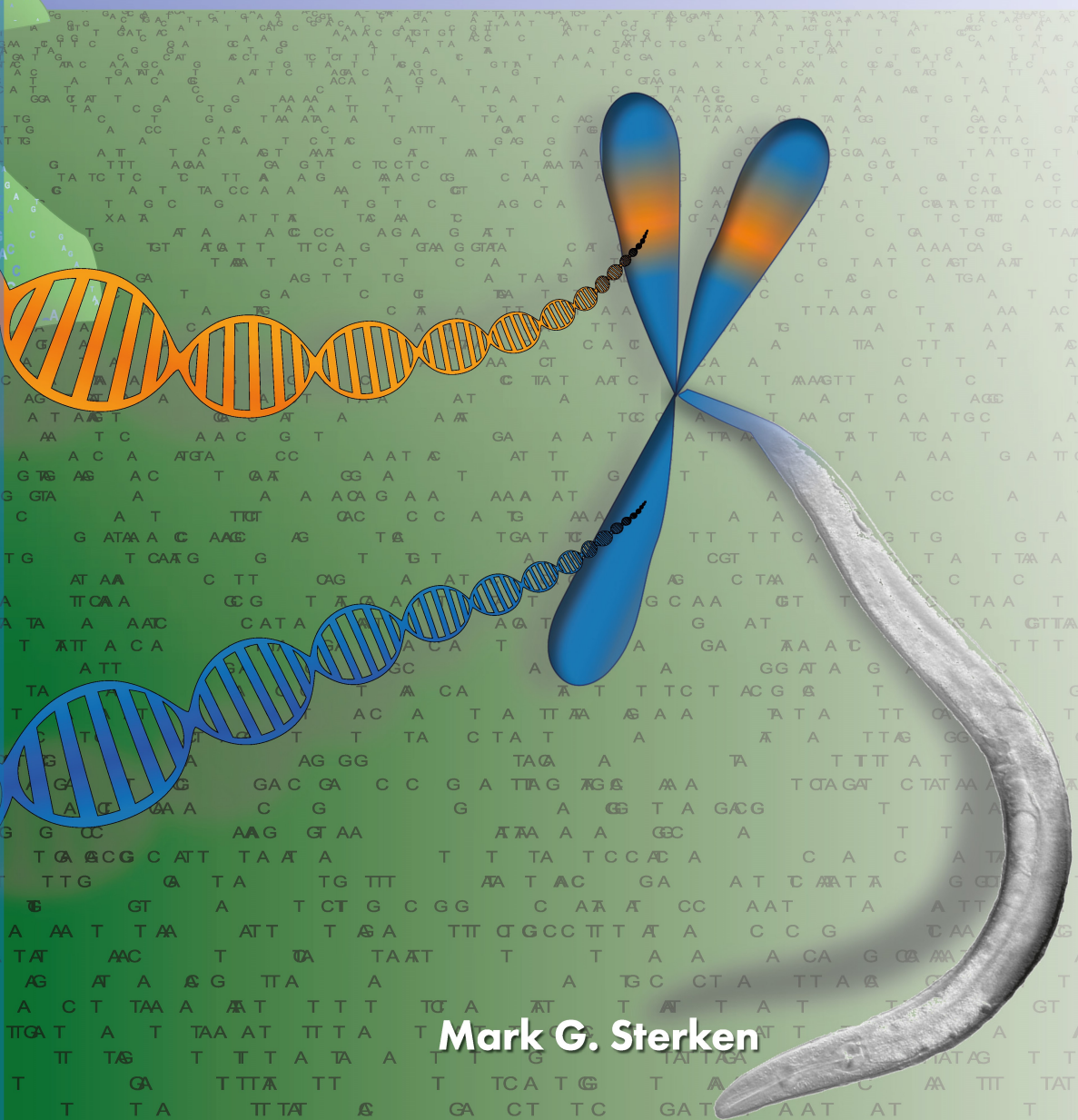
Building towards a multi-dimensional genetic architecture in *Caenorhabditis elegans*



Mark G. Sterken

Building towards a multi-dimensional genetic architecture in *Caenorhabditis elegans*

Mark G. Sterken

Thesis committee

Promotor
Prof. Dr J.E. Kammenga
Personal chair at the Laboratory of Nematology
Wageningen University

Co-promotors
Prof. Dr J. Bakker
Professor of Nematology
Wageningen University

Dr G.P. Pijlman
Associate professor, Laboratory of Virology
Wageningen University

Other members
Prof. Dr B.J. Zwaan, Wageningen University
Prof. Dr A. Hajnal, University of Zurich, Switzerland
Prof. Dr H. Schulenburg, Christian-Albrechts-Universität zu Kiel, Germany
Dr R.H. Houtkooper, Academic Medical Center, Amsterdam

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC)

Building towards a multi-dimensional genetic architecture in *Caenorhabditis elegans*

Mark G. Sterken

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 19 October 2016
at 1:30 p.m. in the Aula.

Mark G. Sterken
Building towards a multi-dimensional genetic architecture in *Caenorhabditis elegans*
168 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

ISBN 978-94-6257-869-2
DOI 10.18174/386549

Table of contents

Chapter 1

General introduction

Mark G. Sterken

The nematode *C. elegans*

The free-living bacteriovorous nematode *Caenorhabditis elegans* is a versatile and widely used model organism. It made its claim to fame by providing geneticists with a simple metazoan organism [1, 2]. Major advances in biology have been made using this organism, for example in understanding apoptosis, small RNAs, and aging [3-5]. *C. elegans* boosted research leading to 3 Nobel prizes and over 10,000 scientific publications, in a community encompassing over 2,000 researchers. Obviously, this animal forms the cornerstone of a vibrant research community.

The success of *C. elegans* relates back to its simple biology and small genome (**Figure 1A**), but also to some major technological advances made throughout its career in the laboratory. The biological aspects that make *C. elegans* interesting include: a short life-cycle (only 3.5 days at 20°C), clonal propagation (the most abundant state is that as a self-fertilizing hermaphrodite), possibility for out-crossing (males do occur, albeit not often, <1% of the offspring under normal conditions), and abundant offspring (a self-fertilizing hermaphrodite can produce ~300 offspring) [1]. These biological aspects make this animal a very attractive genetic model organism. The technical advances made by researchers working on this species greatly add to its appeal. The major advantages of this organism include: the ability for cryopreservation [1], a well annotated genome [6], and abundantly available tools (*e.g.* RNAi libraries and protocols) and mutant/transgenic strains [7]. Many of these tools, information, and protocols are accessible via on-line resources, such as WormBook (www.wormbook.org), WormBase (www.wormbase.org) and WormAtlas (www.wormatals.org) [8, 9].

Importantly, these advances have mainly been made in a single strain, the N2 strain, isolated in Brisol (UK) in 1950. While the N2 strain is one of the strengths of the *C. elegans* field, it is also its weakness. Basing all findings on a single strain makes comparative work easier, however, it severely limits the outlook on 'natural responses' in this animal. The actual implications of N2 on *C. elegans* research will be discussed in-depth in **Chapter 2** of this thesis.

Wild isolates and the ecology of *C. elegans*

The advances in our understanding of the genetics and biology of the N2 strain stand in stark contrast with the knowledge gathered on wild strains. Until a decade ago, knowledge on the ecology and whereabouts of this nematode were very limited. More recent investigations provide a clearer picture of its natural habitat and life-cycle (reviewed by [10, 11]) . It is now clear that this nematode can be isolated from rotten material, such as petrifying plant-stems, compost heaps, and rotten fruits [12-14]. Furthermore, it has been found associated with small invertebrates such as isopods, millipedes, and snails [14]. Still, many researchers working on *C. elegans* still refer to it as a 'soil dwelling nematode'.

The locations where *C. elegans* can be isolated give some insights in the natural lifecycle of this animal. It is thought that *C. elegans* populations go through rapid expansions and contractions. As soon as a nematode colonizes a new substrate (probably in a long living stage called 'dauer'), it rapidly procreates, producing many offspring that feed on the bacteria present on the substrate. Subsequently, as the food is depleted, new offspring will arrest their development in the dauer-stage and either reach a new substrate and start a novel population, or die while searching [10, 11]. The association with small invertebrates could be due to phoresis, transporting the nematodes to new colonisable substrates [14, 15]. These 'boom and burst' cycles take place all throughout summer and autumn in Northern Europe [12]. It still remains unknown where and how this animal overwinters.

The current samplings of *C. elegans* show that, in general, there is little genetic variation between isolates. Remarkably, little genetic population structure was found related to geographic site of isolation [16]. Still, intensive sampling at two sites shows that variation linked to site can be found [17]. The global pattern of low variation is contributed to a selective sweep that took place in recent history (~100 years in the past) [16], while the local patterns could point towards selection [17]. However, considering the global variation, local variation could also be attributed to a founder effect and genetic drift. The variations that are observed (*e.g.* single nucleotide polymorphisms, insertions, and deletions) are mostly located on the distal ends of the chromosomes. This pattern is observed consistently over *C. elegans* isolates and correlate with the recombination frequencies [16-18]. In combination, these polymorphisms and the recombination pattern can be the basis of an ideal genotyping strategy (see the marker design in **Chapter 3**).

Through sampling of wild *C. elegans* strains, co-occurring organisms and natural pathogens of this animal were discovered. This led to the discovery of microsporidia, fungi, bacteria, and a virus that infect *C. elegans* [15, 19-21]. The discovery of the Orsay virus (OrV) was a major leap in the possibilities for this model system and is now opening new venues to study host-virus interactions [19]. In my thesis, I used OrV to study the influence of natural genetic variation on development of the infection in the host (**Chapter 4** and **5**), and to identify genes involved in viral susceptibility (**Chapter 5**).

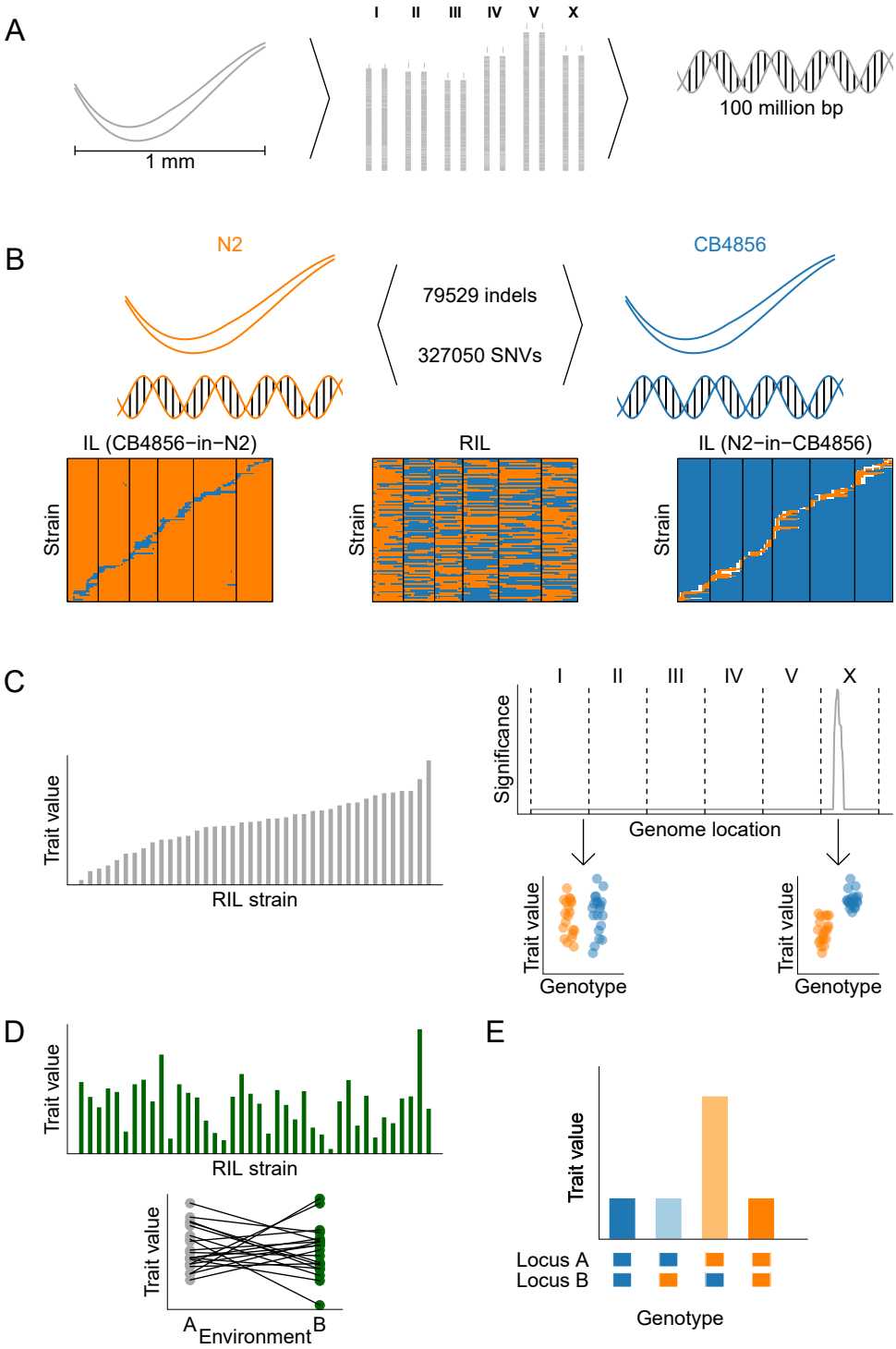
Relating genotypic variation to trait variation

Most of the quantitative genetics studies in *C. elegans* in the last decade have been conducted in inbred populations constructed from crosses between N2 and CB4856 (**Figure 1B**). The CB4856 strain was isolated in Hawaii and is one of the most genotypically divergent isolates compared to N2 [16, 22]. Several inbred populations were created with these strains, including two recombinant inbred line populations (RILs) and two genome-wide introgression line (IL) populations [18, 23, 24; **Chapter 3**]. Recombinant inbred lines are genetic mosaics that contain

an equal contribution of each parental strain. In contrast, introgression lines contain a single locus of one strain introgressed in a background of another strain. Both types of populations can be generated by crossing hermaphrodites with males and nearly homozygous strains can be generated by selecting offspring of self-fertilizing hermaphrodites for several generations. For more information on the construction of inbred panels, see **Chapter 3**, which describes the construction of the second genome-wide IL population in *C. elegans*.

Inbred panels can be used to relate genotypic variation to trait variation by quantitative trait locus (QTL) mapping (for an explanation, see **Figure 1C**). There are several steps that can be taken in such an analysis, some of which are discussed here. Without any genotypic information about the inbred strains, repeated measurements of a trait can be used to infer the amount of trait variation that can be attributed to genotypic variation. The ratio of genotype induced variation to total trait variation is called the heritability. The heritability gives an upper bound for the trait variation that can be explained by QTL. Heritability is environment dependent, that means that the heritability of the same trait can have a different value if measured in another environment. If additional genotypic information is available for the inbred panel, such as a genetic map of the population, QTL can be mapped. A QTL contains information about the location of the loci that is associated with the trait and the effect of that regulation. Any quantifiable trait can be investigated using this method, including the regulation of gene expression (genetical genomics) [25]. In the case of gene expression, expression QTL (eQTL) are found. Together, the information on the location of the QTL and the location of the regulated transcript (gene) gives rise to a wealth of trait-regulation information (see **Chapters 3**, **6**, and **7**).

Figure 1: The nematode *Caenorhabditis elegans* and the various inbred populations and analysis techniques used in this thesis. (**A**) *C. elegans* is a 1 mm long nematode, with a genome consisting of 5 autosomes and 1 sexual chromosome (X) [1]. The genome is ~100 million base pairs in length and contains ~20000 genes [6]. (**B**) In this thesis the Bristol N2, the Hawaii CB4856 strain, and derived IL and RIL populations are used. N2 and CB4856 are among the most genetically distant *C. elegans* isolates and differ in many single nucleotide variatiants (SNVs), insertions and deletions (indels), which are mostly occurring on the chromosome arms (not shown) [16, 18, 22]. From these strains several derived populations were constructed: a CB4856-in-N2 IL population [24], two RIL populations [18, 23], and an N2-in-CB4856 IL population [**Chapter 3**]. The figures show (a selection of) each population type. On the x-axis the location on the genome is plotted (from the start of chromosome I to the end of chromosome X) and on the y-axis strains are plotted. The colours indicate the genotype at a specific genomic location: N2 in orange and CB4856 in blue. (**C**) Quantitative trait locus (QTL) mapping. If a quantifiable trait is measured in a RIL panel, and there are genetic differences influencing the trait value, a pattern as shown on the left is generally seen. The trait values of each RIL strain, combined with the genetic map of each strain, can be used to associate the genotype at each genome location with the trait value (right). If the genome location does not explain the trait variation, there is no significant association with the trait value. However, sometimes a genome location can be found that explains the trait variation, and a QTL is found (the peak at chromosome X). Thus, a QTL indicates a location where natural variation affects the trait value. (**D**) In different environments trait values can change (top). These changes can have a genetic component, leading to different reaction norms if environment A and environment B are compared within strains (bottom). (**E**) Yet another source of trait variation are loci-loci interactions. On the x-axis the genotype is shown (N2 in orange and CB4856 in blue) and on the y-axis the trait value. In case of an N2 genotype on locus A and a CB4856 genotype on locus B, the trait value becomes higher than predicted from the separate loci; there is a genetic interaction.


A
I
II
III
IV
V
X
1 mm
100 million bp
B
N2
CB4856
79529 indels
327050 SNVs
IL (CB4856−in−N2)
RIL
IL (N2−in−CB4856)
Strain
C
Trait value
RIL strain
Significance
Genome location
Genotype
D
E
Locus A
Locus B
A
Environment
B

In nature, trait architectures range from very simple (monogenic) to very complex (polygenic) [26, 27]. The expected architectures can differ per trait and organism and it remains a point of consideration what a relevant effect size is; how relevant is a QTL that explains 1% of the trait variation? Although QTL mapping is done with standing genetic variation in a population, ultimately QTL can have an impact on the allele frequencies within a population. This impact can be in such an extent that one of the alleles is removed from the population. In this case, the QTL confers a selective advantage. In order to affect selection, the effect of the QTL can be very small, even in such an extent that it might explain less than 1% of trait variation. However, QTL mapping in such resolution requires tremendous efforts in terms of inbred strains required and level of replication in the experiments [27, 28]. Therefore, most QTL studies focus on traits with a few strong QTL (explaining a considerable amount of the variation). Examples can be found in **Chapter 5** and **6**, the ultimate focus is to find the allele(s) that underlie the QTL. In the end, these studies provide information about alleles and the molecular mechanism resulting in trait variation.

Adding dimensionality by perturbations

One way of expanding our understanding of trait architecture comes from perturbing the system by influencing the environment (**Figure 1D**). Such perturbations may expose trait variation that otherwise remains hidden [23, 29-31]. In general, this perturbation is environmental in nature, but can be extended to genetic perturbations using knock-out mutations in different genetic backgrounds [32, 33]. Examples of studies with environmental perturbations can be found in **Chapter 6**, in which a RIL and an IL population are exposed to two different treatments (control and heat-shock), after which the transcriptome was quantified.

While additional variation is uncovered, there are also patterns that remain similar over environments. One of the reliably replicated gene expression traits are *cis*-eQTL (applied in **Chapter 3**, **6**, and **7**). These eQTL have a genetic regulator near the gene that is affected. This often means that these genes carry a polymorphism affecting their expression (*e.g.* gene deletion or promotor polymorphisms) [22]. Another type of eQTL is regulated *in trans*, meaning the regulator is not near the affected gene. In all organisms investigated by genetical genomics so-far, *trans*-eQTL form *trans*-bands, where the eQTL of many genes map to [23, 29, 30, 34-38]. *Trans*-bands often reflect a specific response. For example, a *trans*-band observed by Rockman et al. in a genetical genomics study in *C. elegans* [36], is caused by a starvation response due to a polymorphism in the gene *npr-1* [39].

Expanding the dimensionality to multiple loci

It is widely recognized that trait variation in natural populations is complex, and in many cases governed by many loci (**Figure 1E**) [26, 28, 40]. Unfortunately, it is experimentally and statistically challenging to study (highly) polygenic traits [26]. There are some approaches that can make headway, such as extremely large RIL populations [27, 28], IL populations (see **Chapter 3**) [24, 41-43], and specific tests for two interacting loci (see **Chapter 7**) [44]. There are many developments in computational approaches for epistasis, including epistasis in eQTL [45-47]. However, most of the approaches on interactions in eQTL require assumptions on involved loci in order to obtain statistically significant findings.

In light of the complexity expected for trait architectures, it is important to get a grip on how these complex architectures function. Therefore, we mapped the loci-loci interactions in a published dataset of RILs between N2 and CB4856 (**Chapter 7**) [36]. This was followed-up by constructing populations containing two introgression in an N2 background to verify the findings from the eQTL mapping. This study is yet another indication that epistasis is pervasive and likely affects many traits.

Scope of this thesis

In my thesis, I study the contribution of genetic and environmental factors on trait variation using the model system *C. elegans*. The goal is to understand genetic architectures and the role of environmental effects on these architectures. Together, insights in this multi-dimensional relation (genotype by genotype and genotype and environment) in *C. elegans* can translate to expectations for less tractable systems.

References

1. **Brenner, S.**, *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94**.
2. **Ankeny, R.A.**, *The natural history of Caenorhabditis elegans research.* **Nat Rev Genet, 2001. 2(6): p. 474-9**.
3. **Ellis, R.E., J.Y. Yuan, and H.R. Horvitz**, *Mechanisms and functions of cell death.* **Annu Rev Cell Biol, 1991. 7: p. 663-98**.
4. **Fire, A., et al.**, *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* **Nature, 1998. 391(6669): p. 806-11**.
5. **Kenyon, C., et al.**, *A C-Elegans Mutant That Lives Twice as Long as Wild-Type.* **Nature, 1993. 366(6454): p. 461-464**.
6. **Consortium, C.e.S.**, *Genome sequence of the nematode C. elegans: a platform for investigating biology.* **Science, 1998. 282(5396): p. 2012-8**.
7. **Kamath, R.S., et al.**, *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* **Nature, 2003. 421(6920): p. 231-7**.
8. **Stein, L., et al.**, *WormBase: network access to the genome and biology of Caenorhabditis elegans.* **Nucleic Acids Research, 2001. 29(1): p. 82-86**.
9. **Altun, Z.F., Herndon, L.A., Crocker, C., Lints, R. and Hall, D.H**. *Wormatlas*. **2002-2016**.
10. **Petersen, C., P. Dirksen, and H. Schulenburg**, *Why we need more ecology for genetic models such as C. elegans.* **Trends Genet, 2015**.
11. **Frezal, L. and M.A. Felix**, *C. elegans outside the Petri dish.* **Elife, 2015. 4**.
12. **Petersen, C., et al.**, *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: p. 4**.
13. **Kiontke, K.C., et al.**, *A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits.* **BMC Evol Biol, 2011. 11: p. 339**.
14. **Felix, M.A. and C. Braendle**, *The natural history of Caenorhabditis elegans.* **Curr Biol, 2010. 20(22): p. R965-9.**
15. **Felix, M.A. and F. Duveau**, *Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae.* **BMC Biol, 2012. 10: p. 59**.
16. **Andersen, E.C., et al.,** *Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nat Genet, 2012. 44(3): p. 285-90**.
17. **Volkers, R.J., et al.,** *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93**.
18. **Rockman, M.V. and L. Kruglyak**, *Recombinational landscape and population genomics of Caenorhabditis elegans.* **PLoS Genet, 2009. 5(3): p. e1000419**.
19. **Felix, M.A., et al.**, *Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses.* **PLoS Biol, 2011. 9(1): p. e1000586**.
20. **Troemel, E.R., et al.**, *Microsporidia are natural intracellular parasites of the nematode Caenorhabditis elegans.* **PLoS Biol, 2008. 6(12): p. 2736-52**.
21. **Maguire, S.M., et al.**, *The C. elegans touch response facilitates escape from predacious fungi.* **Curr Biol, 2011. 21(15): p. 1326-30**.
22. **Thompson, O.A., et al.**, *Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856.* **Genetics, 2015. 200(3): p. 975-89**.
23. **Li, Y., et al.**, *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222**.
24. **Doroszuk, A., et al.**, *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110**.
25. **Jansen, R.C. and J.P. Nap**, *Genetical genomics: the added value from segregation.* **Trends Genet, 2001. 17(7): p. 388-91**.
26. **Mackay, T.F.,** *Epistasis and quantitative traits: using model organisms to study gene-gene interactions.* **Nat Rev Genet, 2014. 15(1): p. 22-33**.
27. **Bloom, J.S., et al.**, *Finding the sources of missing heritability in a yeast cross.* **Nature, 2013. 494(7436): p. 234-7**.

28. **Bloom, J.S., et al.**, *Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.* **Nat Commun, 2015. 6: p. 8712**.
29. **Keurentjes, J.J., et al.**, *Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.* **Proc Natl Acad Sci U S A, 2007. 104(5): p. 1708-13**.
30. **Vinuela, A., et al.**, *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37**.
31. **Francesconi, M. and B. Lehner**, *The effects of genetic variation on gene expression dynamics during development.* **Nature, 2014. 505(7482): p. 208-11**.
32. **Duveau, F. and M.A. Felix**, *Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans.* **PLoS Biol, 2012. 10(1): p. e1001230**.
33. **Schmid, T., et al.**, *Systemic Regulation of RAS/MAPK Signaling by the Serotonin Metabolite 5-HIAA.* **Plos Genetics, 2015. 11(5)**.
34. **Brem, R.B., et al.**, *Genetic interactions between polymorphisms that affect gene expression in yeast.* **Nature, 2005. 436(7051): p. 701-3**.
35. **Brem, R.B. and L. Kruglyak**, *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* **Proc Natl Acad Sci U S A, 2005. 102(5): p. 1572-7**.
36. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak**, *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6**.
37. **Kelly, S.A., et al.**, *Functional genomic architecture of predisposition to voluntary exercise in mice: expression QTL in the brain.* **Genetics, 2012. 191(2): p. 643-54**.
38. **Swanson-Wagner, R.A., et al.**, *Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids.* **Science, 2009. 326(5956): p. 1118-20**.
39. **Andersen, E.C., et al.**, *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156**.
40. **Rockman, M.V.**, *The QTN program and the alleles that matter for evolution: all that's gold does not glitter.* **Evolution, 2012. 66(1): p. 1-17**.
41. **Keurentjes, J.J., et al.**, *Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.* **Genetics, 2007. 175(2): p. 891-905**.
42. **Green, J.W., et al.**, *Highly Polygenic Variation in Environmental Perception Determines Dauer Larvae Formation in Growing Populations of Caenorhabditis elegans.* **PLoS One, 2014. 9(11): p. e112830**.
43. **Green, J.W., et al.**, *Genetic mapping of variation in dauer larvae development in growing populations of Caenorhabditis elegans.* **Heredity (Edinb), 2013. 111(4): p. 306-13**.
44. **Gaertner, B.E., et al.**, *More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in Caenorhabditis elegans.* **Genetics, 2012. 192(4): p. 1533-42**.
45. **Zhang, Y., et al.**, *Bayesian models for detecting epistatic interactions from genetic data.* **Ann Hum Genet, 2011. 75(1): p. 183-93**.
46. **Zhang, W., et al.,** *A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules.* **PLoS Comput Biol, 2010. 6(1): p. e1000642**.
47. **Hu, Z., et al.**, *Genomic value prediction for quantitative traits under the epistatic model.* **BMC Genet, 2011. 12: p. 15**.

Chapter 2

The laboratory domestication of *Caenorhabditis elegans*

Mark G. Sterken,
L. Basten Snoek,
Jan E. Kammenga,
Erik C. Andersen

This chapter has been published in a modified form:

Sterken, M.G.; Snoek, L.B.; Kammenga, J.E.; Andersen, E.C. (2015). The laboratory domestication of Caenorhabditis elegans. *Trends in genetics* 31(5): 224-231

Abstract

Model organisms are of great importance to understanding basic biology and to making advances in biomedical research. However, the influence of laboratory cultivation on these organisms is underappreciated, especially how that environment can affect research outcomes. Recent experiments led to insights into how the widely used laboratory reference strain of the nematode *Caenorhabditis elegans* compares to natural strains. Here, we describe potential selective pressures that led to fixation of laboratory-derived alleles for the genes *npr-1*, *glb-5*, and *nath-10*. These alleles influence a large number of traits, resulting in behaviors that affect experimental interpretations. Furthermore, strong phenotypic effects caused by these laboratory-derived alleles hinder the discovery of natural alleles. Additionally, we highlight strategies to reduce the influences of laboratory-derived alleles and to harness the full power of *C. elegans*.

Model organisms pave the road to biological discovery

Sustained progress in the biological sciences is facilitated by discoveries using organisms that are amenable to laboratory investigation. They have large numbers of offspring, are small, and easy to maintain. Many different species have these attributes. For example, the single-cell eukaryote *Saccharomyces cerevisiae* (baker's yeast) is immensely powerful as a genetic model organism for conserved cellular processes [1] and for quantitative genetics using large populations [2]. The fruit fly *Drosophila melanogaster* contributed extensively to our understanding of signal-transduction pathways and developmental patterning [3]. The free-living nematode *Caenorhabditis elegans* is a widely used model organism in studies of development [4, 5], mechanistic neurobiology [6], aging [7], and small RNAs [8, 9]. When the results from experimental studies of model organisms are tabulated, it is obvious that they facilitated much of what we know about conserved biological processes.

Quantitative geneticists often use tractable model organisms to identify loci and (sometimes) genetic variants that influence phenotypic differences among populations. To elucidate the underlying genetic basis of complex traits, recombinant offspring are generated and their traits measured. Organisms that give rise to large (preferably clonal) populations and are easy to experimentally manipulate enable these approaches. These attributes make *S. cerevisiae* the most powerful eukaryotic organism for quantitative genetics [2, 10]. However, as a metazoan genetic system, *C. elegans* is unmatched [11]. It has an extremely rapid life cycle (3.5 days at 20°C), produces 200-300 offspring per hermaphrodite individual, possesses a small and well-annotated genome, can be cryopreserved, and transgenic strains are easily obtained. Wild strains isolated from nature can be phenotyped and genotyped to perform genome-wide association studies (GWAS, see Glossary) [12-15]. This combination of studies on natural allelic variation paired with analyses of mutations using the laboratory strain offers a powerful approach to broaden our understanding of how genetic background contributes to phenotype. However, characterization of the behaviors and genomes of wild *C. elegans* strains led to suspicions about laboratory adaptation in the widely used N2 laboratory strain [16]. Indeed, wild-type strains used in other model organisms have laboratory-derived variants that result in large pleiotropic effects, including cell clumping in *S. cerevisiae* [17, 18] and plant growth in *A. thaliana* [19].

Here, we review documented examples of *C. elegans* laboratory-derived alleles in the commonly used Bristol (or N2) strain and the effects on its phenotype. We first describe how the laboratory strain N2 is different from wild strains of *C. elegans* and discuss the laboratory history of this nematode to gain insight into genetic bottlenecks and possible laboratory selection. Then, we discuss three known laboratory-derived alleles and their effects on *C. elegans* biology, including implications for the interpretation of observations that can confound experimental outcomes.

N2 is distinct from all wild strains of *C. elegans*

Since its introduction to the research community by Sydney Brenner in 1974 [20], the Bristol (or N2) strain has been used in many laboratories and became the canonical wild-type strain (for more history, see **Box 1**). From the Laboratory of Molecular Biology in Cambridge, N2 spread across the world via trainees from the Brenner laboratory, resulting in massive clonal amplification of N2 around the world. However, before its dissemination and cryogenic preservation, the N2 strain was propagated for many generations leading to the accumulation and selection of random mutations. We do not know how often the strain was transferred to new cultures during this early propagation. Conservatively, the strain could have been passaged every two months. At the other extreme, the strain could have been passaged every four days. Therefore, the strain underwent approximately 300 to 2000 generations from 1951 to 1969 (**Box 1**). Given the germline mutation rate of $2.7x10^{-9}$ mutations per site per generation [21], up to a thousand neutral mutations could have accumulated before cryogenic preservation. Furthermore, after dispersal of this strain around the world, additional genetic differences between N2 strains from different laboratories arose, causing differences in the phenotypes of these standard wild-type strains [22, 23].

Box 1: *C. elegans*: the journey from nature to the bench

Most *C. elegans* research laboratories use the strain named N2, which was collected from mushroom compost collected in Bristol, England in 1951. Like most model organisms, the journey from nature to the laboratory was circuitous (**Figure 1**). The compost was collected by L.N. Staniland, who brought the sample to a short course on plant nematology organized by the British Ministry of Agriculture and Fisheries [16]. From this sample, Bristol *C. elegans* was isolated by Warwick Nicholas who cultured the animal first on petri dishes containing nutrient agar with bacterial contaminants as food [16]. Later, Warwick Nicholas developed axenic liquid cultures from the nutrient agar cultures, as these required less frequent sub-culturing [65]. In 1957 the nematodes were transported to the laboratory of Ellsworth Dougherty at the Kaiser Foundation Research Institute in Richmond, California in a liquid axenic culture [16]. In the Dougherty laboratory, Eder Hansen cultured the nematodes. Two types of cultures were established: nutrient agar slants seeded with *E. coli* in test tubes and liquid axenic culture based on liver extract [64].

Concurrently, Sydney Brenner sought an organism suitable for neurobiology research [67]. He corresponded with Ellsworth Dougherty and even isolated nematodes from his own garden [68]. This nematode culture was called the N1 strain. Sydney Brenner requested the Bristol strain from Ellsworth Dougherty, and it was sent in 1963 [20, 24, 67]. In the Brenner laboratory, the liquid axenic culture was transferred to agar plates containing *E. coli*. After several passages of a population containing both males and hermaphrodites, a single hermaphrodite was selected. This strain, which was used for all subsequent work, was called N2 [68]. The populations were kept in culture on *E. coli* monoxenic agar plates, and the hermaphrodite strain was eventually frozen in 1969 by John Sulston [69].

The cultivation history of the Bristol N2 strain provides only a few opportunities to identify mutations that accumulated during the early culturing period (**Box 1; Figure 1**). Clues could come from a strain that diverged from N2 sometime before 1963 while in the Dougherty laboratory [24], up to 12 years after initial isolation from nature. This strain was mislabeled

as *C. briggsae*, a mistake that was corrected later [25]. In 1995 and 2009, hermaphrodites were removed from axenic culture and frozen as the LSJ1 and LSJ2 strains [25, 26], respectively. Comparison by sequencing revealed an estimate of approximately 100 accumulated variants in N2 [25]. However, no strains are currently known that diverged from N2 before the LSJ1 and LSJ2 strains diverged. Therefore, it is impossible to identify the mutations that accumulated in the initial decade after isolation.

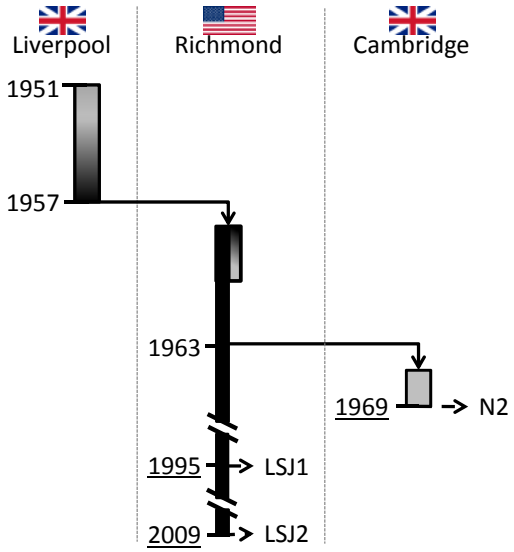


Figure 1:**The history of the Bristol (N2) lineage**. Monoxenic (gray) or axenic (black) cultures are denoted by colored boxes. The gradients indicate uncertainty in when the culture type was switched. The dates show the year of isolation or when the strain was moved to another laboratory. Underscored dates mark the dates of cryogenic preservation. In 1951, Bristol was isolated by L.N. Staniland and Warwick Nicholas. The strain was kept in Liverpool first as a monoxenic culture and later as an axenic culture [16]. In 1957, it was shipped to the Kaiser Foundation Research Institute in Richmond, California. During early laboratory propagation, both axenic and monoxenic cultures were maintained. Later, the monoxenic culture was discontinued and the axenic culture continued [24]. The LSJ1/LSJ2 strains originate from this axenic culture [16, 26], as does N2. It is unclear when exactly the LSJ1/LSJ2 lineage split from N2 [25]. In 1963, Brenner received an axenic culture containing *C. elegans* [24], which was cryogenically preserved by John Sulston in 1969 [69]. The LSJ1 strain was cryogenically preserved in 1995 [16, 26] and the LSJ2 strain in 2009 [16].

From analysis of the genotypes and phenotypes of wild strains, we understand a great deal about variation in nature. Notably, many *C. elegans* strains reported to be isolated from nature were contaminated by the N2 strain (**Table 1**) [16]. Initial characterizations of natural phenotypic variation were confounded by these contaminated strains [27, 28]. Fortunately, recent sampling and genotyping of true wild strains have made it possible to study natural variation in *C. elegans* [12, 29, 30]. The genomes of most wild *C. elegans* strains isolated from nature are highly related, sharing nearly two thirds of the genome. This high degree of sharing is likely the effect of advantageous alleles that swept through the population reducing linked variation [12] and background selection that eliminates variation linked to deleterious alleles [31]. However, we

still cannot identify all of the alleles that accumulated during laboratory propagation of the N2 strain even with these genotype data. The N2 genome contains private alleles found in the strain when isolated from nature along with mutations that accumulated during laboratory propagation. Together, this mix of alleles makes it impossible to identify laboratory-derived alleles from sequence information alone.

Table 1: A large number of "wild" *C. elegans* strains are actually mislabeled N2 strains or recombinant strains derived from N2.

Strain	Genotype*	Ref.
CB3191	N2	[16, 71]
CB3192	N2	[16, 71]
CB3193	N2	[16, 71]
CB3194	N2	[16, 71]
CB3195	N2	[16, 71]
CB4507	N2	[16]
CB4555	N2xCB4851 recombinant	[16]
DH424	N2xCB4851 recombinant	[16]
DR1349	N2xCB4851 recombinant	[16]
PX176	N2	[16]
TR388	N2	[16, 71]
TR389	N2	[16, 71]
TR403	N2xCB4851 recombinant	[16]

*: All strains with the N2 genotype have N2 markers at 1,453 of 1,454 markers spread throughout the genome. The N2xCB4851 recombinant strains are largely or completely N2 for chromosomes I, II, III, and X and CB4851 for chromosomes IV and V. Strain TR389 has N2 for 1,453 markers but harbors the CB4856-like *glb-5* deletion allele.

Selective pressures: nature versus laboratory

To understand how laboratory conditions could influence *C. elegans*, we need to know more about its natural habitat and ecology. Although progress has been made in recent years, the ecology of *C. elegans* is still largely unknown [15, 32, 33]. Despite frequent use of the statement in the literature, it is unlikely that *C. elegans* is a soil nematode. Soil samples harbor *C. elegans* only when in close proximity to rotting vegetation or fruit [15, 34] and recent successful sampling suggests that its natural habitat is rotting material. Wild strains were successfully isolated from rotting hogweed [15, 29], rotting fruits [15, 29, 32, 34], and compost [29, 32]. Additionally, strains have been isolated from 'carrier' species, such as snails or terrestrial isopods [15]. Current observations indicate that *C. elegans* occupies short-lived microbiota-rich habitats. In this niche, it establishes a population quickly and is thought to compete for bacterial food with other species [15, 32]. When food is limiting and population density is high, *C. elegans* enters a long-lived alternative larval stage called dauer. These dauers likely endure periods without food while dispersing to new habitats [35]. By contrast, laboratory cultivation provides a much more constant environment (**Box 2**).

Box 2: Living conditions of C. *elegans* in the laboratory

The life cycle of *C. elegans* consists of an embryonic stage, followed by four larval stages (L1-L4) and an adult stage. The N2 strain completes one generation every 3.5 days at 20ºC. Alternatively, *C. elegans* can enter a long-term survival stage (dauer) as an alternative to the standard L3 larval stage [66].

Axenic culture

Axenic cultures do not contain other organisms as a food source and can be chemically defined or contain extracts of organic material (*e.g.* liver). Such cultures can be either in a solid state (*e.g.* nutrient agar) or in liquid.

Nowadays, axenic cultures are not often used for keeping *C. elegans* with the exception of transport into space [70]. In the early days of *Caenorhabditis* sp. research, much time was invested to establish a defined axenic medium to grow nematodes [64, 65] for two major reasons. First, it required a lower frequency of sub-culturing. Before the cryopreservation method was developed, infrequent sub-culture requirements were a great advantage. Second, axenic culture offered the ability to chemically define the medium, which allows the researcher to alter components and investigate nutritional requirements.

Monoxenic culture

Monoxenic cultures contain one organism as a food source. In the case of *C. elegans*, the nematode is almost exclusively cultured on media containing *E. coli*.

There are two main methods for monoxenic culture of *C. elegans*: either in liquid or on solid medium. In liquid culture, animals are grown with agitation in solution. On solid media, the animals are kept on nematode growth medium (NGM) agar plates seeded with an *E. coli* strain [20].

When animals are removed from their natural environments and transported to the laboratory, species undergo strong selective pressures that ultimately can change the organism. The impact of a laboratory environment on an organism is significant: environmental conditions are kept nearly constant; breeding regimes are strictly enforced; and food is readily available (**Box 2**). Additionally, researchers impose novel pressures by the culturing system, *e.g.* transferring individual animals to start a new culture (bottlenecks). The substrate on which animals are grown should be considered. Agar plates offer a two-dimensional substrate, whereas rotting fruit is a three-dimensional environment [15]. This laboratory propagation results in evolution through artificial selection, which inevitably affects genotypic and phenotypic characteristics of model organisms, including *C. elegans*.

From two studies, it is clear that the N2 genotype exhibits higher fitness in laboratory conditions than wild strains [26, 36]. The phenotype of N2 is distinct from wild strains in several ways, including aggregation behavior, maturation time, fecundity, body size, and many other traits [26, 27, 29, 36-43]. The atmospheric oxygen concentration on agar plates is substantially higher than levels preferred by wild strains [44-46], and laboratory oxygen concentration is a strong selective pressure on the organism. This oxygen concentration affects the growth and physiology of the

animal profoundly because many behaviors are altered, including how the animals consume bacterial food. These oxygen-dependent effects are so profound that two out of three confirmed laboratory-derived alleles are associated with altered behaviors at higher oxygen concentrations [16, 41]. The effects of these alleles and possibly other laboratory-derived alleles are pleiotropic, so they could have been selected by additional unexplained pressures.

Laboratory-derived alleles in the *C. elegans* N2 strain and their functional consequences

During the first 18 years that the N2 strain was grown in the laboratory, many mutations arose that might not have conferred any selective advantage [21]. However, we know that laboratory propagation of this strain led to the fixation of several alleles that confer a strong selective advantage under these conditions [26, 36]. Laboratory-derived alleles distinguish themselves from random mutations by increasing the fitness of the organism in laboratory conditions. At least three genes in the N2 strain have laboratory-derived variation: *npr-1*, *glb-5*, and *nath-10* [16, 36, 41]. For each of these genes, the N2 genome contains a different variant than found among all *bona fide* wild strains. Furthermore, the two N2-diverged strains, LSJ1 and LSJ2 (**Figure 1**), carry the same alleles as wild strains. These results provide further evidence for the laboratory origin of the alleles, because LSJ1 and LSJ2 were separated from the N2 strain at least 6 years before cryopreservation [16, 25, 26].

The neuropeptide receptor encoding gene *npr-1*: laboratory adaptation abnormally represses the *C. elegans* nervous system

A seven transmembrane neuropeptide receptor encoded by *npr-1* was first identified as a master regulator of a behavioral dimorphism where animals either aggregate or remain solitary in the presence of bacterial food [27]. This aggregation behavior mapped to an amino-acid substitution within the third intracellular loop of the NPR-1 receptor. Wild strains of *C. elegans* contain the 215F allele, with which the NPR-1 receptor responds to the neuropeptide FLP-21. By contrast, the laboratory strain N2 contains the 215V allele, which leads to a neomorphic gain-of-function sensitivity of NPR-1 to FLP-18 in addition to sensitivity to FLP-21 [47]. This gain-of-function sensitivity creates an abnormally repressed neural circuit through inactivation of the RMG interneuron [48], affecting a large number of behaviors (**Table 2**) [16, 27, 31, 39, 41, 44-56].

A modified aerotaxis response is one of the central drivers of the behavioral differences caused by variation in NPR-1 (**Figure 2**). Wild-type *C. elegans* strongly prefer oxygen concentrations lower than ambient levels [44-46]. On agar plate cultures, this behavior manifests as taxis to oxygen concentrations of approximately 10%, which is often found at the border of the bacterial lawn [44-46]. Aggregation of animals decreases the local oxygen concentration even further [46]. This

reduction in oxygen concentration caused by aggregation reinforces the further formation of aggregates, which in turn decreases available food as animals compete in close proximity. The reduction in growth rate and offspring production observed in wild *C. elegans* strains is likely caused by a mild starvation state in aggregates [41]. Additionally, these animals could experience higher levels of pheromones, potentially signaling a stress state that reduces growth rate and offspring production [41]. The attraction of wild strains to the border of the lawn increases the exposure to bacteria [39]. When these bacteria are pathogenic, strains with the 215F allele will be exposed more extensively to the pathogen and succumb faster to infection than the N2 strain with the 215V allele [41, 49, 50].

Table 2: The laboratory-derived allele of *npr-1* causes a large number of phenotypic effects.

Trait	Phenotypic effect	Related to aerotaxis?	Ref.
Aggregation	Lower	Yes	[27, 44, 47, 48]
Taxis to low oxygen	Lower	Yes	[41, 44-46]
Pathogen avoidance	Higher	Yes	[41, 49, 50]
Lifetime fecundity	Higher	Yes	[41]
Body size	Larger	Yes	[41]
Gene-expression regulation	NA	Yes	[31, 41]
Ethanol tolerance	Lower	Not tested	[51]
Carbon dioxide avoidance	Higher	Not tested	[16, 52]
Heat avoidance	Higher	Not tested	[53]
Hermaphrodite leaving	Lower	Yes	[39]
Pheromone responses	Repulsed	Not tested	[48, 54]
Lethargus quiescence	Higher	Yes	[55]
Crawling speed	Lower	Not tested	[56]

Most traits related to *npr-1* variation are linked to aerotaxis behavior (**Table 2**). Traits not linked to aerotaxis include heat avoidance [53], ethanol tolerance [51], carbon dioxide avoidance [16, 52], and pheromone response [48, 54]. However, these traits still might be linked to aerotaxis via the RMG neuron, but these connections have not been characterized extensively. For example, the ethanol response might be regulated through FLP-18 [51] or might be sensed in the nociception neuron ASH, which is connected to RMG [48]. RMG is also connected to the pheromone sensing ADL neuron. As the aggregation observed in wild strains likely causes higher pheromone exposure and lower oxygen concentrations, it is difficult to distinguish the contributions of both factors [41]. Many traits where *npr-1* variation is implicated in the behavior have not been directly connected to aerotaxis behaviors by empirical evidence. Most of these traits, however, are likely caused by variation in the aerotaxis responses and differences in food consumption mediated by RMG through its role as a "hub and spoke" neuron [48].

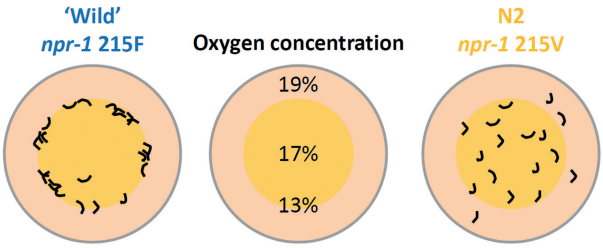


Figure 2: **The aerotaxis effects of *npr-1*.** Natural *C. elegans* strains aggregate at the edges of the bacterial lawn (orange) when propagated in laboratory conditions on monoxenic agar plates (left diagram). The edges of the bacterial lawn have lower than ambient oxygen concentrations (approximately 13%, center diagram). Wild *C. elegans* strains respond to this oxygen gradient and prefer lower oxygen concentrations in the presence of bacterial food [44]. The abnormal N2 strain is less sensitive to oxygen concentrations and does not aggregate at the edges of the bacterial lawn (right diagram). This difference in aerotaxis or oxygen preference leads to different aggregation and lawn leaving behaviors [27, 39, 44, 47, 48]. Because of these behavioral changes, strains also differ in exposure to pathogens [41, 49, 50]. Additionally, the aggregation behavior causes a chronic mild starvation state, which results in a reduced growth rate [41], reduced fecundity [41], altered gene expression [31, 41], increased crawling speed [56], and reduced quiescence during lethargus [55].

Other laboratory-derived alleles found in the N2 strain affect *nath-10* and *glb-5*

Together with NPR-1, the neuronal globulin domain protein GLB-5 affects a behavioral response to changes in carbon dioxide and oxygen concentrations. The causal variant in *glb-5* is a duplication/insertion of 765 base pairs, leading to a 179 amino-acid truncation and a 40 amino-acid substitution in the N2 strain [16, 56]. The combination of laboratory-adapted alleles at the *glb-5* and *npr-1* loci leads to opposing responses to changes in carbon dioxide concentration, as compared to the wild alleles. Wild strains move more quickly and make more turns when they sense a simultaneous decrease in carbon dioxide concentration and increase in oxygen concentration. By contrast, N2 animals move more quickly and make more turns when they sense an increase in carbon dioxide concentration. Furthermore, the N2 allele of *glb-5* desensitizes the URX neuron (which is also connected to RMG) to small fluctuations in oxygen levels, leading to reduced responses to in oxygen concentrations [16, 56]. The *npr-1* and *glb-5* alleles exhibit a genetic interaction. A strain with the wild-type alleles at both loci display a different phenotype than the strains with only one allele. If only the N2 allele of *npr-1* is present, animals will react to oxygen in a concentration dependent manner, whereas the N2 *glb-5* allele on its own renders them insensitive to fluctuations in oxygen concentrations. If animals carry both alleles, however, they react strongly to minor shifts in oxygen and carbon dioxide concentrations around the atmospheric oxygen concentration [16, 56]. These discoveries related to oxygen and carbon dioxide preferences led to original observations of the derived nature of the N2 strain [16].

Variation in the human N-acetyltransferase homolog gene (*nath-10*) causes variation in vulval cell-fate specification and shows pleiotropic effects on fecundity and growth rate [36]. The laboratory-derived allele encodes a putative substitution of methionine 746 with isoleucine in a highly conserved region of the N-acetyltransferase domain. This laboratory-derived allele was identified because of specific effects on variation in vulval cell-fate specification. Variation in *nath-10* causes visible effects on vulval development only when additional mutations sensitize the *let-60* Ras pathway activity. The laboratory-derived allele of *nath-10* partially suppresses a lower level of vulval cell-fate induction caused by a reduction-of-function mutation in the gene encoding an EGF receptor (*let-23*) and enhances the level of vulval cell-fate induction caused by a gain-of-function mutation in the gene encoding Ras (*let-60*), indicating that the laboratory-derived allele of *nath-10* stimulates Ras pathway activity. This allele also affects the age at maturity, brood size, and egg-laying speed through an increase in the production of sperm. Given this large effect on fitness, the N2 allele of *nath-10* causes a selective advantage when animals are grown in laboratory competition assays [36].

The effects of natural allelic variation is obscured by propagation of strains in the laboratory

To investigate the effects of laboratory alleles, we analyzed the *C. elegans* linkage mapping results from the last decade for linkage to *npr-1*, *glb-5*, and *nath-10* genomic regions (**Table 3**). A large number of linkage mapping experiments detected a quantitative trait locus (QTL) with a confidence interval that includes the *npr-1* locus, including dauer formation [57, 58], body size [38, 41], lifespan [59], and vulval index [36]. Laboratory-derived alleles have large effects when strains are grown in laboratory conditions. To estimate this effect, we compared the broad-sense heritability (H^2) and the variance explained by the *npr-1* QTL. This comparison indicates how much of the genetic differences among strains are influenced by the *npr-1* QTL. Variation at the *npr-1* locus explains 30-82% of the variance contributed by genetic factors [38, 41] – a large phenotypic effect. However, not all traits consistently detect a QTL at the *npr-1* locus. For example, one expression QTL study detected a *trans*-band at *npr-1* [31], but three other studies did not [60-62]. Similarly, one study detected the *npr-1* QTL for fecundity [41], but two other studies did not [37, 38]. We suggest that seemingly inconsistent studies that failed to detect a QTL nearby *npr-1* likely had different laboratory culture conditions, including the number of animals in the culture and the assay temperature. The differences in population density and variation at *npr-1* interact with large effects on the growth and physiology of the organism [41]. With increased culture density comes more crowding of animals at the edge of the bacterial lawn. This crowding causes a chronic low-level starvation state, which has large phenotypic effects. Additionally, differences in rearing temperature result in altered growth rates and population densities with similar effects.

Table 3: Many linkage mapping genetic studies identified the *npr-1*, *glb-5*, or *nath-10* locus.

Strains	Trait	Interval detected*			Identified Causal gene	Ref.
		nath-10	*glb-5*	*npr-1*		
N2xBO	Lifespan		yes			[72]
	Oxidative stress response		yes			
N2xCB4856	Age at maturity, 24°C	yes				[37]
N2xCB4856	Body mass, 12°C			yes		[38]
	Body mass, 24°C			yes		
N2xDR1350	Dauer formation, high food, 19°C			yes		[57]
	Dauer formation, food, plasticity			yes		
N2xCB4856	Pathogen susceptibility			yes	*npr-1*	[50]
N2xCB4856	Lifespan			yes		[59]
N2xCB4856	Carbon dioxide upshift, oxygen downshift		yes	yes	*glb-5*, *npr-1*	[16]
	Carbon dioxide downshift, oxygen upshift		yes	yes	*glb-5*, *npr-1*	
N2xCB4856	Oxygen sensing and response		yes	yes	*glb-5, npr-1*	[56]
N2# xCB4856	Male tail phenotype, 13°C	yes				[73]
N2xCB4856	Gene expression, L4, 24°C	yes				[61]
	Gene expression, L4 and reproductive, 24°C	yes				
N2xCB4856	Gene expression, young adult, 20°C			yes		[31]
N2xCB4856	Population growth on RNAi (8/11 genes)	yes			*ppw-1*	[74]
N2xCB4856	Lawn leaving			yes	*tyra-3*	[39]
N2xCB4856	Heat avoidance			yes	*npr-1*	[53]
JU605xJU606	Vulval induction, 20°C			yes		[36]
	Vulval induction, 25.5°C	yes			*nath-10*	
	Vulval induction, plasticity	yes			*nath-10*	
N2xCB4856	Bordering		yes		*exp-1*	[40]
N2xCB4856	Thermal preference			yes		[75]
	Isothermal dispersion			yes		
N2xCB4856	Dauer formation	yes		yes		[58]
N2xCB4856	Gene expression			yes		[76]
N2xCB4856	Lifetime fecundity			yes	*npr-1*	[41]
	Adult body size		yes	yes	*npr-1*	
	Susceptibility to *S. aureus*			yes	*npr-1*	
	Gene expression			yes	*npr-1*	
N2xCB4856	Dauer formation (pheromone exposure)	yes	yes			[58]
	Dauer formation (food exhaustion)	yes			*nath-10*	
N2xCB4856	Embryonic development	yes				[77]

*: In case these intervals are not detected, it can be due to this particular interval not involved in the trait or technical reasons (presence of markers, statistical power, etc)

#: The strain used for constructing the recombinant inbred population was CB5362, a strain containing the *tra-2(ar221)* and the *xol-1(y9)* mutations in an N2 background

The *nath-10* locus has been associated with several traits, including age at maturity [36, 37], an expression QTL *trans*-band [61], and dauer formation [58, 63]. It is difficult to assess the effect of the locus in general, as only one study reports the contribution to heritable variation (52%) [37]. The *trans*-band associated with the *nath-10* locus was measured in L4 and reproductive animals at 24°C. Therefore, it is likely that the developmental differences caused by *nath-10* result in gene-expression differences. Because *nath-10* is a pleiotropic locus implicated in fecundity and growth rate, these correlated QTL are not surprising.

In summary, the contribution of laboratory-derived alleles to heritable variation is large (30-82%) and seems to be environment-dependent. It is important to consider the context in which traits are measured. Given that both *npr-1* and *glb-5* affect behavior at atmospheric oxygen concentrations [16, 41, 44], many behavioral studies using the N2 strain in standard laboratory conditions might be difficult to interpret with respect to a normal behavioral circuit and natural behaviors.

Where do we go from here?

C. elegans is an essential model organism used to understand human biology. However, we need to be aware of the large and pleiotropic phenotypic effects caused by laboratory-derived alleles, especially those alleles present in the reference strain N2. These alleles can influence our conclusions and could alter the interpretations of results for understanding human biology, as it alters the natural physiology of *C. elegans*. However, investigators should not abandon the N2 strain. The large experimental toolkit and the decades of results obtained by study of this one strain are invaluable. These advantages need to be tempered with the knowledge that the N2 strain has been bred in a single environment for a long time prior to cryopreservation. Laboratories whose research focuses exclusively on the N2 strain and mutant derivatives should consider expanding to more natural *C. elegans* strains, especially when the focus includes traits that are influenced by population density (*e.g.* metabolism).

Newly isolated *C. elegans* that are cryopreserved as soon as possible after arrival in the laboratory are an untapped resource of genetic variants to expand the experimental power of *C. elegans* and its applicability to humans. For example, these strains can be added to the panel currently used for GWAS [12-14]. Additionally, new recombinant inbred line collections can be constructed using natural strains, which will greatly benefit quantitative genetic studies. One of the major strengths of *C. elegans* is the combination of wild strains and the accumulated knowledge from the study of the laboratory strain N2. This combination allows for rapid screening of causal genes to understand evolutionary and ecological genetics along with making a larger impact on biomedical science. The *C. elegans* research community is ready for the next round of rapid and important progress once natural strains are integrated into the existing genetic toolkit.

Acknowledgements

We thank Rachel Ankeny, Bob Horvitz, Patrick McGrath, John Sulston, and Diana Wall for discussions, further documentation, and additional information about the history of the *C. elegans* N2 strain. We are grateful to Roel Bevers, Daniel Cook, Patrick McGrath, Anne Morbach, Lisa van Sluijs, Robyn Tanny, and Stefan Zdraljevic for editorial comments on the manuscript.

Glossary

Axenic culture: Conditions where organisms are grown in completely synthetic media. In the case of *C. elegans*, media is based on a liver extract [64, 65].
Dauer: At high culture density, low food abundance, and high temperature, second larval stage (L2) animals enter this alternate larval stage. [66] These L3 dauer larvae can survive stressful conditions and are thought to disperse to new locations in nature.
Ecological niche: The specific environment in which an organism lives and competes for resources. *C. elegans* are most often found in decaying material and not in soil [15, 32].
FLP-18: One of two FMRFamide neuropeptides encoded by the *C. elegans* genome that can bind to the NPR-1 neuropeptide receptor. FLP-18 can activate the NPR-1(215V) allele (found in N2 animals) but not the NPR-1(215F allele) (found in all wild strains) [16, 47].
FLP-21: The second of two FMRFamide neuropeptides encoded by the *C. elegans* genome that can bind to the NPR-1 neuropeptide receptor. FLP-21 is the natural ligand of NPR-1 [16, 41]
GLB-5: A globin domain protein that modifies behavioral responses to oxygen and oxygen/carbon dioxide stimuli [16, 56]
GWAS: Genome-wide association study, a technique used on natural populations to identify genomic regions correlated with differences in phenotypic traits [12-14].
Heritability: The amount of trait variation in a population that can be explained by genetic factors.
NATH-10: A vertebrate N-acetyltransferase homolog that has been shown to affect vulval induction in *C. elegans*. The 746I allele (N2) also results in faster maturation and fitness under laboratory conditions [36].
Neomorphic: Describes a type of phenotype caused by an alteration of gene function that is novel and different from the normal function of the gene.
NPR-1: a G protein-coupled neuropeptide receptor normally activated by FLP-21. The 215V allele (N2) gained the ability to respond to FLP-18. The 215V allele affects many different traits [16, 27, 31, 39, 41, 44-56].
Private alleles: Specific alleles occurring only in a single strain.
QTL: Quantitative trait locus, a locus correlated with quantitative trait variation. For example, the *npr-1* QTL is correlated with variation in aggregation behavior.
QTN: Quantitative trait nucleotide, the variant site that causes variation in the quantitative trait.
RMG interneuron: The central neuron involved in most behaviors mediated by NPR-1 [48].
***Trans*-band**: In expression QTL studies, when variation of expression in many genes is correlated with the same genomic locus [31, 60-62].
Vulval cell induction: When any of six hypodermal cells located on the ventral surface of the hermaphrodite are specified and divide to become vulval cells.

References

1. **Dunham, M.J. and D.M. Fowler**, *Contemporary, yeast-based approaches to understanding human genetic variation.* **Curr Opin Genet Dev, 2013. 23(6): p. 658-64**.
2. **Bloom, J.S., et al.**, *Finding the sources of missing heritability in a yeast cross.* **Nature, 2013. 494(7436): p. 234-7**.
3. **Ejsmont, R.K. and B.A. Hassan**, *The Little Fly that Could: Wizardry and Artistry of Drosophila Genomics.* **Genes (Basel), 2014. 5(2): p. 385-414**.
4. **Jackson, B.M. and D.M. Eisenmann**, *beta-catenin-dependent Wnt signaling in C. elegans: teaching an old dog a new trick.* **Cold Spring Harb Perspect Biol, 2012. 4(8): p. a007948**.
5. **Felix, M.A. and M. Barkoulas**, *Robustness and flexibility in nematode vulva development.* **Trends Genet, 2012. 28(4): p. 185-95**.
6. **Husson, S.J., A. Gottschalk, and A.M. Leifer**, *Optogenetic manipulation of neural activity in C. elegans: from synapse to circuits and behaviour.* **Biol Cell, 2013. 105(6): p. 235-50**.
7. **Kenyon, C.**, *The first long-lived mutants: discovery of the insulin/IGF-1 pathway for ageing.* **Philos Trans R Soc Lond B Biol Sci, 2011. 366(1561): p. 9-16**.
8. **Fire, A., et al.**, *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* **Nature**, **1998. 391(6669): p. 806-1**1.
9. **Kaufman, E.J. and E.A. Miska**, *The microRNAs of Caenorhabditis elegans.* Semin **Cell Dev Biol, 2010. 21(7): p. 728-37**.
10. **Ehrenreich, I.M., J.P. Gerke, and L. Kruglyak**, *Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross.* **Cold Spring Harb Symp Quant Biol, 2009. 74: p. 145-53**.
11. **Gaertner, B.E. and P.C. Phillips**, *Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation.* **Genet Res (Camb), 2010. 92(5-6): p. 331-48**.
12. **Andersen, E.C., et al.**, *Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nat Genet, 2012. 44(3): p. 285-90**.
13. **Rockman, M.V. and L. Kruglyak**, *Recombinational landscape and population genomics of Caenorhabditis elegans.* **PLoS Genet, 2009. 5(3): p. e1000419**.
14. **Ashe, A., et al.**, *A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity.* **Elife, 2013. 2: p. e00994**.
15. **Felix, M.A. and C. Braendle**, *The natural history of Caenorhabditis elegans.* **Curr Biol, 2010. 20(22): p. R965-9**.
16. **McGrath, P.T., et al.**, *Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors.* **Neuron, 2009. 61(5): p. 692-**9.
17. **Yvert, G., et al.**, *Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* **Nat Genet, 2003. 35(1): p. 57-64**.
18. **Brem, R.B., et al.**, *Genetic dissection of transcriptional regulation in budding yeast.* **Science, 2002. 296(5568): p. 752-5**.
19. **van Zanten, M., et al.**, *The many functions of ERECTA.* **Trends Plant Sci, 2009. 14(4): p. 214-8**.
20. **Brenner, S.**, *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94**.
21. **Denver, D.R., et al.**, *A genome-wide view of Caenorhabditis elegans base-substitution mutation processes.* **Proc Natl Acad Sci U S A, 2009. 106(38): p. 16310-**4.
22. **Vergara, I.A., et al.**, *Polymorphic segmental duplication in the nematode Caenorhabditis elegans.* **BMC Genomics, 2009. 10: p. 329**.
23. **Gems, D. and D.L. Riddle**, *Defining wild-type life span in Caenorhabditis elegans.* **J Gerontol A Biol Sci Med Sci, 2000. 55(5): p. B215-9**.
24. **Dougherty, E.C.** *Letter from Ellsworth C. Dougherty to Sydney Brenner*. **Sydney Brenner Collection (1927-2010) [Letter] 1963 10-22-1963 [cited 2015 January 5, 2015]; 1**].
25. **McGrath, P.T., et al.**, *Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes.* **Nature, 2011. 477(7364): p. 321-5**.
26. **Weber, K.P., et al.**, *Whole genome sequencing highlights genetic changes associated with laboratory domestication of C. elegans.* **PLoS One, 2010. 5(11): p. e13922**.

27. **de Bono, M. and C.I. Bargmann**, *Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans.* **Cell, 1998. 94(5): p. 679-89**.
28. **Reiner, D.J., et al.**, *C. elegans anaplastic lymphoma kinase ortholog SCD-2 controls dauer formation by modulating TGF-beta signaling.* **Curr Biol, 2008. 18(15): p. 1101-9**.
29. **Volkers, R.J., et al.**, *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93**.
30. **Thompson, O., et al**., *The million mutation project: a new approach to genetics in Caenorhabditis elegans.* Genome **Res, 2013. 23(10): p. 1749-62**.
31. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak**, *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6**.
32. **Petersen, C., et al.**, *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: p. 4**.
33. **Petersen, C., P. Dirksen, and H. Schulenburg**, *Why we need more ecology for genetic models such as C. elegans.* **Trends Genet, 2015**.
34. **Kiontke, K.C., et al.**, *A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits.* **BMC Evol Biol, 2011. 11: p. 339**.
35. **Lee, H., et al.**, *Nictation, a dispersal behavior of the nematode Caenorhabditis elegans, is regulated by IL2 neurons.* **Nat Neurosci, 2012. 15(1): p. 107-12**.
36. **Duveau, F. and M.A. Felix**, *Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans.* **PLoS Biol, 2012. 10(1): p. e1001230**.
37. **Gutteling, E.W., et al.**, *Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in Caenorhabditis elegans.* **Heredity (Edinb), 2007. 98(1): p. 28-37**.
38. **Gutteling, E.W.**, et al., *Environmental influence on the genetic correlations between life-history traits in Caenorhabditis elegans.* **Heredity (Edinb), 2007. 98(4): p. 206-13**.
39. **Bendesky, A., et al.**, *Catecholamine receptor polymorphisms affect decision-making in C. elegans.* **Nature, 2011**. **472(7343): p. 313-8**.
40. **Bendesky, A., et al.**, *Long-range regulatory polymorphisms affecting a GABA receptor constitute a quantitative trait locus (QTL) for social behavior in Caenorhabditis elegans.* **PLoS Genet, 2012. 8(12): p. e1003157**.
41. **Andersen, E.C., et al.**, *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156**.
42. **Kammenga, J.E., et al.**, *A Caenorhabditis elegans wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3.* **PLoS Genet, 2007. 3(3): p. e34**.
43. **Snoek, L.B., et al.,** *A rapid and massive gene expression shift marking adolescent transition in C. elegans.* **Sci Rep, 2014. 4: p. 3912**.
44. **Gray, J.M., et al.,** *Oxygen sensation and social feeding mediated by a C. elegans guanylate cyclase homologue.* **Nature, 2004. 430(6997): p. 317-22**.
45. **Chang, A.J., et al.**, *A distributed chemosensory circuit for oxygen preference in C. elegans.* **PLoS Biol, 2006. 4(9): p. e274**.
46. **Rogers, C., et al.**, *Behavioral motifs and neural pathways coordinating O2 responses and aggregation in C. elegans.* **Curr Biol, 2006. 16(7): p. 649-59**.
47. **Rogers, C., et al.**, *Inhibition of Caenorhabditis elegans social feeding by FMRFamide-related peptide activation of NPR-1.* **Nat Neurosci, 2003. 6(11): p. 1178-85**.
48. **Macosko, E.Z.**, et al., *A hub-and-spoke circuit drives pheromone attraction and social behaviour in C. elegans.* **Nature, 2009. 458(7242): p. 1171-5**.
49. **Styer, K.L., et al.**, *Innate immunity in Caenorhabditis elegans is regulated by neurons expressing NPR-1/GPCR.* **Science, 2008. 322(5900): p. 460-4**.
50. **Reddy, K.C., et al.**, *A polymorphism in npr-1 is a behavioral determinant of pathogen susceptibility in C. elegans.* **Science, 2009. 323(5912): p. 382-4**.
51. **Davies, A.G., et al.**, *Natural variation in the npr-1 gene modifies ethanol responses of wild strains of C. elegans.* **Neuron, 2004. 42(5): p. 731-43**.

52. **Hallem, E.A. and P.W. Sternberg**, *Acute carbon dioxide avoidance in Caenorhabditis elegans.* **Proc Natl Acad Sci U S A, 2008. 105(23): p. 8038-43.**
53. **Glauser, D.A., et al.,** *Heat avoidance is regulated by transient receptor potential (TRP) channels and a neuropeptide signaling pathway in Caenorhabditis elegans.* **Genetics, 2011. 188(1): p. 91-103**.
54. **Jang, H., et al.**, *Neuromodulatory state and sex specify alternative behaviors through antagonistic synaptic pathways in C. elegans.* **Neuron, 2012. 75(4): p. 585-92.**
55. **Choi, S., et al.,** *Analysis of NPR-1 reveals a circuit mechanism for behavioral quiescence in C. elegans.* **Neuron, 2013. 78(5): p. 869-80**.
56. **Persson, A., et al.,** *Natural variation in a neural globin tunes oxygen sensing in wild Caenorhabditis elegans.* **Nature, 2009. 458(7241): p. 1030-3**.
57. **Harvey, S.C., A. Shorto, and M.E. Viney**, *Quantitative genetic analysis of life-history traits of Caenorhabditis elegans in stressful environments.* **BMC Evol Biol, 2008. 8: p. 15**.
58. **Green, J.W., et al.**, *Genetic mapping of variation in dauer larvae development in growing populations of Caenorhabditis elegans.* **Heredity (Edinb), 2013. 111(4): p. 306-13**.
59. **Doroszuk, A., et al.**, *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e11**0.
60. **Li, Y., et al.,** *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222**.
61. **Vinuela, A., et al.**, *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37**.
62. **Li, Y., et al.**, *Global genetic robustness of the alternative splicing machinery in Caenorhabditis elegans.* **Genetics, 2010. 186(1): p. 405-10.**
63. **Green, J.W., et al.**, *Highly Polygenic Variation in Environmental Perception Determines Dauer Larvae Formation in Growing Populations of Caenorhabditis elegans.* **PLoS One, 2014. 9(11): p. e112283**0.
64. **Hansen, E.L.Y., E.A.; Nicholas, W.L.; Sayre, F.W.,** *Differential nutritional requirements for reproduction of two strains of caenorhabditis elegans in axenic culture.* **Nematologica, 1960. 5: p. 27-31**.
65. **Nicholas, W.L. and E.M. Mc**, *A technique for obtaining axenic cultures of rhabditid nematodes.* **J Helminthol, 1957. 31(3): p. 135-44**.
66. **Altun, Z.F., Herndon, L.A., Crocker, C., Lints, R. and Hall, D.H.** *Wormatlas*. **2002-2016**.
67. **Ankeny, R.A.**, *The natural history of Caenorhabditis elegans research.* **Nat Rev Genet, 2001. 2(6): p. 474-9**.
68. **Brown, A.**, *In the beginning was the worm : finding the secrets of life in a tiny hermaphrodite*. 2003, **New York: Columbia University Press. 244 p**.
69. **Riddle, D.L.**, *C. elegans II*. Cold Spring Harbor monograph series,. 1997, Plainview, N.Y.: **Cold Spring Harbor Laboratory Press. xvii, 1222** p.
70. **Adenle, A.A., B. Johnsen, and N.J. Szewczyk**, *Review of the results from the International C. elegans first experiment (ICE-FIRST).* **Adv Space Res, 2009. 44(2): p. 210-216**.
71. **Haber, M., et al.**, *Evolutionary history of Caenorhabditis elegans inferred from microsatellites: evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding.* **Mol Biol Evol, 2005. 22(1): p. 160-73**.
72. **Shmookler Reis, R.J., P. Kang, and S. Ayyadevara**, *Quantitative trait loci define genes and pathways underlying genetic variation in longevity.* **Exp Gerontol, 2006. 41(10): p. 1046-5**4.
73. **Chandler, C.H.**, *Cryptic intraspecific variation in sex determination in Caenorhabditis elegans revealed by mutations.* **Heredity (Edinb), 2010. 105(5): p. 473-82**.
74. **Elvin, M., et al.**, *A fitness assay for comparing RNAi effects across multiple C. elegans genotypes.* **BMC Genomics, 2011. 12: p. 510**.
75. **Gaertner, B.E., et al**., *More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in Caenorhabditis elegans.* **Genetics, 2012. 192(4): p. 1533-42**.
76. **Francesconi, M. and B. Lehne**r, *The effects of genetic variation on gene expression dynamics during development.* **Nature, 2014. 505(7482): p. 208-1**1.
77. **Snoek, L.B., et al.**, *Widespread genomic incompatibilities in Caenorhabditis elegans.* **G3 (Bethesda), 2014. 4(10): p. 1813-23**.

Chapter 3

A complementary introgression line population in *Caenorhabditis elegans*

Mark G. Sterken,
Jelle W. van Creij,
Joost A.G. Riksen,
Beatrice Tan,
L. Basten Snoek,
Jan E. Kammenga

Abstract

A quantitative genetics study is as good as the mapping population used. One of the most powerful mapping population designs is the introgression line (IL) design. ILs contain a single genetic locus introgressed in an otherwise homogeneous background. For the nematode *Caenorhabditis elegans*, one such population is available, containing loci of Hawaii CB4856 in a Bristol N2 background. Here we present the first genome-wide complementary IL population, it contains loci of N2 in a CB4856 background. The combination of both panels retains the power of the IL design, while adding the possibility to study the effect of the genetic background.

The ILs were created by backcrossing six N2 x CB4856 recombinant inbred lines (RILs) with CB4856. We started with two backcrosses. First, a CB4856 hermaphrodite was crossed with a RIL male. Second, the F1 was backcrossed to a CB4856 male. This crossing scheme assures that the mitochondria are derived from CB4856 and the N2 *peel-1/zeel-1* incompatibility locus on chromosome I can be removed. After these two crosses, the strains were selected using 41 insertions-deletions between the N2 and CB4856 genomes as markers and additional backcrosses were made when needed. These steps resulted in a population of 154 strains covering ~97% of the *C. elegans* genome, with a median introgression size of 3.9 Mb. To demonstrate the complementarity of the N2-in-CB4856 ILs with the CB4856-in-N2 ILs, two traits were measured in three strains of each IL population: (i) resistance to heat stress, and (ii) transcriptional response to heat stress. For these experiments, synchronized populations of ILs were grown for 46 hours at 20°C and either grown for an additional 2 hours at 20°C or exposed to 35°C for 2 hours. Both experiments, albeit small, showed that for ~100 gene expression traits and one heat-stress survival trait there is evidence for background interactions.

Construction of the genome-wide IL population and its coarse genetic map has laid the foundation for whole-genome trait mapping respective to the genetic background. Future plans include measuring several life-history traits in both complementary populations and the construction of a high-resolution genetic map by sequencing. The combination of both panels can resolve complex trait architecture in relation to the genetic background.

Introduction

In the last decade many advances were made in quantitative genetics using *Caenorhabditis elegans*. The most frequently used strains for such studies are the N2 and CB4856 strains (for an overview, see [1] and **Chapter 2**). Even though more wild isolates have become available, N2 and CB4856 remain among the most divergent strain pairs [2-4]. These two strains differ in 327,050 single nucleotide variants and 79,529 insertions-deletions [4]. These polymorphisms also translate to differences in traits, of which the trait variation originating from a single amino acid substitution in the neuropeptide receptor *npr-1* is arguably the best-studied (see also **Chapter 2**) [5, 6]. Furthermore, crossing populations between N2 and CB4856 so far identified polymorphisms affecting the genes *zeel-1*, *tra-3*, *plg-1*, *glb-5*, *tyra-3*, *glc-1*, *exp-1*, and *ppw-1* as causal [7-15]. These results mark the nematode *C. elegans* as a well-studied quantitative genetics model.

Most quantitative genetics studies focussing on gene-identification take a two-step approach whereby first recombinant inbred lines (RILs) are used, followed by introgression lines (ILs). RILs consists of lines which form a genetic mosaic of both parental genomes, whereas ILs consists of small segments of one parental strain introgressed in the background of the other parental strain. In the RIL-IL approach, first a quantitative trait locus (QTL) is identified using RILs, followed by confirmation and fine-mapping using ILs (for examples see [8, 11]). Although this paradigm is used in most QTL studies, different approaches do exist. For example, approaches depending on chromosome-substitution populations, or whole-genome IL populations [16-20]. In studies using whole-genome IL populations, it has generally been noted that ILs are more sensitive for small-effect QTLs compared to RILs [17, 21, 22]. The reason ILs are more sensitive lies in the low amount of noise; there is no varying genetic background as in RILs. However, depending on the size of the introgressed locus, ILs generally are less informative on the exact location of the QTL. For mapping purposes in *C. elegans*, several inbred populations have been constructed using N2 and CB4856, including two RIL populations, a CB4856-in-N2 chromosome-substitution population, and two CB4856-in-N2 IL populations [16, 19, 23, 24]. However, an N2-in-CB4856 IL population is still missing.

Here we constructed an N2-in-CB4856 IL population, to complement the existing IL population. We took advantage of the recently constructed CB4856 genome assembly [4]. Our genotyping strategy used insertions-deletions between the N2 and the CB4856 strain for swift genotyping for selection during the crossing process. This population can be used complementary to the previously constructed CB4856-in-N2 population, allowing for further dissection of quantitative traits.

The added value of a complementary population is the possibility to detect loci-background interactions. In a single-background introgression line population the locus-background interactions are entangled per definition. Such interactions can only be detected when adjacent ILs overlap, but cannot be distinguished from closely-linked additive effects or interactions.

Adding a population with the complementary genetic background can further elucidate the genetic architecture of traits, as has already been demonstrated for specific QTLs [6, 11, 25]. Here we use a limited set of strains to exemplify the application of the combined IL populations. We measured two traits: heat-shock survival and the transcriptional response to heat-shock. For these traits we have advanced insight in the variation between N2 and CB4856 from other studies [26] [Snoek & Sterken, unpublished]. Therefore, these traits form an excellent benchmark for testing the combination of the complementary ILs.

Material and methods

Strains

The introgression lines were constructed from recombinant inbred lines with N2 and CB4856 parents, namely WN001, WN007, WN025, WN068, WN071, and WN110 [23, 27]. The original CB4856 parent was used for the crosses. By crossing these lines, 154 novel introgression lines were created (See **Figure 2** and **Supplementary file 1**). The introgression lines have been cryopreserved after at least 10 generations of inbreeding and the parental strains have been cryopreserved as well [28].

For the heat stress experiment the parental strains N2 and CB4856, the CB4856-in-N2 strains WN246, WN248 and WN251 [19], and the newly created introgression lines CBN068, CBN077, and CBN082 were used.

Strain maintenance

Strains were kept using standard culturing conditions. All strains were kept on NGM plates seeded with *Escherichia coli* OP50 and culturing temperatures used during the crosses were 12°C, 16°C, or 20°C, depending on the desired speed of population growth [28, 29].

Crossing scheme

To generate the CB4856-background introgression lines two crossing stages were used. The first cross stage was used for most loci, where a RIL male was back-crossed to a CB4856 hermaphrodite, to ensure the presence of CB4856 mitochondria in the F1 (making the background completely CB4856). This was followed-up by a second cross against CB4856 males, this resulted in homozygous CB4856 genotypes at the *peel-1/zeel-1* incompatibility locus on chromosome I [30]. After this step, selected genotyping was conducted in the F2 (4-6 markers), screening for a high number of CB4856 loci and (in most cases) absence of the N2 genotype at the *peel-1/zeel-1* locus. Selected strains were inbred further to obtain as many homozygous CB4856 loci as possible.

The second stage of crossing was conducted after the first round of selection was finished and multiple N2 loci remained. These strains were back-crossed against CB4856 males and selected further until only one detectable N2 locus remained in an otherwise CB4856 genotype background.

Genotyping Primers

Initial genotyping was done using newly developed primers utilizing insertions/deletions between the CB4856 and N2 genomes [4]. In total 41 primers have been developed, with a bias for covering loci with a high-recombination frequency (for a list, see **Supplementary file**

2). The selection criteria were: (i) the deletion occurred in CB4856, (ii) the deletion is larger than 25 bases and shorter than 150 bases, and (iii) is not located in a repetitive region. All primers have been developed with Primer3 (primer3-win-bin-2.3.6) on the 1000 bases up- and downstream of the deletion [31]. Primer3 was used with standard settings, selecting 3 primers in the size ranges of: 100-150bp, 200-250bp, 300-350bp, 400-450bp, 500-550bp, 600-650bp, 700-750bp, and 800-850bp. The annealing temperature was selected between 58°C and 60°C. The specificity of the primers was tested using BLAST (ncbi-blast 2.2.28 win64) against WS230 (settings: blastn –word_size 7 –reward 1 –penalty -3) [32]. Only primers with fewer than 5 hits were considered for further selection. Final selection of the primers was based on application of the primers (**Supplementary figure 1**).

Genotyping PCR

DNA was isolated from single adults that had generated offspring. The genotype of the parent gave information about the genotypes expected in the offspring. Nematodes were lysed at 65°C for 30 minutes using a custom lysis buffer [33], followed by 5 minutes at 99°C. Genotyping PCRs were performed with GoTaq using the manufacturers recommendations. The annealing temperature used was 58°C (30 seconds), with an elongation time of 1 minute, for 40 cycles. All samples were run on 1.5% agarose gels stained with Ethidium Bromide.

Genetic map construction

Based on the insertion-deletion markers a rough genetic map was constructed. This map was expanded by estimation of breakpoints based on the recombination frequencies measured in a *C. elegans* RIL population [24]. Between markers used in the RIL population, the recombination likelihoods were extrapolated linearly, using custom scripts written in R (3.2.2 x64). Confidence intervals were estimated based on the recombination frequencies over the expanded locations.

Heat-shock survival experiment

The experiment was started by transferring a starved population to a new 9 cm NGM dish, there the population was allowed to develop for ~60 h. At that time, the population consisted of egg-laying adults, which were isolated and bleached for synchronization (day 0) [28]. These synchronized populations were grown for 48 h at 20°C. Subsequently (day 2), 20-40 nematodes were transferred to 6 cm NGM plates containing FUdR, which inhibits cell division [34]. Per strain two plates were generated, one as a control (remaining on 20°C), one receiving a 4 h 35°C heat-shock immediately after transfer [26]. On day 3, 4 and 11 of the experiment the number of surviving and dead nematodes were counted. Three biological replicates were conducted.

For the two parental strains and the CB4856-in-N2 strains, additional replicates were included which were measured before the N2-in-CB4856 strains were ready. Five additional parental and two additional IL replicates were included. For the analysis we used an ANOVA model,

$$S_x = B_x + G_x + B_x \times G_x + e$$

where the survival (S) of strain *x* was explained over the genetic background (B, either N2 or CB4856) and the introgression (G, either N2 or CB4856) and the interaction between B and G, and residual variation, e.

Transcriptional response to heat-shock

The stress survival experiment was started by transferring a starved population to a new 9 cm NGM dish, where the population was allowed to develop for ~60 h. Subsequently, the population was bleached for synchronization. These populations were grown on 9 cm NGM dishes, two per treatment per strain. The control population was grown at 20°C for 48 h and the heat-shock population was grown at 20°C for 46 h, followed by 2 h at 35°C. At the end of the experiment, the population was washed off the plate using M9 and flash frozen in liquid nitrogen and stored at -80°C until processed further.

Transcriptional profiling using microarray

Transcriptional profiling was performed as described in [35]. In short, RNA was isolated using Promega's Maxwell® 16 AS2000 with the Maxwell® 16 LEV simply RNA tissue kit. The RNA was used in the 'Two-Color Microarray-Based Gene Expression Analysis' protocol from Agilent, with the *C. elegans* (V2) 4x44K slides from Agilent. The microarrays were scanned by an Agilent High Resolution C Scanner, and the data was extracted with Agilent Feature Extraction software (version 10.7.11). Data was normalized using Limma in R (3.2.2 x64). As recommended, the data was not background corrected before normalization [36]. Data was normalized within-array using Loess, and between-array using Quantile [37].

Statistical analysis of transcription data

All analyses were conducted in R (3.2.2 x64). First the quality of the transcriptome data was assessed, using correlation analysis. This analysis revealed no outlier profiles and when grouped, it separated the transcription profiles first on treatment (control and heat-shock), and secondly on the main genetic background (N2 and CB4856). Visual inspection of the expression of heat shock marker genes, *hsp-16.2* and *hsp-16.41* confirmed that the heat-shock samples received a heat-shock.

To confirm the strain identity, we used *cis*-eQTL mapped in the study of Rockman *et al*, 2010 to correlate the expression of the experimental strains with [38]. Per treatment the intensities were transformed to the mid-parental mean

$$R_{x,i} = \frac{Y_{x,i}}{0.5 * (Y_{N2,i} + Y_{CB4856,i})}$$

where Y stands for the untransformed intensities of strain *x* and spot *i* (1, 2, 3, ..., 45220). These values were correlated with the *cis*-eQTL effects per 20 genes (**Supplementary file 3**), which confirmed the genotypes of the strains used.

The effects of the introgressions and background were calculated using the ANOVA model

$$E_{x,i} \sim B_{x,i} + G_{x,i} + B_{x,i} \times G_{x,i} + e$$

where E stands for the log2 transformed intensities of strain *x* and spot *i* (1, 2, 3, ..., 45220). B is the genetic background (either N2 or CB4856), G is the introgression (either N2 or CB4856), and e is the residual variation. The outcome of this model was adjusted for multiple testing using the p.adjust function with the "BH" method [39].

Enrichment analyses on differentially expressed genes were conducted using a hypergeometric test. The following categories were investigated for enrichment: gene ontology [40, 41], anatomy terms [40, 41], protein domains [40, 41], gene classes [40, 41], KEGG [42], Transcription factor binding sites [43, 44], and RNAi phenotypes [40, 41].

Results

Primer selection and validation

Before setting out to generate an introgression line population, we devised a novel genotyping strategy based on the CB4856 genome sequence [4] and the recombination frequencies observed on *C. elegans* chromosomes [24].

We developed genetic markers that allow for fast and reliable screening. The comparative analysis of the genome sequence of CB4856 versus N2 identified several insertion/deletion polymorphisms between the strains. Amplification by PCR over insertions/deletions of sufficient size will yield clearly distinguishable amplicons and surpasses the need for an additional restriction enzyme digestion as needed for single nucleotide polymorphisms [19, 23, 24]. In essence, this approach results in a higher throughput and is more cost-efficient.

The preferred genomic locations of the markers were determined by the recombination frequencies of the *C. elegans* chromosomes. *C. elegans* has six chromosomes that display different recombination frequencies depending on the location at the chromosome. The tips and centres show a low recombination frequency, whereas the arms show a high recombination frequency [24]. Therefore, most recombination events can be covered by focussing on the arms of the chromosomes.

Finally, we used 41 markers for breeding of the introgression lines (**Figure 1**, **Supplementary file 2**). These were verified by re-genotyping of various recombinant inbred lines and introgression lines (see **Supplementary figure 1**), which showed that these markers can be used successfully.

Construction of a CB4856-background IL population

A population of 154 introgression lines containing an N2 segment in a CB4856 background was constructed. As for the reciprocal population, this set was created by back-crossing a limited set of six recombinant inbred lines [19]. In contrast to the reciprocal population, these strains were selected in-between by genotyping with a limited set of markers and additional back-crosses. Most lines containing only a single N2 segment were obtained after 1-3 additional back-crosses.

This set of lines covers the entire genome of *C. elegans* (**Figure 2**). The map was expanded by estimating the recombination events, based on a RIL population [24]. Furthermore, uncertainty estimates for the introgression sizes were made using the recombination frequencies of this population. The estimated median introgression size is 3.9 Mb (95% confidence interval: 1.2-5.5 Mb), with the smallest segment spanning 0.7 Mb (95% confidence interval: 0.4-1.7 Mb) and the largest 16 Mb (95% confidence interval: 11.6-16.4 Mb). These estimates place the introgression sizes in the same size range as the N2-background IL population [19]. We estimate that 2.8 Mb (95% confidence interval: 0.3-12.1 Mb) of the genome is not covered by N2 introgressions in this population. Therefore, the introgressions will cover ~97% of the *C. elegans* genome.

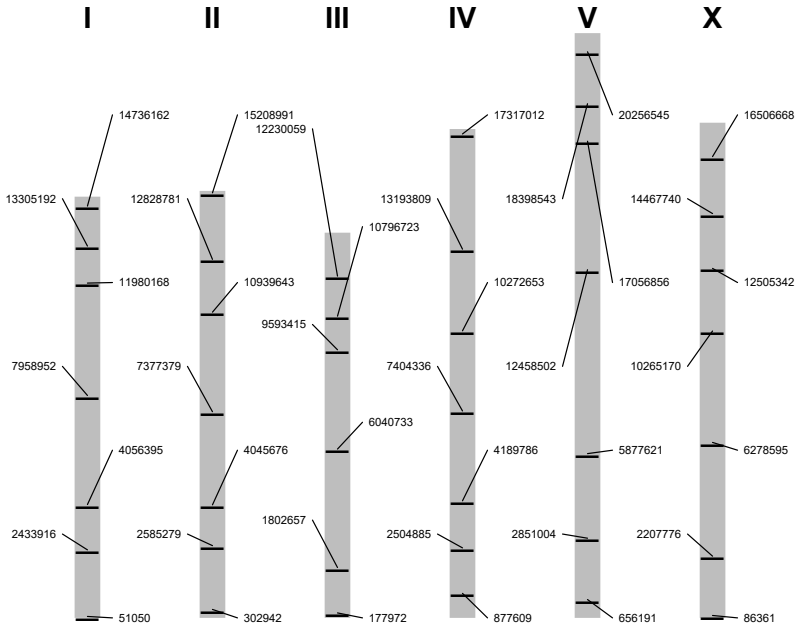


Figure 1. The insertion-deletion markers developed for breeding and initial genotyping of the introgression lines. In general, the most distal markers on each chromosome are located on or very near the tip of the chromosome. The second most distal markers are located on the middle of the arms. The third most distal markers are located near the border of the arms and the chromosome centres. The middle marker is located near the centre of the chromosome. Only for chromosome III no adequate marker was developed for the left-arm/centre boarder.

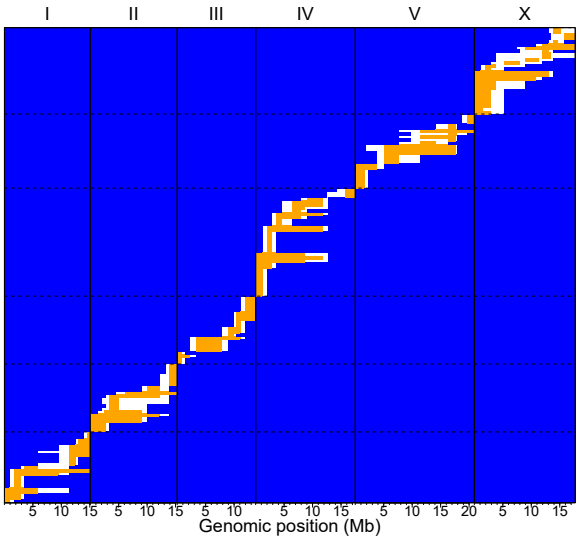


Figure 2. The genomes of the 154 N2-in-CB4856 introgression lines. Blue represents the CB4856 genotype and orange the N2 genotype, white space represents uncertainty in-between markers (p>0.25 as determined based on RIL recombination frequencies [24]). Chromosome I-V and X are covered by 23, 22, 22, 35, 24, and 28 lines, respectively. The mitochondrial DNA in these strains is inherited from CB4856.

Detection of heat-shock survival QTL on the top of chromosome IV using complementary ILs

We measured survival after heat-shock in the parental strains N2 and CB4856, three N2-background ILs and three CB4856-background ILs. The six ILs contained introgressions in their respective backgrounds on the top of chromosome IV; in previous work a heat-stress-resistance QTL was mapped to this location (see **Chapter 6**) [26]. Starting from the genetic map, it is likely that all these ILs share the introgression site in-between 1.0-2.5 Mb (mind, it cannot be ruled out that one of the novel ILs, CBN068, does not contain this site at all). For the measured traits we focus on this overlapping locus.

The locus on the top of chromosome IV influences heat-shock survival as determined previously (see **Chapter 6**) [26]. We found that in control conditions none of the strain types differed in survival on day 3 and 4 (as there were no deaths occurring at those days; ANOVA, $p >> 0.05$). However, on day 11 after hatching a weak effect from the genetic background was detected (ANOVA, $p < 0.05$); the N2 background showed an 8% increase in survival compared to CB4856. However, upon exposure to a heat-shock of 35°C for 4 hours during the L4 stage stronger effects were detected. The introgression significantly determined the survival rate in the introgression lines in the first 2 days after the heat-shock (ANOVA, $p < 0.01$). The N2-introgressions increased survival with 16% compared to the CB4856 introgression (**Figure 3A**). This is in line with previous observations on introgression lines covering this locus [**Chapter 6**].

The long-term survival after heat-shock depends on a locus-background interaction. While survival only weakly depends on the genetic background under control conditions, under heat-shock conditions a background-locus interaction is a major determinant of survival (ANOVA, $p < 0.01$, $R^2 = 0.30$). Although there seems to be quite some residual variation (**Figure 3B**), the observation is indicative of a genetic interaction of the locus with the background genotype.

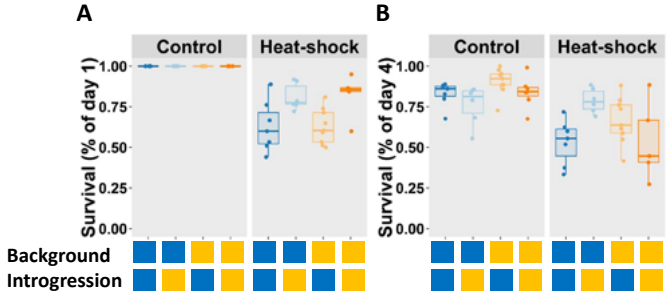


Figure 3. Survival after heat-shock, measured in three CB4856-in-N2 introgression lines (light orange), 3 N2-in-CB4856 (light blue) introgression lines and the two parental strains: N2 (orange) and CB4856 (blue). **(A)** Survival at an age of 4 days (two days post heat-shock), $n \geq 5$ for all measurements. The introgression determines the trait value under heat-shock (ANOVA, $p < 0.05$). **(B)** Survival at an age of 11 days, expressed as % of day 4. The survival is determined by an interaction between the introgression and the genetic background (ANOVA, $p < 0.01$).

Expression QTL analysis in introgression lines

The strains were also assayed for gene expression after heat-shock by microarray. Using eQTL data generated on *C. elegans* [38], we were able to confirm both the genetic backgrounds and the presence of the introgressions (**Supplementary file 3**). When we explained the gene expression over genetic background, introgression, and the interaction between the two, we found that the largest number of genes are affected by the genetic background (**Table 1**, **Supplementary file 4**). This is as expected given the large number of genes polymorphic between N2 and CB4856, genes differently expressed between N2 and CB4856 [45], and previously detected eQTLs [38, 46]. Via enrichment analysis we found that indeed those genes are the ones affected, e.g *math*, *bath*, and *fbx* genes (**Supplementary file 5**) [4]. Not many genes were affected by the introgression or the interaction between background and introgression (< 100 in both conditions at FDR = 0.05, **Table 1**). Still, some genes displayed patterns that indicate the presence of *cis*-eQTL (such as C23H5.8), or even an interaction with the genetic background (*clec-62*, *dod-21*, and *scrm-4*), see **Figure 4**. The expression levels of C23H5.8 are mainly affected by the introgression on chromosome IV. This gene lies on the introgression under study (2.1Mb) and is polymorphic between N2 and CB4856 [4]. However, the other three genes are not located near the introgression, and are therefore affected *in trans*. Furthermore, these genes display an expression pattern that is determined by both the introgression and the genetic background. Especially the *clec-62* is a strong example of a genetic interaction between the background and the introgression as the extreme phenotype is found in the crossing population.

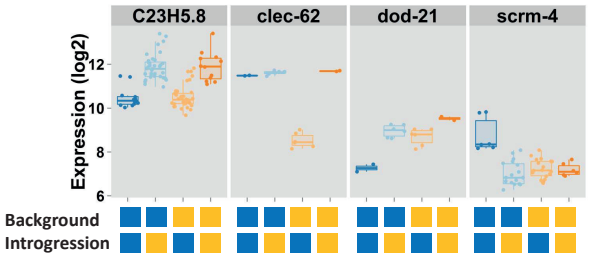


Figure 4. Four examples of genes affected by the introgressions (FDR = 0.05), without treatment playing a role. C23H5.8 displays characteristics of a *cis*-eQTL, where the introgression determines the expression levels. The other three genes: *clec-62*, *dod-21*, and *scrm-4* are examples of interactions between the background and the introgression; the combination of background and introgression determines the expression levels.

Table 1. The number of genes affected per condition (FDR = 0.05), and the variation explained (R^2).

	Affected genes	R^2 (min)	R^2 (median)	R^2 (max)
Control				
Background	1873	0.06	0.92	1.00
Introgression	93	0.14	0.43	0.86
Background x Introgression	85	0.19	0.41	0.78
Heat-shock				
Background	1909	0.06	0.91	1.00
Introgression	40	0.10	0.42	0.70
Background x Introgression	39	0.32	0.43	0.95

Discussion

In this paper we presented a novel whole-genome IL population in *C. elegans*. This new N2-in-CB4856 IL population is complementary to the previous CB4856-in-N2 population constructed by our group [19]. The construction of this novel population was straightforward due to the availability of a high-quality CB4856 genome [4], allowing for the selection of insertions/deletions as genetic markers. These markers provide a cost-effective and fast alternative to SNP-markers, making high-throughput genotyping and therefore several rapid rounds of selection for introgressions possible. However, the limited set of markers used provides a poor genetic map and has to be complemented with low-coverage whole genome sequencing to construct a genetic map with an increased resolution. For example, with the current markers, it cannot be ruled out that additional introgressions exist in-between the markers.

The crossing scheme used to obtain the IL population depended on results obtained from quantitative genetics studies in *C. elegans*. The scheme accounted for the *peel-1*/*zeel-1* locus [30] leading to marker distribution distortions in N2xCB4856 RIL populations [4, 23, 24]. A double back-cross with CB4856 was used to remove the N2 *peel-1/zeel-1* allele, which failed to segregate otherwise (data not shown). As the background of the developed IL population is CB4856, it lacks the N2 laboratory derived alleles, such as *nath-10* [47], *glb-5* [11], and *npr-1* [6, 11]. Therefore, these strains may be especially useful if the studied trait might be affected by these pleiotropic and large-effect alleles [**Chapter 2**].

For many species genome-wide IL populations have now been constructed, including barley [18], *Arabidopsis* [17, 20], tomato [48], maize [49], rice [50], and mice [51]. Our N2-in-CB4856 IL population will likely be useful for fine-mapping of complex traits. The population contains novel breakpoints compared to the CB4856-in-N2 population, which will aid in pinpointing a QTL to a more narrow locus. Furthermore, our new IL population can serve as a resource for the generation of ILs with smaller introgressions. One of the main strengths of IL populations in comparison to RIL populations lies in the detection of small effects. RIL populations are hampered by the residual variance induced by the segregation of multiple QTL [17]. In strains with an homogeneous background all QTL are fixed, except the introgressed locus, reducing the residual variance per QTL to a minimum (the experimental variation). In many studies it has been observed that ILs resolve more QTL than RILs [17, 21, 51, 52]. However, one of the pitfalls of ILs is the homogeneous background; since all QTL are in context of a single background. This leads to different effect size (or direction) estimations compared to RIL estimates.

Complementary IL populations can place the QTL effect in the context of the genetic background. Although there are many IL populations, to our knowledge there are no genome-wide complementary IL populations. In an IL panel with an homogeneous background, the QTL-background interaction is confounded by definition. The homogenous background can lead to different estimations of QTL effect sizes compared to RILs (as reviewed [53]). The cause of this

effect is due to the frequency of the genotype at the second locus. In an ideal RIL population the loci are unlinked and therefore both genotypes affect the main effect at the QTL. However, in ILs, the loci are linked and the QTL main effects are therefore affected by the background interaction [11, 25, 54]. All these examples come from ILs generated for specific loci. The availability of two complementary IL populations makes QTL dissection on a larger scale possible.

There are still some analytical challenges to overcome in mapping using the combined IL populations. The primary obstacle is the lack of perfect complementary between both populations. The complementarity issue is hard to solve since it is practically impossible to make the exact complementary ILs. The effect of this is that it will be difficult to distinguish between closely linked loci and interacting loci. Good examples of such traits in *C. elegans* are olfactory preference for *Serratia marcescens* and dauer larvae formation which are regulated by multiple QTL, some closely linked [21, 22]. It is still hard to predict the power of the combined populations, since the current genetic map is to coarse to determine the individual breakpoints with any accuracy. Once a high-resolution genetic map is available, it will be possible to determine the added value of the individual population and the power of the combined populations.

We provide two examples of trait mapping using strains from both IL populations, one involved transcriptome analysis. It should be mentioned here that the absence of a high resolution genetic map could be mitigated by use of expression QTL. The eQTL comparison with Rockman [38] shows that the strains have the expected background-effects and possess introgressions at the expected locations. Although most *cis*-eQTL matched up with expectations, only sequencing will resolve whether the strains carry any additional introgressions. Several eQTL were detected, either caused by the introgressions or an interaction between background and introgression. However, to detect the effects more reliably, the power of the experiments can be increased by adding more replications per strain, rather than imposing a single overlapping locus. When executed fully, expression analysis over both panels would provide tremendous insight into complex trait architectures. The appeal of expression lies in the unbiased choice of traits and the variation in the degree of complexity over these traits. Therefore, this trait qualifies for future exploration.

The second experiment was survival after heat-shock. The chromosome IV locus was implicated in previous experiments and a complex architecture was suspected (see **Chapter 6**) [26]. The same strains as for the transcriptome analysis were assayed and the results demonstrate that a QTL-background interaction is likely. In particular, a QTL for survival at 11 days after heat-shock depends on such an interaction. This experiment can also be replicated further, the survival trait will be explored over all ILs (of both populations).

These results provide a first insight in the influence of background interactions on trait variation, which are thought to be pervasive [51, 52, 55]. As discussed above, from many studies it became clear that ILs capture additional or different QTL compared to RIL populations. The complementary IL populations will place QTL in their respective background and capture both extremes genotypic extremes of the background influence.

Conclusion

This paper promises considerable advances for QTL mapping with introgression lines in N2xCB4856 crosses. The novel N2-in-CB4856 introgression lines are an important addition to the already existing CB4856-in-N2 population. The combined populations can provide insight into trait variation contributed by loci-background interactions. To reach this point, two important steps need to be made: (i) the population needs to be sequenced, and (ii) the complete combined populations need to be measured for traits. We plan to study life-history related traits, such as pumping, heat-shock survival, and dauer formation, for which the CB4856-in-N2 population has already been studied exhaustively. Upon completion of these goals, the combination of both panels can resolve complex trait architecture in relation to the genetic background.

Acknowledgements

The authors thank Jasmijn Schouten, Lisa van Sluijs, and Myrthe Walhout for technical support.

Author contributions

Conceived and designed the experiments: MGS, LBS, JEK. Performed the experiments: MGS, JWVK, JAGR, BT. Analysed the data: MGS. Wrote the paper: MGS, LBS, JEK.

Supplementary figures and files

The supplementary files and figures are deposited at: http://marksterken.nl, under 'PhD thesis'.

Supplementary figure 1: The electrophoresis patterns of the insertion/deletion markers. The markers are organized per chromosome and location. For each marker the two parental strains (N2 and CB4856) and the six recombinant inbred lines used for crossing are shown.

Supplementary file 1: The genetic map of the N2-in-CB4856 introgression lines, as determined by the insertion/deletion markers. The genotypes are as follows: CB4856 (0), N2 (2), heterozygous (1). If the genotype is inferred from a previous generation .1 is added, for example: 0.1 means that the genotype is CB4856 as determined in a previous generation.

Supplementary file 2: The primer sequences for the insertion/deletion markers.

Supplementary file 3: (tab 1, overview) The predicted introgression locations for the introgression lines used in the transcriptomics experiment. (**tab 2, cisQTL_correlations**) the correlations of the gene expression with the *cis*-eQTL from [38]. The p-value given is in –log10(p), and the R-squared is the variance explained by ANOVA.

Supplementary file 4: The outcome from the full interaction model per treatment and for the combination of the two treatments. Only the spots with significant differences are listed.

Supplementary file 5: Enrichment analysis of the significantly differentially expressed genes, as determined by the full interaction model. The annotation group is given (*e.g.* Anatomy or Gene class), the group (*e.g. btb*-genes within Gene class), the number of genes in the group, the overlap with the differentially expressed genes, and the significance of the overlap (in –log10(p)).

References

1. **Gaertner, B.E. and P.C. Phillips,** *Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation.* **Genet Res (Camb), 2010. 92(5-6): p. 331-48.**
2. **Andersen, E.C., et al.,** *Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nat Genet, 2012. 44(3): p. 285-90.**
3. **Volkers, R.J., et al.,** *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93.**
4. **Thompson, O.A., et al.,** *Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856.* **Genetics, 2015. 200(3): p. 975-89.**
5. **de Bono, M. and C.I. Bargmann,** *Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans.* **Cell, 1998. 94(5): p. 679-89.**
6. **Andersen, E.C., et al.,** *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156.**
7. **Seidel, H.S., et al.,** *A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in C. elegans.* **PLoS Biol, 2011. 9(7): p. e1001115.**
8. **Kammenga, J.E., et al.,** *A Caenorhabditis elegans wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3.* **PLoS Genet, 2007. 3(3): p. e34.**
9. **Palopoli, M.F., et al.,** *Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans.* **Nature, 2008. 454(7207): p. 1019-22.**
10. **Persson, A., et al.,** *Natural variation in a neural globin tunes oxygen sensing in wild Caenorhabditis elegans.* **Nature, 2009. 458(7241): p. 1030-3.**
11. **McGrath, P.T., et al.,** *Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors.* **Neuron, 2009. 61(5): p. 692-9.**
12. **Bendesky, A., et al.,** *Catecholamine receptor polymorphisms affect decision-making in C. elegans.* **Nature, 2011. 472(7343): p. 313-8.**
13. **Ghosh, R., et al.,** *Natural variation in a chloride channel subunit confers avermectin resistance in C. elegans.* **Science, 2012. 335(6068): p. 574-8.**
14. **Bendesky, A., et al.,** *Long-range regulatory polymorphisms affecting a GABA receptor constitute a quantitative trait locus (QTL) for social behavior in Caenorhabditis elegans.* **PLoS Genet, 2012. 8(12): p. e1003157.**
15. **Tijsterman, M., et al.,** *RNA helicase MUT-14-dependent gene silencing triggered in C. elegans by short antisense RNAs.* **Science, 2002. 295(5555): p. 694-7.**
16. **Glauser, D.A., et al.,** *Heat avoidance is regulated by transient receptor potential (TRP) channels and a neuropeptide signaling pathway in Caenorhabditis elegans.* **Genetics, 2011. 188(1): p. 91-103.**
17. **Keurentjes, J.J., et al.,** *Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.* **Genetics, 2007. 175(2): p. 891-905.**
18. **Schmalenbach, I., N. Korber, and K. Pillen,** *Selecting a set of wild barley introgression lines and verification of QTL effects for resistance to powdery mildew and leaf rust.* **Theor Appl Genet, 2008. 117(7): p. 1093-106.**
19. **Doroszuk, A., et al.,** *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110.**
20. **Fletcher, R.S., et al.,** *Development of a next-generation NIL library in Arabidopsis thaliana for dissecting complex traits.* **BMC Genomics, 2013. 14: p. 655.**
21. **Glater, E.E., M.V. Rockman, and C.I. Bargmann,** *Multigenic natural variation underlies Caenorhabditis elegans olfactory preference for the bacterial pathogen Serratia marcescens.* **G3 (Bethesda), 2014. 4(2): p. 265-76.**
22. **Green, J.W., et al.,** *Highly Polygenic Variation in Environmental Perception Determines Dauer Larvae Formation in Growing Populations of Caenorhabditis elegans.* **PLoS One, 2014. 9(11): p. e112830.**
23. **Li, Y., et al.,** *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222.**
24. **Rockman, M.V. and L. Kruglyak,** *Recombinational landscape and population genomics of Caenorhabditis elegans.* **PLoS Genet, 2009. 5(3): p. e1000419.**

25. **Gaertner, B.E., et al.,** *More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in Caenorhabditis elegans.* **Genetics, 2012. 192(4): p. 1533-42.**
26. **Rodriguez, M., et al.,** *Genetic variation for stress-response hormesis in C. elegans lifespan.* **Exp Gerontol, 2012. 47(8): p. 581-7.**
27. **van der Velde, K.J., et al.,** *WormQTLHD--a web database for linking human disease to natural variation data in C. elegans.* **Nucleic Acids Res, 2014. 42(Database issue): p. D794-801.**
28. **Brenner, S.,** *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94.**
29. **Altun, Z.F., Herndon, L.A., Crocker, C., Lints, R. and Hall, D.H.** *Wormatlas***. 2002-2016.**
30. **Seidel, H.S., M.V. Rockman, and L. Kruglyak,** *Widespread genetic incompatibility in C. elegans maintained by balancing selection.* **Science, 2008. 319(5863): p. 589-94.**
31. **Untergasser, A., et al.,** *Primer3--new capabilities and interfaces.* **Nucleic Acids Res, 2012. 40(15): p. e115.**
32. **Altschul, S.F., et al.,** *Basic local alignment search tool.* **J Mol Biol, 1990. 215(3): p. 403-10.**
33. **Vervoort, M.T., et al.,** *SSU ribosomal DNA-based monitoring of nematode assemblages reveals distinct seasonal fluctuations within evolutionary heterogeneous feeding guilds.* **PLoS One, 2012. 7(10): p. e47555.**
34. **Hosono, R.,** *Sterilization and growth inhibition of Caenorhabditis elegans by 5-fluorodeoxyuridine.* **Exp Gerontol, 1978. 13(5): p. 369-74.**
35. **Snoek, L.B., et al.,** *A rapid and massive gene expression shift marking adolescent transition in C. elegans.* **Sci Rep, 2014. 4: p. 3912.**
36. **Zahurak, M., et al.,** *Pre-processing Agilent microarray data.* **BMC Bioinformatics, 2007. 8: p. 142.**
37. **Smyth, G.K. and T. Speed,** *Normalization of cDNA microarray data.* **Methods, 2003. 31(4): p. 265-73.**
38. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak,** *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6.**
39. **Benjamini, Y. and Y. Hochberg,** *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* **Journal of the Royal Statistical Society Series B-Methodological, 1995. 57(1): p. 289-300.**
40. **Stein, L., et al.,** *WormBase: network access to the genome and biology of Caenorhabditis elegans.* **Nucleic Acids Research, 2001. 29(1): p. 82-86.**
41. **Yook, K., et al.,** *WormBase 2012: more genomes, more data, new website.* **Nucleic Acids Res, 2012. 40(Database issue): p. D735-41.**
42. **Ogata, H., et al.,** *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Res, 1999. 27(1): p. 29-34.**
43. **Niu, W., et al.,** *Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans.* **Genome Res, 2011. 21(2): p. 245-54.**
44. **Gerstein, M.B., et al.,** *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* **Science, 2010. 330(6012): p. 1775-87.**
45. **Capra, E.J., S.M. Skrovanek, and L. Kruglyak,** *Comparative developmental expression profiling of two C. elegans isolates.* **PLoS One, 2008. 3(12): p. e4055.**
46. **Vinuela, A., et al.,** *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37.**
47. **Duveau, F. and M.A. Felix,** *Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans.* **PLoS Biol, 2012. 10(1): p. e1001230.**
48. **Monforte, A.J. and S.D. Tanksley,** *Development of a set of near isogenic and backcross recombinant inbred lines containing most of the Lycopersicon hirsutum genome in a L. esculentum genetic background: a tool for gene mapping and gene discovery.* **Genome, 2000. 43(5): p. 803-13.**
49. **Szalma, S.J., et al.,** *QTL mapping with near-isogenic lines in maize.* **Theor Appl Genet, 2007. 114(7): p. 1211-28.**
50. **Li, Z.K., et al.,** *Genome-wide introgression lines and their use in genetic and molecular dissection of complex phenotypes in rice (Oryza sativa L.).* **Plant Molecular Biology, 2005. 59(1): p. 33-52.**
51. **Gale, G.D., et al.,** *A genome-wide panel of congenic mice reveals widespread epistasis of behavior quantitative trait loci.* **Mol Psychiatry, 2009. 14(6): p. 631-45.**
52. **Edwards, A.C. and T.F. Mackay,** *Quantitative trait loci for aggressive behavior in Drosophila melanogaster.* **Genetics, 2009. 182(3): p. 889-97.**

53. **Mackay, T.F.,** *Epistasis and quantitative traits: using model organisms to study gene-gene interactions.* **Nat Rev Genet, 2014. 15(1): p. 22-33.**
54. **Kroymann, J. and T. Mitchell-Olds,** *Epistasis and balanced polymorphism influencing complex trait variation.* **Nature, 2005. 435(7038): p. 95-8.**
55. **Shao, H., et al.,** *Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis.* **Proc Natl Acad Sci U S A, 2008. 105(50): p. 19910-4.**

Chapter 4

A heritable antiviral RNAi response limits Orsay virus infection in *Caenorhabditis elegans* N2

Mark G. Sterken,
L. Basten Snoek, Kobus
J. Bosman, Jikke Daamen,
Joost A.G. Riksen,
Jaap Bakker,
Gorben P. Pijlman,
Jan E. Kammenga

This chapter has been published in a modified form:

Sterken, M.G.; Snoek, L.B.; Bosman, K.J.; Daamen, J.; Riksen, J.A.G.; Bakker, J.; Pijlman, G.P.; Kammenga, J.E. (2014). A heritable antiviral RNAi response limits Orsay virus infection in *Caenorhabditis elegans* N2. *PLOS ONE* 9(2): e89760.

Abstract

Orsay virus (OrV) is the first virus known to be able to complete a full infection cycle in the model nematode species *Caenorhabditis elegans*. OrV is transmitted horizontally and its infection is limited by antiviral RNA interference (RNAi). However, we have no insight into the kinetics of OrV replication in *C. elegans*. We developed an assay that infects worms in liquid, allowing precise monitoring of the infection. The assay revealed a dual role for the RNAi response in limiting Orsay virus infection in *C. elegans*. Firstly, it limits the progression of the initial infection at the step of recognition of dsRNA. Secondly, it provides an inherited protection against infection in the offspring. This establishes the heritable RNAi response as anti-viral mechanism during OrV infections in *C. elegans.* Our results further illustrate that the inheritance of the anti-viral response is important in controlling the infection in the canonical wild type Bristol N2. The OrV replication kinetics were established throughout the worm life-cycle, setting a standard for further quantitative assays with the OrV-*C. elegans* infection model.

Introduction

The nematode *Caenorhabditis elegans* is an important model species for human biology. Research on this roundworm contributed to the understanding of cancer [1], aging [2], development [3], physiology [4] and the immune system [5]. Recently, a virus able to naturally infect this nematode was discovered, which infects the intestine inducing abnormal intestinal morphology [6, 7]. The causative agent, Orsay virus (OrV), was identified as a plus-strand RNA virus and putative member of the family *Nodaviridae*. The closest relatives to OrV are the co-discovered Santeuil virus [7] and Le Blanc virus [8], both of which infect *C. briggsae*. OrV persistently infects the *C. elegans* wild isolate JU1580 and is horizontally transmitted through the population in the laboratory [7].

The discovery of OrV coincides with increased sampling efforts and studies to expand the knowledge about natural variation in *C. elegans* [9-11]. Studying genotype-phenotype relations enhance our understanding of the ecological niche of *C. elegans* [10, 12, 13]. It is clear that this nematode thrives on decaying organic material. These short-lived and nutrient-rich environments mean that populations have the intrinsic property to grow fast [14]. It also places *C. elegans* in a complex web of inter-species interactions, where pathogens will often be encountered [10, 14, 15]. The variation in susceptibility to OrV in different genotypes is particularly interesting, as these can either be the result of adaptation from the side of the host (antiviral responses) or the virus (immune suppression). The availability of genetic variation in *C. elegans* can be combined with the powerful molecular tools also available for this model organism [16, 17].

The ability of OrV to complete a full replication cycle within its natural host *C. elegans* enables detailed studies on virus-host interactions [7]. This will lead to a better understanding of host specificity and identification of crucial genetic factors determining host susceptibility and/or resistance to viruses. In particular, the RNA interference (RNAi) response [18] plays a crucial role in the antiviral immune response of *C. elegans* [19-21]. Furthermore, the importance of antiviral RNAi is underscored by the fact that it is transmitted to the next generation, likely providing an advantage to the population as a whole [22]. Potent RNAi activity against OrV has also been observed in the canonical *C. elegans* N2 Bristol strain but less so in the natural isolate JU1580 [7]. This has been attributed to a mutation in the gene *drh-1* (a RIG-I like helicase), likely involved in the recognition of non-self RNA, including viral RNA [23-26]. However, nothing is known about the effects of mixed populations, prolonged virus exposure and the relative contribution of the trans-generationally inherited RNAi response. These unknown factors may also contribute to the progression of the infection.

To provide a deeper understanding of natural OrV infection, we set out to develop an infection procedure, which i) exposes *C. elegans* to OrV in liquid, ii) uses a defined viral dose, iii) times the exposure to infectious virus, and iv) enables larval stage-dependent infection kinetic studies. Using this procedure, we show the existence of genotype-dependent differences in the progression of OrV infection between *C. elegans* strains. In addition, a heritable RNAi response in the canonical N2 strain is identified as an important antiviral mechanism to convey resistance to its offspring.

Materials and methods

C. elegans culturing

The strains N2, JU1580, WM29 (*rde-2*, ne221), and WM49 (*rde-4*, ne301) were kept at 12°C on 6 cm Petri dishes containing Nematode Growth Medium (NGM), seeded with *Escherichia coli* strain OP50 [27]. Before onset of the experiments, single worms were picked of each genotype and grown into a new population. They were grown at 20°C on 9 cm dishes. For synchronization the populations were bleached [28]. A virus free JU1580 population was created by bleaching an infected strain of JU1580 [7]. This uninfected strain was used as starting material for the experiments.

Generating stocks of Orsay virus

Orsay virus was isolated from persistently infected nematodes of strain JU1580. The nematodes were kept at 16°C on NGM plates and transferred to new NGM plates every 14 days. Virus stocks were generated by isolation from *C. elegans* as previously described by [7]. PBS was used for isolating the virus from the worms. Virus stocks were flash frozen in liquid nitrogen and stored at -80°C. Before use in experiments the stocks were tested by infecting virus free JU1580 with different volumes (1, 10, 50, and 100 µL) and using an RT-PCR to confirm the establishment of viral replication. This yielded an estimate for the infectious dose, as no other assays are available.

Infection procedure

Before infection the *C. elegans* strains were synchronized. Populations were grown at 20°C until the desired larval stage was reached: 20h for L1, 26h for L2, 40h for L3, or 48h for L4. To infect the synchronized population, the worms were collected by rinsing the plate with M9 buffer and centrifuged shortly to pellet the worms. Thereafter the M9 buffer was removed and 500 µL of infection solution (370 µL of M9, 30 µL of virus stock and 100 µL of OP50 in LB) or mock solution (400 µL of M9 and 100 µL of OP50 in LB) was added. The worms were incubated in infection solution for 1h in Eppendorf tubes at room temperature and regularly mixed to infect them with OrV. Next the worms were pelleted by centrifugation and the supernatant was removed. The worms were washed three times with 1 mL of M9 buffer to remove virus from the supernatant and thereafter plated on a fresh NGM plate containing OP50.

RNA isolation and RT-qPCR

The RNA of infected *C. elegans* was isolated using the QIAGEN RNeasy Micro kit, following the prescribed protocol. cDNA was made using the SuperScript III kit from Invitrogen following the prescribed protocol with random hexanucleotides. Per RT-reaction 1 µg of isolated RNA

was used. For the qPCR reaction the cDNA was diluted 1/50 and qPCR was performed with Absolute QPCR SYBR Green Fluorescein Mixes (Thermo scientific). Viral RNA was detected using two primer pairs, both annealing to the start of the RNA1 coding region (HM030970.1) (pOrV-RNA1.1F: 5'ATACTCTACGACCTTGTCGG 3', pOrV-RNA1.1R: 5'CTCGGTTGATGTTCTTCCAG 3', pOrV-RNA1.2F: 5'AACCAGGAAACACTACTCCG 3', pOrV-RNA1.2R: 5'GTTGTGATATCGCTTGGTGG 3'). Two reference genes (Y37E3.8 and *rpl-6*) were selected based on stable expression, even during stress condition, in transcriptomics data generated by microarray (pY37E3.8F: 5'GCGTTTGTGGTCTCTTGTC 3', pY37E3.8R: 5'CTCTGGGAGGAGTCCTTTTC 3', pRPL6-F: 5'TGTCACTCTCCGCAAGAC 3', pRPL6-R: 5'TGATCTTGTGTGGTCCAGTG 3').

The primer pairs were designed for an optimal annealing temperature of 62°C [29], which was verified by testing the primers on a temperature gradient. Furthermore, specificity was checked by measuring the melting curve and efficiency by performing an RT-qPCR reaction on serially diluted template. The primer pairs do not generate unspecific products within 40 cycles and the measured efficiency was in-between 90% and 110% (100% being the product doubles every cycle).

Data normalization

Between biological replicates the RT-qPCR data was checked using the reference genes, outliers per biological replicate were identified as having Ct values for the reference genes that fell outside μ +/- 2*σ of all the measurements. Outlier samples typically showed low RNA quality (e.g. partial degradation or contaminations). The expression of the reference genes was checked for genotype and larval stage effects. This to ensure that differences found were not the result from different reference gene expression (**Supplementary file 1**).

The data was transformed with

$$Q_{gene} = 2^{40.3\text{-}Ct_{gene}}$$

where Q is the expression of the gene and Ct is the measured Ct value of the gene. The number 40.3 indicates the level of the 5% highest Ct values in mock infected samples (based on 104 mock infected samples). Thereafter the viral measurements were normalized to the expression of the control genes

$$E = \frac{Q_V}{0.5 * ((Q_{rpl6} / \overline{Q}_{rpl6}) + (Q_{Y37E3.8} / \overline{Q}_{Y37E3.8}))}$$

where E is the relative expression, Q is the transformed expression (v indicates either one of the viral genes, *rpl-6* and *Y37E3.8* are reference genes). All data was normalized together, to allow for direct comparison.

Statistical analysis

Statistical tests were executed using custom written scripts in R (version x64 2.13.1, www.r-project.org). Pairwise testing was done using a two-sample independent t-test not assuming equal variances (Welch's *t*-test), as provided by R. Testing over multiple samples was done by ANOVA, as provided by R.

Logistic curve fitting was performed using a non-linear model, fitting to a basic logistic curve with the function

$$E_t = \frac{A}{1 + e^{(B-t)/C}}$$

where E is the relative expression at time point t, A is the fitted upper asymptote, B is the fitted inflection point and C is the lower asymptote. To be able to fit the function, the relative expression had to be transformed so the lower asymptote approached 0, otherwise the SSlogis function [30] in R was not able to correctly estimate the inflection point. Confidence intervals were calculated using the predict.nls function of R. The goodness of fit was calculated as

$$R^2 = \frac{\sum_t (E_t - \overline{E})^2}{\sum_t (E_t - P_t)^2}$$

where R^2 is the coefficient of determination, E is the relative expression at time point t, and P is the predicted value at time point t.

Multiple testing over the descriptive values obtained from the sigmoidal curve fitting was done using linear regression, by fitting the data to

$$F_{i,j} = L_i + S_j$$

where F is the descriptive value obtained from the sigmoidal curve fitting, L is the larval stage i (L1, L2, L3, or L4), and S is the strain j (JU1580 or N2). When testing within one strain, the model was simplified by excluding the strain as an influence.

Results

Worm stage affects OrV infection development in *C. elegans* JU1580

First, the exposure time needed to infect a population was established by incubation of JU1580 with OrV in liquid. Liquid infections have the following advantages: the dosage is the same for every worm, the infection is timed, it can be executed at a defined larval stage, and the worms can be washed to remove the non-internalized virions. A starved mixed stage population (mainly adults and L2 present) of JU1580 was exposed to OrV in infection solution for 0.5, 1, 2 and 4h. Relative viral loads were measured at 48 h post infection by quantitative (q)PCR. The results show that OrV infections can be established following 1h exposure (**A**) and do not increase with longer exposure to the virus (ANOVA, P = 0.70).

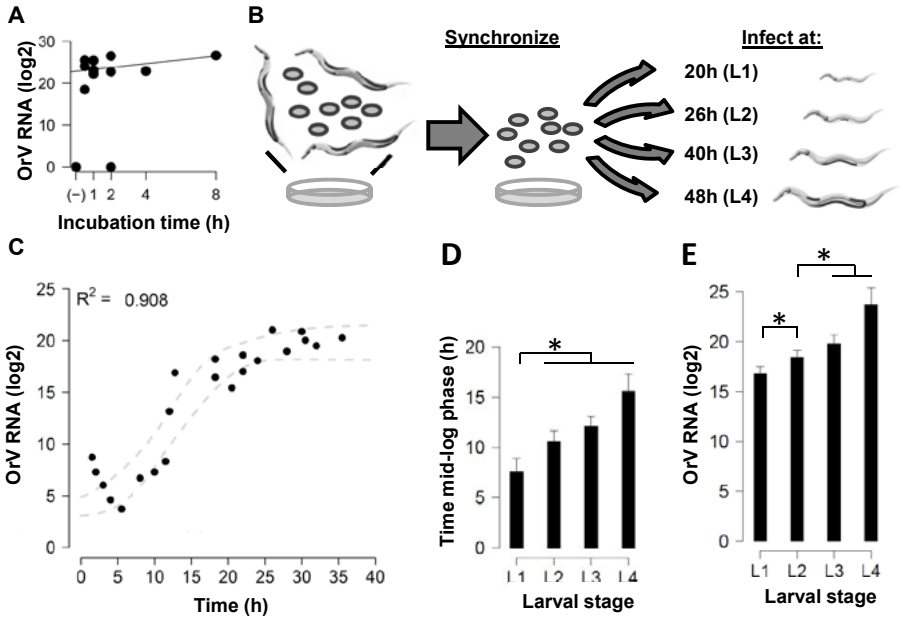


Figure 1: Infections in genotype JU1580. **(A)** JU1580 was exposed to 30µL virus in solution for 0.5, 1, 2, 4 and 8h. The relative viral load measured by RT-qPCR 48h post infection is shown. There are no differences between the viral loads after different exposure times (ANOVA, P=0.70). **(B)** The design of the experiments to measure the infection progress in different larval stages. First populations were synchronized by bleaching and subsequently grown until the desired larval stage was reached. At the indicated time points the larvae were infected by exposure to virus for 1h, after which the infection was allowed to develop and at different time points populations were isolated. **(C)** An example of the data, the outcome of the experiments performed in JU1580 when infected in the L3 stage (bullet points indicate sample points used for the fit). The dashed grey line indicates the SD around the fitted curve. **(D)** The time needed to reach the inflection point as determined by the curve fit is shown for JU1580. There is a significant difference between L1 and the other three larval stages (Two-sided *t*-test, P≤0.05). **(E)** The maximum viral load reached (the asymptote of the curve) in JU1580. Here the load of L1 versus the other three larval stages (Two-sided *t*-test, P≤0.05) and L2 versus the other larval stages (Two-sided *t*-test, P≤0.05) is significantly lower.

To investigate the relative susceptibility of *C. elegans* larval stages to OrV, virus infections were carried out in JU1580 synchronized populations of L1, L2, L3, and L4 larval stages (**Figure 1B**). An example of the data retrieved and a sigmoidal-curve fit for JU1580 infected at the L3 stage is shown in **Figure 1C** (curves for L1, L2 and L4 are shown in **Figure S1)**. For all stages the fit explained >80% of the variation. Within the first 3h after exposure, the viral levels decreased in all larval stages (see also **Figure S1**). Since the route of infection is oral uptake and the infection takes place in the intestine [7], this initial decrease most likely represents an overload of virus which is leaving the intestine but still measurable by qPCR. After the initial decrease a steady level (lag phase) is reached after which replication starts (log phase).The speed at which the infection develops is dependent on the larval stage (**Figure 1D**). In L1 larvae the inflection point (mid-log phase) is reached after 7.5h, whereas in the other stages it takes >10.5h (**Figure 1D**, significant difference, Two-sided *t*-test, P≤0.05). The maximum viral load is larval-stage dependent; significantly higher levels are reached in older larvae (**Figure 1E**). For instance, the difference in maximum viral load is 6.9 log2 units (>100-fold) between L1 versus L4 (Two-sided *t*-test, P≤0.05). In conclusion, in older larvae the OrV infection progresses slower, but reaches higher maximum viral loads.

Comparative infections in *C. elegans* JU1580 and N2

We first investigated if the reported difference in susceptibility between the two wild types JU1580 and N2 [7] could be the result of differences between the dose-response relationship. The effect of viral dose was studied in the two genotypes, using an exposure time of 1h (**Figure 2A**). Since it was previously found that after long-term infection, *C. elegans* N2 strain displayed ~100-fold lower viral levels compared to the wild strain JU1580 [7], we expected to see a comparable difference. Genotype JU1580 could be infected with a smaller dose to reach maximum viral load levels compared to N2. JU1580 could be infected by a dose as little as 10µL of virus stock in 80% of the experiments, whereas N2 was only infected (at a very low level) in 33% of the experiments in which the dose-response was determined. However when more virus (>30 µL) was used, N2 was productively infected as well, and the differences became smaller, up to the point that JU1580 and N2 had comparable infection levels (ANOVA, P = 0.09). Concluding, JU1580 and N2 are comparable in susceptibility when exposed to higher viral doses.

The dose-response experiments showed that there was no difference between maximum viral levels in JU1580 and N2, provided that larvae were exposed to sufficient amounts of OrV. This puts earlier findings [7] in a new perspective. Previously it was shown that N2 was less susceptible than JU1580 to OrV infection. The main differences between our experiments and the latter are the mode of infection (in liquid vs. on agar), the exposure time, and the population dynamics during the experiment. Viral loads in our experiments were measured up to 48h post infection as compared to 4-7d after exposure. In addition, our experiments predominantly involve virus infections within a single generation, in contrast to experiments on infected adults which in turn spread the infection to their offspring. Therefore, the apparent discrepancy might originate from these differences, exposure duration and/or re-infection.

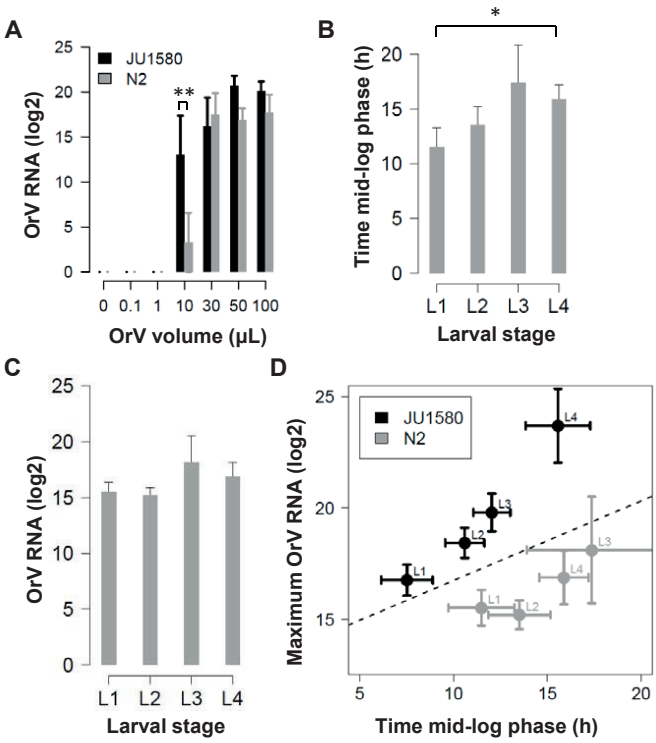


Figure 2: Infections in N2 compared with JU1580. **(A)** JU1580 and N2 were exposed to different amounts of virus (0, 0.10, 1.0, 10, 50, and 100µL) in infection solution for 1h. The relative viral load measured by RT-qPCR 48h post infection is shown (+/- SEM), for 3 independent repeats in duplo per experiment (except 10, 50 and 100µL, which were replicated 5 times in JU1580). JU1580 reaches maximum viral load after exposure to 10µL (ANOVA, P>0.05), whereas N2 after exposure to 30µL (ANOVA, P>0.05). **(B)** The time needed to reach the inflection point as determined by the curve fit is shown for N2. There is a small delay (approximately 4h) in reaching the mid-log phase between L3 and L4 versus L1 (Two-sided *t*-test, P≤0.05). **(C)** The maximum viral load reached (the asymptote of the curve) for N2. Here there are no larval-stage dependent effects (*t*-test, P>0.05). **(D)** A comparison of the values obtained for JU1580 (figure 2) and N2. The dashed line is a fitted linear function used to compare the maximum viral load and mid-log phase, showing that JU1580 has higher viral loads and/or a faster developing infection (ANOVA, P≤0.01).

To further compare the two genotypes our infection experiment as described in **Figure 1B** was also conducted in N2 in all four larval stages. The data obtained from the curve fits can be seen in **Figure 2B and C** (the separate curves showing all data points can be found in **Supplementary Figure 1**). Like JU1580, the curve fits for N2 explained >80% of the variation. On average, the explained variation was a bit lower in N2 compared to JU1580. In N2 the mid-log phase of OrV infection is reached earlier in younger larvae and ~4.5h faster in L1 versus L4 (two-sided *t*-test, P≤0.05). Maximum viral loads were similar in all N2 larval stages (two-sided *t*-test, P>0.05), in contrast to JU1580 for which there was an age-dependent increase in maximum viral load.

The initial infection kinetics of JU1580 and N2 were also compared (**Figure 2D**). It was found that, in general, the infection progresses at an equal pace in JU1580 and N2 (two-sided *t*-test, P≤0.05). The maximum viral load however, was significantly different between JU1580 and N2 for the larval stages L2 and L4 (P≤0.05). Overall, there was a trend that the maximum load in JU1580 is 10-fold higher (3.3 log2 units). However, these differences are not large enough to be the main source of the previously reported 100 fold difference in OrV infection levels between JU1580 and N2 populations [7]. The difference in maximum load between the two genotypes in the experiments described in this paper arises mostly in the L4 stage. In our experiment the worms were infected at an age of 48 hours, and the last hours of the experiment the first larvae are observed. Since we show that N2 and JU1580 are equally susceptible (not-previously exposed and at high viral dosages) and Félix *et al.* (2011) found a 100-fold difference in maximum viral load between the two after multiple generations we suspected that mechanisms, like the inheritance of an antiviral response, could play a role.

The antiviral RNAi response suppresses the progression of infection in *C. elegans* N2

Since the antiviral RNAi response was shown to be an important factor in OrV replication [7], we investigated whether or not this response influences the initial infection. In parallel with the experiments described in **Figure 1B**, experiments were carried out in L3 larvae of *rde-2* and *rde-4* mutants in an N2 background. RDE-2 is involved in heritable silencing of RNA only and functions after initiation of the original RNAi response [31] and functions in concert with *mut-7* [32]. RDE-4 is involved in siRNA production from exogenous dsRNA, in concert with DCR-1. Mutants of *rde-4* are impaired in siRNA production and thus cannot initiate antiviral RNAi nor pass on to their offspring the heritable silencing signals [33-35]. Both the *rde-2* and *rde-4* mutants were shown to increase OrV infection to the level observed in JU1580 [7].

Both *rde-2* and *rde-4* mutants displayed a JU1580 phenotype regarding maximum viral load reached (**Figure 3, Supplementary Figure 1**), however compared to N2 these differences were not significant due to a larger variation in N2 (Two-sided *t*-test, P>0.05 in both cases). Importantly, the time until the mid-log phase was similar between the *rde-2* and N2 (Two-sided *t*-test, P>0.05). Whereas in the *rde-4* mutant the infection developed significantly faster than in N2 (Two-sided *t*-test, P≤0.05), showing that the initial recognition and cleavage of dsRNA is an important step in the progression of the infection, and therefore in the antiviral response in *C. elegans*. The effect on maximum viral load in both RNAi mutants points to JU1580 being compromised in provoking an effective antiviral RNAi response which is in line with recent findings [23, 24]. However, overall there is not a large difference between N2 and the genotypes impaired in the RNAi response: JU1580 and the RNAi mutants. This is in agreement with our previous findings, comparing only N2 and JU1580, which suggest that the main difference between N2 and JU1580 does not lie in the primary infection.

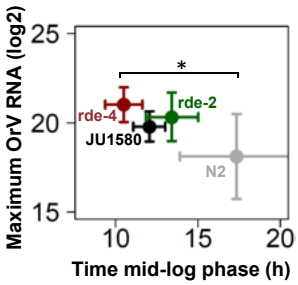


Figure 3: Infection assays in L3 of N2, JU1580, *rde-2* and *rde-4*. The maximum viral load reached and the time till the inflection point was reached is shown for four genotypes. The *rde-4* mutant reaches the mid-log phase significantly faster than N2 (Two-sided *t*-test, P≤0.05), whereas the other two genotypes are similar to N2 (Two-sided *t*-test, P>0.05). However, there is no significant difference between maximum viral load reached between the genotypes (Two-sided *t*-test, P>0.05).

Role of heritable RNAi in susceptibility to OrV infection

The combined results of the OrV infection experiments in N2 and JU1580 show that the initial infection development only contributes to a ~10-fold difference in viral levels, which falls short to explain the difference in viral load between JU1580 and N2 reported previously [7]. Given that inheritance of RNAi is a well-studied phenomenon in *C. elegans* in general [18], as well as in the context of viral replication [22], the differences with previously published results [7] could be caused by an inheritable antiviral RNAi response.

In order to test this hypothesis an experiment was designed to determine the trans-generational effects of the RNAi response on the development of OrV infection in JU1580, N2 and both the *rde-2* and *rde-4* mutants (**Figure 4A**). In short, worms were synchronized and exposed to OrV at 26h (L2 stage). At 72h, infected worms were sampled and either transferred or bleached (to synchronize and get rid of OrV infection). The transferred worms were sampled again at 72h after transfer and the same process was repeated once more. The bleached group was re-infected at 26h and again sampled at 72h after bleaching. This cycle was also repeated for a third time. The worms in the transferred group are expected to show trans generational silencing effects, but not as severe as in the bleached group since the population is more mixed and contains the primary infected worms. However, if there is a strong negative effect in spread of the infection (as N2 needs a higher dose for establishing the infection), it will be seen in this group. In particular in the two RNAi mutants in the N2 background, these should then show lower infection levels independent of the RNAi effect. The bleached group will show the RNAi effect in particular as they are re-exposed to OrV every generation.

When the viral loads were compared upon OrV infection, only in N2 a significant difference in infection was found between the pre-exposed and naïve exposed populations (**Figure 4B and 4C, Supplementary Figure 2**). A >10-fold decrease was found in the subsequent generations compared to the first generation (ANOVA, P≤0.05) (**Figure 4B**). Also in the transferred group

there was a >3-fold decrease (ANOVA, P≤0.001) (**Figure 4C**). As expected, neither of the RNAi mutants (*rde-2, rde-4*) displayed a trans-generational effect of pre-exposure to OrV replication in the offspring, since no differential susceptibility to OrV infection was observed between naïve and pre-exposed worms. This result suggests that the effect seen in N2 is linked to the formation of a heritable RNAi response. Furthermore, since JU1580 does not show decreased replication of OrV in the offspring may indicate that this genotype cannot mount an effective heritable RNAi response.

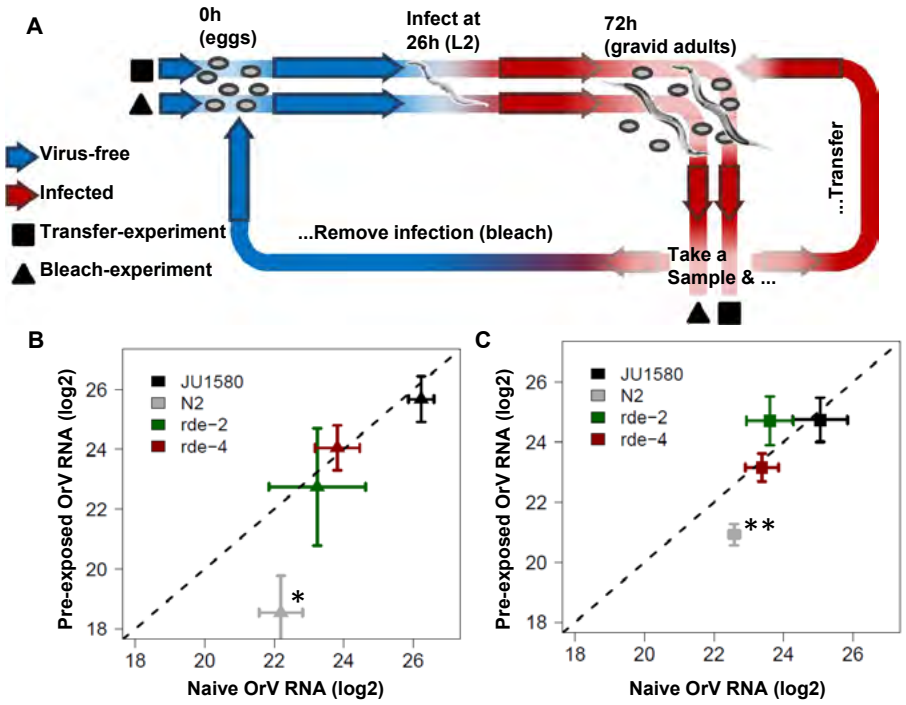


Figure 4: Trans-generational inheritance of antiviral RNAi response. **(A)** Shows the outline of the experiment, worms were synchronized and 26h thereafter exposed to OrV. 72h past bleaching the worms have laid eggs, at this point the experiment progresses in two parts (indicated by the triangles and squares), either re-synchronized and re-infected populations (triangles) or populations that were only transferred (squares). **(B)** The outcome of the experiments (n=6) for the re-infected populations. Only in N2 there was a significant reduction in virus titer in the pre-exposed populations compared to the naive populations (ANOVA, P≤0.05). **(C)** The outcome of the experiment where the populations were transferred (so no re-infection). Also here only N2 showed a significant reduction in virus titer in the following generations compared to the first generation (ANOVA, P≤0.01).

Discussion

The power of *C. elegans* as a model species initially prompted the development of artificial systems in which virus-host interactions could be studied in the worm. This research used vesicular stomatitis virus (VSV) and Flock house virus (FHV). Despite their wide host range, both viruses could only replicate within *C. elegans* by either using embryonic derived cells or using a transgenic system [19-21]. An important outcome from these studies was that the RNAi pathway had the capacity to limit viral replication via the involvement of the Argonaute RDE-1 [19-21], the nucleotidyltransferase MUT-2 [21], and the dsRNA binding protein RDE-4 [20, 21]. FHV was used to further identify genes involved in antiviral RNAi [25] and discover trans-generational inheritance of the antiviral small-interfering RNAs (viRNAs) [22]. The inheritance of an RNAi response was first established in the landmark paper by Fire *et al.* (1998), where gene silencing was induced in the progeny of worms injected with dsRNA [18]. Subsequent research identified the genes involved in initial silencing and in the transfer of the silencing response to the offspring [35, 36]. It became clear that the induction of siRNAs and inheritance of an RNAi response are different mechanisms [35] and that the inheritance requires formation of secondary siRNAs [21].

The *C. elegans* RNAi response is also important in limiting OrV replication, which was convincingly shown by experimental infection of a range of mutant *C. elegans* strains [7]. Also the phenotypic differences found between N2 and JU1580 have been linked to a polymorphism in the *drh-1* gene [23] and it seems that this gene is also involved in limiting OrV infection among wild isolates [24]. OrV persists in the natural JU1580 population due to efficient horizontal transmission from the infected worms to other worms and its offspring. To analyse the development of OrV infection with higher resolution we developed a quantitative infection assay. By infecting worm cohorts at different time points with OrV, progression of infection was monitored through quantification of viral RNA using qPCR. The influence of worm age and genotype on viral replication was determined using a defined viral dose for a defined incubation time.

In JU1580 larvae the mid-log phase was reached earlier in younger JU1580 larvae and coincided with a lower maximum infection load. These observations may be linked to the development of the intestine, the site of OrV infection [6, 7]. At the end of each larval stage the volume of the intestinal cells expand, as does the ploidity of the middle 10 intestinal cells [37]. Therefore, the maximum viral load measured could be limited by available space in the nematode. Next to that, the speed at which the infection develops is slower at older age. The cause of these observations remains unclear but could be elucidated by detailed immunofluorescence studies following the development of the infection.

In N2 there was a trend for higher susceptibility in younger larvae, but the effect was not as strong as in JU1580. The most striking difference was found in maximum viral load, which did

not differ significantly between infections started in the respective larval stages in N2. This is in contrast with JU1580, where the maximum viral load increases with age. The reason for this could lie in a more limited infection in N2 [7] or a stronger antiviral response, persisting over time, to begin with. We found that in experiments carried out within one generation, where no offspring was present (infection at L1, L2 and L3 experiments), the differences between N2 and JU1580 were relatively small. However, when offspring was present (infection at L4 and the heritable RNAi experiment), we observed that the differences in viral load increased. The cause of this is unclear as it seems unlikely that larvae are already (highly) infected at this stage.

The difference in viral load phenotype of JU1580 and both the *rde-4* and *rde-2* mutants relative to N2 was small compared to the reported 100-fold differences in viral load by Félix *et al.* (2011) and prompted the hypothesis that trans-generational effects may play a role. Therefore, several subsequent generations exposed to virus were re-infected to determine if trans-generational effects could be observed in infections in *C. elegans* populations. These experiments showed that the RNAi response has a dual role in limiting infection; i) an RNAi response to limit OrV replication in the individual worm, combined with ii) a trans-generational effect rendering offspring of infected N2 less susceptible to viral replication. Consequently, N2 populations might lose the infection after a limited number of generations, whereas JU1580 populations remain infected.

This heritable RNAi response present in N2 appears absent or severely compromised in the wild isolate JU1580, which may be linked to the recently detected polymorphism in its *drh-1* gene [23, 24]. All the *C. elegans* strains that were isolated from the two sites where the *Caenorhabditis*-infecting viruses were found (Orsay and Santeuil) are polymorphic for *drh-1* [10]. *DRH-1* is a homologue of mammalian RIG-I and most likely a molecule with activity in the RNAi pathway involved in sensing non-self (e.g. viral) RNA [25, 26]. We found that the OrV infection develops faster in the *rde-4* mutant than in the *rde-2* and both are similar to JU1580. This, along with abnormalities of the small antiviral RNA response against OrV [7, 24], show that JU1580 cannot mount an effective early RNAi response. The finding that JU1580 does not show a trans-generational effect is indicative of an abrogated function in or upstream of secondary siRNA generation.

To conclude, we report a quantitative study of OrV replication and the discovery of trans-generational effects of antiviral RNAi. Dose-response analysis of different larval stages revealed that the progression speed of OrV infection decreased with subsequent larval stages (L1-L4) and higher maximum viral loads were reached in the older stages. Surprisingly, hitherto presumed OrV sensitive strain JU1580 showed similar susceptibility as N2 at exposure to higher viral doses in liquid inoculum. In contrast to JU1580, viral infection in N2 is controlled by a heritable RNAi response. Consequently, offspring of infected N2 is less susceptible to viral replication. We present a new quantitative infection assay using *C. elegans* which allows for studying the molecular details of OrV replication, thus facilitating virus-host interaction studies in a genetically tractable model organism.

Acknowledgements

The authors would like to thank M.A. Félix for the infected JU1580 strain. The two mutant strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We would like to thank M. Rodriguez, P. Mooijman and S. Van Elsen for help with the set-up of the qPCR experiments.

Author contributions

Conceived and designed the experiments: MGS, LBS, GPP, and JEK . Performed the experiments: MGS, KJB, JD, and JAGR. Analysed the data: MGS and LBS. Wrote the paper: MGS, LBS, JB, GPP, and JEK.

Supporting Information

The supplementary files and figures are deposited at: http://marksterken.nl, under 'PhD thesis'.

Supplementary Figure 1: Logistic curve fits. All the curve fits obtained for JU1580 (infected in L1, L2, L3 and L4), N2 (infected in L1, L2, L3 and L4), WM29 (*rde-2*, infected in L3) and WM49 (*rde-4*, infected in L3). The time is time post infection. Individual data points are shown in dots. Identified outliers are shown with an x instead of a dot. The sigmoidal curve fit +/- SD is shown in the dashed grey lines. The calculated inflection point and calculated asymptote are also shown. As is the R^2 of the curve-fit.

Supplementary Figure 2: Heritable RNAi experiment. All the individual data points for the heritable RNAi experiment (6 independent experiments) are shown, for the genotypes JU1580, N2, WM29 (*rde-2*) and WM49 (*rde-4*). The mean +/- SE are shown.

Supplementary File 1: Reference genes. The mean Ct-values +/- SD for the reference genes per genotype and stage (NA means none available; indicates in which stages no experiments were done).

References

1. **Kirienko, N.V., K. Mani, and D.S. Fay,** *Cancer models in Caenorhabditis elegans.* **Dev Dyn, 2010. 239(5): p. 1413-48.**
2. **Kenyon, C.J.,** *The genetics of ageing.* **Nature, 2010. 464(7288): p. 504-12.**
3. **Snoek, L.B., et al.,** *A rapid and massive gene expression shift marking adolescent transition in C. elegans.* **Sci Rep, 2014. 4: p. 3912.**
4. **Jager, T., et al.,** *Modelling nematode life cycles using dynamic energy budgets.* **Functional Ecology, 2005. 19(1): p. 136-144.**
5. **Marsh, E.K. and R.C. May,** *Caenorhabditis elegans, a model organism for investigating immunity.* **Appl Environ Microbiol, 2012. 78(7): p. 2075-81.**
6. **Franz, C.J., et al.,** *Orsay, Santeuil and Le Blanc viruses primarily infect intestinal cells in Caenorhabditis nematodes.* **Virology, 2014. 448: p. 255-64.**
7. **Felix, M.A., et al.,** *Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses.* **PLoS Biol, 2011. 9(1): p. e1000586.**
8. **Franz, C.J., et al.,** *Complete genome sequence of Le Blanc virus, a third Caenorhabditis nematode-infecting virus.* **J Virol, 2012. 86(21): p. 11940.**
9. **Andersen, E.C., et al.,** *Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nat Genet, 2012. 44(3): p. 285-90.**
10. **Volkers, R.J., et al.,** *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93.**
11. **Kiontke, K.C., et al.,** *A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits.* **BMC Evol Biol, 2011. 11: p. 339.**
12. **Grishkevich, V., et al.,** *A genomic bias for genotype-environment interactions in C. elegans.* **Mol Syst Biol, 2012. 8: p. 587.**
13. **Elvin, M., et al.,** *A fitness assay for comparing RNAi effects across multiple C. elegans genotypes.* **BMC Genomics, 2011. 12: p. 510.**
14. **Felix, M.A. and C. Braendle,** *The natural history of Caenorhabditis elegans.* **Curr Biol, 2010. 20(22): p. R965-9.**
15. **Pujol, N., et al.,** *Anti-fungal innate immunity in C. elegans is enhanced by evolutionary diversification of antimicrobial peptides.* **PLoS Pathog, 2008. 4(7): p. e1000105.**
16. **Kammenga, J.E., et al.,** *Beyond induced mutants: using worms to study natural variation in genetic pathways.* **Trends Genet, 2008. 24(4): p. 178-85.**
17. **Vinuela, A., et al.,** *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37.**
18. **Fire, A., et al.,** *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* **Nature, 1998. 391(6669): p. 806-11.**
19. **Lu, R., et al.,** *Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans.* **Nature, 2005. 436(7053): p. 1040-3.**
20. **Wilkins, C., et al.,** *RNA interference is an antiviral defence mechanism in Caenorhabditis elegans.* **Nature, 2005. 436(7053): p. 1044-7.**
21. **Schott, D.H., et al.,** *An antiviral role for the RNA interference machinery in Caenorhabditis elegans.* **Proc Natl Acad Sci U S A, 2005. 102(51): p. 18420-4.**
22. **Rechavi, O., G. Minevich, and O. Hobert,** *Transgenerational inheritance of an acquired small RNA-based antiviral response in C. elegans.* **Cell, 2011. 147(6): p. 1248-56.**
23. **Sarkies, P., et al.,** *Competition between virus-derived and endogenous small RNAs regulates gene expression in Caenorhabditis elegans.* **Genome Res, 2013. 23(8): p. 1258-70.**
24. **Ashe, A., et al.,** *A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity.* **Elife, 2013. 2: p. e00994.**
25. **Lu, R., et al.,** *An RIG-I-Like RNA helicase mediates antiviral RNAi downstream of viral siRNA biogenesis in Caenorhabditis elegans.* **PLoS Pathog, 2009. 5(2): p. e1000286.**

26. **Guo, X., et al.,** *Homologous RIG-I-like helicase proteins direct RNAi-mediated antiviral immunity in C. elegans by distinct mechanisms.* **Proc Natl Acad Sci U S A, 2013. 110(40): p. 16085-90.**
27. **Brenner, S.,** *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94.**
28. **Emmons, S.W., M.R. Klass, and D. Hirsh,** *Analysis of the constancy of DNA sequences during development and evolution of the nematode Caenorhabditis elegans.* **Proc Natl Acad Sci U S A, 1979. 76(3): p. 1333-7.**
29. **Le Novere, N.,** *MELTING, computing the melting temperature of nucleic acid duplex.* **Bioinformatics, 2001. 17(12): p. 1226-7.**
30. **Bates, D.M.C., J.M.,** *Nonlinear models***, in** *Statistical Models in S.***, J.M.H. Chambers, T.J., Editor. 1992, Wadsworth & Brooks/Cole: London.**
31. **Tijsterman, M., et al.,** *RNA helicase MUT-14-dependent gene silencing triggered in C. elegans by short antisense RNAs.* **Science, 2002. 295(5555): p. 694-7.**
32. **Sundaram, P., et al.,** *Caenorhabditis elegans ABCRNAi transporters interact genetically with rde-2 and mut-7.* **Genetics, 2008. 178(2): p. 801-14.**
33. **Blanchard, D., et al.,** *On the nature of in vivo requirements for rde-4 in RNAi and developmental pathways in C. elegans.* **RNA Biol, 2011. 8(3): p. 458-67.**
34. **Parrish, S. and A. Fire,** *Distinct roles for RDE-1 and RDE-4 during RNA interference in Caenorhabditis elegans.* **RNA, 2001. 7(10): p. 1397-402.**
35. **Grishok, A.,** *Genetic Requirements for Inheritance of RNAi in C. elegans.* **Science, 2000. 287(5462): p. 2494-2497.**
36. **Vastenhouw, N.L., et al.,** *Gene expression: long-term gene silencing by RNAi.* **Nature, 2006. 442(7105): p. 882.**
37. **Altun, Z.F.a.H., D.H.** *Alimentary system, intestine***. 2009.**

Chapter 5

Natural variation in Orsay virus infection links to chromosome IV in *Caenorhabditis elegans*

Mark G. Sterken,
Lisa van Sluijs,
Yiru Wang,
Wannisa Ritmahan,
Frederik Pankok,
Joost A.G. Riksen,
Rita J.M. Volkers,
L. Basten Snoek,
Gorben P. Pijlman,
Jan E. Kammenga

Abstract

Host-pathogen interactions play a major role in evolutionary selection and in shaping natural genetic variation. Recent identification of viral infection in *C. elegans* has prompted research into understanding the underlying pathways of Orsay virus (OrV) infection in natural populations. Here we report the dissection of the genetic architecture of OrV infection. We found that the *C. elegans* wild type Hawaii CB4856 strain was more resistant to OrV infection than the canonical Bristol N2 strain. To gain insight into the genetic architecture of resistance, 52 fully sequenced recombinant inbred lines (CB4856 x N2 RILs) were exposed to OrV using our recently developed quantitative infection assay. This led to the identification of two distinct loci on chromosome IV associated with OrV resistance. Both loci were associated with a lower viral load in the CB4856 genotype. Strikingly, these loci do not harbour the recently identified *drh-1* locus, which encodes a RIG-I like helicase that plays an important role in infection control via antiviral RNAi. To verify the RIL results and gain additional insight into the genetic architecture controlling virus infection, a panel of 18 introgression lines (ILs) (together covering chromosome IV entirely) was exposed to OrV. Both loci were found back in ILs, also leading to more resistance against OrV infection. Analysis of the transcriptome revealed that ubiquitination plays an important role in the response to OrV infection. Natural variation in the *cul-6* gene might result in the lower viral load in CB4856.

Introduction

Genetic variation plays a major role in the arms race between pathogen and host [1-3]. The interaction between host genetic background and pathogens can shape natural variation by imposing a strong selection regime on the affected population. For example, rare allele variants have been found to convey resistance to bubonic plague in humans [4]. Host genetic variation is also important in more recent viral outbreaks as illustrated by studies that correlate outcome of infection with Hepatitis, HIV, and Ebola to the host's genetic background [5-7]. These interplays between virus and host showcase the role of natural variation in the relation between pathogen and host. Studying host-virus interactions in model systems can uncover the genetic networks determining viral susceptibility [7].

The nematode *Caenorhabditis elegans* encounters a variety of pathogens in its natural habitat, which include: bacteria, microsporidia, fungi, and viruses [8-12]. Only recent advances in the understanding of *C. elegans* ecology have helped to appreciate the lush microbial world this nematode lives in [13, 14]. So far, only one virus has been discovered that naturally infects *C. elegans*, the Orsay virus (OrV).

Several pathways have been implicated in defence against viral infections in *C. elegans*. Of which the RNAi response is the best studied and arguably the most important antiviral pathway. The RNAi response is a highly adaptive and diverse pathway that plays a role in many processes in an organism, for example in development and antiviral responses in invertebrates [15]. In OrV infection, it recognizes the double stranded RNA replication intermediate, which ultimately leads to the production of small interfering RNAs (siRNAs) that target the viral RNA for degradation [16]. Mutants defective for various genes in the RNAi pathway display higher viral susceptibility upon infection (see also **Chapter 4**) [9, 15-17].

The ubiquitination pathways is also involved in defence against infections, targeting viral proteins for degradation. This response is mediated by an SCF complex, which consists of several proteins, including an F-box protein that recognizes thesubstrate and an SKR protein that functions as an adapter for the F-box protein, binding it to a CULLIN protein at the N-terminus. The C-terminus of the cullin protein contains an RBX binding site. RBX functions as an adapter for E2 ligase proteins, that can ligate ubiquitin to the substrate [18-20]. This is a highly diverse response; the *C. elegans* genome contains over 300 F-box proteins, 21 SKR proteins, and 22 E2 ligases [19, 21-23]. Furthermore, the F-box gene family is highly divergent between *C. elegans* isolates [24]. One CULLIN protein, CUL-6, has been implicated in defence against microsporidia and OrV infections in *C. elegans* [25].

Natural variation plays a role in the susceptibility to OrV infections. Initially, it was observed that the natural *C. elegans* isolate JU1580 is more susceptible to infection with OrV than the wild-type strain N2 [9]. This difference has been linked to a natural polymorphism in *drh-1* (a RIG-I like protein) affecting the anti-viral RNAi response. Among *C. elegans* strains isolated around

the globe, there is a variety in *drh-1* alleles. Interestingly, many strains contain the allele shown to be dysfunctional in JU1580. This frequent presence of a deleterious allele has been attributed to hitchhiking with a favourable allele, supported by the observation of linkage of the *drh-1* locus with the surrounding loci [16]. These findings establish RNAi as an important adaptive response against OrV infection. Next to the RNAi response, the ubiquitination pathway has been implicated as an anti-viral pathway against OrV infection [25]. Given the strong selection pressure on proteins involved in immunity, it is likely there are many large-effect polymorphisms present in the *C. elegans* meta-genome [1, 2].

In this study, we investigated whether the wild-type strains N2 and CB4856 react differently to OrV infection. We indeed found such a difference, CB4856 is less susceptible compared to N2. To uncover causal loci, we used inbred panels constructed from these strains. The CB4856 and N2 strain are very polymorphic, with more than 400,000 polymorphisms, including insertions/deletions and single nucleotide variants [26]. Over the last decade, both strains have been jointly used in many quantitative genetics studies in *C. elegans*, focused on traits like: aging, stress tolerance, and pathogen avoidance [27-31]. Most of these studies have been conducted on one of the two available recombinant inbred line (RIL) panels [32, 33] or on the introgression line (IL) population which contains fragments of CB4856 in a background of N2 [28].

Here we set out to investigate the loci involved in the phenotypic differences between the Bristol N2 strain and the Hawaii CB4856 in response to OrV infection. We characterized the traits affected in these strains by characterizing viral replication in a dose- and time-dependent manner. Subsequently, we exposed a panel of 52 RILs to OrV and measured the viral load. We identified two QTL associated with differences in viral load on chromosome IV. A subsequent analysis in an IL population (with an N2 genetic background) covering chromosome IV and a transcriptome analysis led to the identification of a candidate gene involved in ubiquitination.

Material and methods

C. elegans Strains

C.elegans strains Bristol N2 and Hawaii CB4856 were used and strains derived from crosses between these two wild-type strains. In this paper 52 recombinant inbred lines and 18 CB4856-in-N2 introgression lines covering chromosome IV were used (**Supplementary file 1**) [28, 30, 32]. Most of these strains (except for 8 of the ILs) have been genotyped by sequencing [26].

C. elegans culturing

The nematodes were kept at 12°C in-between experiments on 6 cm NGM plates seeded with *E. coli* OP50 [34]. Bleaching was used to synchronize populations and to remove bacterial or fungal contaminations [34]. Before experiments, a population without males was created by picking single worms in the L1/L2 stage and transferring hermaphrodite populations to fresh 9 cm NGM plates. New experiments were started by bleaching an egg-laying population grown at 20°C.

Orsay virus stock preparation

A virus stock was generated by isolating OrV from a persistently infected JU1580 culture (see also **Chapter 4**) [9]. Over 50 JU1580 populations were grown on 9 cm NGM plates containing twice the amount of agar. The nematodes were collected by washing the animals off the plate with M9 and collecting the suspension in an Eppendorf tube. The suspension was flash frozen in liquid nitrogen to break the nematodes which were slowly thawed and kept at 4°C thereafter. The suspension was centrifuged for 5 minutes at 10,000 rpm to pellet the bacteria and nematodes. The supernatant was collected and passed through a 0.2 µm filter. The obtained virus stock was divided in aliquots, flash frozen in liquid nitrogen, and stored at -80°C until use. Specific infectivity of the virus stock was tested by serial dilution infections on the natural host for OrV, *C. elegans* JU1580 (see also **Chapter 4**) .

Infection experiments

The infection assay was conducted as described in **Chapter 4**. Populations were synchronized (t=0 hours) and grown at 20°C on 9 cm NGM plates. At the moment of infection, the strains were washed off the plate with M9 buffer, spun down in a centrifuge for 10 seconds at 10,000 rpm. The supernatant was removed and the strains were exposed to OrV in liquid for 1 hour. The worms were washed 3 times with M9 and placed on a fresh 9 cm NGM plate. All experiments were conducted on animals in the L2 stage (26 hours after synchronization).

For the viral dosage experiments, N2 and CB4856 animals were exposed in different concentrations

of OrV. The concentrations used were, 0, 10, 20, 50, 100, and 200 µL of OrV stock per 500 µL of infection solution. The experiment was conducted in 6 independent biological replicates, all using the same OrV stock solution. For the replication kinetics experiments on N2 and CB4856, the animals were harvested 2-35 hours post infection. This experiment was conducted in 8 independent biological replicates, each evenly covering the time-series. For the viral load experiments on the RIL and IL panels, the animals were harvested 30 hours post infection. The experiment in the RIL panel was conducted on 3 independent biological replicates. The experiment in the IL panel was conducted on 7 independent biological replicates for the whole chromosome IV set and an additional 12 replicates for four ILs that covered identified QTLs (WN247, WN248, WN252, and WN256). The transcriptome experiment was conducted on 16 replicates: 4 infected N2, 4 mock-infected N2, 4 infected CB4856, and 4 mock-infected CB4856. The infection of the transcriptome experiment was stopped at 30 hours post infection.

Egg laying delay and fecundity experiments

Nematode populations were kept at 20°C and stage synchronized by bleaching (t=0 hours). The eggs were hatched in M9 and the next day, 17 hours post bleaching, the nematodes were (mock) infected with OrV. At 24 hours post infection single animals were picked a placed in 12 wells plates. For each batch 24 animals per genotype per treatment were scored hourly for egg laying, starting at 46 hours post infection. The number of eggs was scored as: 0 (no eggs), 1 (1-10 eggs), 2 (10-20 eggs), 3 (20-50 eggs), and 4 (over 50 eggs). This experiment was repeated 3 times.

RNA isolation

The RNA was isolated using a Maxwell® 16 AS2000 instrument with a Maxwell® 16 LEV simply RNA Tissue Kit (both Promega) following the recommended protocol, except the addition of 10 mg of proteinase k (5prime) at the lysis step. The lysate was incubated in a shaker for 10 minutes at 65°C at 1,000 rpm. After isolation the quality and quantity of the RNA was determined via NanoDrop (Thermo Scientific).

qPCR: cDNA preparation and qPCR

cDNA was synthesized using the GoScript Reverse Transcriptase kit (Promega) following the recommended protocol with random hexanucleotides (Thermo Scientific) and 1 µg of total RNA as starting material. The cDNA was diluted 1/50 with nuclease free water and quantified by qPCR (MyIQ, Biorad) using Absolute QPCR SYBR Green Fluorescein Mixes (Thermo Scientific) following the recommended protocol. The samples were quantified with two technical replicates for two primer combinations amplifying OrV RNA-1 (HM030970.1): pOrV-RNA1.1F (5'ATACTCTACGACCTTGTCGG 3') plus pOrV-RNA1.1R (5'CTCGGTTGATGTTCTTCCAG 3') and pOrV-

RNA1.2F (5'AACCAGGAAACACTACTCCG 3') plus pOrV-RNA1.2R (5'GTTGTGATATCGCTTGGTGG 3'). The samples were also quantified in two technical replicates for two primer combinations amplifying two reference genes: Y37E3.8 (by pY37E3.8F: 5'GCGTTTGTGGTCTCTTGTC 3' plus pY37E3.8R: 5'CTCTGGGAGGAGTCCTTTTC 3') and *rpl-6* (by pRPL6-F: 5'TGTCACTCTCCGCAAGAC 3' plus pRPL6-R: 5'TGATCTTGTGTGGTCCAGTG 3') [**Chapter 4**].

qPCR: Data normalization

The data was processed using R (x64 3.2.2), as described before [**Chapter 4**]. In short, before normalization, the qPCR measurements were transformed by

$$Q_{gene} = 2^{40-Ct_{gene}}$$

where Q is the expression of the gene and Ct is the measured Ct value of the gene. The viral expression was normalized by the two reference genes, using the formula

$$E = \frac{Q_V}{0.5 * ((Q_{rpl6} / \overline{Q}_{rpl6}) + (Q_{Y37E3.8} / \overline{Q}_{Y37E3.8}))}$$

where E is the normalized viral load, Q_v is the expression of the viral RNA and Q_{rpl6} is the expression of reference gene *rpl-6* and $Q_{Y37E3.8}$ is the expression of reference gene Y37E3.8. An overview of the normalized data for the RIL and IL panels is attached in **Supplementary figure 2**.

From the replicate measurements in the RIL panel, several traits could be derived for QTL mapping over the RIL population. The following were derived including all measurements: mean viral load, median viral load, maximum viral load, minimum viral load, variation, and infection success. Furthermore, we excluded the unsuccessful infections (as these could also arise due to technical failures) and calculated the mean viral load (Mean_infected) and the median viral load (Median_infected). A matrix of these (derived) values per strain, and the original measurements can be found in **Supplementary file 2**.

Quantitative trait locus mapping in theRIL population

Single locus QTL mapping was done using a linear model (R, x64 3.2.2) to explain viral load and derived traits (see **Supplementary file 2**) over the markers by

$$E_i \sim x_{i,j} + \varepsilon_{i,j}$$

where E is the viral load of RIL i (1, 2, …, 52) and x is the marker of RIL i at location j (a set of 729 sequenced markers was used, **Chapter 6**). For E the outcome of each replicate of the experiment was used separately, as well as the average over all three experiments. For the mean viral load, the minor peak on chromosome IV was mapped by regressing the major peak out of the data based on the linear model outcome and re-mapping using a single marker model.

For all mappings, the statistical threshold was determined via a permutation analysis, where the values measured for E were randomly distributed over the genotypes. The same model as for the mapping was used and this analysis was repeated 1,000 times. The 950th highest p-value was taken as the threshold p-value for a false discovery rate of 0.05.

Heritability and variance explained

The broad sense heritability (H^2) was calculated by ANOVA, explaining the viral load over the genotype

$$H^2 = \frac{V_G}{(V_G + V_e)}$$

where V_G is the genotypic variation and V_e is the residual variation.

The variation explained by a QTL peak was calculated by

$$V_{Explained} = \frac{R^2_{QTL}}{H^2}$$

where R^2 is the R^2 from the fit of the peak marker to the trait as calculated by an ANOVA model and H^2 is the heritability of the trait. In case of multiple markers, the R^2 was calculated over the full model.

Introgression line analysis

The viral loads obtained for the introgression lines were analyzed in two ways: individually against N2 and by bin-mapping [28]. For both the analysis the data was batch corrected.

The ILs were compared against N2 via a two sided t-test assuming unequal variance, as provided by R (x64 3.2.2). The values obtained over the independent biological replicates were used, excluding the experiments where no virus was detected.

The ILs were used for bin-mapping by applying a linear model (R, x64 3.2.2)

$$E_i \sim x_{i,j} + \varepsilon_{i,j}$$

where E is the viral load of IL i (1, 2, …, 52) and x is the marker of IL i at location j (a subset of the 729-marker map was used, using the markers covering chromosome IV [26, 28]).

Microarray: cDNA and cRNA synthesis, hybridization and scanning

RNA was isolated as for the qPCR samples, the cDNA and cRNA were constructed according to the 'Two-Color Microarray-Based Gene Expression Analysis; Low Input Quick Amp Labeling' protocol (Agilent Technologies). The microarrays were scanned using an Agilent High Resolution C Scanner, using the recommended settings. The data was extracted using the Agilent Feature Extraction software (version 10.7.11).

Microarray: statistical analysis

All data processing and analysis was done in R (3.2.2). The Limma package was used to normalize within arrays with the 'loess' method and between arrays using the quantile method [35, 36]. The obtained log2 normalized intensities and the log2 ratio with the mean (per spot) were used in subsequent analysis. By correlation and PCO analysis we detected a batch effect from the labelling process. This batch effect, and the dye effect were removed by subtracting the effects obtained by regression analysis (data not shown). One infected N2 sample was excluded from further analysis due to a strong mismatch with the other samples.

The log2 ratio with the mean values were used in a principal component analysis to determine which axis contributed to variation. Most of the variation was contributed by genotype (38.1%), and the 4th axis divided infected from non-infected samples (8.4% of variation, **Supplementary figure 5**).

The data was analyzed using a linear model,

$$Y_i \sim G + T + \varepsilon$$

where Y is the log2 normalized intensity of spot i (1, 2, ..., 45220), which was explained over Genotype (either N2 or CB4856), treatment (either infected or mock), and error term ε. The significance threshold was determined by the p.adjust function, using the Benjamini & Hochberg correction (FDR < 0.05) [37]. The outcome of the linear model is plotted in **Supplementary figure 6** and a list of significant genes per category can be found in **Supplementary file 5**.

Microarray: Gene enrichment analysis

Gene group enrichment analysis was done using a hypergeometric test and several databases with annotations. The databases used were: The WS220 GO-annotation, anatomy terms, protein domains, and gene classes [38, 39]; the MODENCODE release 32 transcription factor binding sites (www.modencode.org) [40, 41], which were mapped to transcription start sites (according to[42]); the KEGG pathway release 65.0 (www.genome.jp/kegg/) [43].

Enrichments were selected based on the following criteria: hypergeometric test p-value < 0.001, size of the category n>3, size of the overlap n>2. Enrichments were calculated based on gene-names, not on spots. The outcome for the eQTL enrichment analysis can be found in **Supplementary file 7**.

Protein structure analysis

Protein sequences from the human CUL1, *Saccharomyces cerevisiae* CDC53 and *C. elegans* CUL-1, CUL-6 N2 allelic variant and CUL-6 CB4856 allelic variant were aligned in ClustalX (version 2.1) using the standard settings (**Supplementary file 8**) [44]. A structural model for the N2 and CB4856 allelic variant was predicted using the human CUL1 protein structure as a template in

the SWISS-MODEL ExPASy web server. The standard search parameters were used, based on the SWISS-MODEL template library (version 14-01-2015) and the protein data bank (version 09-01-2015) [45-50]. The obtained models for N2 and CB4856 CUL-6 were compared in SwissPDBViewer (v. 4.1.0) [47].

Results

CB4856 displays resistance to OrV infection

We found that the two wild-type genotypes N2 and CB4856 react differently to OrV infection. Upon exposing these strains with different amounts of OrV (**Figure 1A**) [**Chapter 4**], higher dosage leads to a higher viral load (**Figure 1B**, ANOVA, $p < 1*10^{-5}$). Furthermore, N2 developed a viral load 3.4 units higher than CB4856 after OrV exposure (**Figure 1B**, ANOVA, $p < 1*10^{-5}$). This difference could arise due to a slower developing infection, a difference in the stationary phase of the infection, or a difference in the number of infected individuals.

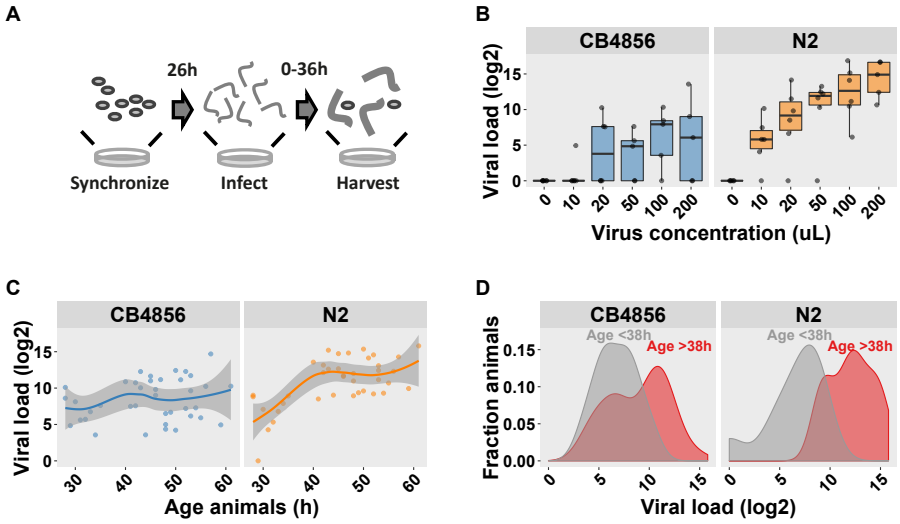


Figure 1: The differences in OrV infection between N2 and CB4856. (**A**) The infection assay. Nematodes are grown until the L2 stage (26 hours) and exposed for 1 hour to OrV in liquid. After that, the worms are grown on NGM dishes until 0-36 hours post infection. Subsequently the population is isolated and the viral load is measured. (**B**) Exposure of N2 and CB4856 to a concentration range of OrV. While both strains show a dose-dependent relation between virus concentration and viral load, the infection develops to a higher level in N2 compared to CB4856. (**C**). Infection development over time. Both strains start out with similar viral loads in early infection and develop – on average - higher viral loads as time progresses. CB4856 shows more variation, in viral load in the stationary phase than N2. (**D**) A density plot of the viral load measurements over time, divided in two groups: early infection (up to 12 hours post infection, grey) and late infection (after 12 hours post infection, red).

To investigate this we measured the infection development over time. By infecting both strains at an age of 26h and measuring the viral load at multiple time points (**Figure 1C**). Whereas the infection developed steadily in N2, a large variation was observed in CB4856. In some experiments the infection developed to a similar (but lower level) compared to N2, however in other experiments the infection did not develop beyond levels reached in the lag-phase of the infection. This is visible as a bifurcation in the distribution of the viral load after 12 hours of infection (**Figure 1D**). In this time series experiment, a significant amount of the variation

was explained by genotype (ANOVA, $p < 1*10^{-4}$). N2 developed a 3.2 units higher viral load than CB4856, in concordance with the observation in the dosage experiment (**Figure 1C**). The time passed since infection was also explaining variation in viral load (**Figure 1C**, ANOVA, $p < 1*10^{-6}$). Next to this, we observed that infection could be established in N2 more often than in CB4856 (76% (n=121) versus 61% (n=115) success rate). Together, these observations showed that CB4856 developed a lower viral load and can control the infection more successfully compared to N2.

We also measured whether progeny production was affected, since this has been reported previously in JU1580 versus N2 [9, 16]. We found that the onset of progeny production was slightly delayed in infected CB4856 versus mock-infected CB4856 (102 minutes, ANOVA, $p < 0.01$). This was not the case in N2, where the onset of progeny production was equal in the mock-infected and the infected nematodes (ANOVA, $p = 0.38$, **Supplementary figure 1**). Our results on N2 were consistent with previous findings [16]. The delay in CB4856 is counter intuitive given that the viral load is lower in this strain. Therefore, there might be other mechanisms leading to the CB4856 progeny production phenotype after OrV infection.

Two loci on chromosome IV are linked to resistance against OrV

To find the causal loci underlying the differences between N2 and CB4856 in viral load, recombinant inbred lines (RILs) constructed from a cross between these strains were infected with OrV (**figure 2A**) [26, 30, 32]. The RILs were infected in the L2 stage (at an age of 26h) and the infection was continued for 30h, after which the viral load was measured (**Supplementary figure 2**) [51]. We found a broad-sense heritability (the fraction of trait variation explained by genotype) of 0.45 based on all three infection replicates, which is reasonable given the biological variation of this trait.

Linkage analysis identified two QTL on chromosome IV (**Figure 2B**). The linkage analysis was started on multiple traits derived from three independent replicates of the RIL panel (see materials and methods; **Supplementary file 2**). Correlation analysis of these derived traits shows that two main groups could be discerned: traits linked to the lowest measured viral load per strain and traits linked to the highest measured viral load per strain (**Supplementary figure 3**). All of the traits mapped showed a link to chromosome IV, and significant QTL were detected on both arms of the chromosome (**Figure 2B** and **Supplementary figure 4**). The mean viral load was explained by a major QTL between 12.5 and 14.9 Mb ($QTL_{IV:12.5\text{-}14.9,}$ $R^2 = 0.34$) and a minor QTL between 2.5 and 4.3 Mb ($QTL_{IV:2.5\text{-}4.3}$, $R^2 = 0.08$). For both peaks, the CB4856 genotype was linked to a lower viral load. When the other traits derived from the replicates were mapped, it became clear that the peaks on the left side of chromosome IV were linked to the success of infection, whereas the peaks on the right side of chromosome IV were linked to the height of the viral load measured (**Supplementary figure 4**). This could indicate that each locus influences another aspect of OrV infection.

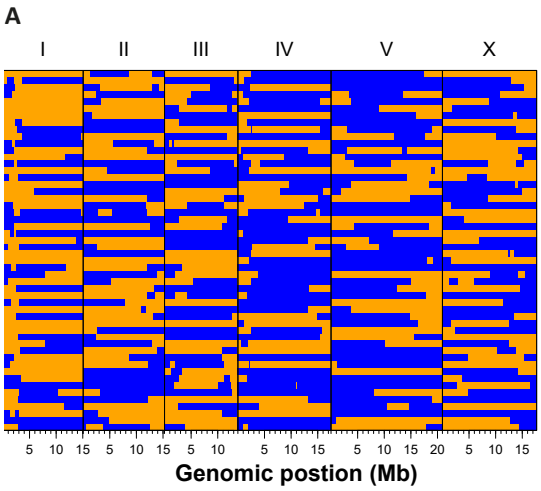


Figure 2: OrV infections in segregating populations of N2 and CB4856. (**A**) The 52 RIL strains used for the infection assays. (**B**) The QTL profile for mean viral load, the major peak is found at the end of chromosome IV at 13.3 Mb (1.5 LOD-drop interval from 12.5-14.9 Mb). A minor peak is found at the start of chromosome IV, at 2.7 Mb (1.5 LOD-drop interval from 2.5-4.3 Mb). (**C**) The 18 IL strains used to verify the chromosome IV QTL. (**D**) The result of bin-mapping with the introgression lines. The strongest association found is –log10(p) = 2.8, from 12.8-13.3 Mb.

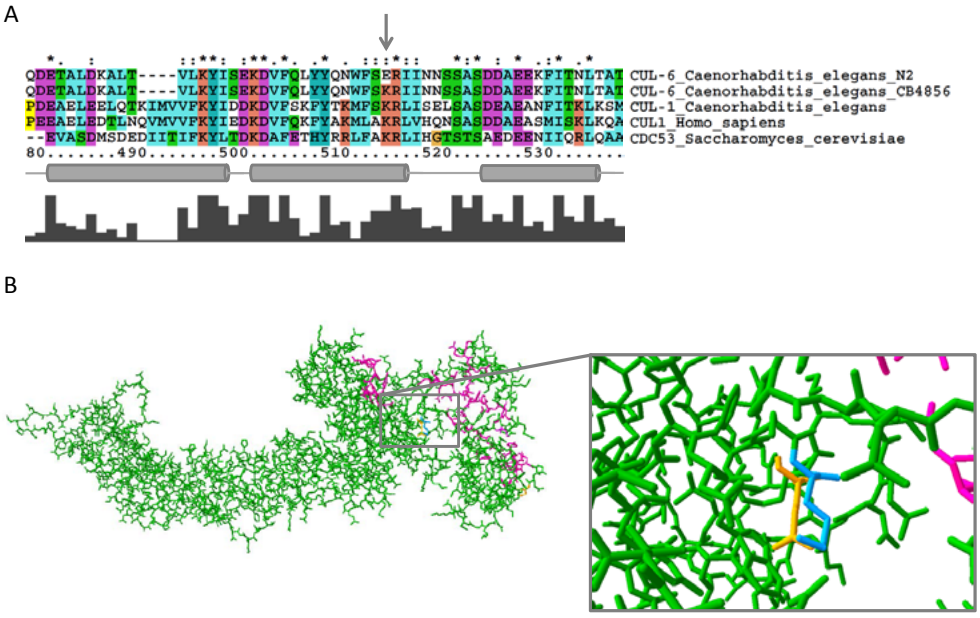


Figure 3: The gene *cul-6* contains a polymorphism between N2 and CB4856 at a conserved site. (**A**) Part of the sequence alignment between *Homo sapiens* CUL1, *Saccharomyces cerevisiae* CDC53, *C. elegans* CUL-1 and the *C. elegans* N2 and CB4856 allelic variants of CUL-6. The location of the N2 and CB4856 polymorphism is indicated with an arrow. The amino acid conservation is indicated by the grey bars at the bottom and by the annotations on top (single dot: weakly conserved, double dot: strongly conserved, asterisk: completely conserved). Colours are based on the amino acid properties. Locations of alpha-helices are indicated by cylinders [60]. (**B**) Predicted structure of *C. elegans* CUL-6. The lysine present in the CB4856 allelic variant is shown is blue, the glutamic acid present in the N2 allelic variant in orange. The RBX-1 binding domain is shown in purple.

Verification of the QTL locus by introgression lines

Analysis of introgression lines (ILs) covering chromosome IV (**Figure 2C**) confirmed the major QTL found with the RILs. ILs contain small fragments of one strain (CB4856 in this case) in the genetic background of another strain (N2 in this case) [28]. For each IL at least 7 independent experiments were conducted, and four ILs covering the two peaks were repeated an additional 12 times. The ILs were analysed on two levels: individual and as group via bin-mapping [28]. As for the RILs, also in the IL measurement large within genotype variation was found, which seems to reflect the dichotomy seen in the CB4856 replication. Four ILs displayed a lower viral load than N2, namely WN248, WN252, WN254, and WN256 (two-sided t-test, $p < 0.05$, **Supplementary file 3**). These strains cover both the $QTL_{IV:2.5\text{-}4.3}$ (WN248 and WN252) and $QTL_{IV:12.5\text{-}14.9}$ (WN252, WN254, and WN256). However, in the bin-mapping analysis, we only found a peak in the region of the major peak $QTL_{IV:12.5\text{-}14.9}$ (**Figure 2D**, **Supplementary file 4**). Combined with the outcome of the RIL mapping, the locus of interest is likely to be between 12.5 and 13.4 Mb, which is still a 900 Kb region.

The differences per individual IL show that the genetic architecture of viral load is likely to be complex. Although the introgression of WN256 is also covered by WN255 and WN258, both of these strains display an N2 trait level. Furthermore, the minor locus $QTL_{IV:2.5\text{-}4.3}$ is verified by WN248 and WN252, but not all of the other ILs that contain that QTL display a lower viral load (*e.g.* WN249 if there is no QTL in the first 2.7 Mb of chromosome IV, **Supplementary file 3**). On the one hand this could be the result of measurement error, since the experimental variation is large, on the other hand, this could indicate closely linked QTL. To dissect this trait further, we searched for potential regulators.

OrV infection affects the transcriptome of *C. elegans*

To identify the processes affected by OrV infection, we analysed the transcriptome of N2 and CB4856 after OrV infection by microarray. As for the experiments in the RILs and ILs, the strains were measured at an age of 56 hours (30 hours after infection). First, the expression data was analysed by principal components analysis. We found that the majority of expression variation was attributable to genotype (38.1%) and the axis that separated mock-infection from infection captured 8.4% of the expression variation (**Supplementary figure 5**). Second, changes in gene expression were explained over genotype (N2 or CB4856) and treatment (mock or infection) by an additive linear model. As expected, genotype showed the largest effects, leading to differential expression in 6810 genes (FDR = 0.05). The OrV infection affected 223 genes (**Supplementary Figure 6** and **Supplementary file 5**). Five of these genes are located at the QTL on chromosome IV: *rfc-3*, *ttr-45*, *pcn-1*, Y45F10B.8, and F49E11.2 (**Supplementary file 6**). Therefore, these genes could be interesting candidates. Analysis of the literature on these genes did however not yield clear links, only *rcf-3* and *pcn-1* are studied in some detail and are both linked to DNA damage repair [52, 53]. Compared to the genotypic effects, the number of genes affected by the treatment was small. This was to be expected, as a study on JU1580 also found modest differences in gene expression due to infection [54].

The differentially expressed genes were analysed by enrichment analysis on gene annotations, both for genotype as well as treatment (**Supplementary file 7**). Genotype affected the *btb* gene class, which is expected [24, 26]. Treatment-affected genes showed an overrepresentation of genes with F-box domains, EGF signalling domains, proteinase domains, and phospholipase domains. The differentially expressed genes in the phospholipase pathway are probably linked to the subcellular symptoms manifesting during OrV infection. The virus infection severely distorts the intracellular membranes of the infected cells [9, 55]. For all positive-strand RNA viruses the genome replication occurs on the host intracellular membranes [56, 57]. The OrV and its related viruses (Le Blanc and Santeuil virus) are putatively classified as Nodaviruses, having a similar genome organization and sequence homology with known *Nodaviridae* [9, 58]. Nodaviruses create membrane invaginations with a size of 50-80 nm that are associated with RNA replication [59]. The differential regulation of phospholipase genes is therefore probably a symptom of this remodelling of intracellular membranes.

The other two enriched terms, F-box proteins and EGF signalling proteins, are both playing a role in the ubiquitination pathway. EGF signalling is known to be able to activate this pathway and F-box proteins control ubiquitination of specific targets [18-20]. The ubiquitination pathway has been previously implicated in OrV infection, and one of the key players in the anti-viral ubiquitination pathway is the gene *cul-6* [25]. This gene lies directly under the mapped QTL, and is polymorphic between N2 and CB4856 (**Supplementary file 8**). The amino acid at the polymorphism location is highly conserved from humans to yeast in the closely related CUL1 and CBC53 proteins (amino acid conservation between the *C. elegans* CUL-1 and CUL-6 is 47%) [60]. The *cul-1* proteins contain a positively charged lysine at the location 428 of the polypeptide. However, the N2 strain contains a K428E polymorphism, whereas the CB4856 allele contains the lysine (**Figure 3A**). Glutamic acid is negatively charged and the position is close to the RBX-1 binding site of *cul-6*. Therefore, this polymorphism might affect RBX-1 binding efficiency as binding sites for this protein are located closely to the polymorphism location (**Figure 3B**) [60]. Therefore, *cul-6* is one of the candidates for the difference in viral load between N2 and CB4856.

Discussion

The OrV infection phenotypes of CB4856 and N2

Our experiments show that there is a difference in OrV infections between N2 and CB4856, contrary to previous studies in which differences in susceptibility have not been found [9, 16]. It is possible that this is caused by differences in the infection assay [9, 16, 51]. The infection in [9] is conducted on NGM plates, by pipetting the virus stock on the bacterial lawn. Thereafter, the infection was maintained over a longer period. Similarly, in [16], the infection was conducted in the same way and continued for a period of 7 days. Both methods infect at 23°C. One paper compared the methods, and found that the assays on agar are more reproducible given different virus concentrations [17]. However, those results are not in accordance with our observations for dilution ranges as shown here and in **Chapter 4**.

There are several possibilities for the observed differences in OrV infection between N2 and CB4856. Three of these possibilities are: i) individual nematodes in the CB4856 population are less likely to be infected, ii) in CB4856 a lower number of cells is infected, iii) or the infection topology in CB4856 is different from N2. It seems unlikely that the intestinal exposure to OrV is lower in CB4856, because CB4856 is consistently observed to have a higher pharyngeal pumping rate at younger age [28, 61]. Since we observe that CB4856 populations are less likely to be successfully infected, it is likely that the phenotypic difference is due to a difference in viral entry or any process downstream of that event. The observation that the progeny production in CB4856 is delayed upon infection might indicate a different topology of the infection. However, it is unlikely the virus infects the germline in CB4856, since the infection can be removed from a CB4856 population by bleaching [9].

Chromosome IV is implicated in natural variation in OrV infection

By exposing RILs and ILs to OrV infection, we identified two QTL on chromosome IV that are implicated in a lower viral load due to the CB4856 allele. Interestingly, a genome wide association study (GWAS) on OrV infection in *C. elegans* also implicated chromosome IV [16]. Unlike these authors, did not find a peak near the *drh-1* locus, but between N2 and CB4856 only two polymorphisms are found in the introns [26]. Still, the more distal associations uncovered by the GWAS could potentially result from the same allelic variation as $QTL_{IV:12.5\text{-}13.4}$. The GWAS identified five peaks on chromosome IV, in-between 5 and 13 Mb. This region overlaps with the fine-mapped QTL identified in this study. Therefore, the allelic variation in the wild-type strains segregating for the more distal QTL in the GWAs study should be compared to the allelic variation between N2 and CB4856.

From QTL to causal gene

In order to further dissect the QTL on chromosome IV, the research will be continued by measuring viral loads in the N2-in-CB4856 IL panel and by generating smaller introgression lines covering the major QTL on chromosome IV. Both approaches can further narrow down the QTL and enable a targeted search for the causal genetic variation. Furthermore, the N2-in-CB4856 IL panel will also identify if the QTL is more complex and is affected by interactions with the genetic background (see **chapter 3**).

In parallel, a CRISPR strategy will be followed, creating knock-out mutants of *drh-1* and *cul-6* in N2 and CB4856. Although *drh-1* is not a candidate gene, it provides a good benchmark for a strong viral load phenotype. The gene *cul-6* on the other hand is a candidate for the causal allelic variation. In combination with (smaller) ILs, a *cul-6* knock-out in both genetic backgrounds can form the basis for quantitative complementation assays that can assess the role of the allelic variation in these genes related to differences in viral load between N2 and CB4856.

Conclusion

Here we present that the wild-type strain Hawaii CB4856 displays a lower viral load upon OrV infection than the Bristol N2 strain. This result is surprising, as previous studies showed little or no phenotypic variation for viral load between these two strains [9, 16]. Using an N2xCB4856 RIL population, we mapped the trait variation to two loci on chromosome IV, and were able to robustly replicate the effect of the major locus in an IL population. Based on gene expression analysis and literature research a possible candidate for the major locus is the ubiquitination pathway gene *cul-6*. The *cul-6* gene is not the only possible candidate, and only by further dissecting the loci on chromosome IV can we make headway in finding the causal allele(s).

Acknowledgements

The authors want to thank all the people that contributed to the OrV reseach: Kobus Bosman, Henrikje Smits, Jikke Daamen, Koen Semeijn, and Yahya Zakaria Abdou Gaafar. We want to thank Marie-Anne Félix for providing us with the OrV.

Author contributions

Conceived and designed the experiments: MGS, LBS, GPP, and JEK. Performed the experiments: MGS, LvS, YW, WR, FP, RV, and JAGR. Analysed the data: MGS and LvS. Wrote the paper: MGS, LvS, GPP, and JEK.

Supplementary figures and files

The supplementary files and figures are deposited at: http://marksterken.nl, under 'PhD thesis'.

Supplementary figure 1: The time until the first egg was laid as measured in mock-infected and infected N2 and CB4856 nematodes. The egg laying only delays in CB4856 when infected with OrV. The effect is small (102 minutes), but significant (ANOVA, $p < 0.01$).

Supplementary figure 2: Overview of the normalized viral load of the RIL and the IL experiments. All of the viral load measurements are shown, organized by population type (RIL in green and IL in purple). The parental strains included in the panels (N2 in orange and CB4856 in blue), were measured in the same batches as the inbred panels shown.

Supplementary figure 3: The Pearson correlations between the viral load traits from the RIL panel (**Supplementary file 2**). Two main groups can be discerned: those traits related to the maximum viral load per strain and those related to the minimum viral load per strain. The variation groups with neither.

Supplementary figure 4: The QTL maps for all of the traits derived from the RIL panel. On the x-axis the genome position is indicated and on the y-axis the significance (in –log10(p)). The significance of association per trait is indicated by the solid black line. The horizontal segmented line indicates the threshold for FDR = 0.05. Red boxes at the bottom of the plot indicate significant QTL and the confidence interval of the QTL. Only one QTL per chromosome was allowed.

Supplementary figure 5: Principal component analysis of gene expression in mock or OrV infected N2 and CB4856. The first 6 axis of the PCO analysis are shown, the amount of variation explained is indicated on the axis annotation. The circles represent N2 samples and the triangles represent CB4856 samples. Mock samples are indicated in grey and infected samples in red. The first axis (PCO1), which captures 38.1% of the variation, separates samples on genotype. The 4th axis separates the mock from the infected samples (8.4% of variation).

Supplementary figure 6: Volcano plot of the linear model outcome. On the x-axis the effect is plotted (log2) and on the y-axis the significance is plotted (-log10(p)). The largest effect came from genotype (N2 or CB4856), affecting 10564 spots (FDR = 0.05), whereas the treatment (mock or infected) affected 275 spots (FDR = 0.05). Significant spots are indicated in red.

Supplementary file 1: The strains used in this study. A matrix of the strains used in this study is given, together with the genotypes. The strains can be found in the columns and the genotypes in the rows.

Supplementary file 2: The aggregated (derived) traits for the RIL panel. For each of the RILs (columns) the value is given for: mean viral load, median viral load, maximum viral load, minimum viral load, the number of successful infections, the mean viral load of the successful infections (Mean_infected), the median viral load of the successful infections (Median_infected), the variation in viral load and the viral loads per replication experiment (RIL_1, RIL_2, and RIL_3).

Supplementary file 3: The outcome of the analysis of individual ILs versus N2. The strain is shown, with the effect versus N2 (log2 viral load), the significance of the effect (-log10(p)), and the fit (R^2). Also the introgression location (CB4856 in an N2-background) is given per IL.

Supplementary file 4: The outcome bin-mapping the viral load in the IL population. The marker name and location (Name, Chr, and Pos) is given, The effect (log2 viral load), significance (-log10(p)), and R^2 of the fit is shown.

Supplementary file 5: Outcome of the linear model explaining gene-expression over genotype and treatment. The trait (array spot) and the annotation is given (gene name, Wormbase ID, sequence name, chromosome, and position). Per term in the linear model (genotype or treatment), the outcome of the linear model is given. The effect indicates the difference between N2 and CB4856 for genotype (positive is higher in N2), and the difference between infected and mock in treatment (positive is higher in mock). The significance is given in –log10(p).

Supplementary file 6: List of the differentially expressed genes that are located at the QTL.

Supplementary file 7: Outcome of the gene enrichment analysis, organized by genotype and treatment. The differentially expressed genes (**Supplementary file 5**) were tested for enrichment and this file shows the categories that were enriched. The annotation group is given (*e.g.* Anatomy or Gene class), the group (*e.g. btb*-genes within Gene class), the number of genes in the group, the overlap with the differentially expressed genes, and the significance of the overlap (in –log10(p)).

Supplementary file 8: The alignment between the Cullins. The polymorphism between N2 and CB4856 is located at position 516 of the alignment.

References

1. **Obbard, D.J., et al.,** *Natural selection drives extremely rapid evolution in antiviral RNAi genes.* **Curr Biol, 2006. 16(6): p. 580-5.**
2. **Vasseur, E., et al.,** *The selective footprints of viral pressures at the human RIG-I-like receptor family.* **Human Molecular Genetics, 2011. 20(22): p. 4462-4474.**
3. **Enard, D., et al.,** *Viruses are a dominant driver of protein adaptation in mammals.* **Elife, 2016. 5.**
4. **Galvani, A.P. and M. Slatkin,** *Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele.* **Proceedings of the National Academy of Sciences of the United States of America, 2003. 100(25): p. 15276-15279.**
5. **Khakoo, S.I., et al.,** *HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection.* **Science, 2004. 305(5685): p. 872-4.**
6. **Dean, M., et al.,** *Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study.* **Science, 1996. 273(5283): p. 1856-62.**
7. **Rasmussen, A.L., et al.,** *Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance.* **Science, 2014. 346(6212): p. 987-91.**
8. **Samuel, B.S., et al.,** *Caenorhabditis elegans responses to bacteria from its natural habitats.* **Proc Natl Acad Sci U S A, 2016.**
9. **Felix, M.A., et al.,** *Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses.* **PLoS Biol, 2011. 9(1): p. e1000586.**
10. **Felix, M.A. and F. Duveau,** *Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae.* **BMC Biol, 2012. 10: p. 59.**
11. **Troemel, E.R., et al.,** *Microsporidia are natural intracellular parasites of the nematode Caenorhabditis elegans.* **PLoS Biol, 2008. 6(12): p. 2736-52.**
12. **Maguire, S.M., et al.,** *The C. elegans touch response facilitates escape from predacious fungi.* **Curr Biol, 2011. 21(15): p. 1326-30.**
13. **Frezal, L. and M.A. Felix,** *C. elegans outside the Petri dish.* **Elife, 2015. 4.**
14. **Petersen, C., P. Dirksen, and H. Schulenburg,** *Why we need more ecology for genetic models such as C. elegans.* **Trends Genet, 2015.**
15. **Lu, R., et al.,** *Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans.* **Nature, 2005. 436(7053): p. 1040-3.**
16. **Ashe, A., et al.,** *A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity.* **Elife, 2013. 2: p. e00994.**
17. **Ashe, A., et al.,** *Antiviral RNA Interference against Orsay Virus Is neither Systemic nor Transgenerational in Caenorhabditis elegans.* **J Virol, 2015. 89(23): p. 12035-46.**
18. **Kipreos, E.T.,** *Ubiquitin-mediated pathways in C. elegans.* **WormBook, 2005: p. 1-24.**
19. **Kipreos, E.T. and M. Pagano,** *The F-box protein family.* **Genome Biol, 2000. 1(5): p. REVIEWS3002.**
20. **Papaevgeniou, N. and N. Chondrogianni,** *The ubiquitin proteasome system in Caenorhabditis elegans and its regulation.* **Redox Biol, 2014. 2: p. 333-47.**
21. **Nayak, S., et al.,** *The Caenorhabditis elegans Skp1-related gene family: diverse functions in cell proliferation, morphogenesis, and meiosis.* **Curr Biol, 2002. 12(4): p. 277-87.**
22. **Yamanaka, A., et al.,** *Multiple Skp1-related proteins in Caenorhabditis elegans: diverse patterns of interaction with Cullins and F-box proteins.* **Curr Biol, 2002. 12(4): p. 267-75.**
23. **Jones, D., et al.,** *Functional and phylogenetic analysis of the ubiquitylation system in Caenorhabditis elegans: ubiquitin-conjugating enzymes, ubiquitin-activating enzymes, and ubiquitin-like proteins.* **Genome Biol, 2002. 3(1): p. RESEARCH0002.**
24. **Volkers, R.J., et al.,** *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93.**

25. **Bakowski, M.A., et al.,** *Ubiquitin-mediated response to microsporidia and virus infection in C. elegans.* **PLoS Pathog, 2014. 10(6): p. e1004200.**
26. **Thompson, O.A., et al.,** *Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856.* **Genetics, 2015. 200(3): p. 975-89.**
27. **Vinuela, A., et al.,** *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37.**
28. **Doroszuk, A., et al.,** *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110.**
29. **Reddy, K.C., et al.,** *A polymorphism in npr-1 is a behavioral determinant of pathogen susceptibility in C. elegans.* **Science, 2009. 323(5912): p. 382-4.**
30. **Rodriguez, M., et al.,** *Genetic variation for stress-response hormesis in C. elegans lifespan.* **Exp Gerontol, 2012. 47(8): p. 581-7.**
31. **Harvey, S.C., A. Shorto, and M.E. Viney,** *Quantitative genetic analysis of life-history traits of Caenorhabditis elegans in stressful environments.* **BMC Evol Biol, 2008. 8: p. 15.**
32. **Li, Y., et al.,** *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222.**
33. **Rockman, M.V. and L. Kruglyak,** *Recombinational landscape and population genomics of Caenorhabditis elegans.* **PLoS Genet, 2009. 5(3): p. e1000419.**
34. **Brenner, S.,** *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94.**
35. **Zahurak, M., et al.,** *Pre-processing Agilent microarray data.* **BMC Bioinformatics, 2007. 8: p. 142.**
36. **Smyth, G.K. and T. Speed,** *Normalization of cDNA microarray data.* **Methods, 2003. 31(4): p. 265-73.**
37. **Benjamini, Y. and Y. Hochberg,** *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* **Journal of the Royal Statistical Society Series B-Methodological, 1995. 57(1): p. 289-300.**
38. **Stein, L., et al.,** *WormBase: network access to the genome and biology of Caenorhabditis elegans.* **Nucleic Acids Research, 2001. 29(1): p. 82-86.**
39. **Yook, K., et al.,** *WormBase 2012: more genomes, more data, new website.* **Nucleic Acids Res, 2012. 40(Database issue): p. D735-41.**
40. **Niu, W., et al.,** *Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans.* **Genome Res, 2011. 21(2): p. 245-54.**
41. **Gerstein, M.B., et al.,** *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* **Science, 2010. 330(6012): p. 1775-87.**
42. **Tepper, R.G., et al.,** *PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity.* **Cell, 2013. 154(3): p. 676-90.**
43. **Ogata, H., et al.,** *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Res, 1999. 27(1): p. 29-34.**
44. **Larkin, M.A., et al.,** *Clustal W and Clustal X version 2.0.* **Bioinformatics, 2007. 23(21): p. 2947-8.**
45. **Altschul, S.F., et al.,** *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* **Nucleic Acids Res, 1997. 25(17): p. 3389-402.**
46. **Remmert, M., et al.,** *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* **Nat Methods, 2012. 9(2): p. 173-5.**
47. **Guex, N. and M.C. Peitsch,** *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* **Electrophoresis, 1997. 18(15): p. 2714-23.**
48. **Sali, A. and T.L. Blundell,** *Comparative protein modelling by satisfaction of spatial restraints.* **J Mol Biol, 1993. 234(3): p. 779-815.**
49. **Benkert, P., M. Biasini, and T. Schwede,** *Toward the estimation of the absolute quality of individual protein structure models.* **Bioinformatics, 2011. 27(3): p. 343-50.**
50. **Mariani, V., et al.,** *Assessment of template based protein structure predictions in CASP9.* **Proteins, 2011. 79 Suppl 10: p. 37-58.**
51. **Sterken, M.G., et al.,** *A heritable antiviral RNAi response limits Orsay virus infection in Caenorhabditis elegans N2.* **PLoS One, 2014. 9(2): p. e89760.**

52. **Deichsel, A., J. Mouysset, and T. Hoppe,** *The ubiquitin-selective chaperone CDC-48/p97, a new player in DNA replication.* **Cell Cycle, 2009. 8(2): p. 185-90.**
53. **Kim, S.H. and W.M. Michael,** *Regulated proteolysis of DNA polymerase eta during the DNA-damage response in C. elegans.* **Mol Cell, 2008. 32(6): p. 757-66.**
54. **Sarkies, P., et al.,** *Competition between virus-derived and endogenous small RNAs regulates gene expression in Caenorhabditis elegans.* **Genome Res, 2013. 23(8): p. 1258-70.**
55. **Franz, C.J., et al.,** *Orsay, Santeuil and Le Blanc viruses primarily infect intestinal cells in Caenorhabditis nematodes.* **Virology, 2014. 448: p. 255-64.**
56. **Ahlquist, P.,** *Parallels among positive-strand RNA viruses, reverse-transcribing viruses and double-stranded RNA viruses.* **Nature Reviews Microbiology, 2006. 4(5): p. 371-382.**
57. **Miller, S. and J. Krijnse-Locker,** *Modification of intracellular membrane structures for virus replication.* **Nat Rev Microbiol, 2008. 6(5): p. 363-74.**
58. **Franz, C.J., et al.,** *Complete genome sequence of Le Blanc virus, a third Caenorhabditis nematode-infecting virus.* **J Virol, 2012. 86(21): p. 11940.**
59. **Miller, D.J., M.D. Schwartz, and P. Ahlquist,** *Flock house virus RNA replicates on outer mitochondrial membranes in Drosophila cells.* **J Virol, 2001. 75(23): p. 11664-76.**
60. **Zheng, N., et al.,** *Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex.* **Nature, 2002. 416(6882): p. 703-9.**
61. **Andersen, E.C., et al.,** *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156.**

Chapter 6

Natural variation of gene expression reveals cross-talk between longevity and stress tolerance in *Caenorhabditis elegans*

L. Basten Snoek*, Mark G. Sterken*,
Roel P. J. Bevers, Rita J. M. Volkers,
Arjen van't Hof, Rachel Brenchley,
Joost A. G. Riksen, Eva van Cleef,
Jorik Bot, Olga Valba,
Sreenivas Chavali, Benjamin Lang,
Mirko Francesconi, Sergei Nechaev,
Owen Thompson, Olga Vasieva,
M. Madan Babu, Ben Lehner,
Andrew Cossins, Jan E. Kammenga

*:Authors contributed equally

This chapter has been submitted for publication in a modified form.

Abstract

The genetics of aging has primarily been studied using laboratory-derived single gene mutations in model animals. These mutants often display extreme lifespan accompanied by increased stress resistance. Yet, the identification of genes that underlie natural variation in lifespan and stress tolerance remains a major challenge due to the polygenic character of these traits. Here we leveraged the detection of causal genes by analyzing regulatory variation of gene expression (eQTL) in heat stressed *C. elegans*. We detected a *trans*-band on chromosome *IV* that affects the expression of over 200 genes. Fine mapping using introgression lines confirmed the location of the *trans*-band. Expression profiling of mutants of candidate genes, showed that *egl-4*, *eor-1,* and *cmk-1* are involved in transcriptional regulation of genes mapping to this location. Phenotyping revealed that this locus is linked to lifespan and stress tolerance in which *egl-4*, *eor-1,* and *cmk-1* play a regulatory role. Together our results show that the genetics of gene expression provides a powerful tool for detecting genes and expression patterns underlying natural variation in lifespan and stress tolerance.

Introduction

Since the early 1990's evidence has accumulated showing the rate of biological aging is modulated by genes involved in stress responses. Stress and aging are strongly linked; animals displaying stress tolerance often have a prolonged lifespan. Intracellular signaling pathways, like the insulin-like pathway, play a role in this connection by modulating the expression of stress response genes [1, 2]. The primary model animals to study genetic control of aging are laboratory-derived mutants of nematodes, fruit-flies and mice [3, 4]. These mutants often display extreme longevity accompanied by increased stress resistance [5-8]. For example, mutations in the insulin/IGF-1-like receptor protein DAF-2 doubled lifespan in the nematode *Caenorhabditis elegans* [9] and depend on a FOXO-like transcription factor DAF-16 that is required for stress responses [10]. In *Drosophila melanogaster* mutations in c-Jun N-terminal kinase lead to reduced oxidative damage and a 50% increase in lifespan compared to wild-type flies [11]. Mice mutated in the insulin-like growth factor type 1 receptor (IGF-1R) live up to 30% longer and display increased resistance to metabolic stress [12].

However, these phenotypes fall well beyond the variation in lifespan typically observed in natural populations, which varies between 10-25% [13]. Although exceptional differences in longevity within species have been documented, extreme lifespan elongation displayed by laboratory mutants are not commonly observed in natural populations. Moreover, despite the identification of hundreds of mutations in genes leading to extended lifespan in model organisms, the identification of the genetic variants that underlie lifespan variation in nature remains a major challenge due to the highly polygenic character of this trait [14, 15].

C. elegans lifespan and stress resistance are both quantitative traits influenced by natural genetic variation [16-19]. To genetically dissect the causes of this variation, we used a population of homozygous recombinant inbred lines (RILs) derived from a cross between two highly divergent wild type strains Bristol N2 and Hawaii CB4856 [16, 20]. We used the stress-induced transcriptional architecture (expression-QTL or eQTL) as the basis for detecting the causal regulators of both lifespan and stress tolerance. eQTLs are polymorphic loci associated with variation in gene expression and have been found to be characteristic for age differences in *C. elegans* [21].

First, we mapped eQTLs that affect the expression of hundreds of genes in response to a heat stress compared to control conditions. A heat-stress specific *trans*-band was detected on the left arm of chromosome *IV*. We confirmed this locus using introgression lines. By expression profiling of mutants of candidate genes, we found that the genes *egl-4*, *eor-1*, and *cmk-1* genes are regulators of the genes with an eQTL at this locus. Phenotyping revealed that this locus is linked to lifespan and stress tolerance in which *egl-4*, *eor-1,* and *cmk-1* play a regulatory role. Together our results show that the genetics of gene expression provides a powerful tool for detecting genes and expression patterns underlying natural variation in lifespan and stress tolerance

Materials and Methods

Strains used

The wild-types N2 and CB4856, 54 CB4856 x N2 RILs and 57 CB4856 x N2 ILs were used [16, 20, 22]. A matrix with the strain names and the genetic map can be found in **Supplementary file 1**. All 16 mutant strains were provided by the CGC, *elt-6*, VC1534(gk723); *gcy-37,* RB626(ok384); *clp-7*, RB2084(ok2750); *egl-4*, FK223(ks60); *nhr-92*, VC3265(gk3168); *clp-6*, RB1509(ok1779); *ced-2*, CB3257(e1752); *egl-18*, JR2370(ok290); *cmk-1*, PY1589(oy21); *nhr-122*, VC1232(gk560); *nhr-287*, VC1758(gk1014); C54E4.2, VC2213(gk1003); *lin-1*, MT7567(sy254); *eri-5,* WM171 (tm1705); *eel-1*, VC1100 (ok1575); *eor-1*, RB1166 (ok1127).

Nematode culturing

The strains were kept on 6 cm Nematode Growth Medium (NGM) dishes containing Escherichia coli strain OP50 as food source [23]. Strains were kept in maintenance culture at 12°C, the standard growing temperature for experiments was 20°C. Fungal and bacterial infections were cleared by bleaching [23]. The strains were cleared of males prior to the experiments by selecting L2 larvae and placing them individually in a well in a 12 wells plate at 20°C. The strains were screened for male offspring after 3 days and only the 100% hermaphrodites populations were transferred to fresh 9 cm NGM dishes containing E. coli OP50 and grown until starved.

Control and heat stress experiments for studying transcriptomics

The experiments were started by transferring a starved population to a fresh 9 cm NGM dish. This population was grown for 60 hours at 20°C to obtain egg laying adults. These populations were bleached for synchronization and the eggs were placed on a fresh 9 cm NGM dish. Under control conditions these populations were grown for 48 hours at 20°C, for heat stress conditions the populations were grown for 46 hours at 20°C and 2 hours at 35°C. The populations were collected after exactly 48 hours by washing off the plate with M9 buffer and flash freezing in liquid nitrogen.

Lifespan measurements

The experiments were also started by transferring a starved population to a fresh 9 cm NGM dish. This population was grown for 60 hours at 20°C to obtain egg laying adults. These populations were bleached for synchronization and the eggs were placed on a fresh 9 cm NGM dish. These populations were grown for 48 hours at 20°C until the L4 stage was reached. 30-40 nematodes were then collected with M9 buffer and transferred to a 9 cm NGM dish containing FUDR [24]. For all experiments a total number of animals tested per genotype per condition ranged from 86 to 316. The populations undergoing heat stress treatment were exposed to 35°C for 4 hours and

thereafter placed at 20°C. For a better detection of any survival effect after application of heat-stress, we used 4 hours instead of 2 hours (used in the transcriptomics experiment) because this resulted in a more severe effect on survival [16]. The number of dead and live nematodes was counted manually every day. Death was confirmed by a lack of mobility on prodding the animals with a worm pick.

Lifespan curves were analyzed using the survival package in R. Survival curves were compared to N2 within each treatment (control or heat shock). The curves were also compared by treatment with in the same genotype. The comparison was done for the total curve, from day 0-4, from day 5-13, and from day 14 onwards by the log-rank test. Day 13 was chosen as it is the mean lifespan in the control condition and therefore represents the first ~50% of the survival curve.

High density map from whole genome sequencing data

DNA preparation

Bleach-sterilized nematode strains were grown at 20°C on 2% agar-NGM plates seeded with *Escherichia coli* NA22 strain. The animals were transferred to M9 medium and kept at 4°C allowing them to sink into a pellet, as a result of reduced activity. The supernatant containing the (slower sinking) bacteria was removed once the nematodes settled. The nematodes were allowed to settle twice more in fresh M9 to minimize the bacterial proportion of sequence reads downstream. The nematodes were dissolved and digested in a lysis buffer containing a high concentration of SDS (1%) and proteinase K. DNA was extracted using a standard Phenol Chloroform Isoamyl alcohol (Sigma-Aldrich, St Louis, MO, USA) procedure and additionally cleaned with Ampure beads (Beckman Coulter, Indianapolis, IN, USA).

Sequence method and library preparations

Whole genome sequencing was performed on a SOLiD 4 platform (Life Technologies) using the manufacturers 'Barcoded Fragment Library Preparation' instructions and recommended reagents and equipment. Barcoded libraries were size selected for 250bp fragments and sequenced in a pool of 20 per slide. Data were processed using the standard SOLiD 4 pipeline.

Alignment and SNP calling

SOLiD reads from the 106 lines were mapped to the N2 reference genome (version WS220), obtained from Wormbase, using the BWA (version 0.5.9) "aln" followed by "samse" commands [25-27]. Default parameters for "aln" were used except for "-q 20", controlling read trimming based on quality. Each strain was mapped separately, and labelled using the read group tags (RG), and then non-mapping and non-uniquely mapping reads were filtered out. BAM files were combined using the SAMtools (version 0.1.18) "merge" command [27]. SNPs were identified using the Genome Analysis ToolKit (GATK) (version 1.2.4) [28, 29]. Firstly, reads were re-aligned around INDELs using the "IndelRealigner" command to create an improved BAM file. Then, GATK's

"UnifiedGenotyper" command identified the SNPs and the results were quality filtered using the GATK "VariantFiltration" command.

As the 106 lines contained only certain regions of the CB4856 genome, further filtering of the identified SNPs was performed. The reason being most reads covering each part of the genome were from N2 and only a few strains at a given region would show SNPs belonging to CB4856. SNPs were called in each strain separately, and even though coverage over each strain was very low for normal variant detection (average of 5.1X for RILs and 5.6X for ILs), differences were considered to be true SNPs when they were present in >1 strain. This allowed true polymorphisms to be separated from any sequencing errors.

Genotypes and genetic map

The SNPs detected by the alignment to the N2 reference genome were filtered against the SNP data from the million mutations project to obtain only the reliable SNPs [30]. Of the detected SNPs 96.6% was found back in the CB4856 polymorphisms of the million mutations project (WS230 release), leaving 77,545 unique SNPs for constructing the genetic map. The SNP density was determined per 10kb bins and recombination events were recognized as transition of an area where there were no SNPs in 10 consecutive bins into an area where there were SNPs and the other way around. It was not allowed to have two recombination events within 10 consecutive bins (100kb). The 10kb bin where the first SNPs were detected was marked as the recombination event. Before use in genotyping, the map was filtered for informative markers – that is - markers indicating a recombination event in at least one of the lines. This resulted in a map of 729 informative markers, each indicating the location of the recombination events within 20 kb (**Supplementary figure 1 and 2**).

Transcript profiling

RNA isolation

The RNA of the RIL and IL samples was isolated using the RNeasy Micro Kit from (Qiagen, Hilden, Germany). The 'Purification of Total RNA from Animal and Human Tissues' protocol was followed, with a modified lysing procedure. The frozen pellets were lysed in 150 µl RLT buffer, 295 µl RNAse-free water, 800 µg/ml proteinase K and 1% ß-mercaptoethanol. The suspension was incubated at 55°C and 1000 rpm in a Thermomixer (Eppendorf, Hamburg, Germany) for 30 minutes or until the sample was clear. After this step the prescribed protocol was followed.

The RNA of the mutants strains was isolated using a Maxwell® 16 AS2000 instrument with a Maxwell® 16 LEV simplyRNA Tissue Kit (both Promega Corporation, Madison, WI, USA). The prescribed protocol was followed, only the lysis step was modified. Here 200 µl homogenization buffer, 200 µl lysis buffer and 500 µg/ml proteinase K were added to each sample. This suspension was incubated at 65°C and 1000 rpm in a Thermomixer (Eppendorf, Hamburg, Germany) for

10 minutes. Before adding the samples to the cartridges they were cooled on ice for 1 minute and thereafter the standard protocol was followed.

cDNA synthesis, labelling and hybridization

The 'Two-Color Microarray-Based Gene Expression Analysis; Low Input Quick Amp Labeling' -protocol, version 6.0 from Agilent (Agilent Technologies, Santa Clara, CA, USA) was followed, starting from step 5. The *C. elegans* (V2) Gene Expression Microarray 4X44K slides, manufactured by Agilent were used. Before starting cDNA synthesis the quality and quantity of the RNA was measured on the NanoDrop, and the integrity of the RNA was determined by loading 3 µL of sample RNA on a 1% agarose gel.

Data extraction and normalization

The microarrays were scanned by an Agilent High Resolution C Scanner with the recommended settings. The data was extracted with Agilent Feature Extraction Software (version 10.5), following manufacturers' guidelines. Normalization of the data was executed in two parts, first the RILs and the ILs, second the mutant strains. For normalization "R" (version 3.0.2 x 64) with the Limma package was used. The data was not background corrected before normalization (as recommended by [31]). Within-array normalization was done with the Loess method and between-array normalization was done with the Quantile method [32]. The obtained single channel normalized intensities were log2 transformed and used for further analysis.

Transcriptional response to heat stress

The transcriptional response to heat stress was determined by explaining the gene expression over the treatment with a linear model,

$$y_i \sim T + e_i$$

where y is the log2 normalized intensity as measured by microarray of spot i (i = 1, 2, ..., 45220) and T is the treatment (either control or heat shock). This analysis ignored genotype.

A significance threshold was determined by permutation analysis. The data was permutated and 1000 times, and in each permutation the most significant p-value was notated. These were ordered and we used the 5% highest value as threshold, -log10(p) >5.82. This strict threshold ensures only the very strongly associated spots are selected (outcome of the analysis, see **Supplementary file 2**).

Expression quantitative trait locus analysis

eQTL mapping and threshold determination

The eQTL mapping was done in "R" (version 3.0.2 x 64). Only measurements that passed quality control were used for mapping. The gene-expression data was fitted to the linear model,

$$y_{i,j} \sim x_j + e_j$$

where y is the log2 normalized intensity as measured by microarray of spot i (i = 1, 2, ..., 45220) of RIL j. This is explained over the genotype (either CB4856 or N2) on marker location x (x = 1, 2, ..., 729) of RIL j.

The genome wide significance threshold was determined via permutation, where the log2 normalized intensities were randomly distributed per gene over the genotypes. The randomized data was tested using the same model as for the eQTL mapping. This was repeated for 10 randomized datasets. A false discovery rate was used to determine the threshold (as recommended for multiple testing under dependency) [33],

$$\frac{FDS}{RDS} \leq \frac{m_0}{m} q \cdot \log(m)$$

where FDS is the outcome of the permutations and RDS is the outcome of the eQTL mapping at a specific significance level. The value of m_0, the number of true null hypotheses tested, was 45220-RDS and for the value of m, the number of hypotheses tested, the number of spots (45220) was taken. The q-value was set at 0.025. For the control set the threshold was –log10(p) > 3.9 and for the heat-shock set the threshold was –log10(p) > 3.8.

eQTL analysis

The eQTL were divided in *cis*- and *trans*-regulated eQTL, the window for *cis* regulation was set at 1 Mb, at either side of the gene.

The presence of condition specific *trans*-bands was determined in the eQTL data at the FDR=0.05 significance level. The number of *cis*- and *trans*-eQTL were counted per bin of 0.5Mb and tested versus the number of expected eQTL (if these were evenly distributed). This enrichment was tested using a chi-squared test (R, version 3.0.2 x64), a *trans*-band was called at p < 0.01 (**Supplementary figure 4**). The reported *trans*-bands were robustly detected over a number of bin-sizes (0.25-1.5 Mb). Differences in the mapped location of *cis*-eQTL become more apparent at smaller bin-sizes; at 1.5Mb no enrichments for *cis*-eQTL are detected anymore.

eQTL validation by ILs

Thirteen different introgression lines covering the chromosome II (WN225, WN226, and WN227), chromosome III (WN228, WN229, WN230, and WN232), and chromosome IV (WN245, WN246, WN248, WN250, WN251, and WN252) *trans*-bands under heat-stress were taken and used in the same experimental setup as the RIL panel. To determine the effect of the introgression (and interactions of the background with the introgression) specifically, the expression was normalized against the mean expression as measured in the background strain (N2), by

$$R_{i,j} = \log 2(\frac{y_{i,j}}{\bar{y}_{i,N2}})$$

where R is the log2 relative expression of spot i (i = 1, 2, ..., 45220) in line j (IL) and y is the

intensity (not the log2-transformed intensity) of spot i in line j. This is divided by the average intensity y as calculated in the N2 strains (n=3 in heat-shock).

The relative expression was correlated with the expected QTL effect (based on the genotype of the strain at the location of the QTL), therefore a positive correlation indicates the expression in that strain is in accordance with the expected QTL pattern. A correlation around 0 means that the expression pattern is not in accordance with the expected QTL pattern and a negative correlation means that there is an inverse relation between the expression in the strain and the expectation from the model. For the correlation analysis the Pearson correlation was calculated and the significance of the correlation (**Supplementary file 6**).

To test for the *trans*-band, the eQTL from the *trans*-band location were tested by correlation analysis. The eQTL effects measured in the RIL panel for the heat-shock *trans*-bands (II:12.0-13.5 Mb, III:0.5-2.0 Mb, and IV:1.0-2.5 Mb) were correlated with the log2 relative expression per IL. The *cis*- and *trans*-eQTL effects were tested separately. A correlation was called significant if $p < 1*10^{-10}$, based on the highest significance found in the N2 controls. The analysis was robust for the selection of eQTL with larger effects in the ILs ($|R| > 0.5$, 1.0, and 1.5). As such effects were not seen in the N2 controls, there was no reference for by-chance effects. Furthermore, we also analysed the overlap in eQTL by enrichment analysis over the overlap ($|R| > 1$ in the ILs), which gave a qualitatively the same result (data not shown).

Candidate regulators on chromosome IV

The chromosome IV *trans*-band locus (1.0-2.5 Mb) contained 206 genes, of which 28 had one or more mutants available at the *Caenorhabditis* Genetics Center (CGC) containing a defined mutation in a single gene (based on information from Wormbase WS220). The first selection was made by considering candidate genes highly connected to the genes mapping to the *trans*-band (eQTL) in WormNet (V.2) [34]. Direct connections were counted, as well as connections via one, two, or three steps to the candidate gene. Of the 206 candidate genes on the *trans*-band locus, 108 were present in WormNet and of the 276 genes mapping to the *trans*-band locus (eQTL), 186 were present in WormNet. The candidate genes present in WormNet are connected to all (or almost all) of the 186 genes with an eQTL present in WormNet. The largest difference in connections was found for genes connected via exactly one other gene (one step). Based on this score we made the first selection which led to the selection of: *cmk-1, elt-6, lin-1*, *clp-7, egl-18, clp-6, egl-4,* C54E4.2 (*test-1*), and *ced-2*. We also selected two genes with a low connection: *eri-5*, and *gcy-37* and three genes that were not present in WormNet, but could be regulatory due to their transcription factor function: *nhr-92*, *nhr-122*, and *nhr-287* (**Supplementary file 7**).

Next to the WormNet analysis based on the *trans*-band locus we also focussed on the larger region of the *trans*-band (0-6.0 Mb). Based on literature research we also tested the genes *eel-1* and *eor-1*. We selected *eor-1* based on the involvement in the Ras signalling pathway [35], and

eel-1 was selected because it regulates SKN-1, which is connected to the DAF-2 insulin signalling pathway [36, 37].

eQTL enrichment in laboratory induced mutants

To determine the effect of the mutation specifically, the expression was normalized against the mean expression as measured in the background strain (N2), by

$$R_{i,j} = \log 2(\frac{y_{i,j}}{\bar{y}_{i,N2}})$$

where R is the log2 relative expression of spot i (i = 1, 2, ..., 45220) in line j (mutant) and y is the intensity (not the log2-transformed intensity) of spot i in line j. This is divided by the average intensity y as calculated in the N2 strains (n=7 over 4 batches).

First, we applied correlation analysis to determine the involvement of the single mutants in the chromosome IV *trans*-band. However, we noticed that the eQTL effect directions played a strong role in these correlations. Therefore, we investigated the variation in gene expression in the mutants by principal component analysis using the genes mapping to the *trans*-band on chromosome IV (1.0-2.5Mb). This PCO investigation was combined with an enrichment analysis of *trans*-eQTLs mapping to the *trans*-band as fraction of all *trans*-eQTLs found in the differentially expressed genes compared to N2. Both analyses were robust for different selection thresholds (based on the log2 ratio of expression with N2), ranging from 0.5-2.0.

The most significantly associated mutants, *cmk-1*, *egl-4*, and *eor-1*, were re-tested in a separate experiment with three biological replicas. The analysis was conducted in the same way as for the first set of mutants.

Gene enrichment analysis

Gene group enrichment analysis was done using a hypergeometric test and several databases with annotations. The databases used were: The WS220 GO-annotation, anatomy terms, protein domains, and gene classes [25, 26]; the MODENCODE release 32 transcription factor binding sites (www.modencode.org) [38, 39], which were mapped to transcription start sites (according to [40]); the KEGG pathway release 65.0 (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/) [41].

Enrichments were selected based on the following criteria: hypergeometric test p-value < 0.001, size of the category n>3, size of the overlap n>2. Enrichments were calculated based on gene-names, not on spots.

Results

RIL sequencing and transcriptome analyses

The genomes of 49 RILs were fully sequenced to construct a high-resolution genetic map consisting of 729 markers informative of recombination events (**Figure 1A**, **Supplementary figure 1** and **2**). To gain insight into the transcriptional architecture of stress resistance, we measured genome-wide gene expression in the RILs at the L4 stage of development in control conditions (20 °C) and following a heat stress (2h at 35 °C). In response to the heat shock, 2642 genes were down-regulated and 2125 genes were up-regulated (**Supplementary figure 3** and **Supplementary file 2**; -log10(p)>5.82). A description of the functional analysis of these genes can be found in **Supplementary file 3** and **4**.

eQTL analyses and *trans*-band confirmation

To identify regulatory loci, we mapped expression quantitative trait loci (eQTL) in both the control condition and following heat stress, identifying 1886 and 2382 genes with an eQTL, respectively (**Figure 1B**, **Supplementary file 5**). The *cis*-eQTL were very similar between the two conditions, but the *trans*-eQTL were mainly condition specific (**Table 1**). *Trans*-eQTL comprised 38-47% of all eQTLs and were comprised of five specific *trans*-bands, loci that regulate the abundance of many transcripts (**Supplementary figure 4**, **Table 2**). Two major *trans*-bands influenced variation in gene expression in the control condition and three major *trans*-bands influenced gene expression following heat stress. The eQTL *trans*-band on the left arm of chromosome *IV* (*IV*: 1.0-2.5 Mb) was specific to the heat stress condition and affected the expression levels of the largest number of genes. In total 276 genes with an eQTL were mapped to this locus, 14 *cis*- and 262 *trans*-eQTLs. Of the 262 *trans*-eQTLs, 234 (89%) showed an increase in expression linked to the CB4856 genotype.

Table 1: Number of genes with an eQTL

	Control[1]	Heat stress[2]	Overlap[3]
cis-eQTL	1171	1264	796
trans-eQTL	715	1118	170
Total	1886	2382	966

[1]: The eQTL threshold for control was –log10(p) > 3.9 (FDR = 0.05).

[2]: The eQTL threshold for heat stress was –log10(p) > 3.8 (FDR =0.05).

[3]: As with determining local-eQTL, a QTL within a 2 Mb window is considered the same locus.

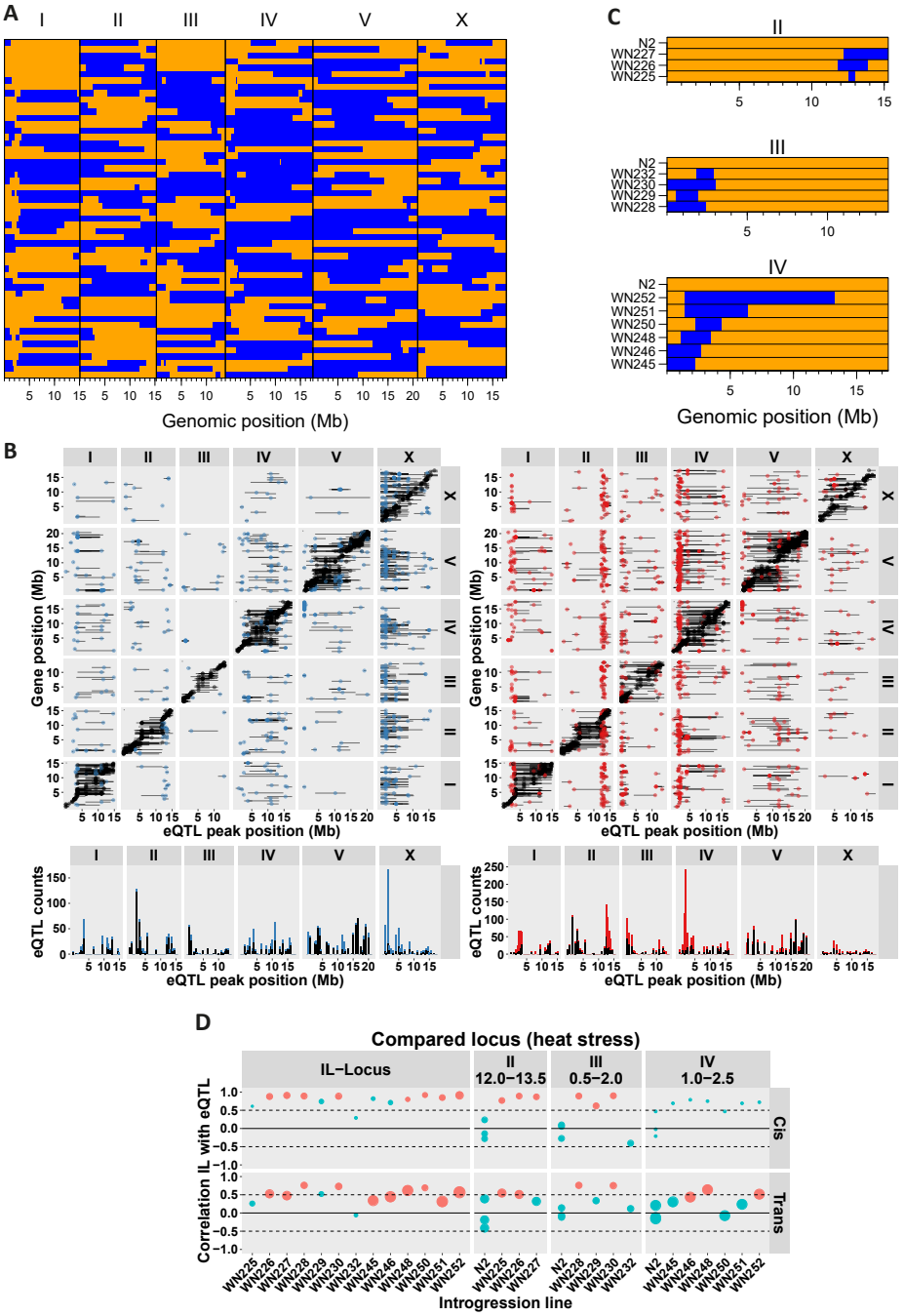


Figure 1: Genome wide results of eQTL mapping**.** (**A**) The RIL population used for eQTL mapping in control and heat stress condition. (**B**) The eQTL mapped in control (left) and heat stress (right), the false discovery rate was 0.05 (-log10(p) > 3.9 in control and –log10(p) > 3.8 in heat stress). The eQTL peak position shown on the x-axis, gene position shown on the y-axis. The *cis*-eQTL (within 1Mb of the gene) are shown in black and the *trans*-eQTL in blue (control) or red (heat stress).

The horizontal bars indicate the confidence interval of the eQTL as determined by a 1.5 LOD-drop. The chromosomes are indicated on the top and right of the plot. The histogram under the plot shows the eQTL density per 0.5 Mb bin. (**C**) The ILs used for confirmation of the heat stress *trans*-bands. (**D**) The correlation analysis of the expression in the ILs and the eQTL effects. On the x-axis the strain is shown and on the y-axis the correlation. The size of the dots indicates the number of spots the analysis is based on (ranging from n=19 *cis*-eQTL mapping to WN225 to n=477 *trans*-eQTL mapping to WN252). Color indicates significance (red, $p < 1*10^{-10}$). On top of the panels the tested locus is indicated.

To validate the *trans*-band locus, we used a panel of 55 homozygous introgression lines (ILs) with CB4856 segments in an otherwise N2 background, all of which were also genotyped by whole genome sequencing (**Supplementary file 1**) [22]. We set out to validate the heat stress specific *trans*-bands, by selecting ILs covering each of the three *trans*-bands and measuring the expression during heat stress (**Figure 1C**). In order to test the replicability of the eQTL, we correlated the expression in the ILs with the eQTL effects derived from the RILs (see materials and methods). First we determined the correlations and effects for the eQTL at the IL locus, which showed that both the *cis*- and *trans*-eQTL effects were replicable in the ILs (**Supplementary file 6**). Next, in order to verify the location of the *trans*-bands, the ILs were correlated with the eQTL mapping to either of the three *trans*-bands. If the introgression affects the genes in the *trans*-band, it will show a positive correlation, whereas if the introgression does not affect those genes, the correlation will be around 0 (**Figure 1D)**.

Each *trans*-band could be replicated in at least two of the ILs covering the locus ($p < 1*10^{-10}$, **Table 2**). The major *trans*-band on chromosome IV was significantly confirmed by 3 ILs, and only one IL (WN250) did not show a positive correlation. From the ILs confirming this *trans*-band, its location could be estimated at 1.41-2.73 Mb at the left arm of chromosome IV. Although this result verified the location of the *trans*-band, it also shows the actual architecture is likely to be more complex than inferred from the RILs alone. Although the correlations are significant, not all individual eQTL effects correlate positively; the *trans*-eQTL effect sizes in the ILs are only 47% of what was measured in the RILs (**Supplementary file 6**). This indicates that the effect sizes are probably over-estimated in the RILs. We focused on the left arm of chromosome IV in search of potential regulators.

Table 2: Condition specific eQTL *trans*-bands.

Condition	Chromosome	Position (Mb)	Confirming ILs[1]	Genes[2]
Control	IV	10.5-11.0	NA	41
	X	0.5-2.0	NA	179
Heat stress	II	12.0-13.5	WN225, WN226	242
	III	0.5-2.0	WN228, WN230	159
	IV	1.0-2.5	WN246, WN248, WN252	276

[1]: *Trans*-band was confirmed if Pearson correlation significance was $p < 1*10^{-10}$.

[2]: Genes with an eQTL within the indicated *trans*-band position.

Identifying polymorphic genes underlying the *trans*-eQTL on chromosome IV

We used literature and WormNet to find candidate genes that could be involved in the regulation of the chromosome IV *trans*-band (see materials and methods) [34]. We selected the genes: *clp-6*, *clp-7*, *ced-2*, *nhr-122*, *nhr-287*, *nhr-92*, *cmk-1*, *egl-4*, *egl-18*, *elt-6*, *test-1*, *gcy-37*, *lin-1*, *eri-5*, *eel-1*, and *eor-1* as candidates for a regulatory role in the chromosome IV *trans*-band.

The heat-shock induced transcription profiles of the 16 selected mutants were used to identify the potential gene(s) involved in regulating the genes with an eQTL on the left arm of chromosome IV(1-2.5Mb). By principal component analysis we found *eor-1*, *cmk-1*, and *egl-4* separate from N2 (**Figure 2**). Interestingly, *cmk-1* and *eor-1* capture different transcriptional variation, as they were found separately from *egl-4*. The transcription profiles of all three genes were enriched with *trans*-eQTLs mapping to the *trans*-band (IV:1-2.5Mb). Three other genes, Y67D8C.5 (*eel-1*), *elt-6*, and *nhr*-287, were also enriched for *trans*-eQTLs, yet with less genes than *eor-1*, *cmk-1*, and *egl-4* (**Figure 2**).

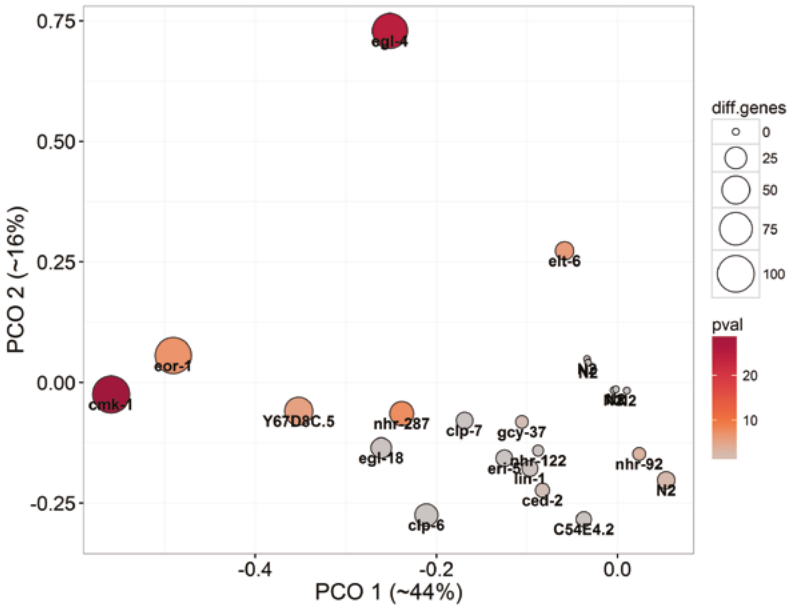


Figure 2: Variation partitioning by principal component analysis of based on the chromosome IV *trans*-band. The x-axis shows PCO 1 explaining ~44% of the variation and the y-axis shows PCO 2 explaining ~16%. The color of the circles indicates significance (in -10log(p)) based on enrichment of *trans*-eQTL differentially expressed between mutants of candidate regulators and N2 during heat-shock. The threshold for differential expression was |R| > 1.5. The size of the circle indicates the number of differentially expressed genes. The transcription profiles of mutants of three genes were found to be most different from N2, *eor-1* and *cmk-1* as one group and separate from that *egl-4*. Three other genes were found to be enriched with *trans*-eQTLs, Y67D8C.5 (*eel-1*), *elt-6*, and *nhr-287*.

To confirm the involvement of *eor-1*, *cmk-1*, and *egl-4* in regulation of the genes mapping to IV:1-2.5Mb we measured gene expression during heat-stress in three additional biological replicates. All confirmed the separation form N2 in a PCO analysis and were enriched with *trans*-eQTLs of the chromosome IV *trans*-band (**Supplementary figure 5**)**.** As we did not conduct a complementation assay nor made transgenic lines, we cannot confirm these genes as causal polymorphic regulators underlying the *trans*-band. However, the enrichment of *trans*-eQTLs placed these genes in the transcriptional response affecting the eQTL at this location. The genes *cmk-1* and *eor-1* regulated a different subset compared to *egl-4* (**Supplementary file 8**). Where the genes *eor-1* and *cmk-1* showed strong overlap in differentially expressed genes, this overlap is less with *egl-4*. Following the overlaps in differentially expressed genes, it seems that *eor-1* and *cmk-1* were upstream of the genes *eel-1*, *elt-6*, and *nhr-287*.

The chromosome IV *trans*-band is linked to a survival response

We expect that the *trans*-band on chromosome IV is an indication of a physiological response. Therefore we set out to measure survival after heat stress. We evaluated the survival curves under control and heat stress conditions (4h at 35°C) in N2, CB4856, 6 ILs which have introgressions on the left arm of chromosome IV, and the mutants *eor-1*, *cmk-1*, and *egl-4*. The heat-shock always strongly affected the survival within the first two days after application of the heat-shock (**Supplementary file 9**, log-rank test, $p < 1*10^{-5}$), except in the *eor-1* mutant strain (log-rank test, $p = 0.019$). However, this mutant already showed poor survival in control conditions (mean lifespan 10.0 days, versus 13.6 days in N2). Under control conditions, most strains showed a mean lifespan lower than, or equal to, N2. Only the ILs WN248 and WN252 lived longer than N2 (**Figure 3A**, log-rank test, $p < 0.01$). However, after heat stress all strains showed an average lifespan as long as N2 after heat stress; the ILs WN245, WN246, WN250, and WN252 even lived longer on average (**Figure 3A**, log-rank test, $p < 0.05$). These results imply that the ILs on the top of chromosome IV are more resistant to heat stress than both parental strains. Furthermore, two of the tested mutants, *cmk-1* and *eor-1*, show a relative improvement in lifespan after heat-shock compared to N2. Where *ckm-1* and *eor-1* lived 1.24 days and 3.47 days shorter in control conditions (log-rank test, $p < 0.01$), both strains lived slightly longer than N2 after heat-stress (**Figure 3B**).

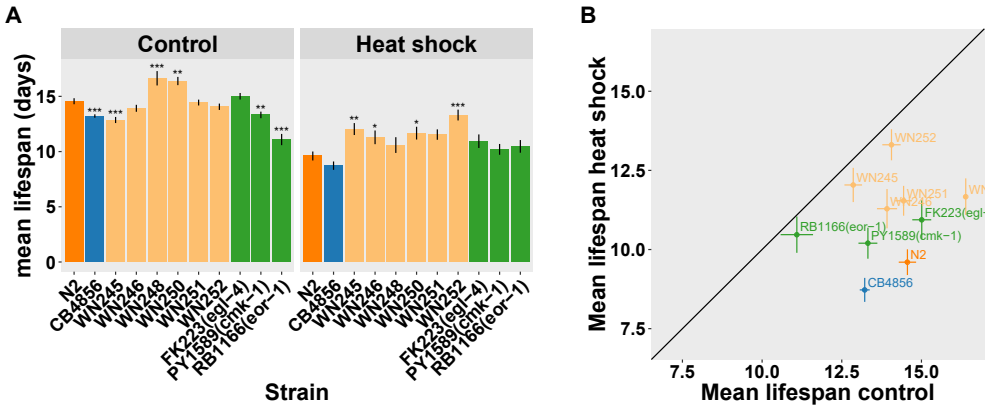


Figure 3: Lifespan measurements in the ILs and mutants. (**A**) The mean lifespan (±standard error) in the strains after control (left) and heat shock (right) treatment. All strains are compared to N2 and the significance of the difference is indicated by asterisks (* $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$). (**B**) Comparison between the lifespan in the control and heat shock experiment. The diagonal line is added as a visual aid. The strains WN245, WN246, WN252, and *eor-1* did not show a decrease in average lifespan after heat-shock (log-rank test, $p > 0.05$), all the others strains did.

Closer inspection of the survival curves after heat-shock shows that the increased lifespan compared to N2 is mainly due to survival in the first half of the survival curve. Namely, those strains show a higher heat-shock survival rate in the first two days or recover better (as measured from day 5-13), see **Supplementary file (9)**. This shows that the survival is not due to long term effects on lifespan (heterosis), but due to a better recovery. Therefore, it is likely that the *trans*-band on chromosome IV in the heat-shock condition is linked to stress-recovery.

Discussion

We have identified genes involved in the heat stress induced transcriptional response affected by natural variation between N2 and CB4856. Especially the genes *eor*-1, *cmk-1*, and *egl-4* affect the same subset of genes also affected by natural variation. This transcriptional network is implicated in stress survival, as shown by an increased recovery from heat-shock compared to N2.

Trans-band validation by introgression lines

The eQTL *trans*-bands mapped in a RIL population can be confirmed using introgression lines. The approach makes use of the multiple observations per IL, each corresponding to an eQTL mapped to the IL locus. The IL locus effect is derived by comparing the IL expression to the genetic background strain (N2 in this case). For confirmation, the obtained effects are correlated with the eQTL effects obtained from the RIL model.

We found that *cis*-eQTL are highly replicable, both by higher correlation values and a better replication of the effect sizes that are measured. This reflects the monogenic architecture of *cis*-eQTL regulation [42, 43]. Although some *cis*-eQTLs arise due to miss-hybridization [44, 45], many are due to other polymorphisms affecting the expression of these genes [42]. We found that the *trans*-eQTL were more environment specific, which is in line with previous findings [20, 21, 46, 47]. Furthermore, *trans*-eQTL were less replicable in the screened ILs compared to *cis*-eQTL. Both the correlation was lower, and effect size estimation was less accurate.

Trans-bands are indicative of a physiological response

Dissection of the genetic architecture by eQTL mapping informs on how a trait is regulated [20, 21, 42, 43, 48]. Here we show that variation in the transcriptional response to heat stress was regulated by a locus on chromosome IV, within an eQTL *trans*-band specific for heat stress. This polymorphic regulator was verified using ILs with an introgressions covering the *trans*-band locus.

The left arm of chromosome IV has been implicated in heat stress-related traits, which is consistent with our finding that the chromosome IV locus regulates variation of both stress and aging [16]. The comparison of the effects of genetic variation on transcription with transcription variation in induced mutants, allows for estimation of the allelic effects [48, 49]. This method can also be used to implicate genes that act in the transcriptional response that is affected by natural variation. In this way, we were able to implicate *cmk-1*, *egl-4*, and *eor-1* as regulators in this pathway. We used a lifespan assay to link these genes to genetic variation affecting the stress survival.

Lifespan variation in natural populations is less extreme than in induced mutants

Our lifespan findings in the ILs are in stark contrast to the effect of induced (loss-of-function) mutations. For instance, *daf-2* mutants in *C. elegans* display a two-fold lifespan prolongation and the recessive mutant allele *age-1(hx546)* results in an increase in mean life span averaging 40% [9, 50]. Such large-effect mutations were also found in other species, *e.g.* in fruit flies (two-fold increases) and mice (1.7 fold increases) [51, 52]. Yet, such extreme lifespan extensions are rarely observed in natural populations. Natural strains from *C. elegans* differ maximally by 25% in lifespan under constant laboratory conditions and even lower variation is found between wild populations of fruit flies [13, 53].

The lifespan of the CB4856 strain did not exceed N2 in either control or heat stress conditions, which is in line with previous studies [16, 21]. However, we did observe an extended lifespan in introgression lines compared to N2 under both control and heat stress conditions. Four of the six tested ILs showed an increased stress resistance, whereas the lifespan under standard conditions did not exceed N2. By testing induced mutants in the genes *cmk-1*, *egl-4*, and *eor-1* we make it plausible that this is related to the observed transcriptional variation. Under control conditions *egl-4* lives slightly longer than N2, but not as long as reported in literature [54], *cmk-1* lives slightly shorter than N2 (which is consistent with previous observations [55]. The gene *eor-1* on the other hand is compromised in lifespan under control conditions [56]. However, after exposure to heat stress, the mutants of *cmk-1* and *eor-1* show a relatively increased lifespan compared to N2. The measured effects are due to a better recovery, not due to an increased lifespan because of heterosis. As a side note, for CB4856 and the two ILs with introgressions on the far left of chromosome IV we did observe heterosis [16]. Combined, the observations in the ILs and in the mutants indicate that the *trans*-band on chromosome IV can be linked to increased heat stress recovery.

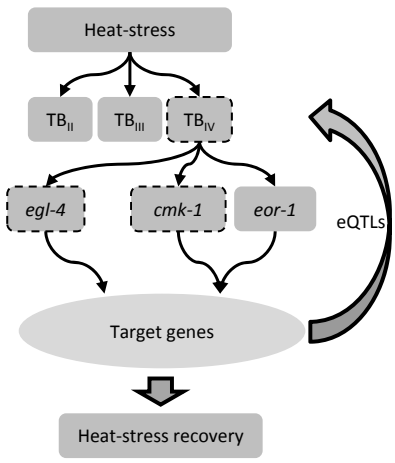


Figure 4: A possible regulatory hierarchy based the results obtained in this study. Three *trans*-bands in heat stress conditions indicate natural variation in the transcriptional response to the heat stress. For the *trans*-band on chromosome IV, we implicate three genes that affect (a part of) the same transcriptional response: *egl-4*, *cmk-1*, and *eor-1*. Where *egl-4* affects a different subset of genes than *cmk-1* and *eor-1*. The lifespan phenotypes in the ILs and the mutants indicate a link between the transcriptional response and recovery from heat stress.

Regulatory order of *eor-1*, *cmk-1*, and *egl-4*

Based on the gene expression, we identified a possible regulatory order of the mutant genes in the transcriptional regulation. First of all, we identified three *trans*-bands, linked to natural variation in the transcriptional response to heat stress. For the *trans*-band on chromosome IV, we identified six mutants that show significant overlap with the *trans*-eQTL. The strongest overlap is seen with *egl-4*, *cmk-1*, and *eor-1* (**Figure 4**). The *egl-4* mutant affects a different subset of genes compared to *cmk-1* and *eor-1*. Further downstream of *cmk-1*/*eor-1* the genes *nhr-287* and *eel-1* could play a role, but these genes only affect a small subset of the genes. The gene *el-6* on the other hand could act downstream of *egl-4*.

The gene *cmk-1* encodes a Ca2+/calmodulin-dependent protein kinase I that is required for thermotaxis and regulates gene expression in thermosensory neurons as well as the function of these neurons [57, 58]. The gene *eor-1* encodes a BTB-zinc finger protein and is involved involvement in the Ras signalling pathway [35, 59]. The gene *egl-4* encodes a cyclic GMP-dependent protein kinase and is expressed in the head neurons [54]. All of these genes encode regulatory proteins and are therefore likely actors in a transcriptional response.

Author contributions

MGS, RPJB, RJMV, JAGR, EVC, JB conducted the experiments. LBS, MGS, RPJB conducted transcriptome and main analyses. LBS, MGS, JK, BL, AC wrote the manuscript. OT conducted the genotypic analyses of the wild types. AVH, RB conducted the sequencing of the strains. OV, SC, BLe, MF, SN, OV and MMB aided in the analyses.

Supplementary figures and files

The supplementary files and figures are deposited at: http://marksterken.nl, under 'PhD thesis'.

Supplementary figure 1: Location of the 729 markers, the marker locations indicated are based on WS220.

Supplementary figure 2: marker correlations between the 729 markers.

Supplementary figure 3: The effect (x-axis) and significance (-log10(P); y-axis) of a linear model explaining variation in gene expression between control and heat stress treated RILs. Horizontal dotted line indicates the significance threshold (-log10(p) > 5.82).

Supplementary figure 4: Enrichment of condition specific eQTL per genomic location. On the x-axis the genome location is plotted (divided in 0.5Mb bins) and on the y-axis the significance of a chi-squared test between eQTL found in the control and heat stress condition is shown. A blue dot indicates control has significantly more eQTL then heat stress, whereas a red dot indicates the heat stress condition has more eQTL then the control condition.

Supplementary figure 5: Re-testing of the candidate regulators. Variation partitioning by principal component analysis of based on the chromosome IV *trans*-band, as in **Figure 2**. The x-axis shows PCO 1 explaining ~43% of the variation and the y-axis shows PCO 2 explaining ~15%. The color of the circles indicates significance (in -10log(p)) based on enrichment of *trans*-eQTL differentially expressed between mutants of candidate regulators and N2 during heat-shock. The threshold for differential expression was |R| > 1.5. The size of the circle indicates the number of differentially expressed genes.

Supplementary file 1: matrix with the strain names and genotypes of the inbred strains used in this study. The genotypes are based on the genome sequencing (see materials and methods).

Supplementary file 2: Transcriptional effects of heat stress. The outcomes of the linear model are shown. Per microarray spot the significance (-log10(p)) and the effect (log2 units) of the treatment on the expression is given. It is indicated whether the effect is significant, and the direction ('up' is higher expressed in heat stress). Also information about the gene is given (chromosome, location, names).

Supplementary file 3: Description of the transcriptional response to heat stress.

Supplementary file 4: Enrichment analysis on the genes with transcriptional responses by heat stress. The database used for enrichment (Annotation) and the category (Group), and the number of genes on the array that are in the group (Genes_in_group) is also indicated. Furthermore, the overlap with the cluster (Overlap) and the significance of that overlap (sig) is shown.

Supplementary file 5: The mapped eQTL in the control and heat stress condition. The mapped trait is annotated by spot on the microarray (trait), location of the gene (Chr_gene and BP_gene), gene identity (WBID, Sequence_name, Public_name). The eQLT is described by location (Chr, BP, marker, Peak), the boundaries of the location (marker_left, BP_left, marker_right, BP_right), the effect (Eff), and significance (P.val). Furthermore the relation between the eQTL and the transcript is given (Type), as well as the condition in which the eQTL was mapped.

Supplementary file 6: The correlation analysis of the expression effects in ILs with eQTL mapped in RILs. Each plot shows a separate IL tested for specific types of eQTL. On the x-axis the eQTL effect is indicated and on the y-axis the expression ratio of the IL with N2. The fitted lines are derived from a linear model fit. The condition is indicated on the top and the eQTL type on the right. The table on the right shows values derived from the fit.

Supplementary file 7: Candidate selection based on Wormnet and mutant availability.

Supplementary file 8: Overlap in differentially expressed genes with a *trans*-eQTL in the chromosome IV *trans*-band. The threshold for differential expression was |R| > 1.5. The genes in bold were significantly enriched for the eQTL, and the underscored genes were re-tested. The borders indicate the *cmk-1*/*eor-1* group and the *egl-4* group.

Supplementary file 9: Lifespan curves. Each comparison is plotted on a new page. First the total survival curve is shown, then the curve for the first 4 days, followed by the curve for day 5 up till day 14, followed by the curve from day 15 onwards. The p-value in the top of the curve is derived from the log-rank test on the displayed curve. On the top of each page the tested genotype and the condition is given.

References

1. **Epel, E.S. and G.J. Lithgow,** *Stress biology and aging mechanisms: toward understanding the deep connection between adaptation to stress and longevity.* **J Gerontol A Biol Sci Med Sci, 2014. 69 Suppl 1: p. S10-6.**
2. **Kennedy, B.K., et al.,** *Geroscience: linking aging to chronic disease.* **Cell, 2014. 159(4): p. 709-13.**
3. **Fontana, L., L. Partridge, and V.D. Longo,** *Extending Healthy Life Span-From Yeast to Humans.* **Science, 2010. 328(5976): p. 321-326.**
4. **Kenyon, C.J.,** *The genetics of ageing.* **Nature, 2010. 464(7288): p. 504-12.**
5. **Gems, D. and L. Partridge,** *Stress-response hormesis and aging: "that which does not kill us makes us stronger".* **Cell Metab, 2008. 7(3): p. 200-3.**
6. **Westerheide, S.D., et al.,** *Stress-Inducible Regulation of Heat Shock Factor 1 by the Deacetylase SIRT1.* **Science, 2009. 323(5917): p. 1063-1066.**
7. **Zuin, A., et al.,** *Living on the edge: stress and activation of stress responses promote lifespan extension.* **Aging (Albany NY), 2010. 2(4): p. 231-7.**
8. **Lithgow, G.J. and G.A. Walker,** *Stress resistance as a determinate of C-elegans lifespan.* **Mechanisms of Ageing and Development, 2002. 123(7): p. 765-771.**
9. **Kenyon, C., et al.,** *A C-Elegans Mutant That Lives Twice as Long as Wild-Type.* **Nature, 1993. 366(6454): p. 461-464.**
10. **Essers, M.A., et al.,** *Functional interaction between beta-catenin and FOXO in oxidative stress signaling.* **Science, 2005. 308(5725): p. 1181-4.**
11. **Wang, M.C., D. Bohmann, and H. Jasper,** *JNK signaling confers tolerance to oxidative stress and extends lifespan in Drosophila.* **Dev Cell, 2003. 5(5): p. 811-6.**
12. **Holzenberger, M., et al.,** *IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice.* **Nature, 2003. 421(6919): p. 182-7.**
13. **Gems, D. and D.L. Riddle,** *Defining wild-type life span in Caenorhabditis elegans.* **J Gerontol A Biol Sci Med Sci, 2000. 55(5): p. B215-9.**
14. **De Luca, M., et al.,** *Dopa decarboxylase (Ddc) affects variation in Drosophila longevity.* **Nat Genet, 2003. 34(4): p. 429-33.**
15. **de Magalhaes, J.P.,** *Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research.* **Cell Cycle, 2014. 13(17): p. 2671-3.**
16. **Rodriguez, M., et al.,** *Genetic variation for stress-response hormesis in C. elegans lifespan.* **Exp Gerontol, 2012. 47(8): p. 581-7.**
17. **Vinuela, A., et al.,** *Aging Uncouples Heritability and Expression-QTL in Caenorhabditis elegans.* **G3 (Bethesda), 2012. 2(5): p. 597-605.**
18. **Ebert, R.H., 2nd, et al.,** *Longevity-determining genes in Caenorhabditis elegans: chromosomal mapping of multiple noninteractive loci.* **Genetics, 1993. 135(4): p. 1003-10.**
19. **Shook, D.R. and T.E. Johnson,** *Quantitative trait loci affecting survival and fertility-related traits in Caenorhabditis elegans show genotype-environment interactions, pleiotropy and epistasis.* **Genetics, 1999. 153(3): p. 1233-43.**
20. **Li, Y., et al.,** *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222.**
21. **Vinuela, A., et al.,** *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37.**
22. **Doroszuk, A., et al.,** *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110.**
23. **Brenner, S.,** *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94.**
24. **Hosono, R.,** *Sterilization and growth inhibition of Caenorhabditis elegans by 5-fluorodeoxyuridine.* **Exp Gerontol, 1978. 13(5): p. 369-74.**
25. **Stein, L., et al.,** *WormBase: network access to the genome and biology of Caenorhabditis elegans.* **Nucleic Acids Research, 2001. 29(1): p. 82-86.**

26. **Yook, K., et al.,** *WormBase 2012: more genomes, more data, new website.* **Nucleic Acids Res, 2012. 40(Database issue): p. D735-41.**
27. **Li, H., et al.,** *The Sequence Alignment/Map format and SAMtools.* **Bioinformatics, 2009. 25(16): p. 2078-9.**
28. **DePristo, M.A., et al.,** *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* **Nat Genet, 2011. 43(5): p. 491-8.**
29. **McKenna, A., et al.,** *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* **Genome Res, 2010. 20(9): p. 1297-303.**
30. **Thompson, O., et al.,** *The million mutation project: a new approach to genetics in Caenorhabditis elegans.* **Genome Res, 2013. 23(10): p. 1749-62.**
31. **Zahurak, M., et al.,** *Pre-processing Agilent microarray data.* **BMC Bioinformatics, 2007. 8: p. 142.**
32. **Smyth, G.K. and T. Speed,** *Normalization of cDNA microarray data.* **Methods, 2003. 31(4): p. 265-73.**
33. **Benjamini, Y. and Y. Hochberg,** *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* **Journal of the Royal Statistical Society Series B-Methodological, 1995. 57(1): p. 289-300.**
34. **Lee, I., et al.,** *A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans.* **Nat Genet, 2008. 40(2): p. 181-8.**
35. **Rocheleau, C.E., et al.,** *A lin-45 raf enhancer screen identifies eor-1, eor-2 and unusual alleles of Ras pathway genes in Caenorhabditis elegans.* **Genetics, 2002. 161(1): p. 121-31.**
36. **Page, B.D., et al.,** *EEL-1, a Hect E3 ubiquitin ligase, controls asymmetry and persistence of the SKN-1 transcription factor in the early C. elegans embryo.* **Development, 2007. 134(12): p. 2303-14.**
37. **Tullet, J.M., et al.,** *Direct inhibition of the longevity-promoting factor SKN-1 by insulin-like signaling in C. elegans.* **Cell, 2008. 132(6): p. 1025-38.**
38. **Gerstein, M.B., et al.,** *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* **Science, 2010. 330(6012): p. 1775-87.**
39. **Niu, W., et al.,** *Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans.* **Genome Res, 2011. 21(2): p. 245-54.**
40. **Tepper, R.G., et al.,** *PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity.* **Cell, 2013. 154(3): p. 676-90.**
41. **Ogata, H., et al.,** *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Res, 1999. 27(1): p. 29-34.**
42. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak,** *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6.**
43. **Keurentjes, J.J., et al.,** *Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.* **Proc Natl Acad Sci U S A, 2007. 104(5): p. 1708-13.**
44. **Alberts, R., et al.,** *Sequence polymorphisms cause many false cis eQTLs.* **PLoS One, 2007. 2(7): p. e622.**
45. **West, M.A., et al.,** *High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis.* **Genome Res, 2006. 16(6): p. 787-95.**
46. **Snoek, L.B., et al.,** *Genetical Genomics Reveals Large Scale Genotype-By-Environment Interactions in Arabidopsis thaliana.* **Front Genet, 2012. 3: p. 317.**
47. **Lowry, D.B., et al.,** *Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis.* **Plant Cell, 2013. 25(9): p. 3266-79.**
48. **Andersen, E.C., et al.,** *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156.**
49. **Terpstra, I.R., et al.,** *Regulatory network identification by genetical genomics: signaling downstream of the Arabidopsis receptor-like kinase ERECTA.* **Plant Physiol, 2010. 154(3): p. 1067-78.**
50. **Friedman, D.B. and T.E. Johnson,** *A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility.* **Genetics, 1988. 118(1): p. 75-86.**
51. **Tatar, M., et al.,** *A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function.* **Science, 2001. 292(5514): p. 107-10.**
52. **Bartke, A., et al.,** *Extending the lifespan of long-lived mice.* **Nature, 2001. 414(6862): p. 412.**
53. **Sgro, C.M., et al.,** *Complexity of the genetic basis of ageing in nature revealed by a clinal study of lifespan and methuselah, a gene for ageing, in Drosophila from eastern Australia.* **Mol Ecol, 2013. 22(13): p. 3539-51.**

54. **Hirose, T., et al.,** *Cyclic GMP-dependent protein kinase EGL-4 controls body size and lifespan in C elegans.* **Development, 2003. 130(6): p. 1089-99.**
55. **Xiao, R., et al.,** *A genetic program promotes C. elegans longevity at cold temperatures via a thermosensitive TRP channel.* **Cell, 2013. 152(4): p. 806-17.**
56. **Liu, G., et al.,** *EGF signalling activates the ubiquitin proteasome system to modulate C. elegans lifespan.* **EMBO J, 2011. 30(15): p. 2990-3003.**
57. **Satterlee, J.S., W.S. Ryu, and P. Sengupta,** *The CMK-1 CaMKI and the TAX-4 Cyclic Nucleotide-Gated Channel Regulate Thermosensory Neuron Gene Expression and Function in C. elegans.* **Current Biology, 2004. 14(1): p. 62-68.**
58. **Ghosh, R., et al.,** *Multiparameter behavioral profiling reveals distinct thermal response regimes in Caenorhabditis elegans.* **BMC Biol, 2012. 10: p. 85.**
59. **Howell, K., et al.,** *EOR-2 is an obligate binding partner of the BTB-zinc finger protein EOR-1 in Caenorhabditis elegans.* **Genetics, 2010. 184(4): p. 899-913.**

Chapter 7

Epistatic eQTL are clustered across the genome and affect evolutionary conserved genes

Mark. G. Sterken,
Jan E. Kammenga,
and L. Basten Snoek

Abstract

A major challenge in genetics is to elucidate the contribution of genetic interactions (epistasis) to trait variation. Although the causal loci associated with monogenic traits can be readily detected, unravelling the epistatic architecture of complex traits forms a major challenge. We found that epistatic eQTL cluster together, which is key for their statistical identification. This approach is conceptually different from widely established approaches which aim to reduce the number of testable interactions because of the multiple testing problems. Here we show that epistatic interactions in gene expression are clustered over the genome. By mapping two-loci interactions in a gene-expression dataset on 203 recombinant inbred lines, we detected over 100 locus-pairs affecting 5956 expression traits in total. These locus-pairs were strongly clustered and we experimentally verified 5 clusters by generating two-locus introgression strains and measuring the gene expression. Epistatic eQTL identification by clustering shows that epistasis is pervasive and affects evolutionary conserved genes. By identifying regulatory clustering as epistatic fingerprint, these findings present opportunities for detecting epistasis in other organisms, including humans.

Introduction

Over the last decades it has become increasingly clear that there is a gap between known allelic variants affecting traits and how much of the trait variance these alleles explain [1]. This observation is known as 'missing heritability' and one of the causes is thought to be pervasive genetic interactions between loci (epistasis, as reviewed by [2]). Genetic interactions between loci can be recognized by 'lack of additivity'. For example, if pairwise mutants are tested and the resulting phenotype of the double mutant is not the summation of the two individual mutants, then these genes interact [3]. Experimental evidence from pairwise mutant designs [4-6], genetic background effects on mutations [7, 8], and introgression line panels [9-13] all show that epistasis is pervasive. Yet, experiments in yeast indicate that genetic architectures are predominantly consisting of many additive small effect loci, with epistasis only contributing to a small part of the trait variation [14-16].

The most commonly used mapping population for linkage analysis of complex traits are recombinant inbred lines (RILs). However, the number of statistical tests required and the problems in estimating true effect sizes make it extremely challenging to map epistatic interactions in RIL populations [16]. RILs are genetic mosaics derived from a cross between two or more parental strains and have been developed for many species to elucidate trait architectures [15, 17-20]. Although mapping of genetic interaction can be undertaken using a RIL panel, the multiple testing burden makes it hard to separate signal from noise [16, 21, 22]. Furthermore, in RIL panels with smaller sizes, QTL effects become skewed, making it hard to estimate the true effect of a QTL [23]. Therefore, the standard RIL approach (few traits in many strains) is hard to apply on epistatic interactions.

A more powerful approach for analysing epistatic interactions, consists of studying many (linked) traits in fewer strains. Genetical genomics studies are especially amendable for this kind of analysis. Genetical genomics has been performed in many species, including yeast, *A. thaliana*, *C. elegans*, mice, and even humans [14, 17-19, 24-27]. In this approach, the whole transcriptome is measured and for each transcript the association with the genetic markers is calculated, where a significant association identifies an expression QTL (eQTL) [28]. Because an eQTL map represents the transcriptome regulation affected by natural variation, single eQTL do not stand by themselves, but can be sharing regulators. For example, *trans*-bands can be observed, which are locations on the genome that affect many different transcripts thought to be affected by a common regulator. Therefore, epistatic eQTL could also share regulators.

Here we investigate two-marker genetic interaction in the round-worm *Caenorhabditis elegans*. The strains used were derived from crosses between the Bristol N2 strain and the Hawaii CB4856 strain [20, 29, 30]. These two strains are among the most genetically divergent strains in *C. elegans* and thereby capture much of the genetic variation in this species [31-33]. We started out with the hypothesis that it would be likely that genetic interactions behave in a similar way

as eQTL derived from additive models; a pair of regulators can affect many genes [18, 20, 34]. Therefore, it is to be expected that large groups of genes sharing the same pair of regulators occur as clusters, analogous to *trans*-bands in single marker mappings. This hypothesis was tested on a genetical genomics experiment in *C. elegans* (206 RILs, with 1455 markers) [20]. We found 192 putative interacting locus-pairs and verified these experimentally, the experimental steps are summarised in **Supplementary figure 1**. Our results indicate that epistatic eQTL affect evolutionlary conserved genes, in contrast to eQLT derived from additive models.

Material and Methods

Dataset retrieval

The dataset from Rockman *et al.* [20] was retrieved from WormQTL [35]. The WormQTL dataset consisting of the log2 normalized *C. elegans* micro array spot intensity from Rockman *et al.* (2010) was retrieved from GEO (GSE23857) [20], and the genetic map from Rockman & Kruglyak (2009) [29]. The strains belonging to the expression profiles labels were confirmed by correlation with *cis*-eQTL effects [35].

Influence of linkage on interaction mapping

Before setting out to map eQTL using a full two-way interaction model, the genetic map of the RIL population was analysed regarding linkage and power.

Markers in strong linkage disequilibrium, or markers that are physically linked, will result in a low number of genotypes with recombination between the markers. While overall there is a low correlation between individual markers [29], not all of the four possible locus combinations exist for 18585 (1.76% of 1057785) locus-pairs, and for 160071 (15.1% of 1057785) locus-locus combinations the lowest occurring marker combination occurs less than 2 times (**Supplementary figure 2**). When for every marker-pair the lowest occurring combinations are counted, the median is 25 occurrences. There is strong bias against the occurrence of the CB4856-CB4856 combination [29], in 60.5% of the marker-pairs this combination is the rarest. This is followed by the CB4856 and N2 combination, which occurs in 36.7% of the combinations.

Interaction mapping was simulated to determine the effect sizes detectable given the skewed marker distributions. The interaction simulation was based on normally distributed data N(0,1), with 10000 traits over 203 strains (the amount of strains in the retrieved data). Effect sizes of simulated interactions were in the range of 0-3 standard deviations, with 1-50 of the strains containing the interaction (to simulate the range of skewedness in the marker combinations). These simulations give an idea of the effect sizes that can be detected in case of perfect linkage and no additive effects. For example, if log10(p) > 5 is taken as threshold, more than 25% of the 1-sigma eQTL with 50 strains containing the peak can be detected. With a median minimum coverage of 25 strains, we are expected to find 50% of the 1.4-sigma eQTL at a threshold of –log10(p) > 5 (**Supplementary figure 3**). This indicates there is limited power to detect interactions; only large effects can be detected if the outcome of a linear model is the only consideration.

eQTL mapping

The expression quantitative trait loci (eQTL) were mapped following three models, all executed in R (Version 3.2.2 x64) using custom written scripts. The first model used was a single marker model, according to the linear model

$$I_{i,j} \sim G_{x,j} + \varepsilon_{i,j}$$

where I is the log2 normalized spot intensity of spot i (1, 2, 3, ..., 44381) of strain j (1, 2, 3, ..., 203) explained by the genotype G (N2 or CB4856) at marker x (1, 2, 3, ..., 1455) and an error term ε.

The second model used was a linear model that accounted for additive effects

$$I_{i,j} \sim G_{x1,j} + G_{x2,j} + \varepsilon_{i,j}$$

where the same parameters were used as in the single marker model. This model differs from the single marker model in the fact that additive effects between two markers (x1 and x2) are taken into account.

The third model used was a linear model that also accounted for interactions

$$I_{i,j} \sim G_{x1,j} + G_{x2,j} + G_{x1,j} * G_{x2,j} + \varepsilon_{i,j}$$

where the same parameters were used as in the additive and the single marker model. The difference compared to the additive model is the addition of a term for the interaction effects between two markers.

Permutation

For all three models permutation analysis was conducted to correct for multiple testing. The permutation used the same genetic map, but with the gene-expression values per gene randomly distributed over the genotypes.

The single marker model was permutated 10 times. A false discovery rate (FDR) of 0.05 was determined using the Benjamini-Yekutieli correction [36]. This gave a FDR = 0.05 threshold at –log10(p) > 4.5. The resulting eQTL can be found in **Supplementary file 1**.

Because of the computational load and the observation that one permutation in the in the single marker mapping already gave a good indication of the FDR, the two-marker additive model was permutated once. Here also a FDR of 0.05 was calculated using the Benjamini-Yekutieli correction, based on the lowest p-value of the marker pair. For the additive mapping a FDR = 0.05 was found at -log10(p) > 4.8. The resulting eQTL can be found in **Supplementary file 2**.

As for the two-marker additive model, the interaction model was also permutated once. An FDR of 0.05 was calculated using the Benjamini-Yekutieli correction, based on the p-value of the interaction term. For individual eQTL an FDR = 0.05 was found at -log10(p) > 9.6 (**Supplementary file 3)**

Threshold determination of interacting locus-pairs

Interacting locus-pairs were selected based on physical location and on p-value. First, the locus-pair with the most significant interaction term was selected per spot. The p-values from the interaction term are almost identical when derived from the real expression data and from the permutated dataset (**Supplementary figure 4A**). However, the distribution of the data over the genome is non-random, whereas the permutated set is (**Supplementary figure 4B**). Without any selection on p-value, the highest number of peaks per locus-pair (*e.g.* locus 10 and locus 42) in the permutated dataset was 13 (occurring once), followed by 10 (occurring thrice). Thus, in permutated data only a few false-positives per locus-pair were identified.

To identify interacting locus-pairs, we applied two thresholds, a minimum p-value for the interaction (we took $-\log10(p) \geq 4.8$, which yields an FDR=0.05 for the additive term) and a threshold for number of significant genes per locus-pair. Based on the peak distribution in the permutated data (red dots in **Supplementary figure 4B**), we excluded locus-pairs for which less than 10 genes had the most significant association mapped to that location. Together, the p-value and location thresholds yielded 195 locus-pairs (0 in our permutated set). For these 195 locus-pairs we ran 1000 permutations each to assess the FDR at $-\log10(p) \geq 4.8$. We kept locus-pairs with an FDR < 0.1 (based on the mean number of spots found over the 1000 permutations at $-\log10(p) \geq 4.8$), which left us with 192 locus-pairs (**Supplementary file 4**). These were grouped in 16 clusters based on distance (**Supplementary file 5**).

Calculation of variance explained

For calculation of the heritability, we employed the same approach as in Rockman *et al.* [20], using the realised identity by descent (IBD) kinship matrix. This approach allows for estimation of the environmental and genotypic variance based on the population structure [37]. A mixed model approach using Efficient Mixed-Model Association (EMMA) was employed to estimate the variance components based on restricted maximum likelihood (REML) [38]. The realised kinship matrix was calculated on a map with interpolated markers estimated at every 10 kb. These interpolated markers were assigned based on the linear extrapolation of recombination frequencies measured in the full RIL panel [29]. The chromosome ends were assigned the genotype of the most distal marker. Narrow sense heritability was calculated as

$$h^2 = \frac{V_G}{V_G + V_E}$$

where V_G is the genetic variance and V_E is the residual variance, as estimated by REML.

For all three models used, we determined the fraction of variance explained (R^2) of the significant associations by fitting the model (single marker, additive, and interaction) for which the linkage was derived. For the interactions we specifically calculated the R^2 of the interaction term.

Chromosomal distribution of eQTL

The chromosomal distribution of the genes with an eQTL was tested based on the chromosomal domains as defined by Rockman [29]. The differences in distribution were compared using a chi-squared test versus the complete probe set to ensure testing for deviations from the array design (results in **Supplementary file 7**).

Enrichment analysis on gene sets

Enrichment analysis was done using a hypergeometric test and as threshold we used categories that contained more than 3 genes, overlapped for more than 2 genes, and had a significance -log10(p) > 3. We analysed all 195 initially identified sites, of which 177 were enriched for one of the classes investigated according to the aforementioned thresholds (**Supplementary file 6**). The databases used were: the GO-annotation, anatomy terms, protein domains and gene classes (WS220, [39, 40]; Genes from Wormbook (2012 version, www.wormbook.org); transcription factor binding sites from MODENCODE (release #32 [41]), mapped according to [42]; KEGG pathways from the Kyoto Encyclopedia of Genes and Genomes (Release 65.0, www.genome.jp/kegg/) [43].

Strains and crossing scheme for eQTL validation

For experimental verification of the interaction clusters, we used introgression lines from a sequenced genome-wide introgression line (IL) library of CB4856 segments in an N2 background. The strains used were WN207, WN217, WN222, WN226, WN242, WN248, WN268, WN270 WN276, WN277, WN281, and WN288 [30, 31].

We used the strains to generate crosses that cover 7 pairs of interacting loci, see **Figure 2A**, **2B** and **Supplementary file 8** for an overview of the selected loci). To combine each loci-pair, the two ILs were crossed and four types of offspring were collected (as to control for any background mutations): completely N2, CB4856 for loci 1, CB4856 for loci 2 and CB4856 for both loci. Furthermore, each cross was conducted in two ways where we switched the IL acting as male. Each of these 14 crosses was conducted 6-fold with one hermaphrodite and two males, the offspring was assessed for the fraction of males in the resulting offspring (~50%) to confirm success. A complete overview of each of the resulting offspring can be found in **Supplementary file 8**. From each of these 14 crosses, four strains (one of each type) were selected for further experiments and formed an interaction panel.

Genotyping

Genotyping during the crossing was done using primers developed based on insertions/deletions between the CB4856 and N2 genomes [31]. We developed a set of 21 primers to track the introgression fragments during the cross (also see **chapter 3**). These primers were selected based on: (i) detecting a deletion in CB4856, (ii) deletion size between 25-150 bp, and (iii) deletion

occurring outside a repetitive region. Primer3 (primer3-win-bin-2.3.6) was used with standard settings to develop the primers on the area 1kb up- and downstream of the deletion, with an amplicon size in the range of 100-800bp, and an annealing temperature between 58°C and 60°C [44]. The specificity of the primers was first assessed by BLAST (ncbi-blast 2.2.28 win64) against WS230 (settings: blastn –word_size 7 –reward 1 –penalty -3) [45]. Final selection of the primers was based on application (for a list see **Supplementary file 9**).

We genotyped handpicked hermaphrodite adults that established an F1 population by self-fertilization, which reflects the genotypic variation in the offspring. Nematodes were lysed at 65°C for 30 minutes using a custom lysis buffer [46], followed by 5 minutes at 99°C. Genotyping PCRs were performed with GoTaq using the manufacturer's recommendations. The annealing temperature used was 58°C (30 seconds), with an elongation time of 1 minute, for 40 cycles. All samples were run on 1.5% agarose gels stained with Ethidium Bromide.

Confirmation experiment

For each cross, the selected interaction panel (consisting of strains completely N2, CB4856 for loci 1, CB4856 for loci 2, and CB4856 for both loci) were used in an experiment as described in Rockman *et al.* [20, 47, 48]. In short, the strains were synchronized by bleaching and maintained for 3 generations as growing populations before entering the experiment. In the third generation, once the population consisted of egg-laying adults, the population was bleached for synchronization [49]. The eggs (~400) were placed on a fresh 9 cm NGM dish, containing *E. coli* OP50 and grown for 60 hours at 20°C. At the age of 60 hours, the nematodes were washed off the plate with M9 buffer and flash frozen in liquid nitrogen. These flash frozen samples were used for RNA isolation.

RNA isolation and gene expression measurement by microarray

RNA was isolated using a Maxwell® 16 AS2000 instrument with a Maxwell® 16 LEV simplyRNA Tissue Kit (both Promega Corporation, Madison, WI, USA), following the procedure as described in [50].

The isolated RNA was labelled and hybridized to a microarray following the 'Two-Color Microarray-Based Gene Expression Analysis; Low Input Quick Amp Labeling' -protocol, version 6.0 from Agilent (Agilent Technologies, Santa Clara, CA, USA). We used *C. elegans* (V2) Gene Expression Microarray 4X44K slides, manufactured by Agilent. After hybridization was finished, the microarrays were scanned by an Agilent High Resolution C Scanner, using the recommended settings. The intensities were extracted with Agilent Feature Extraction (10.7.11).

The extracted intensities were normalized using the Limma package in R (3.2.2 x64) [REF]. As recommended, no background correction was applied [51] and we used the within-array normalization method Loess and the between-array normalization method Quantile [52]. The obtained normalized intensities were used in subsequent analysis.

Statistical analysis of confirmation experiment

In order to compare the obtained expression with the re-mapped eQTL data, the expressions were transformed to the log2 ratio of the mean per experiment by

$$R_{i,j} = \log 2\left(\frac{I_{i,j}}{\bar{I}_{i,z}}\right)$$

where R is the log2 ratio of the mean of spot i (1, 2, 3, ..., 45220) in strain j (1, 2, 3, ... 56) and I is the normalized intensity of spot i in strain j and $\bar{I}$ is the average intensity of the spots of the strains originating from the same set of parents, z (1, 2, 3, ..., 7). The obtained ratios were correlated with the eQTL derived from the single marker model over the areas covered by the introgressions. We only used the spots that were present on both array designs (19926 of the *C. elegans* V2 array spots). The result can be found in **Supplementary file 10**. As an extra confirmation of the genotype, we assured that the *cis*-eQTL confirmed the introgressions in the interaction panel.

The same approach was used for confirmation of epistatic eQTL. First, the expected loci effects in the RIL population were calculated (for all four loci combinations) on strains without breakpoints in the targeted regions. The same was done for the newly generated interaction panel (also for all four loci combinations), based on the log2 ratio of the mean per experiment. Subsequently, the expected effects from the RILs were correlated with the observed effects in the interaction panel, per spot per targeted cluster.

We permutated this test 100 times by randomly assigning the strains labels per spot in the RIL data. Subsequently, these values were used to construct the expected interaction model, which was tested against the unaltered measurements in interaction panel by correlation analysis. Therefore, these permutations gave us the random distribution for the correlations per interaction cluster. Per permutation and per cluster, this random distribution of correlations was tested against the observed distribution using the Kolmogorov-Smirnov test for continuous distributions. First, it was tested whether the distributions were equal, followed up by a test for an enrichment of larger or smaller values compared to expected. The least significant p-value over 100 permutations per loci-pair is reported.

Gene conservation

The conservation of the genes with an eQTL was tested based on induced mutations (as measured in the million mutation project) [53], natural variation (as measured in 40 wild isolates) [53], and occurrence of RNAi phenotypes [REF wormbase]. For the induced mutations and natural variation, the reported values are based on mutations leading to coding changes in the exon and to mutations in the UTRs of the genes. These results do not differ qualitatively from including all the mutations (*e.g.* synonymous mutations and mutations in introns; data not shown). The differences in distribution were tested versus genes that did not fell in the tested category (*e.g.* eQTL derived from a single marker model) using a chi-squared test (results in **Supplementary file 11**).

Results & Discussion

Epistatic eQTL are clustered across the genome

Epistatic eQTL are highly clustered, meaning that loci-pairs affect multiple transcripts. This characteristic facilitates the separation of signal from noise (false discoveries). One of the classical hindrances in whole genome epistatic QTL mapping is the multiple testing problem and therefore the chance of many false discoveries. By solely depending on statistical significance for detection of individual epistatic eQTL (an interacting loci-pair), only 27 significant interactions were detected (**Supplementary file 3**, FDR = 0.05). However, by taking into account that the false-positive epistatic eQTL are randomly distributed over the genome (see methods), we were able to identify 192 two-locus locations affecting gene expression by interaction (FDR < 0.1, **Supplementary file 4** and **5**, **Figure 1**). Of these loci-pairs, 177 could be linked to putative biological functions via enrichment analysis of the linked genes (**Supplementary file 6**). To estimate the gene expression variation explained by the interaction term per spot, we calculated the contribution of the interaction term by ANOVA. The per spot interaction contribution to the total variance explained by the epistatic model (median of 9.6%) was low compared to the two-locus additive model (median of 28.5%, **Table 1**).

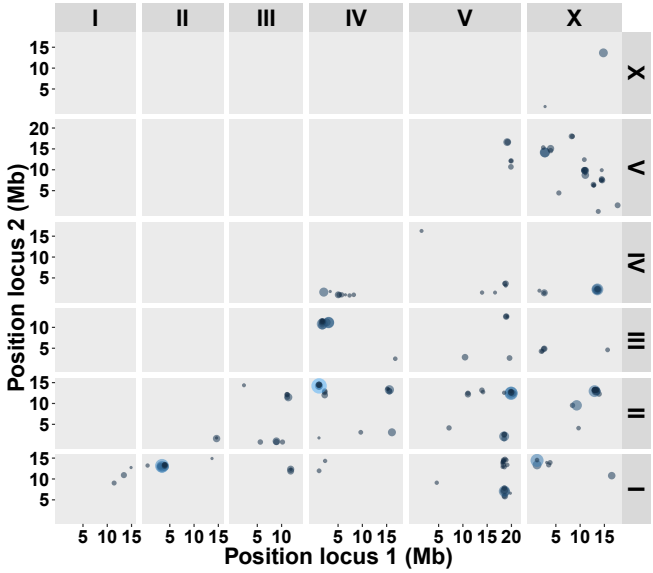


Figure 1. Two-loci interactions across the genome as mapped in the RIL population. The lightness of the colour and the size of the dot correspond to the number of spots affected by the pair of loci. There are 191 loci-pairs affecting gene expression by interaction (FDR < 0.1), affecting 1 to 138 unique genes each. These loci-pairs are strongly clustered and are found predominantly on the chromosome arms.

Table 1. The variation explained per eQTL per model. The *cis*-eQTL have one marker that is near the regulated gene, whereas the *trans*-eQTL are regulated only by markers distant from the gene (see **Supplementary table 1, 2, and 4** for the individual values). The n reported is the number of spots.

Model	eQTL type	n	Minimum R^2	Median R^2	Maximum R^2
Single marker	*Cis*	4268	0.083	0.212	0.972
	Trans	2378	0.083	0.108	0.968
	Total	6646	0.083	0.153	0.972
Additive	*Cis*	232	0.112	0.411	0.973
	Trans	239	0.102	0.208	0.959
	Total	471	0.102	0.285	0.973
Interaction*	*Cis*	221	0.034	0.094	0.131
	Trans	5735	0.086	0.096	0.178
	Total	5956	0.034	0.096	0.178

*: the 192 locus-pair epistatic eQTL were used.

As stated above, many of these epistatic interacting loci-pairs are clustered; 16 clusters representing >100 regulated genes could be detected (**Supplementary file 5**). As for the eQTL derived from the additive models, also interacting loci-pairs are mostly found on the chromosome arms (see also **Chapter 6**) [20, 24]. The chromosome arms are highly enriched for epistatic eQTL (82% occur at the arms, Chi-squared test, $P < 1*10^{-16}$, **Supplementary file 7**). These locations display the highest recombination rates in the *C. elegans* genome [29] and harbour most of the natural variation [31-33]. As a result, these sites have the largest potential to affect traits.

Our findings are in concordance with results obtained from RIL populations in yeast, where total interactions detected also explained ~10% of the trait variation [16]. However, it is not certain whether RIL derived variance estimations are correct. For example, evidence from introgression lines show entirely different effects, sometimes implying many and very strong interactions [9-13]. Most interactions derived from trait mapping in ILs result in a less-then-additive phenotype (the parental trait difference is smaller than the sum of the inferred loci). Such findings point towards a strong confounding effect of the genetic background in RIL panels. This can lead to (upwardly) biased effect sizes [2, 23], leading to wrong variance estimations and low replicability of (interaction) eQTL.

Experimental confirmation of eQTL derived from a single marker model

To verify the eQTL mapped in the RIL population, ILs were crossed covering seven two-loci locations affecting the expression of >100 genes. For each loci combination we generated all four types of inbred lines originating from the same initial cross. Furthermore, each cross was conducted in both ways with either of the two ILs acting as the paternal strain (**Figure 2A**). The resulting seven interaction panels (**Figure 2B** and **Supplementary file 8**) were grown to an age of 60 hours (as in [20]) and subsequently the transcriptome was measured by microarray.

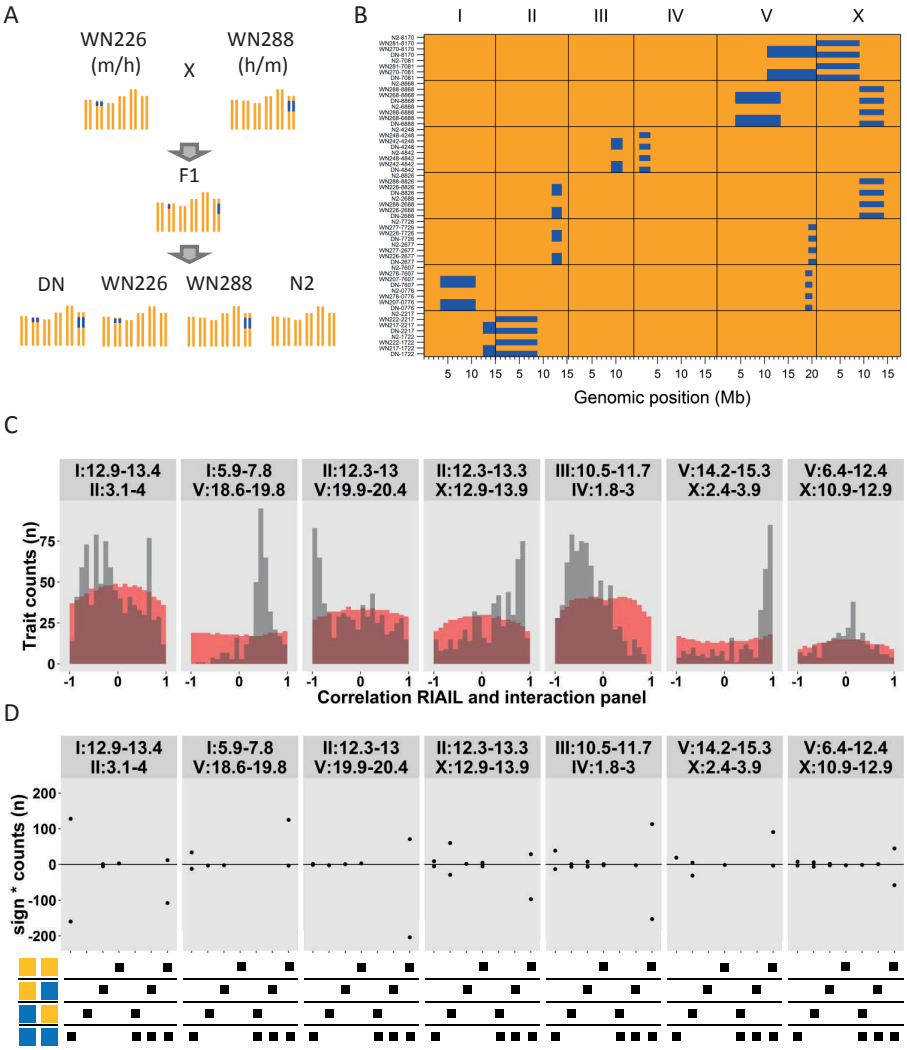


Figure 2. Interaction cluster confirmation experiment. **(A)** An example of the crossing scheme, we crossed seven different introgression lines (alternative crosses per line, h = hermaphrodite, m = male). For each cross we collected all of the four possible recombinants (double introgressions "DN", the two types of ILs, and an "N2" type), of which the genotypes were confirmed by PCR and by transcriptomics. **(B)** The map of the resulting interaction panels used for transcriptomics. **(C)** The correlations between the gene expression traits with an interaction in the RILs and the expression measurements in the seven interaction panels. In grey the correlations and in red the distribution of permutated data. Only the interaction panels covering (I:12.9-13.4, II:3.1-4.0) and (V:6.4-12.4, X:10.9-12.9) did not deviate from the permutated data ($p > 0.01$; NS, not significant). **(D)** The patterns of the interacting eQTL correlated with the two-loci genotypes (orange is N2 and blue is CB4856). The most significantly correlated pattern was counted. For example, the model indicated on the left of each panel places the two loci CB4856 at the extreme. If the trait is positively correlated with that model, it means the expression is high in the 'CB4856' and low in the other three combinations. However, if the trait is negatively correlated with that model, the expression is low in 'CB4856' compared to the other three combinations.

We compared the effect of genetic variation in the interaction panel with the single marker eQTL effects determined in the RIL population by correlation analysis. Only the per-locus eQTL were compared, not the entire eQTL set. For example, this means that of the 4268 *cis*-eQTL and the 2378 *trans*-eQTL only 77 and 60 were used for the correlation analysis of strain DN-0776 for the 12.4-15.1 Mb locus on chromosome I (**Supplementary file 10**). When compared to the single marker model, we found that the loci effects could be recapitulated; the *cis*-eQTL effects had a median Pearson correlation of 0.70 with the relative expression measured in the interaction panels. This is to be expected, since most *cis*-eQTL result from gene deletions, polymorphisms in regulatory elements, and high genetic variation [20, 31]. In general, there are no other factors involved in the gene expression variation and therefore the single *cis*-eQTL fully explains the trait variation, resulting in a high replicability.

The *trans*-eQTL were replicated to a lower extent, showing a median Pearson correlation of 0.40 with the relative expression in the introgression lines (**Supplementary file 10**). The regulation of *trans*-eQTL is more likely to be complex since *trans*-eQTL reflect genotype-environment interactions (see also **Chapter 6**) [19, 24, 34]. As *trans*-eQTL reflect specific responses, for example the *npr-1 trans*-band in these RILs reflects a mild starvation response [34], they can be harder to replicate. Furthermore, since the genes with a *trans*-eQTL react to a specific environment, the expression of those genes is more likely to be affected by multiple other genes. Therefore, *trans*-eQTL are more likely to have complex genetic architectures governed by multiple loci.

Single marker eQTL effects, especially *cis*-eQTL, are highly replicable in an independently derived introgression line population. However, we find that more complex trait architectures (*cis* versus *trans*) are less replicable. It is likely that the more complex architectures are affected by environmental effects, differences in the genetic background, and interactions of the multiple individual eQTL with that background [2, 19, 24]. Furthermore, if multiple loci are involved in trait regulation, the effect sizes of the mapped eQTL are more likely to be upwardly biased [23, 54]. This would explain the differences found between *cis*- and *trans*-eQTL replication. It is therefore to be expected that epistatic eQTL effects are hard to replicate.

Experimental confirmation of interaction eQTL

Five out of seven interaction clusters showed strong genetic interaction in the interaction panel. As for the single-marker derived eQTL, we correlated the predicted expression from the interaction model in the RILs with the measured expression in the interaction panels per spot. To test the significance of the correlations, we generated interaction effects from permutated data and tested the correlations per spot (**Figure 2C**). The correlations derived from these random distributions were tested against the observed distribution and out of the seven loci-pairs, five were clearly different from the random distribution (Kolmogorov-Smirnov test, $p < 1*10^{-5}$). In 3/7 loci-pairs, we found a clear overrepresentation for positively correlated interaction effects (Kolmogorov-Smirnov test, $p < 1*10^{-8}$). Interestingly, 2/7 loci pairs showed an effect opposite of expectation;

an overrepresentation for negatively correlated traits (Kolmogorov-Smirnov test, $p < 1*10^{-5}$). One of the possible explanations is the difference in genetic background; any background effect or higher-order interactions are linked to the N2 genotype (as reviewed by [2]).

Most of the genes at a single interaction cluster follow the same interaction pattern. We tested which pattern fits the data best (by correlation) and whether the correlation coefficient is positive or negative. For example, if the trait is high in both ancestral loci combinations (the "N2" and "CB4856"), but low when the loci are recombined (N2-CB4856 or CB4856-N2), it will correlate positively with the model that places both parents at the extreme (model on the right of each panel, **Figure 2D**). On the other hand, if the ancestral loci combinations are low and the recombined loci high, then the correlation will be negative. Either way, the recombined loci are at the phenotypic extreme, which was the most commonly observed pattern (**Figure 2D**). In other words, the trait value observed in the recombined strains goes beyond the trait value in the ancestral loci combination. This would be in line with buffering epistasis, where the ancestral loci combination leads to a fixed trait value, which is disturbed by recombination [2]. From studies in IL populations, it seems plausible that such interactions are commonplace [9-13].

Interaction eQTL mainly affect evolutionary conserved genes

We used data from the million mutation project to assess the genetic conservation of the genes with an epistatic eQTL in natural populations and populations carrying induced mutations. We found that genes with an epistatic eQTL are more mutagenizable in a laboratory setting than genes without (Chi-squared test, $P < 0.001$, **Supplementary file 11**), but are also less likely to carry naturally occurring polymorphisms (Chi-squared test, $P < 1*10^{-5}$, **Supplementary file 11**) [53]. Furthermore, these genes are more likely to result in a lethal phenotype when knocked down by RNAi (Chi-squared test, $P = 0.01$, **Supplementary file 11**) [39, 40]. This means that these genes are under higher purifying selection than genes without epistatic eQTL. In contrast, eQTL derived from additive models act on less conserved genes (**Supplementary file 11**) [20].

Conclusion

We show that epistasis is pervasive in eQTL, but the explained variation per individual gene seems small. Yet, we found evidence for a multitude of regulator genes affecting many different transcripts. The genes affected by epistatic eQTL differ qualitatively from eQTL derived from additive models. Most importantly, epistatic eQTL affect genes that are more conserved in natural populations. It is possible that the genes detected here are mostly balanced by epistasis, maintaining the expression at the desired level. One of the pieces of evidence for this hypothesis is the observation that the expression levels of the recombined loci are at the extreme of the scale for many of these genes. Since *C. elegans* is an inbreeding rather than an outbreeding animal, these disruptions of loci could be at the basis of the lower fitness observed in crossed strains [55, 56].

By nature, epistasis is hard to detect due to the large number of tests it requires [16]. Our observation that epistatic eQTL mainly regulate *in trans* whereas additive eQTL mainly interact *in cis*, this challenges some of the assumptions in models used to limit the test burden. Studies on trait mapping should therefore not assume that interactions occur between QTL mapped using a single marker model, rather it is more likely to find interacting loci by doing a full two-dimensional scan. Unfortunately, this greatly increases the test burden. However, in a larger mapping population, strong effect interactions can be detected in this way.

We found that many epistatic eQTL are regulated by the same loci-pair, equivalent to *trans*-bands detected in single-marker eQTL mappings. The loci-pairs are clustered across the genome. We also show that while these interactions can mostly be reproduced in an independent experiment, the effect sizes cannot. While we found strong positive correlations for the effects for three of the clusters, two showed strong negative correlations. These negative correlations indicate that the selected genes are interacting, however it is likely that these interactions are affected by another interactor in the genetic background. Still, our findings present opportunities for the discovery of genetic interactions in eQTL in other (model) organisms, including humans.

Acknowledgements

The authors want to thank Jelle van Creij, Joost Riksen, and Lisa van Sluijs for technical support and Jaap Bakker, Katharina Jovic, and Lisa van Sluijs for comments on the manuscript.

Author contributions

Conceived and designed the experiments: MGS, LBS, and JEK. Performed the experiments: MGS. Analysed the data: MGS. Wrote the paper: MGS, LBS, and JEK.

Supplementary figures and files

The supplementary files and figures are deposited at: http://marksterken.nl, under 'PhD thesis'.

Supplementary figure 1: Schematic overview of the steps and findings presented in this paper.

Supplementary figure 2: The two-locus marker distributions, where the lowest occurring combination of the four possible combinations was counted. The lowest occurring marker pairs are mostly CB4856-CB4856, which is expected given the skewed distribution of the map [29].

Supplementary figure 3: Simulation of epistatic interactions in eQTL. The median –log10(p) of the QTL peak as factors of the number of strains carrying the peak (x-axis) and the peak size in sigma (y-axis). Only higher effect size peaks covered by multiple strains can be detected reliably.

Supplementary figure 4: (A) The –log10(p) values ordered by size for both real (black) and permutated (red) data. As expected, only a few spots show interactions with a significance above permutated values. **(B)** The distribution of the most significant interactions over loci x loci. On the x-axis the number of interaction peaks per site and on the y-axis the number of sites with that amount of peaks. In the real data there are more loci x loci harbouring multiple interaction peaks (black) compared to the permutated data (red). In other words, in the real data interactions are localized together, whereas in the permutated data these are randomly distributed.

Supplementary file 1: Outcome of the single marker model mapping. The mapped trait is annotated by spot on the microarray (trait), location of the gene (Chr_gene and BP_gene), gene identity (WBID, Sequence_name, Public_name). The eQLT is described by location (Chr, BP, marker, Peak), the boundaries of the location (marker_left, BP_left, marker_right, BP_right). The eQLT is further described by effect (Eff), significance (P.val) and its relation to the gene (Type). Furthermore the explained variation (r2_sm) and the statistics for deriving it (sum of squares of the marker, SS_mrk, and sum of squares of the residual (SS_res), and the narrow-sense heritability (h2) are given.

Supplementary file 2: Outcome of the two-marker additive model mapping. The annotation is similar as in Supplementary file 1, only the eQTL is defined by two peaks and two p-values, and the explained variation is defined for the additive effects and the interaction effect of the two markers.

Supplementary file 3: Outcome of the two-marker interaction model mapping, without selection for loci-pairs with significant interactions. The annotation is similar as in Supplementary file 1 and 2, only the p-values and effects for the interaction QTL are included.

Supplementary file 4: Outcome of the two-marker interaction model mapping, selected for loci-pairs with significant interactions. The annotation is similar as in Supplementary file 1, 2 and 3, only the two-locus information is added, the false discovery rate (FDR) of the number of affected spots, the total number of affected spots (number_spots) and the number of affected genes (number_genes).

Supplementary file 5: The two-locus locations grouped based on genome location. The cluster letter is given and the location is shown (Chr_1, Loc_1, Chr_2, and Loc_2). Per locus pair, the number of affected spots (number_spots), unique genes (number_genes), and the false discovery rate (FDR) is shown. Furthermore, all the affected spots and the wormbase identifiers (WBiD) are listed per locus pair.

Supplementary file 6: List of the annotation enrichments found per locus pair. The locus pairs are indicated by number (as in **Supplementary file 4)**, the location (Chr_1, Loc_1, Chr_2, and Loc_2). The database used for enrichment (Annotation) and the category (Group), and the number of genes on the array that are in the group (Genes_in_group) is also indicated. Furthermore, the overlap with the locus pair (Overlap) and the significance of that overlap (sig) is shown.

Supplementary file 7: Distribution of the epistatic eQTL over the chromosomal domains [29].

Supplementary file 8: The strains obtained from the crossings for the 7 interaction panels. The target cluster (see Supplementary file 5) is indicated, and its location (Target_locus_1, Target_locus_2). The strains used for the cross

are shown (Male_IL and Hermaphrodite_IL), and the location of the introgressions. The cross information is given (Strain_code_full, Strain_code_short) and the strain name, type, and whether the strain was used in the confirmation experiment. After that, the genotypes of the strains are listed as determined by PCR.

Supplementary file 9: List of the primers and markers used for genotyping. Most of these markers were developed in **Chapter 3**.

Supplementary file 10: Outcome of correlation analysis of *cis*- and *trans*-eQTL mapped in Rockman *et al.*, 2010 and the expression in the 7 interaction panels. Correlations are only determined for the introgressed loci per strain (for an overview see **Supplementary file 8**). The chromosome of the introgression is indicated, as is the correlation and the number of spots upon which the correlation was based.

Supplementary file 11: Gene conservation of genes with epistatic eQTL and genes with eQTL derived from the additive models. The first tab shows the tests for occurrence of mutations in the genes in the million mutation project. The second tab shows the occurrence of mutations in the genes in 40 wild isolates. The third tab shows whether the genes have an RNAi phenotype.

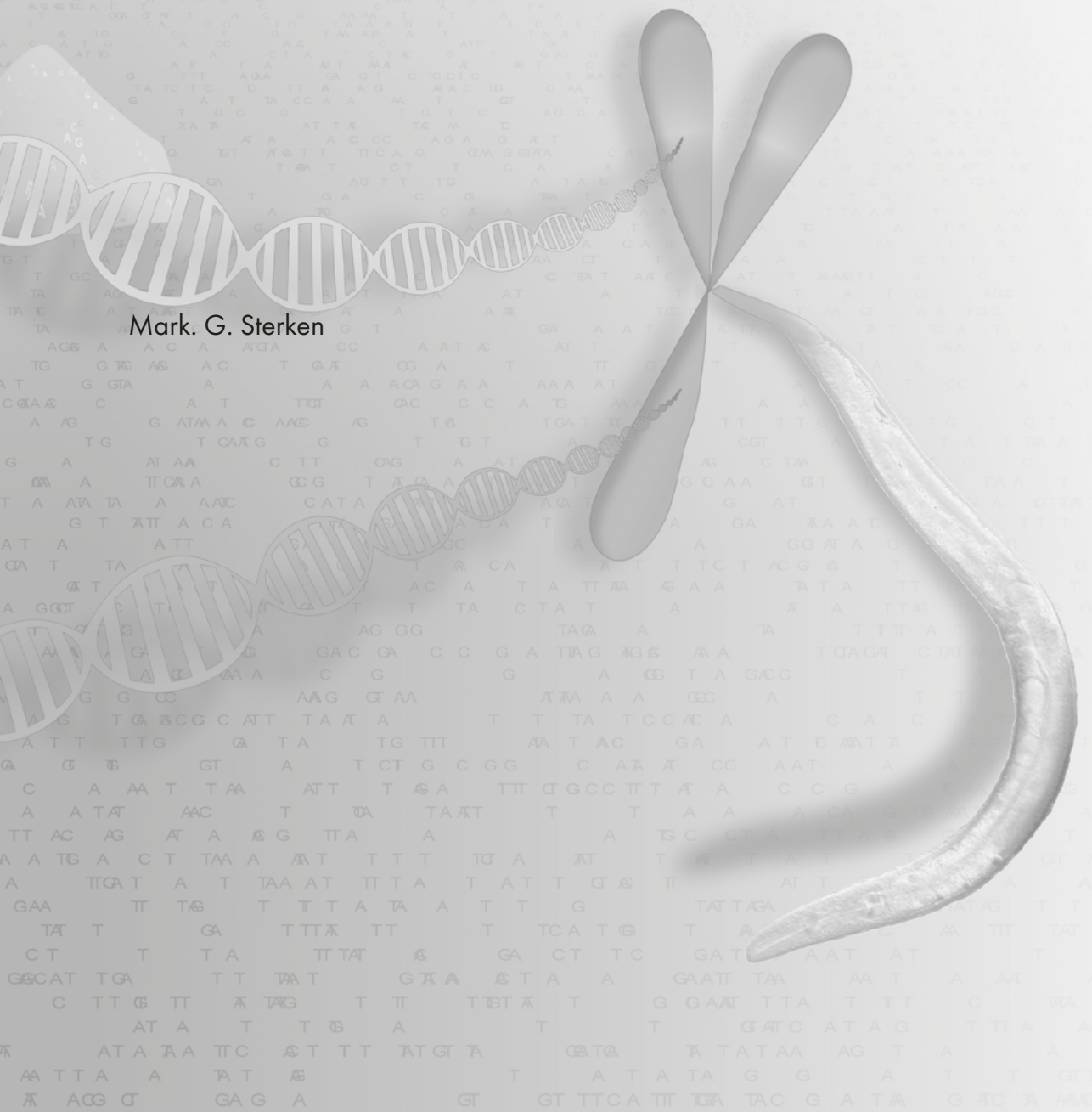
References

1. **Manolio, T.A., et al.**, *Finding the missing heritability of complex diseases.* **Nature, 2009. 461(7265): p. 747-53**.
2. **Mackay, T.F.**, *Epistasis and quantitative traits: using model organisms to study gene-gene interactions.* **Nat Rev Genet, 2014. 15(1): p. 22-33**.
3. **Elena, S.F. and R.E. Lenski**, *Test of synergistic interactions among deleterious mutations in bacteria.* **Nature, 1997. 390(6658): p. 395-398**.
4. **Tong, A.H.Y., et al.**, *Global mapping of the yeast genetic interaction network.* **Science, 2004. 303(5659): p. 808-813**.
5. **Lehner, B., et al.**, *Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways.* **Nat Genet, 2006. 38(8): p. 896-903**.
6. **Horn, T., et al.**, *Mapping of signaling networks through synthetic genetic interaction analysis by RNAi.* **Nat Methods, 2011. 8(4): p. 341-6**.
7. **Duveau, F. and M.A. Felix**, *Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans.* **PLoS Biol, 2012. 10(1): p. e1001230**.
8. **Schmid, T., et al.**, *Systemic Regulation of RAS/MAPK Signaling by the Serotonin Metabolite 5-HIAA.* **Plos Genetics, 2015. 11(5).**
9. **Keurentjes, J.J., et al.**, *Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.* **Genetics, 2007. 175(2): p. 891-905**.
10. **Gale, G.D., et al.**, *A genome-wide panel of congenic mice reveals widespread epistasis of behavior quantitative trait loci.* **Mol Psychiatry, 2009. 14(6): p. 631-45**.
11. **Shao, H., et al.**, *Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis.* **Proc Natl Acad Sci U S A, 2008. 105(50): p. 19910-4**.
12. **Edwards, A.C. and T.F. Mackay**, *Quantitative trait loci for aggressive behavior in Drosophila melanogaster.* **Genetics, 2009. 182(3): p. 889-97**.
13. **Glater, E.E., M.V. Rockman, and C.I. Bargmann**, *Multigenic natural variation underlies Caenorhabditis elegans olfactory preference for the bacterial pathogen Serratia marcescens.* **G3 (Bethesda), 2014. 4(2): p. 265-76**.
14. **Brem, R.B., et al.**, *Genetic interactions between polymorphisms that affect gene expression in yeast.* **Nature, 2005. 436(7051): p. 701-3**.
15. **Bloom, J.S., et al.**, *Finding the sources of missing heritability in a yeast cross.* **Nature, 2013. 494(7436): p. 234-7.**
16. **Bloom, J.S., et al.**, *Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.* **Nat Commun, 2015. 6: p. 8712**.
17. **Brem, R.B. and L. Kruglyak**, *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* **Proc Natl Acad Sci U S A, 2005. 102(5): p. 1572-7**.
18. **Keurentjes, J.J., et al.**, *Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.* **Proc Natl Acad Sci U S A, 2007. 104(5): p. 1708-1**3.
19. **Li, Y., et al.**, *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222**.
20. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak**, *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6**.
21. **Zhang, W., et al.**, *A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules.* **PLoS Comput Biol, 2010. 6(1): p. e1000642**.
22. **Becker, J., et al.**, *A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals.* **Eur J Hum Genet, 2012. 20(1): p. 97-101**.
23. **Xu, S.**, *Theoretical basis of the Beavis effect.* **Genetics, 2003. 165(4): p. 2259-68**.
24. **Vinuela, A., et al.**, *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37**.
25. **Doss, S., et al.**, *Cis-acting expression quantitative trait loci in mice.* **Genome Res, 2005. 15(5): p. 681-91**.
26. **Schadt, E.E., et al.**, *Mapping the genetic architecture of gene expression in human liver.* **PLoS Biol, 2008. 6(5): p. e107**.
27. **Brown, A.A., et al.**, *Genetic interactions affecting human gene expression identified by variance association mapping.* **Elife, 2014. 3: p. e01381**.

28. **Jansen, R.C. and J.P. Nap**, *Genetical genomics: the added value from segregation.* **Trends Genet, 2001. 17(7): p. 388-91**.
29. **Rockman, M.V. and L. Kruglyak**, *Recombinational landscape and population genomics of Caenorhabditis elegans.* **PLoS Genet, 2009. 5(3): p. e1000419**.
30. **Doroszuk, A., et al.,** *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110**.
31. **Thompson, O.A., et al.**, *Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856.* **Genetics, 2015. 200(3): p. 975-89**.
32. **Andersen, E.C., et al.**, *Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nat Genet, 2012. 44(3): p. 285-90**.
33. **Volkers, R.J., et al.**, *Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations.* **BMC Biol, 2013. 11: p. 93**.
34. **Andersen, E.C., et al.**, *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156**.
35. **Snoek, L.B., et al.**, *WormQTL--public archive and analysis web portal for natural variation data in Caenorhabditis spp.* **Nucleic Acids Res, 2013. 41(Database issue): p. D738-43**.
36. **Benjamini, Y. and Y. Hochberg**, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* **Journal of the Royal Statistical Society Series B-Methodological, 1995. 57(1): p. 289-300**.
37. **Visscher, P.M.**, *Whole genome approaches to quantitative genetics.* **Genetica, 2009. 136(2): p. 351-8**.
38. **Kang, H.M., et al.**, *Efficient control of population structure in model organism association mapping.* **Genetics, 2008. 178(3): p. 1709-23**.
39. **Stein, L., et al.**, *WormBase: network access to the genome and biology of Caenorhabditis elegans.* **Nucleic Acids Research, 2001. 29(1): p. 82-86**.
40. **Yook, K., et al.**, *WormBase 2012: more genomes, more data, new website.* **Nucleic Acids Res, 2012. 40(Database issue): p. D735-41**.
41. **Gerstein, M.B., et al**., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* **Science, 2010. 330(6012): p. 1775-87**.
42. **Tepper, R.G., et al.**, *PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity.* **Cell, 2013. 154(3): p. 676-90**.
43. **Ogata, H., et al.**, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* **Nucleic Acids Res, 1999. 27(1): p. 29-34**.
44. **Untergasser, A., et al.**, *Primer3--new capabilities and interfaces.* **Nucleic Acids Res, 2012. 40(15): p. e115**.
45. **Altschul, S.F., et al.**, *Basic local alignment search tool.* **J Mol Biol, 1990. 215(3): p. 403-1**0.
46. **Vervoort, M.T., et al**., *SSU ribosomal DNA-based monitoring of nematode assemblages reveals distinct seasonal fluctuations within evolutionary heterogeneous feeding guilds.* **PLoS One, 2012. 7(10): p. e47555**.
47. **Capra, E.J., S.M. Skrovanek, and L. Kruglyak**, *Comparative developmental expression profiling of two C. elegans isolates.* **PLoS One, 2008. 3(12): p. e4055**.
48. **Kirienko, N.V., K. Mani, and D.S. Fay,** *Cancer models in Caenorhabditis elegans.* **Dev Dyn, 2010. 239(5): p. 1413-48**.
49. **Brenner, S.,** *The genetics of Caenorhabditis elegans.* **Genetics, 1974. 77(1): p. 71-94**.
50. **Snoek, L.B., et al.**, *A rapid and massive gene expression shift marking adolescent transition in C. elegans.* **Sci Rep, 2014. 4: p. 3912**.
51. **Zahurak, M., et al.**, *Pre-processing Agilent microarray data.* **BMC Bioinformatics, 2007. 8: p. 142**.
52. **Smyth, G.K. and T. Speed**, *Normalization of cDNA microarray data.* **Methods, 2003. 31(4): p. 265-73**.
53. **Thompson, O., et al.**, *The million mutation project: a new approach to genetics in Caenorhabditis elegans.* **Genome Res, 2013. 23(10): p. 1749-62**.
54. **Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard**, *Prediction of total genetic value using genome-wide dense marker maps.* **Genetics, 2001. 157(4): p. 1819-1829**.
55. **Dolgin, E.S., et al.**, *Inbreeding and outbreeding depression in Caenorhabditis nematodes.* **Evolution, 2007. 61(6): p. 1339-52**.
56. **Snoek, L.B., et al.**, *Widespread genomic incompatibilities in Caenorhabditis elegans.* **G3 (Bethesda), 2014. 4(10): p. 1813-23**.

Chapter 8

General discussion, the QTL that failed to replicate



Mark. G. Sterken

Introduction

The goal of quantitative genetics is coupling genetic variation to phenotypic variation, ultimately understanding the various mechanisms that contribute to and shape the genotype-phenotype landscape. It is mainly studied using many (model) organisms, including *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans, Mus musculus*, and *Homo sapiens* [1-11]. However, quantitative genetics is not only an academic pursuit, it is also applied in multi-billion dollar industries for breeding the best crops and animals and identifying potential human drug targets. Therefore, an understanding of the basic mechanisms determining the genotype-phenotype landscape is essential to many fields of interest.

Quantitative trait locus mapping in *C. elegans* has led to the discovery of many genes with allelic variation associated with trait variation: *tra-3*, *npr-1*, *zeel-1*, *plg-1*, *scd-2*, *glb-5*, *tyra-3*, *ppw-1*, *nath-10*, *exp-1*, *drh-1*, *glc-1*, *srg-36*, and *srg-37* [12-26]. It also yielded information on the heritability of many traits, ranging from gene expression, to fecundity, to viral load upon infection with the Orsay virus [14, 27-31; **Chapter 5**]. Although many traits are highly heritable, in many cases quantitative trait loci (QTL) have not been discovered [7, 29, 32]. Considering this, it seems that the undiscovered QTLs represent alleles of genes which do not follow a relatively simple Mendelian architecture. Therefore we do not sample the full spectrum of genetic architectures. This leaves many open questions regarding genetic architectures, some of which have been debated for decades. The main questions relate to the distribution of effect sizes, the polygenicity of genetic architectures, and how regulators act in concert (as reviewed by [33, 34]).

The current trait mapping paradigm depends mainly on recombinant inbred line (RIL) populations to map QTL (**Figure 1A**), which are subsequently verified by using introgression lines (IL) containing the locus of the QTL in an otherwise homogeneous background to replicate its effect (**Figure 1B**, for practical examples, see **Chapter 3**, **5** and **6**). Here I will define 'replication' as confirming a QTL in an independent experiment. For example, the viral load QTLs in **Chapter 5** can be replicated in some of the tested IL strains. However, the tested ILs also reveal that the QTLs are complex, and probably the result of interactions between closely linked genes/loci. In another instance, the eQTL for C23H5.8 can be replicated in two genetic backgrounds and is therefore highly replicable (see **Chapter 3**). On the other hand, the gene *clec-62* does not have an eQTL using a single marker model in a RIL population [**Chapter 6** and **7**], but there is an eQTL in the IL population [**Chapter 3**; Sterken *et al.*, unpublished]. In the case of *clec*-62, we therefore detect a complex trait architecture which is not very replicable. There can be many reasons for failure of QTL replication in IL strains, and here I will argue that an important reason for this apparent lack of reliability stems from the complexity of trait regulation [3, 6, 8, 9, 35-37].

In this discussion, I will address different types of genetic architectures, how these can be replicated over experiments, and how they affect trait variation. Furthermore, I will discuss the

influence of environment and genetic interactions in more detail, as these add dimensionality to the observed trait variation. Finally, I want to discuss the added value of whole-genome IL panels for the detection of complex trait architectures.

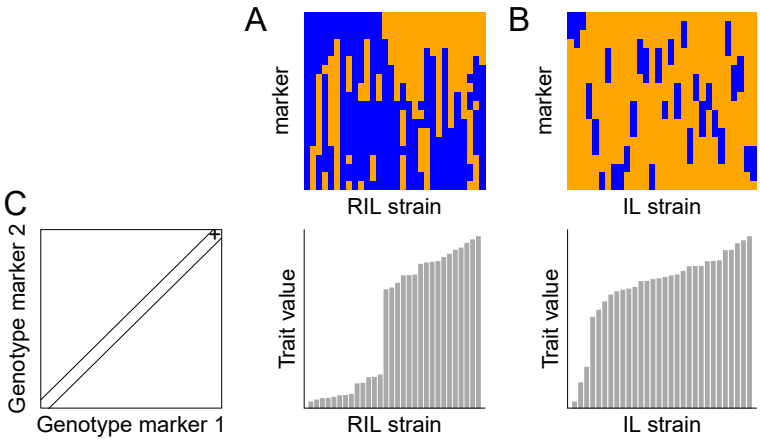


Figure 1: Monogenetic trait architectures in recombinant inbred lines and introgression lines. (**A**) the map of a simulated RIL population, consisting of 30 strains (x-axis) with 20 markers (y-axis). The colour indicates the genotype at the marker location. (**B**) The map of a simulated IL population, also consisting of 30 strains (x-axis) with 20 markers (y-axis). (**C**) Trait values for a simulated monogenic trait. On the left the simulated monogenic trait architecture, the x-axis indicates the marker 1 genotype, the y-axis the marker 2 genotype. The diagonal band therefore indicates the possible locations of QTL with no interactions. The plus-sign indicates the simulated QTL. Based on the simulated trait the trait value distribution in the RIL population and the IL population is shown.

Genetic architectures and their replicability

There are many different types of genetic architectures, and the discussion on the accuracy of the underlying models has been elaborate (as reviewed by [33, 34, 38]). Currently, there is a strong focus on mapping quantitative trait nucleotides (QTN), the exact allelic variation underlying a QTL. It has been argued by others that this pursuit of QTN should not be a goal in itself and does not provide us with the complete picture [38]. The question remains: what are the different genetic architectures and how good are we at replicating them? In this section, I will describe three types of genetic architectures and their replicability.

First, the most simple trait architecture: monogenic traits. Here a polymorphism in a single gene affects the trait of interest, thus the trait distribution in a RIL population would be (almost) binary (**Figure 1C**). Virtually all the aforementioned mapped QTN in *C. elegans* are examples of such traits. A prime example being the laboratory allele of *npr-1* [**Chapter 2**]. Also in expression QTL (eQTL) mapping, or genetical genomics, many monogenic traits can be found. Genetical genomics provides a wealth of trait information linked to genetic variation [4, 5, 7, 32, 39-41]

[**Chapter 6** and **7**]. The eQTL mapped in such studies can either be regulated *in cis* (the QTL is near the affected gene) or *in trans* (the QTL is distant of the affected gene). The *cis*-eQTL are highly replicable across (microarray) platforms (for an application of these, see **Chapter 3**, **6**, and **7**) [Snoek *et al.*, personal communication]. Some of these *cis*-eQTL are of technical origin (*e.g.* hybridization artefacts due to sequence variation) [4, 42, 43], but many are genuine [4]. Most *cis*-eQTLs have a monogenic trait architecture caused by polymorphisms in or near the affected gene. The phenotypic variations of these monogenic traits have in common that they are easily replicated over RIL and IL populations [**Chapter 3**, **6**, and **7**].

Second, a more complex genetic architecture can consist of many (>> 10) additive QTLs that affect trait variation (**Figure 2A**). Large effect QTL (QTL that capture a large amount of the trait variation) can be part of such architectures. Interestingly, due to biases in QTL mapping, traits with many equally sized QTL distributed across the genome can appear to follow an L-shaped QTL effect distribution due to biases in effect size estimation. Effect sizes are biased upwardly due to selection based on a threshold; only effects that (randomly) breach the threshold are detected as QTL [44]. In case of *C. elegans*, mapping with RIL panels can identify several QTL per trait (for example see **Chapter 7**), but the main limiting step is the number of strains used and the size (number of recombination events) of the genetic map[1, 2, 45]. Evidence for models assuming a high number of additive loci (opposed to models identifying trait architectures locus by locus) comes from the successful in explaining complex trait variation both in theory [46] and in practice [11, 32, 47]. In these examples, although individually insignificant, an aggregate of many markers of equal effect is more predictable than more simple QTL models. Traits that are regulated by an (apparent) polygenic additive architecture are highly replicable, explaining the success of genomic selection in breeding programmes (as reviewed by [48]).

Third, more complexity can arise in genetic architectures due to interactions between loci. In this case allelic variation at two loci acts in concert and the combination produces effects that go beyond additive effects. Genetic interactions (or epistasis) are thought to be pervasive and affect many traits (reviewed by [49]). Also in *C. elegans* epistasis has been observed, for example in thermal preference [30] and between the laboratory alleles *glb-5* and *npr-1* [18]. In general, the estimations for the contribution of epistatic interactions to trait variation are very diverse. In an extensive study on 20 quantitative traits in a RIL population in yeast, ~10% of the total trait variation was contributed to interactions [2] and a previous study reported interactions contribute to ~30% of the heritable genetic variation [1]. Studies in *D. melanogaster* estimate that trait variation due to genetic interactions explains 9-61% of the heritable trait variation [49]. Epistasis affects many traits, also in gene expression [50; **Chapter 3** and **7**] , but it is notoriously difficult to find due to the huge multiple testing problem.

In general, it seems that epistasis is pervasive, but RIL populations of impractically large size are required to test this observation. Furthermore, these traits are difficult to replicate as the complexity makes it likely these are not detected in most RIL populations in use. Another reason for failure to replicate is the effect of the genetic background on the trait variation [18; **Chapter**

3]. In other words, if these traits are measured in another inbred population the QTL cannot be detected, because the genetic background in which the QTL is embedded is different.

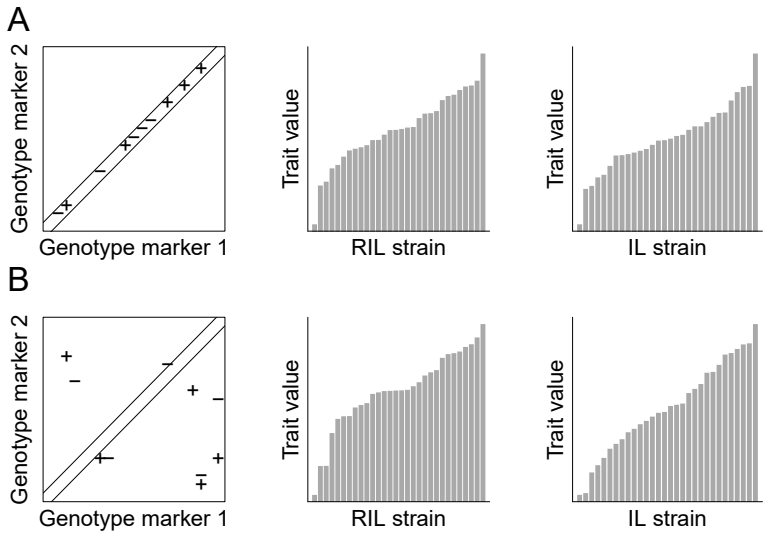


Figure 2: Simulated polygenic trait architectures. (**A**) A polygenic additive trait architecture with 10 equally sized QTL, of which five have a positive effect and five have a negative effect. On the left the simulated trait architecture is shown, the x-axis indicates the marker 1 genotype, the y-axis the marker 2 genotype. The diagonal band therefore indicates the possible locations of additive QTL. Based on the simulated trait the trait value distribution in the RIL population and the IL population is shown. (**B**) The same as in (**A**), but then for a trait architecture consisting of interactions. Notice the trait value distributions in the segregating strains are qualitatively similar compared to the additive model.

Environmental perturbations reveal additional trait dimensions

The genetic architecture only forms the first dimension of trait variation, as traits are also affected by the environment. An organism exists in a specific series of environments throughout its life, which is unique for each individual. In case of *C. elegans* it is possible to precisely control the environment in the laboratory. Standardized culturing conditions, controlling the substrate, food, population density, and temperature reduce environmental variation for the individual. However, in contrast to the constant environment in the laboratory, in nature environments fluctuate and organisms have to respond to these environmental changes. The responses can be different per genotype and can be detected in QTL approaches as plastic QTL (or genotype times environment QTL). Plasticity can be extensive and in *C. elegans* it is particularly well investigated in responses to temperature and heat-stress [12, 22, 27, 28, 40, 51-54; **Chapter 6**]. Amongst others, these researches led to the identification of the *tra-3* gene that affects body size and the *nath-10* gene that affects vulva induction at high temperatures [12, 22]. The role of these genes only becomes visible when the organism is taken outside of the 'standardized culturing conditions'.

Plasticity in QTL can be an important contributor to QTL replicability problems, because unexpected environmental influences can affect the QTL detection. For example, the trait variation linked to *npr-1* can almost be negated by controlling oxygen concentration or population density [14; **Chapter 2**]. If the oxygen concentration is lowered from the atmospheric concentration of 21% to only 10%, there are no differences between the laboratory and wild-type *npr-1* alleles. Furthermore, if the population density on the culturing plate is below or beyond standard densities, the allelic effects can also be negated. Although a different environment can be used to uncover trait variation that otherwise remains hidden, it can also obscure trait variation. Especially if environments are not well defined and controlled, the analysis of the genetic contribution to trait variation can be severely hampered. In some experiments the environment cannot be tightly controlled, such as field tests in plant breeding. In such cases, accounting for the environment can be challenging [55, 56].

A powerful study design to investigate plasticity is genetical genomics, since it allows the measurement and mapping of many linked traits [57]. Expression QTL mapping reveals that the underlying genetic regulation of gene expression is strongly shifted upon exposure to heat-stress, revealing genetic variation that is not detected beforehand [**Chapter 6**]. But already rearing *C. elegans* nematodes in different non-stressful temperatures reveals altered eQTL patterns [40]. On top of these obvious environmental responses, also age and development affect the detection of eQTL [5, 58]. Genetical genomics makes use of transcriptomics, which produces a wealth of information, which can be used to infer the environmental or developmental state of the sample [59, 60]. In turn, this inferred information can be used to uncover eQTL linked to the environment [58].

All genetical genomics experiments have in common that *trans*-bands appear that are specific to the experimental environment [5, 40; **Chapter 6**]. Next to being environment specific, *trans*-bands link to a physiological response in the animal [**Chapter 6**]. For example, the *npr-1 trans*-band results from a mild starvation response [4, 14] and the Chromosome IV *trans*-band induced by heat-shock is linked to heat stress resistance [**Chapter 6**]. This means that *trans*-bands are markers of a physiological response, they are indicative of physiological differences due to natural variation at the *trans*-band locus. Therefore, if transcriptional responses in different environments are sampled, only those genes that respond to environmental variation are likely to result in a *trans*-eQTL.

Genetic interactions are the norm for complex traits

So far different genetic architectures have been discussed, as well as the influence of the environment on trait variation. What have we learned from all these studies and what is to be expected for genetic architectures in general? For all the (mainly) monogenic allelic variation that has been uncovered and linked to their respective QTN in *C. elegans*, there are also many traits that do not follow this simple trait architecture. These traits, which are truly complex traits, are regulated by many genes and can either be consisting of many additive loci or a combination of both additive and

interacting loci (as discussed above). Although there are many successes with treating polygenetic traits as a combination of many additive loci in breeding using genomic selection [48], the apparent additivity could be an emergent property of underlying epistasis [49]. One of the strongest lines of evidence for this point comes from introgression line populations.

If introgression lines are used to dissect the genetic architecture, usually more QTLs are found compared to mapping efforts in RIL populations [6, 8, 31, 37]. Furthermore, analysis of quantitative traits in IL populations uncovers substantial epistasis for many traits [3, 6, 8, 9, 35-37; **Chapter 3**]. It is difficult to directly detect epistasis in IL populations, as ILs only contain a single introgression per line. Therefore, interactions have to be inferred from overlapping introgression lines, or by summation of the identified QTL effects over a genome wide IL screen [3, 6, 8, 9]. Such estimates show that epistasis can contribute substantially to trait variation. In this light, it seems paradoxical that large RIL populations uncover smaller contributions of epistasis to trait variation [1, 2]. It is possible that epistasis is masked in such populations due to a high number of loci stabilizing the final trait level [9; **Chapter 7**].

Pervasive epistasis means that the road from genotype to phenotype can be spurious. Also in human genetics, this has strong implications. In a recent study healthy individuals were identified that carried causal Mendelian disease linked mutations [61]. Furthermore, although genomic predictions are successful in explaining phenotypic variation in GWAS [11, 47], the prediction accuracy of such models is strongly influenced by incorporating individuals in the training set that are genetically related to the predicted individual [11]. It is therefore possible that the modelled additive effects of such predictions arise from epistatic interactions. Another possibility is the existence of environmental plasticity. Genetically related individuals (also known as family members), are more likely to be exposed to the same environment than unrelated individuals. If there are strong plastic effects, it is possible that the inclusion of family members accounts for some of these. In the end, it is likely that a combination of these and other factors play a role in shaping complex trait variation.

The added value of genome wide IL populations for complex trait mapping

Currently RIL populations are the *de facto* standard in complex trait mapping. However, if introgression lines can be constructed for a species, these form a very powerful design for dissecting complex traits. Not only can they confirm the location of QTLs [**Chapter 5** and **6**], ILs can also be used to find polygenic interactions [3, 6, 8, 9, 35-37; **Chapter 3**]. ILs can even be used to estimate how polygenic a trait can be. Since an IL only contains a single locus of one parent in the background of another parent, trait variation in a genome-wide IL population will only be (relatively) high if the trait is affected by many loci (**Figure 2**). On the other hand, traits affected by only a few or even a single gene will result in low trait variation in a genome-wide IL population

(**Figure 1C**). Therefore, polygenic traits will give rise to large trait variation in a genome-wide IL population, whereas the variation attributable to monogenic effects will be very small. Furthermore, if the variation in a genome-wide IL population exceeds the variation in a RIL population, there is strong evidence for genetic interactions. However, IL populations can only detect interactions relative to the genetic background in which the introgressions are placed. In other words, a single IL population is likely to underestimate the total contribution of genetic interactions.

Ultimately, the combination of RIL and IL panels over multiple environments will allow the placement of genetic architectures in the plasticity landscape. If it can be determined which architectures are more likely to act genetically and which are more likely to act in a plastic response, hypotheses can be formulated for systems where such experimental designs are not possible (*e.g.* humans).

Where do we go from here?

Evolution through natural selection acts on standing allelic variation and novel mutations arising in populations. This selection ultimately shapes genomes; collections of genes that together build the phenotype of an organism. Our current models are especially able to dissect simple trait architectures and understand how the alleles underlying these traits contribute to genetic variation. However, we are only beginning to explore the mechanisms driving more complex trait architectures.

The typical alleles we uncover - those pleiotropic mutations with large effects - seem to be the result of strong selection pressures [18, 26, 62, 63; **Chapter 2**]. One such example is a mutation in the *npr-1* gene in *C. elegans*, which arose while culturing this nematode on agar dishes [18; **Chapter 2**]. The mutation strongly affects the response to oxygen and carbon dioxide in this animal [13, 14, 18, 64], and it is likely that the culturing method selected this particular variant [**Chapter 2**]. Interestingly, the loci on the human genome that are implicated in many traits and have large(r) effect sizes are also those loci that have been under strong recent selection pressure. The best example of such a locus is the MHC region, which is linked to many different traits [65].

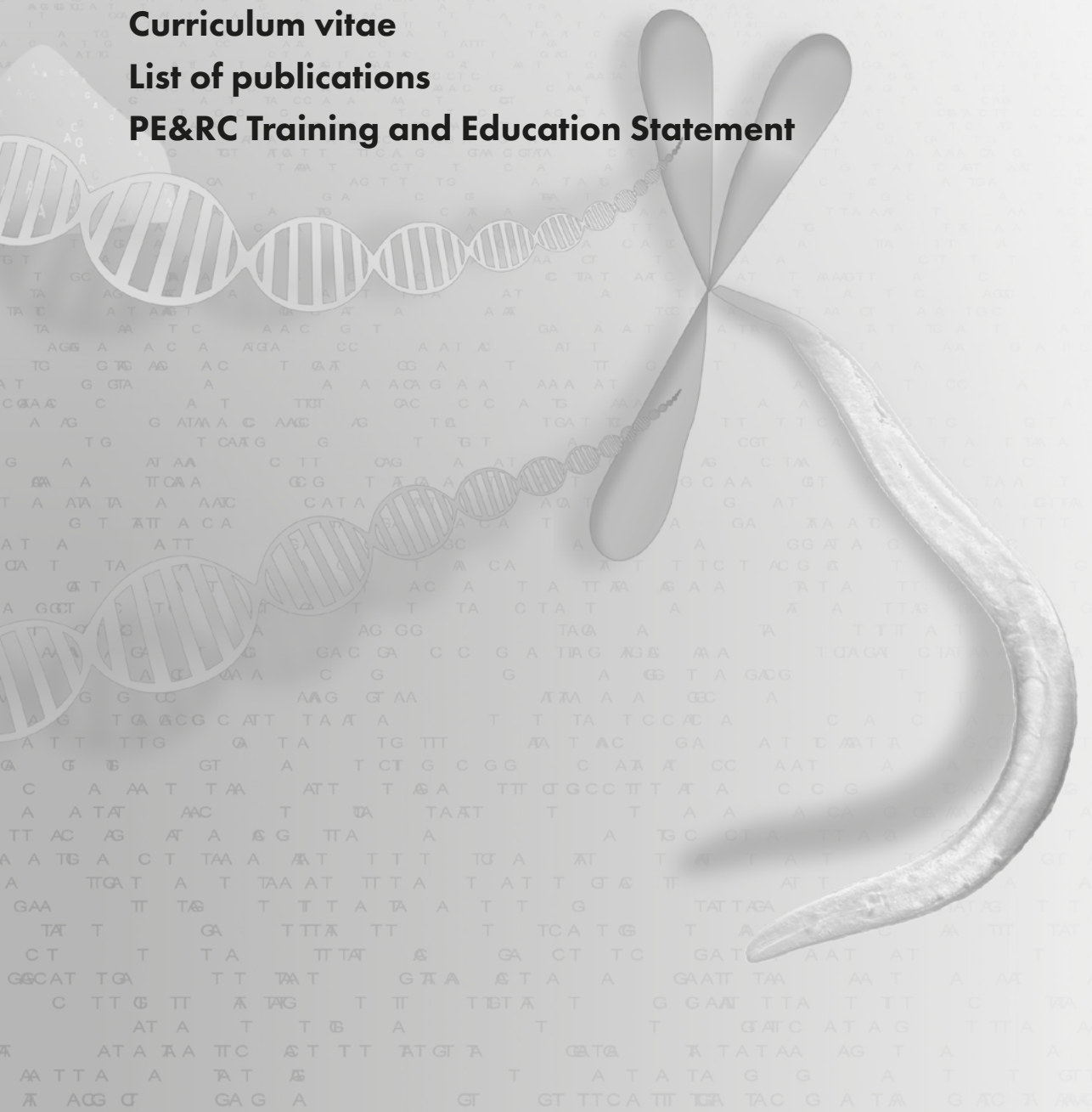
It is important to realize that we mostly only sample the monogenic trait architectures; there are many dimensions still open for exploration. Evidence is mounting for pervasive epistasis, and if these more complex trait architectures are commonplace, then genome wide IL population screens should be able to provide us with estimates on how pervasive these architectures are. Knowledge on the occurrence of such architectures will be important to improve population genetics models and genomic prediction models [49, 66]. Especially experimental designs addressing trait architectures over a wide variety of traits with different types of genetic architectures (*e.g.* expression QTL) will be very informative on the occurrence of architectures. When measured over multiple environments, we can uncover the dimensionality of genetic architectures.

References

1. **Bloom, J.S., et al.,** *Finding the sources of missing heritability in a yeast cross.* **Nature, 2013. 494(7436): p. 234-7.**
2. **Bloom, J.S., et al.,** *Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.* **Nat Commun, 2015. 6: p. 8712.**
3. **Edwards, A.C. and T.F. Mackay,** *Quantitative trait loci for aggressive behavior in Drosophila melanogaster.* **Genetics, 2009. 182(3): p. 889-97.**
4. **Rockman, M.V., S.S. Skrovanek, and L. Kruglyak,** *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science, 2010. 330(6002): p. 372-6.**
5. **Vinuela, A., et al.,** *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome Res, 2010. 20(7): p. 929-37.**
6. **Keurentjes, J.J., et al.,** *Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.* **Genetics, 2007. 175(2): p. 891-905.**
7. **Keurentjes, J.J., et al.,** *Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.* **Proc Natl Acad Sci U S A, 2007. 104(5): p. 1708-13.**
8. **Gale, G.D., et al.,** *A genome-wide panel of congenic mice reveals widespread epistasis of behavior quantitative trait loci.* **Mol Psychiatry, 2009. 14(6): p. 631-45.**
9. **Shao, H., et al.,** *Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis.* **Proc Natl Acad Sci U S A, 2008. 105(50): p. 19910-4.**
10. **Brown, A.A., et al.,** *Genetic interactions affecting human gene expression identified by variance association mapping.* **Elife, 2014. 3: p. e01381.**
11. **Makowsky, R., et al.,** *Beyond missing heritability: prediction of complex traits.* **PLoS Genet, 2011. 7(4): p. e1002051.**
12. **Kammenga, J.E., et al.,** *A Caenorhabditis elegans wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3.* **PLoS Genet, 2007. 3(3): p. e34.**
13. **de Bono, M. and C.I. Bargmann,** *Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans.* **Cell, 1998. 94(5): p. 679-89.**
14. **Andersen, E.C., et al.,** *A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology.* **PLoS Genet, 2014. 10(2): p. e1004156.**
15. **Seidel, H.S., M.V. Rockman, and L. Kruglyak,** *Widespread genetic incompatibility in C. elegans maintained by balancing selection.* **Science, 2008. 319(5863): p. 589-94.**
16. **Palopoli, M.F., et al.,** *Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans.* **Nature, 2008. 454(7207): p. 1019-22.**
17. **Reiner, D.J., et al.,** *C. elegans anaplastic lymphoma kinase ortholog SCD-2 controls dauer formation by modulating TGF-beta signaling.* **Curr Biol, 2008. 18(15): p. 1101-9.**
18. **McGrath, P.T., et al.,** *Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors.* **Neuron, 2009. 61(5): p. 692-9.**
19. **Persson, A., et al.,** *Natural variation in a neural globin tunes oxygen sensing in wild Caenorhabditis elegans.* **Nature, 2009. 458(7241): p. 1030-3.**
20. **Bendesky, A., et al.,** *Catecholamine receptor polymorphisms affect decision-making in C. elegans.* **Nature, 2011. 472(7343): p. 313-8.**
21. **Tijsterman, M., et al.,** *PPW-1, a PAZ/PIWI protein required for efficient germline RNAi, is defective in a natural isolate of C-elegans.* **Current Biology, 2002. 12(17): p. 1535-1540.**
22. **Duveau, F. and M.A. Felix,** *Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans.* **PLoS Biol, 2012. 10(1): p. e1001230.**
23. **Bendesky, A., et al.,** *Long-range regulatory polymorphisms affecting a GABA receptor constitute a quantitative trait locus (QTL) for social behavior in Caenorhabditis elegans.* **PLoS Genet, 2012. 8(12): p. e1003157.**
24. **Ashe, A., et al.,** *A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity.* **Elife, 2013. 2: p. e00994.**

25. **Ghosh, R., et al.,** *Natural variation in a chloride channel subunit confers avermectin resistance in C. elegans.* **Science, 2012. 335(6068): p. 574-8.**
26. **McGrath, P.T., et al.,** *Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes.* **Nature, 2011. 477(7364): p. 321-5.**
27. **Gutteling, E.W., et al.,** *Environmental influence on the genetic correlations between life-history traits in Caenorhabditis elegans.* **Heredity (Edinb), 2007. 98(4): p. 206-13.**
28. **Gutteling, E.W., et al.,** *Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in Caenorhabditis elegans.* **Heredity (Edinb), 2007. 98(1): p. 28-37.**
29. **Vinuela, A., et al.,** *Aging Uncouples Heritability and Expression-QTL in Caenorhabditis elegans.* **G3 (Bethesda), 2012. 2(5): p. 597-605.**
30. **Gaertner, B.E., et al.,** *More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in Caenorhabditis elegans.* **Genetics, 2012. 192(4): p. 1533-42.**
31. **Snoek, L.B., et al.,** *Widespread genomic incompatibilities in Caenorhabditis elegans.* **G3 (Bethesda), 2014. 4(10): p. 1813-23.**
32. **Brem, R.B. and L. Kruglyak,** *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* **Proc Natl Acad Sci U S A, 2005. 102(5): p. 1572-7.**
33. **Phillips, P.C.,** *Testing hypotheses regarding the genetics of adaptation.* **Genetica, 2005. 123(1-2): p. 15-24.**
34. **Barton, N.H. and P.D. Keightley,** *Understanding quantitative genetic variation.* **Nat Rev Genet, 2002. 3(1): p. 11-21.**
35. **Green, J.W., et al.,** *Genetic mapping of variation in dauer larvae development in growing populations of Caenorhabditis elegans.* **Heredity (Edinb), 2013. 111(4): p. 306-13.**
36. **Doroszuk, A., et al.,** *A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic Acids Res, 2009. 37(16): p. e110.**
37. **Glater, E.E., M.V. Rockman, and C.I. Bargmann,** *Multigenic natural variation underlies Caenorhabditis elegans olfactory preference for the bacterial pathogen Serratia marcescens.* **G3 (Bethesda), 2014. 4(2): p. 265-76.Rockman, M.V.,** *The QTN program and the alleles that matter for evolution: all that's gold does not glitter.* **Evolution, 2012. 66(1): p. 1-17.**
38. **Jansen, R.C. and J.P. Nap,** *Genetical genomics: the added value from segregation.* **Trends Genet, 2001. 17(7): p. 388-91.**
39. **Li, Y., et al.,** *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS Genet, 2006. 2(12): p. e222.**
40. **Li, Y., et al.,** *Global genetic robustness of the alternative splicing machinery in Caenorhabditis elegans.* **Genetics, 2010. 186(1): p. 405-10.**
41. **Alberts, R., et al.,** *Sequence polymorphisms cause many false cis eQTLs.* **PLoS One, 2007. 2(7): p. e622.**
42. **West, M.A., et al.,** *High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis.* **Genome Res, 2006. 16(6): p. 787-95.**
43. **Xu, S.,** *Theoretical basis of the Beavis effect.* **Genetics, 2003. 165(4): p. 2259-68.**
44. **Huang, Y.F., et al.,** *The genetic architecture of grain yield and related traits in Zea maize L. revealed by comparing intermated and conventional populations.* **Genetics, 2010. 186(1): p. 395-404.**
45. **Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard,** *Prediction of total genetic value using genome-wide dense marker maps.* **Genetics, 2001. 157(4): p. 1819-1829.**
46. **Yang, J., et al.,** *Common SNPs explain a large proportion of the heritability for human height.* **Nat Genet, 2010. 42(7): p. 565-9.**
47. **Hayes, B.J., H.A. Lewin, and M.E. Goddard,** *The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation.* **Trends Genet, 2013. 29(4): p. 206-14.**
48. **Mackay, T.F.,** *Epistasis and quantitative traits: using model organisms to study gene-gene interactions.* **Nat Rev Genet, 2014. 15(1): p. 22-33.**
49. **Brem, R.B., et al.,** *Genetic interactions between polymorphisms that affect gene expression in yeast.* **Nature, 2005. 436(7051): p. 701-3.**

50. **Shook, D.R., A. Brooks, and T.E. Johnson,** *Mapping quantitative trait loci affecting life history traits in the nematode Caenorhabditis elegans.* **Genetics, 1996. 142(3): p. 801-17.**
51. **Harvey, S.C. and M.E. Viney,** *Thermal variation reveals natural variation between isolates of Caenorhabditis elegans.* **Journal of Experimental Zoology Part B-Molecular and Developmental Evolution, 2007. 308B(4): p. 409-416.**
52. **Chandler, C.H.,** *Cryptic intraspecific variation in sex determination in Caenorhabditis elegans revealed by mutations.* **Heredity (Edinb), 2010. 105(5): p. 473-82.**
53. **Rodriguez, M., et al.,** *Genetic variation for stress-response hormesis in C. elegans lifespan.* **Exp Gerontol, 2012. 47(8): p. 581-7.**
54. **Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk,** *The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis.* **Front Physiol, 2013. 4: p. 44.**
55. **Marigorta, U.M. and G. Gibson,** *A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects.* **Front Genet, 2014. 5: p. 225.**
56. **Li, Y., R. Breitling, and R.C. Jansen,** *Generalizing genetical genomics: getting added value from environmental perturbation.* **Trends Genet, 2008. 24(10): p. 518-24.**
57. **Francesconi, M. and B. Lehner,** *The effects of genetic variation on gene expression dynamics during development.* **Nature, 2014. 505(7482): p. 208-11.**
58. **Snoek, L.B., et al.,** *A rapid and massive gene expression shift marking adolescent transition in C. elegans.* **Sci Rep, 2014. 4: p. 3912.**
59. **van der Bent, M.L., et al.,** *Loss-of-function of beta-catenin bar-1 slows development and activates the Wnt pathway in Caenorhabditis elegans.* **Sci Rep, 2014. 4: p. 4926.**
60. **Chen, R., et al.,** *Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases.* **Nat Biotechnol, 2016. 34(5): p. 531-8.**
61. **Orr, H.A.,** *The population genetics of adaptation: The distribution of factors fixed during adaptive evolution.* **Evolution, 1998. 52(4): p. 935-949.**
62. **Stern, D.L. and V. Orgogozo,** *Is Genetic Evolution Predictable?* **Science, 2009. 323(5915): p. 746-751.**
63. **Bretscher, A.J., K.E. Busch, and M. de Bono,** *A carbon dioxide avoidance behavior is integrated with responses to ambient oxygen and food in Caenorhabditis elegans.* **Proc Natl Acad Sci U S A, 2008. 105(23): p. 8044-9.**
64. **Johnson, A.D. and C.J. O'Donnell,** *An open access database of genome-wide association results.* **BMC Med Genet, 2009. 10: p. 6.**
65. **Messer, P.W., S.P. Ellner, and N.G. Hairston, Jr.,** *Can Population Genetics Adapt to Rapid Evolution?* **Trends Genet, 2016. 32(7): p. 408-18.**



Summary: Building towards a multi-dimensional genetic architecture in *Caenorhabditis elegans*

Trait variation within species is shaped by the genotype and the environment an individual is exposed to. Genomic information is inherited from the parents and forms the basis of the phenotype of an organism. The genetic variation between parents becomes differently distributed between their offspring, leading to trait variation in the offspring. Each trait can be affected by many genes, therefore the genetic architecture can be complex. In complex traits, multiple loci contribute to the ultimate trait value. However, complex traits are shaped not only by genetic variation but also by the environment and the interaction between genotype and environment. The interplay between genetic and environmental variation can affect the fitness of an organism.

Chapter 2 discusses how genotype and environment have shaped the phenotype of the nematode *Caenorhabditis elegans*, the model species used in this thesis, resulting in a laboratory adapted domesticized strain known as Bristol N2. Bristol N2 has been cultivated in the laboratory for over a decade, leading to the fixation of novel mutations in several genes that strongly affect its phenotype. Genotypic variation arisen by novel mutations in the genes *npr-1*, *glb-5*, and *nath-10* was fixed in N2 due to the laboratory environment. The allelic variation in *npr-1* affects the behaviour of this animal in an environment dependent manner, showcasing the interplay between genotype and environment. However, the altered behaviour warrants caution for interpretation of results obtained in the N2 strain.

The genotypic effects on trait variation can be large, and one of the more powerful population designs to study these effects are introgression lines. In **Chapter 3** the construction of a genome-wide introgression line (IL) panel between the N2 and the CB4856 strain is described. This panel contains loci of N2 introgressed in a homogeneous CB4856 background. It is demonstrated that together with CB4856-in-N2 ILs this new genome-wide introgression line library strongly facilitates the dissection of genetic interactions.

Chapter 4 and **5** investigate natural variation in infection with Orsay virus, a natural pathogen of the nematode *C. elegans*. In **Chapter 4** an assay is developed and tested on two wild-type strains (N2 and JU1580) and two mutant strains with mutations in the RNAi pathway. The development of the virus infection in the separate strains can be traced and the influence of genotype and age on the progression of the infection can be quantified. Furthermore, it is demonstrated that heritable RNAi plays a role in the viral load upon Orsay virus infection, an example of an epigenetically inherited environmental influence. In **Chapter 5** the assay is applied on an N2xCB4856 recombinant inbred line (RIL) population, after observing a lower viral load in CB4856 compared to N2. The RIL analysis resulted in the identification of two QTL on chromosome IV. These quantitative trait loci (QTL) were verified by CB4856-in-N2 ILs, but the IL analysis also indicated that there could be genetic interactions affecting the QTL. By a transcriptome analysis and a candidate gene search, the gene *cul-6* was identified as a candidate underlying the allelic variation between the N2 and CB4856 strain.

Chapters 6 and **7** investigate the influence of genetic interactions and the environment on the genetic architecture of gene expression. In **Chapter 6** a N2xCB4856 RIL population was exposed to heat stress, leading to the identification of a *trans*-band on the top of chromosome IV. By analysis of candidate genes, *cmk-1*, *egl-4*, and *eor-1* were implicated as contributing to the heat-stress induced transcriptional response affected by natural variation between N2 and CB4856. Furthermore, the genes with an expression QTL on the *trans*-band were indicative of a stress response phenotype. By analysis of CB4856-in-N2 ILs, it was found that this locus leads to increased recovery from stress. In **Chapter 7** two-loci genetic interactions were mapped for gene expression in a N2xCB4856 RIL panel. These epistatic interactions were confirmed by measuring gene expression in a novel population of inbred line containing the full set of loci combinations. It was found that genetic interactions in gene expression can be identified by clustering and are pervasive. These genetically interacting loci affect evolutionary conserved genes.

In conclusion, this thesis unveils the mechanisms underlying the genetic architecture of complex traits in *C. elegans* resulting from genotype and interactions between genotype and environment. It provides tools to unravel these interactions in *C. elegans*, by providing the community with new resources such as the N2-in-CB4856 introgression lines. Although *C. elegans* has been a very powerful platform for quantitative trait dissection, we need to expand our mechanistic understanding of polygenic traits.

Acknowledgements

There are many people that stand along the road I took from MSc to PhD that I would like to thank. Yet, there is also beauty in briefness; therefore I will keep it to the essentials.

Gorben, you were my first guide into the world of academic research and I learned a lot from you. I always enjoyed working with you and I hope we get around doing some crazy sfRNA work together. Jaap, I am still glad for that fateful walk from the Forum building towards Radix. You are a wonderful organizer to have on the backseat during a PhD. Professor Jan, during one afternoon in 2009 you got me hooked on eQTL and it has still not bored me. You are a clever scientist and incredibly resilient. I am really glad that you and Gorben created the opportunity to combine my interest in virology and genetics when the Orsay virus was discovered in *C. elegans*.

Swimming together in the data and the numbers was a privilege, Basten. I learned a lot from you and I hope we can still create a couple more of those nice stories and drink some beers. Roel, this also extends to you, it was really good to team up in those early GRAPPLE days and I hope to attend your defence sometime in the coming years!

Down in the proverbial salt mines, Joost, you are the master of the worm. Without your help half of this book would not be there (the other half would vanish if you had not taught me how to work with these animals). Therefore, I am really happy you are one of my paranimfs. Rita, I am also grateful for your help and ideas during the experiments. Katharina, Yiru, and Lisa, all of you were outstanding students and now awesome colleagues. I really liked (and will continue to like) working with you and I hope you can soon reap the fruits from your labour. Of course Nematology cannot exist without the support of Rikus, Sven, Debby, Casper oi!, Lisette, and Christel. Thanks for letting everything run smoothly. Corinne and Marleen, the same extends to you for Virology.

Erik, thank you so much for having me visit your lab at Northwestern, it was great! You taught me a whole lot in that time and I am really glad that led to the second chapter in this thesis. Dan, Stefan, Mustafa, Robyn, Bryn, Sarah, and Sam thank you for being awesome and making my time there very nice (yes, Erik, you were also awesome)! Where do we go from here? Well, I want to thank the various people that collaborated with me during the thesis. Esther Schnettler for taking the sfRNA work at Virology further, resulting in two nice papers. Ben Lehner, Andrew Cossins, Olga Vasieva, Madan Babu, and Sergei Nechaev for collaboration in the GRAPPLE project, we are almost there. Owen Thompson and Bob Waterston for taking the lead in building the CB4856 reference sequence. Erik Andersen and Simon Harvey for sequencing and phenotyping work on the new introgression line library.

During my PhD I have supervised many thesis students: Katharina (twice), Kobus, Henrikje, Jikke, Aliki, Daniël, Eva, Yiru, Koen, Yahya, Myrthe, Arie, Wannisa, Frederik, Jelle, Marloes, Lisa, Judith, Yvonne, Cas, Kilian, Irene, Beatrice, Jasmijn, and Linda. Thank you for all the work you have done and the things you taught me!

Work does not solely consist of experiments, coding, and papers. Fortunately, I could enjoy activities with two socially active laboratories. There have been and still are many wonderful people in both groups and I had a good time with both of them. Koen, thank you for commenting on some of the chapters, I hope I can return the favour in the future! A special thanks to Paula for being an awesome swimming buddy (twice as strong as I am), a beer buddy, and of course for being my paranimf! It was also very interesting and nice to be a member of the PE&RC Phd council. I especially want to thank Claudius, Theo, and Lennart for running such a nice graduate school (again top-grades for the next peer review)! And I also want to thank my PE&RC day buddies: Claudio, Janna, Paolo, Masha, Natalie, and Jelle for all the fun we had planning those events.

There are also people in my personal life who I like to thank. First my Vechtdal friends: Jelmer, Rienco, Gerlin, Kim, and Lotte (and their significant others) for the hiking trips and weekends at random bungalow parks. It is really nice to still be in touch and hang out! Also my study friends, Björn & Emily, Tom, Anneke, Francine, Niek, Malaika, Erik-Jan, Nick, Marloes, Koen, Merel, and -of course- SIB for all the nice times. Also my family, Oma, thanks for your nice stories. Arjan, thanks for being interested, even though there is no worm deity. Also, keep pursuing your dreams! Paulien, I enjoyed you visiting and I hope you will come over more often in the future. Pa and ma, thanks for always supporting me in everything. Pa, I guess you can keep the shovel.

Lastly, Anne, thank you for being you. I am excited about embarking on an adventure together!

Curriculum vitae

Marten Gerko (Mark) Sterken was born on the 12th of November 1987 in Zwolle, the Netherlands. After finishing pre-scientific education in 2006, he moved to Wageningen to study Biology at Wageningen University. In 2009 he completed his BSc with the specialization Cell Biology and a minor in Physical Chemistry and continued with an MSc in Biology. His first MSc thesis was on the structure-function relations of the West Nile Virus 3' untranslated region, under supervision of Dr. G.P. Pijlman. This work led to two publications and was later awarded with the Professor Van der Want thesis award for best virology thesis in the period 2010-2012. After the thesis in Virology, Mark started a second thesis at the laboratory of Nematology under the supervision of Prof. Dr. J.E. Kammenga and Dr. L.B. Snoek on genetical genomics in two types of *C. elegans* inbred populations. This work so-far led to two publications and one submitted thesis chapter. In February 2011 Mark moved to Marburg (Germany) on an Erasmus scholarship to work in the group of Prof. Dr. F. Weber on interactors of the non-structural protein of Sandfly fever Sicilian virus. In November 2011 he completed his MSc *with distinction* and started a PhD at the laboratory of Nematology and the laboratory of Virology.

During his PhD, Mark worked on quantitative genetics with *C. elegans* as model system. He presented his work at nine international conferences, including a presentation in the main session of the 4th International Conference on Quantitative Genetics and organizing a workshop for the 20th International *C. elegans* meeting. He was awarded an EMBO Travel Grant and a PE&RC publication award for work during his PhD. The latter supported his visit of the laboratory of Prof. Dr. E.C. Andersen at Northwestern University (USA), which led to publication of a review on the laboratory domestication of *C. elegans*. Mark enjoyed supervising 13 MSc and 13 BSc thesis students and was involved in teaching activities for six courses. He was also a member of the PE&RC PhD council for over three years. During this time he was chair of the PhD council and PE&RC board member from January 2014 till October 2015, and helped organizing the annual PE&RC symposium three times.

Currently Mark has a post-doc position at the Laboratory of Nematology and works on *C. elegans* and *A. thaliana* genetics.

List of publications

Peer reviewed articles

Kamkina, P; Snoek, L.B.; Grossmann, J.; Volkers, R.J.M.; Sterken, M.G.; Daube, M.; Roschitzki, B.; Fortes, C.; Schlapbach, R.; Roth, A.; Mering, C. von; Hengartner, M.O.; Schrimpf, S.P.; Kammenga, J.E. (2016). Natural genetic variation differentially affects the proteome and transcriptome in *C. elegans*. *Molecular & Cellular proteomics* 15(5): 1670-1680.

Valba, O.V.; Nechaev, S.K.; Sterken, M.G.; Snoek, L.B.; Kammenga, J.E.; Vasieva, O. (2015). On predicting regulatory genes by analysis of functional networks in *C. elegans*. *BioData Mining 8: 33.*

Hedil, M.; Sterken, M.G.; Ronde, D. de; Lohuis, D.; Kormelink, R. (2015). Analysis of Tospovirus NSs proteins in suppression of systemic silencing. *PLos one* 10(8): e0134517.

Thompson, O.A.; Snoek, L.B.; Nijveen, H.; Sterken, M.G.; Volkers, R.J.M.; Brenchley, R.; Hof, A.E. van 't; Bevers, R.P.J.; Cossins, A.; Yanai, I.; Hajnal, A.; Schmid, T.; Spencer, J.D.; Kruglyak, L.; Andersen, E.C.; Moerman, D.G.; Hillier, L.W.; Kammenga, J.E.; Waterston, R.H. (2015). Remarkably divergent regions punctuate the genome assembly of the *Caenorhabditis elegans* Hawaiian strain CB4856. *Genetics 200(3): 975-989.*

Sterken, M.G.; Snoek, L.B.; Kammenga, J.E.; Andersen, E.C. (2015). The laboratory domestication of *Caenorhabditis elegans*. *Trends in genetics* 31(5): 224-231.

Schnettler, E.; Tykalová, H.; Watson, M.; Sharma, M.; Sterken, M.G.; Obbard, D.; Lewis, S.; McFarlane, M.; Bell-Sakyi, L.; Barry, G.; Weisheit, S.; Best, S.; Kuhn, R.; Pijlman, G.; Chase-Topping, M.; Gould, E.A.; Grubhoffer, L.; Fazakerley, J.; Kohl, A. (2014). Induction and suppression of tick cell antiviral RNAi responses by tick-borne flaviviruses. *NAR 42 (14): 9436-9446.*

Van der Bent, M.L.; Sterken, M.G.; Volkers, R.J.M.; Riksen, J.A.G.; Schmid, T.; Hajnal, A., Kammenga, J.E.; Snoek, L.B. (2014). Loss-of-function of the Wnt Associated β-catenin *bar-1* affects transcription and developmental timing in *Caenorhabditis elegans*. *Scientific Reports* 4: 4926.

Hoogstrate, S.W.; Volkers, R.J.M.; Sterken, M.G.; Kammenga, J.E.; Snoek, L.B. (2014). Nematode endogenous small RNA pathways. *Worm 3(1): e28234.*

Sterken, M.G.; Snoek, L.B.; Bosman, K.J.; Daamen, J.; Riksen, J.A.G.; Bakker, J.; Pijlman, G.P.; Kammenga, J.E. (2014). A heritable antiviral RNAi response limits Orsay virus infection in *Caenorhabditis elegans* N2. *PLOS ONE 9(2): e89760.*

Snoek, L.B.; Sterken, M.G.; Volkers, R.J.M.; Klatter, M.; Bosman, K.J.; Bevers, R.P.J.; Riksen, J.A.G.; Smant, G.; Cossins, A.R.; Kammenga, J.E. (2014). A rapid and massive gene expression shift marking adolescent transition in *C. elegans*. *Scientific Reports* 4:3912.

Volkers, J.M.; Snoek, L.B.; Hellenberg Hubar, C.J. van; Coopman, R.; Chen, Wei; Yang, Wentao; Sterken, M.G.; Schulenburg, H.; Braeckman, B.; Kammenga, J.E. (2013). Gene-environment and protein degradation signatures characterize genomic and phenotypic diversity in wild *Caenorhabditis elegans* populations. *BMC Biology* 11:93.

Schnettler, E.; Sterken, M.G.; Leung, J.Y.; Metz, S.W.; Geertsema, C.; Goldbach, R.W.; Vlak, J.M.; Kohl, A.; Kromykh, A.A.; Pijlman, G.P. (2012). Noncoding flavivirus RNA displays RNA interference suppressor activity in insect and mammalian cells. *Journal of Virology, 86(24): 13486 – 13500.*

Patent

Westerhof, L.B.; Bakker, J.; Wilbers, R.H.P.; Schots, A.; Smant, G.; Goverse, A.; Johannes, H.; Sterken, M.G.; Snoek, L.B.; Kammenga, J.E. (09/06/2016). Optimisation of coding sequence for functional protein expression. WO 2016/086988

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (6 ECTS)

- The laboratory domestication of *Caenorhabditis elegans* (2015)

Writing of project proposal

- Mechanisms of nematode-virus interactions in natural *C. elegans* populations (2011)

Post-graduate courses (5.3 ECTS)

- Linear models; WGS (2012)
- Modelling critical transitions in nature and society; PE&RC, SENSE (2014)
- Springschool host microbe interactions; WGS (2014)
- Genotype by environment interactions, uniformity and stability; EPS, WIAS, PE&RC (2015)

Laboratory training and working visits (10.5 ECTS)

- Quantitative genetics of *Caenorhabditis elegans*; Northwestern University (2014)

Invited review of (unpublished) journal manuscript (1 ECTS)

- BMC Genomics: impact of environment on the *C. elegans* transcriptome (2015)

Competence strengthening / skills courses (6.3 ECTS)

- Afstudeervak organiseren en begeleiden; WGS (2012)
- Techniques for writing and presenting a scientific paper; WGS (2012)
- How to write a world-class paper (2013)
- Writing grant proposals; WGS (2015)
- 3rd WGS PhD Workshop carousel; WGS (2016)
- LS@W Accelerator: the venture challenge; Lifesciences@work (2016)

PE&RC Annual meetings, seminars and the PE&RC weekend (2.7 ECTS)

- PE&RC Introduction weekend (2012)
- PE&RC Day: extreme life (2012)

- PE&RC Day: biomimicry: unlocking nature's secrets (2013)
- PE&RC Day: optimization of science: pressure & pleasure (2014)
- PE&RC Day: one's waste...another's treasure? (2015)
- PE&RC Last year's weekend (2016)

Discussion groups / local seminars / other scientific meetings (6.0 ECTS)

- Netherlands genomics initiative: life science momentum – science for society (2011)
- GRAPPLE Project meeting; Barcelona CRG (2011)
- Experimental evolution discussion group; oral presentation in 2012 (2011-2014)
- Dutch worm meeting (2012)
- Dutch annual virology symposium (2012-2016)
- GRAPPLE project meeting; oral presentation; Liverpool (2012)
- Dutch worm meeting; oral presentation (2013)
- Symposium on experimental evolution (2013)
- Van der Want award ceremony; oral presentation (2013)
- Wageningen PhD symposium: healthy food and living environment; oral presentation (2013)
- VECTORIE training course; anti-vector, anti-virus, pro-one health: tackling emerging viral vector-borne diseases in Europe (2014)
- Current topics in plant biotechnology; oral presentation (2015)
- Research funding by government and industry: our ticket to societal relevance or the end of independent science? (2015)
- Dutch worm meeting; oral presentation (2015)

International symposia, workshops and conferences (19.2 ECTS)

- Cold Spring Harbor Laboratory, Evolution of *Caenorhabditis* and other nematodes; oral presentation (2012)
- Quantitative genetics meeting; oral presentation; Edinburgh (2012)
- 19th International *C. elegans* meeting; poster presentation (2013)
- 5th EMBO Meeting; poster and oral presentation (2013)
- European worm meeting; poster and oral presentation; Berlin (2014)
- Hinxton, volution of *Caenorhabditis* and other nematodes; poster and oral presentation (2014)
- 20th International *C. elegans* meeting; poster presentation (2014)
- Molecular biology of aging; poster presentation (2015)
- Cold Spring Harbor Laboratory, Evolution of *Caenorhabditis* and other nematodes; oral presentation (2016)

Lecturing / supervision of practicals / tutorials (16.5 ECTS)

- Introduction environmental sciences (2012-2013)
- Molecular virology (2012-2014)
- Cell biology and health (2012-2015)
- Ecophysiology (2013, 2015)
- Ecology (2015)

Supervision of 13 MSc students

- Quantitative genetics of Orsay virus infection in *Caenorhabditis elegans*
- Quantitative genetics of stress resistance and its relation to lifespan in *Caenorhabditis elegans*
- Targeted mutagenesis using the CRISPR/Cas9 system

This research was conducted at the Laboratory of Nematology and the Laboratory of Virology at Wageningen University (Wageningen, The Netherlands). The research in this thesis was financially supported by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) and ERASysbio-plus ZonMW project GRAPPLE - Iterative modelling of gene regulatory interactions underlying stress, disease and ageing in *C. elegans* (90201066).

Cover design: Anne Morbach, *C. elegans* photograph by Mark Sterken and Joost Riksen
Lay-out design: Iliana Boshoven-Gkini (Agilecolor.com)
Printed by: Ridderprint BV, the Netherlands (Ridderprint.nl)