways during innate immune responses in *Drosophila*. Dev. Cell 3, 711–722.
Brown, G.D., Gordon, S., 2001. Immune recognition: a new receptor for β-glucans. Nature 413, 36–37.
Chen, Z., Hagler, J., Palombella, V.J., Melandri, F., Scherer, D., Ballard, D., et al., 1995. Signal-induced site-specific phosphorylation targets I kappa B alpha to the ubiquitin-proteasome pathway. Genes Dev. 9, 1586–1597.
Couillault, C., Pujol, N., Reboul, J., Sabatier, L., Guichou, J.-F., Kohara, Y., et al., 2004. TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human SARM. Nat. Immunol. 5, 488–494.
Couillault, C., Fourquet, P., Pophillat, M., Ewbank, J.J., 2012. A UPR-independent infection-specific role for a BiP/GRP78 protein in the control of antimicrobial peptide expression in *C. elegans* epidermis. Virulence 3, 299–308.
de Hoon, M.J., Imoto, S., Nolan, J., Miyano, S., 2004. Open source clustering software. Bioinformatics 20, 1453–1454.
De Gregorio, E., Spellman, P.T., Rubin, G.M., Lemaitre, B., 2001. Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. PNAS 98, 12590–12595.
De Gregorio, E., Spellman, P.T., Tzou, P., Rubin, G.M., Lemaitre, B., 2002. The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. EMBO J. 21, 2568–2579.
Dramsi, S., Cossart, P., 1998. Intracellular pathogens and the actin cytoskeleton. Annu. Rev. Cell Dev. Biol. 14, 137–166.
Dustin, M.L., Cooper, J.A., 2000. The immunological synapse and the actin cytoskeleton: molecular hardware for T cell signaling. Nat. Immunol. 1, 23–29.
Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48.
Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.
Encinas, P., Rodriguez-Milla, M.A., Novoa, B., Estepa, A., Figueras, A., Coll, J., 2010. Zebrafish fin immune responses during high mortality infections with viral haemorrhagic septicemia rhabdovirus. A proteomic and transcriptomic approach. BMC Genomics 11, 518.
Engelmann, I., Griffon, A., Tichit, L., Montanana-Sanchis, F., Wang, G., Reinke, V., et al., 2011. A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans. PLoS ONE 6, e19055.
Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., et al., 2011. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 7, e1001393.
Griffitts, J.S., Haslam, S.M., Yang, T., Garczynski, S.F., Mulloy, B., Morris, H., et al., 2005. Glycolipids as receptors for *Bacillus thuringiensis* crystal toxin. Science 307, 922–925.
Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720–1730.
Halaschek-Wiener, J., Khattra, J.S., McKay, S., Pouzyrev, A., Stott, J.M., Yang, G.S., et al., 2005. Analysis of long-lived *C. elegans* daf-2 mutants using serial analysis of gene expression. Genome Res. 15, 603–615.
Hasshoff, M., Böhnisch, C., Tonn, D., Hasert, B., Schulenburg, H., 2007. The role of *Caenorhabditis elegans* insulin-like signaling in the behavioral avoidance of pathogenic *Bacillus thuringiensis*. FASEB J. 21, 1801–1812.
Hooper, L.V., Wong, M.H., Thelin, A., Hansson, L., Falk, P.G., Gordon, J.I., 2001. Molecular analysis of commensal host-microbial relationships in the intestine. Science 291, 881–884.
Hooper, L.V., Stappenbeck, T.S., Hong, C.V., Gordon, J.I., 2003. Angiogenins: a new class of microbicidal proteins involved in innate immunity. Nat. Immunol. 4, 269–273.
Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., Lempicki, R.A., 2003. Identifying biological themes within lists of genes with EASE. Genome Biol. 4, R70.
Huffman, D.L., Abrami, L., Sasik, R., Corbeil, J., van der Goot, F.G., Aroian, R.V., 2004. Mitogen-activated protein kinase pathways defend against bacterial pore-forming toxins. Proc. Natl. Acad. Sci. U.S.A. 101, 10995–11000.
Iatsenko, I., Sinha, A., Rödelsperger, C., Sommer, R.J., 2013. New role for DCR-1/dicer in *Caenorhabditis elegans* innate immunity against the highly virulent bacterium *Bacillus thuringiensis* DB27. Infect. Immun. 81, 3942–3957.
Irazoqui, J.E., Troemel, E.R., Feinbaum, R.L., Luhachack, L.G., Cezairliyan, B.O., Ausubel, F.M., 2010. Distinct pathogenesis and host responses during infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. PLoS Pathog. 6, e1000982.
Irving, P., Troxler, L., Heuer, T.S., Belvin, M., Kopczynski, C., Reichhart, J.-M., et al., 2001. A genome-wide analysis of immune responses in *Drosophila*. PNAS 98, 15119–15124.
Izzi, L., Attisano, L., 2004. Regulation of the TGFβ signalling pathway by ubiquitin-mediated degradation. Oncogene 23, 2071–2078.
Janeway, C.A., Jr., Medzhitov, R., 2002. Innate immune recognition. Annu. Rev. Immunol. 20, 197–216.
Kao, C.-Y., Los, F.C., Huffman, D.L., Wachi, S., Kloft, N., Husmann, M., et al., 2011. Global functional analyses of cellular responses to pore-forming toxins. PLoS Pathog. 7, e1001314.
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.
Kurz, C.L., Ewbank, J.J., 2003. Caenorhabditis elegans: an emerging genetic model for the study of innate immunity. Nat. Rev. Genet. 4, 380–390.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.
Mallo, G.V., Kurz, C.L., Couillault, C., Pujol, N., Granjeaud, S., Kohara, Y., et al., 2002. Inducible antibacterial defense system in *C. elegans*. Curr. Biol. 12, 1209–1214.
Marsh, E.K., May, R.C., 2012. Caenorhabditis elegans, a model organism for investigating immunity. Appl. Environ. Microbiol. 78, 2075–2081.
McElwee, J., Bubb, K., Thomas, J.H., 2003. Transcriptional outputs of the *Caenorhabditis elegans* forkhead protein DAF-16. Aging Cell 2, 111–121.
McElwee, J.J., Schuster, E., Blanc, E., Thomas, J.H., Gems, D., 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, 44533–44543.
Mochii, M., Yoshida, S., Morita, K., Kohara, Y., Ueno, N., 1999. Identification of transforming growth factor-β-regulated genes in *Caenorhabditis elegans* by differential hybridization of arrayed cDNAs. PNAS 96, 15020–15025.
Mukherjee, S., Zheng, H., Derebe, M.G., Callenberg, K.M., Partch, C.L., Rollins, D., et al., 2013. Antibacterial membrane attack by a pore-forming intestinal C-type lectin. Nature 505, 103–107.
Mukhopadhyay, D., Riezman, H., 2007. Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science 315, 201–205.
Murphy, C.T., McCarroll, S.A., Bargmann, C.I., Fraser, A., Kamath, R.S., Ahringer, J., et al., 2003. Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. Nature 424, 277–283.
O'Rourke, D., Baban, D., Demidova, M., Mott, R., Hodgkin, J., 2006. Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res. 16, 1005–1016.
Pujol, N., Zugasti, O., Wong, D., Couillault, C., Kurz, C.L., Schulenburg, H., et al., 2008. Anti-fungal innate immunity in *C. elegans* is enhanced by evolutionary diversification of antimicrobial peptides. PLoS Pathog. 4, e1000105.
Ragno, S., Romano, M., Howell, S., Pappin, D.J., Jenner, P.J., Colston, M.J., 2001. Changes in gene expression in macrophages infected with *Mycobacterium tuberculosis*: a combined transcriptomic and proteomic approach. Immunology 104, 99–108.
Richardson, C.E., Kooistra, T., Kim, D.H., 2010. An essential role for XBP-1 in host protection against immune activation in *C. elegans*. Nature 463, 1092–1095.
Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25.
Sahu, S.N., Lewis, J., Patel, I., Bozdag, S., Lee, J.H., LeClerc, J.E., et al., 2012. Genomic analysis of immune response against *Vibrio cholerae* hemolysin in *Caenorhabditis elegans*. PLoS ONE 7, e38200.

Saijo, S., Ikeda, S., Yamabe, K., Kakuta, S., Ishigame, H., Akitsu, A., et al., 2010. Dectin-2 recognition of α-mannans and induction of Th17 cell differentiation is essential for host defense against *Candida albicans*. Immunity 32, 681–691.
Saldanha, A.J., 2004. Java Treeview–extensible visualization of microarray data. Bioinformatics 20, 3246–3248.
Scherl, A., François, P., Charbonnier, Y., Deshusses, J.M., Koessler, T., Huyghe, A., et al., 2006. Exploring glycopeptide-resistance in *Staphylococcus aureus*: a combined proteomics and transcriptomics approach for the identification of resistance-related markers. BMC Genomics 7, 296.
Schulenburg, H., Müller, S., 2004. Natural variation in the response of *Caenorhabditis elegans* towards *Bacillus thuringiensis*. Parasitology 128, 433–443.
Schulenburg, H., Hoeppner, M.P., Weiner, J., III, Bornberg-Bauer, E., 2008. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. Immunobiology 213, 237–250.
Shapira, M., Hamlin, B.J., Rong, J., Chen, K., Ronen, M., Tan, M.-W., 2006. A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. PNAS 103, 14086–14091.
Sherman, B.T., Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.
Simonsen, K.T., Møller-Jensen, J., Kristensen, A.R., Andersen, J.S., Riddle, D.L., Kallipolitis, B.H., 2011. Quantitative proteomics identifies ferritin in the innate immune response of *C. elegans*. Virulence 2, 120–130.
Sinha, A., Rae, R., Iatsenko, I., Sommer, R.J., 2012. System wide analysis of the evolution of innate immunity in the nematode model species *Caenorhabditis elegans* and *Pristionchus pacificus*. PLoS ONE 7, e44255.
Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., Spieth, J., 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. 29, 82–86.
Stiernagle, T., 2006. Maintenance of C. elegans. In: WormBook. The C. elegans Research Community. WormBook.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550.
Team, R.C., 2012. R: a language and environment for statistical computing.
Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31, 46–53.
Treitz, C., Cassidy, L., Höckendorf, A., Leippe, M., Tholey, A., 2015. Quantitative proteome analysis of *Caenorhabditis elegans* upon exposure to nematicidal *Bacillus thuringiensis*. J. Proteomics 113, 337–350.
Troemel, E.R., Chu, S.W., Reinke, V., Lee, S.S., Ausubel, F.M., Kim, D.H., 2006. p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. PLoS Genet. 2, e183.
van Die, I., van Vliet, S.J., Nyame, A.K., Cummings, R.D., Bank, C.M., Appelmelk, B., et al., 2003. The dendritic cell–specific C-type lectin DC-SIGN is a receptor for Schistosoma mansoni egg antigens and recognizes the glycan antigen Lewis x. Glycobiology 13, 471–478.
Vierstraete, E., Verleyen, P., Baggerman, G., D'Hertog, W., Van den Bergh, G., Arckens, L., et al., 2004. A proteomic approach for the analysis of instantly released wound and immune proteins in *Drosophila melanogaster* hemolymph. Proc. Natl. Acad. Sci. U.S.A. 101, 470–475.
Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227–232.
Wang, J., Nakad, R., Schulenburg, H., 2012. Activation of the *Caenorhabditis elegans* FOXO family transcription factor DAF-16 by pathogenic *Bacillus thuringiensis*. Dev. Comp. Immunol. 37, 193–201.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Weis, W.I., Taylor, M.E., Drickamer, K., 1998. The C-type lectin superfamily in the immune system. Immunol. Rev. 163, 19–34.
Ziegler, K., Kurz, C.L., Cypowyj, S., Couillault, C., Pophillat, M., Pujol, N., et al., 2009. Antifungal innate immunity in *C. elegans*: PKCδ links G protein signaling and a conserved p38 MAPK cascade. Cell Host Microbe 5, 341–352.

Note: the supplementary tables are too large to be attached here and they are available at http://doi:10.1016/j.dci.2015.02.010

Chapter V

Contrasting invertebrate immune defense behaviors caused by a single gene, the *Caenorhabditis elegans* neuropeptide receptor gene *npr-1*

Rania Nakad[1,2], Basten L. Snoek[3], Wentao Yang[1], Sunna Ellendt[1], Franziska Schneider[1], Timm G. Mohr[1], Lone Rösingh[1], Anna C. Masche[1], Philip C. Rosenstiel[4], Katja Dierking[1], Jan E. Kammenga[3], Hinrich Schulenburg[1]*

[1] Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, 24098 Kiel, Germany

[2] Cologne Excellence Cluster for Cellular Stress Responses in Ageing-Associated Diseases (CECAD) and Systems Biology of Ageing, University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany.

[3] Laboratory of Nematology, Wageningen University, Wageningen 6708 PB, The Netherlands

[4] Institute for Clinical Molecular Biology, University of Kiel. 24098 Kiel, Germany

* Correspondence: hschulenburg@zoologie.uni-kiel.de

Nakad *et al. BMC Genomics* (2016) 17:280
DOI 10.1186/s12864-016-2603-8

BMC Genomics

RESEARCH ARTICLE Open Access

Contrasting invertebrate immune defense behaviors caused by a single gene, the *Caenorhabditis elegans* neuropeptide receptor gene *npr-1*

Rania Nakad[1,2], L. Basten Snoek[3], Wentao Yang[1], Sunna Ellendt[1], Franziska Schneider[1], Timm G. Mohr[1], Lone Rösingh[1], Anna C. Masche[1], Philip C. Rosenstiel[4], Katja Dierking[1], Jan E. Kammenga[3] and Hinrich Schulenburg[1*]

Abstract

Background: The invertebrate immune system comprises physiological mechanisms, physical barriers and also behavioral responses. It is generally related to the vertebrate innate immune system and widely believed to provide nonspecific defense against pathogens, whereby the response to different pathogen types is usually mediated by distinct signalling cascades. Recent work suggests that invertebrate immune defense can be more specific at least at the phenotypic level. The underlying genetic mechanisms are as yet poorly understood.

Results: We demonstrate in the model invertebrate *Caenorhabditis elegans* that a single gene, a homolog of the mammalian neuropeptide Y receptor gene, *npr-1*, mediates contrasting defense phenotypes towards two distinct pathogens, the Gram-positive *Bacillus thuringiensis* and the Gram-negative *Pseudomonas aeruginosa*. Our findings are based on combining quantitative trait loci (QTLs) analysis with functional genetic analysis and RNAseq-based transcriptomics. The QTL analysis focused on behavioral immune defense against *B. thuringiensis*, using recombinant inbred lines (RILs) and introgression lines (ILs). It revealed several defense QTLs, including one on chromosome X comprising the *npr-1* gene. The wildtype N2 allele for the latter QTL was associated with reduced defense against *B. thuringiensis* and thus produced an opposite phenotype to that previously reported for the N2 *npr-1* allele against *P. aeruginosa*. Analysis of *npr-1* mutants confirmed these contrasting immune phenotypes for both avoidance behavior and nematode survival. Subsequent transcriptional profiling of *C. elegans* wildtype and *npr-1* mutant suggested that *npr-1* mediates defense against both pathogens through p38 MAPK signaling, insulin-like signaling, and C-type lectins. Importantly, increased defense towards *P. aeruginosa* seems to be additionally influenced through the induction of oxidative stress genes and activation of GATA transcription factors, while the repression of oxidative stress genes combined with activation of Ebox transcription factors appears to enhance susceptibility to *B. thuringiensis*.

Conclusions: Our findings highlight the role of a single gene, *npr-1*, in fine-tuning nematode immune defense, showing the ability of the invertebrate immune system to produce highly specialized and potentially opposing immune responses via single regulatory genes.

Keywords: *Caenorhabditis elegans*, Pathogen avoidance behavior, Innate immunity, Immune specificity, QTL analysis

* Correspondence: hschulenburg@zoologie.uni-kiel.de
[1]Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, 24098 Kiel, Germany
Full list of author information is available at the end of the article

Background

In contrast to higher vertebrates, which have adaptive immune response systems, invertebrates exclusively rely on the innate immune system (the immune system is here defined *sensu lato* as the organism's defense against infection, including avoidance behavior, physical barriers, and physiological processes). For a long time it was assumed that only the adaptive system is capable of mounting highly specific defense responses. However, evidence is accumulating that invertebrates have surprisingly complex immune systems that in theory may have the potential to produce similar specificities [1–3]. Yet, to date, we possess only little information on the genetic and molecular mechanisms underlying such specificities. First insights into these mechanisms were previously obtained for the model nematode *C. elegans*, an important invertebrate system for studying immune defense [2, 4, 5]. For example, loss of function of the prolylhydroxylase encoding gene *egl-9* enhances susceptibility to *Staphylococcus aureus* [6] but resistance to *Pseudomonas aeruginosa* (PA01) [7] and *Bacillus thuringiensis* toxins [8]. Similarly, a loss of function of the Toll-like receptor gene *tol-1* increases susceptibility to *Salmonella enterica* but resistance to *Enterococcus faecalis* [9], even though the general importance of *tol-1* in worm immunity is unclear [5, 10].

Such specificities may not only be expressed by the nematode's physiological immune system, but could also be expected for behavioral defenses. Such behaviors are a central component of immune defense *sensu lato* - next to protective barriers and physiological processes - and are likely to represent a highly economic immune defense strategy because they simultaneously reduce pathogen contact, and thus the risk of tissue damage, and also the necessity to activate the energetically costly physiological and cellular response [11]. *C. elegans* colonizes microbe-rich habitats in nature where it feeds on bacteria and yeasts [12–15]. Since these habitats also contain many pathogenic microorganisms, *C. elegans* has evolved distinct types of behavioral responses including physical avoidance, associative learning and reduced oral uptake of pathogens [4, 16–22].

Previous studies revealed the presence of substantial genetic variation among wild isolates of *C. elegans* in their behavioral response towards different pathogens [17, 19, 23–27]. In one case, namely the defense response against the Gram-negative bacterium *Pseudomonas aeruginosa*, this variation could be linked to the polymorphic neuropeptide receptor *npr-1* locus on the X chromosome. The gene *npr-1* was proposed to regulate *C. elegans'* immunity against PA14 either through controlling the aerotaxis response [17], or through controlling both aerotaxis response and physiological immune defense [18]. *npr-1* is a homolog of the mammalian neuropeptide Y receptor gene and it is found in two different isoforms in *C. elegans* that result from a single amino acid change at position 215 (valine in isoform 215 V; phenylalanine in isoform 215 F) [28]. These isoforms do not only influence pathogen defense but also foraging behavior in response to oxygen concentrations [28, 29] and leaving behavior from lawns with the laboratory food bacterium *Escherichia coli* [30, 31].

The apparent complexity of the *C. elegans* defense against pathogens [1–3, 5] raises the question whether single pathways or genes can also fine-tune the behavioral defense response towards specific pathogens. To address this question we studied the genetic architecture of behavioral immune defense of *C. elegans* towards the Gram-positive pathogen *Bacillus thuringiensis*. This pathogen is likely to coexist with *C. elegans* in nature [15]. Some strains are nematocidal, whereby the host is infected by the oral uptake of spore-toxin mixtures. Infection of the gut is followed by toxin-mediated cellular damage of the intestinal epidermis, germination of spores and subsequent proliferation of vegetative cells, including expression of various virulence factors, ultimately resulting in nematode death [32–36]. Nematocidal *B. thuringiensis* induces pronounced behavioral responses in *C. elegans* [21, 23, 37, 38].

Here we explored genetic variation in *C. elegans* and used quantitative trait locus (QTL) analysis to characterize the genetic basis of behavioral immune defense against two pathogenic *B. thuringiensis* strains, whereby one strains (BT B-18679) is known to be more pathogenic than the other (BT B-18247) [39, 40]. Our QTL analysis was based on a panel of 200 recombinant inbred lines (RILs) and 90 introgression lines (ILs), derived from a cross between the *C. elegans* strains N2 and CB4856 [41, 42]. Our QTL analysis identified *npr-1* as one of the candidate genes, though with an opposite effect on avoidance behavior to that previously reported towards *P. aeruginosa* [17, 18]. Therefore, we further characterized the function of the *npr-1* gene in producing contrasting pathogen defense responses. Using *npr-1* mutants, we assessed the influence of the gene on both avoidance behavior and survival towards the two pathogen species, *B. thuringiensis* and *P. aeruginosa*. Moreover, we used RNAseq to identify differences in the pathogen-dependent transcription of *npr-1* down-stream targets. The functional importance of such differences was assessed through enrichment analysis of gene ontology (GO) categories, customized nematode-specific gene sets, which we collated from previous gene expression analyses, and transcription factor binding motifs.

Results and discussion

Two *C. elegans* wild-type strains differ in bacterial lawn leaving behavior

The standard laboratory strain N2 and the Hawaiian strain CB4856 showed significant variation in lawn-leaving behavior towards nematocidal *B. thuringiensis* strains (B-

18679 & B-18247) and non-nematocidal strains (DSM350 & *E. coli* OP50). Lawn leaving served as a proxy for behavioral defense and was based on an assay (Additional file 1) related to those previously used to characterize *C. elegans* avoidance behavior [10, 16, 19, 21, 31, 37], in this case using peptone-free medium (PFM) to prevent *B. thuringiensis* spore germination outside the host (see Methods). Lawn leaving behavior was significantly higher for CB4856 compared to N2 on all tested bacterial strains and for all exposure time periods (Fig. 1; Sheet 1 in Additional files 2, 3 and 4). For both *C. elegans* strains, we observed a significant increase in leaving across time (Fig. 1). For both, the avoidance response towards the most pathogenic strain (B-18679) was higher than that towards the less pathogenic strain (B-18247) (Fig. 1c, d).

Our results confirm previously reported higher avoidance behavior and resistance of CB4856 compared to N2 towards one of the pathogens used in the current study, *B. thuringiensis* B-18247 [23]. Our findings are also consistent with two previous studies that demonstrated a higher OP50-patch leaving behavior [31] and a higher microsporidia resistance of CB4856 compared to N2 [43]. Interestingly, the opposite phenotype has been reported regarding the nematode's response to two other pathogens, *P. aeruginosa* and *Serratia marcescens*. In these cases, N2 rather than CB4856 produced higher resistance and behavioral avoidance towards *P. aeruginosa* [17, 19], and higher avoidance towards *S. marcescens* [26]. Moreover, as the more pathogenic B-18679 was more strongly avoided than the less virulent B-18247 (Fig. 1c, d), *C. elegans* appears to be able to differentiate between different levels of pathogenicity of the same bacterial species. In this case, the difference in pathogenicity is likely due to expression of different Cry toxins that result in different infection patterns [37]. Based on our results we expect that the N2 and CB4856-derived RIL and IL populations are likely to contain sufficiently high levels of variation for a QTL analysis of avoidance behaviors towards the four chosen bacterial strains.

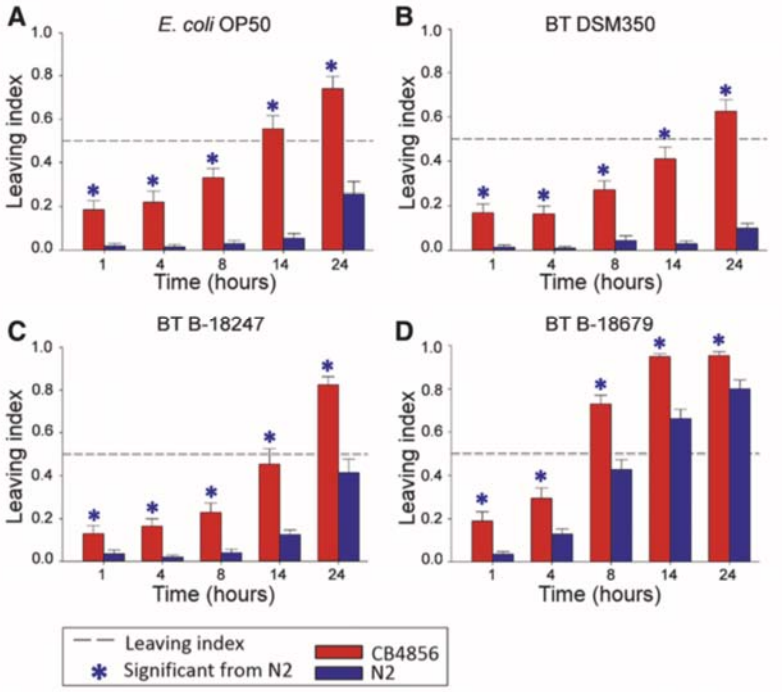


Fig. 1 Variation in lawn leaving behavior between N2 and CB4856 towards different bacterial strains. Lawn leaving behavior towards: (**a**) the non-pathogenic *E. coli* strain OP50; (**b**) the non-nematocidal *B. thuringiensis* strain DSM350; (**c**) the nematocidal *B. thuringiensis* strain B-18247; and (**d**) the highly nematocidal *B. thuringiensis* strain B-18679. The Y axis shows the leaving index for N2 (*blue*) or CB4856 (*red*), calculated as the proportion of nematodes which left the bacterial lawn; the X axis indicates the different exposure time points. Asterisks denote the treatments for which leaving behavior of CB4856 is significantly different from N2. The dotted reference line shows a leaving index of 0.5. Error bars represent standard error of the mean. Raw data are given in sheet 1 of Additional file 2. Detailed statistical results are shown in Additional file 3

Multiple QTLs and their interactions account for variation in avoidance behavior

We performed QTL analyses on *C. elegans* pathogen defense and revealed the genetic architecture of pathogen avoidance behavior to (i) be polygenic, (ii) include epistatically interacting loci, and (iii) incorporate general as well as pathogen-specific avoidance loci. In particular, our study simultaneously assessed the behavioral response of 200 RILs and 90 ILs [41, 42] towards four bacterial strains (two nematocidal *B. thuringiensis* and two non-nematocidal controls) at two exposure time points (14 h and 24 h) and with three replicates per treatment combination, using the same lawn leaving assays as above for N2 and CB4856. Below we present our results of a main-effect QTL analysis of the RIL population and an analysis of interaction effects among loci for the RIL population.

The main effect QTL analysis uncovered five main regions associated with avoidance: (i) one large region in the middle of chromosome II, for which the CB4856 allele(s) increase(s) leaving behavior of the non-pathogenic bacteria only at the 24 h time point (Fig. 2a, b); (ii) a region on chromosome II, for which the N2 allele(s) specifically increase(s) avoidance of the more pathogenic strain B-18679 at the 24 h time point (Fig. 2d); (iii) a region on the left arm of chromosome IV, for which the N2 allele(s) increase(s) leaving behavior towards controls and pathogens (Fig. 2a-d); (iv) a region on the right arm of chromosome IV, for which the CB4856 allele(s) increase(s) avoidance of pathogens at both time points (Fig. 2c, d); and (v) a region on chromosome X with the strongest effect on leaving behavior towards controls and pathogens (Fig. 2a-d), mediated by the CB4856 allele(s) and including a significant time effect on the response to the pathogenic bacteria (Fig. 2c, d). This very strong X chromosome effect was confirmed by the ILs (Additional file 5).

Our analysis of interaction effects among QTLs, using a standard interaction model (phenotype ~ time + marker1 * marker2), revealed several significant intra-genomic associations with an influence on lawn leaving behavior. For *E. coli*, significant interactions were found for at least two cases (Fig. 3a): (i) between the beginning of chromosome I and the first half of chromosome X; and (ii) between the end of chromosome II and almost the entire IV chromosome. For DSM350, we identified interaction effects between: (i) the end of chromosome IV and the first quarter of chromosome X; and (ii) the end of chromosome II and almost the entire IV chromosome (Fig. 3b). The latter seems to be specific for food patch leaving behavior as it was identified for both of the non-pathogenic bacteria (Fig. 3a, b). For the nematocidal B-18247, we found interaction effects at least between: (i) the beginnings of chromosome II and V; and (ii) the second quarter of chromosome II and the beginning of chromosome X (Fig. 3c). For the highly nematocidal B-18679, several interaction effects were identified including: (i) between

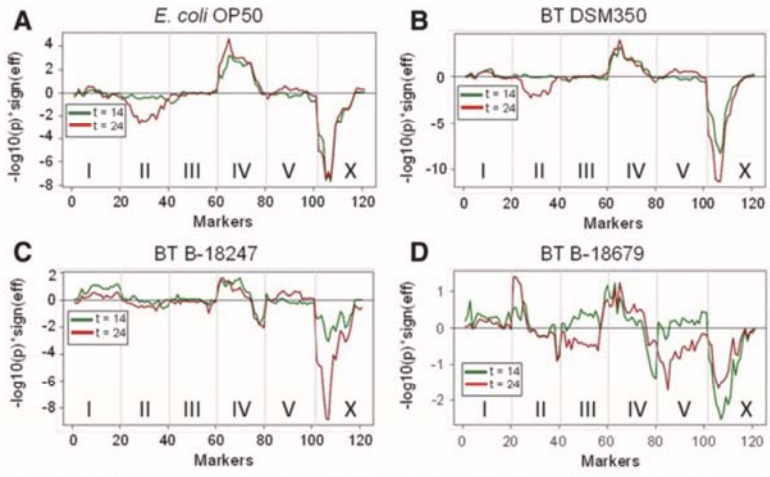


Fig. 2 QTL profiles for single marker mapping of avoidance behavior. QTL analysis of avoidance of: (**a**) *E. coli* OP50; (**b**) the non-nematocidal *B. thuringiensis* DSM350; (**c**) nematocidal B-18247; and (**d**) highly nematocidal B-18679. The *X-axis* shows the markers along the five autosomes and the X chromosome. Vertical light gray lines denote the boundaries between chromosomes. The *Y-axis* indicates the association between the chromosomal markers and variation in avoidance. Significance is indicated by −log10 of the *p*-value obtained from the linear model, which is multiplied by the sign of the effect to indicate the N2 allelic effect on avoidance. A value above 0 indicates an increase in leaving caused by the N2 allele; a value below 0 indicates an increase in leaving caused by the CB4856 allele. By convention, values above +2 or below -2 are considered to indicate a significant influence. *Green* and *red* lines show the results after either 14 h or 24 h exposure, respectively

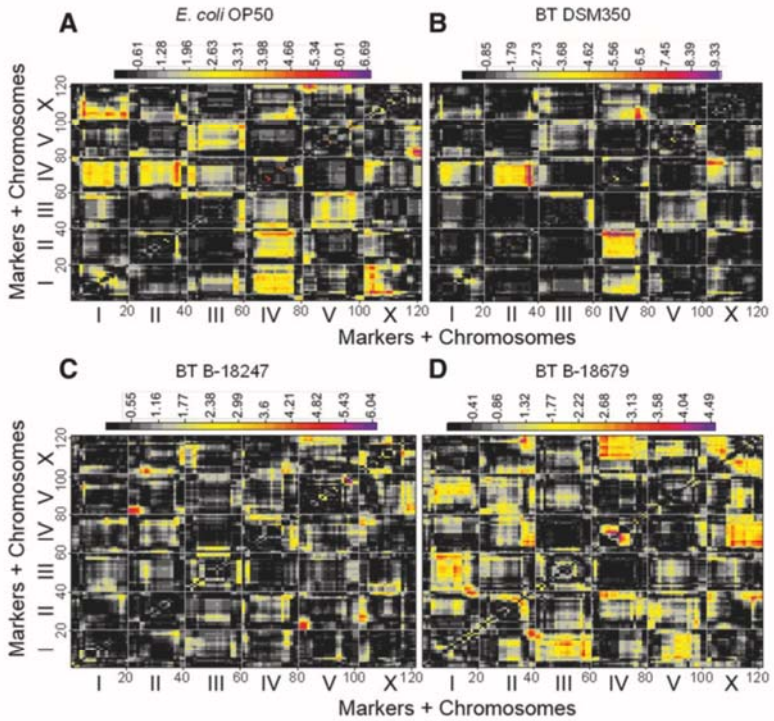


Fig. 3 Heat-map of interaction effects for avoidance behavior. Results for avoidance of: (**a**) *E. coli* OP50; (**b**) DSM350; (**c**) nematocidal B-18247; and (**d**) nematocidal B-18679. The distribution of markers across the genome is shown on both axes. The chromosome boundaries are indicated by the thin vertical and horizontal lines. A color legend for significance of the interaction between markers is shown on top of the panels. Significance is in −log10(p). High significance values are given in "warm" colors with purple and red indicating the highest significances, followed by orange and then yellow. For example, in panel (**c**) showing the results for B-18247, the thin red area towards the bottom right indicates a significant interaction between loci from the beginning of the X chromosome (*horizonal axis*) and the middle of chromosome II (*horizontal axis*). The other slightly larger red area in the bottom right of this panel points to a significant interaction between the beginning of chromosome V (*horizontal axis*) and the beginning of chromosome II (*vertical axis*)

the ends of chromosome I and II; and (ii) between the second quarter and the middle of chromosome IV (Fig. 3d). Interestingly, the X chromosome region with the strongest influence in the main effect model (Fig. 2) only contributed to very few significant interaction effects. One of these is an interaction with a chromosome I region,mediating avoidance of *E. coli* OP50 (Fig. 3a), and another one with a chromosome IV region, influencing avoidance of DSM350 (Fig. 3b).

Taken together, our results demonstrate that pathogen avoidance has a complex genetic architecture in *C. elegans*, which overlaps with, but differs from the response to non-pathogenic microbes. In particular, pathogen defense traits are related to the response to non-pathogenic bacteria, because they are affected by the same loci. Defense is thus in part determined by the general response to microbes, whereby pathogenicity of the bacteria may simply elevate the response mediated by a particular locus, as indicated for the X chromosome QTL (Fig. 2). Moreover, our results for the pathogen-specific QTLs are consistent with the previous finding that pathogen defense in invertebrate animals seems to rely on few loci and involve epistatic interactions among them [44, 45], possibly as a consequence of reciprocal coevolution among host and pathogens [44]. It will be a rewarding challenge for the future to characterize the genes underlying the pathogen-specific QTLs. Interestingly, the main effect QTL on the X chromosome was previously implicated in lawn leaving behavior with similar allelic effects towards the non-pathogenic *E. coli* OP50 (i.e., the CB4856 allele increases avoidance; [31]) but with opposite allelic effects towards *P. aeruginosa* (i.e., the N2 allele increases avoidance; [17–19]). In these cases, the QTL effect on chromosome X could be associated with variation in the gene *npr-1* [17–19, 31] and, at least towards *E. coli*, additionally the catecholamine receptor gene *tyra-3* [31].

The *npr-1* gene affects defense against *B. thuringiensis*

The gene *npr-1* was previously linked to the strong-effect QTL on chromosome X [17], which we found in the current study to influence *C. elegans* defense towards *B. thuringiensis* in the opposite way than that towards *P. aeruginosa*. We therefore specifically tested whether this gene is indeed responsible for the contrasting phenotypes in pathogen avoidance and resistance.

We first studied the role of *npr-1* in lawn leaving behavior on *E. coli* and *B. thuringiensis* using two mutants (*npr-1(ad609)*, *npr-1(ur89)*), which were both previously shown to decrease pathogen avoidance behavior against *P. aeruginosa* [17, 18]. Another gene from the left arm of chromosome X was previously demonstrated to influence food patch leaving behavior, namely *tyra-3*, which encodes a tyramine receptor homologue [31]. Therefore, we further tested its involvement in pathogen defense with the knock-out mutant *tyra-3(ok325)*.

Analysis of the two mutant *npr-1* alleles (*npr-1(ad609)*, *npr-1(ur89)*) yielded different results in avoidance and resistance (Figs. 4 and 5; sheets 2–4 of Additional files 2, 6, 7 and 8). In particular, avoidance behavior of *npr-1(ur89)* was similar to CB4856 but higher than that of N2 (Fig. 4, Additional file 6). In contrast, *npr-1(ad609)* always produced a low leaving rate, similar to N2 and clearly different from CB4856. On the highly pathogenic B-18679, avoidance behavior was extremely high at both time points without any significant differences among the *C. elegans* strains (Fig. 4d). In addition, the *tyra-3* mutant consistently showed a similar behavioral response to N2, irrespective of the bacterium and the exposure time (Fig. 4, Additional file 6).

The *npr-1* alleles produced similarly contrasting effects on survival rate, which is often used as a proxy for nematode immunity. For the RIL/IL parental strains, we found that CB4856 showed significantly higher resistance than N2 on both nematocidal *B. thuringiensis* strains (Fig. 5; Additional files 7, 8 and 9). Moreover, the *npr-1(ur89)* mutant was significantly more resistant than N2 and as resistant as CB4856 on both nematocidal pathogens, whereas

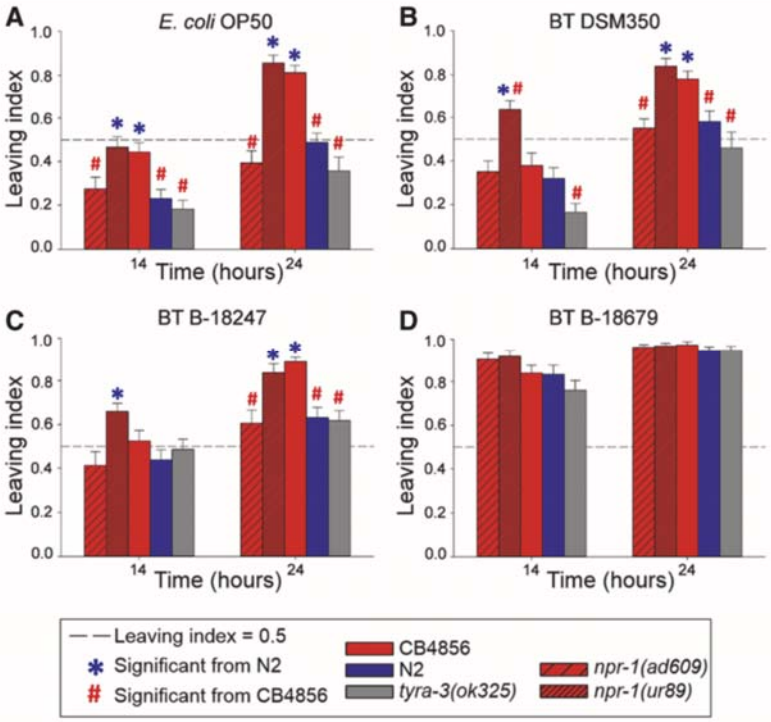


Fig. 4 Lawn leaving behavior for N2, CB4856, and different mutants towards *E. coli* and *B. thuringiensis*. Results for avoidance of: (**a**) *E. coli* OP50; (**b**) non-nematocidal *B. thuringiensis* DSM350; (**c**) nematocidal *B. thuringiensis* B-18247; and (**d**) highly nematocidal B-18679. The *Y axis* shows the proportion of escaped worms (i.e., leaving index); the *X axis* shows the two time points at which leaving was scored. The asterisk (*) is shown for the treatments which differ significantly from N2 whereas the hash sign (#) indicates significant differences from CB4856. The dotted reference line shows a leaving proportion of 0.5 for orientation among sub-panels. Error bars represent standard error of the mean. Raw data are given in sheet 2 of Additional file 2; detailed statistical results in Additional file 6

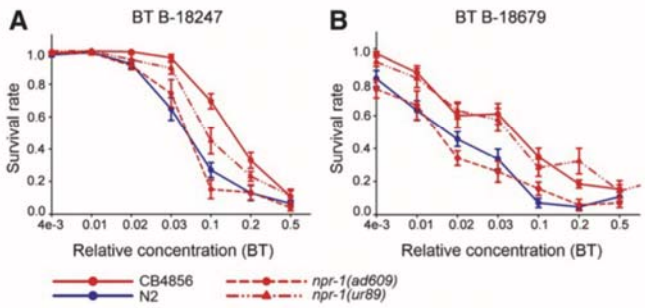


Fig. 5 Survival of N2, CB4856, *npr-1(ad609)* and *npr-1(ur89)* on nematocidal *B. thuringiensis*. Results for: (**a**) nematocidal B-18247; and (**b**) highly nematocidal B-18679. Survival on the *Y axis* was plotted against relative *B. thuringiensis* concentration on the *X axis*. Error bars represent standard error of the mean. The raw data are provided in sheet 4 of Additional file 2; the statistical results in Additional file 8

npr-1(ad609) was as susceptible as N2 on both pathogenic strains (Fig. 5, Additional file 8). None of the *C. elegans* strains showed any mortality under control conditions (results not shown).

We conclude that the two mutant *npr-1* alleles produce opposite effects on both behavioral avoidance of the four bacterial strains and also resistance against nematocidal *B. thuringiensis*. Consequently, variation in *npr-1* may only partially explain the strong main effect QTL on the X chromosome. The difference between the two *npr-1* alleles in avoidance and resistance of *B. thuringiensis* is surprising, because both alleles behaved similarly in previous studies investigating resistance against *P. aeruginosa* [17, 18]. Yet the two alleles carry different mutations: *npr-1(ad609)* two in exons 2 and 3, whereas the mutation of *npr-1(ur89)* falls into exon 3 (http://www.wormbase.org). The exact reasons for the different effects of these alleles clearly deserve further investigation in the future, ideally including additional loss-of-function and also reduced-function *npr-1* alleles in combination with a tissue-specific analysis of the mutational effects. We further conclude that the *tyra-3* gene does not appear to influence the assayed phenotypes (Figs. 4 and 5), including avoidance of *E. coli* OP50, which was however previously demonstrated in a separate study [31]. The difference in results could be due to variation in experimental approaches. For example, we directly characterized leaving behavior, whereas the previous study scored activity as a proxy for leaving behavior [31].

Contrasting effect of *npr-1* on defense against *Pseudomonas aeruginosa*

We sought to confirm the previously published finding [17–19] that the wild-type N2 produces higher resistance and stronger avoidance behavior towards the pathogen *P. aeruginosa* PA14 than the Hawaiian strain CB4856 and two *npr-1* mutants [17–19], thus contrasting with our above results for *B. thuringiensis*. Here, we specifically re-evaluated these previous results under our laboratory conditions and assay protocols, using the peptone-rich NGM plates required for expression of *P. aeruginosa* virulence. We first used the lawn leaving assay to assess the avoidance response against PA14 at different exposure time points. Consistent with previous findings, the *npr-1* mutants and CB4856 showed significantly lower PA14 pathogen avoidance than N2 across all time points (Fig. 6; sheets 5 and 6 of Additional files 2, 10, 11 and 12). For the 48 h time point the mutant *npr-1(ad609)* even had a lower leaving response than CB4856 (Fig. 6b). On OP50, leaving behavior was similar for all *C. elegans* strains at all time points except at time point 14 h, when the mutant *npr-1(ad609)* showed a more pronounced leaving behavior than N2 (Fig. 6a, Additional file 12). In this assay, we also included a *tyra-3* mutant, which expressed a similar leaving response to N2 under all conditions except at time point 48 h, where its leaving response against PA14 was reduced in comparison to N2, but still significantly higher than that of the remaining strains (Fig. 6b, Additional file 12).

We next evaluated the effect of *npr-1* on resistance against PA14 using standard *C. elegans* survival assays. All strains survived less on the pathogen PA14 than on the control OP50 (Fig. 7; sheet 7 of Additional files 2 and 13). On the pathogen, N2 was significantly more resistant than all other tested strains (Fig. 7b, Additional file 14).

We conclude that *npr-1* directly influences avoidance of PA14, in agreement with previous work [17–19]. In these studies, *npr-1* was suggested to affect PA14 resistance either as a consequence of hyperoxia avoidance behavior only (proposed by Reddy et al., [17]) or through both hyperoxia avoidance and the regulation of physiological immune responses (proposed by Styer et al., [18]).

Our results further demonstrate that the two *C. elegans* wild-type strains express opposite phenotypes on the two tested pathogens and that this contrast may be mediated at least partially by *npr-1*, as one of the *npr-1* alleles also produces an opposite phenotype relative to N2. Such opposite effects in the wild-type strains indicate specific

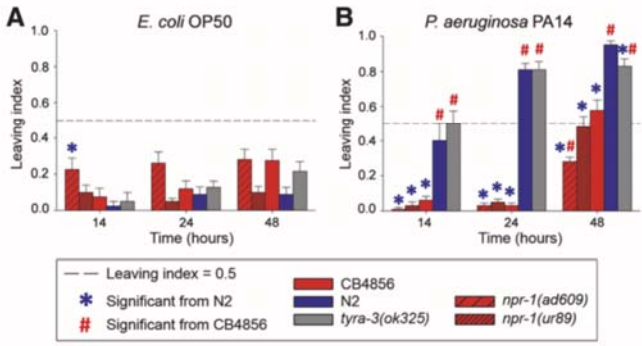


Fig. 6 Lawn leaving behavior for N2, CB4856 and different mutants towards *E. coli* and *P. aeruginosa*. Results for: (**a**) *E. coli* OP50; and (**b**) *P. aeruginosa* strain PA14. The *Y axis* shows the proportion of escaped worms; the *X axis* shows the three time points at which leaving was scored (12 h, 24 h, and 48 h). The asterisk (*) denotes the treatments which differ significantly from N2, whereas the hash sign (#) indicates significant differences from CB4856. The dotted reference line shows a leaving proportion of 0.5. Error bars represent standard error of the mean. The raw data are provided in sheet 6 of Additional file 2; detailed statistical results in Additional file 12

interactions with pathogens. The underlying genetics for such specificities have not yet been explored for behavioral immune defense. Some information is available for physiological and cellular immune specificities. In the higher vertebrates, such specificities can be mediated by components of the adaptive immune system such as the highly variable receptors of the major histocompatibility complex or the highly variable T and B cell receptors. Similar specificities have also been recorded in invertebrates [1, 46], where they may be due to different immune signaling cascades. For example, in *Drosophila*, the immune deficiency pathway appears to be more important in the systemic response to Gram-negative bacteria, whereas the Toll pathway is more important towards Gram-positive bacteria and fungi [47]. Moreover, in *C. elegans*, mutations in the *egl-9* and *tol-1* gene enhance resistance against some pathogens, while simultaneously increasing susceptibility to other pathogens (see introduction and [5–8, 10]). Our study thus provides one of the few examples which demonstrate that a single gene, in this case the neuropeptide Y receptor homolog gene *npr-1*, produces contrasting pathogen specificities in an invertebrate.

At the same time, it is less clear how exactly *npr-1* causes these contrasting phenotypes. Previous work on nematode social behavior demonstrated that *npr-1* influences worm aggregation, lawn bordering and clumping through its effect on aerotaxis behavior. The two tested *npr-1* alleles and also that of the CB4856 strain result in a preference towards lower oxygen concentrations usually found at the edge of the bacterial lawn [48, 49], whereas the N2 allele shows no such preference. A similar difference in aerotaxis behavior may also explain the reduced *P. aeruginosa* avoidance and resulting higher susceptibilities of the CB4856 and *npr-1* mutant strains, which remain in longer contact with the harmful pathogen, because the bacterial lawn boundaries show the preferred lower oxygen concentrations [17]. An involvement of such aerotaxis behavior in the *B. thuringiensis*

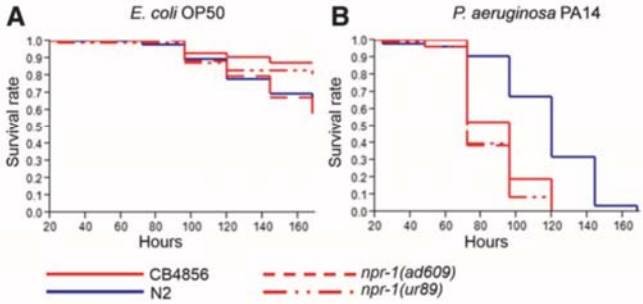


Fig. 7 Survival of N2, CB4856, *npr-1(ad609)* and *npr-1(ur89)* on *E. coli* and *P. aeruginosa*. Results for: (**a**) *E. coli* OP50; and (**b**) *P. aeruginosa* strain PA14. Survival (*Y axis*) was plotted against time in hours (*X axis*). We used the Kaplan-Meier method to calculate survival fractions. The raw data are given in sheet 7 of Additional file 2; statistical results in Additional files 13 and 14

response of CB4856 and one of the *npr-1* mutants would then require that the preferred oxygen concentration is outside of the bacterial lawn, which is unlikely to be the case (assuming higher oxygen concentrations outside, where no oxygen is consumed by proliferating bacteria). Thus, it is conceivable that CB4856 and the one *npr-1* mutant are directly responding to a compound produced by *B. thuringiensis*, and that this response is less pronounced in the N2 strain and the other *npr-1* mutant. In this context, it is worth noting that the high variation between N2 and CB4856 in their leaving response towards the control *E. coli* OP50 was only observed on peptone free PFM but not the peptone-rich NGM assay plates (Figs. 1a and 4a versus Figs. 6a and Additional file 10). This is most likely explained by bacterial proliferation, which is possible on NGM but not PFM assay plates. In turn, the lack of proliferation on the PFM plates is unlikely to coincide with large variation in oxygen concentration, such that an aerotaxis response should be less pronounced under these conditions. Yet, a non-proliferating, static bacterial population may produce particular metabolites, which then could have induced the CB4856 avoidance response.

npr-1 influences the transcriptomic response to *B. thuringiensis* and *P. aeruginosa*

To explore the mechanisms underlying the *npr-1* mediated contrasting effects on immune defense, we assessed whether *npr-1* differentially affects gene expression in the presence of either of the two pathogens. Using RNAseq we compared the transcriptomes of the N2 and *npr-1(ur89)* strains exposed to either the nematocidal *B. thuringiensis* B-18247, the pathogenic *P. aeruginosa* PA14, or the control *E. coli* OP50. We chose the mutant *npr-1(ur89)* because it showed differential leaving behavior and survival on both pathogens compared to N2 (Figs. 4 and 6). Exposure experiments were performed on Agar plates fully covered with bacterial lawns, thus reducing possible avoidance behaviors and producing comparable levels of lawn occupancy for the worms from the various treatment combinations. RNA transcript levels were characterized at two time points, 12 h and 24 h of pathogen exposure. We used principal component analysis (PCA) to explore which experimental factors generated different transcriptional responses. The first principal component indicated that the two nematode strains vary in their transcriptional signature to all three bacteria (Fig. 8). The second principal component highlights variation across several additional factors. The strongest effect stems from exposure time (light versus dark colors; Fig. 8). Additional influences can be seen for pathogen exposure versus the corresponding control, especially at the later time point (filled versus open symbols of the same type; Fig. 8) and also a clearly distinct signal after 24 h exposure to PA14 compared to all other conditions

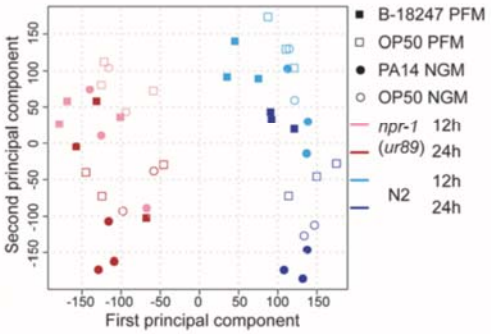


Fig. 8 Principal component analysis of transcriptomic variation. Variation is assessed for *npr-1(ur89)* and N2 upon exposure to pathogens (B-18247 and PA14) and the control bacterium OP50. N2 is shown in blue whereas *npr-1(ur89)* is shown in red. *Light* and *dark* colors indicate the early and late time point, respectively (12 h versus 24 h). Filled and open symbols denote exposure to pathogens (B-18247 and PA14) and control bacteria (OP50), respectively. NGM is the nematode growth medium enriched with peptone for the PA14 assay plates, PFM refers to peptone-free NGM used for the B-18247 assay plates

(filled dark colored circles towards the bottom of the graph; Fig. 8). These latter differences are more pronounced for N2 than the *npr-1* mutant, especially as N2 produces clearly distinct treatment signatures at the later 24 h time point (i.e., clearly separated dark blue open and filled circles and squares; Fig. 8). One possible reason for lower differentiation in the *npr-1* mutant may be a lower number of differentially expressed genes compared to the N2 strain. This was indeed the case, especially upon pathogen exposure (Table 1), suggesting that mutations in *npr-1* somehow compromise the signalling response to pathogen infection.

To identify groups of co-regulated genes, we next performed K-means clustering on the significant gene sets. The resulting eight clusters confirm that the transcriptional response is influenced by the *C. elegans* strain, the pathogen strain and also the exposure time point (Fig. 9; Additional file 15). In detail, clusters 1, 2, 3, and 4 refer to genes with strong differential expression upon exposure to only pathogenic *B. thuringiensis* B-18247 in only the *C. elegans* N2 strain, but neither the *npr-1(ur89)* mutant on the same pathogen nor any of the other treatments with *P. aeruginosa* (e.g., the stronger the color intensity in Fig. 9, the stronger the expression difference between pathogen versus non-pathogen exposure). This result again highlights that the *npr-1* mutant shows generally lower responsiveness in inducible gene expression (i.e., most clusters do not show high color intensity in Fig. 9). Clusters 1, 2, 3 and 4 only responded to the pathogen B-18247, and clusters 7 and 8 only or at least predominantly to PA14. Two clusters are specific to expression variation at the 12 h time point, in both cases upon

Table 1 Number of up- and down-regulated genes in the N2 and *npr-1(ur89)* strains

	B-18247		PA14	
C. elegans	12 h	24 h	12 h	24 h
Up-regulated				
N2	1393	1094	222	529
npr-1(ur89)	179	191	188	352
Down-regulated				
N2	2384	458	81	1233
npr-1(ur89)	117	70	58	370

RNAseq was used to assess variation in gene expression among the *C. elegans* N2 and *npr-1(ur89)* mutant strain in response to exposure to nematocidal *B. thuringiensis* B-18247 and *P. aeruginosa* PA14, always relative to the respective *E. coli* OP50 control. Gene expression variation was studied at two time points, 12 h and 24 h after initial exposure. The results are shown separately for the up- and down-regulated genes (top and bottom part of the table, respectively)

exposure to B-18247 (i.e., clusters 1 and 3), whereas four clusters indicate a more pronounced response at the later 24 h time point, either towards only B-18247 (clusters 2 and 4) or only PA14 (clusters 7 and 8). The remaining two clusters highlight patterns of early or continuous transcriptional response towards B-18247 and late transcriptional response towards PA14 (clusters 5 and 6; Fig. 9). None of the identified clusters showed an opposite gene expression pattern between either N2 and the *npr-1* mutant (e.g. up in N2 and down in the *npr-1* mutant) or the two pathogens (e.g. up after *B. thuringiensis* but down after *P. aeruginosa* exposure). Taken together, clusters 5 and 6 appear to encompass a general defense response against both pathogens, whereas the clusters 1, 2, 3, and 4 define the specific response to B-18247 and clusters 7 and 8 that to PA14. Therefore, the latter two groupings are likely to account for the observed *npr-1* dependent defense differences towards the two pathogens. We thus conclude that the considered mutation in the *npr-1* gene causes a decreased transcriptomic response to the two pathogens, which induce overlapping and distinct sets of differentially expressed genes.

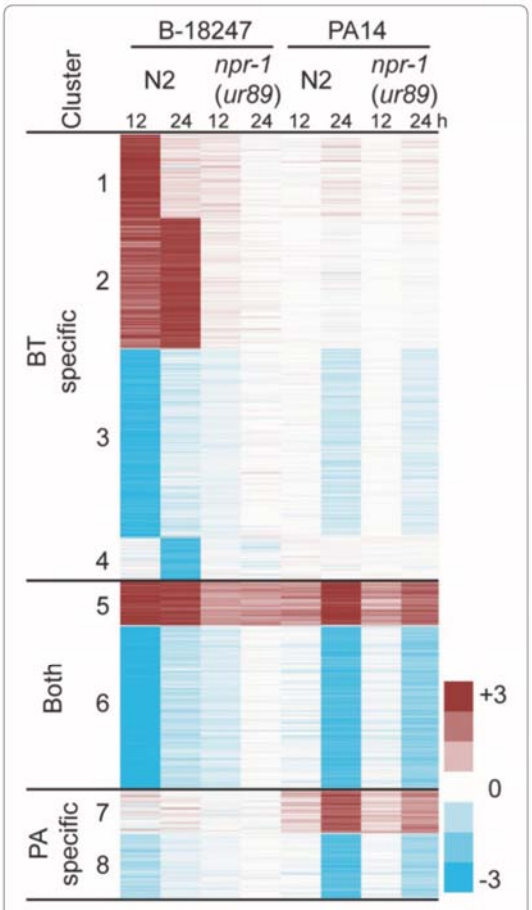


Fig. 9 Co-regulation of the differentially expressed genes. K-means cluster analysis yielded eight clusters of co-regulated genes, as indicated by the numbers on the left. The *top* of the graph shows the different treatment conditions, including *C. elegans* strain, pathogen strain and exposure time point. *Red* refers to up-regulated genes, whereas blue to down-regulated genes, always upon pathogen exposure relative to the corresponding *E. coli* OP50 control. High color intensities indicate strong expression differences to the control (see legend in the *bottom right* corner). The complete results are given in Additional file 15

Different functions and signaling processes are affected by the pathogen-dependent *npr-1*-specific transcriptome

We used enrichment analysis as a statistical tool to explore the possible functions of the differentially regulated gene clusters. Four types of enrichment analyses were performed, which aim to identify significant over-representation of (i) genes with a specific gene ontology (GO) term (GO term analysis); (ii) customized nematode-specific gene sets, inferred from previous gene expression analyses and based on the program EASE (EASE analysis); (iii) genes with specific transcription factor-binding motifs (Motif analysis), and (iv) expression QTLs (eQTLs). The customized enrichment analysis with the program EASE [50] was based on a large database of all previous *C. elegans* transcriptome studies, WormExp [51], which we collated from published work. These studies investigated differential gene expression (i) across development, (ii) in specific tissues, (iii) in worms with defined mutations or subjected to RNAi-knockdown of specific genes, or (iv) upon exposure to environmental stimuli such as pathogens, heavy metals, and other chemical compounds [51]. The GO term and Motif analyses were based on published methods, such as DAVID [52, 53] and AMD [54]. Analysis of eQTL enrichment expression differences, using the eQTL database collated from different previous eQTL analyses, all based on RIL panels

derived from N2 and CB4856 as parental lines [41, 55–60], thus potentially allowing us to link the identified QTLs to the expression variation inferred against pathogens. The results are summarized in Figs. 10 and 11, Table 2, Additional files 16 and 17, and explained in more detail below.

We first focus on the general defense response against the two pathogens, which is defined by two clusters (i.e, clusters 5 and 6; Fig. 9). These clusters are enriched for genes previously implicated in *C. elegans* pathogen defense. These include genes involved in carbohydrate binding (Fig. 10a), most likely mediated by C type lectin-like genes, many of which underlie this GO term (Additional file 17 under GO) and which are up-regulated across treatment conditions (Fig. 11a) and have repeatedly been implicated in *C. elegans* immunity, possibly as pathogen recognition receptors or antimicrobials [11, 61–64]. These two clusters are also enriched for genes which were previously shown to respond to exposure to the same pathogens and other types of stressors, such as heavy metals, osmotic stress, or pesticides (Fig. 10b). The upregulated genes appear to be

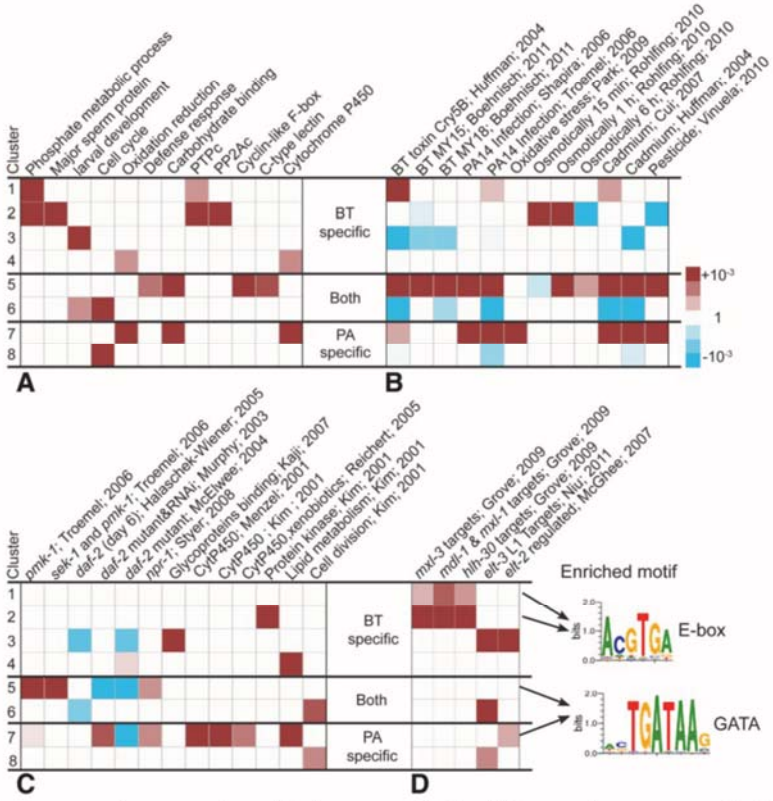


Fig. 10 Functional consequences of gene expression variation between *npr-1(ur89)* and N2 upon pathogen exposure. **a** Enrichment of gene ontology (GO) terms. The shown terms were significant with FDR < 0.05 (Additional file 17). **b** Overview of enrichment of pathogen- and stress-induced gene sets, inferred from EASE analysis on the various clusters of co-regulated genes. The significantly enriched gene sets are indicated on the top and include - from *left* to *right* - differentially expressed genes upon exposure to (i) *B. thuringiensis* Cry5B toxin [89]; (ii)-(iii) the same B-18247 strain used in the current study in *C. elegans* isolate MY15 or MY18 [90]; (iv)-(v) the same PA14 strain used here [66, 91]; (vi) Oxidative stress response [92]; (vii)-(ix) Osmotic induction [93]; (x)-(xi) Heavy metal Cadmium dysregulated genes [89, 94]; (xii) Pesticide influence [58]. **c** Overview of enriched gene sets for selected immunity pathways and general categories, including (i)-(ii) the p38 MAPK pathway (*pmk-1* and *sek-1* targets; [91]); (iii)-(v) insulin signalling (*daf-2* targets; [95–97]); (vi) *npr-1* targets [18]; (vii) Glycoproteins; (viii)-(x) Cytochrome P450 [98–100]; (xi) Protein kinase [98]; (xii) Lipid metabolism [98]; (xiii) Cell division [98]. **d** Enrichment of Ebox and GATA motifs and their transcriptional targets. The enriched gene sets were inferred with EASE and are indicated at the top, including differentially expressed genes in mutants of (i)-(iii) E-box transcriptions factors [101] or (iv)-(v) GATA transcription factors ELT-2 and ELT-3 [102, 103]. Enriched transcription factor binding motifs were inferred with AMD and are shown on the right (Additional file 17). In all panels, the clusters are given on the very left and are identical to those in Fig. 9. *Red* color indicates an enrichment for up-regulated genes per gene set, blue that for down-regulated genes per gene set. Color intensity corresponds to the significance level, inferred by EASE analysis (see scale at the *right side*)

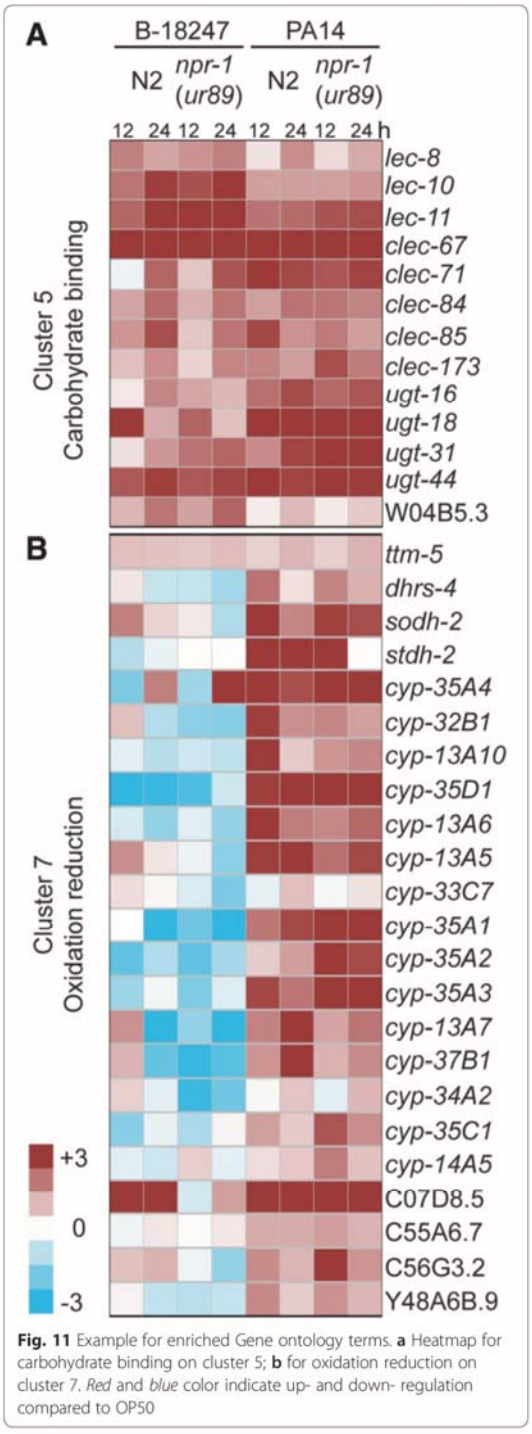


Fig. 11 Example for enriched Gene ontology terms. **a** Heatmap for carbohydrate binding on cluster 5; **b** for oxidation reduction on cluster 7. *Red* and *blue* color indicate up- and down- regulation compared to OP50

Table 2 eQTL enrichment analysis on identified expression clusters

Expression cluster	Location of eQTL		eQTL set
	Chromosome	Approximate position	
1	V	3.7 M	Rockman
2	IV	6.6 M	Viñuela (old)
3	X	15.5 M	Rockman
4	-	-	-
5	X	2.3 M	Rockman
	X	5.8 M	Rockman
	X	10.9 M	Rockman
	IV	6.3 M	Rockman
	III	1.9 M	Rockman
6	I	5.0 M	Viñuela (juvenile)
7	X	15.7 M	Rockman
	V	12.3 M	Rockman
8	I	3.9 M	Rockman

eQTL enrichment analysis was performed to identify significant overlaps between the genes underlying a specific cluster in our analysis (first column; see also Fig. 9) and previously characterized gene sets that define particular eQTLs (last column). Such overlaps can then be linked to the specific QTL regions within the genome (second and third column), which may then contain regulatory elements important for the expression variation in our study and which may also have been identified as QTLs for the observed variation in behavioral immune defense (Figs. 2 and 3). For further details see Additional file 16. No significant enrichment was found for cluster 4

controlled by two of the main *C. elegans* immunity signalling cascades, the p38 MAPK and the insulin-like receptor pathways (Fig. 10c) [2], and also the *npr-1* gene (Fig. 10c) [18]. They also show an enrichment in their promotor sequences for a GATA binding motif, although not for known targets of the GATA transcription factors ELT-2 and ELT-3 (Fig. 10d). They are also enriched for gene sets defined by eQTLs on chromosome I, III, the middle of chromosome IV, and the left arm and the middle of chromosome X (Table 2). One of the enriched X chromosome eQTLs encompasses the *npr-1* gene, another the gene *sek-1* of the p38 MAPK cascade, and the one on chromosome IV may include the MAPK gene *jnk-1* or the p38 homolog *pmk-1*, additionally supporting the role of these genes in the nematode's expression response. The enriched eQTLs from chromosome IV and the left arm of the X chromosome also lie within the QTLs identified to influence behavioral defense against *B. thuringiensis* (Fig. 2). Taken together, we conclude that clusters 5 and 6 comprise the components of a general defense response, apparently active not only against pathogens but also other stressors and mediated by central stress and immune response pathways. In the *npr-1* mutant, this defense response is strongly reduced towards both pathogens.

The specific response to *P. aeruginosa* is captured by two clusters (i.e., clusters 7 and 8; Fig. 9). They are enriched for eQTLs on chromosome I, V, and X (Table 2), although in

chromosomal regions without any known regulator of immune defense. These clusters, especially the upregulated cluster, are however enriched for genes previously shown to respond to infection by the same pathogen (Fig. 10b) and also those involved in the response to oxidative stress and xenobiotics, including cadmium and pesticides (Fig. 10a, b). The response to oxidative stress and xenobiotics is dominated by an up-regulation of cytochrome P450 genes (Fig. 11b; Additional file 17). The up-regulated cluster is further influenced by the two main immunity pathways, the *npr-1* gene, and also a GATA transcription factor (Figs. 10c, d). All of the latter components have previously been shown to be central for immune defense against *P. aeruginosa*, especially the p38 MAPK signalling cascade [65] and the GATA transcription factor ELT-2 [66]. The transcriptomic response to *P. aeruginosa* infection is additionally strongly influenced by cytochrome P450 expression, possibly as part of a general stress response to reduce oxidative stress (Fig. 10c). Because these responses are activated more strongly in the N2 strain, they are likely to mediate the observed higher resistance and behavioral defense for this strain compared to the *npr-1(ur89)* mutant (Figs. 6 and 7 and model in Fig. 12).

The specific response to *B. thuringiensis* is defined by four clusters, two up-regulated and two down-regulated groups of genes (i.e., clusters 1–4; Fig. 9). They are enriched for eQTLs in the middle of chromosome IV (Table 2), which could contain known defense regulators against *B. thuringiensis*, the MAPK genes *jnk-1* and *pmk-1* [67], and which lies within the QTL above identified to contribute to behavioral defense against this pathogen (Fig. 2). Enriched eQTLs are additionally found on the left arm of chromosome V and the right arm of the X chromosome (Table 2), in both cases without a link to any known immune regulator. Moreover, the upregulated clusters show an over-representation of genes involved in metabolic processes and phosphatase activity (Fig. 10a). They also include genes known to respond to the same pathogen and cadmium, as well as genes that are usually down- rather than up-regulated upon osmotic and pesticide stress (Fig. 10b). These two upregulated clusters are enriched for an Ebox transcription factor binding motif and the corresponding targets (Fig. 10d). The two down-regulated clusters show an enrichment for oxidation reduction and developmental genes (Fig. 10a), genes responsive to *B. thuringiensis* and cadmium (Fig. 10b), and genes controlled by insulin-like signalling, including glycoproteins (Fig. 10c). One of the down-regulated clusters also shows an enrichment of the GATA transcription factor targets (Fig. 10d).

Taken together, the results of our enrichment analyses allow us to propose a possible mechanistic basis for the contrasting defense effects of the *npr-1* gene. It is worth

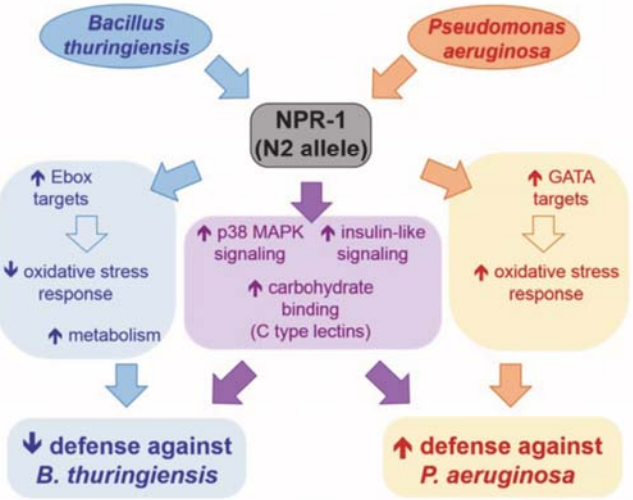


Fig. 12 Model of *npr-1* dependent effects on pathogen defense in the N2 *C. elegans* strain. Exposure to both pathogens leads to an *npr-1* dependent activation of carbohydrate-binding factors, such as C-type lectin-like proteins, and also two central immune signaling cascades, the p38 MAPK and the insulin-like pathways, which could all enhance pathogen resistance (*middle* part of the graph). Upon exposure to *P. aeruginosa* PA14 (*right side*), *npr-1* also influences the activation of a general stress response, via one or several GATA transcription factors(s), which increases oxidative stress resistance and thus resistance to the pathogen. The response to *B. thuringiensis* (*left side*) is mediated by *npr-1* through one or several Ebox transcription factors, resulting in a reduced oxidative stress response and increased metabolic activity as a possible cause of enhanced pathogen susceptibility. *Arrows* with light colors indicate uncertain connections

reiterating that N2 and the mutant only differ in the presence of a mutation in the *npr-1* gene. Therefore, the differences observed between strains must be influenced by the allelic variation in this gene. The higher resistance and avoidance behavior of N2 towards *P. aeruginosa* is likely influenced by the activation of GATA and/or p38 MAPK targets and/or the induced oxidative stress response (see above; Fig. 12). The situation is less clear for the response to *B. thuringiensis*. Because N2 produces lower defenses against *B. thuringiensis* than the *npr-1(ur89)* mutant (Figs. 4 and 5), and because any differential gene expression is repressed in the *npr-1(ur89)* mutant (Table 1; Fig. 9), the specific activation of certain genes in N2 (i.e., for the two up-regulated clusters 1 and 2) and/or the suppression of other gene groups in N2 (i.e., the down-regulated clusters 3 and 4) must account for the observed lower resistance and avoidance response against this pathogen. We hypothesize that this may possibly be mediated by one of the following two processes or a combination thereof (see model in Fig. 12): (i) the lower oxidative stress response in N2 could lead to increased susceptibility towards *B. thuringiensis*, in analogy to the effect recently described towards *E. faecalis* [68] and assuming that *B. thuringiensis* toxicity causes oxidative stress (which is however currently unknown); and/or (ii) an activation of metabolic processes could be disadvantageous during pathogen infection, because it may reduce availability of energetic resources that can be invested in immune defense and because metabolic products may be exploited as a source of nutrition by the pathogen. Any of the other implicated functions (Fig. 10) may also contribute to enhanced susceptibility in an as yet unknown manner. These processes then seem to be influenced by *npr-1* through Ebox-specific transcription factors. Interestingly, the higher susceptibility is apparently caused by an activation of specific functions and signalling processes rather than their absence or at least reduced activity. This may indicate a sub-optimal response to this specific pathogen in the N2 strain or represent a consequence of pathogen-mediated manipulation of host responses, which are widespread among pathogens [69] and which have also been shown for another *Bacillus* species, *Bacillus nematocida*, to change *C. elegans* behavior and intestinal responses [70]. At the moment, it is unclear in what way the indicated processes influence either behavioral or physiological responses or both simultaneously. This represents a challenging topic for future research.

Conclusion

Our study revealed a complex genetic architecture comprising several epistatically interacting QTLs associated with variation in *C. elegans* pathogen avoidance behavior. The most significant QTL encompassed the gene *npr-1*. Our functional analyses of this gene revealed a contrasting effect of *npr-1* on *C. elegans* immune defense, as assessed through both behavioral and also survival phenotypes. In particular, the CB4856 allele was associated with faster lawn leaving behavior and higher survival than the N2 allele on *B. thuringiensis*, whereas it was associated with lower lawn leaving behavior and lower survival on *P. aeruginosa*. A further characterization of the exact role of *npr-1* suggested that it mediates differential regulation of defense genes via either GATA transcription factors leading to increased immune defense towards *P. aeruginosa* or Ebox transcription factors leading to decreased immune defense towards *B. thuringiensis*. Our study thus demonstrates that a single gene in *C. elegans* mediates contrasting pathogen-specific defense responses.

Methods

C. elegans and bacterial strains

C. elegans strains: (i) the standard wild-type strains N2 and CB4856; (ii) 200 Recombinant Inbred Lines (RILs) and 90 Introgression lines (ILs) generated from crosses between N2 and CB4856 [41, 42, 55, 60]; and (iii) two distinct mutant alleles of *npr-1*, *npr-1(ur89)* X (strain IM222) and *npr-1(ad609)* X (strain DA609), and also the *tyra-3* knock-out allele *tyra-3(ok325)* X (strain VC125). The three mutant strains were obtained from the Caenorhabditis Genetics Center (CGC; http://www.cbs.umn.edu/CGC/) and were all generated in the N2 background. All worm strains were maintained at 20 °C on Nematode Growth Medium (NGM) plates with the non-pathogenic *E. coli* OP50 as an *ad libitum* food according to standard protocols [71]. All mutants were backcrossed at least three times to N2. Presence of the target mutation was confirmed for the two *npr-1* mutants by sequencing the *npr-1* gene at the location of the mutations and for the knock-out mutant *tyra-3* by polymerase chain reaction (PCR) analysis of the deleted region.

Bacterial strains: (i) two nematocidal strains of *B. thuringiensis*, NRRL B-18247 and NRRL B-18679, originally provided by the Agriculture Service Patent Culture Collection (United States Department of Agriculture, Peoria, Illinois, USA); (ii) the non-nematocidal *B. thuringiensis* strain DSM350, originally obtained from the German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, DSMZ, Braunschweig, Germany); (iii) the pathogenic *P. aeruginosa* strain PA14, obtained from Dennis H. Kim, Boston, USA; and (iv) the non-pathogenic *E. coli* OP50. Before the start of this study, the three *B. thuringiensis* strains were cultured in large quantities as previously described [32]. The cultures consisted mainly of spores associated to nematocidal toxins in the case of B-18679 and B-18247, and non nematocidal toxins in the case of the control DSM 350. All cultures were set to a concentration of 1.5×10^{10} particles/ml,

assessed through standard Thoma counting chambers and microscopic analysis. The cultures were cryo-preserved in aliquots at −20 °C; spore viability and pathogenicity are not affected by freezing under these conditions [38, 40, 72]. Usage of viable spore aliquots from the same starting culture allowed us to minimize variation across experiments, thus enhancing comparability of the data from different study approaches. In all cases, aliquots were thawed directly before each experiment. The bacterial cultures were then diluted, as indicated below in the description of the various assays.

Lawn leaving assays

The assay was designed for 9 cm petri dishes with either Peptone free NGM (PFM) for *B. thuringiensis* assays or peptone-rich NGM for PA14 assays. 30 μl of the test bacterium were spotted onto the center of the plate and 80 μl of *E. coli* OP50 only were additionally placed in the shape of a ring (Additional file 1: Figure S1). This ring of OP50 protects escaping worms from starvation, minimizing their return to the lawn in the center. The test bacterium consisted of either *B. thuringiensis* diluted with OP50 at 1:250 from a stock with a concentration of 1.5×10^{10} particles/ml, or PA14 diluted with OP50 in a 4:1 ratio. 10 hermaphroditic fourth instar larvae (L4) were picked onto the test lawn. Experiments were performed at 20 °C.

Leaving behavior was recorded by counting the number of worms on the lawn at different time points of exposure and calculated according to the following formula:

$$\text{Leaving index} = \frac{10 - \text{number of worms on the lawn}}{10}$$

The screens of RILs (200 lines) and ILs (90 lines) were done using a randomized block design on 17 dates, always including the parental strains of these lines, N2 and CB4856, as internal controls. Each *C. elegans* strain-bacteria-time point combination was assayed in three replicates, resulting in a total of 35040 individual data points. The screens of *npr-1* mutants included 12 replicates of each treatment.

We would like to note that the leaving assay for PA14 was performed at 20 °C and thus at a different temperature than the standard PA14 survival assays at 25 °C (see below and [19]). The reason is that the 25 °C temperature led to increased bacterial growth on the assay plate, which caused enhanced dispersal of bacterial colonies through the crawling worms, thus compromising reliable scoring of the leaving behavior. Such a bias was not observed at 20 °C. As our results did confirm previously published data on *C. elegans* avoidance behavior towards PA14 [19], our assay conditions allowed us to characterize a robust behavioral response against this pathogen.

Survival assays

For survival analysis with *B. thuringiensis*, 6 cm peptone free NGM plates were inoculated with 100 μl of a mixture of *B. thuringiensis* with *E. coli* OP50. Mixtures were prepared in seven dilutions: 1:2, 1:5, 1:10, 1:30, 1:50, 1:100, and 1:250 (equivalent to the relative concentration given in the main text). Plates were left to dry overnight (9–15 h) at 20 °C. 30 L4 hermaphrodites were picked onto each assay plates. After 24 h, survival was recorded by counting alive, dead and missing worms. Each treatment group was replicated 8 times across 8 runs (one replicate per run).

Analysis of *P. aeruginosa* effects was based on 3 cm peptone-rich NGM plates, which were inoculated with 5 μl of an overnight culture at 37 °C of either PA14 or OP50. Seeded plates were incubated overnight at 37 °C and then at 25 °C. 30 L4 hermaphrodites were picked onto each assay plate and stored at 25 °C in the dark. Alive and dead worms were scored every 24 h and surviving worms were transferred to new assay plates every 48 h. Each treatment group was replicated 10 times across two runs.

Statistical analysis of phenotypic data

We used the non-parametric Kruskall-Wallis test to assess differences in leaving behavior between the *C. elegans* strains, and a Bonferroni based adjustment to correct for multiple testing. We used Kaplan Meier analysis applying the Log Rank test to assess differences in *C. elegans*' survival on PA14. A Bonferroni based adjustment was used to correct for multiple testing. We used GLM ordinal logistic regression analysis to assess differences in survival between the *C. elegans* strains across the concentration range of *B. thuringiensis*, using *C. elegans* strains, BT concentration and the interaction between the two as factors. A Bonferroni based adjustment was used to correct for multiple testing. All escape and survival assays data were analyzed using the program JMP version 9.0 (SAS Institute Inc.), while graphic illustrations were produced with the program SIGMAPLOT version 12.0 (Systat Software Inc.).

Quantitative trait locus (QTL) analysis

The QTL analysis was performed on the average of three replicates per genotype/line of the calculated proportion "leave" (see assay method) of 200 Recombinant Inbred Lines (RILs) [41, 55–58, 73–76]. QTLs were calculated by single marker mapping using a linear model (trait ~ marker + error) for each marker using a custom written script in the statistical programming language "R" [77]. Significance levels were estimated from 1000 permutations

of the data. The analysis calculated the significance of the linkage between the genetic marker and the trait one by one for each bacterium and exposure time point separately. We furthermore evaluated epistatic interactions between each two markers across the genome for each bacterium separately using the following model: phenotype ~ time + marker1 * marker2. BIN mapping in the set of 90 ILs was done as described in [42, 56].

RNA isolation and sequencing

N2 and *npr-1(ur89)* worms were exposed to either PA14, BT B-18247 or OP50 for either 12 or 24 h of exposure. The experiment had 3 replicates of each treatment combination (a total of 36 samples across treatment combinations and replicates). Exposure experiments were performed on large Agar plates (15 cm diameter), which were fully covered with a bacterial lawn, thus minimizing escape responses and resulting in comparable occupancy rates across the treatment combinations. At both time points (12 & 24), worms were washed off the exposure plates with PBS containing 0.3 % Tween20® and resuspended in TRIzol® (Life Technologies) reagent. Prior to RNA extraction, worm suspensions were treated five times with a freeze-thaw cycle using liquid nitrogen and a thermo block at 45 °C. RNA extraction was performed using a NucleoSpin® mRNA extraction kit (Macherey-Nagel). RNA samples were treated with DNAse, and then stored at −80 °C. RNA libraries were prepared for sequencing using standard Illumina protocols. Libraries were sequenced on an Illumina HiSeq™ 2000 sequencing machine with a paired-end strategy and read length of 100 nucleotides.

Statistical analysis of transcriptomic data

RNAseq reads were mapped to the *C. elegans* genome from Wormbase version WS235 (www.wormbase.org) by Tophat2 [78] using option *–b2-very-sensitive*, other default settings and without a transcriptome reference. Tophat2 aligns RNAseq reads to a genome based on the ultra-fast short read mapping program Bowtie [79]. Estimation of transcript abundance and significantly differentially expressed genes were identified by Cuffdiff [80] using the quartile normalization method [81]. Cuffdiff is a program from the Cufflinks package and aims to find significant changes in transcript expression in consideration of possible formation of isoforms for a particular gene. The raw data is available from the GEO database [82, 83] under the GSE number GSE60063.

For clustering and visualization, transcripts with a significant change between different conditions (adjusted p-value < 0.01 by the false discovery rate, FDR) were treated as signature for each comparison. Due to the biological variation of the replicates, the p-value, instead of fold-change, of those genes were firstly log10-transformed and ordered according to increasing or decreasing expression and then taken as input for k-means cluster analysis using cluster 3.0 [84] with a k of 8. A heatmap was generated by TreeView version 1.1.4r3 [85]. Principal component analysis (PCA) was carried out on log-transformed gene expression profiles using a probabilistic PCA algorithm [86] from R package pcaMethods [87], which links PCA to the probability density of patterns. Dimensionality of the samples was reduced from 57165 (total isoforms) to three dimensions (PCs). Motif analysis was carried out on the promoter regions, -600 bp and 250 bp relative to transcription start sites (TSS), of genes in each group. De novo motif discovery was performed by AMD [54].

GO and EASE analysis

Gene ontology (GO) and a gene set enrichment analysis was carried out on each group of genes from the K means cluster analysis. GO analysis was performed using DAVID [52, 53] with a cut-off of p-value < 0.05, adjusted by FDR. For the gene set analysis, we used EASE [50], a free, stand-alone software package from DAVID bioinformatics resources (http://david.abcc.ncifcrf.gov/). As recently described [51, 64], we constructed a *C. elegans*-specific gene set database, WormExp [51], from published data and also using the previously established data sets collected by Ilka Engelmann et al., [88]. Based on this data set, we performed the EASE analysis and selected the results with a Bonferroni adjusted p-value < 0.05.

eQTL enrichment analysis

eQTL enrichment was done using a hypergeometric test in R. The eQTL sets [41, 57–59] were obtained from WormQTL.org [55, 60]. The eQTLs at a specific locus were compared to the genes in a specific expression cluster, as identified from the above described K-means cluster analysis.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets, supporting the conclusions of this article, is available in case of the QTL analysis from WormQTL [55, 60], in case of the phenotypic analysis of N2, CB4856, and the *npr-1* mutants in Additional file 2 of this article, and in case of the transcriptome analysis from the GEO database [82, 83] under the GSE number GSE60063. The *C. elegans* strains are available from the Caenorhabditis Genetic Center (CGC), which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). All bacterial strains are available from the corresponding author upon request.

Additional files

Additional file 1: Illustration of the lawn leaving assay. 10 hermaphrodites at the L4 stage were transferred by picking onto 9 cm peptone free NGM plates containing a lawn of the tested bacteria, each mixed with *E. coli* OP50 and surrounded by a ring of 80 μl of OP50. (TIF 144 kb)

Additional file 2: Raw data for the analysis of N2, CB4856, and the npr-1 mutants. Sheet 1, Lawn leaving behavior of N2 and CB4856 from Bacillus thuringiensis - Raw data for Fig. 1; sheet 2, Lawn leaving behavior of mutants from Bacillus thuringiensis - Raw data for Fig. 4; sheet 3, Surival of N2 and CB4856 on Bacillus thuringiensis - Raw data for Additional file 9; sheet 4, Surival of mutants on Bacillus thuringiensis - Raw data for Fig. 5; sheet 5, Lawn leaving behavior of N2 and CB4856 from Pseudomonas aeruginosa PA14 - Raw data for Additional file 10; sheet 6, Lawn leaving behavior of mutants from Pseudomonas aeruginosa PA14 - Raw data for Fig. 6; sheet 7, Surival of mutants on Pseudomonas PA14 - Raw data for Fig. 5. (XLS 482 kb)

Additional file 3: Table on the statistical results for the comparison between N2 and CB4856 leaving behavior towards *B. thuringiensis* and *E. coli*. (PDF 83 kb)

Additional file 4: Table on the statistical results for the pairwise comparisons of the leaving response on the *E.coli* OP50 control versus the *B. thuringiensis* strains. (PDF 87 kb)

Additional file 5: Figure on the leaving phenotypes of the introgression lines (ILs) plotted against the introgression position along the chromosomes. (A) Results for *E. coli* strain OP50; (B) non-nematocidal *B. thuringiensis* strain DSM350; (C) nematocidal *B. thuringiensis* B-18247; and (D) highly nematocidal *B. thuringiensis* B-18679. Green and red lines show the results after either 14 h or 24 h exposure, respectively. Position of markers is given along the X axis. Light gray vertical lines indicate boundaries of the chromosomes. (TIF 1279 kb)

Additional file 6: Table on the statistical results for the comparison of the N2 and CB4856 leaving behavior with that of the mutant strains towards *B. thuringiensis* and *E. coli*. (PDF 96 kb)

Additional file 7: Table on the statistical results for the separate comparison between *C. elegans* N2 and CB4856 survival on nematocidal *B. thuringiensis*. (PDF 74 kb)

Additional file 8: Table on the statistical results for the pairwise comparison of survival of the *C. elegans* strains on nematocidal *B. thuringiensis*. (PDF 78 kb)

Additional file 9: Figure on the separate analysis of N2 and CB4856 survival in the presence of nematocidal *B. thuringiensis*. (A) Survival on *B. thuringiensis* strain B-18247; and (B) B-18679. Survival on the Y axis is plotted against BT concentration on the X axis. Error bars represent standard error of the means. The statistical results are given in Additional file 6. (TIF 106 kb)

Additional file 10: Figure on the separate analysis of lawn leaving behavior of N2 and CB4856 towards *E. coli* and *P. aeruginosa*. (A) Results for avoidance of *E. coli* strain OP50; and (B) *P. aeruginosa* strain PA14. The asterisk (*) points to a significant difference to N2. The dotted reference line indicates the 0.5 avoidance response. Statistical results are shown in Additional file 10. (TIF 97 kb)

Additional file 11: Table on the statistical results for the comparison between N2 and CB4856 leaving behavior towards *E. coli* and *P. aeruginosa*. (PDF 76 kb)

Additional file 12: Table on the statistical results for the comparison of the N2 and CB4856 leaving behavior with that of the mutant strains towards *P. aeruginosa* and *E. coli*. (PDF 90 kb)

Additional file 13: Table on the statistical results for the comparison of *C. elegans* survival rate on *P. aeruginosa* PA14 versus the control *E. coli* OP50. (PDF 74 kb)

Additional file 14: Table on the statistical results for the pairwise comparisons of *C. elegans* survival rates on *P. aeruginosa* PA14. (PDF 72 kb)

Additional file 15: Table with the list of differentially expressed genes after exposure of the *C. elegans* N2 wildtype or *npr-1* mutant to pathogenic *B. thuringiensis* B-18247, pathogenic *P. aeruginosa* PA14, or non-pathogenic *E. coli*. (XLS 2096 kb)

Additional file 16: Results of the eQTL enrichment analysis of the K means clusters of differentially expressed genes. (XLS 160 kb)

Additional file 17: Results of the GO term enrichment analysis (first sheet), the motif analysis (second sheet), and the *C. elegans*-specific EASE enrichment analysis (third sheet) of the K means clusters of differentially expressed genes. (XLS 316 kb)

Abbreviations

AMD: automated motif discovery; DAVID: database for annotation, visualization and integrated discovery; EASE: expression analysis systematic explorer; eQTL: expression quantitative trait locus; FDR: false discovery rate; GEO: gene expression omnibus; GLM: generalized linear model; GO: gene ontology; PCA: principal component analysis; QTL: quantitative trait locus.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RN conceived the study, performed and participated in all experiments, analyzed the data, and drafted the manuscript. LBS conceived and performed QTL analysis and drafted the manuscript. WY performed transcriptomic analysis and drafted the manuscript. SE, FS, TGM, LR, ACM, KD helped performing phenotypic and functional genetic experiments. PCR conceived and participated in the transcriptomic analysis. JEK provided the RIL and IL libraries, conceived the QTL analysis, and drafted the manuscript. HS conceived the study, analyzed the data, and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to all members of the Schulenburg and Kammenga labs for support and advice. We thank particularly Christiana Anagnostou, Daniela Haase, Leila Masri, Barbara Pees, Joost Riksen, Stefanie Rohwer, and Anna Sheppard. We also thank the Kiel ICMB sequencing team (especially Markus Schilhabel, Melanie Friskovec, Melanie Schlapkohl) for technical assistance. The *C. elegans* strains were provided by the Caenorhabditis Genetic Center (CGC), which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We thank the CGC for providing these strains and Dennis Kim for providing the *P. aeruginosa* strain PA14.

Funding

We are most grateful for financial support from the German Science Foundation to HS (DFG grants SCHU 1415/6 and SCHU 1415/9); Kiel University to HS and KD; infra-structural funds from the DFG Excellence Cluster Inflammation at Interfaces to HS and PR; the Netherlands Organisation for Scientific Research to LBS and JEK (grants 823.01.001 and 855.01.151); the HFSP to JEK; and the International Max Planck Research School (IMPRS) for Evolutionary Biology to WY.

Author details

[1]Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, 24098 Kiel, Germany. [2]Cologne Excellence Cluster for Cellular Stress Responses in Ageing-Associated Diseases (CECAD) and Systems Biology of Ageing, University of Cologne, Joseph-Stelzmann-Str. 26, 50931 Cologne, Germany. [3]Laboratory of Nematology, Wageningen University, Wageningen 6708 PB, The Netherlands. [4]Institute for Clinical Molecular Biology, University of Kiel, 24098 Kiel, Germany.

Received: 7 November 2015 Accepted: 25 March 2016
Published online: 11 April 2016

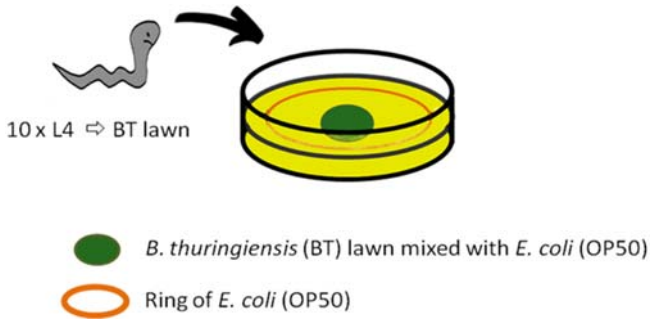
References

1. Schulenburg H, Boehnisch C, Michiels NK. How do invertebrates generate a highly specific innate immune response? Mol Immunol. 2007; 44(13):3338–44.
2. 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.
3. Buchon N, Silverman N, Cherry S. Immunity in *Drosophila melanogaster* - from microbial recognition to whole-organism physiology. Nat Rev Immunol. 2014;14(12):796–810.

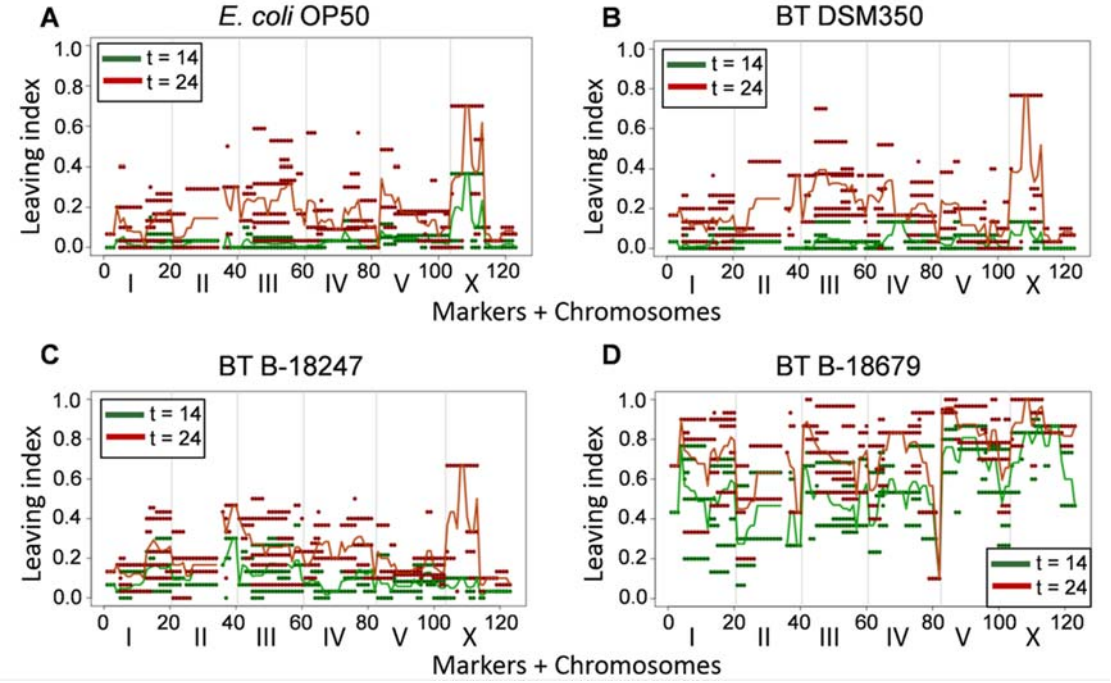
4. Meisel JD, Kim DH. Behavioral avoidance of pathogenic bacteria by *Caenorhabditis elegans*. Trends Immunol. 2014;35(10):465–70.
5. Ewbank JJ, Pujol N. Local and long-range activation of innate immunity by infection and damage in *C. elegans*. Curr Opin Immunol. 2016;38:1–7.
6. Luhachack LG, Visvikis O, Wollenberg AC, Lacy-Hulbert A, Stuart LM, Irazoqui JE. EGL-9 controls *C. elegans* host defense specificity through prolyl hydroxylation-dependent and-independent HIF-1 pathways. PLoS Pathog. 2012;8(7):e1002798.
7. Shao Z, Zhang Y, Ye Q, Saldanha JN, Powell-Coffman JA. *C. elegans* SWAN-1 Binds to EGL-9 and regulates HIF-1-mediated resistance to the bacterial pathogen Pseudomonas aeruginosa PAO1. PLoS Pathog. 2010;6(8): e1001075.
8. Bellier A, Chen C-S, Kao C-Y, Cinar HN, Aroian RV. Hypoxia and the hypoxic response pathway protect against pore-forming toxins in *C. elegans*. PLoS Pathog. 2009;5(12):e1000689.
9. Tenor JL, Aballay A. A conserved Toll-like receptor is required for *Caenorhabditis elegans* innate immunity. EMBO Rep. 2008;9(1):103–9.
10. Pujol N, Link EM, Liu LX, Kurz CL, Alloing G, Tan M-W, Ray KP, Solari R, Johnson CD, Ewbank JJ. A reverse genetic analysis of components of the Toll signaling pathway in *Caenorhabditis elegans*. Curr Biol. 2001; 11(11):809–21.
11. Schulenburg H, Ewbank JJ. The genetics of pathogen avoidance in *Caenorhabditis elegans*. Mol Microbiol. 2007;66(3):563–70.
12. Petersen C, Dirksen P, Schulenburg H. Why we need more ecology for genetic models such as *C. elegans*. Trends Genet. 2015;31:120–127.
13. Felix MA, Braendle C. The natural history of *Caenorhabditis elegans*. Curr Biol. 2010;20(22):R965–9.
14. Felix M-A, Duveau F. Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and *C. briggsae*. BMC Biol. 2012;10:59.
15. Petersen C, Dirksen P, Prahl S, Strathmann EA, Schulenburg H. The prevalence of *Caenorhabditis elegans* across 1.5 years in selected North German locations: the importance of substrate type, abiotic parameters, and Caenorhabditis competitors. BMC Ecol. 2014;14:4.
16. Pradel E, Zhang Y, Pujol N, Matsuyama T, Bargmann CI, Ewbank JJ. Detection and avoidance of a natural product from the pathogenic bacterium *Serratia marcescens* by *Caenorhabditis elegans*. Proc Natl Acad Sci U S A. 2007;104(7):2295–300.
17. Reddy KC, Andersen EC, Kruglyak L, Kim DH. A polymorphism in *npr-1* is a behavioral determinant of pathogen susceptibility in *C. elegans*. Science. 2009;323(5912):382–4.
18. Styer KL, Singh V, Macosko E, Steele SE, Bargmann CI, Aballay A. Innate immunity in *Caenorhabditis elegans* is regulated by neurons expressing NPR-1/GPCR. Science. 2008;322(5900):460–4.
19. Chang HC, Paek J, Kim DH. Natural polymorphisms in *C. elegans* HECW-1 E3 ligase affect pathogen avoidance behaviour. Nature. 2012;480(7378):525–9.
20. McMullan R, Anderson A, Nurrish S. Behavioral and immune responses to infection require Gαq- RhoA Signaling in *C. elegans*. PLoS Pathog. 2012;8(2): e1002530.
21. Hasshoff M, Boehnisch C, Tonn D, Hasert B, Schulenburg H. The role of *Caenorhabditis elegans* insulin-like signaling in the behavioral avoidance of pathogenic *Bacillus thuringiensis*. FASEB J. 2007;21(8):1801–12.
22. Zhang Y, Lu H, Bargmann CI. Pathogenic bacteria induce aversive olfactory learning in *Caenorhabditis elegans*. Nature. 2005;438(7065): 179–84.
23. Schulenburg H, Muller S. Natural variation in the response of *Caenorhabditis elegans* toward *Bacillus thruingiensis*. Parasitology. 2004;128:433–43.
24. Volkers RJM, Snoek LB, van Hellenberg Hubar CJ, Coopman R, Chen W, Yang W, Sterken MG, Schulenburg H, Braeckman BP, Kammenga JE. Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild *Caenorhabditis elegans* populations. BMC Biol. 2013;11(1):93.
25. Sicard M, Hering S, Schulte R, Gaudriault S, Schulenburg H. The effect of *Photorhabdus luminescens* (Enterobacteriaceae) on the survival, development, reproduction and behaviour of *Caenorhabditis elegans* (Nematoda: Rhabditidae). Environ Microbiol. 2007;9(1):12–25.
26. Glater EE, Rockman MV, Bargmann CI. Multigenic natural variation underlies *Caenorhabditis elegans* olfactory preference for the bacterial pathogen *Serratia marcescens*. G3. 2014;4(2):265–76.
27. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Felix M-A, Kruglyak L. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat Genet. 2012;44(3):285–90.
28. de Bono M, Bargmann CI. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. Cell. 1998;94(5):679–89.
29. Rogers C, Persson A, Cheung B, de Bono M. Behavioral motifs and neural pathways coordinating O2 responses and aggregation in *C. elegans*. Curr Biol. 2006;16(7):649–59.
30. Weber KP, De S, Kozarewa I, Turner DJ, Babu MM, de Bono M. Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans. PLoS One. 2010;5(11):e13922.
31. Bendesky A, Tsunozaki M, Rockman MV, Kruglyak L, Bargmann CI. Catecholamine receptor polymorphisms affect decision-making in *C. elegans*. Nature. 2012;472:313–318.
32. Borgonie G, Van Driessche R, Leyns F, Arnaut G, De Waele D, Coomans A. Germination of *Bacillus thuringiensis* spores in bacteriophagous nematodes (Nematoda: Rhabditida). J Invertebr Pathol. 1995;65(1):61–7.
33. Borgonie G, Claeys M, Leyns F, Arnaut G, De Waele D, Coomans AV. Effect of a nematicidal *Bacillus thuringiensis* strain on free-living nematodes. 3. Characterization of the intoxication process. Fundam Appl Nematol. 1996; 19:523–8.
34. Wei JZ, Hale K, Carta L, Platzer E, Wong C, Fang SC, Aroian RV. *Bacillus thuringiensis* crystal proteins that target nematodes. Proc Natl Acad Sci U S A. 2003;100(5):2760–5.
35. Griffitts JS, Aroian RV. Many roads to resistance: how invertebrates adapt to Bt toxins. Bioessays. 2005;27(6):614–24.
36. Nielsen-LeRoux C, Gaudriault S, Ramarao N, Lereclus D, Givaudan A. How the insect pathogen bacteria *Bacillus thuringiensis* and Xenorhabdus/ Photorhabdus occupy their hosts. Curr Opin Microbiol. 2012;15(3):220–31.
37. Wang J, Nakad R, Schulenburg H. Activation of the *Caenorhabditis elegans* FOXO family transcription factor DAF-16 by pathogenic *Bacillus thuringiensis*. Dev Comp Immunol. 2012;37(1):193–201.
38. Schulte RD, Makus C, Hasert B, Michiels NK, Schulenburg H. Host–parasite local adaptation after experimental coevolution of *Caenorhabditis elegans* and its microparasite *Bacillus thuringiensis*. Proc Royal Soc B. 2011;278(1719): 2832–9.
39. Schulte RD, Makus C, Hasert B, Michiels NK, Schulenburg H. Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. Proc Natl Acad Sci. 2010;107(16):7359–64.
40. Masri L, Branca A, Sheppard AE, Papkou A, Laehnemann D, Guenther PS, Prahl S, Saebelfeld M, Hollensteiner J, Liesegang H. Host-pathogen coevolution: the selective advantage of *Bacillus thuringiensis* virulence and its cry toxin genes. PLoS Biol. 2015;13(6):e1002169.
41. Li Y, Álvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, Hazendonk E, Prins P, Plasterk RHA, Jansen RC, et al. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genet. 2006;2(12):e222.
42. Doroszuk A, Snoek LB, Fradin E, Riksen J, Kammenga J. A genome-wide library of CB4856/N2 introgression lines of *Caenorhabditis elegans*. Nucleic Acids Res. 2009;37(16):e110.
43. Balla KM, Troemel ER. *Caenorhabditis elegans* as a model for intracellular pathogen infection. Cell Microbiol. 2015;15(8):1313–22.
44. Wilfert L, Gadau J, Schmid-Hempel P. The genetic architecture of immune defense and reproduction in male *Bombus terrestris* bumblebees. Evolution. 2007;61(4):804–15.
45. Luijckx P, Fienberg H, Duneau D, Ebert D. A matching-allele model explains host resistance to parasites. Curr Biol. 2013;23(12):1085–8.
46. Schulenburg H, Kurtz J, Moret Y, Siva-Jothy MT. Introduction. Ecological immunology. Philo Transac Royal Soc B. 2009;364(1513):3–14.
47. Ferrandon D, Imler J-L, Hetru C, Hoffmann JA. The Drosophila systemic immune response: sensing and signalling during bacterial and fungal infections. Nat Rev Immunol. 2007;7(11):862–74.
48. Cheung BHH, Cohen M, Rogers C, Albayram O, de Bono M. Experience-dependent modulation of *C. elegans* behavior by ambient oxygen. Curr Biol. 2005;15(10):905–17.
49. Chang AJ, Chronis N, Karow DS, Marletta MA, Bargmann CI. A distributed chemosensory circuit for oxygen preference in *C. elegans*. PLoS Biol. 2006; 4(9):e274.
50. 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.
51. Yang W, Dierking K, Schulenburg H. WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. Bioinformatics. 2015; [Epub ahead of print].

52. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008; 4(1):44–57.
53. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
54. Shi J, Yang W, Chen M, Du Y, Zhang J, Wang K. AMD, an automated motif discovery tool using stepwise refinement of gapped consensuses. PLoS One. 2011;6(9):e24576.
55. Snoek LB, Van der Velde KJ, Arends D, Li Y, Beyer A, Elvin M, Fisher J, Hajnal A, Hengartner MO, Poulin GB, et al. Worm QTL—public archive and analysis web portal for natural variation data in Caenorhabditis spp. Nucleic Acids Res. 2013;41(D1):D738–43.
56. Snoek LB, Orbidans HE, Stastna JJ, Aartse A, Rodriguez M, Riksen JAG, Kammenga JE, Harvey SC. Widespread genomic incompatibilities in *Caenorhabditis elegans*. G3. 2014;4(10):1813–23.
57. Li Y, Breitling R, Snoek LB, Velde KJ, Swertz MA, Riksen JAG, Jansen RC, Kammenga JE. Global genetic robustness of the alternative splicing machinery in *Caenorhabditis elegans*. Genetics. 2010;186:405–10.
58. Viñuela A, Snoek LB, Riksen JAG, Kammenga JE. Genome-wide gene expression regulation as a function of genotype and age in *C. elegans*. Genome Res. 2010;20(7):929–37.
59. Rockman MV, Skrovanek SS, Kruglyak L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. Science. 2010; 330(6002):372–6.
60. Snoek LB, Joeri van der Velde K, Li Y, Jansen RC, Swertz MA, Kammenga JE. Worm variation made accessible: Take your shopping cart to store, link, and investigate! In: Worm. Abingdon, UK: Taylor & Francis; 2014: e28357.
61. Miltsch SM, Seeberger PH, Lepenies B. The C-type lectin-like domain containing proteins CLEC-39 and CLEC-49 are crucial for *Caenorhabditis elegans* immunity against *Serratia marcescens* infection. Dev Compar Immunol. 2014;45(1):67–73.
62. Irazoqui JE, Troemel ER, Feinbaum RL, Luhachack LG, Cezairliyan BO, Ausubel FM. Distinct pathogenesis and host responses during infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. PLoS Pathog. 2010;6(7):e1000982.
63. O'Rourke D, Baban D, Demidova M, Mott R, Hodgkin J. Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res. 2006;16(8):1005–16.
64. Yang W, Dierking K, Esser D, Tholey A, Leippe M, Rosenstiel P, Schulenburg H. Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* toward pathogenic *Bacillus thuringiensis*. Dev Compar Immunol. 2015;51(1):1–9.
65. Kim DH, Feinbaum R, Alloing G, Emerson FE, Garsin DA, Inoue H, Tanaka-Hino M, Hisamoto N, Matsumoto K, Tan M-W, et al. A conserved p38 MAP kinase pathway in *Caenorhabditis elegans* innate immunity. Science. 2002; 297(5581):623–6.
66. Shapira M, Hamlin BJ, Rong J, Chen K, Ronen M, Tan M-W. A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. Proc Natl Acad Sci. 2006;103(38):14086–91.
67. Kao C-Y, Los FCO, Huffman DL, Wachi S, Kloft N, Husmann M, Karabrahimi V, Schwartz J-L, Bellier A, Ha C, et al. Global functional analyses of cellular responses to pore-forming toxins. PLoS Pathog. 2011;7(3):e1001314.
68. Feng N, Zhi D, Zhang L, Tian J, Ren H, Li C, Zhu H, Li H. Molecular mechanisms of resistance to human pathogenic bacteria in *Caenorhabditis elegans* by MEV-1 mediated oxidative stress. Biochem Biophys Res Commun. 2015;459(3):481–7.
69. Schmid-Hempel P. Parasite immune evasion: a momentous molecular war. Trends Ecol Evol. 2008;23(6):318–26.
70. Niu Q, Huang X, Zhang L, Xu J, Yang D, Wei K, Niu X, An Z, Bennett JW, Zou C. A Trojan horse mechanism of bacterial pathogenesis against nematodes. Proc Natl Acad Sci. 2010;107(38):16631–6.
71. Stiernagle T. Maintenance of *C. elegans*. WormBook. The *C. elegans* research community. In: WormBook. 2006.
72. Leyns F, Borgonie G, Arnaut G, De Waele D. Nematicidal activity of *Bacillus thuringiensis* isolates. Fundam Appl Nematol. 1995;18(3):211–8.
73. Gutteling EW, Doroszuk A, Riksen JAG, Prokop Z, Reszka J, Kammenga JE. Environmental influence on the genetic correlations between life-history traits in *Caenorhabditis elegans*. Heredity. 2007;98(4):206–13.
74. Viñuela A, Snoek LB, Riksen JAG, Kammenga JE. Aging uncouples heritability and expression-QTL in *Caenorhabditis elegans*. G3. 2012;2(5):597–605.
75. Elvin M, Snoek L, Frejno M, Klemstein U, Kammenga J, Poulin G. A fitness assay for comparing RNAi effects across multiple *C. elegans* genotypes. BMC Genomics. 2011;12(1):1–14.
76. Rodriguez M, Snoek LB, Riksen JAG, Bevers RP, Kammenga JE. Genetic variation for stress-response hormesis in *C. elegans* lifespan. Exp Gerontol. 2012;47(8):581–7.
77. R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2015, http://www.R-project.org/. 2015.
78. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
79. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
80. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53.
81. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
82. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10.
83. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M. NCBI GEO: archive for functional genomics data sets update. Nucleic Acids Res. 2013;41(D1):D991–5.
84. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004;20(9):1453–4.
85. Saldanha AJ. Java Treeview–extensible visualization of microarray data. Bioinformatics. 2004;20(17):3246–8.
86. Tipping ME, Bishop CM. Probabilistic principal component analysis. J Royal Stat Soc Ser B. 1999;61(3):611–22.
87. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods–a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007;23(9):1164–7.
88. 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.
89. Huffman DL, Abrami L, Sasik R, Corbeil J, van der Goot FG, Aroian RV. Mitogen-activated protein kinase pathways defend against bacterial pore-forming toxins. Proc Natl Acad Sci U S A. 2004;101(30):10995–1000.
90. Boehnisch C, Wong D, Habig M, Isermann K, Michiels NK, Roeder T, May RC, Schulenburg H. Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. PLoS One. 2011;6(9):e24619.
91. Troemel ER, Chu SW, Reinke V, Lee SS, Ausubel FM, Kim DH. p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. PLoS Genet. 2006;2(11):e183.
92. Park SK, Tedesco PM, Johnson TE. Oxidative stress and longevity in *Caenorhabditis elegans* as mediated by SKN-1. Aging Cell. 2009;8(3):258–69.
93. Rohlfing A-K, 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.
94. Cui Y, McBride SJ, Boyd WA, Alper S, Freedman JH. Toxicogenomic analysis of *Caenorhabditis elegans* reveals novel genes and pathways involved in the resistance to cadmium toxicity. Genome Biol. 2007;8(6):R122.
95. Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H, Kenyon C. Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. Nature. 2003;424(6946):277–83.
96. 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.
97. Halaschek-Wiener J, Khattra JS, McKay S, Pouzyrev A, Stott JM, Yang GS, Holt RA, Jones SJM, Marra MA, Brooks-Wilson AR. Analysis of long-lived *C. elegans daf-2* mutants using serial analysis of gene expression. Genome Res. 2005;15(5):603–15.
98. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for *Caenorhabditis elegans*. Science. 2001;293(5537):2087–92.

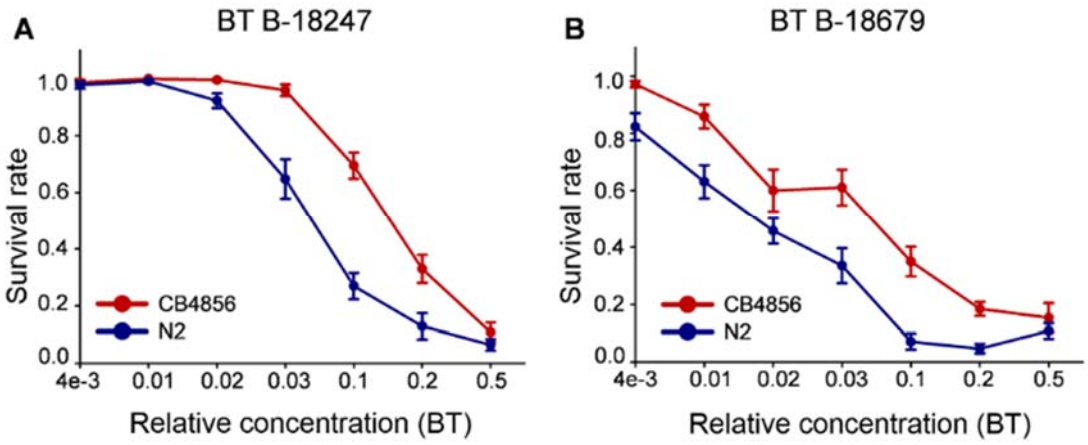
99. Menzel R, Bogaert T, Achazi R. A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. Arch Biochem Biophys. 2001;395(2):158–68.
100. Reichert K, Menzel R. Expression profiling of five different xenobiotics using a *Caenorhabditis elegans* whole genome microarray. Chemosphere. 2005; 61(2):229–37.
101. Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJM. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. Cell. 2009;138(2):314–27.
102. McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. Dev Biol. 2007;302(2):627–45.
103. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, Janette J, Chen C, Alves P, Preston E. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. Genome Res. 2011;21(2):245–54.



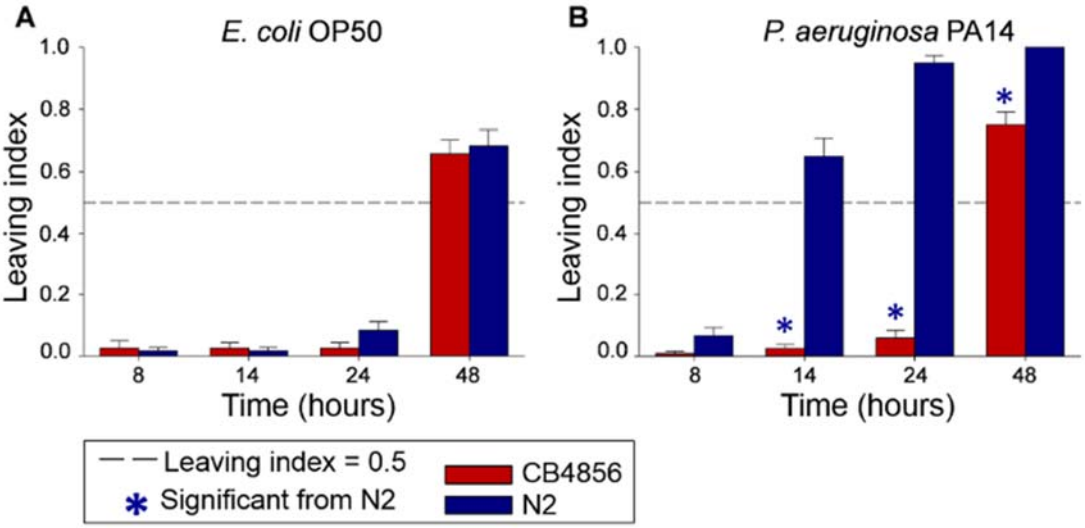
Additional file 1: Illustration of the lawn leaving assay. 10 hermaphrodites at the L4 stage were transferred by picking onto 9 cm peptone free NGM plates containing a lawn of the tested bacteria, each mixed with E. coli OP50 and surrounded by a ring of 80 µl of OP50.



Additional file 5: Figure on the leaving phenotypes of the introgression lines (ILs) plotted against the introgression position along the chromosomes. (A) Results for E. coli strain OP50; (B) non-nematocidal *B. thuringiensis* strain DSM350; (C) nematocidal *B. thuringiensis* B-18247; and (D) highly nematocidal *B. thuringiensis* B-18679. Green and red lines show the results after either 14 h or 24 h exposure, respectively. Position of markers is given along the X axis. Light gray vertical lines indicate boundaries of the chromosomes.



Additional file 9: Figure on the separate analysis of N2 and CB4856 survival in the presence of nematocidal *B. thuringiensis*. (A) Survival on *B. thuringiensis* strain B-18247; and (B) B-18679. Survival on the Y axis is plotted against BT concentration on the X axis. Error bars represent standard error of the means.



Additional file 10: Figure on the separate analysis of lawn leaving behavior of N2 and CB4856 towards E. coli and *P. aeruginosa*. (A) Results for avoidance of E. coli strain OP50; and (B) *P. aeruginosa* strain PA14. The asterisk (*) points to a significant difference to N2. The dotted reference line indicates the 0.5 avoidance response.

Note: the additional tables are too large to be attached here and they are available at https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2603-8

Chapter VI

GATA transcription factor as a likely key regulator of the *Caenorhabditis elegans* innate immune response against gut pathogens

Wentao Yang[1], Katja Dierking[1], Philip Rosenstiel[2] and Hinrich Schulenburg[1]*

[1] Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, 24098 Kiel, Germany

[2] Institute for Clinical Molecular Biology, University of Kiel, 24098 Kiel, Germany

* Correspondence: hschulenburg@zoologie.uni-kiel.de

Zoology 119 (2016) 244–253

Contents lists available at ScienceDirect

Zoology

journal homepage: www.elsevier.com/locate/zool

GATA transcription factor as a likely key regulator of the *Caenorhabditis elegans* innate immune response against gut pathogens☆

Wentao Yang[a], Katja Dierking[a], Philip C. Rosenstiel[b], Hinrich Schulenburg[a,*]

[a] Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, D-24098 Kiel, Germany
[b] Institute for Clinical Molecular Biology, University of Kiel, D-24098 Kiel, Germany

ARTICLE INFO

Article history:
Received 11 November 2015
Received in revised form 19 April 2016
Accepted 27 May 2016
Available online 27 May 2016

Keywords:
Caenorhabditis elegans
Innate immunity
GATA transcription factor
RNA-Seq
WormExp

ABSTRACT

Invertebrate defence against pathogens exclusively relies on components of the innate immune system. Comprehensive information has been collected over the last years on the molecular components of invertebrate immunity and the involved signalling processes, especially for the main invertebrate model species, the fruitfly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*. Yet, the exact regulation of general and specific defences is still not well understood. In the current study, we take advantage of a recently established database, WormExp, which combines all available gene expression studies for *C. elegans*, in order to explore commonalities and differences in the regulation of nematode immune defence against a large variety of pathogens versus food microbes. We identified significant overlaps in the transcriptional response towards microbes, especially pathogenic bacteria. We also found that the GATA motif is overrepresented in many microbe-induced gene sets and in targets of other previously identified regulators of worm immunity. Moreover, the activated targets of one of the known *C. elegans* GATA transcription factors, ELT-2, are significantly enriched in the gene sets, which are differentially regulated by gut-infecting pathogens. These findings strongly suggest that GATA transcription factors and particularly ELT-2 play a central role in regulating the *C. elegans* immune response against gut pathogens. More specific responses to distinct pathogens may be mediated by additional transcription factors, either acting alone or jointly with GATA transcription factors. Taken together, our analysis of the worm's transcriptional response to microbes provides a new perspective on the *C. elegans* immune system, which we propose to be coordinated by GATA transcription factor ELT-2 in the gut.

1. Introduction

Pathogens are ubiquitous. They comprise different species and genotypes that vary among locations and/or across time. They may even specifically adapt to host defences. In turn, host organisms are expected to adapt to these pathogen challenges, for example through the evolution of counter-adaptations to coevolving pathogen varieties and/or by tight regulation of immune responses, which allows individual hosts to react flexibly and quickly to an unpredictable pathogen threat. As a consequence, the immune system is expected to have a complex architecture that is fine-tuned to perceive a variety of pathogen-related signals in order to mount an appropriate and possibly highly specific defence response (Hughes and Nei, 1988; Sackton et al., 2007; Schulenburg et al., 2009). Comprehensive information is available on the complex organization of the immune system of higher vertebrates, which consists of innate and adaptive responses (Janeway et al., 2001). For invertebrates, the most detailed data sets are currently available for two model species, the fruitfly *Drosophila melanogaster* and the roundworm *Caenorhabditis elegans*, highlighting the involvement of several signalling cascades and various immune effectors (Buchon et al., 2014; Cohen and Troemel, 2015). Yet to date it is still unclear how exactly invertebrate immune responses are coordinated by general and possibly also by more specific regulators. To address this topic, we here focus on the nematode *C. elegans* and explore a recently established database, WormExp, which encompasses all gene expression studies for this organism (Yang et al., 2015b).

The nematode *C. elegans* has become a central model for dissecting the genetics of invertebrate immunity. It can be infected by various pathogens via several distinct infection routes (reviewed in Powell and Ausubel, 2008; Engelmann and Pujol, 2010; Marsh

☆ This article is part of a special issue entitled "Host-parasite coevolution - rapid reciprocal adaptation and its genetic basis".

* Corresponding author.
E-mail address: hschulenburg@zoologie.uni-kiel.de (H. Schulenburg).

http://dx.doi.org/10.1016/j.zool.2016.05.013

and May, 2012; Clark and Hodgkin, 2014), and a comprehensive toolbox is available for functional genetic analysis (Irazoqui et al., 2010b). The previous studies revealed that the nematode immune system is based on several signalling pathways conserved across invertebrates and vertebrates, including the p38 mitogen-activated protein kinase (MAPK), c-Jun N-terminal kinase (JNK) MAPK, extracellular signal-regulated kinase (ERK) MAPK, transforming growth factor-β (TGF-β), and also the insulin-like receptor (ILR) pathways (Engelmann and Pujol, 2010; Irazoqui et al., 2010a; Pukkila-Worley and Ausubel, 2012). Several transcription factors have been found to contribute to *C. elegans* immune defence, such as the GATA transcription factor ELT-2 (Shapira et al., 2006), the basic-region leucine zipper (bZIP) transcription factors ATF-7 (Shivers et al., 2010), ATFS-1 (Pellegrino et al., 2014), ZIP-2 (Estes et al., 2010), and SKN-1 (Papp et al., 2012), the basic helix-loop-helix (bHLH) transcription factor HLH-30 (Visvikis et al., 2014), the signal transducer and activator of transcription (STAT)-like transcription factor STA-2 (Dierking et al., 2011), and the activator protein 1 (AP-1) transcription factor dimer JUN-1/FOS-1 (Kao et al., 2011). Moreover, pathogen elimination involves certain antimicrobial peptides (reviewed in Dierking et al., 2016), including, for example, the caenacins and related peptides (Couillault et al., 2004), the caenopores (Mysliwy et al., 2010; Roeder et al., 2010), and additionally the generation of reactive oxygen species (ROS) (Chávez et al., 2009; Van Der Hoeven et al., 2011). While it remains unclear if and how pathogens are directly recognized by *C. elegans*, nematode defence can also be activated indirectly through a cellular surveillance system and/or damage signals, allowing the worms to respond to the cellular disturbance caused by an infection (Melo and Ruvkun, 2012; Zugasti et al., 2014; Ewbank and Pujol, 2016).

Although most studies on *C. elegans* immunity rely on functional genetic approaches, numerous transcriptomic and proteomic analyses were additionally performed to explore the set of genes which are activated or repressed upon pathogen exposure. These expression analyses considered a total of 21 pathogens and 6 non-pathogenic bacterial strains (see Section 3.1 and Table S1 in the supplementary online Appendix). Several of these studies involved more than one pathogen and showed an overlapping signature in the response to the various bacteria (Wong et al., 2007; Irazoqui et al., 2010a), indicating the presence of a common regulatory mechanism in the worm's immune system. Moreover, several expression analyses explored the downstream targets of immunity pathways, allowing us to assess the presence of shared or divergent signatures in the differentially expressed gene sets. Such expression analyses have been performed with mutants of the p38 MAPK, JNK MAPK, TGF-β and the ILR cascades (for example *pmk-1*(*km25*) for p38 MAPK (Bond et al., 2014), *jun-1*(*gk557*) for JNK (Uno et al., 2013), *dbl-1*(*ctls40*) overexpression for TGF-β (Roberts et al., 2010) and *daf-16*(*mu86*) for ILR (Murphy et al., 2003)). Similar expression analysis was performed for mutants of GATA transcription factors, including a mutant for the GATA transcription factor gene *elt-2*, which is known to contribute to immunity against some gut-infecting pathogens such as *Pseudomonas aeruginosa, Salmonella typhimurium*, *Enterococcus faecalis, Cryptococcus neoformans* or *Burkholderia pseudomallei* (Kerry et al., 2006; Shapira et al., 2006; Lee et al., 2013). ELT-2 is the major transcriptional regulator in the *C. elegans* intestine, controlling the constitutive expression of most of the genes necessary for maintenance of intestinal function (McGhee et al., 2009). It was recently suggested to act as a master regulator for additionally activated immune defence responses in the intestine (Block and Shapira, 2015). The availability of numerous immunity-related gene expression studies provides a unique opportunity to explore the regulation of common and specific responses to pathogens.

The aim of the current article is to use the available gene expression studies to identify common and specific regulators of the *C. elegans* immune response to pathogens. To date, such an integrative analysis, which considers all available microbe- and immunity-related transcriptome data sets, has not yet been performed. It allows us to assess the following questions: (i) How similar or how different are the inducible responses to various microbes? (ii) How similar or how different are the responses to the gene sets controlled by distinct immunity regulators and pathways? These assessments may yield novel insights into the regulation of the worm's response to pathogens and also to non-pathogenic microbes. We specifically consider transcription factors and characterize the presence of transcription factor binding motifs in the promoter regions of differentially expressed genes. We take advantage of a recently established database, WormExp, in which we collated all of the more than 1800 available *C. elegans* gene expression data sets (Yang et al., 2015b). We used this database to extract gene sets that are differentially regulated upon exposure to pathogens and non-pathogenic bacteria. We further assessed gene expression data sets that comprise the downstream targets of transcription factors or of known immunity pathway components. The presence of common or specific regulators is analysed with the help of different statistical methods, including gene set enrichment (Subramanian et al., 2005) and de novo transcription factor-binding motif analyses (Shi et al., 2011).

2. Materials and methods

2.1. Data sets

All transcriptome and proteome data sets are available from WormExp (http://wormexp.zoologie.uni-kiel.de/wormexp/) (Yang et al., 2015b). We specifically focused on 30 studies in which the gene expression response to microbes was analysed (see Table S1, links to references in column I). We additionally considered gene sets from gene expression analysis of mutants of known immunity regulators and immunity pathway components, including, for example, mutants for the GATA transcription factors *elt-1*, *elt-2*, and *elt-3* (all based on microarrays). Information on confirmed transcription factor binding sites, inferred through ChIP-Seq (Furey, 2012), were obtained from the modencode database (http://www.modencode.org/) (Gerstein et al., 2010).

2.2. Analysis

Gene set enrichment analyses were performed with the help of Fisher exact tests, as implemented in WormExp. To identify possible transcription factors that mediate differential expression, we searched for enriched motifs in the promoter regions of a particular gene set with the program AMD (Shi et al., 2011). We focused on the core promoter region, which is widely assumed to range from 600 bp (base pair) upstream up to 250 bp downstream of the transcription start site (TSS) (Michaloski et al., 2006; Michaloski et al., 2011; Tabach et al., 2007; Trinklein et al., 2003) and which is also often most conserved across genes (Cheung et al., 2007) and thus suitable for comparative sequence analysis. We did not consider a larger region because the core promoter region should be most relevant for general regulatory elements and also because extraction of larger intergenic regions is not always straightforward in *C. elegans* due to the comparatively high gene density in the nematode genome (Hillier et al., 2005). The core promoter region was thus assessed for all of the genes per gene set, using *C. elegans* genome version WS235 from wormbase (http://www.wormbase.org) (Stein et al., 2001). ChIP-Seq data of transcription factor binding sites were directly used in motif analysis, based on genome version WS220. Motif logos are produced by Weblogo 3.0 (Crooks et al., 2004).

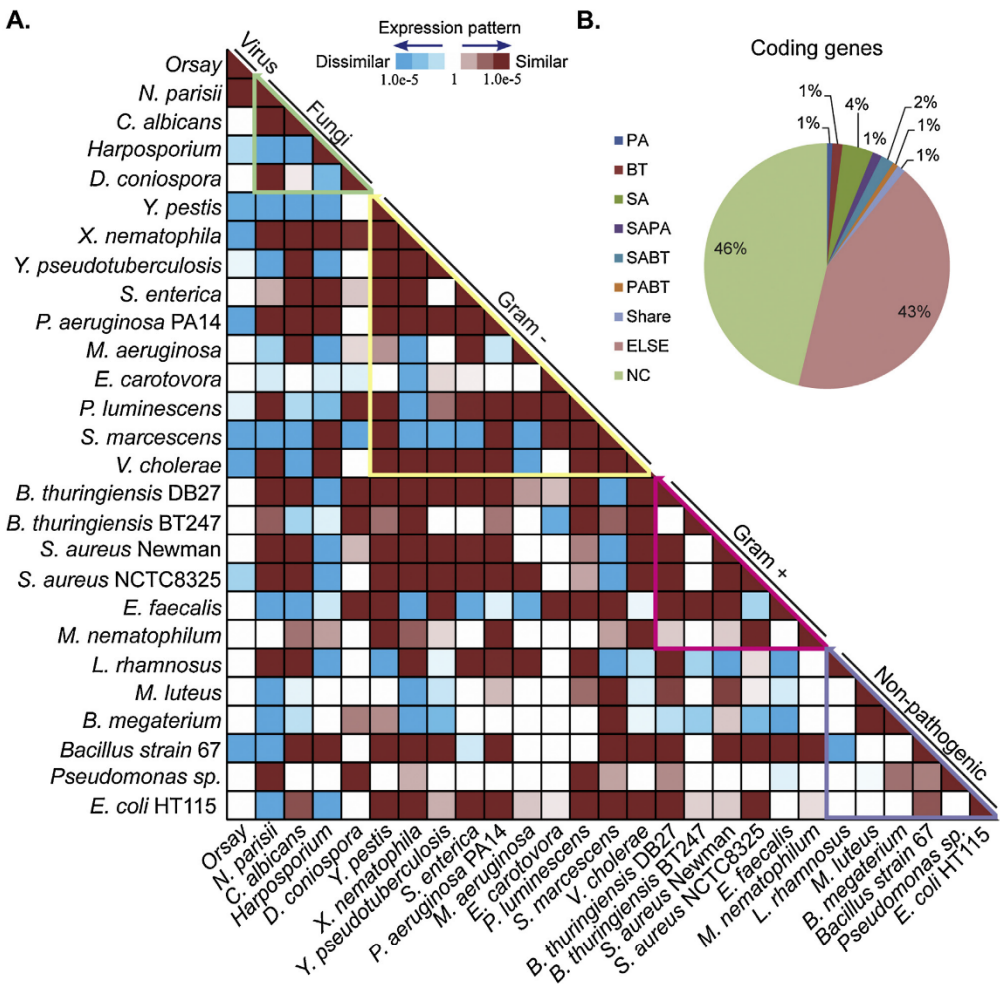


Fig. 1. Overlap among differentially expressed genes in response to pathogens and food bacteria. (A) Heat map for the significance of overlapping gene sets responding to virus, fungi, Gram-negative and Gram-positive bacteria, and also to non-pathogenic bacteria. The significance of overlapping gene sets is measured by a Fisher exact test with Bonferroni correction and indicated by colour intensity (see legend on top). Red indicates all overlaps for which genes show identical up- or downregulation, whereas blue represents all overlaps with opposite expression patterns (up in one case, down in the other). This illustration only shows the results for one data set per microbe strain in order to keep the overview manageable. For each pathogen, that data set has been included which produced the largest overlap with other data sets, in order to emphasise the presence of similar transcriptomic responses. Full results are shown in Fig. S1. (B) Percentage of all *C. elegans* coding genes influenced by pathogens. Number of coding genes is based on Wormbase version WS235. The colours highlight different pathogens or different combinations of pathogens, as indicated by the abbreviations in the legend on the left. For example, the orange colour is labelled PABT and indicates the overlapping gene set differentially regulated by both *Pseudomonas aeruginosa* and *Bacillus thuringiensis*. Abbreviations: NC, not changed by pathogens; PA, *P. aeruginosa*; BT, *B. thuringiensis*; SA, *Staphylococcus aureus*; Share, overlap among SA, PA and BT.

3. Results

3.1. *Overview of data sets used for analysis of the* C. elegans *response towards microbes*

Our analysis focused on 30 previously published gene expression studies (Table S1). Of these, 23 considered the response to pathogens (virus, bacteria, and fungi) and 7 to non-pathogenic bacteria. From these studies, we extracted a total of 127 individual transcriptome data sets (in several cases, more than one set from the same study), for which the transcriptional response was examined at different time points after initial exposure (Table S1). These studies compared the response towards the test microbe with that towards a control bacterium, usually the standard laboratory food strain *Escherichia coli* OP50, or, in a few cases, a non-pathogenic strain of the pathogen species under study. Most of these expression data sets were directly taken from the original studies, while in a few cases the significantly differentially expressed gene sets were inferred by us from re-analysis of the raw data (Table S1). The number of significantly differentially expressed genes varied among studies from 12 to 4051 (Table S1). Several independent studies used the same pathogen strain (e.g., *Pseudomonas aeruginosa* PA14,

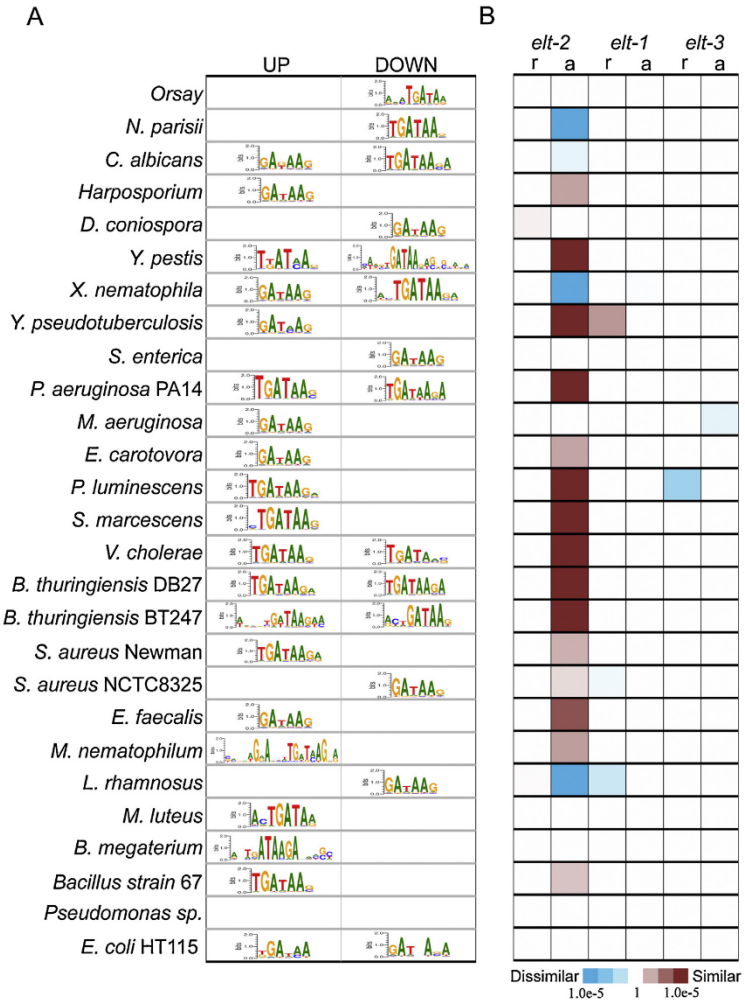


Fig. 2. Enriched GATA motifs and GATA transcription factor target genes for microbe-responsive gene sets. (A) The GATA motif is enriched for most microbe-related gene sets (either upregulated or downregulated). Microbes are presented in the same order as in Fig. 1. Further results are given in Table S1. (B) Only the gene set activated by the GATA transcription factor gene *elt-2* (but not *elt-1* or *elt-3*) produces a significant overlap with the microbe-responsive genes. Colour scale shows the significance of Bonferroni-corrected Fisher exact probabilities and the direction of enrichment, as in Fig. 1. Abbreviations: a, activated genes; r, repressed genes.

Bacillus thuringiensis BT247, and *Staphylococcus aureus* NCTC8325; Table S1), thus allowing us to evaluate consistency of inducible gene expression against specific bacteria. For our analysis, we excluded all data sets based on proteomic approaches, because these are not really comparable to transcriptomic data due to methodological biases and usually much lower coverage and thus a much smaller number of significantly differentially expressed genes (for example Yang et al., 2015a). We also excluded one transcriptomic study which assessed the response to *Bacillus subtilis* and which did not yield a single significantly differentially expressed gene under the here considered adjusted *p*-value (adjusted to take account of multiple testing using the false discovery rate, fdr). Moreover, for the analysis of transcriptome overlaps and motif inference we only show the results for one data set per microbe strain in order to keep figures as simple and thus as accessible as possible to the reader. For these cases, we chose the data set which produced the largest number of overlaps with other data sets in order to emphasize the presence of similar transcriptomic responses to microbes. Nevertheless, we provide the full results, based on all of the data sets, in Fig. S1 in the supplementary online Appendix.

3.2. *Pathogen exposure produces overlapping gene expression signatures in* C. elegans

We first assessed the percentage of coding genes in the *C. elegans* genome that are influenced by pathogen exposure (i.e., changed expression in response to at least one pathogen). Based on all data sets available, we found that 54% of all *C. elegans* coding genes are differentially expressed upon exposure to one of the considered pathogens (Fig. 1B). 11% are still differentially expressed in at least

one of the data sets if we only consider the three bacterial pathogens *P. aeruginosa* (indicated by PA), *B. thuringiensis* (BT), and *S. aureus* (SA), for which several independent data sets are available. Overall, a surprisingly large number of *C. elegans* genes are responsive to pathogenic microbes.

We next asked to what extent exposure to microbes causes similar or divergent expression responses in *C. elegans*. We found numerous significant overlaps among the considered pathogens and microbes (Fig. 1A; see full results in Fig. S1). In many comparisons, the overlap in gene expression goes in the same direction (i.e., either upregulated or downregulated by both of the two compared pathogens, as indicated by red colour in Fig. 1A). For example, the only two intracellular pathogens, the Orsay virus and the microsporidian *Nematocida parisii*, both seem to overlap in the up- and downregulated gene sets, confirming previous observations (Bakowski et al., 2014). A similar same-direction overlap is also found for those cases for which multiple independent data sets are available for the same pathogen species, which is the case for *B. thuringiensis*, *P. aeruginosa*, and also *S. aureus* (Fig. S2 in the supplementary online Appendix). These results strongly suggest that these pathogens produce a very robust transcriptome response that is independent of the platform and laboratory involved. In contrast, in other comparisons between different pathogens the changes in gene expression are in opposite directions (i.e., up in one pathogen but down in the other, indicated by blue colour in Fig. 1A). For example, the genes upregulated by the Orsay virus seem to be downregulated by *Yersinia pestis* and vice versa. Such opposite gene expression patterns may indicate specific responses to different pathogen types and/or infection routes.

In general, we found the largest number of same-direction overlaps among pathogenic bacteria (108 out of 210 comparisons among pathogens are significant at an adjusted p value < 0.05 and go into the same direction, Fig. 1A). The opposite pattern is particularly common for comparisons involving the Orsay virus (7 out of 20 comparisons between the virus and other pathogens). In fact, the virus only produces a same-direction overlap with the only other intracellular pathogen, the microsporidian *Nematocida parisii* (Fig. 1A). Other cases with a large number of opposite-direction overlaps include, for example, comparisons with the Gram-negative pathogen *Serratia marcescens* (11 out of 20 comparisons) and the Gram-positive pathogen *Enterococcus faecalis* (6 out of 20 comparisons).

Interestingly, certain non-pathogenic bacteria show significant same-direction overlaps with each other and also with pathogenic bacteria (for example, comparison between *E. coli* HT115 and *P. aeruginosa* PA14). As the non-pathogenic bacteria usually serve as food sources for *C. elegans*, this overlap may indicate that digestion and metabolism shape, at least partially, the response to pathogens. Alternatively, the food microbes which account for the observed overlap have at least some pathogenic effect, which was previously shown for *E. coli* OP50 (Gems and Riddle, 2000; Garigan et al., 2002).

3.3. *GATA motifs and GATA transcription factor targets are enriched in* C. elegans *pathogen response genes*

As an overlapping transcriptome signature may result from the action of the same regulator (i.e., a transcription factor), we next asked whether the core promoter regions of the different microbe-related gene sets are enriched for certain transcription factor binding motifs. Using a *de novo* motif analysis we demonstrate that the main motif enriched in the microbe-related gene sets is a GATA motif. Most pathogen-related gene sets are enriched for this motif, either in the up- or the downregulated gene sets and in several cases in both (Fig. 2A; full results are given in Table S1). GATA motifs are also enriched for non-pathogenic bacteria except for *Pseudomonas* sp. (Fig. 2A). Interestingly, the enriched

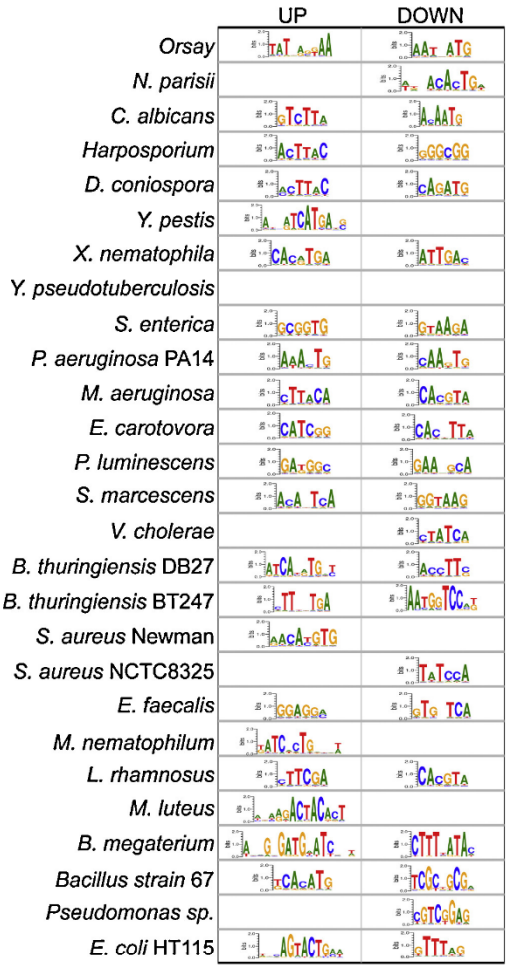


Fig. 3. Enriched non-GATA motifs for the microbe-responsive gene sets. One of the most enriched non-GATA motifs in the promoter regions of microbe-related gene sets is shown for the up- and downregulated genes. CACGTG is the Ebox motif enriched in the response to *X. nematophila*, *P. aeruginosa* PA14, *M. aeruginosa*, *S. aureus* Newman and *L. rhamnosus*. Full results are shown in Table S1.

GATA motifs can be divided into two distinct sub-families (Zhang et al., 2009): GATAAG and TGATAA. These two types are found for both pathogen- and non-pathogen-related gene sets and also for both up- and downregulated genes (Fig. 2A).

To further explore the particular importance of GATA transcription factor binding sites, we used the WormExp database to examine the overlap between the microbe-related gene sets and the sets of genes controlled in their expression by the three GATA transcription factor genes *elt-1*, *elt-2*, and *elt-3* (i.e., sets of differentially expressed genes in these mutants inferred from microarrays). While *elt-1* and *elt-3* regulate epidermal specification and differentiation, *elt-2* controls differentiation and function of the intestine (McGhee et al., 2009). An overrepresentation in the microbe-related data sets is mostly found for the *elt-2*-regulated genes and this enrichment is almost exclusively restricted to the

response to pathogenic taxa (Fig. 2B). Moreover, enrichment is also almost exclusively observed for the genes that are activated but not repressed by *elt-2* (Fig. 2B). The overlaps are either consistent (activated by *elt-2* and pathogens or vice versa) or opposite (activated by *elt-2* but repressed by pathogens or vice versa). Taken together, these results suggest that *elt-2* might serve as a central regulator in the *C. elegans* immune response to pathogens.

In addition to the GATA motif, we identified a few other motifs that are enriched in the microbe-related gene sets and possess palindromic or at least non-repetitive structures typical for transcription factor binding sites (Sorek et al., 2008) (Fig. 3, complete results in Table S1). One example is an Ebox motif (CACGTG) in the gene sets of *Xenorhabdus nematophila*, *P. aeruginosa* PA14, *Microcystis aeruginosa* and *Lactobacillus rhamnosus* (Fig. 3, Table S1). These additional motifs vary in the responses to different microbes, suggesting that they might mediate microbe-specific defences, possibly in interaction with a GATA transcription factor.

3.4. *GATA motifs are enriched in the targets of other immune-related transcription factors and signalling cascades*

We next hypothesised that the particular role of GATA-dependent transcription in *C. elegans* immunity is additionally reflected by an interaction between GATA transcription factors and other transcription factors previously implicated in the worm's immune system. Similarly, GATA transcription factors may also influence expression of the targets of the nematode's immunity-related signalling cascades.

To test the first hypothesis, we performed a *de novo* motif analysis on the transcription factor binding regions, which were previously identified through ChIP-Seq for the immune-related transcription factor genes *zip-2*, *skn-1*, *hlh-30*, *jun-1/fos-1*, and *daf-16*. Intriguingly, the GATA motif is enriched in the target binding regions of all of these transcription factors (Fig. 4A). In addition to the GATA motif, these binding regions are also enriched for other motifs, usually the motif predicted or validated for the respective transcription factor (Fig. 4A, right column). For instance, the *daf-16* ChIP-Seq data contains a significant signal for the consensus motif GTAAACA, which is indeed the confirmed binding site for the DAF-16 FOXO transcription factor (Tepper et al., 2013). Similarly, the *hlh-30* ChIP-Seq data is enriched for an Ebox motif, which also represents the previously reported binding sequence for this transcription factor (Visvikis et al., 2014). Such an Ebox motif is also overrepresented in the data for *zip-2*.

To further test the above hypotheses, we assessed the presence of the GATA motif in the promoter regions of the transcriptional targets of immunity regulators, including additional transcription factors such as *xbp-1* (Henis-Korenblit et al., 2010), *elt-2* (Block et al., 2015; Shapira et al., 2006), and *hif-1* (Bishop et al., 2004), or immune signalling pathway components such as *pmk-1* (Bond et al., 2014), *dbl-1* (Roberts et al., 2010), and *dkf-2* (Ren et al., 2009), or other types of regulators such as *npr-1* (Andersen et al., 2014). The targets were inferred from transcriptome data for these genes, as available from WormExp. A *de novo* motif analysis was then performed on the respective promoter regions of these targets. Promoter regions of *elt-2* activated genes showed enrichment for the GATA motif, confirming the accuracy of the motif analysis approach (Fig. 4B). The promoters of activated targets of the other considered genes (outside the circle) except *vhl-1* and *hif-1* (inside the circle) were similarly enriched for a GATA motif (Fig. 4B). Moreover, promoter regions of genes repressed by *npr-1* also contained the GATA motif, while we found the GATA enrichment for all the other considered genes always in the promoter region of activated targets.

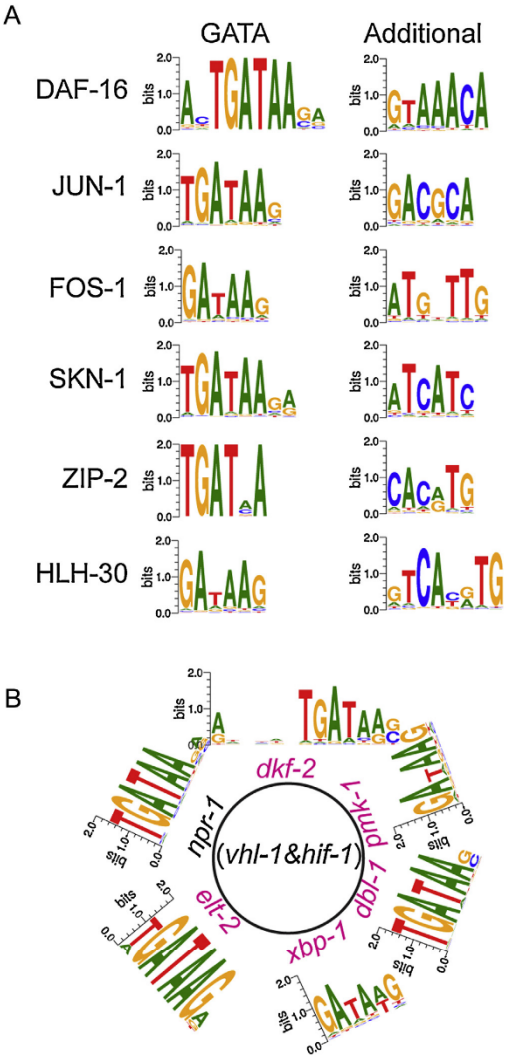


Fig. 4. GATA is enriched in binding sites in the targets of known regulators of the nematode immune response. (A) Enriched motifs in known targets of immunity-related transcription factors which were previously identified through ChIP-Seq. The enriched GATA motif is presented in the first column, while the enriched binding motifs for the considered transcription factors are shown in the second column. (B) Enriched motifs in the targets of various known immunity regulators, as inferred from previous transcriptomic studies with mutants or RNAi knockdown of the indicated genes. Pink regulator gene names (e.g., *elt-2*) indicate that only the activated target genes are enriched for the GATA motif. The *npr-1* regulator in black shows GATA enrichment in both activated and repressed genes. The targets of both *vhl-1* and *hif-1* (inside the circle) are not enriched for GATA.

3.5. *A set of 82 genes is regulated by both* elt-2 *and pathogens*

The above motif and gene set enrichment analyses identified a GATA transcription factor as the central regulator of the nematode's response to pathogens that likely accounts for the large overlap in the pathogen response genes. As *elt-2* is expressed exclu-

Table 1
Genes activated by *elt-2* and induced by at least five pathogens.*

Sequence name	Gene name	# Pathogens	Annotation
C50F7.5		12	No annotation
F55G11.5	*dod-22*	9	GO:0045087 innate immune response
Y46C8AL.2	*clec-174*	9	C type lectin domain- (CTLD-) containing proteins
F53A9.8		9	GO:0050830 response to Gram-positive bacterium
T28H10.3		9	Cysteine-type endopeptidase activity
ZC443.5	*ugt-18*	8	UDP glycosyltransferase
T01D3.6		7	GO:0005509 calcium ion binding
Y47H9C.1		7	No annotation
C09F12.1	*clc-1*	7	GO:0045087 innate immune response
Y39G10AR.6	*ugt-31*	6	UDP glycosyltransferase
ZK896.5		6	Epoxide hydrolase
C49C8.5		5	GO:0045087 innate immune response
F35C5.6	*clec-63*	5	C type lectin domain- (CTLD-) containing proteins
F35C5.9	*clec-66*	5	C type lectin domain- (CTLD-) containing proteins
F53B2.8		5	No annotation
F35E12.7	*dct-17*	5	GO:0045087 innate immune response
Y51A2D.13		5	Phospholipase
C32H11.12	*dod-24*	5	GO:0045087 innate immune response
C32H11.4		5	GO:0045087 innate immune response
H02F09.3		5	Serine/arginine repetitive matrix

* The full list of genes activated by *elt-2* and responsive to at least one pathogen is provided in Table S2 in the supplementary online Appendix.

sively in the intestine where it is the main regulator of constitutive transcription (McGhee et al., 2007), we finally asked what type of genes are both activated by *elt-2* and changed in inducible expression by at least one of the considered pathogens. The resulting list consists of 82 genes (Table S2 in the supplementary online Appendix), including 36 genes without any annotation and 15 putative immunity-related genes such as three lysozyme genes, three genes encoding caenopores/saposin domain-containing proteins, three galectin genes, five genes encoding C type lectin domain (CTLD)-containing proteins and one *nlp* gene (*nlp-40*). One of these 82 genes, namely *C50F7.5*, which currently has no functional annotation, is induced by 12 different pathogens (Table 1). 19 additional genes are induced by at least five different pathogens (Table 1), including several with a likely immune function such as the CTLD genes *clec-174*, *clec-63*, and *clec-66*.

4. Discussion

Microbes are central for our understanding of *C. elegans* biology, as the worm uses them as a food source and/or experiences them in its natural habitat as pathogens or commensals (Petersen et al., 2015). Therefore, we expect the response to microbes to be tightly regulated and this tight regulation should be manifested in the microbe-related gene expression data sets. Taking advantage of the recently established database WormExp with all available *C. elegans* expression data sets (Yang et al., 2015b), we demonstrate that the response to microbes and especially to gut-infecting pathogenic bacteria shows significant overlaps and is dominated by one of the *C. elegans* GATA transcription factors, ELT-2. We further demonstrate an enrichment of the GATA motif in the binding regions of previously identified immune-related transcription factors and also in the promoter regions of targets of other immune regulators and signalling cascades. ELT-2 is the central transcription factor controlling constitutive gene expression in the *C. elegans* intestine (McGhee et al., 2009), where infection with most bacterial pathogens and the corresponding immune responses are localised. Therefore, we propose that the ELT-2 GATA transcription factor is the primary regulator of the nematode's inducible intestinal immune defence, which is specifically activated in response to gut-infecting pathogens – in addition to the constitutively expressed genes involved in intestinal development and maintenance. The activated immune response is most likely regulated by ELT-2 in interaction with other transcription factors and/or other regulatory processes. In the following, we will discuss three main aspects of our results: (i) the similarity between the transcriptomic responses towards pathogens and food microbes; (ii) the central importance of the GATA transcription factor in regulating the response to gut pathogens; and (iii) possible mechanisms which could mediate a more specific response to different pathogens.

Although the strongest transcriptome overlap is found for the various responses against pathogens, these also show significant similarities to the expression changes induced by non-pathogenic bacteria and thus potential food organisms (Figs. 1 and S1). This may indicate that the worm's immune response incorporates components of the digestive machinery. Many of the studied pathogens cause an infection of the gut and thus the site of digesting food. Therefore, it may represent a highly economic strategy for the worm to use the same enzymes to process both types of microbes, including for example the members of the lysozyme family previously proposed to contribute to food digestion and immunity (Schulenburg and Boehnisch, 2008). This idea is supported by the observation that the transcriptional response to the non-pathogenic bacteria shows the smallest overlap and/or an opposite pattern to that found towards pathogens with a clearly distinct infection etiology, such as the intracellular Orsay virus, the intracellular microsporidian *N. parisii*, or the fungus *D. coniospora* which infects through the cuticle and not via the gut (Fig. 1) (Pujol et al., 2008; Marsh and May, 2012).

A non-exclusive alternative explanation may be that the available transcriptome studies with pathogens go beyond depicting only the nematode's immune response. In fact, most of these studies use the standard laboratory food *E. coli* OP50 as a control and often take any gene with expression differences to this bacterium to be part of the immune response. However, such expression differences may in these cases also be caused by exposing the worm to microbes with distinct characteristics (e.g., Gram-positive versus Gram-negative bacteria). Therefore, the observed expression differences could at least in part be determined by the worm's differential response to distinct microbial taxa, irrespective of their pathogenicity. A more precise characterisation of the immune system-specific expression response may in the future benefit from using several controls, including non-pathogenic varieties of the same pathogen taxon, as previously already used for *P. aeruginosa* (Troemel et al., 2006), *M. nematophilum* (O'Rourke et al., 2006), and *B. thuringiensis* (Boehnisch et al., 2011; Yang et al., 2015a).

The large overlaps in the microbe-induced expression responses were associated with the presence of the GATA motif in the promoter regions of differentially expressed genes. Intriguingly, only the gene set representing activated targets of one of the analysed *C. elegans* GATA transcription factors, ELT-2, was significantly enriched for genes responsive to gut-infecting pathogens, but not for genes differentially regulated by intracellular pathogens or those infecting via the cuticle (e.g., the Orsay virus or the fungus *D. coniospora*; Fig. 2B), and if so, only in the opposite direction (the microsporidian *N. parisii*; Fig. 2B). The same applies to the non-pathogenic microbes (Fig. 2B).

The ELT-2 GATA transcription factor is known to be primarily expressed in intestinal cells and to be the predominant transcription factor regulating differentiation and maintenance of the intestine (Fukushige et al., 1998; McGhee, 2013). ELT-2 is also known to regulate the inducible defence response to *P. aeruginosa* in the *C. elegans* intestinal epithelium by interacting with the two p38-activated transcription factors ATF-7 and SKN-1 (Shapira et al., 2006; Block et al., 2015). Moreover, worms in which *elt-2* is knocked down by RNAi are more susceptible to the bacterial pathogens *S. typhimurium* and *E. faecalis*, as well as the fungal pathogen *C. neoformans* (Kerry et al., 2006). In addition, ELT-2 was identified as a target for the manipulation of the host's immune defence by *Burkholderia pseudomallei* (Lee et al., 2013), emphasising its role in regulating pathogen-inducible responses. Furthermore, ELT-2 also controls non-infection stress responses in the intestine, such as the response to osmotic stress (Rohlfing et al., 2010), TOR-dependent hypoxia responses (Schieber and Chandel, 2014), and the response to high dietary zinc (Roh et al., 2015). ELT-2 also contributes to life span extension in calorically restricted *eat-2* mutants (Zhang et al., 2013) and survival of *rpn-10* mutants, which exhibit proteasome dysfunction (Keith et al., 2016). Based on these findings, Block and Shapira proposed a model in which ELT-2 functions as a master regulator of inducible responses in the intestine – in cooperation with different signal-activated transcription factors (Block and Shapira, 2015). Our data extends the previously available information on infection-induced responses by highlighting the involvement of ELT-2 in the activation of the response to a larger variety of pathogens, particularly those infecting the nematode gut (Fig. 3). As the transcriptional response to most pathogens is measured relative to control conditions with the standard food bacterium *E. coli* OP50, the role of ELT-2 in activating this defence response goes beyond its function in coordinating the constitutive expression of genes involved in development and maintenance of the intestine. Taken together, this strongly suggests that ELT-2 is the *C. elegans* GATA transcription factor that is centrally involved in the inducible response to gut-infecting pathogens. To date, we cannot yet exclude that another intestinal GATA transcription factor, such as ELT-7, for which transcriptional targets have not been characterised, additionally contributes to the pathogen response.

Such a central role of a GATA transcription factor (or several of these) is further supported by our finding of the GATA-binding motif in the targets of other regulators of the worm's immune response, for example various other transcription factors, such as ATF-7 (Shivers et al., 2010), ZIP-2 (Estes et al., 2010), SKN-1 (Papp et al., 2012), HLH-30 (Visvikis et al., 2014), JUN-1/FOS-1 (Kao et al., 2011), and DAF-16 (Garsin et al., 2003). The DAF-16 associated element (DAE) is a known GATA-like DNA motif found in promoters of DAF-16 target genes (Tepper et al., 2013). Zhang and colleagues showed that ELT-2 can bind to the DAE/GATA site and collaborates with DAF-16 to control tissue-dependent expression of multiple target genes, contributing to lifespan extension in the insulin-like receptor mutant *daf-2(e1370)* (Zhang et al., 2013). In addition, Block et al. highlighted that ELT-2 may cooperate with the SKN-1 and ATF-7 transcription factors to regulate p38 MAPK-dependent immune responses against the pathogen *P. aeruginosa* (Block et al., 2015).

The interaction of GATA transcription factors with other immune regulators may also contribute, at least in part, to more specific defence responses to different pathogens. Such more specific responses are indicated by those cases for which differential gene expression does not show any overlap or goes in the opposite direction (indicated by white and blue colours, respectively, in Figs. 1 and S1). Our finding of additional enriched binding motifs in the promoter regions of the pathogen-related gene sets (Fig. 3) suggests that the corresponding transcription factors could then mediate the more specific immune responses, either alone or in interaction with a GATA transcription factor. The latter is, for example, indicated in the cases with an enrichment of both the GATA motif and the Ebox motif (e.g., the genes activated by *X. nematophilum* or the genes repressed by *P. aeruginosa*; Figs. 2 A and Fig. 33), especially because the two transcription factors using the Ebox motif, ZIP-2 and HLH-30, both harbour the GATA motif in their validated binding regions (Fig. 4) (Estes et al., 2010; Visvikis et al., 2014).

5. Conclusion

Based on the analysis of transcriptomic responses to pathogenic and non-pathogenic microbes, we here propose that GATA transcription factors play a central role in regulating the *C. elegans* immune response. The ELT-2 GATA transcription factor seems to be particularly important in the defence against gut-infecting pathogens. GATA-dependent transcription may also mediate more specific immune responses against distinct pathogen taxa, possibly in interaction with other transcription factors. In future work it may be particularly interesting to assess how exactly ELT-2 interacts with other transcription factors and different immune signalling cascades to coordinate immune responses in the gut and to what extent other GATA transcription factors may additionally contribute to the response to pathogens and non-pathogenic microbes both inside and outside of the gut. Considering that pathogens may be detected through their effect on cellular homeostasis (Ewbank and Pujol, 2016) and considering that *elt-2* targets include genes important for the cytoprotective response in *C. elegans* (Shore et al., 2012), it may be particularly rewarding to find out how GATA-dependent transcription may relate to damage signals and the cellular surveillance system. Finally, our analysis provides a list of pathogen-responsive, ELT-2-regulated genes (Table 1) which could be important for mediating general and specific responses to pathogens. Their exact role, especially that of the gene *C50F7.5* without any current annotation, deserves particular attention using the available toolbox for functional genetic analysis in *C. elegans*.

Acknowledgements

We thank the members of the Schulenburg lab for advice and support. We are grateful for funding from the German Science Foundation within the Priority Program SPP 1399 on host–parasite coevolution (grants SCHU 1415/9 to H.S. and RO 2994/3 to P.C.R.). W.Y. was additionally supported by the International Max Planck Research School for Evolutionary Biology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.zool.2016.05.013.

References

Andersen, E.C., Bloom, J.S., Gerke, J.P., Kruglyak, L., 2014. A variant in the neuropeptide receptor npr-1 is a major determinant of *Caenorhabditis elegans* growth and physiology. PLoS Genet. 10, e1004156.

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

Bishop, T., Lau, K.W., Epstein, A.C., Kim, S.K., Jiang, M., O Rourke, D., Pugh, C.W., Gleadle, J.M., Taylor, M.S., Hodgkin, J., 2004. Genetic analysis of pathways regulated by the von Hippel–Lindau tumor suppressor in *Caenorhabditis elegans*. PLoS Biol. 2, 1549–1560.

Block, D.H., Shapira, M., 2015. GATA transcription factors as tissue-specific master regulators for induced responses. Worm 4, e1118607.

Block, D.H., Twumasi-Boateng, K., Kang, H.S., Carlisle, J.A., Hanganu, A., Lai, T.Y.-J., Shapira, M., 2015. The developmental intestinal regulator ELT-2 controls p38-dependent immune responses in adult *C. elegans*. PLoS Genet. 11, e1005265.

Boehnisch, C., Wong, D., Habig, M., Isermann, K., Michiels, N.K., Roeder, T., May, R.C., Schulenburg, H., 2011. Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. PLoS One 6, e24619.

Bond, M.R., Ghosh, S.K., Wang, P., Hanover, J.A., 2014. Conserved nutrient sensor O-GlcNAc transferase is integral to *C. elegans* pathogen-specific immunity. PLoS One 9, e113231.

Buchon, N., Silverman, N., Cherry, S., 2014. Immunity in *Drosophila melanogaster* – from microbial recognition to whole-organism physiology. Nature Rev. Immunol 14, 796–810.

Chávez, V., Mohri-Shiomi, A., Garsin, D.A., 2009. Ce-Duox1/BLI-3 generates reactive oxygen species as a protective innate immune mechanism in *Caenorhabditis elegans*. Infect. Immun 77, 4983–4989.

Cheung, L., Andersen, M., Gustavsson, C., Odeberg, J., Fernández-Pérez, L., Norstedt, G., Tollet-Egnell, P., 2007. Hormonal and nutritional regulation of alternative CD36 transcripts in rat liver – a role for growth hormone in alternative exon usage. BMC Mol. Biol. 8, 60.

Clark, L.C., Hodgkin, J., 2014. Commensals: probiotics and pathogens in the *Caenorhabditis elegans* model. Cell. Microbiol. 16, 27–38.

Cohen, L.B., Troemel, E.R., 2015. Microbial pathogenesis and host defense in the nematode *C. elegans*. Curr. Opin. Microbiol. 23, 94–101.

Couillault, C., Pujol, N., Reboul, J., Sabatier, L., Guichou, J.-F., Kohara, Y., 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. Nature Immunol. 5, 488–494.

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Dierking, K., Polanowska, J., Omi, S., Engelmann, I., Gut, M., Lembo, F., Ewbank, J.J., 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.

Dierking, K., Yang, W., Schulenburg, H., 2016. Antimicrobial effectors in the nematode *C. elegans* – an outgroup to the Arthropoda. Phil. Trans. R. Soc. B 371, 20150299.

Engelmann, I., Pujol, N., 2010. Innate immunity in *C. elegans*. In: Söderhäll, K. (Ed.), Invertebrate Immunity. Springer. New York. pp. 105–121.

Estes, K.A., Dunbar, T.L., Powell, J.R., Ausubel, F.M., Troemel, E.R., 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, 2153–2158.

Ewbank, J.J., Pujol, N., 2016. Local and long-range activation of innate immunity by infection and damage in *C. elegans*. Curr. Opin. Immunol. 38, 1–7.

Fukushige, T., Hawkins, M.G., McGhee, J.D., 1998. The GATA-factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. Dev. Biol. 198, 286–302.

Furey, T.S., 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Rev. Genet. 13, 840–852.

Garigan, D., Hsu, A.-L., Fraser, A.G., Kamath, R.S., Ahringer, J., Kenyon, C., 2002. Genetic analysis of tissue aging in *Caenorhabditis elegans*: a role for heat-shock factor and bacterial proliferation. Genetics 161, 1101–1112.

Garsin, D.A., Villanueva, J.M., Begun, J., Kim, D.H., Sifri, C.D., Calderwood, S.B., Ruvkun, G., Ausubel, F.M., 2003. Long-lived *C. elegans* daf-2 mutants are resistant to bacterial pathogens. Science 300, 1921–1921.

Gems, D., Riddle, D.L., 2000. Genetic, behavioral and environmental determinants of male longevity in *Caenorhabditis elegans*. Genetics 154, 1597–1610.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330, 1775–1787.

Henis-Korenblit, S., Zhang, P., Hansen, M., McCormick, M., Lee, S.-J., Cary, M., Kenyon, C., 2010. Insulin/IGF-1 signaling mutants reprogram ER stress response regulators to promote longevity. Proc. Natl. Acad. Sci. U. S. A. 107, 9730–9735.

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., Waterston, R.H., 2005. Genomics in *C. elegans*: so many genes, such a little worm. Genome Res. 15, 1651–1660.

Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170.

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

Irazoqui, J.E., Urbach, J.M., Ausubel, F.M., 2010b. Evolution of host innate defence: insights from *Caenorhabditis elegans* and primitive invertebrates. Nature Rev. Immunol. 10, 47–58.

Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.J., 2001. Immunobiology: The Immune System in Health and Disease. Churchill Livingstone, London.

Kao, C.-Y., Los, F., Huffman, D.L., Wachi, S., Kloft, N., Husmann, M., Karabrahimi, V., Schwartz, J.-L., Bellier, A., Ha, C., 2011. Global functional analyses of cellular responses to pore-forming toxins. PLoS Pathog. 7, e1001314.

Keith, S.A., Maddux, S.K., Zhong, Y., Chinchankar, M.N., Ferguson, A.A., Ghazi, A., Fisher, A.L., 2016. Graded proteasome dysfunction in *Caenorhabditis elegans* activates an adaptive response involving the conserved SKN-1 and ELT-2 transcription factors and the autophagy-lysosome pathway. PLoS Genet. 12, e1005823.

Kerry, S., TeKippe, M., Gaddis, N.C., Aballay, A., 2006. GATA transcription factor required for immunity to bacterial and fungal pathogens. PLoS One 1, e77.

Lee, S.-H., Wong, R.-R., Chin, C.-Y., Lim, T.-Y., Eng, S.-A., Kong, C., Ijap, N.A., Lau, M.-S., Lim, M.-P., Gan, Y.-H., 2013. *Burkholderia pseudomallei* suppresses *Caenorhabditis elegans* immunity by specific degradation of a GATA transcription factor. Proc. Natl. Acad. Sci. U. S. A. 110, 15067–15072.

Marsh, E.K., May, R.C., 2012. *Caenorhabditis elegans*: a model organism for investigating immunity. Appl. Environm. Microbiol. 78, 2075–2081.

McGhee, J.D., 2013. The *Caenorhabditis elegans* intestine. Wiley Interdiscip. Rev. Dev. Biol. 2, 347–367.

McGhee, J.D., Sleumer, M.C., Bilenky, M., Wong, K., McKay, S.J., Goszczynski, B., Tian, H., Krich, N.D., Khattra, J., Holt, R.A., 2007. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. Dev. Biol. 302, 627–645.

McGhee, J.D., Fukushige, T., Krause, M.W., Minnema, S.E., Goszczynski, B., Gaudet, J., Kohara, Y., Bossinger, O., Zhao, Y., Khattra, J., 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine from embryo to adult. Dev. Biol. 327, 551–565.

Melo, J.A., Ruvkun, G., 2012. Inactivation of conserved *C. elegans* genes engages pathogen- and xenobiotic-associated defenses. Cell 149, 452–466.

Michaloski, J.S., Galante, P.A., Malnic, B., 2006. Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. Genome Res. 16, 1091–1098.

Michaloski, J.S., Galante, P.A., Nagai, M.H., Armelin-Correa, L., Chien, M.-S., Matsunami, H., Malnic, B., 2011. Common promoter elements in odorant and vomeronasal receptor genes. PLoS One 6, e29065.

Murphy, C.T., McCarroll, S.A., Bargmann, C.I., Fraser, A., Kamath, R.S., Ahringer, J., Li, H., Kenyon, C., 2003. Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. Nature 424, 277–283.

Mysliwy, J., Dingley, A.J., Stanisak, M., Jung, S., Lorenzen, I., Roeder, T., Leippe, M., Grötzinger, J., 2010. Caenopore-5: the three-dimensional structure of an antimicrobial protein from *Caenorhabditis elegans*. Dev. Comp. Immunol. 34, 323–330.

O'Rourke, D., Baban, D., Demidova, M., Mott, R., Hodgkin, J., 2006. Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res. 16, 1005–1016.

Papp, D., Csermely, P., Soti, C., 2012. A role for SKN-1/Nrf in pathogen resistance and immunosenescence in *Caenorhabditis elegans*. PLoS Pathog. 8, e1002673.

Pellegrino, M.W., Nargund, A.M., Kirienko, N.V., Gillis, R., Fiorese, C.J., Haynes, C.M., 2014. Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection. Nature 516, 414–417.

Petersen, C., Dirksen, P., Schulenburg, H., 2015. Why we need more ecology for genetic models such as *C. elegans*. Trends Genet. 31, 120–127.

Powell, J.R., Ausubel, F.M., 2008. Models of *Caenorhabditis elegans* infection by bacterial and fungal pathogens. Methods Mol. Biol. 415, 403–427.

Pujol, N., Cypowyj, S., Ziegler, K., Millet, A., Astrain, A., Goncharov, A., Jin, Y., Chisholm, A.D., Ewbank, J.J., 2008. Distinct innate immune responses to infection and wounding in the *C. elegans* epidermis. Curr. Biol. 18, 481–489.

Pukkila-Worley, R., Ausubel, F.M., 2012. Immune defense mechanisms in the *Caenorhabditis elegans* intestinal epithelium. Curr. Opin. Immunol. 24, 3–9.

Ren, M., Feng, H., Fu, Y., Land, M., Rubin, C.S., 2009. Protein kinase D is an essential regulator of *C. elegans* innate immunity. Immunity 30, 521–532.

Roberts, A.F., Gumienny, T.L., Gleason, R.J., Wang, H., Padgett, R.W., 2010. Regulation of genes affecting body size and innate immunity by the DBL-1/BMP-like pathway in *Caenorhabditis elegans*. BMC Dev. Biol. 10, 61.

Roeder, T., Stanisak, M., Gelhaus, C., Bruchhaus, I., Grötzinger, J., Leippe, M., 2010. Caenopores are antimicrobial peptides in the nematode *Caenorhabditis elegans* instrumental in nutrition and immunity. Dev. Comp. Immunol. 34, 203–209.

Roh, H.C., Dimitrov, I., Deshmukh, K., Zhao, G., Warnhoff, K., Cabrera, D., Tsai, W., Kornfeld, K., 2015. A modular system of DNA enhancer elements mediates tissue-specific activation of transcription by high dietary zinc in *C. elegans*. Nucleic Acids Res. 43, 803–816.

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

Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., Clark, A.G., 2007. Dynamic evolution of the innate immune system in *Drosophila*. Nature Genet. 39, 1461–1468.

Schieber, M., Chandel, N.S., 2014. TOR signaling couples oxygen sensing to lifespan in *C. elegans*. Cell Rep. 9, 9–15.

Schulenburg, H., Boehnisch, C., 2008. Diversification and adaptive sequence evolution of *Caenorhabditis* lysozymes (Nematoda: Rhabditidae). BMC Evol. Biol. 8, 114.

Schulenburg, H., Kurtz, J., Moret, Y., Siva-Jothy, M.T., 2009. Introduction. Ecological immunology. Phil. Trans. R. Soc. B 364, 3–14.

Shapira, M., Hamlin, B.J., Rong, J., Chen, K., Ronen, M., Tan, M.-W., 2006. A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. Proc. Natl. Acad. Sci. U. S. A. 103, 14086–14091.

Shi, J., Yang, W., Chen, M., Du, Y., Zhang, J., Wang, K., 2011. AMD, an automated motif discovery tool using stepwise refinement of gapped consensuses. PLoS One 6, e24576.

Shivers, R.P., Pagano, D.J., Kooistra, T., Richardson, C.E., Reddy, K.C., Whitney, J.K., Kamanzi, O., Matsumoto, K., Hisamoto, N., Kim, D.H., 2010. Phosphorylation of the conserved transcription factor ATF-7 by PMK-1 p38 MAPK regulates innate immunity in *Caenorhabditis elegans*. PLoS Genet. 6, e1000892.

Shore, D.E., Carr, C.E., Ruvkun, G., 2012. Induction of cytoprotective pathways is central to the extension of lifespan conferred by multiple longevity pathways. PLoS Genet. 8, e1002792.

Sorek, R., Kunin, V., Hugenholtz, P., 2008. CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol. 6, 181–186.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., Spieth, J., 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. 29, 82–86.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Tabach, Y., Brosh, R., Buganim, Y., Reiner, A., Zuk, O., Yitzhaky, A., Koudritsky, M., Rotter, V., Domany, E., 2007. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. PLoS One 2, e807.

Tepper, R.G., Ashraf, J., Kaletsky, R., Kleemann, G., Murphy, C.T., Bussemaker, H.J., 2013. PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. Cell 154, 676–690.

Trinklein, N.D., Aldred, S.J.F., Saldanha, A.J., Myers, R.M., 2003. Identification and functional analysis of human transcriptional promoters. Genome Res. 13, 308–312.

Troemel, E.R., Chu, S.W., Reinke, V., Lee, S.S., Ausubel, F.M., Kim, D.H., 2006. p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. PLoS Genet. 2, e183.

Uno, M., Honjoh, S., Matsuda, M., Hoshikawa, H., Kishimoto, S., Yamamoto, T., Ebisuya, M., Yamamoto, T., Matsumoto, K., Nishida, E., 2013. A fasting-responsive signaling pathway that extends life span in *C. elegans*. Cell Rep. 3, 79–91.

Van Der Hoeven, R., McCallum, K.C., Cruz, M.R., Garsin, D.A., 2011. Ce-Duox1/BLI-3 generated reactive oxygen species trigger protective SKN-1 activity via p38 MAPK signaling during infection in *C. elegans*. PLoS Pathog. 7, e1002453.

Visvikis, O., Ihuegbu, N., Labed, S.A., Luhachack, L.G., Alves, A.-M.F., Wollenberg, A.C., Stuart, L.M., Stormo, G.D., Irazoqui, J.E., 2014. Innate host defense requires TFEB-mediated transcription of cytoprotective and antimicrobial genes. Immunity 40, 896–909.

Wong, D., Bazopoulou, D., Pujol, N., Tavernarakis, N., 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.

Yang, W., Dierking, K., Esser, D., Tholey, A., Leippe, M., Rosenstiel, P., Schulenburg, H., 2015a. Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* toward pathogenic *Bacillus thuringiensis*. Dev. Comp. Immunol. 51, 1–9.

Yang, W., Dierking, K., Schulenburg, H., 2015b. WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. Bioinformatics 15, 943–945.

Zhang, P., Judy, M., Lee, S.-J., Kenyon, C., 2013. Direct and indirect gene regulation by a life-extending FOXO protein in *C. elegans*: roles for GATA factors and lipid gene regulators. Cell Metab. 17, 85–100.

Zhang, Y., Wu, W., Cheng, Y., King, D.C., Harris, R.S., Taylor, J., Chiaromonte, F., Hardison, R.C., 2009. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. Nucleic Acids Res. 37, 7024–7038.

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

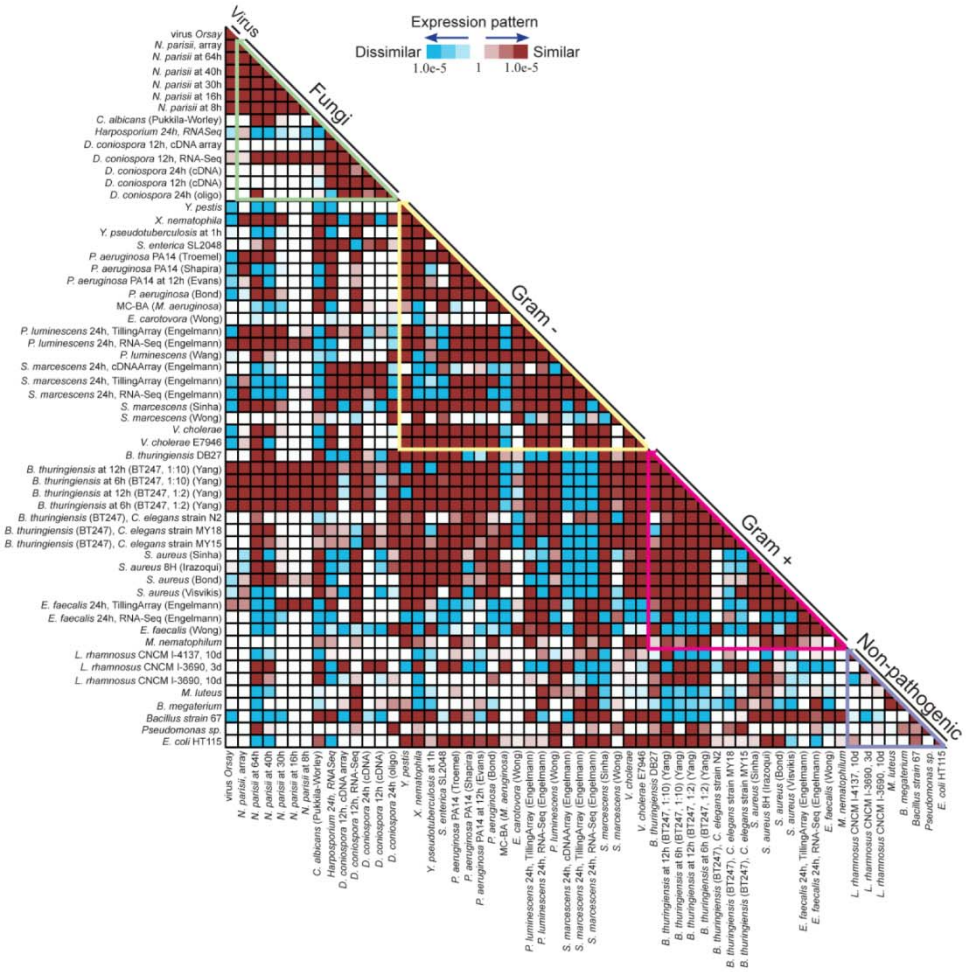


Fig. S1. Full results of overlap among the differentially expressed genes in response to pathogens and food bacteria. The significance of the overlap, inferred through a Fisher exact test with Bonferroni adjustment, is indicated by the heat map (see legend on top). Red indicates an overlap in the same direction (both up- and/or downregulated), while blue indicates gene expression in opposite directions. Table S1 provides information on the data sets considered.

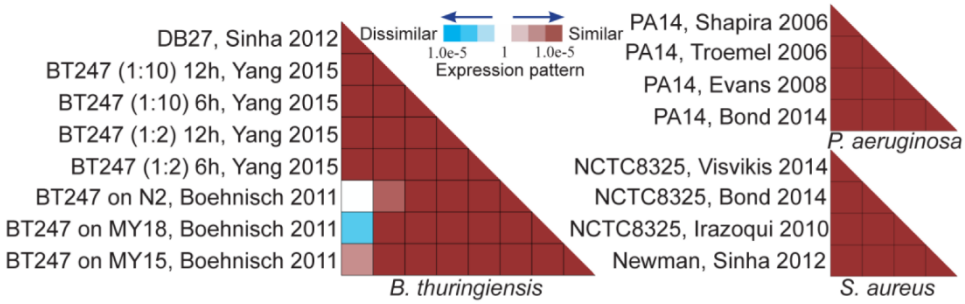


Fig. S2. Overlap among differentially expressed genes in response to the same pathogens. Significant congruence was found among independent gene expression studies with the same pathogen species and even pathogen strains, as shown here for *B. thuringiensis* (strains DB27 and BT247), *P. aeruginosa* (strains PA14), and also *S. aureus* (strains NCTC8325 and Newman). Significance level is indicated by colour intensity. Red indicates overlap with same-direction gene expression and blue with the opposite expression pattern.

References

Boehnisch, C., Wong, D., Habig, M., Isermann, K., Michiels, N.K., Roeder, T., May, R.C., Schulenburg, H., 2011. Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. PloS One 6, e24619.

Bond, M.R., Ghosh, S.K., Wang, P., Hanover, J.A., 2014. Conserved nutrient sensor O-GlcNAc transferase is integral to *C. elegans* pathogen-specific immunity. PloS One 9, e113231.

Evans, E.A., Kawli, T., Tan, M.W., 2008. *Pseudomonas aeruginosa* suppresses host immunity by activating the DAF-2 insulin-like signaling pathway in *Caenorhabditis elegans*. PLoS Pathog. 4, e1000175.

Irazoqui, J.E., Urbach, J.M., Ausubel, F.M., 2010. Evolution of host innate defence: insights from *Caenorhabditis elegans* and primitive invertebrates. Nat. Rev. Immunol. 10, 47–58.

Shapira, M., Hamlin, B.J., Rong, J., Chen, K., Ronen, M., Tan, M.-W., 2006. A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. Proc. Natl. Acad. Sci. U.S.A. 103, 14086–14091.

Sinha, A., Rae, R., Iatsenko, I., Sommer, R.J., 2012. System wide analysis of the evolution of innate immunity in the nematode model species *Caenorhabditis elegans* and *Pristionchus pacificus*. PloS One 7, e44255.

Troemel, E.R., Chu, S.W., Reinke, V., Lee, S.S., Ausubel, F.M., Kim, D.H., 2006. p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. PLoS Genet. 2, e183.

Visvikis, O., Ihuegbu, N., Labed, S.A., Luhachack, L.G., Alves, A.-M.F., Wollenberg, A.C., Stuart, L.M., Stormo, G.D., Irazoqui, J.E., 2014. Innate host defense requires TFEB-mediated transcription of cytoprotective and antimicrobial genes. Immunity 40, 896–909.

Yang, W., Dierking, K., Esser, D., Tholey, A., Leippe, M., Rosenstiel, P., Schulenburg, H., 2015. Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode *Caenorhabditis elegans* toward pathogenic *Bacillus thuringiensis*. Dev. Comp. Immunol. 51, 1–9.

Note: the supplementary tables are too large to be attached here and they are available at https://doi.org/10.1016/j.zool.2016.05.013

Chapter VII.

Antimicrobial effectors in the nematode *C. elegans* – an outgroup to the Arthropoda.

Katja Dierking [1], Wentao Yang[1], and Hinrich Schulenburg[1]*

[1] Department of Evolutionary Ecology and Genetics, Zoological Institute, University of Kiel, 24098 Kiel, Germany

* Correspondence: hschulenburg@zoologie.uni-kiel.de

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Dierking K, Yang W, Schulenburg H. 2016 Antimicrobial effectors in the nematode *Caenorhabditis elegans*: an outgroup to the Arthropoda. *Phil. Trans. R. Soc. B* **371**: 20150299.
http://dx.doi.org/10.1098/rstb.2015.0299

Accepted: 27 February 2016

One contribution of 13 to a theme issue 'Evolutionary ecology of arthropod antimicrobial peptides'.

Subject Areas:
immunology, evolution

Keywords:
Caenorhabditis elegans, antimicrobial peptides, lysozymes, caenopores, caenacins, reactive oxygen species

Author for correspondence:
Hinrich Schulenburg
e-mail: hschulenburg@zoologie.uni-kiel.de

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2015.0299 or via http://rstb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Antimicrobial effectors in the nematode *Caenorhabditis elegans*: an outgroup to the Arthropoda

Katja Dierking, Wentao Yang and Hinrich Schulenburg

Department of Evolutionary Ecology and Genetics, University of Kiel, Kiel 24098, Germany

Nematodes and arthropods likely form the taxon Ecdysozoa. Information on antimicrobial effectors from the model nematode *Caenorhabditis elegans* may thus shed light on the evolutionary origin of these defences in arthropods. This nematode species possesses an extensive armory of putative antimicrobial effector proteins, such as lysozymes, caenopores (or saposin-like proteins), defensin-like peptides, caenacins and neuropeptide-like proteins, in addition to the production of reactive oxygen species and autophagy. As *C. elegans* is a bacterivore that lives in microbe-rich environments, some of its effector peptides and proteins likely function in both digestion of bacterial food and pathogen elimination. In this review, we provide an overview of *C. elegans* immune effector proteins and mechanisms. We summarize the experimental evidence of their antimicrobial function and involvement in the response to pathogen infection. We further evaluate the microbe-induced expression of effector genes using WormExp, a recently established database for *C. elegans* gene expression analysis. We emphasize the need for further analysis at the protein level to demonstrate an antimicrobial activity of these molecules both *in vitro* and *in vivo*.

This article is part of the themed issue 'Evolutionary ecology of arthropod antimicrobial peptides'.

1. Brief overview of the *Caenorhabditis elegans* immune system

Arthropods and nematodes are likely members of related evolutionary lineages and form the clade of the moulting animals, the Ecdysozoa [1,2]. Owing to their common ancestry, comparison between these lineages may enhance our understanding of the evolutionary origin of specific traits in either of the taxa. Such comparisons have previously revealed important differences in the general organization of nematode and insect immune systems (see below). They complement broader evolutionary comparisons, for example between insects and more distantly related animal lineages such as molluscs or cnidarians [3]. In the current review, we focus on the model nematode *Caenorhabditis elegans*, which has become a central model organism for studying the genetics of invertebrate immunity. First work in this field was published in 1999 [4,5]. Since then, the nematode's response to a variety of pathogens has been assessed. These include Gram-negative bacteria (e.g. *Pseudomonas aeruginosa*, *Serratia marcescens*, *Salmonella enterica*), Gram-positive bacteria (e.g. *Microbacterium nematophilum*, *Leucobacter* sp., *Enterococcus faecalis*, *Staphylococcus aureus* and *Bacillus thuringiensis*), fungi (e.g. *Drechmeria coniospora*, *Nematocida parisii*, *Candida albicans*) and also a nodavirus (Orsay virus) [6]. Some of the used pathogens interact with *C. elegans* in nature, especially the microsporidian *N. parisii* [7], Orsay virus [8], *P. aeruginosa* [9], *Leucobacter* sp. [10], and possibly also *M. nematophilum* and *B. thuringiensis*. These are thus likely to elicit more specific defence responses. Most of the used pathogens infect the worm's intestine. Thus, the immune response mediated by intestinal epithelial cells has been examined in detail. Nevertheless, a few pathogens show different infection characteristics. The microsporidian *N. parisii* and Orsay virus are intracellular pathogens [7,8]. The fungus *D. coniospora* and the bacteria *Leucobacter* sp. and *M. nematophilum* infect via the

Table 1. Overview of putative antimicrobial effector gene families in the nematode *Caenorhabditis elegans*.[a]

gene family	abbr.	comment	no. of genes	antimicrobial examples[b]
caenacins/neuropeptide-like proteins	*cnc/nlp*	short peptides, rich in glycine and aromatic acid	12	NLP-31
caenopores	*spp*	SAPLIPs (with saposin domain)	23	SPP-1, SPP-3, SPP-5, SPP-12
lysozymes	*lys/ilys*	2 lysozyme types (entamoeba- and invertebrate-types)	16	none tested
defensin-like AMPs	*abf*	sequence homology to insect and mammalian defensins	6	ABF-2
C-type lectin domain-containing proteins	*clec*	diverse family with C-type lectin domain	283	none found
fungal-induced peptides and *fip*-related peptides	*fip/fipr*	short peptides induced upon fungal exposure	36	none tested
thaumatin-like proteins	*thn*	homologies to anti-fungal thaumatins from plants	8	none tested

[a]The overview only includes the gene families with members for which an antimicrobial function was demonstrated experimentally or which show homologies to known antimicrobial effectors from other taxa or for which an antimicrobial function was proposed. In addition to these gene families, antimicrobial effector mechanisms covered by this review also include the production of ROS and autophagy.
[b]Examples, for which an antimicrobial function was demonstrated *in vitro* at the protein/peptide level.

cuticle; the latter two via the anal region and tail, and the former mainly via the mouth, the vulva and the anus [10–12]. These pathogens therefore target distinct tissues to those infecting the gut.

Analysis of the response of *C. elegans* to these various pathogens revealed the presence of a complex interconnected immune system the structure of which has been described in several reviews (e.g. [13–17]). Thus, for the purpose of the current review, we only indicate some of the main elements. Most curiously, the nematode lacks specialized immune cells and also homologues of central components of insect immune systems such as the transcriptional regulator NFκB or genes belonging to the prophenoloxidase cascade (reviewed in [18]). *Caenorhabditis elegans* immune defence relies on behavioural and physiological responses. Of these, behavioural responses represent an important, fine-tuned first line defence against a variety of bacterial pathogens (reviewed in [15,19]). The physiological immune systems is based on several conserved core signalling cascades, including three mitogen-activated protein kinase (MAPK) pathways (i.e. the p38, the extracellular signal regulated kinase (ERK) and the c-Jun N-terminal kinase (JNK) MAPK pathways), the insulin-like receptor (ILR) pathway, and a transforming growth factor β (TGF-β) cascade. In addition, the RNA interference (RNAi) machinery mediates defence against Orsay virus, which seems to be recognized by the retinoic acid inducible gene I (RIG-I) helicase DHR-1 [20]. Recognition of bacterial and fungal pathogens is less clear. A G protein-coupled receptor was implicated in the indirect recognition of the fungal pathogen *D. coniospora* via the perception of a so-called damage-associated molecular pattern (DAMP) [21]. Indirect pathogen detection also seems to be achieved through a cellular surveillance system, which activates pathogen defence responses when central cellular processes are disrupted [17,22]. While the upstream activators of *C. elegans* immune signalling cascades are less well understood, several downstream transcription factors that regulate the activation of immune effector gene expression following infection with different pathogens have been identified. For example, the basic helix–loop–helix (bHLH) transcription factor HLH-30 (TFEB in mammals) was demonstrated to regulate expression of effector genes in response to *S. aureus* [23], the GATA transcription factor ELT-2 (homologous to human GATA-4, -5 and -6) in response to *P. aeruginosa* [24], and the signal transducer and activator of transcription (STAT)-like transcription factor STA-2 in response to *D. coniospora* [25]. *Caenorhabditis elegans* immune effector mechanisms, which function in pathogen elimination, have been characterized in several cases, as explained in more detail in §2.

The aim of the current review is to provide an overview of antimicrobial immune effectors and effector mechanisms in the model nematode *C. elegans*. We focus on mechanisms, genes and gene families, for which there is evidence of an antimicrobial function in the worm, or which are at least implicated to be part of nematode immunity because of homology with characterized antimicrobial effectors from other organisms. We explore the functional diversity between and within effector types. We summarize the available evidence for their role as effectors, including the demonstration of an antimicrobial effect at the protein or peptide level, an immune phenotype upon experimental manipulation of the gene (knock-out (KO), RNAi knock-down, or gene overexpression), and also induced gene expression after pathogen exposure. For the latter, we take advantage of a recently established database, WormExp, which combines all available gene expression studies for *C. elegans* [26]. We conclude by highlighting promising avenues for future research on this topic.

2. *Caenorhabditis elegans* immune effectors

A variety of immune effectors and mechanisms have been described for *C. elegans*. These include the production of reactive oxygen species (ROS), the process of autophagy, and also the expression of putative antimicrobial peptides and proteolytic enzymes (see overview of the latter in table 1). The evidence for an immune role of these effector types will be discussed in detail below.

(a) Reactive oxygen species

ROS are chemically highly reactive molecules such as superoxide (O_2^-) and hydrogen peroxide (H_2O_2) that are generated by the partial reduction of oxygen, mainly during mitochondrial

oxidative metabolism, but also during cellular response to xenobiotics and cytokines [27]. Superoxide is generated by nicotinamide adenine dinucleotide phosphate (NADPH) oxidase (NOX) enzymes, coenzyme Q10, and complexes I and III of the mitochondrial electron transport chain [28]. While ROS are produced as toxic by-products of normal metabolism, they also function as signalling molecules and represent an efficient, highly conserved effector mechanism to eliminate pathogens in animals and plants. For example, ROS are generated by NOX2 in human macrophages to directly kill ingested pathogens [29]. In *Drosophila melanogaster*, ROS are produced by the NOX enzyme dDuox in the gut to limit microbial proliferation [30]. In *C. elegans*, ROS can activate protective cellular mechanisms to promote longevity, pathogen defence responses and wound healing [31–34]. In addition to ROS functioning as signalling mediators, *C. elegans* responds to infection with *E. faecalis* by producing ROS via the NOX Duox1/BLI-3 in the intestine [35,36]. ROS production by intestinal cells represents a protective antimicrobial response: worms are more susceptible to *E. faecalis* infection when ROS production is impaired by reducing *bli-3* expression via RNAi, or when ROS is eliminated by the addition of antioxidants to the medium [36]. The only other pathogen which is as yet known to induce ROS production in the *C. elegans* intestine is the yeast *Saccharomyces cerevisiae*. As for *E. faecalis*, a decrease in ROS production, in this case in a *bli-3* mutant, leads to enhanced susceptibility to *S. cerevisiae* infection [37]. These studies produced experimental evidence for an important role of ROS as microbicidal effectors in *C. elegans*. It remains, however, unclear if the generation of ROS is a general defence mechanism to protect the worm's intestinal epithelium against pathogenic attack and how exactly the generation of ROS is activated upon pathogen exposure.

(b) Autophagy

Autophagy is not a classical effector mechanism *in sensu stricto*, such as the production of ROS or the expression of antimicrobial peptides (AMPs). It may still be involved in the elimination of pathogens, and thus it can provide an additional immune effector process. In the following, we will summarize the evidence for its contribution to *C. elegans* immune defence. Autophagy is a process during which intracellular material is sequestered within double-membrane vesicles (autophagosomes) and then targeted for lysosomal degradation. In this way, autophagy recycles intracellular components to produce energy during nutrient depletion, or removes potentially toxic material which is generated during, for example, oxidative stress, to prevent cellular damage. Autophagy is thus an important part of the protective surveillance machinery of the cell. In addition, autophagy plays a role in defence against pathogens by delivering intracellular microorganisms to lysosomes in both specialized immune cells and epithelial cells [38].

In *C. elegans*, autophagy is required for host defence against *Salmonella enterica* serovar Typhimurium [39,40], *N. parisii* and *S. aureus* [23], and mitophagy (the specific elimination of mitochondria by autophagy) for resistance to *Pseudomonas aeruginosa* [41]. Its contribution to defence against these four pathogens shows some variation, as explained in more detail in the following: (i) *S. enterica*, a facultative intracellular pathogen, invades the *C. elegans* intestinal epithelial cells only in worms in which the autophagy gene *bec-1* (homologous to human BECN1) is silenced by RNAi. In the *bec-1* RNAi animals *S. enterica* replicates intracellularly, leading to cytoplasmic destruction and premature death of the animal [39]. Similar to *bec-1* RNAi, silencing the expression of two other autophagy genes, *lgg-1* (encoding the homologue of mammalian MAP-LC3) and *atg-7*, resulted in animals being more susceptible to *S. enterica* infection. These data provide evidence that in wild-type animals *S. enterica* is efficiently degraded by the autophagy pathway and thus represents an important effector mechanism against an intracellular pathogen. (ii) Infection with the obligate intracellular pathogen *N. parisii* activates expression of several autophagy genes. Knock-down of *lgg-1* and *atg-18* (encoding the homolog of human WIPI1) by RNAi resulted in increased bacterial load, indicating that autophagy is required for controlling *N. parisii* infection. In addition, examination of transgenic worms carrying a translational GFP reporter for LGG-1 revealed that the autophagy machinery is targeted to *N. parisii* cells [42]. (iii) Infection with the extracellular pathogen *S. aureus* also activates the expression of several autophagy genes and this activation is dependent on the bHLH transcription factor HLH-30, which is required for the expression of 80% of the transcriptional *C. elegans* response to *S. aureus*, including AMP genes [23]. *Staphylococcus aureus* infection-induced autophagosome formation could be observed by analysis of the localization of GFP-tagged LGG-1 in the *C. elegans* intestine [23]. Worms in which the autophagy genes *lgg-1*, *unc-51* (homologous to human ULK1) and *vps-34* (homologous to human phosphoinositide 3-kinaseVPS34) are knocked down by RNAi are more susceptible to *S. aureus*-mediated killing. (iv) In the case of *P. aeruginosa* infection, iron chelation through pyoverdin, an iron-binding virulence factor of PA14, causes mitochondrial damage, which in turn triggers mitophagy. Several genes involved in autophagy (e.g. *bec-1*, *lgg-1*, *mboa-7* (homologous to human MBOAT7)) and mitophagy (e.g. *pink-1* (homologous to human PINK1), and *pdr-1* (homologous to human PARK2)) are required for resistance to PA14 in a liquid-based killing assay [41].

Together, these data indicate that the autophagy machinery is required for *C. elegans* defence against infection with the facultative and obligate intracellular pathogens *S. enterica* and *N. parisii*, respectively, and the extracellular pathogens *S. aureus* and *P. aeruginosa*. The exact protective mechanism of autophagy and mitophagy still needs to be determined. While autophagy might act in the direct elimination of intracellular pathogens (as proposed by [39,42]), both autophagy and mitophagy might help the host in coping with the cellular damage caused by infection (as proposed by Visvikis *et al.* [23]), thus indirectly increasing tolerance against extracellular pathogens. The latter idea is supported by the fact that autophagy was shown to protect *C. elegans* against necrosis during *P. aeruginosa* infection [43].

(c) Caenacins and related peptides

The families of neuropeptide-like proteins (NLPs) and caenacins (CNCs) are antimicrobial peptides that are phylogenetically closely related. They were first discovered in a microarray analysis of the *C. elegans* transcriptional response to infection with the fungus *D. coniospora* [44]. Unlike most other *C. elegans* pathogens, which establish an infection in the intestine, *D. coniospora* spores infect the worm by attaching to its cuticle. After spore germination, the fungus penetrates the cuticle and epidermis to invade the worm. The NLPs and

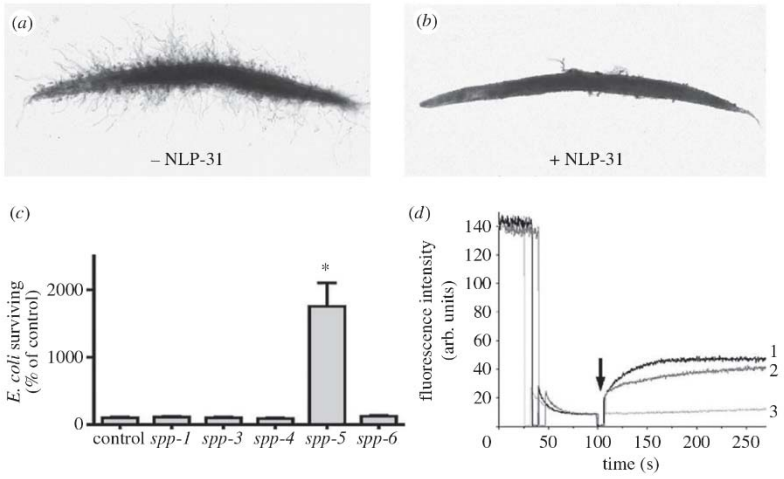


Figure 1. Antimicrobial effects of two *C. elegans* antimicrobial peptides. (*a*,*b*) Inhibitory effect of synthesized NLP-31 on hyphal growth of *D. coniospora in vitro*. (*a*) Worms are overgrown by *D. coniospora* hyphae 48 h post infection. (*b*) The addition of synthetic NLP-31 at 200 µM to infected worms completely inhibits fungal growth. These two images are from [44] and are shown here with permission from Jonathan Ewbank. (*c*) Knock-down of *spp-5* (*caenopores-5*) leads to increased survival of *E. coli* in the *C. elegans* intestine. Homogenates of worms treated with *spp-1*, *spp-3*, *spp-4*, *spp-5*, *spp-6* RNAi or an empty vector control, exposed to ampicillin-resistant *E. coli* for 30 min, and then thoroughly washed, were plated on LB agar plates to count the number of growing bacterial colonies. (*d*) SPP-5 exhibits pore-forming activity. Pore-forming activity was measured by fluorimetrically monitoring the dissipation of a valinomycin-induced diffusion potential in liposomes. The increase in fluorescence over time after addition (arrow) of 0.2 nmol SPP-5 (trace 1) or a control pore-forming peptide (trace 2), but not of the peptide solvent control (trace 3), reflects pore-forming activity. Panels (*c*,*d*) are from [49] and are shown here with permission from Thomas Roeder and Matthias Leippe.

CNCs are small proteins (51–82 amino acids) with signal peptides at their N-terminus. The mature peptides are basic and rich in glycine and aromatic acid [44]. Two AMP genes that are similar to the *C. elegans nlp*s and *cnc*s can be found in *D. melanogaster*; otherwise the gene family seems to be restricted to nematodes. Six *nlp* (*nlp-27*, *nlp-28*, *nlp-29*, *nlp-30*, *nlp-31* and *nlp-34*) and six *cnc* genes (*cnc-1*, *cnc-2*, *cnc-3*, *cnc-4*, *cnc-5* and *cnc-11*) are localized in two separate clusters on chromosome V (named *nlp-29* cluster and *cnc-2* cluster, respectively [45,46]). All *nlp* genes of the *nlp-29* cluster and five of the *cnc* genes in the *cnc-2* cluster (all except *cnc-3*) are highly upregulated after *D. coniospora* infection [44–46]. At the time of discovery, these *nlp* genes had already been annotated as neuropeptide-like genes, as they show some sequence similarity with known *C. elegans* neuropeptides [47]. They form, however, a monophyletic group that is distinct from the other *nlp* genes encoding characterized neuropeptides of the worm's nervous system. Moreover, as described further below, there is evidence for NLP and CNC function as genuine AMPs. Analysis of the genomic distribution and evolutionary history of *C. elegans nlp* genes and their respective orthologues from *C. briggsae* and *C. remanei* provided evidence of recent expansion by gene duplications and of positive selection likely driving the *nlp* gene diversification [45]. Several *nlp* and *cnc* genes, such as *nlp-29*, *nlp-30*, *nlp-31* and *cnc-2* [44–46], were shown to be expressed in the epidermis of the worm where the fungus attacks. Interestingly, the expression of *cnc-2* is only activated by fungal infection, while *nlp-29* expression is also induced by sterile wounding, osmotic stress and in worms with epidermal defects, such as the *dpy-9* and *dpy-10* mutants [45,46,48]. There is experimental evidence from protein and genetic analyses for an important role of NLPs and CNCs in resistance to infection. The chemically synthesized NLP-31 protein inhibited fungal growth in infected worms [44] (figure 1*a*,*b*), although it may not be produced by nematodes themselves at very high concentrations. In addition, overexpression of the complete *nlp-29* or *cnc-2* cluster renders worms more resistant to *D. coniospora* infection [45,46].

The Ewbank lab developed an elegant AMP reporter gene-based approach to decipher the molecular signalling pathways that underlie regulation of *nlp-29* and *cnc-2* expression [46,48]. Using a p*nlp-29*::GFP reporter strain as a tool for candidate gene, forward genetic and RNAi screen approaches, it was possible to characterize a complex signalling pathway, from the upstream infection signal and its receptor to the downstream transcription factor that is required for AMP gene expression. In particular, *nlp-29* expression is upregulated following sterile wounding and fungal infection via detection of the endogenous signal 4-hydroxyphenyllactic acid (HPLA) by the G protein-coupled receptor DCAR-1, which acts upstream of the Gα protein GPA-12, the protein kinase C TPA-1, the TIR-domain adaptor protein TIR-1, a central p38 MAPK cascade and the STAT-like transcription factor STA-2. While the p38 MAPK pathway is indispensable for activation of *nlp-29* expression [48], a non-canonical TGF-β pathway is essential for the induction of *cnc-2* expression [46].

Using WormExp [26], we looked for microorganism-dependent expression of *nlp*/*cnc* genes. For this analysis and also all further assessments below, we considered a total of 111 differentially expressed gene sets from 35 transcriptome studies, which examined the *C. elegans* response to bacteria, fungi and one virus at various exposure time periods (electronic supplementary material, table S1). For convenience, the main figures (figures 2 and 3) show the results for the most responsive gene set per pathogen (i.e. the gene set with the largest number of differentially expressed putative immune effector genes), while the full results across all gene sets are shown in the electronic supplementary material,

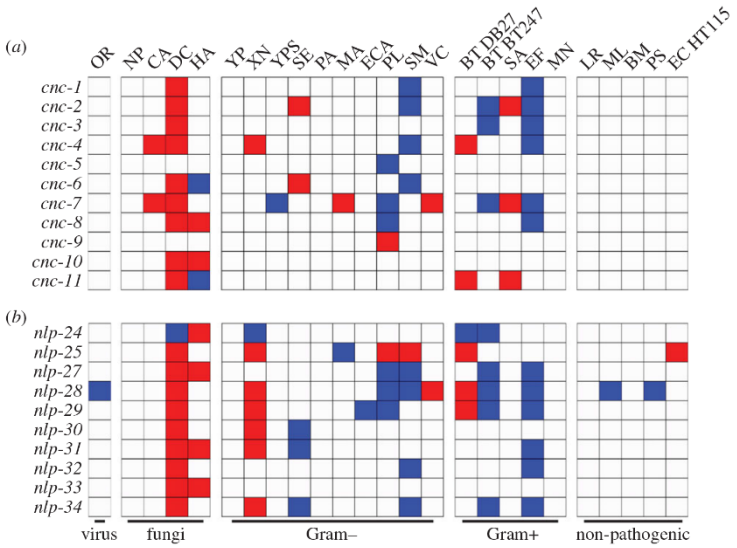


Figure 2. Expression of caenacins and related peptides in response to microbe exposure. Red and blue colours indicate up- and downregulated expression after microbe exposure, respectively. The panels show exemplary results for the considered 25 microbes, chosen from the total of 111 gene sets available for these taxa. The list of all gene sets is given in the electronic supplementary material, table S1 and the full results in the electronic supplementary material, figure S1. OR, Orsay virus; NP, *Nematocida parisii*; CA, *Candida albicans*; DC, *Drechmeria coniospora*; HA, *Harposporium*; YP, *Yersinia pestis*; XN, *Xenorhabdus nematophila*; YPS, *Yersinia pseudotuberculosis*; SE, *Salmonella enterica*; PA, *Pseudomonas aeruginosa*; MA, *Microcystis aeruginosa*; ECA, *Erwinia carotovora*; PL, *Photorhabdus luminescens*; SM, *Serratia marcescens*; VC, *Vibrio cholerae*; BT, *Bacillus thuringiensis*; SA, *Staphylococcus aureus*; EF, *Enterococcus faecalis*; MN, *Microbacterium nematophilum*; LR, *Lactobacillus rhamnosus*; ML, *Micrococcus luteus*; BM, *Bacillus megaterium*; PS, *Pseudomonas* sp.; EC, *Escherichia coli*.

figure S1. Most remarkably, expression of *nlp* and *cnc* genes is mainly induced by infection with the fungal pathogens *D. coniospora* and *Harposporium* sp. (in the case of *cnc-4* and *cnc-7* also with *Candida albicans*), while their expression is mainly downregulated and only upregulated in a few cases towards pathogenic or non-pathogenic bacteria (figure 2). This confirms previous observations [50] and supports the role of NLP and CNC peptides as inducible anti-fungal effectors. One exception is infection with the Gram-positive bacterium *Xenorhabdus nematophila*, which induces the expression of several *cnc* and *nlp* genes (figure 2). Interestingly, *X. nematophila* is one of the few bacterial pathogens that does not infect the *C. elegans* intestine, but adheres to the cuticle and forms biofilms on the head of the worm [51,52]. The tissue affected by the infection is thus the cuticle and the underlying epidermis, as in the case of *D. coniospora*. The fungus *Harposporium* sp., however, establishes an intestinal infection. The expression pattern of the *nlp* genes may thus suggest that both pathogen type (i.e. fungus versus bacterium) and site of infection (i.e. intestine versus epidermis) can influence effector gene expression in the worm [50].

(d) Caenopores

The caenopores belong to an ancient family of antimicrobial peptides with a saposin domain. They share similarities with the saposin-like proteins (SAPLIPs) such as the amoebapores first characterized for *Entamoeba histolytica* [53,54] and the mammalian NK lysin and granulysin [55,56]. In *C. elegans*, the caenopores or saposins (*spp*s) form a gene family of currently 23 members. At the moment, the most comprehensive evidence for an antimicrobial function of *C. elegans* effector molecules is available for this group of peptides. The first caenopore characterized in more detail was SPP-1(T07C4.4) [57]. The recombinant SPP-1 protein was shown to have a helix bundle structure characteristic of saposin-like domains and to exhibit an antibacterial effect on *E. coli* [57]. A very detailed analysis of SPP-5 or caenopore-5 function at the peptide level was performed by Roeder, Leippe and co-workers, including the first and only inference of the structure of a *C. elegans* antimicrobial effector [49,58]. This caenopore is exclusively expressed in the intestine. Silencing *spp-5* by RNAi leads to reduced worm fitness (e.g. reduced offspring production) and highly increased numbers of the food bacterium *E. coli* OP50 in the nematode intestine (figure 1*c*) [49]. This suggests that *spp-5* is required for killing ingested bacteria, which in *C. elegans* can represent both food and pathogenic bacteria. Heterologously expressed SPP-5 was used to demonstrate the protein's ability to permeabilize the membranes of viable bacteria and to induce pores in phospholipid vesicles (figure 1*d*) [49]. Recently, SPP-5 was synthesized through native chemical ligation based on Boc solid-phase peptide synthesis, allowing further structural analysis of the antimicrobial function. This analysis revealed that a 35 residue long N-terminal fragment is sufficient for cell permeability activity [59]. A detailed peptide-level functional analysis was additionally performed for SPP-1, SPP-3 and SPP-12 [60,61]. SPP-1 is expressed in the intestine, SPP-3 in both intestines and a head neuron, while SPP-12 is exclusively expressed in two pharyngeal neurons. Importantly, all three peptides are able to permeabilize a variety of different microorganisms, they can form pores into phospholipid vesicles, and they may therefore contribute to the worm's interaction with microbes. Indeed, when *spp-1* or

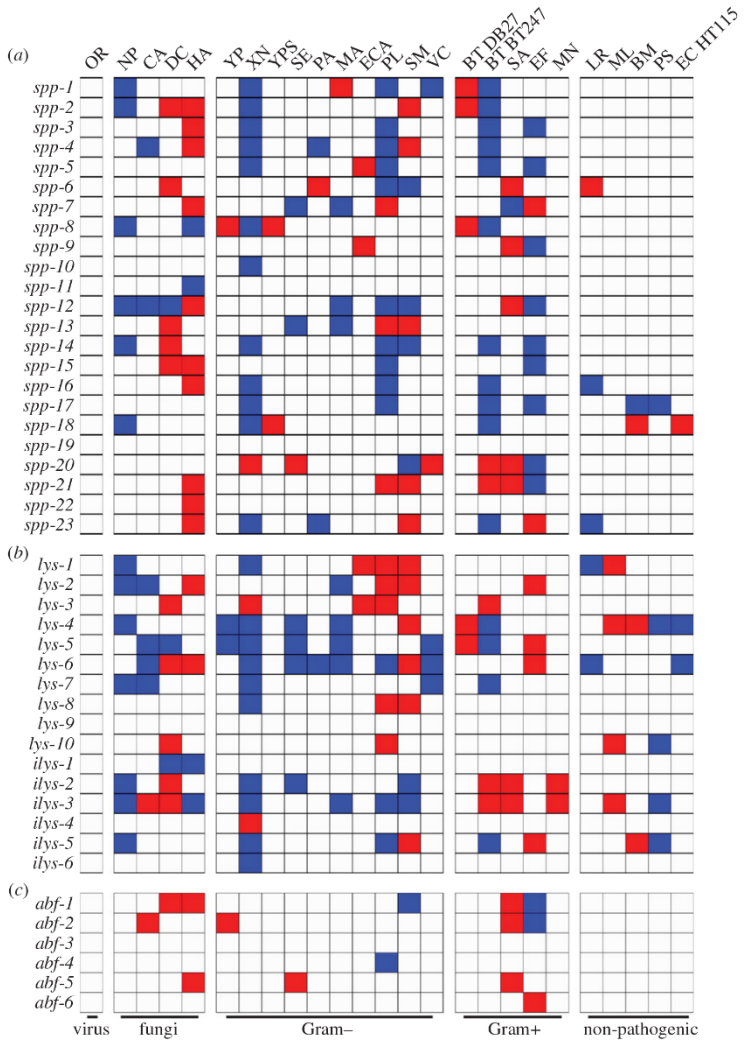


Figure 3. Expression pattern of caenopores, lysozyme and defensin-like AMP genes after microbe exposure. Red and blue colours indicate up- and downregulation of genes, respectively. The figure shows exemplary results of the most responsive gene sets. Full results are given in the electronic supplementary material, figure S1 and the list of considered gene sets in the electronic supplementary material, table S1. OR, Orsay virus; NP, *Nematocida parisii*; CA, *Candida albicans*; DC, *Drechmeria coniospora*; HA, *Harposporium*; YP, *Yersinia pestis*; XN, *Xenorhabdus nematophila*; YPS, *Yersinia pseudotuberculosis*; SE, *Salmonella enterica*; PA, *Pseudomonas aeruginosa*; MA, *Microcystis aeruginosa*; ECA, *Erwinia carotovora*; PL, *Photorhabdus luminescens*; SM, *Serratia marcescens*; VC, *Vibrio cholerae*; BT, *Bacillus thuringiensis*; SA, *Staphylococcus aureus*; EF, *Enterococcus faecalis*; MN, *Microbacterium nematophilum*; LR, *Lactobacillus rhamnosus*; ML, *Micrococcus luteus*; BM, *Bacillus megaterium*; PS, *Pseudomonas* sp.; EC, *Escherichia coli*.

spp-12 expression is knocked down by RNAi, *C. elegans* lifespan on *E. coli* is reduced [62]. Moreover, expression of *spp-1* is induced by infection with *S. enterica* serovar Typhimurium [63]. Furthermore, *S. enterica* strains which lack certain virulence genes and are thus not able to persist in the *C. elegans* intestine, are capable of colonizing *spp-1* RNAi-treated worms.

Despite the considerable knowledge, we have on SPP function on the peptide level, the regulation of *spp* gene expression is much less understood. *spp-1* and *spp-12* were identified as downstream targets of the forkhead box O (FOXO) homologue DAF-16 [62]. *Pseudomonas aeruginosa* is able to manipulate the *C. elegans* immune response by downregulation of immune effector genes, such as *spp-1*, and this repression is dependent on the DAF-2/DAF-16 (ILR/FOXO) pathway [64]. It thus seems that the ILR pathway induces expression of at least some *spp* genes. In addition, expression of several *spp* genes (*spp-1*, *spp-2*, *spp-8*, *spp-15* and *spp-17*) is strongly upregulated in response to *S. aureus* and this upregulation requires the transcription factor HLH-30 [23].

Based on WormExp [26], we further explored differential expression of this gene family and find that all but one of the considered caenopore genes responds to microbe exposure (figure 3*a*; full results given in the electronic supplementary material, figure S1). The only exception refers to *spp-19*. The

strongest gene activation response is shown against fungi, especially the intestinal fungal pathogen *Harposporium* sp., which induces expression of 10 genes. Strong downregulation of *spp* genes is observed after exposure to the Gram-negative bacteria *X. nematophila*, *Photorhabdus luminescens*, the Gram-positive *B. thuringiensis*, *E. faecalis*, and also the intracellular microsporidian fungus *N. parisii*. In general, *spp* genes show differential regulation by all of the considered types of microorganisms, including both pathogenic and also non-pathogenic taxa. The gene *spp-12* appears most responsive to the microbes considered, followed by *spp-1*, *spp- 2*, *spp-4*, *spp-8*, *spp-14*, *spp-20* and *spp-23*. None of the genes shows an identical pattern of differential expression. Taken together, this gene expression pattern suggests that *spp* genes contribute to both digestion and immune defence in a microbe-specific manner.

(e) Lysozymes

Lysozymes are known to contribute to both digestion and immunity in a wide variety of organisms, ranging from bacteria, phages, protists and plants to invertebrate and vertebrate animals [65]. The nematode *C. elegans* is unique in that its 16 lysozyme genes fall into two very distinct lysozyme types with an enormous level of sequence divergence within a single species. Ten of the genes are most closely related to the lysozymes found in protists (*lys*), while the remaining six are of the invertebrate type (*ilys*) [66]. Within these types, the lysozyme genes most likely arose through repeated duplication events across the phylogeny of the genus *Caenorhabditis* and show repeated episodes of positive selection [66]. To date, their exact function in *C. elegans* has not yet been studied at the protein level. Therefore, it is not known whether and how they interact with microbial molecules such as peptidoglycan and thus directly contribute to destruction of microbial membranes. Nevertheless, a likely role in worm immunity was inferred from functional analysis at the gene level, including analysis of knock-out mutants, gene silencing through RNAi, or gene expression analysis in transgenic worms. Most of the genes were shown to be expressed in the intestine, while *lys-1* is additionally expressed in neurons, *lys-7* in larval muscles and *lys-8* in the pharynx (reviewed in [66]). Manipulation of gene activity indicated a role in the interaction with microbes for several lysozyme genes. The overexpression of *lys-1* increased resistance against pathogenic *S. marcescens* [67]. Silencing of either *lys-7* or *ilys-2* led to higher susceptibility towards the pathogen *M. nematophilum* [68]. Gene knock-out of *lys-2*, *lys-5* and *lys-7* enhanced susceptibility to pathogenic *B. thuringiensis*, while overexpression of *lys-5* and *lys-7* but not *lys-2* enhanced resistance relative to a non-pathogenic control [69]. Furthermore, *ilys-2* and *lys-5* RNAi-treated worms are more susceptible to *S. aureus* infection [23].

The expression of several lysozyme genes was shown to be controlled by *C. elegans* immune signalling pathways. For example, *lys-7* and *lys-8* are known targets of DAF-16 (FOXO) [62] and the TGF-β pathway [70,71]. Expression of *lys-2* is regulated by the GATA transcription factor ELT-2 and the p38 MAPK PMK-1 [72]. The transcription factor HLH-30 influences expression of *lys-2*, *ilys-3*, *ilys-4*, *lys-3*, *lys-5* and *lys-10* following *S. aureus* infection [23]. Based on WormExp [26], we find that lysozyme gene expression is differentially regulated by almost all of the considered microbes (figure 3*b*; full results in the electronic supplementary material, figure S1). The only exceptions refer to Orsay virus and the pathogenic bacterium *Yersinia pseudotuberculosis*. Otherwise, lysozymes respond to both pathogenic and also non-pathogenic microorganisms in a highly taxon-specific pattern. The most responsive genes are *lys-6* and *ilys-3*, followed by *lys-4*, *lys-5*, *ilys-2* and *ilys-5* (figure 3*b*). The gene *lys-9* is the only gene which does not show any differential gene expression upon microbe exposure (figure 3*b*). As this gene is very different in sequence from all remaining *C. elegans* lysozymes and found at the end of a long branch within the lysozyme phylogeny [66], *lys-9* may represent a pseudogene or at least show a function unrelated to that of the other lysozymes. Overall, the observed pattern of differential gene expression for the remaining lysozyme genes suggests that they contribute to both defence and digestion in a microbe-specific form.

(f) Defensin-like antimicrobial peptides

The *C. elegans* genome contains six genes with high similarity to the defensin-type antimicrobial peptides, well known from insects and vertebrates to contribute to immune defence [73]. These genes have been named antibacterial factor (*abf*) genes in the nematode. ABF-2 has been characterized at the peptide-level in two studies [74,75]. The heterologously expressed peptide showed high *in vitro* activity against a diversity of microbes, ranging from Gram-negative to Gram-positive bacteria and yeasts. The exact mode of action still requires further examination. *abf-2* is mainly expressed in the pharynx, and, as a secretory peptide, it is likely present in the lumen of the pharynx and the gut. Its expression is induced by infection with *S. enterica* serovar Typhimurium, as monitored by qRT PCR, and *abf-2* knock-down by RNAi resulted in an increased infection load [63]. The signalling pathways that control *abf* gene expression have not been fully deciphered. *abf-2* expression is regulated by the transcription factor HLH-30 following *S. aureus* infection [23]. Upregulation of *abf-1* and *abf-2* by infection with *Cryptococcus neoformans* requires the scavenger receptor CED-1 (homologous to human SCARF1) and C03F11.3 (homologous to human CD36) [76]. Moreover, expression of *abf-1* was demonstrated to be dependent on the gene *npr-1* in the context of *P. aeruginosa* infection [77]. The transcriptome database WormExp [26] revealed that five out of the six *abf* genes respond to microbes and, if so, to only very few microbial taxa (figure 3*c*; electronic supplementary material, figure S1). In particular, *abf-3* does not show any differential gene expression after exposure to the various microorganisms and *abf-4* is downregulated by only *P. luminescens*. The remaining four genes can be activated by a total of seven pathogen strains (including for example *S. aureus*) and repressed by two other pathogens (*E. faecalis and S. marcescens*). Taken together, the *abf* genes appear to play a less prominent role in the inducible defence against pathogens or the inducible response to food microbes than the other above highlighted gene families (also see further discussion in §3). It is still possible that they are important in the constitutively expressed protection of the worm against pathogens and/or enhance the pathogen-specific response mediated by one of the other gene families with antimicrobial functions.

3. Future challenges: functional evidence for worm immune effectors and the involvement of the *Caenorhabditis elegans* microbiota

The model nematode *C. elegans* possesses a large repertoire of potential effector proteins and mechanisms to defend itself

against pathogen attack. Our review assessed the involvement of putative effector gene families through their differential expression upon pathogen exposure, using the database WormExp [26]. To enhance comparability of the very diverse individual transcriptome datasets, this database includes only those genes in the gene sets that were found to be significantly differentially expressed, and it then only uses their presence/absence in the gene sets for all further analyses. This approach may have less sensitivity than analyses based on exact expression fold changes for all nematode genes. Yet, it also reduces the level of noise, often prevalent in transcriptome datasets, and thus it may help to identify the more robust overlaps in gene expression among conditions. A more sensitive analysis of pathogen-induced gene expression should be based on parallel assessment of various pathogens under exactly the same conditions, as previously performed for some pathogenic strains by the Ewbank lab [50]. Similarly, the *C. elegans* transcriptome studies usually use entire worm populations of whole animals, characterized at only one or two time points and usually one or two developmental stages. Although strong expression responses of putative effector genes should be visible in whole animal samples, subtle expression variations that are only shown by certain tissues or life stages may go undetected in such a crude approach. In the future, it would thus be of particular value to perform tissue-specific gene expression analyses across different life stages, taking into account the different modes of infection (intestinal infection, intracellular infections, infection of the cuticle; see §1). Based on our current approach, we can nevertheless conclude that there are at least some differences in microbe-responsiveness among the considered gene families, as discussed in more detail below.

It is additionally important to emphasize that differential expression after pathogen exposure does not suffice to prove an antimicrobial function of these genes. Such functional evidence needs to be obtained at the peptide or protein level. On the one hand, this may be achieved *in vivo* by silencing of the gene. A common approach for this relies on RNAi, which may, however, not always work with high efficiency, especially if genes are expressed in specific cell types such as neurons [78,79]. Another approach is to use knock-out mutants, which are already available for a large number of *C. elegans* genes [80] and which can be produced through application of CRISPR/Cas technology [81,82]. On the other hand, additional information on antimicrobial functions can be obtained through analysis of synthesized proteins and peptides. To date, such a protein/peptide-level analysis has only been performed for six *C. elegans* immune effectors: NLP-31, SPP-1, SPP-3, SPP-5, SPP-12 and ABF-2 [44,49,59–61,74]. To complicate matters further, evidence for an antimicrobial effect is still insufficient for demonstrating the protein's role in *C. elegans* pathogen defence. As this nematode feeds on a variety of microorganisms in the wild [83] and on *E. coli* bacteria in the laboratory, peptides and proteins which can damage or break down bacterial cell walls might also function in digestion. In *C. elegans*, this is most likely the case for those effectors which we here show to be inducible by non-pathogenic microbes (e.g. *lys-4*, *ilys-3* and *spp-18*; figure 3), and also those which are constitutively expressed, as previously shown for *spp-5* [49], *abf-1*, *abf-2* and *abf-3* [63,71,74]. As suggested for *spp-5* [49], constitutively expressed effector genes might function in both defence and digestion, because they could enable the worm to access bacteria-derived nutrients and at the same time eliminate potential pathogens. Such a dual function of effectors is similarly discussed for *Drosophila* fruitflies, which also inhabit microbe-rich environments (see review by Broderick [84]).

Microbe responsive genes are especially found among the *nlp/cnc*, lysozyme and *spp* gene families (figures 2 and 3). These might thus play more general roles in immunity, as most clearly demonstrated thus far for the *nlp* and *cnc* genes and their particular contribution to anti-fungal defence (e.g. [21,44,45]). It is worth noting that none of the considered effector genes show a change in expression after infection with Orsay virus (except downregulation of *nlp-28*). This may suggest that the protective cellular response to viral infection mainly depends on the RNAi pathway in *C. elegans* [8], or that other as yet uncharacterized effector mechanisms contribute to virus elimination. Similarly, infection by the other known intracellular pathogen *N. parisii* only represses but does not activate expression of the considered effector genes. In this case, defence may also rely on as yet uncharacterized effector genes or different types of protective mechanisms (e.g. [16,42]). These observations also suggest that the immune effectors considered here mainly act as secreted proteins targeting extracellular microbes in the intestinal lumen, the pseudocoel or the cuticle. In comparison to the other *C. elegans* effector genes, the *abf* genes are inducible by only a few pathogens (figures 2 and 3). This may suggest that they contribute to defence and/or digestion in a form clearly distinct from the other microbe-inducible effectors. These *abf* genes may thus also have a less prominent role as inducible immune effectors than the homologous defensin-like genes in arthropods (see [3,85,86] in this issue).

Next to the immune effector genes which we have explored in the current review, additional gene families and/or processes may play an as yet undiscovered role in eliminating pathogens. Four gene families have been repeatedly highlighted in this context, although an antimicrobial function of these genes has not yet been demonstrated. One of these families comprises the C-type lectins, originally described as Ca^{2+}-dependent (C-type) glycan binding (lectin) proteins, but now known as a diverse group of proteins which may bind to proteins, lipids and nucleic acids in both Ca^{2+}-dependent and -independent ways. All C-type lectins share a highly conserved domain, the carbohydrate recognition (CRD) or C-type lectin-like domain (CTLD). The CTLD gene family is highly diverse in *C. elegans*, comprising more than 280 genes and being the seventh most abundant gene family in the worm (reviewed in [87]). Although vertebrate CTLD proteins are known to be involved in pathogen recognition, some mammalian CTLD proteins of the RegIII family possess antibacterial activity and function in pathogen elimination [88,89]. Similarly, an *in vitro* bactericidal activity was also described for several crustacean CTLD proteins [87]. In *C. elegans*, the majority of CTLD proteins contain a signal peptide and are thus predicted to be secreted. Moreover, the expression of the majority of *C. elegans* CTLD (*clec*) genes is induced by pathogen infection, showing a highly specific pattern of regulation, as recently evaluated by us with a similar approach to that used here [87]. In addition, several *clec* genes are required for resistance to infection as demonstrated in functional genetic analyses using mutant strains or RNAi (reviewed in [87]). The exact function of CTLD proteins in *C. elegans* immunity is still unclear. To date, only one study has assessed the function of

these genes in the context of an immune response at the protein level, demonstrating that the two CTLD proteins CLEC-39 and CLEC-49 are able to bind to a bacterial pathogen, in this case *S. marcescens* [90]. Although *clec-39* and *clec-49* mutant worms are more susceptible to *S. marcescens* infection, the proteins CLEC-39 and CLEC-49 do not have an inhibitory effect on *S. marcescens* growth *in vitro*. These results suggest that the genes either function in recognition or do not mediate pathogen elimination alone but perhaps in collaboration with other effectors. A more detailed discussion of possible immune functions of these proteins has recently been published elsewhere [87].

Two additional groups of putative effectors are the fungal-induced peptides (*fip*) and *fip*-related peptides (*fipr*). These genes are induced in expression upon infection with fungal pathogens such as *D. coniospora* or *Harposporium* sp. [50]. *Caenorhabditis elegans* possesses seven *fip* genes and 29 *fipr* genes which generally vary in their expression upon pathogen exposure (electronic supplementary material, figures S1 and S2). As they encode proteins that are less than 100 amino acids in size and are predicted to have signal peptides, it is likely that *fip* and *fipr* genes encode AMPs. Experimental evidence for their contribution to *C. elegans* anti-fungal defence or antimicrobial activity is so far missing [91].

Yet another group are the thaumatin-like proteins. These are small proteins around 200 amino acids in length which act as anti-fungal defence proteins in plants. Their anti-fungal properties are likely based on beta-1,3-glucanase activity, alpha-amylase inhibiting properties (reviewed in [92]) and/or membrane-permeabilizing activity [93,94]. There are eight homologues of thaumatin encoding (*thn*) genes in *C. elegans*. While it is not known if *C. elegans* thaumatins exhibit anti-fungal activity, three of the eight *thn* genes (*thn-1*, *thn-2* and *thn-3*) are both induced and repressed by infection with fungal as well as bacterial pathogens, whereas their expression upon exposure to non-pathogenic micro-organisms remains unchanged (electronic supplementary material, figures S1 and S3). A possible immune function of *thn* genes was further suggested by altered resistance to *P. aeruginosa* infection after RNAi knock-down of *thn-1* and *thn-2* [24,64]. Moreover, *thn-2* expression depends on the ILR pathway and can also be directly manipulated by pathogenic *P. aeruginosa* [24,64]. It remains to be determined whether *C. elegans* thaumatins indeed act as *bona fide* antimicrobial effectors. Moreover, it is possible that in future, still other *C. elegans* gene families may be discovered to possess antimicrobial activity.

In conclusion, the nematode *C. elegans* possesses a variety of putative antimicrobial peptides and additional antimicrobial mechanisms, some of which have been characterized in depth at both genetic and protein level, especially certain caenopores. Nevertheless, the exact immune function of most of these putative effectors still needs further clarification. In future, it would be essential to demonstrate at the protein level *in vitro* that these effectors are able to interact with either non-pathogenic and/or pathogenic microbes and break up bacterial cell membranes or inhibit bacterial growth or viability in some other way. Moreover, it is similarly important that the function of the genes is studied *in vivo* in the worm, using the available tools for *C. elegans* gene manipulation at the cellular level in combination with microscopic dissection of the resulting infection pattern produced by various microbes.

In this context, it would also be of particular interest to assess how different immune effectors interact with each other to affect microbe proliferation. Effector molecules are usually studied in isolation, although several effector genes are simultaneously expressed in response to pathogen infection and some effector proteins are known to exert their antimicrobial activity in synergy with other immune effectors to enhance their potency ([95]; see also [85,86,96]). Interestingly, the mixture of expressed effector genes seems to be highly specific (figures 2 and 3). In fact, we do not find identical patterns of co-expressed genes in response to the various microbes, possibly suggesting specifically fine-tuned immune defences. Such fine-tuning is likely orchestrated by interconnected signalling processes which integrate information from various stimuli, including microbial molecules and also the cellular consequences of pathogen infection (i.e. cellular damage; reviewed in [17]). However, the exact regulation of *C. elegans* effector gene expression is largely unexplored. The main exception refers to *nlp-29*, for which an endogenous danger signal triggers gene expression [46] and which is regulated through a complex signalling network (e.g. [21,25,45,46]). Further analysis of the exact regulation of antimicrobial effector genes would be of great value for understanding their role in defence. The screening approach developed by the Ewbank lab, based on an AMP gene reporter strain, may be of particular promise in this context.

Last but not least, it is conceivable that antimicrobial effectors in *C. elegans* are also used to control the worm's microbiome, in analogy to what is known for example for weevils (see review by Masson *et al.* [97]) and proposed for fruitflies (see review by Broderick [84]). At the same time, it is possible that members of the worm's microbial associates themselves produce protective antimicrobial factors, thus increasing the nematode's arsenal of effector molecules. This nematode seems to contain a rich microbial flora [83,98,99], yet the exact species composition and functions of the microbiome of natural *C. elegans* isolates have not yet been published. The production of antimicrobial compounds, for example bacteriocins or specific anti-fungal proteins, is known for microbiota members of various host taxa, ranging for example from Hydra polyps [100] to humans (e.g. [101]). For *C. elegans*, co-cultivation with bacteria such as *Lactobacillus acidophilus* or *Pseudomonas mendocina* enhanced resistance against pathogens [102,103]. In these two cases, it is unclear whether the tested bacteria produce antimicrobial factors themselves or stimulate their production in the worm. Similarly, we lack any information on possible antimicrobial functions of bacteria associated with natural *C. elegans* isolates. Future characterization of the worm's microbiome should thus specifically assess to what extent individual bacterial strains may enhance nematode immune defence.

Data accessibility. All used transcriptome datasets are available from WormExp, http://wormexp.zoologie.uni-kiel.de/.

Authors' contributions. All authors jointly wrote the manuscript.

Competing interests. We have no competing interests.

Funding. We are grateful for support from the German Science Foundation to K.D. (DI 1687/1 and project A1 of CRC 1182) and H.S. (SCHU 1415/8; SCHU 1415/9, and project A1 of CRC 1182). W.Y. is additionally supported by the International Max-Planck Research School (IMPRS) for Evolutionary Biology at the University of Kiel.

Acknowledgements. We thank the members of the Schulenburg group for feedback, especially Barbara Pees and Alejandra Zárate-Potes.

References

1. Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997 Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489–493. (doi:10.1038/387489a0)
2. Dunn CW *et al.* 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749. (doi:10.1038/nature06614)
3. Destoumieux-Garzón D, Rosa RD, Schmitt P, Barreto C, Vidal-Dupiol J, Mitta G, Gueguen Y, Bachère E. 2016 Antimicrobial peptides in marine invertebrate health and disease. *Phil. Trans. R. Soc. B* **371**, 20150300. (doi:10.1098/rstb.2015.0300)
4. Tan M-W, Mahajan-Miklos S, Ausubel FM. 1999 Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model mammalian bacterial pathogenesis. *Proc. Natl Acad. Sci. USA* **96**, 715–720. (doi:10.1073/pnas.96.2.715)
5. Mahajan-Miklos S, Tan M-W, Rahme LG, Ausubel FM. 1999 Molecular mechanisms of bacterial virulence elucidated using a *Pseudomonas aeruginosa–Caenorhabditis elegans* pathogenesis model. *Cell* **96**, 47–56. (doi:10.1016/S0092-8674(00)80958-7)
6. Engelmann I, Pujol N. 2010 Innate immunity in *C. elegans*. In *Invertebrate immunity* (ed. K Söderhäll), pp. 105–121. Berlin, Germany: Springer.
7. Troemel ER, Félix M-A, Whiteman NK, Barrière A, Ausubel FM. 2008 Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. *PLoS Biol.* **6**, e309. (doi:10.1371/journal.pbio.0060309)
8. Félix M-A *et al.* 2011 Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. *PLoS Biol.* **9**, e1000586. (doi:10.1371/journal.pbio.1000586)
9. Grewal P, Hand P. 1992 Effects of bacteria isolated from a saprophagous rhabditid nematode *Caenorhabditis elegans* on the mycelial growth of *Agaricus bisporus*. *J. Appl. Bacteriol.* **72**, 173–179. (doi:10.1111/j.1365-2672.1992.tb01820.x)
10. Hodgkin J, Félix M-A, Clark LC, Stroud D, Gravato-Nobre MJ. 2013 Two *Leucobacter* strains exert complementary virulence on *Caenorhabditis* including death by worm-star formation. *Curr. Biol.* **23**, 2157–2161. (doi:10.1016/j.cub.2013.08.060)
11. Hodgkin J, Kuwabara PE, Corneliussen B. 2000 A novel bacterial pathogen, *Microbacterium nematophilum*, induces morphological change in the nematode *C. elegans*. *Curr. Biol.* **10**, 1615–1618. (doi:10.1016/S0960-9822(00)00867-8)
12. Jansson H-B. 1994 Adhesion of conidia of *Drechmeria coniospora* to *Caenorhabditis elegans* wild type and mutants. *J. Nematol.* **26**, 430.
13. Pukkila-Worley R, Ausubel FM. 2012 Immune defense mechanisms in the *Caenorhabditis elegans* intestinal epithelium. *Curr. Opin. Immunol.* **24**, 3–9. (doi:10.1016/j.coi.2011.10.004)
14. Ermolaeva MA, Schumacher B. 2014 Insights from the worm: the *C. elegans* model for innate immunity. *Semin. Immunol.* **26**, 303–309. (doi:10.1016/j.smim.2014.04.005)
15. Meisel JD, Kim DH. 2014 Behavioral avoidance of pathogenic bacteria by *Caenorhabditis elegans*. *Trends Immunol.* **35**, 465–470. (doi:10.1016/j.it.2014.08.008)
16. Cohen LB, Troemel ER. 2015 Microbial pathogenesis and host defense in the nematode *C. elegans*. *Curr. Opin. Microbiol.* **23**, 94–101. (doi:10.1016/j.mib.2014.11.009)
17. Ewbank JJ, Pujol N. 2016 Local and long-range activation of innate immunity by infection and damage in *C. elegans*. *Curr. Opin. Immunol.* **38**, 1–7. (doi:10.1016/j.coi.2015.09.005)
18. Schulenburg H, Léopold Kurz C, Ewbank JJ. 2004 Evolution of the innate immune system: the worm perspective. *Immunol. Rev.* **198**, 36–58. (doi:10.1111/j.0105-2896.2004.0125.x)
19. Schulenburg H, Ewbank JJ. 2007 The genetics of pathogen avoidance in *Caenorhabditis elegans*. *Mol. Microbiol.* **66**, 563–570. (doi:10.1111/j.1365-2958.2007.05946.x)
20. Ashe A, Bélicard T, Le Pen J, Sarkies P, Frézal L, Lehrbach NJ, Félix M-A, Miska EA. 2013 A deletion polymorphism in the *Caenorhabditis elegans* RIG-I homolog disables viral RNA dicing and antiviral immunity. *Elife* **2**, e00994. (doi:10.7554/eLife.00994)
21. Zugasti O, Bose N, Squiban B, Belougne J, Kurz CL, Schroeder FC, Pujol N, Ewbank JJ. 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. (doi:10.1038/ni.2957)
22. Melo JA, Ruvkun G. 2012 Inactivation of conserved *C. elegans* genes engages pathogen- and xenobiotic-associated defenses. *Cell* **149**, 452–466. (doi:10.1016/j.cell.2012.02.050)
23. Visvikis O, Ihuegbu N, Labed SA, Luhachack LG, Alves A-MF, Wollenberg AC, Stuart LM, Stormo GD, Irazoqui JE. 2014 Innate host defense requires TFEB-mediated transcription of cytoprotective and antimicrobial genes. *Immunity* **40**, 896–909. (doi:10.1016/j.immuni.2014.05.002)
24. Shapira M, Hamlin BJ, Rong J, Chen K, Ronen M, Tan M-W. 2006 A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. *Proc. Natl Acad. Sci. USA* **103**, 14 086–14 091. (doi:10.1073/pnas.0603424103)
25. Dierking K, Polanowska J, Omi S, Engelmann I, Gut M, Lembo F, Ewbank JJ, Pujol N. 2011 Unusual regulation of a STAT protein by an SLC6 family transporter in *C. elegans* epidermal innate immunity. *Cell Host Microb.* **9**, 425–435. (doi:10.1016/j.chom.2011.04.011)
26. Yang W, Dierking K, Schulenburg H. 2016 WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. *Bioinformatics* **32**, 943–945. (doi:10.1093/bioinformatics/btv667)
27. Ray PD, Huang B-W, Tsuji Y. 2012 Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell. Signal.* **24**, 981–990. (doi:10.1016/j.cellsig.2012.01.008)
28. Johnston AD, Ebert PR. 2012 The redox system in *C. elegans*, a phylogenetic approach. *J. Toxicol.* **2012**, 1–20. (doi:10.1155/2012/546915)
29. Roos D, van Bruggen R, Meischl C. 2003 Oxidative killing of microbes by neutrophils. *Microb. Infect.* **5**, 1307–1315. (doi:10.1016/j.micinf.2003.09.009)
30. Ha E-M, Oh C-T, Bae YS, Lee W-J. 2005 A direct role for dual oxidase in *Drosophila* gut immunity. *Science* **310**, 847–850. (doi:10.1126/science.1117311)
31. Xu S, Chisholm AD. 2014 *C. elegans* epidermal wounding induces a mitochondrial ROS burst that promotes wound repair. *Dev. Cell* **31**, 48–60. (doi:10.1016/j.devcel.2014.08.002)
32. Lee S-J, Hwang AB, Kenyon C. 2010 Inhibition of respiration extends *C. elegans* life span via reactive oxygen species that increase HIF-1 activity. *Curr. Biol.* **20**, 2131–2136. (doi:10.1016/j.cub.2010.10.057)
33. Hwang AB *et al.* 2014 Feedback regulation via AMPK and HIF-1 mediates ROS-dependent longevity in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **111**, E4458–E4467. (doi:10.1073/pnas.1411199111)
34. Van Der Hoeven R, McCallum KC, Cruz MR, Garsin DA. 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**, e1002453. (doi:10.1371/journal.ppat.1002453)
35. Chávez V, Mohri-Shiomi A, Maadani A, Vega LA, Garsin DA. 2007 Oxidative stress enzymes are required for DAF-16-mediated immunity due to generation of reactive oxygen species by *Caenorhabditis elegans*. *Genetics* **176**, 1567–1577. (doi:10.1534/genetics.107.072587)
36. Chávez V, Mohri-Shiomi A, Garsin DA. 2009 Ce-Duox1/BLI-3 generates reactive oxygen species as a protective innate immune mechanism in *Caenorhabditis elegans*. *Infect. Immun.* **77**, 4983–4989. (doi:10.1128/IAI.00627-09)
37. Jain C, Yun M, Politz SM, Rao RP. 2009 A pathogenesis assay using *Saccharomyces cerevisiae* and *Caenorhabditis elegans* reveals novel roles for yeast AP-1, Yap1, and host dual oxidase BLI-3 in fungal pathogenesis. *Eukaryot. Cell* **8**, 1218–1227. (doi:10.1128/EC.00367-08)
38. Randow F, Youle RJ. 2014 Self and nonself: how autophagy targets mitochondria and bacteria. *Cell Host Microb.* **15**, 403–411. (doi:10.1016/j.chom.2014.03.012)
39. Jia K, Thomas C, Akbar M, Sun Q, Adams-Huet B, Gilpin C, Levine B. 2009 Autophagy genes protect against *Salmonella typhimurium* infection and mediate insulin signaling-regulated pathogen resistance. *Proc. Natl Acad. Sci.USA* **106**, 14 564–14 569. (doi:10.1073/pnas.0813319106)

40. Curt A, Zhang J, Minnerly J, Jia K. 2014 Intestinal autophagy activity is essential for host defense against *Salmonella typhimurium* infection in *Caenorhabditis elegans*. *Dev. Comp. Immunol.* **45**, 214–218. (doi:10.1016/j.dci.2014.03.009)
41. Kirienko NV, Ausubel FM, Ruvkun G. 2015 Mitophagy confers resistance to siderophore-mediated killing by *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **112**, 1821–1826. (doi:10.1073/pnas.1424954112)
42. Bakowski MA, Desjardins CA, Smelkinson MG, Dunbar TA, Lopez-Moyado IF, Rifkin SA, Cuomo CA, Troemel ER. 2014 Ubiquitin-mediated response to microsporidia and virus infection in *C. elegans*. *PLoS Pathog.* **10**, e1004200. (doi:10.1371/journal.ppat.1004200)
43. Zou C-G, Ma Y-C, Dai L-L, Zhang K-Q. 2014 Autophagy protects *C. elegans* against necrosis during *Pseudomonas aeruginosa* infection. *Proc. Natl Acad. Sci. USA* **111**, 12 480–12 485. (doi:10.1073/pnas.1405032111)
44. Couillault C, Pujol N, Reboul J, Sabatier L, Guichou J-F, Kohara Y, Ewbank JJ. 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. (doi:10.1038/ni1060)
45. Pujol N, Zugasti O, Wong D, Couillault C, Kurz CL, Schulenburg H, Ewbank JJ. 2008 Anti-fungal innate immunity in *C. elegans* is enhanced by evolutionary diversification of antimicrobial peptides. *PLoS Pathog.* **4**, e1000105. (doi:10.1371/journal.ppat.1000105)
46. Zugasti O, Ewbank JJ. 2009 Neuroimmune regulation of antimicrobial peptide expression by a noncanonical TGF-β signaling pathway in *Caenorhabditis elegans* epidermis. *Nat. Immunol.* **10**, 249–256. (doi:10.1038/ni.1700)
47. Nathoo AN, Moeller RA, Westlund BA, Hart AC. 2001 Identification of neuropeptide-like protein gene families in *Caenorhabditis elegans* and other species. *Proc. Natl Acad. Sci. USA* **98**, 14 000–14 005. (doi:10.1073/pnas.241231298)
48. Pujol N, Cypowyj S, Ziegler K, Millet A, Astrain A, Goncharov A, Jin Y, Chisholm AD, Ewbank JJ. 2008 Distinct innate immune responses to infection and wounding in the *C. elegans* epidermis. *Curr. Biol.* **18**, 481–489. (doi:10.1016/j.cub.2008.02.079)
49. Roeder T, Stanisak M, Gelhaus C, Bruchhaus I, Grötzinger J, Leippe M. 2010 Caenopores are antimicrobial peptides in the nematode *Caenorhabditis elegans* instrumental in nutrition and immunity. *Dev. Comp. Immunol.* **34**, 203–209. (doi:10.1016/j.dci.2009.09.010)
50. Engelmann I, Griffon A, Tichit L, Montañana-Sanchis F, Wang G, Reinke V, Waterston RH, Hillier LW, Ewbank JJ. 2011 A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in *C. elegans*. *PLoS ONE* **6**, e19055. (doi:10.1371/journal.pone.0019055)
51. Drace K, Darby C. 2008 The hmsHFRS operon of *Xenorhabdus nematophila* is required for biofilm attachment to *Caenorhabditis elegans*. *Appl. Environ. Microbiol.* **74**, 4509–4515. (doi:10.1128/AEM.00336-08)
52. Couillault C, Ewbank JJ. 2002 Diverse bacteria are pathogens of *Caenorhabditis elegans*. *Infect. Immun.* **70**, 4705–4707. (doi:10.1128/IAI.70.8.4705-4707.2002)
53. Leippe M, Ebel S, Schoenberger OL, Horstmann RD, Müller-Eberhard HJ. 1991 Pore-forming peptide of pathogenic *Entamoeba histolytica*. *Proc. Natl Acad. Sci. USA* **88**, 7659–7663. (doi:10.1073/pnas.88.17.7659)
54. Leippe M, Andrä J, Nickel R, Tannich E, Muller-Eberhard HJ. 1994 Amoebapores, a family of membranolytic peptides from cytoplasmic granules of *Entamoeba histolytica*: isolation, primary structure, and pore bacterial cytoplasmic membranes. *Mol. Microbiol.* **14**, 895–904. (doi:10.1111/j.1365-2958.1994.tb01325.x)
55. Leippe M. 1995 Ancient weapons: NK-lysin, is a mammalian homolog to pore-forming peptides of a protozoan parasite. *Cell* **83**, 17–18. (doi:10.1016/0092-8674(95)90229-5)
56. Leippe M. 1999 Antimicrobial and cytolytic polypeptides of amoeboid protozoa-effector molecules of primitive phagocytes. *Dev. Comp. Immunol.* **23**, 267–279. (doi:10.1016/S0145-305X(99)00010-5)
57. Bányai L, Patthy L. 1998 Amoebapore homologs of *Caenorhabditis elegans*. *Biochim. Biophys. Acta* **1429**, 259–264. (doi:10.1016/S0167-4838(98)00237-4)
58. Mysliwy J, Dingley AJ, Stanisak M, Jung S, Lorenzen I, Roeder T, Leippe M, Grötzinger J. 2010 Caenopore-5: the three-dimensional structure of an antimicrobial protein from *Caenorhabditis elegans*. *Dev. Comp. Immunol.* **34**, 323–330. (doi:10.1016/j.dci.2009.11.003)
59. Medini K, Harris PW, Hards K, Dingley AJ, Cook GM, Brimble MA. 2015 Chemical synthesis of a pore-forming antimicrobial protein, Caenopore-5, by using native chemical ligation at a Glu-Cys site. *ChemBioChem* **16**, 328–336. (doi:10.1002/cbic.201402513)
60. Hoeckendorf A, Stanisak M, Leippe M. 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**, 205–212. (doi:10.1042/BJ20112102)
61. Hoeckendorf A, Leippe M. 2012 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.* **38**, 181–186. (doi:10.1016/j.dci.2012.05.007)
62. Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H, Kenyon C. 2003 Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424**, 277–283. (doi:10.1038/nature01789)
63. Alegado RA, Tan MW. 2008 Resistance to antimicrobial peptides contributes to persistence of *Salmonella typhimurium* in the *C. elegans* intestine. *Cell. Microbiol.* **10**, 1259–1273. (doi:10.1111/j.1462-5822.2008.01124.x)
64. Evans EA, Kawli T, Tan M-W. 2008 *Pseudomonas aeruginosa* suppresses host immunity by activating the DAF-2 insulin-like signaling pathway in *Caenorhabditis elegans*. *PLoS Pathog.* **4**, e1000175. (doi:10.1371/journal.ppat.1000175)
65. Jollès P. 1996 *Lysozymes: model enzymes in biochemistry and biology* (ed. P Jollès), EXS no. 75. Basel, Switzerland: Birkhäuser.
66. Schulenburg H, Boehnisch C. 2008 Diversification and adaptive sequence evolution of *Caenorhabditis lysozymes* (Nematoda: Rhabditidae). *BMC Evol. Biol.* **8**, 114. (doi:10.1186/1471-2148-8-114)
67. Mallo GV, Kurz CL, Couillault C, Pujol N, Granjeaud S, Kohara Y, Ewbank JJ. 2002 Inducible antibacterial defense system in *C. elegans*. *Curr. Biol.* **12**, 1209–1214. (doi:10.1016/S0960-9822(02)00928-4)
68. O'Rourke D, Baban D, Demidova M, Mott R, Hodgkin J. 2006 Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. *Genome Res.* **16**, 1005–1016. (doi:10.1101/gr.50823006)
69. Boehnisch C, Wong D, Habig M, Isermann K, Michiels NK, Roeder T, May RC, Schulenburg H. 2011 Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. *PLoS ONE* **6**, e24619. (doi:10.1371/journal.pone.0024619)
70. Mochii M, Yoshida S, Morita K, Kohara Y, Ueno N. 1999 Identification of transforming growth factor-β-regulated genes in *Caenorhabditis elegans* by differential hybridization of arrayed cDNAs. *Proc. Natl Acad. Sci. USA* **96**, 15 020–15 025. (doi:10.1073/pnas.96.26.15020)
71. Alper S, McBride SJ, Lackford B, Freedman JH, Schwartz DA. 2007 Specificity and complexity of the *Caenorhabditis elegans* innate immune response. *Mol. Cell. Biol.* **27**, 5544–5553. (doi:10.1128/MCB.02070-06)
72. Troemel ER, Chu SW, Reinke V, Lee SS, Ausubel FM, Kim DH. 2006 p38 MAPK regulates expression of immune response genes and contributes to longevity in *C. elegans*. *PLoS Genet.* **2**, e183. (doi:10.1371/journal.pgen.0020183)
73. Zasloff M. 2002 Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389–395. (doi:10.1038/415389a)
74. Kato Y, Aizawa T, Hoshino H, Kawano K, Nitta K, Zhang H. 2002 abf-1 and abf-2, ASABF-type antimicrobial peptide genes in *Caenorhabditis elegans*. *Biochem. J.* **361**, 221–230. (doi:10.1042/bj3610221)
75. Tomisawa S *et al.* 2013 Overexpression of an antimicrobial peptide derived from *C. elegans* using an aggregation-prone protein coexpression system. *AMB Express* **3**, 45. (doi:10.1186/2191-0855-3-45)
76. Means TK *et al.* 2009 Evolutionarily conserved recognition and innate immunity to fungal pathogens by the scavenger receptors SCARF1 and CD36. *J. Exp. Med.* **206**, 637–653. (doi:10.1084/jem.20082109)

77. Styer KL, Singh V, Macosko E, Steele SE, Bargmann CI, Aballay A. 2008 Innate immunity in *Caenorhabditis elegans* is regulated by neurons expressing NPR-1/GPCR. *Science* **322**, 460–464. (doi:10.1126/science.1163673)
78. Timmons L, Court DL, Fire A. 2001 Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene* **263**, 103–112. (doi:10.1016/S0378-1119(00)00579-5)
79. Asikainen S, Vartiainen S, Lakso M, Nass R, Wong G. 2005 Selective sensitivity of *Caenorhabditis elegans* neurons to RNA interference. *Neuroreport* **16**, 1995–1999. (doi:10.1097/00001756-200512190-00005)
80. Consortium CEDM. 2012 Large-scale screening for targeted knockouts in the *Caenorhabditis elegans* genome. *G3: Genes Genomes Genetics* **2**, 1415–1425. (doi:10.1534/g3.112.003830)
81. Friedland AE, Tzur YB, Esvelt KM, Colaiácovo MP, Church GM, Calarco JA. 2013 Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat. Methods* **10**, 741–743. (doi:10.1038/nmeth.2532)
82. Dickinson DJ, Ward JD, Reiner DJ, Goldstein B. 2013 Engineering the *Caenorhabditis elegans* genome using Cas9-triggered homologous recombination. *Nat. Methods* **10**, 1028–1034. (doi:10.1038/nmeth.2641)
83. Petersen C, Dirksen P, Schulenburg H. 2015 Why we need more ecology for genetic models such as *C. elegans*. *Trends Genet.* **31**, 120–127. (doi:10.1016/j.tig.2014.12.001)
84. Broderick NA. 2016 Friend, foe or food? Recognition and the role of antimicrobial peptides in gut immunity and *Drosophila*–microbe interactions. *Phil. Trans. R. Soc. B* **371**, 20150295. (doi:10.1098/rstb.2015.0295)
85. Makarova O, Rodriguez-Rojas A, Eravci M, Weise C, Dobson A, Johnston P, Rolff J. 2016 Antimicrobial defence and persistent infection in insects revisited. *Phil. Trans. R. Soc. B* **371**, 20150296. (doi:10.1098/rstb.2015.0296)
86. Marxer M, Vollenweider V, Schmid-Hempel P. 2016 Insect antimicrobial peptides act synergistically to inhibit a trypanosome parasite. *Phil. Trans. R. Soc. B* **371**, 20150302. (doi:10.1098/rstb.2015.0302)
87. Pees B, Yang W, Zárate-Potes A, Schulenburg H, Dierking K. 2015 High innate immune specificity through diversified C-type lectin-like domain proteins in invertebrates. *J. Innate Immun.* **8**, 129–142. (doi:10.1159/000441475)
88. Cash HL, Whitham CV, Behrendt CL, Hooper LV. 2006 Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313**, 1126–1130. (doi:10.1126/science.1127119)
89. Mukherjee S *et al.* 2014 Antibacterial membrane attack by a pore-forming intestinal C-type lectin. *Nature* **505**, 103–107. (doi:10.1038/nature12729)
90. Miltsch SM, Seeberger PH, Lepenies B. 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**, 67–73. (doi:10.1016/j.dci.2014.02.002)
91. Pujol N, Davis PA, Ewbank JJ. 2012 The origin and function of anti-fungal peptides in *C. elegans*: open questions. *Front. Immunol.* **3**, 237. (doi:10.3389/fimmu.2012.00237)
92. Liu C, Liu T, Yuan F, Gu Y. 2010 Isolating endophytic fungi from evergreen plants and determining their antifungal activities. *Afr. J. Microbiol. Res.* **4**, 2243–2248.
93. Anzlovar S, Serra MD, Dermastia M, Menestrina G. 1998 Membrane permeabilizing activity of pathogenesis-related protein linusitin from flax seed. *Mol. Plant-Microbe Interact.* **11**, 610–617. (doi:10.1094/MPMI.1998.11.7.610)
94. Roberts WK, Selitrennikoff CP. 1990 Zeamatin, an antifungal protein from maize with membrane-permeabilizing activity. *J. Gen. Microbiol.* **136**, 1771–1778. (doi:10.1099/00221287-136-9-1771)
95. Yan H, Hancock RE. 2001 Synergistic interactions between mammalian antimicrobial defense peptides. *Antimicrob. Agents Chemother.* **45**, 1558–1560. (doi:10.1128/AAC.45.5.1558-1560.2001)
96. Baeder DY, Yu G, Hozé N, Rolff J, Regoes RR. 2016 Antimicrobial combinations: Bliss independence and Loewe additivity derived from mechanistic multi-hit models. *Phil. Trans. R. Soc. B* **371**, 20150294. (doi:10.1098/rstb.2015.0294)
97. Masson F, Zaidman-Rémy A, Heddi A. 2016 Antimicrobial peptides and cell processes tracking endosymbiont dynamics. *Phil. Trans. R. Soc. B* **371**, 20150298. (doi:10.1098/rstb.2015.0298)
98. Félix M-A, Duveau F. 2012 Population dynamics and habitat sharing of natural populations of *Caenorhabditis elegans* and *C. briggsae*. *BMC Biol.* **10**, 59. (doi:10.1186/1741-7007-10-59)
99. Berg M, Stenuit B, Ho J, Wang A, Parke C, Knight M, Alvarez-Cohen L, Shapira M. In press. Assembly of the *Caenorhabditis elegans* gut microbiota from diverse soil microbial environments. *ISME J.* (doi:10.1038/ismej.2015.253)
100. Fraune S, Anton-Erxleben F, Augustin R, Franzenburg S, Knop M, Schröder K, Willoweit-Ohl D, Bosch TC. 2015 Bacteria–bacteria interactions within the microbiota of the ancestral metazoan Hydra contribute to fungal resistance. *ISME J.* **9**, 1543–1556. (doi:10.1038/ismej.2014.239)
101. Danna P, Urban C, Rahal J, Bellin E. 1991 Role of candida in pathogenesis of antibiotic-associated diarrhoea in elderly inpatients. *Lancet* **337**, 511–514. (doi:10.1016/0140-6736(91)91296-7)
102. Kim Y, Mylonakis E. 2012 *Caenorhabditis elegans* immune conditioning with the probiotic bacterium *Lactobacillus acidophilus* strain NCFM enhances Gram-positive immune responses. *Infect. Immun.* **80**, 2500–2508. (doi:10.1128/IAI.06350-11)
103. Montalvo-Katz S, Huang H, Appel MD, Berg M, Shapira M. 2013 Association with soil bacteria enhances p38-dependent infection resistance in *Caenorhabditis elegans*. *Infect. Immun.* **81**, 514–520. (doi:10.1128/IAI.00653-12)

Discussion

Omics data requires efficient statistical tools to reduce noise and explore the underlying biological functions. In my PhD project I mainly focused on the analysis of transcriptome studies in *C. elegans*. This work describes the development of new approaches for the analysis of this type of data (Chapter I-III), including the inference of underlying biological function (specifically focused on the case study of the innate immune system, Chapter IV-VII). This discussion is separated as three parts: i) DE detection of RNA-Seq; ii) meta-analysis for biological function inference, namely gene set enrichment analysis; iii) innate immunity of *C. elegans*.

1. **RNA-Seq analysis**

RNA-Seq is widely used to study gene expression and inducible responses of model and non-model organisms [1]. The usage of RNA-Seq data requires normalization in order to reduce possible sequencing and biological biases [2, 3]. We propose a new normalization procedure called *qtotal*, which facilities comparison of gene expression across samples via taking into account the overall distribution of reads count per sample (Chapter II). Additional information (e.g, gene length and GC content) might also influence the reads distribution and need to be accommodated in next version [4].

The commonest aim of RNA-Seq studies is to understand inducible biological functions through an analysis of differential gene expression (DE). We show that DE inference should not only rely on fitting the distribution of read counts with probabilistic models, but also on measuring the magnitude of expression difference by absolute counts difference (Chapter I) or fold change (Chapter II). As we show in Chapters I and II, the main question of DE inference for RNA-Seq is to handle the estimation of variance from samples with limited size. Despite the improvements that we proposed, identifying DE at low expression level is a remaining open question, which might be not able to be solved by statistical models and requires specific improvement of RNA-Seq (e.g, targeted

RNA sequencing [5]). Moreover, ABSSeq (Chapter I) and aFold (Chapter II) is currently designed for pairwise comparison, which lacks the ability to handle complex experimental set-up (e.g., time series and multiple treatment) and needs to be extended in future.

2. Meta-analysis

High-throughput molecular technologies usually yield hundreds or thousands of differentially regulated genes or proteins that are not always easy to interpret. Uncovering the underlying organizational principles from such large gene lists requires meta-analysis. We constructed a taxon-specific database by collecting curated high throughput data sets of *C. elegans*, which help interpret new data and uncover underlying biological mechanisms (Chapter III-VII). However, large number of datasets retain redundant information (overlapping with each other) and result in amounts of significances in meta-analysis, which are not easy to interpret. To reduce or summarize the obtained results, an advanced or complex model might be helpful. For instance, a model with Bayesian network will uncover the causality across data sets [6]. Identifying common regulatory elements for the data sets may be also useful to interpret the overlapping and summarize the results from meta-analysis.

3. Innate immunity of *C. elegans*

The immune system that maintains organismal homeostasis and responds against microbe offenses is traditionally divided into two types: innate immunity, which serves as general defense and adaptive immunity that is characterized by the ability to mount specific immune responses according to the pathogen [7]. The nematode *C. elegans* only relies on the innate immune system, which is mainly conserved across animals [8]. Immune response to pathogens reflects as altering gene expression profiling via triggering immune signaling pathway (namely immune related genes). Interestingly, while immune related genes in *C. elegans* to various pathogens are diversity, the promoter of these genes share a common regulator element: the GATAA transcription factor (TF) motif (Chapter VI). It suggests that the conserved immune pathways of *C. elegans*

might crosslink through GATAA bound TFs. However, there are 12 known GATAA-bound TFs (e.g., six members of *elt* family) in *C. elegans* according to Wormbase [9], which function in different tissues and might play a specific role in responding to various pathogens. Interestingly, our study shows that the gene *npr-1* in *C. elegans* exhibits specificity in defense against two pathogens: *P. aeruginosa* strain PA14 and *B. thuringiensis* strain B-18247, suggesting that the innate immune system is also capable of specific responses to pathogens via regulating different sets of genes (Chapter V). In fact, these two pathogens trigger distinct gene expression changes in *C. elegans*, which could be both suppressed by *npr-1* knockout. This indicates that *npr-1* might be involved in different pathways that specifically defend against these two pathogens and could be an interesting topic for a potential future study.

In addition to conserved immune pathways across animals, *C. elegans* also expresses diverse antimicrobial peptides, which have a function in the elimination of pathogens (Chapter VII). The mechanisms by which antimicrobial peptides are activated to clear pathogens rely on signaling pathways which are not yet to be uncovered. Moreover, antimicrobial peptides potentially display high redundancy through multigene families, which hinders the investigation of their roles in immune response. Uncovering the mechanism of antimicrobial peptides in elimination of pathogens requires a high throughput survey assay and/or meta-analysis of large data sets. Moreover, the immune response might differ at transcript level and protein level (Chapter IV), suggesting that hierarchical information is required for understanding and identifying these immune effectors. Furthermore, a complete omics study of immune response including data at transcript level, protein level and metabolism level might reveal new insights into host-pathogen interaction as we have shown for *B. thuringiensis* and *C. elegans* (i.e., linkage between damage and insulin like signaling pathway via AMPK, Chapter IV).

References – Discussion

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews genetics* 2009, **10**(1):57-63.
2. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome biology* 2010, **11**(3):1.
3. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results**. *Genome biology* 2010, **11**(12):1.
4. Li P, Piao Y, Shon HS, Ryu KH: **Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data**. *BMC bioinformatics* 2015, **16**(1):347.
5. Clark MB, Mercer TR, Bussotti G, Leonardi T, Haynes KR, Crawford J, Brunck ME, Lê Cao K-A, Thomas GP, Chen WY: **Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing**. *Nature methods* 2015, **12**(4):339-342.
6. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nature genetics* 2003, **34**(2):166-176.
7. Medzhitov R, Janeway CA: **Innate immunity: impact on the adaptive immune response**. *Current opinion in immunology* 1997, **9**(1):4-9.
8. Pees B, Yang W, Zárate-Potes A, Schulenburg H, Dierking K: **High innate immune specificity through diversified C-type lectin-like domain proteins in invertebrates**. *Journal of innate immunity* 2016, **8**(2):129-142.
9. Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, Done J, Down T, Gao S, Grove C: **WormBase 2016: expanding to enable helminth genomic research**. *Nucleic acids research* 2016, **44**(D1):D774-D780.

List of Abbreviations

DE: differential expression; CV: coefficient of variation; SD: standard deviation; NB: negative binomial; AUC: area under curve; ROC: receiver operating characteristic; FC: fold change; logFC: log2 of fold change; FDR: false discovery rate; FPR: false positive rate; TPR: true positive rate; RNA-Seq: (high-throughput) sequencing of RNA; TF: transcription factor.

Acknowledgement

I am grateful to the Rechenzentrum of the University of Kiel for access to the Linux cluster and technical support. The study was funded by the German Science Foundation within the priority program SPP1399 on host-parasite coevolution, individual grants SCHU 1415/8 and SCHU1415/9.

Danksagung

Many thanks for the supporting from all members in evoecogen group. Especially thanks to Charlotte Rafaluk, Carsten Makus and Andrei Papkou for helping me out in the beginning of my study and my life in Germany, to Philipp Dirksen, Niels Mahrt and Alejandra Zárate-Potes for helping me find the apartment and other supporting. Great thanks to Hinrich Schulenburg for offering me the opportunity to study in the group and supervise me, and to Katja Dierking for supervising and supporting me. Particular thanks to my wife, Li Fan, who always supports and loves me.

Curriculum Vitae

Personal Information

Name: Wentao Yang

Date of Birth: January 1st, 1984

Place of Birth: Hubei, China

Nationality: Chinese

Education

Since April 2013: PhD student at the Department of Evolutionary Ecology and Genetics Christian Albrechts University, Kiel, Germany

September 2008 – July 2010: Master in Genetics, Shanghai Jiaotong University, China

September 2003 - July 2007: Studies in Biology, Lanzhou University, China