NORTHWESTERN UNIVERSITY

Investigating the genetic mechanisms of phenotypic variation in *Caenorhabditis elegans*

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Biological Sciences

By

Kathryn S. Evans

EVANSTON, ILLINOIS

December 2020

Abstract

Phenotypic diversity underlies life as we know it–a variety of species, each with different roles, are essential for a fully functioning ecosystem just as a range of different crops is necessary to provide different nutritional value. Even within a single species, individual variation promotes diversity, allowing for adaptation to new environments and achievements despite challenges. Characterizing this phenotypic diversity and understanding the factors that drive it is a principal goal of biological research. The fruits of this research could have the power to transform the way farmers cultivate agriculture and breed livestock, the way scientists discover cures, and the way doctors predict, treat, and prevent disease. Scientists have long understood that genetic variation is a main contributor to phenotypic variation and recent advancements in sequencing technologies have improved characterization of genetic variation among species. However, high environmental variation interferes with scientists' abilities to connect phenotypic variation, such as patient response to a chemotherapeutic drug, to a single genetic variant. For this reason, model organisms such as the roundworm nematode *Caenorhabditis elegans* have become forerunners in the field of quantitative genetics. In this dissertation, I investigate how natural genetic variation in *C. elegans* shapes the diverse phenotypes observed across the species. While I mostly focus on the nematode response to a variety of toxins including chemotherapeutics, anthelmintics, and heavy metals, I also explore how genetic variation might play a role in the species' adaptation to new climates. This work highlights the power of our system to connect phenotypic variation to causal genetic variants. Furthermore, this work aims to answer several open questions in the field of quantitative genetics such as understanding the molecular mechanism of phenotypic trait evolution and determining the extent of genetic complexity of phenotypic traits.

Acknowledgements

This dissertation could not have been possible without the support and encouragement from a number of individuals. First and foremost, I would like to acknowledge my advisor, Dr. Erik Andersen. Throughout my five years in graduate school, Erik has been a constant source of support and has always pushed me to be the best scientist I can be. Working with him has vastly improved my genetics and writing skills, taught me about the power of detail in science, helped me to question science but never myself, and instilled within me a hatred for excel I doubt I will ever undo. Despite his schedule getting constantly busier and the lab growing seemingly exponentially as more people began to acknowledge his formidable power in the field of quantitative genetics, I always felt like Erik was looking out for my best interest and doing everything in his power to help me succeed in whatever path I wanted to take. When my main thesis project failed because we lost our reproducible food source, Erik asked me "What can *I* do to help *you*?" If I needed a new project, I could have my pick. If I needed to learn a new skill, he would make sure I had the time and resources to be successful. His passion for science is infectious and I know this is just the beginning for him. I will always be thankful to have been part of the "original" Andersen Lab.

In addition to Erik, Dr. Robyn Tanny is an essential part of the Andersen Lab, without whom we would surely fall into chaos. Robyn works tirelessly to ensure that the lab continues to operate smoothly and doubles as a "lab mother" to all of us–even baking us our favorite treats each year on our birthday. Robyn's kind hospitality and words of encouragement can make anyone feel at home. Most importantly, Robyn introduced me to Buffy the Vampire Slayer and for that, I will always be grateful.

During my rotation in the Andersen Lab, Dr. Shannon Brady and I worked together closely and we quickly became friends outside of the lab. Her optimism, determination (like when she joined Erik and I in the lab at 10 pm one day because a broken sorter was NOT going to stop us from running our assay we had worked so hard on), and empathy for others is admirable and I was constantly impressed with her strong problem solving abilities and clear writing and communication skills. Shannon and I co-authored a manuscript together that started out as a "quick, easy publication" and soon became a monstrous beast. Regardless, the hours we spent coding together side-by-side (because two brains are better than one) or battling the sorter (Sorter Warriors!) are some of the fondest memories I have of

my time in the lab. In August 2017, a group of us went to Hawaii on a lab-sponsored sampling trip and I spent two entire weeks 24/7 with Shannon and our lab technician, Briana Rodriguez (Team MOANA!). Shannon led our team well, bravely driving one of the scariest roads in the U.S. not once, but THREE times while Briana and I stuck our heads out the window, marveling at the natural beauty. The trip was long and tensions were high, but the three of us never tired of each other's company. I learned so much from Shannon and am grateful for her continued friendship as we both move on to life's next adventures.

Dr. Daehan Lee joined the lab about nine months after me and quickly became my "go-to-person" when I needed help. Daehan is incredibly smart and always very patient with me, willing to spend time letting me talk through a difficult problem and offering advice about experiments, careers, and life in general. Daehan's sunny personality and constant encouragement made the tough times bearable, and there is no one better with whom to share good news. I have learned so much from Daehan, including the importance of trusting and advocating for myself. I admire his passion for science; it is an inescapable part of who he is and everything he does. I look forward to watching him grow as a mentor and a scientist and I have no doubt he will continue to accomplish great things.

Dr. Stefan Zdraljevic has always been a constant presence in the Andersen Lab and was instrumental to both the lab's and my own successes. Besides being the one to teach me how to run the sorter (still not sure why it wasn't Shannon's job. . . ?), Stefan is intellectually curious and is constantly pushing the lab to the boundaries of quantitative genetics. Still my favorite memory of Stefan is when he unknowingly walked into the lab late on a Thursday night straight into me, alone and in full panic because I had just broken the arm of the sorter. He immediately jumped in to help, even though neither of us knew what we were doing, and we most definitely made things worse (think: many tiny ball bearings scattered across the lab floor. . .). I look up to Stefan as an example of a great scientist and I am excited to watch his future unfold.

Over the last five years, many people have come and gone from the Andersen Lab, each one contributing to my advancement as a scientist. Dr. Dan Cook was the first person to tell me about the lab and encouraged me to rotate. Dan was always fun to talk to and his dedication to science was impressive. During my rotation, he helped me with my coding projects and inspired me to always

continue learning. Clay Dilks and Loraina Stinson joined the lab as grad students after me and I quickly took upon myself the responsibility of the senior grad student to pass along all the knowledge I had collected about working in the lab, genetics, coding, work-life-balance, and anything else I felt inclined to share with thim, whether they asked for it or not. In addition, they have also served as my sounding board and my confidantes and I am thankful for their friendship and support, especially as stress has doubled during a global pandemic. Whether it's organizing trivia nights, feasting on Taco Diablo at the lake, or trying to solve the infamous mystery of the crystals, I know I can always count on Clay and Loraina. I wish them the best of luck as they tackle the grueling uphill climb that is graduate school.

I also wanted to acknowledge three postdocs who have been a calming and encouraging presence to me during this time. Dr. Tim Crombie has some of the most hilarious stories and is a great person to bounce new ideas off with. His attention to detail is impressive, and he always challenges me to carefully plan my experiments and to go the extra mile to convert ten lines of code into one. However, my favorite memories of Tim are when he would come into the lab late one morning because he had been reading the Harry Potter series for the first time (with my encouragement) and couldn't put it down. He surprisingly knew no spoilers and I loved hearing his predictions, especially when they were completely wrong! Dr. Gaotian Zhang's calm and reassuring attitude was always a refreshing change from the stress of grad school and helped me to gain perspective on life outside of the experiments I was currently running. Whenever I was nervous or upset, his answer would always be the same–a shrug of the shoulders with his classic catchphrase: "It's fine!". I am also thankful for Gaotian's continued dedication to our Harry Potter book club and am happy that I helped yet another person find their way to Hogwarts. Finally, Dr. Janneke Wit's optimistic attitude and willingness to learn is admirable. We bonded over Harry Potter (I am sensing a theme here. . .) and worked together to bring back some of the lab's social nature, even when quarantine made it difficult. In addition to these lovely scientists, I also had the pleasure of working with Joy Nyaanga, Dr. Lewis Stevens, Dr. Dan Lu, Dr. Sam Widmayer, Nicole Roberto, Dr. Ye Wang, Dr. Steffen Hahnel, and a large crew of technicians and undergraduate lab assistants who made my research possible.

My time in the lab would not have been as fulfilling nor as productive without the help of several high school and undergraduate students. My first mentee, Kimberly Collins, taught me a lot about mentoring

another student and made me fully appreciate how much work my undergraduate mentor, Dr. Simon Currie, had put into developing me as a scientist. Grace Park became my mentee unexpectedly and was a delight to work with in the lab. Self-sufficient and determined, Grace continued to come back day after day, fighting the dying food, contamination, and pressure to get results for her project. Ellen Chao is intelligent and a fast learner. I have no doubt she could tackle any task, no matter how difficult, with her smart intuition and dedication. Britney Sun was maybe just a high school student when I met her, but she was incredibly hard working and inquisitive and I admired her dedication and enthusiasm for research. And finally, Katie Introcaso (also known as Katie I. or Katie 2) is highly motivated, intellectually curious, and up for any task (even, and especially, cutting parafilm and chunking). Katie's bubbly personality brought fresh energy into the lab and it was always a joy working with her. As I continued to challenge her to grow, she would continue to rise to the occasion spectacularly, and was never afraid to ask questions, big or small. More than anything else, I have loved watching these baby scientists grow into their full potential. I am so proud of everything they have already accomplished and so excited for everything that they will accomplish in the future.

I am grateful to my committee members, Dr. Richard Morimoto, Dr. Sadie Wignall, and Dr. Thomas O'Halloran, for offering both scientific and career advice and challenging me to think outside the box. We had many engaging conversations during my committee meetings, some of which led to exploring new avenues of research for my projects that I might not have considered otherwise. I am also thankful for my funding sources, the Cellular and Molecular Biology of Disease training program and the NSF-Simons Center for Quantitative Biology.

In addition to all the wonderful people I met during graduate school, I want to take time to thank some of the people who helped me get to where I am today. My undergraduate thesis advisor, Dr. Barbara Graves, saw my potential and took a chance on me, allowing me to do research in her lab at the Huntsman Cancer Institute. Members of her lab, especially Dr. Niraja Bhachech, Dr. Litty Paul, and Dr. Simon Currie, were very involved in my early training as a scientist and were influential in my decision to go to graduate school. In particular, Simon was an amazing mentor to me and displayed enormous amounts of patience and understanding that I couldn't even fully comprehend until I was in his shoes years later. He taught me what it meant to be a mentor and what a responsibility it was: it is

not for the mentor's personal gain but rather for the mentee's. I am thankful that we have kept in touch over the years and I am excited to see what his future holds.

In addition to my fellow scientists, I'd like to thank my friends, family, and anyone else who, over the last five years, was kind enough to listen to me excitedly talk about worms and explain why they are so cool. First and foremost, my parents, Mike and Michelle Evans, and my sister, Megan Evans, have been with me from the very beginning and have helped me celebrate the highest of highs and helped me recover from the lowest of lows. They have always been there for me and continue to push me to be the best version of myself, both personally and professionally. My friends, especially Kristin Johnson, Iva Stojkovska, Dr. Jessica Outhet, Dr. Allyson Brome, Andres Lancheros, and Ariel Leonard, have each, in their own way, enhanced my life, kept me sane during grad school, and helped me remember that it is important to have a life outside of the lab. Finally, my brother, Andrew Evans, is my motivation for writing this dissertation and my hope that science can one day provide a better future.

Dedication

For Andrew Joseph Evans

A shooting star, bright but brief

Table of Contents

List of Tables **19**

List of Figures **21**

1 Introduction **25**

1.1 Quantitative variation . . . 25

1.1.1 Quantitative trait locus mapping . . . 27

1.1.2 From QTL to QTN . . . 29

1.2 *C. elegans* as a model organism . . . 32

1.2.1 Genomics of *C. elegans* . . . 34

1.2.2 High-throughput assays to quantify phenotypic variation in *C. elegans* . . . 35

1.3 QTL mapping in *C. elegans* . . . 36

1.3.1 Tools for QTL mapping in *C. elegans* . . . 37

1.3.2 Overview of quantitative traits mapped in *C. elegans* . . . 40

1.4 Lessons from quantitative genetics in *C. elegans* . . . 44

1.4.1 Benefits of combining linkage and association mapping . . . 45

1.4.2 Mediation analysis: combining expression and phenotypic QTL to identify causal genes . . . 46

1.4.3 QTN that led to a better understanding of the evolution of quantitative variation . . . 47

1.5 Dissertation overview . . . 48

1.6 SUPPLEMENTAL TABLES . . . 49

2 Shared genomic regions underlie natural variation in diverse toxin responses **52**

2.1 PREFACE . . . 52

2.2 ABSTRACT . . . 52

2.3 INTRODUCTION . . . 53

2.4 METHODS . . . 55

2.4.1 Strains . . . 55

2.4.2 High-throughput toxin response assay . . . 55
2.4.3 Toxin-response trait calculations . . . 56
2.4.4 Dose-response assays . . . 57
2.4.5 Principal Component Analysis of RIAILs . . . 57
2.4.6 Linkage mapping . . . 58
2.4.7 Heritability estimates . . . 59
2.4.8 Calculation of hotspots . . . 59
2.4.9 Generation of near-isogenic lines (NIL) . . . 59
2.4.10 Whole-genome sequence library prep and analysis . . . 60
2.4.11 Generation of chromosome substitution strains (CSS) . . . 60
2.4.12 Selection of traits to categorize in CSS and NIL assays . . . 61
2.4.13 Categorization of CSS and NIL results . . . 61
2.4.14 Statistical analysis . . . 63
2.5 RESULTS . . . 63
2.5.1 Identification of QTL underlying variation in responses to 16 diverse toxins . . . 63
2.5.2 Both additive and interactive QTL underlie toxin responses . . . 64
2.5.3 Three QTL hotspots underlie variation in responses to diverse toxins . . . 66
2.5.4 Near-isogenic lines recapitulate some of the predicted QTL effects . . . 67
2.5.5 Chromosome-substitution strains localize QTL underlying transgressive phenotypes . . . 74
2.6 DISCUSSION . . . 77
2.6.1 Pleiotropic regions underlie QTL shared between and among toxin classes . . . 77
2.6.2 A multi-faceted approach suggests that undetected epistatic loci impact toxin responses . . . 79
2.7 FUTURE DIRECTIONS . . . 81
2.8 CONTRIBUTIONS . . . 83
2.9 CITATION . . . 83

3 The gene *scb-1* underlies variation in *Caenorhabditis elegans* chemotherapeutic responses 84

3.1 PREFACE . . . 84

3.2 ABSTRACT . . . 84

3.3 INTRODUCTION . . . 85

3.4 METHODS . . . 86

3.4.1 Strains . . . 86

3.4.2 High-throughput fitness assays for linkage mapping . . . 87

3.4.3 Dose-response assays . . . 88

3.4.4 Linkage mapping . . . 88

3.4.5 Modified high-throughput fitness assay for NIL validation . . . 89

3.4.6 Expression QTL analysis . . . 90

3.4.7 Mediation analysis . . . 90

3.4.8 Statistical analysis . . . 91

3.5 RESULTS . . . 92

3.5.1 Natural variation on chromosome V underlies differences in responses to several chemotherapeutics . . . 92

3.5.2 Expression QTL mapping identifies a hotspot on the center of chromosome V . . . 93

3.5.3 Mediation analysis suggests that *scb-1* expression plays a role in responses to several chemotherapeutics . . . 98

3.5.4 Expression of *scb-1* affects responses to several chemotherapeutics that cause double-strand DNA breaks . . . 100

3.6 DISCUSSION . . . 101

3.6.1 Mediation of drug-response QTL using gene expression to identify causal genes 102

3.6.2 New evidence for the pleiotropic function of *scb-1* . . . 103

3.7 FUTURE DIRECTIONS . . . 105

3.8 CONTRIBUTIONS . . . 106

3.9 CITATION . . . 106

4 Natural variation in the sequestosome-related gene, *sqst-5*, underlies zinc homeostasis in *Caenorhabditis elegans* 108

4.1 PREFACE . . . 108

4.2 ABSTRACT . . . 108

4.3 INTRODUCTION . . . 109

4.4 METHODS . . . 111

4.4.1 Strains . . . 111

4.4.2 Standard high-throughput fitness assay . . . 111

4.4.3 Calculation of zinc-response traits . . . 112

4.4.4 Zinc dose response . . . 113

4.4.5 Linkage mapping . . . 113

4.4.6 Two-dimensional genome scan . . . 114

4.4.7 Construction of near-isogenic lines (NILs) . . . 114

4.4.8 Mediation analysis . . . 115

4.4.9 Generation of deletion strains . . . 115

4.4.10 Modified high-throughput fitness assay . . . 116

4.4.11 Local alignment of the *sqst-5* region using long-read sequence data . . . 116

4.4.12 Assessment of strain relatedness through neighbor-joining tree . . . 117

4.4.13 Genome wide association mapping . . . 117

4.4.14 Statistical analysis . . . 118

4.5 RESULTS . . . 118

4.5.1 Natural genetic variation in response to zinc is complex . . . 118

4.5.2 Near-isogenic lines fractionate the chromosome V QTL into multiple additive loci 120

4.5.3 Analysis of the chromosome III QTL suggests that a sequestosome-related gene, *sqst-5*, contributes to differences in zinc responses . . . 124

4.5.4 Variation in *sqst-5* underlies differences in zinc responses . . . 129

4.5.5 A natural deletion in *sqst-5* is conserved across wild isolates . . . 131

4.6 DISCUSSION . . . 136

4.6.1 A complex genetic architecture underlies differences in zinc responses 137
4.6.2 Common genetic variation underlies differential responses to exogenous zinc . . 139
4.6.3 SQST-5 might function to negatively regulate other sequestosome-related genes 141
4.7 FUTURE DIRECTIONS 142
4.7.1 From QTL to QTG on chromosomes IV, V, and X 142
4.7.2 Investigating the function of *sqst-5* and its mechanism of action in response to zinc 144
4.7.3 Expanding natural variation studies to include more strains and provide evolutionary context of zinc resistance 145
4.8 CONTRIBUTIONS 146
4.9 CITATION 147
4.10 SUPPLEMENTAL FIGURES 148
4.11 SUPPLEMENTAL TABLES 156

5 Gene-by-environment interactions influence docetaxel responses in *Caenorhabditis elegans* 158
5.1 PREFACE 158
5.2 ABSTRACT 158
5.3 INTRODUCTION 159
5.4 METHODS 161
5.4.1 Strains 161
5.4.2 High-throughput phenotyping assay 161
5.4.3 Dose response assay 162
5.4.4 Linkage mapping 163
5.4.5 Genotype-by-environment interactions 163
5.4.6 Modified high-throughput fitness assay 164
5.5 RESULTS 164
5.5.1 Genetic variation on chromosome IV influences animal length in response to docetaxel 164

5.5.2 Environmental variation of food quality dictates drug-response traits 169
5.5.3 The growth-response QTL on chromosome IV responds to environmental variation 171
5.5.4 Modified high-throughput assay does not facilitate candidate gene testing 176
5.6 DISCUSSION 177
5.7 FUTURE DIRECTIONS 178
5.8 CONTRIBUTIONS 179
5.9 SUPPLEMENTAL FIGURES 180

6 Genetic mapping of avermectin resistance in *Caenorhabditis elegans* identifies two novel loci that overlap with resistance loci in *Haemonchus contortus* 184

6.1 PREFACE 184
6.2 ABSTRACT 184
6.3 INTRODUCTION 185
6.4 METHODS 187
6.4.1 Strains 187
6.4.2 High-throughput fitness assays for linkage mapping 188
6.4.3 Abamectin dose response 188
6.4.4 Linkage mapping 189
6.4.5 Genome-wide association mapping 189
6.4.6 Identification of orthologous genes between *C. elegans* and *H. contortus* 190
6.4.7 Modified high-throughput fitness assay 190
6.4.8 Statistical analysis 191
6.5 RESULTS 191
6.5.1 Three additive loci on chromosome V underlie differential responses to abamectin 191
6.5.2 Near-isogenic lines validate effects of all three loci 195
6.5.3 Overlap of *C. elegans* and parasitic nematode candidate genes for avermectin resistance 201
6.5.4 Modified high-throughput assay does not facilitate candidate gene testing 204

6.6 DISCUSSION . . . 205
6.6.1 Linkage and association mapping identify similar, yet different, QTL . . . 205
6.6.2 Power of QTL mapping in *C. elegans* to identify causal genes underlying anthelmintic resistance . . . 206
6.6.3 Shared niches provide the same selective pressures for soil transmitted helminths and *C. elegans* . . . 207
6.7 FUTURE DIRECTIONS . . . 208
6.8 CONTRIBUTIONS . . . 210
6.9 SUPPLEMENTAL FIGURES . . . 211

7 Correlation of genotype with climate parameters suggest *Caenorhabditis elegans* niche adaptation 217
7.1 PREFACE . . . 217
7.2 ABSTRACT . . . 217
7.3 INTRODUCTION . . . 218
7.4 METHODS . . . 220
7.4.1 *C. elegans* wild isolate collection and sequencing . . . 220
7.4.2 Weather and climate data acquisition . . . 220
7.4.3 Association mapping . . . 222
7.4.4 Temperature competition assays . . . 223
7.5 RESULTS . . . 224
7.5.1 Genome-wide association of geographic traits . . . 224
7.5.2 Weather conditions and climate parameters can be determined using the geographic location of the site of strain isolation . . . 225
7.5.3 Genome-wide association of average relative humidity . . . 228
7.5.4 Genome-wide association of average temperature . . . 229
7.5.5 Strains from divergent climates might be adapted to specific temperatures . . . 231
7.6 DISCUSSION . . . 232

7.7 FUTURE DIRECTIONS . . . 236
7.8 CONTRIBUTIONS . . . 236
7.9 CITATION . . . 237

8 Discussion **238**

8.1 Functional validation to identify the causal variant underlying a QTL . . . 238
8.2 Benefits and limitations of high-sensitivity phenotyping . . . 240
8.3 Representing quantitative traits for mapping and validation . . . 241
8.4 Mediation analysis to improve candidate gene prioritization . . . 243
8.5 QTL in divergent regions present new challenges that need to be addressed . . . 244
8.6 A look forward: the future of quantitative genetics . . . 246

A Co-authored publications **278**

A.1 Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in *C. elegans* . . . 278
A.1.1 ABSTRACT . . . 278
A.1.2 CONTRIBUTIONS . . . 279
A.2 Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations . . . 279
A.2.1 ABSTRACT . . . 279
A.2.2 CONTRIBUTIONS . . . 280

B R Shiny application development **281**

B.1 High-throughput assay dilutions . . . 281
B.1.1 PREFACE . . . 281
B.1.2 EXPLANATION OF FUNCTIONALITY . . . 281
B.2 Fine-map QTL with NIL phenotypes . . . 282
B.2.1 PREFACE . . . 282
B.2.2 EXPLANATION OF FUNCTIONALITY . . . 282

B.3 PCR calculator 284
B.3.1 PREFACE 284
B.3.2 EXPLANATION OF FUNCTIONALITY 285

C Expression of the cadmium-responsive gene, *cdr-6*, does not influence responses to high exogenous zinc in *Caenorhabditis elegans* 286

C.1 PREFACE 286
C.2 RESULTS AND DISCUSSION 286

List of Tables

Chapter 1:

Table 1-1 Overview of different N2xCB4856 recombinant panels 39

Table S1-1 List of QTL mapping studies in *C. elegans* 49

Chapter 2:

Table 2-1 Summary of toxins and principal components mapped per hotspot 68

Table 2-2 All traits tested in NIL and CSS assays 69

Chapter 3:

Table 3-1 Main mechanism of action for eight chemotherapeutic drugs 93

Chapter 4:

Table 4-1 Genetic variation in CB4856 for zinc-related genes in *C. elegans* 122

Table 4-2 Genes in QTL intervals for chromosome III and V 129

Table S4-1 Geographic locations of 57 wild isolates with putative structural variation in *sqst-5* 156

Chapter 5:

Table 5-1 Results of ANOVA test of genotype, food quality, and genotype-by-food interaction on length in DMSO 175

Table 5-2 Results of ANOVA test of genotype, food quality, and genotype-by-food interaction on length in docetaxel 175

Table 5-3 Results of ANOVA test of genotype, food quality, drug condition, and various interactions 176

Chapter 6:

Table 6-1 Genomic regions significantly correlated with abamectin resistance 199

Chapter 7:

Table 7-1 Definition of geographic and weather traits 225

List of Figures

Chapter 1:

Figure 1-1 Human height as a quantitative trait 25

Figure 1-2 Refining QTL intervals using near-isogenic lines (NILs) 31

Figure 1-3 Testing causality with genome editing 33

Figure 1-4 High-throughput assay workflow with COPAS BIOSORT 36

Figure 1-5 Toxin-response QTL mapping studies 42

Chapter 2:

Figure 2-1 Diverse genetic architectures are implicated in responses to 16 toxins . . . 65

Figure 2-2 Additive genetic components identified by linkage mapping do not explain all heritable contributions to toxin-response variation 66

Figure 2-3 Three QTL hotspots impact toxin responses 67

Figure 2-4 NIL and CSS trait categorizations 70

Figure 2-5 Results from near-isogenic line (NIL) and chromosome-substitution strain (CSS) tests of recapitulation of QTL effects are categorized based on potential genetic mechanisms implicated in toxin responses 73

Figure 2-6 Categorization of correlated traits 74

Figure 2-7 A model for potential locations of two loci is shown, according to toxin-response phenotypes of near-isogenic lines (NILs) and chromosome-substitution strains (CSSs) 75

Chapter 3:

Figure 3-1 A large-effect QTL on the center of chromosome V underlies responses to several chemotherapeutics 94

Figure 3-2 Near-isogenic lines validate the chromosome V QTL 95

Figure 3-3 Expression QTL mapping identifies several hotspots 97

Figure 3-4 Model for gene expression as a mediator of a drug-response QTL 98

Figure 3-5 Linkage mapping summary for drug-response traits in the set 1 RIAILs . . . 99
Figure 3-6 Mediation analysis for the eQTL hotspot on the center of chromosome V . . 100
Figure 3-7 Testing the role of *scb-1* in drug responses 101

Chapter 4:

Figure 4-1 Dose response with four divergent wild isolates 119
Figure 4-2 Linkage mapping identifies four QTL in response to high dietary zinc 121
Figure 4-3 Two dimensional genome scan for median optical density (median.EXT) in zinc 122
Figure 4-4 Reaction norm shows additive QTL effects between chromosome III and V . 124
Figure 4-5 Validating QTL using near-isogenic lines (NILs) 125
Figure 4-6 NILs identify multiple QTL on chromosome V 126
Figure 4-7 Dose response for modified HTA 127
Figure 4-8 Dominance of chromosome III QTL 128
Figure 4-9 Expression QTL mapping and mediation analysis for *sqst-5* 130
Figure 4-10 The gene *sqst-5*, not *ver-2*, has an eQTL 131
Figure 4-11 Testing the role of *sqst-5* in the zinc response 132
Figure 4-12 Isolating the effect of *sqst-5* in the zinc response 133
Figure 4-13 The gene *sqst-5* confers zinc sensitivity 134
Figure 4-14 Natural genetic variation in *sqst-5* 135
Figure 4-15 Map of wild isolates 136
Figure 4-16 GWA mapping suggests common variation at *sqst-5* is associated with differences in zinc responses among wild isolates 138
Figure 4-17 Linkage mapping summary for drug-response traits in response to five heavy metals 140
Figure S4-1 Linkage mapping identifies 12 QTL across three traits in response to high zinc 151
Figure S4-2 GWA mapping identifies QTL in response to high zinc 155

Chapter 5:

Figure 5-1 Docetaxel dose response 165

Figure 5-2 Linkage mapping results for animal length in docetaxel 166

Figure 5-3 NILs validate that the N2 allele on chromosome IV confers resistance to docetaxel 167

Figure 5-4 NILs narrow the QTL to a 2 Mb region 168

Figure 5-5 Expression QTL mapping for *rmd-4* 169

Figure 5-6 Nematodes grown in the bad lysate do not reach L4 stage within 48 hours after feeding 170

Figure 5-7 An older aliquot of the same HB101 lysate changes the parental response to docetaxel 171

Figure 5-8 The lysate hypothesis (GxE interactions) 172

Figure 5-9 Growth-response QTL in DMSO control 173

Figure 5-10 Reaction norm plots identify significant GxE interactions 174

Figure 5-11 Modified high-throughput assay does not capture phenotypic differences between N2 and CB4856 strains on chromosome IV 177

Figure S5-1 Linkage mapping results for docetaxel response 183

Chapter 6:

Figure 6-1 Dose response with four divergent wild isolates 192

Figure 6-2 Three large-effect QTL on chromosome V control differences in abamectin responses 194

Figure 6-3 Two-dimensional genome scan for mean optical density (mean.EXT) in abamectin 195

Figure 6-4 The three loci on chromosome V are independent and additive 196

Figure 6-5 Summary of QTL mapping for responses to abamectin 196

Figure 6-6 Chromosome substitution strains validate the existence of one or more resistance loci on chromosome V 197

Figure 6-7 Near-isogenic lines confirmed the additive effects of all three QTL 198
Figure 6-8 Refining QTL positions with NILs 199
Figure 6-9 NILs isolate and narrow the VC QTL 201
Figure 6-10 NILs validate and narrow the VC QTL 202
Figure 6-11 Testing the role of *lgc-54* in the *C. elegans* abamectin response 203
Figure 6-12 Modified high-throughput assay does not capture phenotypic differences between N2 and CB4856 strains 204
Figure 6-13 NILs to isolate and narrow the VL QTL 209
Figure S6-1 Genome-wide association (GWA) mapping identifies 15 QTL across three traits in response to abamectin 213
Figure S6-2 Linkage mapping identifies 10 QTL across three traits in response to abamectin 216

Chapter 7:

Figure 7-1 Global distribution of wild isolates and NOAA weather stations 226
Figure 7-2 Genome-wide association of elevation 227
Figure 7-3 Genome-wide association of humidity traits 229
Figure 7-4 Genome-wide association of temperature 230
Figure 7-5 Two-strain temperature competition assay 233

Appendix:

Figure B-1 Screenshots of the HTA dilutions app 282
Figure B-2 Screenshot of the Finemap QTL Shiny application 283
Figure B-3 Screenshot of the PCR Calculator Shiny application 285
Figure C-1 Testing the role of *cdr-6* in the *C. elegans* zinc response 288

1 Introduction

1.1 Quantitative variation

Phenotypic variability is a common feature in all populations and is necessary for the advancement and evolution of a species. Individuals vary in numerous ways ranging from physical attributes such as height, eye color, or skin color to behavioral attributes such as social habits or sleep patterns. This variation is what drives individuality and allows organisms to adapt to different environments and perform complementary functions within a population [1]. Some of these traits might be qualitative, meaning values can be categorized into discrete bins. For example, humans can have one of eight different blood types (A, B, AB, or O for either RhD positive or negative). However, most traits are quantitative, meaning that values exist on a continuous scale. A classic example of a quantitative trait is human height. Although a majority of adults measure between 155.2 and 170.2 cm, individuals can fall anywhere on this distribution ranging from the shortest man in the world (54.6 cm) to the tallest (272 cm) (**Figure 1-1**) [2]. One of the main goals in scientific research today is to characterize phenotypic diversity and understand its causes.

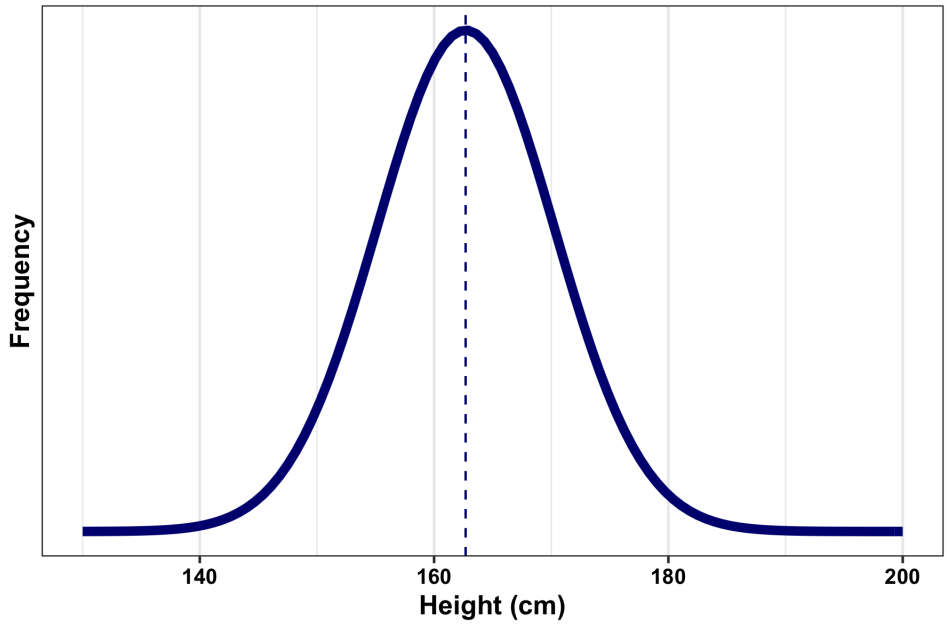


Figure 1-1: Human height as a quantitative trait. Distribution of heights is shown as a quantitative trait. Average human height is 162.7 cm with a standard deviation of 7.5 cm. Data for figure adapted from [2]

All phenotypic variation (V_P) can be described using a combination of genetic (V_G) and environmental (V_E) factors (**Equation 1**). Some traits–often referred to as Mendelian traits, after the Austrian monk Gregor Mendel–are mostly influenced by genetics. Examples of these traits range from color blindness in humans [3] to the classic color and shape of sweet peas observed by Mendel [4]. However, an individual's environment also plays a major role in determining most phenotypic traits. For example, Darwin's finches evolved beaks of different sizes and shapes in response to the divergent environments across the Galapagos islands [5] and a person who smokes cigarettes is more likely to develop lung cancer than a non-smoker [6]. Unfortunately, these environmental factors are often difficult to identify and control within a population. Heritability of a trait (H^2) is often calculated to define the proportion of the total phenotypic variation that can be attributed solely to genetic factors (**Equation 2**). The broad field of genetics aims to describe how this genetic variation translates into phenotypic variation. Understanding how genetic variants influence complex traits will provide information important for predicting disease risk in human populations, increasing the speed of selective breeding programs in agriculturally important plants and animals, and predicting adaptive evolution within species [7].

$$V_P = V_G + V_E \tag{1}$$

$$H^2 = V_G / V_P \tag{2}$$

Since Mendel's pea experiments, scientists have been trying to connect phenotypic differences, from either natural variation within a population or laboratory-induced variation by mutagenesis, to underlying genes. In 1913, 19-year-old geneticist Alfred Sturtevant, working under the supervision of Thomas Hunt Morgan, created the first genetic map showing the approximate locations of six genes on the *Drosophila* X chromosome [8]. This technique, known as genetic mapping, became the foundation of classical forward genetics, an approach by which scientists were able to narrow the position of the variant gene driving phenotypic differences in mutant individuals using known physical markers across the genome. With the advancement of technology, molecular markers such as single nucleotide variants (SNVs), short-tandem repeats (STRs), or insertion/deletion (indel) variants are now commonly used

and allow for more precise gene mapping [9]. In addition to the numerous examples of successes of gene cloning from mutagenesis screens in model organisms [10, 11, 12, 13, 14, 15], this method has been used to identify disease-causing genes in humans for Huntington's disease [16], muscular dystrophy [17, 18], and cystic fibrosis [17, 19, 20].

1.1.1 Quantitative trait locus mapping

Classical genetic mapping is a simple and powerful tool for mapping large-effect Mendelian traits driven by a single gene. However, this technique becomes much more cumbersome as traits become more complex. An alternative method for identifying the genetic basis of complex traits is quantitative trait locus (QTL) mapping. This technique relies on statistical methods to correlate genetic variation with phenotypic variation across a population of genetically diverse individuals. QTL mapping is not a replacement of classical genetic mapping, but rather provides a complementary approach to understanding the genetics of complex traits. There are several advantages to QTL mapping. First, a quantitative approach allows identification of small-effect loci in addition to large-effect loci and can simultaneously map several loci for the same trait. Second, because QTL mapping relies on natural genetic variation between two or more individuals, this approach allows scientists to investigate how evolution could have shaped diverse phenotypes. In particular, QTL mapping has the power to identify SNVs that might alter, rather than break, gene function. Identifying the functional variants that underlie phenotypic differences might lead to a better understanding of how the forces of evolution act to bring about genetic adaptation in nature.

There are two complementary mapping approaches: linkage mapping and genome-wide association (GWA) mapping. Linkage mapping is similar to classical genetic mapping and leverages statistical power from a large number of recombinants generated from a cross between two or more genetically diverged individuals. Alternatively, association mapping takes advantage of the breadth of natural genetic diversity that exists among a large panel of genetically distinct individuals within the species [7]. Both mapping approaches share a common goal of identifying functional variants that contribute to phenotypic diversity and involve two main steps: detection and localization.

QTL can be easily detected using either method for large-effect variants with evenly distributed allele

frequencies across the population (close to 0.5). However, the power to detect QTL will differ between methods depending on the population size, the effect size of the variant, and the allele frequency of the variant within the mapping population. In general, as the effect size becomes smaller and the allele frequency becomes less evenly distributed (further from 0.5), the number of individuals in the mapping population must be increased to maintain sufficient power to detect the QTL. In cases where the allele frequency is extremely skewed (*i.e.* a rare variant), linkage mapping will have greater power to detect QTL than association mapping [7]. This increased power stems from the fact that allele frequencies are more evenly distributed in a population that is from a cross between two individuals, thus magnifying the effect of rare variants. However, it is necessary that both alleles are present within the limited number of individuals that generate the recombinant population. Alternatively, although association mapping might not have the power to detect rare variants, all possible alleles are present in the mapping population, which increases the probability of finding phenotypic variation that is driven by genetic factors.

The power to localize QTL depends on the population structure, which is different between linkage and association mapping panels by design. With a recombinant panel, the precision of QTL mapping depends on the number of recombination events. Higher recombination frequencies and/or more recombinant individuals will identify a smaller QTL interval. Alternatively, association mapping relies on historical recombination between individuals in a population. Species with higher historical recombination (which translates to low linkage disequilibrium, LD) have higher power to localize QTL to smaller regions of the genome without the need for a large population size [7].

Despite the advancement of technology and the detection of QTL underlying hundreds of complex traits, several key questions about quantitative variation still remain heavily debated among scientists. First, what is the underlying genetic architecture of complex traits? Although many studies have identified few large-effect loci that explain much of the heritable phenotypic variation within a population, others contend that many small-effect loci together define complex traits. Advocators for this highly polygenic (or even omnigenic [21]) nature of phenotypic variation argue that biological pathways are complex and evolution acts on a large number of individually undetectable small-effect loci [22]. Second, what is the relationship between QTL? Many believe that the majority of phenotypic variation is composed of additive effects between detected (and undetected) QTL

[23, 24, 25, 26, 27, 28]. However, some argue that genetic epistasis (interactions between two or more loci) play a more significant role in defining complex traits [29, 30, 31, 32, 33]. Finally, where in the genome are functional variants most commonly found? Many known quantitative trait nucleotides (QTNs) are protein-coding genetic variants that cause large-effect Mendelian-like phenotypic variation. However, noncoding genetic variation might be more evolutionarily important, although not always detected in QTL mapping studies because these variants often cause smaller phenotypic effects [34, 22]. These questions, among others, drive scientists in fields from evolution to population genetics to identify functional variants in the hope of explaining the molecular basis of phenotypic evolution.

1.1.2 From QTL to QTN

Linkage and association mapping have identified QTL in numerous diverse species from humans to plants and animals, and even single-celled organisms. However, model organisms are best suited to take full advantage of the goals of QTL mapping and answer questions regarding the molecular basis of complex traits and the evolution and function of quantitative alleles. In contrast to other species, model organisms are cheap and easy to grow in the laboratory and often have compact and defined genomes. High-throughput assays allow for quick and scalable phenotyping to easily test the high numbers of individuals necessary for QTL mapping. Most importantly, model organisms allow for functional validation of quantitative trait genes (QTG) and quantitative trait nucleotides (QTN), which is essential for proving causality. Several methods have been developed to aid researchers in discovering the causal gene or variant that underlies a QTL. Some of the most common methods include fine mapping the QTL interval using additional genetic markers, narrowing the QTL interval with near-isogenic lines, and testing causality with genome editing tools such as the CRISPR-Cas9 machinery.

Refining QTL with fine mapping and near-isogenic lines. Often the first approach to narrow QTL, fine mapping involves testing additional genetic markers or strains to improve the QTL resolution. For a linkage mapping analysis, this often translates to generation of more recombinants within the defined genomic interval. Alternatively, fine mapping for an association mapping analysis relies on

additional information that can be acquired by analyzing all SNVs in the defined genomic interval. This is in contrast to the few genetic markers analyzed during the previous mapping that were selected to represent unique LD blocks.

QTL mapping uses statistical methods to identify regions of the genome that significantly correlate to phenotypic variation. To empirically validate the effect of the discovered QTL, researchers can generate and phenotype near-isogenic lines (NILs) (**Figure 1-2**). NILs (sometimes referred to as introgression lines (ILs) or congenics) are strains where a small genomic interval from one strain is introgressed into the genetic background of another strain [35, 36]. Any differences in phenotype between the NIL and the parental strain with the same genetic background is due to the introgression of the QTL from the opposite genotype (**Figure 1-2**). Once the QTL is isolated and validated, a panel of NILs with smaller introgressions that tile across the genomic interval can be created by backcrossing to the parental strain of the NIL genetic background and selecting for individuals with recombination events within the introgressed region. The phenotypes of these NILs can be used to further refine the QTL interval by identifying the smallest overlapping region within the introgressions that maintains the QTL effect observed in the largest introgression (**Figure 1-2**). In the hypothetical example provided in **Figure 1-2**, NILs 3, 4, and 6 phenocopy Parent 1, suggesting that their introgressed regions (in green) contain the QTL. Alternatively, NILs 5, 7, and 8 phenocopy Parent 2, suggesting that their introgressed regions do not contain the QTL. From these results, we can conclude that the QTL effect observed in NIL 1 can be narrowed to the small genomic region from the overlap of introgressions between NIL 4 and NIL 6.

Testing causality with CRISPR-Cas9 genome editing. Candidate genes can be prioritized from a small genomic region based on their predicted gene function and their extent of genetic variation in the mapping population. Genes with no genetic variation are not likely to be causal and genes with genetic variation that causes a change in either the amino acid sequence of the protein (protein-coding variation) or expression levels of the protein are most likely to be causal. Candidate genes can be tested for causality using genome-editing tools such as the CRISPR-Cas9 system which takes advantage of the RNA-dependent DNA nuclease Cas9 of *Streptococcus pyogenes* [37, 38, 39]. Paired with a specially designed RNA molecule (sgRNA), the Cas9 protein forms targeted double-stranded

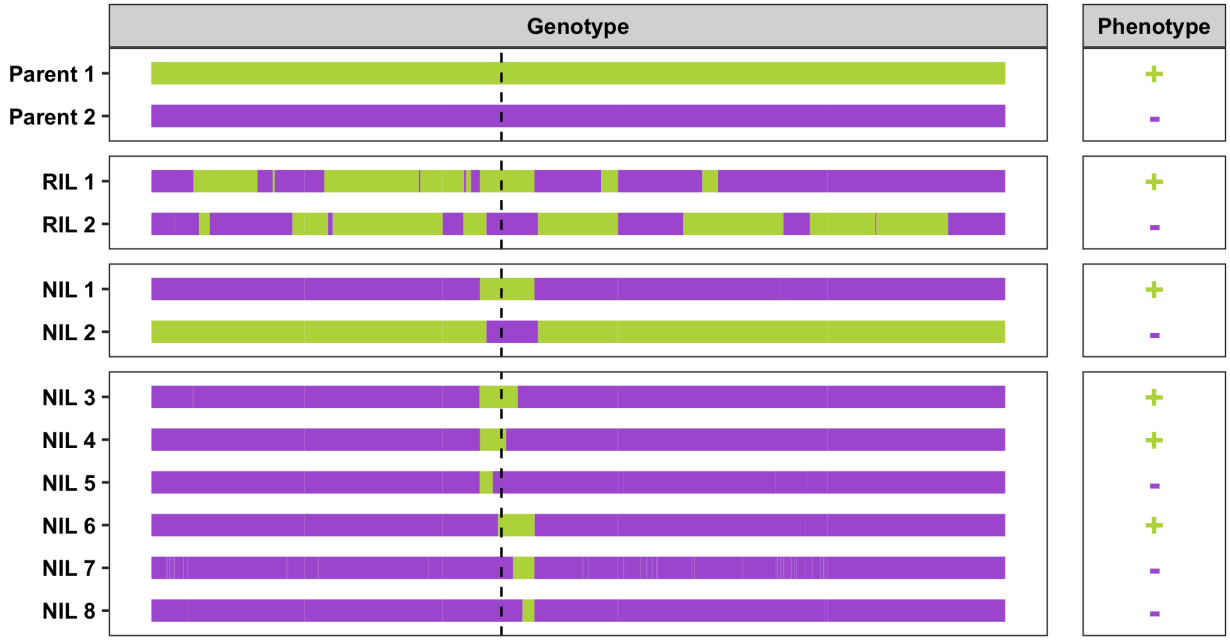


Figure 1-2: Refining QTL intervals using near-isogenic lines (NILs) Parental strains (green and purple) have different phenotypes (plus or minus). Recombinant strains generated from crossing the parental strains can be used to identify quantitative trait loci (QTL) represented here by a dotted vertical line. Recombinants can be backcrossed for several generations to isolate a single introgressed genomic interval. These near-isogenic lines (NILs) can empirically validate the QTL if the genotype of the introgressed region determines the strain phenotype. A NIL with a large introgression can be divided into a panel of NILs with smaller introgressions that tile across the region by backcrossing the NIL and selecting for recombination events. This panel of NILs can be used to narrow a large interval to a smaller interval with a set of candidate genes for causal testing.

DNA breaks that are repaired by either non-homologous end joining (creating a deletion) or homology-driven repair (creating an insertion) [40].

There are several ways CRISPR can be used to identify causal genes or variants. First, CRISPR can be used to identify a causal gene by loss-of-function studies. Targeted gene deletions that take advantage of the cell's non-homologous end joining (NHEJ) pathway can be achieved by designing two sgRNAs, one that targets each end of the gene of interest [39]. NHEJ is the product of blunt re-ligation of the two pieces of DNA and often results in a deletion of one or more nucleotides, sometimes resulting in a reduction or loss of gene function [41]. If a loss-of-function allele in one strain is causing the difference in phenotype, creating a loss-of-function allele in the strain with the functional copy should result in a similar phenotype to the strain with the natural loss-of-function allele (**Figure 1-3A**). However, because loss-of-function alleles can sometimes have negative fitness consequences, a reciprocal hemizygosity

test can be performed to justify the effect of genetic variation within a single gene [42]. In this test, a null mutation is first generated in each genetic background. Mutants are then crossed to the wild-type strain of the opposite genetic background, generating reciprocal hybrids that are genetically identical except for the candidate gene, which is hemizygous. Any differences in phenotype between the two hybrid hemizygous strains can be attributed to the presence of genetic variation within the candidate gene (**Figure 1-3B**) [42]. Finally, for a gene that is shown to contribute to phenotypic variation using this method, CRISPR-Cas9 can then be used to edit a single nucleotide to identify a causal genetic variant. Taking advantage of the homology-driven repair (HDR) pathway, researchers can design a single sgRNA near the desired edit and also supply a repair template with long homologous arms. HDR uses the supplied repair template to incorporate the new desired sequence into the genome [41]. These “allele swaps” where the genetic variant from one strain is introduced into a different genetic background allow for functional tests of a single variant. Phenotypic differences between the single-allele-edited strain and the parental strain suggest that the introduction of this single allele underlies the phenotype of interest (**Figure 1-3C**).

1.2 *C. elegans* as a model organism

In June 1963, ten years after the revolutionary discovery of the double-helical structure of DNA, well esteemed bacterial geneticist Dr. Sydney Brenner began the search for a new model organism. He dreamed of an organism simple enough for thorough genetic analysis yet complex enough to answer his burning scientific questions related to organismal behavior and understanding of the nervous system. He began to scour zoology textbooks in search of a single organism that could fulfill his three basic demands: a quick life-cycle, a simple reproductive cycle and genome, and a small size [43].

Brenner soon settled on the free-living roundworm nematode *Caenorhabditis elegans*. About 1 mm in length (the size of a comma on a printed page), this nematode was named due to the *elegant* way in which it moved. In addition to being microscopic, *C. elegans* are transparent, allowing for visualization of single cells under a microscope. *C. elegans* grow well in the laboratory, either on agar plates or in liquid culture, feeding on *E. coli* bacteria. The hermaphroditic nematode can produce up to 300 identical progeny in three to four days, making it a great system for genetic screens and selections.

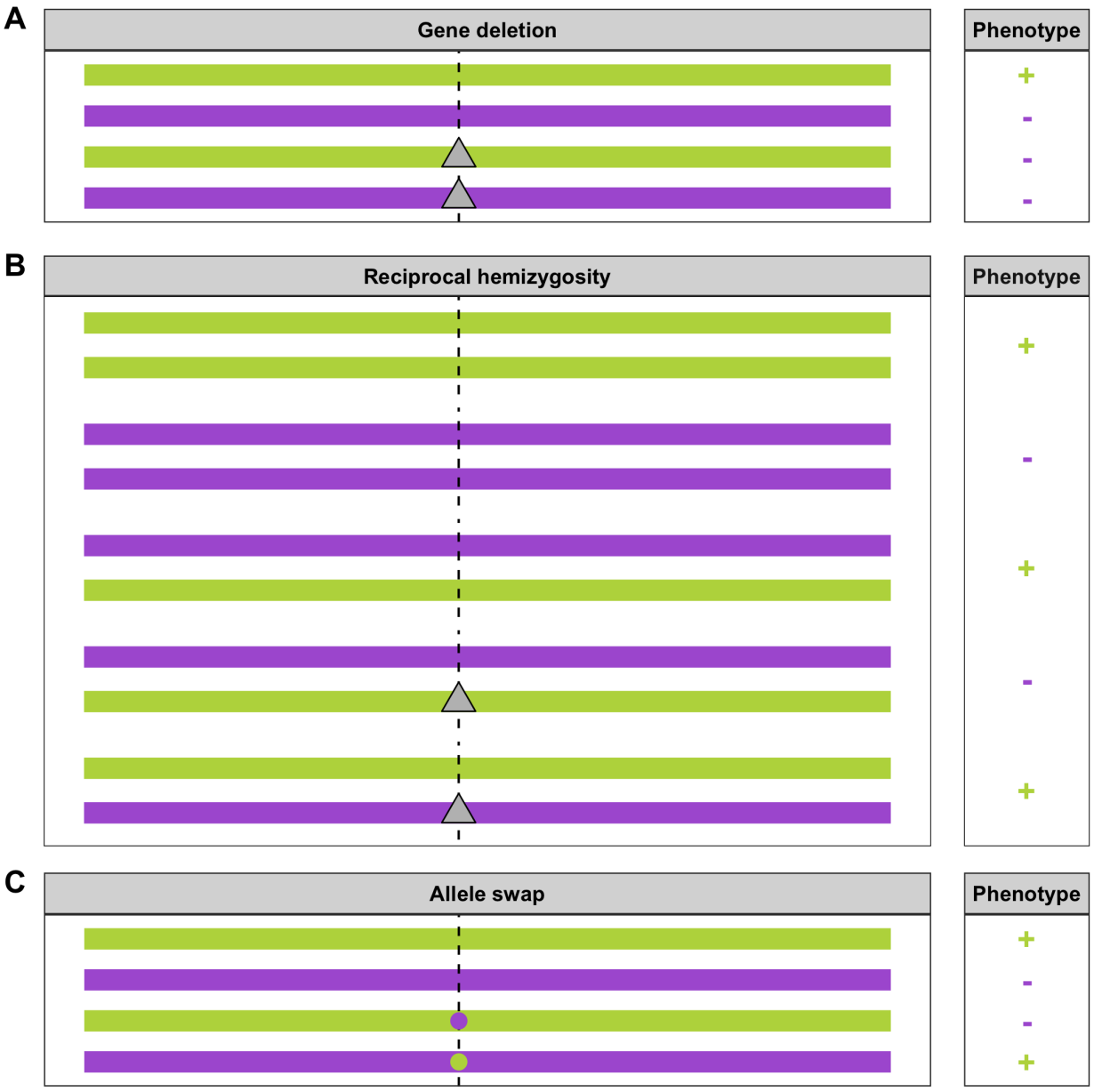


Figure 1-3: Testing causality with genome editing Genotype of parental strains (green and purple) and genome edited strains on the left and phenotype (plus or minus) on the right for gene deletions (A), reciprocal hemizygosity (B), and allele swaps (C). A single rectangle represents a homozygous genome for (A) and (C) while each chromosome of a diploid organism is shown for (B). The QTL position is represented by the dotted vertical line. Grey triangles represent gene deletions and purple or green dots represent a single nucleotide exchange between the parental backgrounds. **(A)** Natural gene loss-of-function in the purple strain leads to a negative phenotype similar to deleting the gene in both parental backgrounds. **(B)** Heterozygous individuals are phenotyped for a reciprocal hemizygous assay. The reciprocal hybrid strains hemizygous for the gene of interest (bottom two) display different phenotypes, suggesting that genetic variation in the gene of interest is causal for the variation in phenotype. **(C)** Replacing a SNV in the green parent with the allele from the purple parent (and vice versa) causes a change in phenotype, suggesting this allele is causal for the variation in phenotype.

Although primarily hermaphroditic, *C. elegans* produce males at a low frequency (< 1% in the laboratory strain) allowing for ease of generation of new recombinant strains via simple genetic crosses [43]. Targeted genome editing provides a complementary approach to generating new strains by deleting specific genes or introducing new genes or alleles to the genome [44]. Importantly, nematodes can be cryopreserved for long-term storage without accumulation of mutations [43].

In 1974, Brenner published the first seminal paper on the genetics of *C. elegans* in which he identified close to 300 mutants affecting behavior and morphology [43]. Although Brenner was not the first to discover *C. elegans* nor the first to study the nematode, he is often credited for establishing the worm as a model organism for cellular and molecular biology, neurobiology, developmental biology, and genetics. Since 1974, over 31,000 papers have been published that mention the nematode *C. elegans*. A team of scientists led by Dr. John Sulston and Dr. Robert Horvitz characterized the entire cell lineage of *C. elegans* from embryo to adult through its four larval stages [45, 46]. The nematode remains to this day the only animal with a comprehensive understanding of every neuronal connection in the brain [47, 48, 49, 50, 51].

1.2.1 Genomics of *C. elegans*

In 1998, *C. elegans* became the first multicellular organism to have its genome fully sequenced. The relatively small 100 Mb genome comprises six chromosomes (five autosomes and one sex chromosome) and approximately 20,000 protein-coding genes [52]. In comparison, the human genome is over 3,000 Mb, has 23 chromosomes, yet still has around 20,000 protein-coding genes [53, 54]. It is estimated that more than two-thirds (13,400+) of the protein-coding genes in *C. elegans* have a human homolog [55]. Many prominent pathways that are misregulated in cancer or targeted by therapeutics including Ras/MAPK, Wnt, Notch, TGF-β, and insulin signaling pathways are conserved in nematodes [56]. In fact, many of these fundamental pathways were first identified and characterized in *C. elegans*, prompting future studies in mammalian systems [57].

Most *C. elegans* research, including Brenner's fundamental studies, uses the single reference strain N2. First isolated from a mushroom compost bin in Bristol, UK in 1951, this strain underwent somewhere between 300 and 2,000 generations before it was first cryopreserved in 1969 [58]. During

this time, the N2 strain was exposed to selective pressures in the laboratory and accumulated several mutations that led to a laboratory-adapted strain drastically different from its wild ancestor [58]. Since that time, hundreds of wild strains have been collected that represent the species in the absence of laboratory adaption and have enabled ecological studies of the species [59]. In particular, a wild isolate from Hawaii, CB4856, has been shown to be remarkably diverged from N2 [60, 61, 62] and quickly became a second reference strain used in many studies. Currently, there are 403 genetically distinct strains, known as isotypes, isolated worldwide that describe relatively high genetic diversity. Information about these strains including the date and location of isolation and the whole-genome sequencing results can be found at the *C. elegans* Natural Diversity Resource (CeNDR; elegansvariation.org) [63]. Researchers can leverage this species-wide diversity to better understand how natural genetic variants underlie phenotypic differences such as variation in toxin responses.

1.2.2 High-throughput assays to quantify phenotypic variation in *C. elegans*

In order to best capture both small and large phenotypic differences between strains, high-throughput assays have been developed using a variety of platforms. Important characteristics of any high-throughput phenotyping assay include high accuracy, precision, and throughput with minimal human interaction to avoid errors. Some common phenotyping platforms include automated fluorescence microscopy [64], microplate readers [65, 66], microfluidic chips [67], and large-particle flow cytometers [68].

In one such example, a large-particle flow cytometer referred to as the “worm sorter” (COPAS BIOSORT, Union Biometrica) measures animal development (length and optical density) and reproductive ability (brood size) in the presence of different toxins [69, 70, 71, 72, 73, 74, 75, 68] or under other environmental conditions [76, 68] (**Figure 1-4**). Nematodes get longer and more optically dense (thicker and denser body composition) as they develop [68]. In the presence of most toxins, nematodes have smaller broods, shorter lengths, and are less dense compared to non-treated animals. This system has the power to distinguish phenotypic effect sizes as small as 5% provided that a high number of replicates are scored [73, 68]. And because the worm sorter can analyze hundreds of animals in each well of a 96-well plate in just over 23 minutes, high replication is rarely an

issue. Lastly, large amounts of data collected with the worm sorter can be easily processed and analyzed using a collection of packages developed and maintained by the Andersen Lab [77, 68].

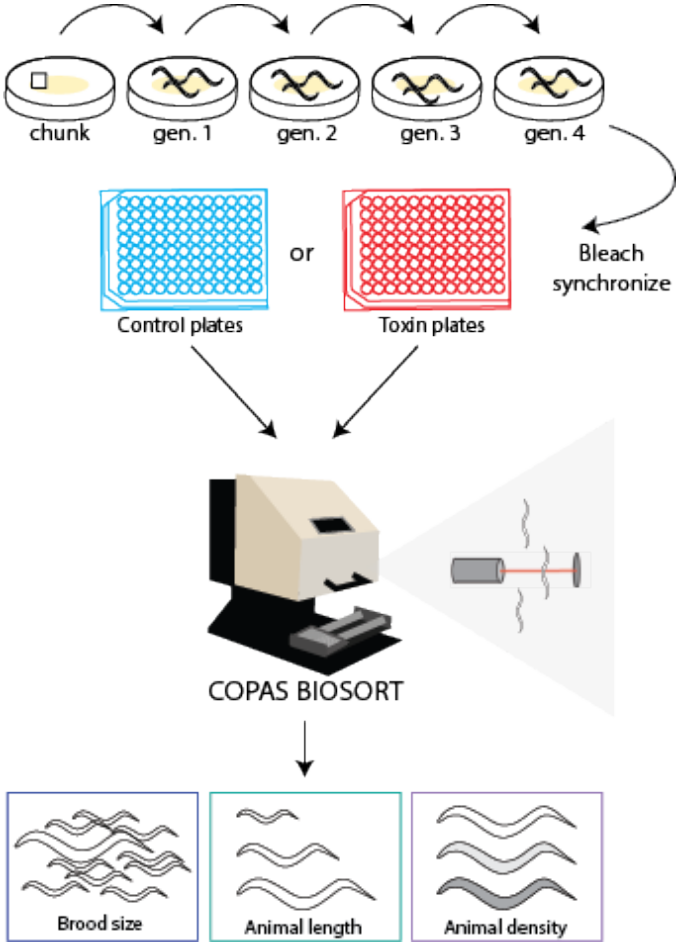


Figure 1-4: High-throughput assay workflow with COPAS BIOSORT Animals are propagated for multiple generations. Embryos are prepped by bleach synchronization and aliquoted to 96-well plates. L1s hatch overnight and are fed to allow development to the L4 larval stage. Three L4s are then sorted into a new well in a 96-well plate containing either drug or control growth media using the COPAS BIOSORT large-particle flow cytometer (Union Biometrica). The animals are grown for four days at 20 ℃, during which time they will mature to adulthood and lay embryos that will subsequently mature. Directly before scoring fitness traits, the animals were treated with sodium azide to straighten their bodies and allow for more accurate length and optical density measurements. Phenotypic measurements obtained by the COPAS BIOSORT include brood size, animal length, and optical density.

1.3 QTL mapping in *C. elegans*

Thanks to recent advancements in technology, the ability to discover QTL is now limited mostly by the number of traits that can be accurately and efficiently phenotyped. However, depending on the

recombination frequency (linkage mapping) or linkage disequilibrium (GWA) between strains, QTL could span over several megabases containing hundreds of genes and thousands of variants. The genetic marker with the highest correlation is not always the causal variant, in fact it usually is not. One major advantage of QTL mapping in *C. elegans* compared to other organisms is the ability to fine-map and narrow large QTL intervals methodically and efficiently. Although QTL studies in some organisms are merely suggestive of a correlative relationship between a phenotype and a genetic marker and all its linked loci, functional studies in *C. elegans* allow researchers to directly test the causal relationship between an allele and phenotype.

1.3.1 Tools for QTL mapping in *C. elegans*

One strong advantage about QTL mapping in *C. elegans* is that strains can be frozen and recovered, allowing for the preservation of mapping sets for future phenotyping. Early association mapping studies used 41 [78] to 96 [79] genetically distinct wild isolate strains. However, there are currently 403 strains in CeNDR available for association mapping. Significance of a genotype-phenotype association can be calculated using several methods including a *t*-test or a Wilcoxon rank-sum test. However, the Andersen Lab developed the R package *cegwas2* (github.com/AndersenLab/cegwas2-nf) for association mapping in *C. elegans* that incorporates a strain-by-strain relatedness matrix K as a random effect in the linear mixed-model with the equation $y = X + Zu + e$ where y is the measured phenotype, X is the genetic variant to be tested for association, Zu is the phenotype y corrected for population structure K, and e are residual effects [70].

Any two (or more) of the 403 strains can be used to generate a recombinant panel for linkage mapping. Phenotypically diverse strains improve the likelihood of detecting QTL for a given trait while genetically diverse strains increase the number of potentially causal alleles, making it more difficult to localize the QTL. The most common set of strains used to perform linkage mapping analysis are N2 and CB4856. A wild isolate from Hawaii, CB4856 has long emerged as a second reference strain that is highly divergent from the laboratory-adapted N2 strain [60, 61, 62]. Over the years, there have been several independently generated panels of recombinants between N2 and CB4856 (**Table 1-1**). The first panel of 80 recombinant inbred lines (Kammenga RILs) was generated in 2006 [80] and later led to

the first gene identified using QTL mapping in *C. elegans* [81]. A few years later, a second panel of 236 recombinants (Rockman RIAILs) was generated that took advantage of an advanced crossing design [82, 83] to generate recombinant inbred advanced intercross lines (RIAILs) [78]. In contrast to RILs that are derived from 20 generations of inbreeding from an F2 cross, RIAILs are the result of ten generations of random pair mating followed by ten generations of selfing [82, 83]. Because of the design differences between the two crosses, RIAILs will contain more sites of recombination and should allow for higher resolution [82]. These RIAILs were first used to map the copulatory plugging phenotype to the gene *plg-1* [84, 78].

However, after its generation, it was discovered that the Rockman RIAIL panel contains a strong skew in allele frequencies favoring the N2 allele at the *peel-1 zeel-1* incompatibility locus on chromosome I [85, 86]. Additionally, the laboratory-adapted *npr-1* locus on chromosome X was defined as a highly pleiotropic gene affecting many different phenotypes [58]. To address these concerns, among others, a second panel of 359 RIAILs (Andersen RIAILs) was generated using CB4856 and QX1430 (a strain with the N2 genetic background but contains a transposon insertion in *peel-1* and an introgression of the CB4856 *npr-1* allele) [68]. This panel was first used in 2015 to map life history traits such as fecundity and body size [68]. Finally, a complementary panel of RILs (Andersen RILs) was recently generated between QX1430 and CB4856 containing almost 600 strains that have been sequenced and could be used for fine mapping QTL [72]. Significance of a genotype-phenotype association using recombinants is calculated as the log of the odds (LOD) ratio using the formula $LOD = log(1 - cor(y, g)2)/2log(10)$ where y is the measured phenotype and g is the genotype information. The Andersen Lab has developed the R package *linkagemapping* (github.com/AndersenLab/linkagemapping) to facilitate detection and localization of QTL in this manner.

Although a lot of work focuses on the genetic variation between the N2 and CB4856 strains, several studies have taken advantage of the genetic variation between other strains in the species. RIL panels have been generated between the N2 and BO strains [87], the N2 and DR1350 strains [88], the N2 and LSJ2 strains [89], the MY10 and JU1395 strains [90], the N2 and MY16 strains [91], the MY14 and CX12311 strains [92], the N2 and AB1 strains [93], and the AB2 and CB4856 strains [94]. Although QTL

Table 1-1: Overview of different N2xCB4856 recombinant panels

Panel Name	Parent strains	Number of strains	Publication	Notes
Kammenga RILs	N2 x CB4856	80	Li *et al.* 2006 [80]	- Many QTL mapped across diverse traits - Genome-wide expression data across different environments
Rockman RIAILs (set 1)	N2 x CB4856	236	Rockman and Kruglyak 2009 [78]	- Genome-wide expression data without environmental perturbations - Many QTL mapped across diverse traits; *peel-1 zeel-1* allele skew on chrI and laboratory adapted allele of *npr-1* on chrX; Highly structured population
Andersen RIAILs (set 2)	N2* (QX1430) x CB4856	359	Andersen *et al.* 2015 [68]	- Many QTL mapped across diverse traits (mostly drug responses) - Transposon insertion in *peel-1* on chrI and wild allele of *npr-1* on chrX
Andersen RILs (set 3)	N2* (QX1430) x CB4856	590	Brady *et al.* 2019 [72]	- No phenotype information yet - Transposon insertion in *peel-1* on chrI and wild allele of *npr-1* on chrX

mapping with one of these RIL panels facilitates detection of QTL that might not be present between the N2 and CB4856 strains, these alternative mapping populations have mostly been generated to answer specific research questions and have yet to be as fully explored as the N2xCB4856 panels. More recently, two multi-parental recombinant panels have been developed to capture more genetic diversity without sacrificing the power of recombinants to detect and localize QTL. The first is a multi-parental experimental evolution (CeMEE) panel generated from crossing 16 diverse wild isolates followed by more than 100 generations of experimental evolution and subsequent inbreeding [95]. The second is a multi-parental RIL panel generated from four parental wild isolates [96]. Recombinant panels that capture a majority of the genetic diversity within the species might provide the most power for QTL mapping, combining the benefits of linkage and association mapping. However, the mapping population (and thus method of QTL mapping) chosen will ultimately depend on the biological question and the quantitative trait.

The generation of all these recombinant panels have been useful for the *C. elegans* quantitative genetics community and have facilitated the discovery of many QTL. However, this method requires a large amount of time, money, and work to generate, cryopreserve, and phenotype such a large

panel of individuals. As an alternative method to phenotyping recombinant individuals, bulk-segregant analysis (BSA) generates a large pool of genetically divergent recombinants that can be genotyped and phenotyped in bulk [97]. Often, pools are exposed to a selective pressure, such as a drug, prior to genotyping. In the absence of selection, allele frequencies are expected to be 50%. However, if a specific allele is associated with drug resistance, this allele will likely be enriched in the selected population. BSA was first used to map QTL in yeast [98] and has been recently adapted to *C. elegans* to study drug resistance, differences in gene expression, and general nematode fitness between the N2 and CB4856 strains [99]. Though fast, powerful, and effective, BSA requires a method for selecting a sub-population based on the phenotypic distribution of the trait and, like all mapping methods, QTL resolution is highly dependent on the nonuniform pattern of genetic recombination [99, 78]

1.3.2 Overview of quantitative traits mapped in *C. elegans*

Some of the most complex phenotypic traits are those related to an organism's life history (reproduction and development) or behavior (response to external stimuli). For this reason, these complex traits are often the most commonly studied in quantitative genetics [87]. Numerous QTL have been detected that underlie life history traits such as reproduction [100, 68, 90, 101, 102, 103, 88, 104, 95, 105, 106], lifespan and aging [104, 96, 107, 35, 108, 109, 110], body size and development [100, 68, 91, 93, 102, 103, 88, 81, 104, 76, 95, 96, 111], and abundance of gene transcripts [112, 113, 80, 108, 114, 115, 116], proteins [117], and metabolites [118]. Behavioral traits studied include pathogen immunity [100, 119, 120, 121], stress responses [68, 122, 72, 69, 73, 123, 120, 75, 124, 74, 71, 70, 88, 96, 125, 116, 126, 99], responses to environmental perturbations such as food [127, 128, 129, 110], oxygen [130], pheromones [92, 76, 89], and temperature [102, 103, 81, 80, 131], and other nematode behaviors [132, 133, 94, 78, 35]. In addition to these traits, genomic features such as telomere length [134] and transposable elements [135] as well as geographical and climate variables [136] have been used as phenotypic traits for QTL mapping. An overview of QTL mapping studies in *C. elegans* can be found in **Table S1-1**.

Toxin responses as quantitative traits. Many human medicines and environmental toxins are designed from compounds naturally secreted by bacteria, fungi, or plants as a mechanism of protection [137]. These compounds usually act to harm a predatory organism by targeting one of several necessary and conserved genetic pathways that control cellular growth [138]. To enable survival, organisms have developed mechanisms of resistance, often in the form of genetic variation in these key molecular pathways [139]. As a result, responses to these compounds are highly variable between individuals based on a variety of both genetic and environmental factors. Nematode behavior, development, and reproductive ability in response to a number of toxic conditions have been extensively studied (**Figure 1-5**). Although these traits are indirect measures of the toxin's effect, these phenotypes have been used to identify numerous QTL [73, 74, 68, 123, 122, 99], causal genes [69, 72], and causal variants [70, 71, 75, 120, 123, 124].

Several of these studies identified causal genes that function in biologically relevant pathways or are known drug targets, suggesting that these QTL studies are useful for learning about the biology of toxin responses [70, 71, 75, 120]. Additionally, studying toxin responses in *C. elegans* might have broader applications in human and veterinary medicine. In one example, Zdraljevic *et al.* (2017) uses *C. elegans* to identify a single residue (Q797M) in the topoisomerase II enzyme TOP-2 that contributes to sensitivity to the double-strand-break-inducing chemotherapeutic drug etoposide. The authors further show that editing the methionine residue in the human ortholog (hTOPIIα) to a glutamine significantly increased etoposide resistance in 293T cells, showing that this residue has the same effect in *C. elegans* and human cells in culture. Although this residue does not vary naturally within the human population, it does differ between the two human isoforms (hTOPIIβ naturally has the glutamine residue), suggesting that variation at this residue contributes to the differential binding of etoposide between the two isoforms hTOPIIα and hTOPIIβ [71]. In another example, Ghosh *et al.* (2012) identified a four-amino acid deletion in the glutamate-gated chloride channel GLC-1 that contributed to variation in response to the anthelmintic compound abamectin in *C. elegans* [120]. Glutamate-gated chloride channels are the known drug targets of abamectin [140, 141] and this study demonstrated that quantitative genetics in *C. elegans* can discover alleles that might also confer anthelmintic resistance in parasitic populations.

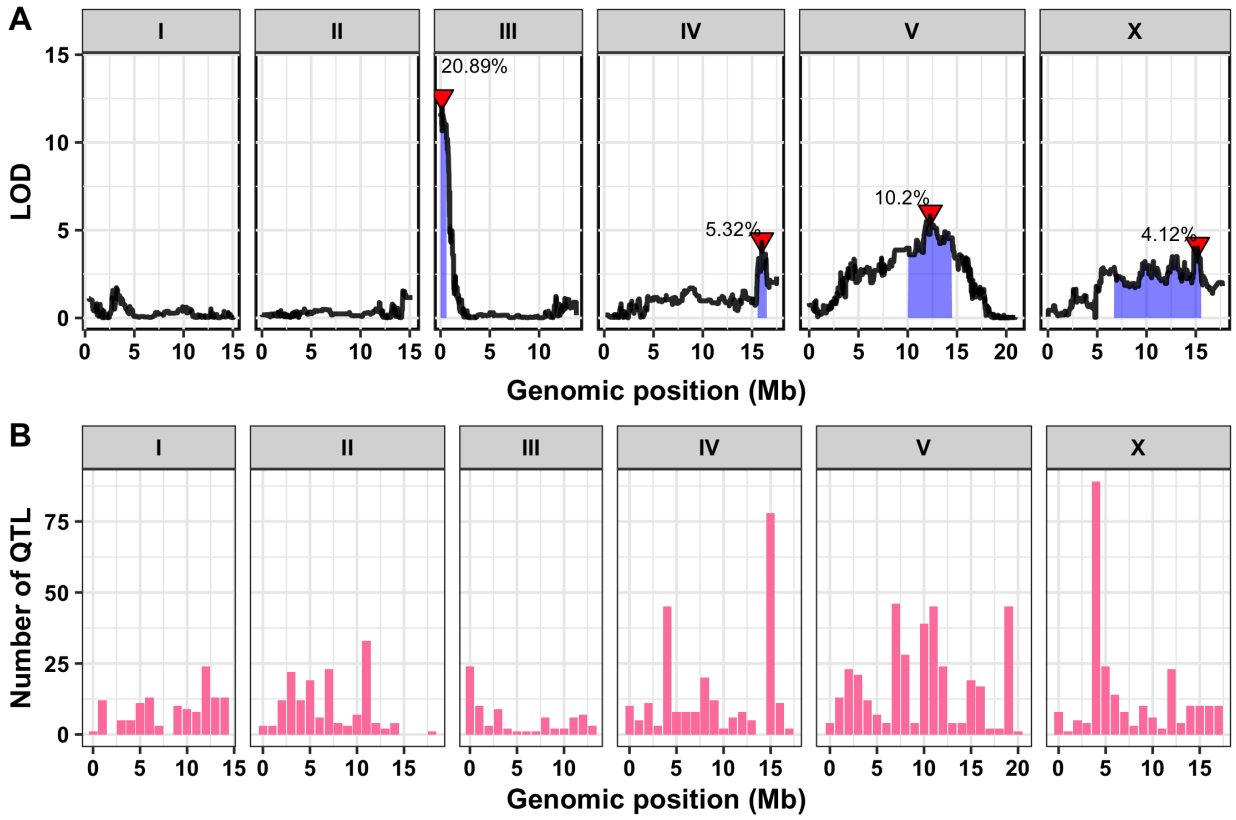


Figure 1-5: Toxin-response QTL mapping studies. (A) An example of a toxin-response QTL mapping (linkage mapping analysis for nematode zinc response [123]). Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis). Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** A histogram showing the distribution (x-axis) and number of toxin-response QTL identified (y-axis) with a bin size of 1 Mb.

Gene expression as a quantitative trait. In addition to organismal phenotypes, gene expression data has been collected across three panels of recombinants in *C. elegans* and used as a phenotypic trait for linkage mapping [80, 108, 115, 69, 113, 116, 109]. QTL mapping of genome-wide gene expression has several goals. First, the analysis of thousands of random, non-correlated traits allows for a generalization of how genetic variation influences complex traits in an unbiased manner [22]. Second, the mapping of intermediate phenotypes such as gene expression often gives much higher resolution compared to organismal phenotypes, making it potentially easier to identify the causal variant [142]. Finally, expression variation, especially in combination with variation in other phenotypes, can facilitate connections between genetic variation and the biology of these traits [143].

More than ten percent of genes in the *C. elegans* genome (2,000+ genes) vary in expression between the N2 and CB4856 strains, and variation of most of the genes is correlated with specific genetic loci [115, 108, 116]. A majority of these expression QTL (eQTL) are defined as *cis*- or local eQTL, meaning the genetic variant that controls gene expression is located near the gene in question (often within 1 Mb) [115]. Interestingly, the genes with local eQTL often overlap between studies, providing more evidence that expression of these genes is genetically determined [113]. The rest of the eQTL that are not mapped locally are classified as *trans*- or distant eQTL. Strikingly, a large fraction of all distant eQTL fall within one of several *trans*-bands or eQTL hotspots, suggesting variation at one pleiotropic locus or several tightly linked loci controls variation in expression of tens or hundreds of genes. However, the location of these eQTL hotspots seem to differ between studies suggesting that they are environmentally controlled. In one powerful example, variation in the neuropeptide receptor gene *npr-1* controls variation in expression of 247 genes across the genome [115, 100]. Furthermore, this variation in *npr-1* has also been associated with numerous behavioral and developmental phenotypes [58]. Because the 247 genes that change expression in response to variation in *npr-1* are enriched for neuropeptide signaling and growth [100], it is possible that expression of these genes together work to produce several pleiotropic phenotypes. Alternatively, it is possible that the pleiotropic variant in *npr-1* alters expression of 247 genes, each acting independently to produce different phenotypes. Either way, this study highlights how expression variation can help bridge the gap between genomic variation and trait biology.

Several expression QTL studies have been performed in *C. elegans* under a variety of different environments to study how expression variation changes between environmental conditions. Li *et al.* (2006) was the first to show that eQTL significantly vary between environments. The authors found hundreds of strain-specific or temperature-specific eQTL but identified significantly less plasticity QTL, defined by the interaction between expression and temperature [80]. The same RILs have been used to show that variation in alternative splicing for only a few genes is controlled by genetic variation [114] and that the integrity of the gene expression network declines with age [108, 109]. More recently, variation in gene expression was studied in combination with the heat shock response where the authors again reiterate that eQTL hotspots are environment-specific while local eQTL are largely

maintained across diverse environments [116]. And finally, a panel of recombinants between MT2124 (strain with the N2 genetic background and a *let-60(gf)* allele) and CB4856 identified eQTL associated with RAS/MAPK signaling [113]. These studies have all provided insight into the genetics of expression variation. However, all eQTL studies to date have been performed with microarray analysis, which is biased towards the N2 reference genome. RNA sequencing would decrease this bias, also facilitating eQTL studies with other RIL or wild isolate panels.

1.4 Lessons from quantitative genetics in *C. elegans*

Over the last decade, QTL mapping in *C. elegans* has worked to address several unanswered questions in quantitative genetics relating to the genetic architecture of complex traits and the evolution of diverse phenotypes. Although many early quantitative genetics studies in *C. elegans* identified mostly single, large-effect loci, advancements in genotyping and phenotyping technologies as well as the collection of more genetically distinct wild isolates have revolutionized the field. Many quantitative traits map to at least two independent loci and some traits have five or more QTL. A study of natural expression variation of over 15,000 transcripts discovered that expression of most genes are defined by a single large-effect locus with only 14% of differentially expressed genes controlled by two or more loci [115, 69]. Alternatively, a large QTL mapping study of nematode responses to 16 diverse toxins identified 82 QTL from 47 non-correlated traits, and more than one third of these traits mapped to two or more loci [73]. Strikingly, 82% of these QTL had small effect sizes, explaining less than 10% of the phenotypic variation in the RIAIL panel. Using NILs, the authors were able to empirically validate a QTL that explained only 7% of the phenotypic variation, demonstrating that small effects can be studied in *C. elegans* with the right tools and a sensitive assay [73].

Most QTL identified appear to have additive effects [123, 122, 73] consistent with what is observed in other highly powered model organisms like yeast and plants [23, 24, 25, 26, 27, 28]. However, empirical evidence of genetic interactions present in only one genetic background show that there is not sufficient power in the RIAIL panels to detect these uni-directional epistatic loci [72, 73]. In fact, several studies have reported that a single detected locus actually comprises multiple tightly linked loci that either act additively or interact to produce the overall phenotypic effect observed [122, 123, 128, 144, 145, 131].

In one thorough example, Bernstein *et al.* (2019) identified a QTL on chromosome X that underlies responses to the heavy metal nickel. The genomic region was divided into 15 intervals of ∼100 kb each using a panel of 16 NILs with increasingly smaller introgressions across the region. The authors discovered that nine of the 15 intervals had a significant effect on the phenotype. However, the effect size and direction of effect varied between intervals suggesting multiple, tightly linked, antagonistic QTL [122]. Given the level of recombination in the current mapping panels, we do not have the statistical power to distinguish tightly linked loci during QTL detection. Regardless, these examples demonstrate that the genetic architecture of complex traits is often even more complex than it appears. Furthermore, these examples provide evidence that the quantitative genetics community is still best suited to identify large-effect Mendelian-like QTN, leaving a biased gap in our understanding of quantitative variation.

1.4.1 Benefits of combining linkage and association mapping

Linkage and association mapping are two complementary approaches to discover causal genetic variation underlying phenotypic traits. Each method has different strengths and weaknesses that might lend one tool to be more powerful than the other for different types of causal variants. Arguably, the best way to identify the QTN would be to combine the strengths of both linkage and association mapping by phenotyping both a panel of recombinants and a panel of wild isolates. If the result is that different QTL are found by the different mapping techniques, this in itself offers more insight into the quantitative variation underlying each QTL than mapping with either recombinants or wild isolates alone. For QTL identified with linkage mapping but not association mapping, it can be concluded that the allele frequency of the causal variant must be rare in the population because association mapping does not have the power to detect rare variants. In this case, the list of potential causal variants present between the parental strains of the recombinant lines can be filtered to remove variants that are also shared among the population. Conversely, for QTL identified with association mapping but not linkage mapping, it can be concluded that although the allele frequency of the causal variant must be relatively common in the population, this causal variant is not present in the parental strains of the recombinants. In this case, the list of potential causal variants can be filtered to remove all variants present in the recombinant strains.

If, however, the result is that the same QTL is identified by both mapping techniques, the combined resources of both techniques can be leveraged to quickly identify the QTN. Under the assumption that the same variant could be causal in both populations, it can be concluded that the causal variant is present in the parental strains of the recombinants and commonly found across the population. In this best-case scenario, the list of potential causal variants can be filtered to remove all common variants that are not present in the recombinant strains. Depending on the extent of genetic diversity between the parental strains, this should significantly decrease the number of potential causal variants. In two examples, this approach narrowed the potential causal variants from thousands to only three [71] or even one single variant [70]. In both cases, the authors were able to conclusively prove that the highlighted variant was causal without the extensive use of quantitative narrowing approaches like NILs and were able to describe the mechanism by which these variants contributed to toxin responses. Although few QTL might fall into this last category, the benefits of this combined approach will likely soon outweigh the extra time and resources needed to phenotype an extra panel of strains. Alternatively, multi-parent recombinant populations like the CeMEE panel were designed to combine the benefits of linkage and association mapping without having to phenotype two different populations [95] and could be leveraged in the future.

1.4.2 Mediation analysis: combining expression and phenotypic QTL to identify causal genes

We are no longer limited to studying large-effect protein-coding genetic variants. Several eQTL studies have discovered thousands of differentially expressed genes that are largely controlled by genetic factors [112, 113, 80, 108, 114, 115, 116]. Co-localization of eQTL and phenotype QTL could suggest that a single genetic variant underlies both QTL. Furthermore, mediation analysis now allows scientists to make statistical connections between genetic variation, variation in gene expression, and variation in other phenotypes. In addition to providing a resource for candidate gene prioritization within a QTL interval, mediation analysis can help to identify the mechanism by which genetic variation causes phenotypic differences. This technique is especially powerful in identifying causal genes distant from their QTL that would otherwise not have been discovered. For example, in Chapter 3, I used mediation analysis to implicate the gene *scb-1* in responses to several double-strand break

chemotherapeutics. Functional analysis with deletions of *scb-1* using CRISPR-Cas9 genome editing further showed that *scb-1* is a pleiotropic gene that influences responses to other chemotherapeutic drugs. This study paired genome-wide gene expression data with drug-response phenotypes to identify the causal gene and provide further evidence for the unknown function of *scb-1* [69].

1.4.3 QTN that led to a better understanding of the evolution of quantitative variation

Functional analysis has facilitated the discovery of many QTN from QTL (**Table S1-1**) and, in combination with mechanistic studies, has provided numerous insights into the evolution of complex traits in *C. elegans*. However, drawing evolutionary conclusions from these studies should be done with caution. Many QTN have been shown to be the direct result of laboratory adaptation of the N2 strain and thus do not represent natural variation within the species [58, 76, 104, 93, 130]. Furthermore, Rockman *et al.* (2010) show that the patterning of eQTL across the genome is best explained by a model of background selection. This result suggests that these QTL might not be the product of adaptive processes, rather the chromosomal location of a gene might affect its evolutionary potential [115].

Nevertheless, the discovery of numerous QTN have led to several important evolutionary conclusions. For example, the insertion of a retrotransposon in a mucin-like gene *plg-1* results in copulatory plugging after mating and is thought to represent a type of sexual selection based on the natural male frequencies in the wild [84, 87]. Additionally, natural variation in the beta-tubulin gene *ben-1* suggests that variation in this gene arose multiple times during the evolutionary history of *C. elegans*, likely in recent years in response to local purifying selective pressures [75]. Conversely, signatures of balancing selection have been implicated in several traits including density-dependent foraging behaviour at the *srx-43* locus [92], embryonic lethality at the *peel-1/zeel-1* [85] and *sup-35/pha-1* incompatibility loci [146], avermectin resistance at the *glc-1* locus [120], and pheromone response at the *srg-37* locus [76]. In conclusion, these results, and others, continue to provide strong evidence that the QTN program in *C. elegans* can identify the alleles that matter for evolution.

1.5 Dissertation overview

This dissertation aims to answer several of the important questions in quantitative genetics outlined above and further aims to discover the QTN underlying complex traits, particularly toxin responses. Chapter 2 details the largest QTL study for phenotypic traits in *C. elegans* to date, showing that most traits have a complex genetic architecture and that additive loci might not explain all the phenotypic variation for toxin responses. Chapter 3 takes advantage of mediation analysis, a technique never before used in *C. elegans*, to better prioritize candidate genes with expression variation in addition to candidate genes with protein-coding genetic variation. Mediation analysis is further leveraged in Chapter 4 to predict a top candidate gene underlying the zinc response and uses functional validation with NILs and CRISPR-Cas9 genome editing to provide new evidence for the function of the sequestosome-related gene, *sqst-5*, as a potential negative regulator of the nematode zinc response. The importance of minimizing environmental variation in quantitative genetics is highlighted in Chapter 5, which suggests that food intake can greatly affect toxin responses. Chapter 6 identifies two novel QTL for avermectin resistance that might be conserved in the parasitic nematode *Haemonchus contortus*, demonstrating the relevance and benefits of studying anthelmintic resistance in the free-living nematode. Chapter 7 correlates genetic variation with environmental variables such as temperature and humidity with the goal of discovering genetic adaptation to specific niches. Finally, Chapter 8 discusses several key challenges for the QTN program in *C. elegans* and offers commentary and advice for the future generation.

1.6 SUPPLEMENTAL TABLES

Table S1-1: List of QTL mapping studies in *C. elegans*

Reference	Experiment	Strains	Phenotype	QTG/QTN
Andersen *et al.* (2014) [100]	RIAILs	N2 x CB4856 (Rockman set)	Lifetime fecundity, body size, susceptibility to S. aureus	NPR-1 (V215F)
Andersen *et al.* (2015) [68]	RIAILs	N2*(QX1430) x CB4856 (Andersen set)	Growth and development; drug response (paraquat)	*NA*
Balla *et al.* (2015) [119]	RIAILs	N2 x CB4856 (Rockman set)	Immunity to microsporidian parasite	*NA*
Bendesky *et al.* (2011) [147]	RIAILs	N2 x CB4856 (Rockman set)	Food foraging responses	*tyra-3*
Bendesky *et al.* (2012) [132]	RIAILs	N2 x CB4856 (Rockman set)	Aggregation/social behavior and bordering	*exp-1*
Bernstein *et al.* (2019) [122]	RIAILs	N2*(QX1430) x CB4856 (Andersen set)	Drug response (nickel)	*NA*
Brady *et al.* (2019) [72]	RIAILs; Association mapping	N2 x CB4856 (Rockman set); N2*(QX1430) x CB4856 (Andersen set); 96 wild isolates	Drug response (bleomycin); gene expression	*scb-1(lf)*
Burga *et al.* (2019) [99]	Bulk-segregant	N2* x CB4856* (*fog-2*)	Drug response (abamectin), stress resistance (H2O2), fitness, gene expression differences,	*NA*
Chandler *et al.* (2010) [148]	RILs	N2* x CB4856* (temperature-sensitive mutants)	Sex determination	*NA*
Cleuren *et al.* (2019) [91]	Association mapping; RILs	96 wild isolates; N2 x MY16	Nrf/Wnt signaling, endoderm development	*NA*
Cook *et al.* (2016) [134]	Association mapping	152 wild isolates	Telomere length	*pot-2*
Doroszuk *et al.* (2009) [35]	RILs	N2 x CB4856 (Kammenga set)	Lifespan, pumping	*NA*
Duveau *et al.* (2012) [93]	RILs	N2*(JU605) x AB1*(JU606)	Vulval induction	*nath-10*
Elvin *et al.* (2011) [126]	RIAILs	N2 x CB4856 (Rockman set)	RNAi response	*ppw-1*
Evans and Andersen (2020) [69]	RIAILs	N2 x CB4856 (Rockman set); N2*(QX1430) x CB4856 (Andersen set)	Drug responses (amsacrine, bleomycin, bortezomib, carmustine, cisplatin, etoposide, puromycin, silver); gene expression	*scb-1(lf)*
Evans and Brady *et al.* (2018) [73]	RIAILs	N2*(QX1430) x CB4856 (Andersen set)	Drug responses (cadmium, carmustine, chlorothalonil, chlorpyrifos, cisplatin, copper, diquat, fluoxetine, FUdR, irinotecan, mechlorethamine, paraquat, silver, topotecan, tunicamycin, vincristine)	*NA*
Evans *et al.* (2017) [136]	Association mapping	152 wild isolates	Geography and climate	*NA*
Evans *et al.* (2020) [123]	RIAILs; Association mapping	N2 x CB4856 (Rockman set); N2*(QX1430) x CB4856 (Andersen set); 96 wild isolates	Drug response (zinc)	*sqst-5(lf)*

Frezal *et al.* (2018) [90]	RILs	MY10 x JU1395	Temperature-induced sterility	*set-24(lf)*
Gaertner *et al.* (2012) [131]	RIAILs	N2 x CB4856 (Rockman set)	Thermal preference behavior	*NA*
Gao and Sterken *et al.* (2018) [118]	RILs	N2 x CB4856 (Kammenga set)	Metabolite levels	*NA*
Ghosh *et al.* (2012) [120]	RIAILs	N2 x CB4856 (Rockman set)	Drug response (abamectin), pathogen immunity	*glc-1(lf)*
Gimond *et al.* (2019) [101]	Association mapping; RILs	97 wild isolates	Sperm size	*NA*
Glater *et al.* (2014) [128]	RIAILs	N2 x CB4856 (Rockman set)	Olfactory preference to bacteria	*NA*
Greene *et al.* (2016) [92]	RILs	MY14 x CX12311	Pheromone sensitivity	*srx-43*
Gutteling *et al.* (2007a) [103]	RILs	N2 x CB4856 (Kammenga set)	Age at maturity, fertility, egg size, and growth rate in response to temperature	*NA*
Gutteling *et al.* (2007b) [102]	RILs	N2 x CB4856 (Kammenga set)	Egg size, egg number, and body mass in response to temperature	*NA*
Hahnel and Zdraljevic *et al.* (2018) [75]	Association mapping	209 wild isolates	Drug response (albendazole)	*ben-1(lf)*
Harvey *et al.* (2008) [88]	RILs	N2 x DR1350	Dauer development, reproduction in response to stress	*NA*
Harvey *et al.* (2009) [88]	RILs	N2 x DR1350	Dispersal between food patches	*NA*
Kammenga *et al.* (2007) [81]	RILs	N2 x CB4856 (Kammenga set)	Body size and temperature	TRA-3 (F96L)
Large *et al.* (2016) [104]	RILs	N2*(CX12311) x LSJ2	Reproductive timing, lifespan, dauer formation, growth rate, and offspring number	*nurf-1(lf)*
Large *et al.* (2017) [106]	RILs	N2*(CX12311) x LSJ2	Egg laying	*NA*
Laricchia *et al.* (2017) [135]	Association mapping	152 wild isolates	Transposable elements	*NA*
Lee *et al.* (2017) [133]	RIAILs	N2 x CB4856 (Rockman set)	Phoretic behavior (nictation)	*prg-1 (piRNA)*
Lee *et al.* (2019) [76]	Association mapping	157 wild isolates	Pheromone response (dauer)	*srg-37(lf)*
Li *et al.* (2006) [80]	RILs	N2 x CB4856 (Kammenga set)	Expression, temperature sensitivity	*NA*
Li *et al.* (2010) [114]	RILs	N2 x CB4856 (Kammenga set)	Expression, alternative splicing	*NA*
McGrath *et al.* (2009) [130]	RIAILs	N2 x CB4856 (Rockman set)	Oxygen sensing and response	NPR-1 (V215F), glb-5
McGrath *et al.* (2011) [89]	RILs	N2 x LSJ2	Dauer pheromone resistance	*srg-36 and srg-37*
NA et al. (2020) [124]	Association mapping	133 wild isolates	Drug response (proprionate)	*glct-3*
Nakad *et al.* (2016) [121]	RILs	N2 x CB4856 (Kammenga set)	Immune defenses	NPR-1 (V215F)
Noble *et al.* (2015) [94]	RILs	QG5 x QX1199*	Male-male plugging behavior	*plep-1*
Noble *et al.* (2017) [95]	CeMEE	16 wild isolates	Fertility and body size	*NA*

Rockman and Kruglyak (2009) [78]	RIAILs, Association mapping	N2 x CB4856 (Rockman set), 41 wild isolates	Copulatory plugging, embryonic lethality	*plg-1*
Rockman *et al.* (2010) [115]	RIAILs	N2 x CB4856 (Rockman set)	Expression	*NA*
Rodriguez *et al.* (2012) [125]	RILs	N2 x CB4856 (Kammenga set)	Heat shock recovery	*NA*
Schmid *et al.* (2015) [112]	RILs	N2*(MT2124) x CB4856	RAS/MAPK signaling output	*amx-2*
Seidel *et al.* (2008) [85]	RIAILs	N2 x CB4856 (Rockman set)	Embryonic lethality	*peel-1;zeel-1*
Singh *et al.* (2016) [117]	RILs	N2 x CB4856 (Kammenga set)	Protein abundance	*NA*
Snoek and Orbidans *et al.* (2014) [149]	RILs	N2 x CB4856 (Kammenga set)	Egg lay defective incompatibility	*NA*
Snoek and Sterken *et al.* (2017) [116]	RILs	N2 x CB4856 (Kammenga set)	Expression, temperature	*NA*
Snoek *et al.* (2019) [96]	mpRILs	JU1511 x JU1926 x JU1931 x JU1941	Lifespan, stress resistance, developmental speed, and population growth in different environments	*NA*
Stast*NA et al.* (2015) [107]	RILs	N2 x CB4856 (Kammenga set)	Diet and lifespan	*NA*
Sterken *et al.* (2017) [113]	RILs	N2*(MT2124) x CB4856*	Expression, ras/mapk modifier	*amx-2*
Vinuela and Snoek *et al.* (2010) [108]	RILs	N2 x CB4856 (Kammenga set)	Aging and expression	*NA*
Vinuela *et al.* (2012) [109]	RILs	N2 x CB4856 (Kammenga set)	Aging and expression	*NA*
Webster *et al.* (2019) [111]	Association mapping	96 wild isolates	Starvation resistance	*NA*
Zamanian *et al.* (2018) [74]	RIAILs	N2*(QX1430) x CB4856 (Andersen set)	Drug responses (albendazole, fenbendazole, mebendazole, thiabendazole)	*NA*
Zdraljevic *et al.* (2017) [71]	RIAILs, Association mapping	N2*(QX1430) x CB4856 (Andersen set), 96 wild isolates	Drug response (etoposide and amsacrine)	TOP-2 (Q797M)
Zdraljevic *et al.* (2019) [70]	RIAILs, Association mapping	N2*(QX1430) x CB4856 (Andersen set), 96 wild isolates	Drug response (arsenic trioxide)	DBT-1 (C78S)
Zhu *et al.* (2015) [110]	RIAILs	N2 x CB4856 (Rockman set)	Mito-nuclear compatibility, lifespan, foraging	*NA*
Zhu *et al.* (2019) [105]	RIAILs	N2 x CB4856 (Rockman set)	Mito-nuclear fecundity	*NA*

2 Shared genomic regions underlie natural variation in diverse toxin responses

2.1 PREFACE

When I first joined the lab in the summer of 2016, we had a large collection of QTL from which I could choose one (or more) to follow up on in more detail. Throughout that summer and fall, I began testing a variety of these toxins to see if there was any "low-hanging fruit." About the same time, Dr. Shannon Brady was beginning a project to analyze a collection of QTL that we may never get the chance to investigate deeply as a lab. We decided to combine our efforts to perform a broad analysis of toxin-response QTL in *C. elegans*, compared with the "normal" Andersen Lab approach of identifying the genetic mechanism of a single toxin. This time holds some of the craziest days of my graduate school career. Running 36 plates (approximately 15 hours) on the sorter (a record high we still hold to this day!), wearing onesies to lab to celebrate the end of the NIL phenotyping, and spending endless hours thinking about genetic interactions, designing flowcharts to categorize traits, trying to understand what the Hadamard product is, and writing code and analyzing data side-by-side in Cook 3118. Our two years of hard work finally paid off with a co-first author manuscript in *Genetics* in 2018 [73] from which this chapter is based.

2.2 ABSTRACT

Phenotypic complexity is caused by the contributions of environmental factors and multiple genetic loci, interacting or acting independently. Studies of yeast and *Arabidopsis* often find that the majority of natural variation across phenotypes is attributable to independent additive quantitative trait loci (QTL). Detected loci in these organisms explain most of the estimated heritable variation. By contrast, many heritable components underlying phenotypic variation in metazoan models remain undetected. Before the relative impacts of additive and interactive variance components on metazoan phenotypic variation can be dissected, high replication and precise phenotypic measurements are required to obtain sufficient statistical power to detect loci contributing to this missing heritability. Here, we used a panel

of 296 recombinant inbred advanced intercross lines of *Caenorhabditis elegans* and a high-throughput fitness assay to detect loci underlying responses to 16 different toxins, including heavy metals, chemotherapeutic drugs, pesticides, and neuropharmaceuticals. Using linkage mapping, we identified 82 QTL that underlie variation in responses to these toxins and predicted the relative contributions of additive loci and genetic interactions across various growth parameters. Additionally, we identified three genomic regions that impact responses to multiple classes of toxins. These QTL hotspots could represent common factors impacting toxin responses. We went further to generate near-isogenic lines and chromosome-substitution strains and then experimentally validated these QTL hotspots, implicating additive and interactive loci that underlie toxin-response variation.

2.3 INTRODUCTION

Rapid advances in sequencing technologies enabled the collection of high-quality genomic datasets for many species [150]. These data, paired with a broad range of high-throughput phenotypic assays, made quantitative genetics a powerful tool in biology. Linkage mapping has been used to identify quantitative trait loci (QTL), leading to profound impacts on human health [151, 152, 153], agriculture and livestock [154, 155, 156, 157], and basic biology [158, 159, 68]. Despite the growing number of detected QTL across numerous traits, these QTL often do not explain the complete heritable component of trait variation [22]. This missing heritability can be attributed to undetected small-effect additive loci and/or interactions between QTL [28]. Although some studies contend that epistatic effects among QTL might explain missing heritability [29, 30, 31, 32, 33], others argue that missing heritability comprises small-effect additive loci that remain undetected in cases where statistical power is too low [23, 24, 25, 26]. Quantitative geneticists have leveraged large numbers of recombinant strains in both yeast and *Arabidopsis* to overcome power limitations and concluded that, when power is sufficient, small-effect additive components can be identified that account for nearly all of the heritability of a given trait [28, 27, 160]. We require a metazoan system with high statistical power to determine whether this predominantly additive-QTL model remains broadly applicable in animals.

One such tractable metazoan is the roundworm nematode *Caenorhabditis elegans.* The genetic variation among a panel of recombinant inbred advanced intercross lines (RIAILs) generated between

the N2 and CB4856 strains of *C. elegans* [78, 68] has been leveraged in many linkage mapping analyses [130, 147, 133, 132, 117, 112, 119, 71, 74, 100, 147, 108, 35, 149, 125, 81, 103, 80, 102, 128, 161, 115, 86, 85]. Additionally, a high-throughput phenotyping platform to rapidly and accurately measure animal fitness could provide the replication and precision required to detect small-effect additive loci and to determine the relative contributions of additive and/or epistatic loci to trait variation [100, 71]. Notably, the combination of this panel and phenotyping platform have facilitated linkage mappings of multiple distinct fitness parameters, resulting in the detection of a single QTL, in fact a single quantitative trait gene (QTG), that underlies several fitness-related traits [100, 71]. This example of pleiotropy suggests that large-scale studies could reveal additional pleiotropic effects.

Such large-scale studies have implicated pleiotropic QTL that impact the expression of a broad range of genes [162, 163, 115, 164]. Variation in the master regulators that are within these expression QTL hotspots have downstream effects on the transcription of many genes. Similarly, other QTL hotspots could impact multiple traits, such as responses to various conditions. In yeast, most chemical-response QTL are thought to be unique to one or a few conditions, whereas few QTL have been found to have pleiotropic effects across many conditions [165, 166, 167]. Although QTL underlying responses to individual conditions have been identified across multiple animal models [168, 169, 170, 171, 172], the existence of QTL hotspots that influence multiple condition responses has yet to be observed broadly in metazoans.

Here, we performed a set of linkage-mapping experiments with a large panel of recombinant lines to identify QTL implicated in responses to 16 different toxins and found three QTL hotspots that underlie many of these responses. We demonstrated how high replication in the high-throughput fitness assay can enable the identification and validation of QTL, even in cases of small phenotypic effects. Additionally, we analyzed relative contributions of additive and epistatic genetic loci in various toxin responses. Finally, we discovered evidence for interactions between loci of the N2 and CB4856 strains that impact several toxin responses and could suggest how large regions of the genome were swept across the species.

2.4 METHODS

2.4.1 Strains

Animals were grown at 20ºC using OP50 bacteria spotted on modified nematode growth medium (NGMA), containing 1% agar and 0.7% agarose to prevent animals from burrowing. For each assay, strains were propagated for five generations after starvation to reduce transgenerational effects of starvation [100]. Recombinant inbred advanced intercross lines (RIAILs) used for linkage mapping were constructed previously [68]. The construction of near-isogenic lines (NILs) and chromosome substitution strains (CSSs) is detailed below. Strains are available upon request.

2.4.2 High-throughput toxin response assay

We used a modified version of the high-throughput fitness assay described previously [71]. Populations of each strain were passaged for four generations, amplified, and bleach-synchronized. Approximately 25 embryos from each strain were then aliquoted to 96-well microtiter plates at a final volume of 50 μL of K medium [173]. Embryos hatched overnight and arrested in the L1 larval stage. The following day, arrested L1 animals were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA; [174]) at a final concentration of 5 mg/mL in K medium and were grown to the L4 larval stage for 48 hours at 20ºC with constant shaking. Three L4 larvae were then sorted using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica, Holliston, MA) into microtiter plates that contained HB101 lysate at 10 mg/mL, K medium, 50 μM kanamycin, and either diluent (1% DMSO or 1% water) or diluent and a toxin of interest. The sorted animals were then grown for 96 hours at 20ºC with constant shaking. During this time, the sorted animals matured to adulthood and laid embryos, yielding a population of parent and progeny in each microtiter well. Prior to the measurement of fitness parameters from the populations, animals were treated with sodium azide (50 mM in M9) to straighten their bodies for more accurate growth-response parameter measurements. Traits that were measured by the BIOSORT include brood size (n), animal length (time of flight, TOF), and optical density (extinction time, EXT).

2.4.3 Toxin-response trait calculations

Phenotypic measurements collected by the BIOSORT were processed using the R package *easysorter*, which was specifically developed for processing this type of data set [77]. Briefly, the function *read_data* imported raw phenotypic data then identified and eliminated bubbles. Next, the *remove_contamination* function discarded wells that contained bacterial or fungal contamination (determined by visual inspection) prior to analyzing population parameters. The *sumplate* function then calculated normalized measurements and summary statistics of the assayed traits for the population of animals in each well. The number of animals in each well was divided by the number of animals sorted into that well, yielding a normalized brood size (norm.n). Additionally, optical density (EXT) of each animal was divided by animal length (TOF), resulting in a normalized optical density (norm.EXT) for each animal in each well. The norm.EXT measurement represents the optical density without conflating variation in body length. The summary statistics calculated for each population include 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles, mean, and median measurements of TOF, EXT, and norm.EXT as well as variance for TOF and EXT. Previously, each of these summary statistics has been shown to reveal distinct genetic architectures underlying trait variation, suggesting values to demonstrate the range of biological phenomena that can be captured using this platform [68]. In total, this analysis resulted in 24 phenotypic measurements for each condition tested. When strains were measured across multiple assay days, the *regress(assay=TRUE)* function was used to fit a linear model with the formula (*phenotype ~ assay*) to account for differences among assays. Next, outliers were removed by eliminating phenotypic values that were outside two standard deviations of the mean (unless at least 5% of the strains were outside this range in the case of RIAIL assays). Finally, toxin-specific effects were calculated using the *regress(assay=FALSE)* function from *easysorter*, which fits a linear model with the formula (*phenotype ~ control phenotype*) to generate residual phenotypic values that account for differences between populations that were present in control conditions. For this reason, strain phenotypes in control conditions can influence regressed toxin effects and trait categorizations (below).

2.4.4 Dose-response assays

For each toxin, a dose-response experiment was performed using quadruplicates of four genetically diverged strains (N2, CB4856, DL238, and JU258). Animals were assayed using the high-throughput fitness assay, and toxin-response trait calculations were performed as described above. The concentration of each toxin that provided a highly reproducible toxin-specific effect with variation between N2 and CB4856 across three distinct traits (brood size - norm.n, mean length - mean.TOF, and mean optical density - mean.norm.EXT) was selected for linkage mapping experiments. The chosen concentrations and diluents of each toxin are as follows: cadmium 100 μM in water, carmustine 250 μM in DMSO, chlorothalonil 250 μM in DMSO, chlorpyrifos 1 μM in DMSO, cisplatin 250 μM in water, copper 250 μM in water, diquat 250 μM in water, fluoxetine 250 μM in DMSO, FUdR 50 μM in water, irinotecan 125 μM in DMSO, mechlorethamine 200 μM in DMSO, paraquat 500 μM in water (50 μM was used for the CSS and NIL assays), silver 150 μM in water, topotecan 400 μM in water, tunicamycin 10 μM in DMSO, and vincristine 80 μM in water. The concentration of paraquat differs the concentration used previously [68], suggesting why the genetic architectures are different between the two studies. Toxins assayed in this manuscript were purchased from Fluka (chlorothalonil, #36791-250MG; chlorpyrifos, #45395-250MG; diquat dibromide monohydrate, #45422-250MG-R), Sigma (vincristine sulfate salt, #V8879-25MG; cisplatin, #479306-1G; silver nitrate, #209139; carmustine, #1096724-75MG; topotecan hydrochloride, #1672257-350MG), Calbiochem (tunicamycin, #654380), Aldrich (mechlorethamine hydrochloride, #122564-5G, cadmium chloride #01906BX), Alfa Aesar (irinotecan hydrochloride trihydrate, #AAJ62370-MD), Bioworld (5-fluoro-2'-deoxyuridine, #50256011), Enzo Life Sciences (fluoxetine, #89160-860), Mallinckrodt (cupric sulfate, #4844KBCK), and Chem Service (paraquat, #ps-366).

2.4.5 Principal Component Analysis of RIAILs

A total of 296 RIAILs were assayed in the high-throughput fitness assay described previously in the presence of each toxin listed above as well as control conditions (water or DMSO). Because some of the 24 population parameters measured by the BIOSORT are highly correlated, a principal component

analysis (PCA) was performed. For each growth-response trait, RIAIL phenotypic measurements were scaled to have a mean of 0 and a standard deviation of 1. The *princomp* function within the *stats* package in R [175] was used to run a principal component analysis for each toxin. For each toxin, the minimum number of principal components (PCs) that explained at least 90% of the total phenotypic variance in the RIAILs was mapped through linkage mapping. A total of 97 PCs were mapped.

2.4.6 Linkage mapping

Linkage mapping was performed on each of the 97 PCs (described above) using the R package *linkagemapping* (www.github.com/AndersenLab/linkagemapping). The genotypic data and residual phenotypic data were merged using the *merge_pheno* function. Quantitative trait loci (QTL) were detected using the *fsearch* function, which scaled phenotypes to have a mean of zero and variance of one, then calculated logarithm of odds (LOD) scores for each marker and each trait as $-n(ln(1-R^2)/2ln(10))$, where r is the Pearson correlation coefficient between RIAIL genotypes at the marker and trait values [27]. We note that this scaling of the data did not impact mappings in that scaled mappings and unscaled mappings were identical. The phenotypic values of each RIAIL were then permuted randomly while maintaining correlation structure among phenotypes 1000 times to calculate a significance threshold based on a genome-wide error rate of 5%. This threshold was set for each mapped PC independently to avoid biases introduced by performing large numbers of mappings. The marker with the highest LOD score was then set as a cofactor and mapping repeated iteratively until no significant QTL were detected. Finally, the *annotate_lods* function was used to calculate the fraction of variation in RIAIL phenotypes explained by each QTL. 95% confidence intervals were defined by markers within a 1.5-LOD drop from the marker with the maximum LOD score. We additionally performed a two-dimensional genome scan using the function *scantwo* in the *qtl* package [176] for all 47 significantly mapped PCs. Significant interactions were determined by permuting the phenotype data for each PC 1000 times and determining the 5% genome-wide error rate for QTL detection.

2.4.7 Heritability estimates

Broad-sense heritability was estimated for each of the 97 PCs using the formula $H^2 = (\sigma_R^2 - \sigma_P^2)/\sigma_R^2$ where σ_R^2 and σ_P^2 are the variance among the RIAIL and parental (N2 and CB4856) phenotypic values, respectively [177]. A variance component model using the R package *regress* was used to estimate the fraction of phenotypic variation explained by additive genetic factors ('narrow-sense' heritability) [178, 28]. The additive relatedness matrix was calculated as the correlation of marker genotypes between each pair of strains. In addition, a two-component variance component model was calculated with both an additive and pairwise-interaction effect. The pairwise-interaction relatedness matrix was calculated as the Hadamard product of the additive relatedness matrix.

2.4.8 Calculation of hotspots

We estimated cM distances from recombination events in the RIAIL panel to account for non-uniform distribution of genetic diversity across the genome. The genome was divided into 65 total bins with each bin containing 26 cM. To determine if the 82 QTL significantly clustered around particular genomic regions, we set a threshold for significant QTL hotspots based on the 99^{th} percentile of a Poisson distribution with a mean of 1.2 QTL (total QTL/total bins).

2.4.9 Generation of near-isogenic lines (NIL)

NILs were generated by crossing selected RIAILs to each parental genotype. For each NIL, eight crosses were performed followed by six generations of propagating isogenic lines to ensure homozygosity of the genome. For each cross, PCR amplicons for insertion-deletion (indel) variants on the left and right of the introgressed region were used to confirm progeny genotypes and select non-recombinants within the introgressed region. NILs were whole-genome sequenced as described below to confirm their genotype. A statistical power calculation was used to determine the minimal number of technical replicates required to observe the predicted phenotypic effect of each QTL at 80% power. The number of technical replicates tested per assay for any given toxin did not exceed 100 because of experimental timing constraints. The principal components that mapped to each NIL region

are those with a QTL with a confidence interval that overlaps with or spans the entire introgressed region in the NILs.

2.4.10 Whole-genome sequence library prep and analysis

DNA was isolated from 100-300 μL of packed animals using Qiagen's Blood and Tissue kit (catalog # 69506). Following the ATL lysis step, 4 μl of 100 mg/mL RNAse was added to each sample and allowed to incubate for two minutes at room temperature. DNA concentration was determined using the Qubit dsDNA BR Assay Kit (catalog # Q32850). For each strain, a total of 0.75 ng of DNA was combined with 2.5 μL transposome (Illumina; kit # FC-121-1011) diluted 35x with 1x Tris Buffer (10x Tris Buffer: 100 mM Tris-HCl pH 8.0, 50 mM $MgCl_2$) in a 10 μL final volume on ice. This reaction was incubated at 55℃ for 10 minutes. The amplification reaction for each strain contained (final concentrations): 1x ExTaq Buffer, 0.2 mM dNTPs, 1 U ExTaq (Takara, catalog # RR001A), 0.2 μM primer 1, 0.2 μM primer 2, and 5 μL of tagmentation material from the previous step in a 25 μL total volume. Each strain had a unique pair of indexed primers. We first made a master mix containing buffer, water, dNTPs, and ExTaq then aliquoted the appropriate volume of this mix into each well. We added the specific primer sets to each well and finally the tagmentation reaction. The amplification reaction was incubated in a thermocycler with the following conditions: 72℃ for three minutes (1x); 95℃ for 30 seconds (1x); 95℃ for 10 seconds, 62℃ for 30 seconds, 72℃ for three minutes (20x); 10℃ on hold. We combined 8 μL from each amplification reaction to generate a pool of libraries. A portion of the libraries was electrophoresed on a 2% agarose gel. DNA was excised and gel purified using Qiagen's Gel Purification Kit (catalog # 28706). The libraries were sequenced on the Illumina HiSeq 2500 platform using a paired-end 100 bp reaction lane. Alignment, variant calling, and filtering were performed as described previously [134]. NIL and CSS genotypes were called using the VCF file and a Hidden Markov Model as described previously [179].

2.4.11 Generation of chromosome substitution strains (CSS)

CSSs were generated by crossing N2 and CB4856 parental strains and mating cross progeny to each parental genotype. For each CSS, eight crosses were performed followed by six generations of

propagating isogenic lines to ensure homozygosity of the genome. For each cross, PCR amplicons for indels on the left and right of the introgressed region were used to confirm progeny genotypes and select non-recombinants within the introgressed region. CSSs were whole-genome sequenced as described above to confirm their genotype. As described for NIL assays, power calculations were performed to determine the number of technical replicates required to observe the predicted phenotypic effect of the CSSs.

2.4.12 Selection of traits to categorize in CSS and NIL assays

Pairwise correlations of RIAIL phenotypes among the 24 growth-response traits measured by the BIOSORT were calculated using the *cor* function within the *stats* package in R with the *use* argument set to “pairwise.complete.obs”. For each toxin, hierarchical clustering was performed using the function *hclust* from the *stats* package (R Core Team 2017). *Cutree* was then used to group the resulting dendrogram into *k* groups, where *k* is equal to the minimum number of principal components that explained at least 90% of the phenotypic variance in the RIAILs. For each principal component that mapped to a hotspot, the growth-response trait that was most correlated to that principal component, as well as all growth-response traits within that cluster of the dendrogram, were assayed in NIL and CSS experiments (**Table 2-2**).

2.4.13 Categorization of CSS and NIL results

Toxin responses for NILs and CSSs were tested using the high-throughput fitness assay for traits correlated with mapped principal components as described above (**Table 2-2**). Complete pairwise statistical analyses of strains was performed for each trait tested in all CSS and NIL assays (Tukey honest significant difference (HSD) test). A p-*value* of $p < 0.05$ was used as a threshold for statistical significance. NIL recapitulation was defined by the significance and direction of effect of the NIL compared to the parental strains. Six categories were defined: 1) no parental difference, 2) recapitulation, 3) no QTL effect, 4) bidirectional interaction, 5) unidirectional interaction, and 6) miscellaneous. Traits for which N2 and CB4856 phenotypes were not statistically different comprise the ‘no parental difference’ category and were not further categorized. Traits in the ‘recapitulation’

category must satisfy the following criteria: significant difference between the parental strain phenotypes, significant difference between phenotypes of each NIL and the parent that shares its background genotype, and both NILs must display the expected direction of effect of the introgressed genotype. Traits with 'no QTL effect' displayed a significant parental phenotypic difference and the phenotype of each NIL was not statistically different from the phenotype of the parent sharing its background genotype. Traits that have a 'bidirectional interaction' must display a significant parental phenotypic difference, the phenotypes of both NILs must be significantly different from phenotypes of both parents, and the phenotypes of both NILs must be transgressive (lie beyond the phenotypic range of the parental strains). Lastly, traits with a 'unidirectional interaction' were categorized similarly to the bidirectional interaction, except only one NIL must display a transgressive phenotype and the other NIL either shows no QTL effect or recapitulation. Traits that did not fit these descriptions were categorized as 'miscellaneous'.

Traits in the chromosome V hotspot were further categorized using the combined data from both the CSS and NIL assays. Seven categories were defined: 1) no parental difference, 2) recapitulation, 3) no QTL effect, 4) external inter-chromosomal interaction (uni- or bidirectional), 5) internal inter-chromosomal interaction (uni- or bidirectional), 6) intra-chromosomal interaction (uni- or bidirectional), and 7) miscellaneous. 'No parental difference' was defined by traits in which the parental strains were either not significantly different from each other or did not have the same direction of effect in both the CSS and NIL assays. 'Recapitulation' and 'no QTL effect' traits were defined by traits that were classified as either recapitulating or no QTL effect, respectively, in both assays. Traits displaying an 'external inter-chromosomal interaction' show evidence for interaction in the CSS but no interaction (either recapitulating or no QTL effect) in the NIL. On the other hand, traits displaying an 'internal inter-chromosomal interaction' showed evidence of the same interaction for both the CSS and the NIL assays. Finally, traits displaying an 'intra-chromosomal interaction' showed evidence of an interaction in the NIL but not in the CSS assay. All other traits that did not fit these descriptions were categorized as 'miscellaneous'.

2.4.14 Statistical analysis

All statistical tests of phenotypic differences in the NIL and CSS assays were performed in R (version 3.3.1) using the *TukeyHSD* function [175] on an ANOVA model with the formula (*phenotype ~ strain*). The p-*values* of individual pairwise strain comparisons were reported, and a p-*value* of $p < 0.05$ was deemed significant. The direction of effect of each NIL was determined by comparing the median phenotypic value of the NIL replicates to that of each parental strain. NILs whose phenotypes were significantly different from both parents and whose median lied outside of the range of the parental phenotype medians were considered hypersensitive or hyper-resistant. Comparing LOD scores and variance explained between traits with no parental effect and traits with a significant parental effect in the NIL assays was performed using a Wilcoxon rank sum test with continuity correction using the *wilcox.test* function in R [175].

2.5 RESULTS

2.5.1 Identification of QTL underlying variation in responses to 16 diverse toxins

Using a high-throughput fitness assay (Methods), we tested variation in 24 fitness-related traits in responses of four genetically divergent strains to different concentrations of 16 toxins, comprising chemotherapeutics, heavy metals, pesticides, and neuropharmaceuticals. A concentration of each toxin was selected that minimized within-strain variation and maximized variation between two of these divergent strains, N2 (the laboratory strain) and CB4856 (a wild isolate from Hawaii). For the selected concentration of each toxin, we assayed 24 growth-response traits for a panel of 296 recombinant inbred advanced intercross lines (RIAILs) generated between the N2 and CB4856 parental genetic backgrounds [68]. Because some of the growth-response traits are highly correlated, we performed principal component analysis (PCA) for each toxin. The minimum number of principal components (PCs) that explained at least 90% of the total phenotypic variance within each toxin was selected for mapping, for a total of 97 PCs across all toxins (minimum of five PCs and a maximum of eight PCs per toxin). We then used linkage mapping to identify quantitative trait loci (QTL) that underlie variation in these 97 PCs.

We detected a total of 82 significant QTL (across 47 PCs) from the 97 PCs tested (**Figure 2-1**). We did not find a single toxin-response QTL shared robustly across all of the various PCs and toxins tested nor across all PCs within any one toxin. However, the majority of QTL on chromosome I were detected in responses to chemotherapeutics. Additionally, almost every toxin (with the exception of FUdR) had QTL that underlie trait variation on at least two different chromosomes, highlighting the diverse architectures implicated across traits, even within a single toxin. Despite the seemingly independent distributions of QTL, we found that the majority of the QTL (61%) mapped to chromosomes IV and V.

2.5.2 Both additive and interactive QTL underlie toxin responses

For each of the PCs that were impacted by the 82 QTL identified using linkage mapping, we calculated the broad-sense heritability, the proportion of broad-sense heritability that could be attributed to additive genetic components (narrow-sense heritability) (**Figure 2-2A**), and the proportion of narrow-sense heritability that was explained by QTL detected through linkage mapping (**Figure 2-2B,** Methods). In many cases, additive genetic components could not explain all of the phenotypic variation predicted to be caused by genetic factors. These results suggest that other additive loci with small effect sizes impact toxin responses, but we failed to detect these QTL by our linkage mapping analyses, potentially because of high complexity and/or insufficient statistical power. Alternatively, this missing heritability could be indicative of genetic interactions [27].

To determine how much of the phenotypic variance comes from additive or interacting genetic components, we fit a linear mixed-effect model to the RIAIL phenotype data for the 47 PCs controlled by the 82 QTL. We observed a range of additive and epistatic components contributing to phenotypic variation across toxin classes. On average, cisplatin, topotecan, and FUdR are primarily explained by additive models. Alternatively, paraquat, irinotecan, vincristine, and mechlorethamine have a larger fraction of their phenotypic variance attributable to genetic interactions than additive effects. To localize potential genetic interactions for these 82 QTL, we scanned the genome for interactions between pairs of markers that might affect the phenotypic distribution of the RIAIL panel (Methods). We identified three significant interactions. This two-factor genome scan was unable to localize all epistatic components identified by the linear mixed-effect model (**Figure 2-2**), perhaps because of

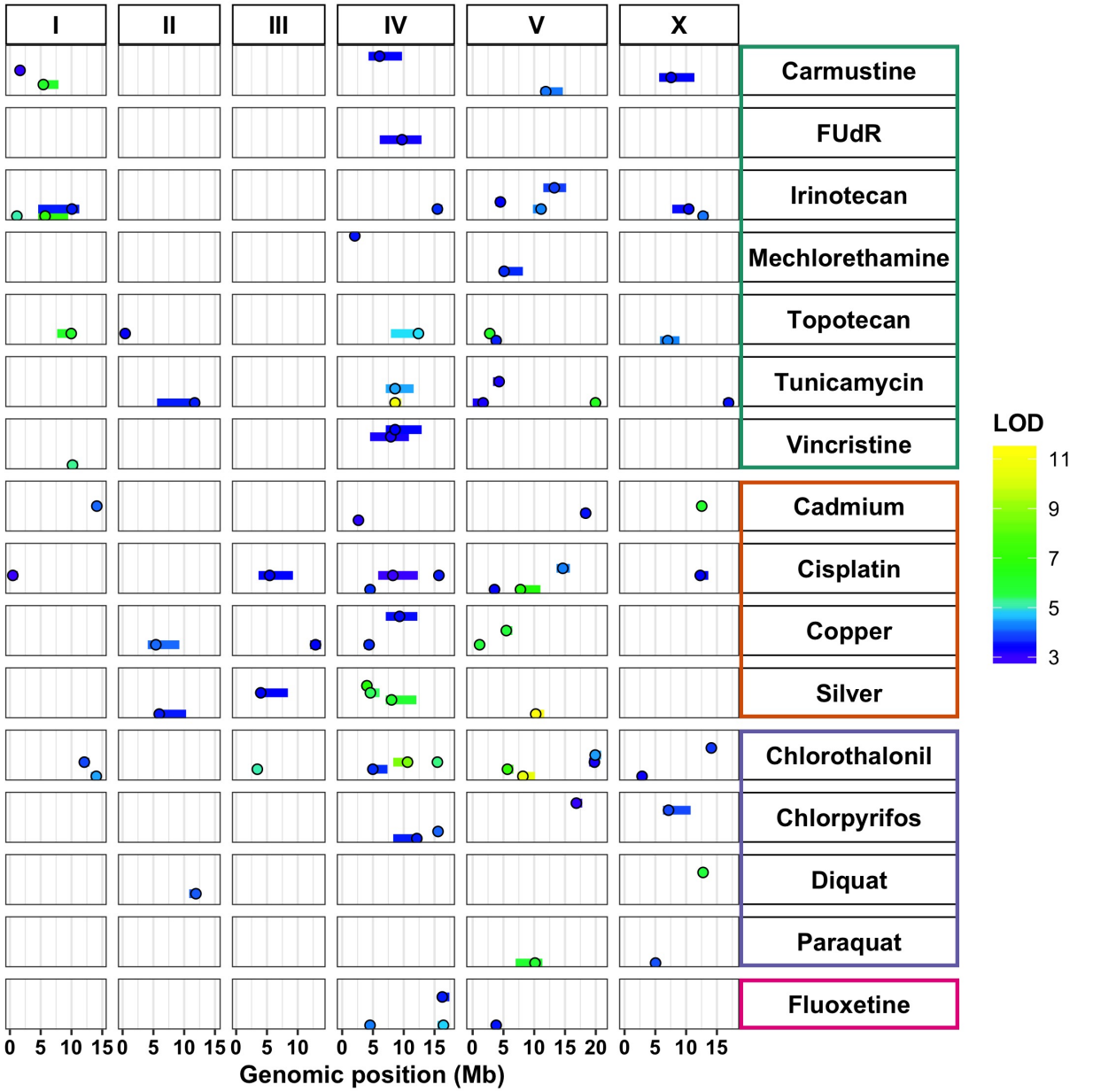


Figure 2-1: Diverse genetic architectures are implicated in responses to 16 toxins. Linkage mapping results for principal components that represent 82 QTL across 16 toxins, comprising chemotherapeutics (teal), heavy metals (orange), pesticides (purple), and neuropharmaceuticals (pink) are plotted. Genomic position (Mb) is shown along the x-axis, split by chromosome, and each of the 47 principal components with a significant QTL is plotted along the y-axis. Each QTL is plotted as a point at the location of the most significant genetic marker and a line indicating the 95% confidence interval. QTL are colored by the logarithm of the odds (LOD) score, increasing in significance from blue to green to yellow.

missing small-effect additive loci in the model and/or insufficient statistical power to identify small-effect interactions.

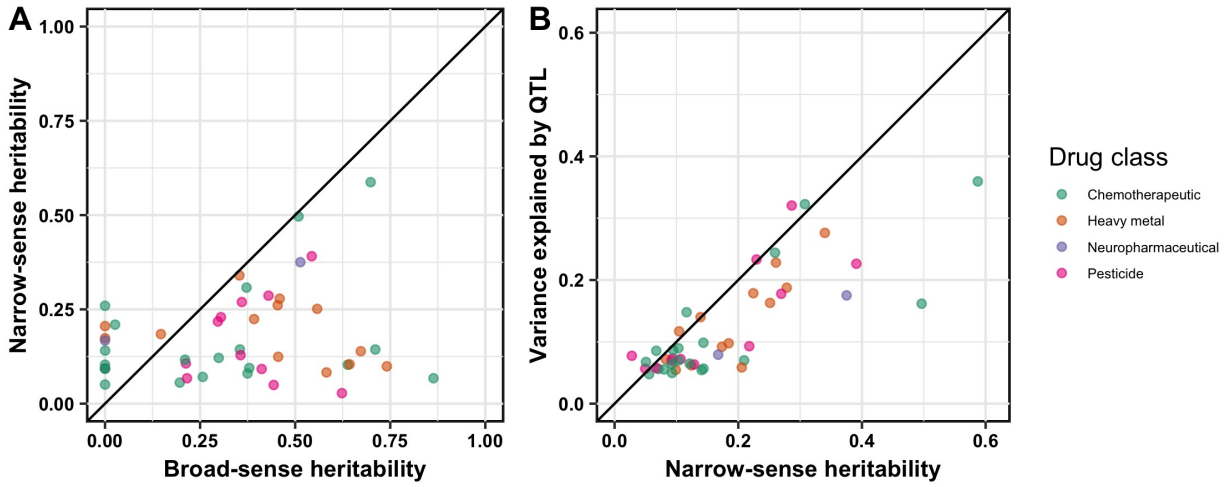


Figure 2-2: Additive genetic components identified by linkage mapping do not explain all heritable contributions to toxin-response variation. For 47 principal components representing the 82 QTL, we compared **(A)** the broad-sense heritability (x-axis) calculated from the RIAIL phenotypic data versus the narrow-sense heritability (y-axis) estimated by a mixed model and **(B)** the narrow-sense heritability (x-axis) versus the variance explained by all QTL detected by linkage mapping (y-axis). In both plots, each principal component is plotted as a point whose color indicates drug class (chemotherapeutic, heavy metal, neuropharmaceutical, or pesticide). The diagonal line represents y = x and is shown as a visual guide.

2.5.3 Three QTL hotspots underlie variation in responses to diverse toxins

The majority of toxin-response QTL cluster on chromosomes IV and V (**Figure 2-1**). We sought to determine if such QTL clustering could be expected by chance or if this clustering is indicative of toxin-response QTL hotspots. To account for the higher rate of recombination, and thus more genetic diversity, on the chromosome arms [78], we divided the genome evenly into 65 bins and calculated the number of QTL that mapped to each bin (**Figure 2-3**, Methods). Three bins with more QTL than expected based on a Poisson distribution [177] were classified as hotspots. These hotspots are located on the center of chromosome IV, the right of chromosome IV, and the center of chromosome V and are hereby denoted as IVL, IVR, and V, respectively. Importantly, these hotspots are not driven by multiple principal components within a single toxin. Instead, hotspots comprise multiple QTL across a variety of principal components and toxins. In fact, 14 of the 16 toxins tested (**Table 2-1**) have a principal component that maps to at least one of the three hotspots. Of the 82 QTL, 18 mapped to IVL, 8 mapped to IVR, and 9 mapped to V. In total, 33 QTL map to a hotspot (note that two QTL have confidence intervals that span both hotspots on chromosome IV). We sought to experimentally

validate the predicted additive and epistatic effects on toxin responses for QTL that mapped to the three hotspots.

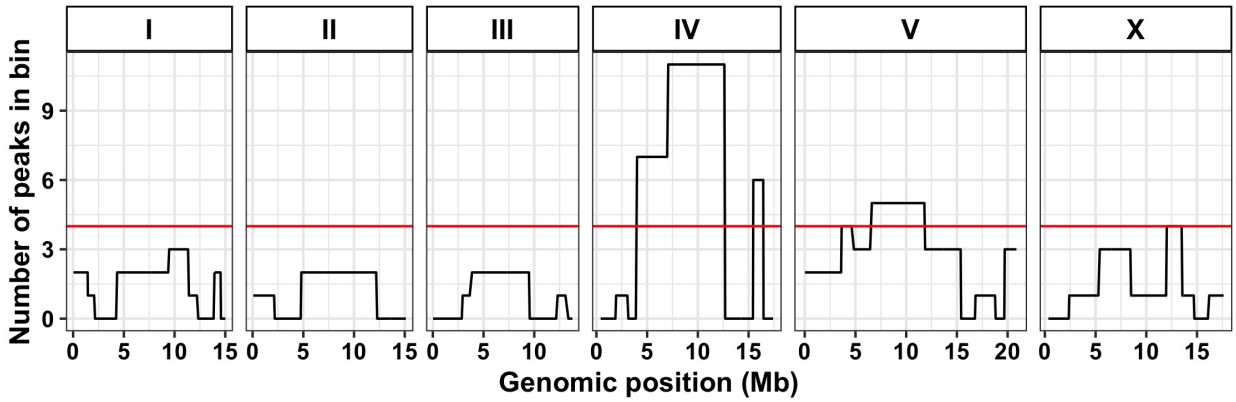


Figure 2-3: Three QTL hotspots impact toxin responses. Each chromosome is divided into equal bins of 26 cM, resulting in a total of 65 bins across the genome. The x-axis shows the genomic position (Mb), and the y-axis shows the number of QTL that lie within the corresponding bin. The red line indicates the 99th percentile of a Poisson distribution with a mean of 1.26 QTL (total QTL/total bins).

2.5.4 Near-isogenic lines recapitulate some of the predicted QTL effects

To experimentally validate the QTL identified from linkage mapping, we created near-isogenic lines (NILs) for the IVL, IVR, and V hotspots. Each NIL has a small genomic region introgressed from one parental strain into the genome of the opposite parental strain (Methods). These NILs were whole-genome sequenced and found to match the expected genotype in the hotspot region; however, occasionally additional breakpoints were observed (Methods). We tested each NIL in our high-throughput fitness assay for a subset of the toxins with a QTL that maps to a given hotspot, choosing QTL with small, medium, and large effect sizes to test our ability to recapitulate various effect sizes (**Table 2-1**, **Table 2-2**). We tested five toxins (ten QTL) with the IVL NILs, three toxins (four QTL) with the IVR NILs, and five toxins (six QTL) with the V NILs. In total, we tested 20 QTL across eight toxins for recapitulation using the NILs.

For each of these 20 QTL, we identified the toxin-response trait that is most correlated with the principal component controlled by that QTL. We then assayed the NILs for that toxin-response trait as well as all toxin-response traits within its same trait cluster, because each principal component

Table 2-1: Summary of toxins and principal components mapped per hotspot[a]

Toxin	Class	PCs in IVL	PCs in IVR	PCs in V
Cadmium	Heavy Metal	0	0	0
Carmustine*	Chemotherapeutic	1*	0	1*
Chlorothalonil*	Pesticide	2*	1*	1*
Chlorpyrifos	Pesticide	1	1	0
Cisplatin*	Chemotherapeutic	2*	1	2*
Copper	Heavy Metal	2	0	0
Diquat	Pesticide	0	0	0
Fluoxetine*	Neuropharmaceutical	1	2*	0
FUdR	Chemotherapeutic	1	1	0
Irinotecan*	Chemotherapeutic	0	1*	2
Mechlorethamine	Chemotherapeutic	0	0	1
Paraquat*	Pesticide	0	0	1*
Silver*	Heavy Metal	3*	0	1*
Topotecan	Chemotherapeutic	1	0	0
Tunicamycin	Chemotherapeutic	2*	0	0
Vincristine	Chemotherapeutic	2	1	0

[a]* denotes a toxin tested with NIL and/or CSS assays

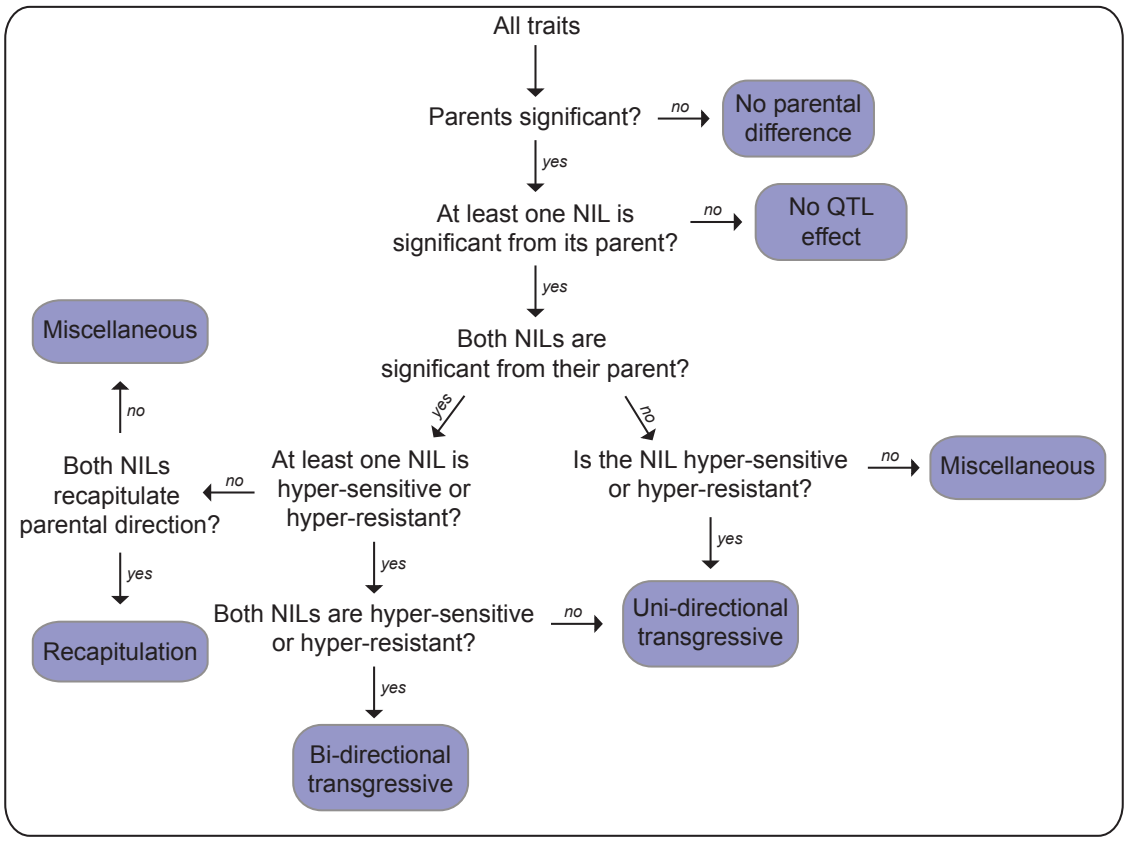
comprises multiple toxin-response traits (**Table 2-2**, Methods). We tested 42 toxin-response traits with the IVL NILs, 12 toxin-response traits with the IVR NILs, and 45 toxin-response traits with the V NILs (**Table 2-2**). In total, we performed 99 tests of recapitulation of QTL effects for toxin-response traits. The results of these tests allowed us to sort QTL effects into six different categories: ‘no parental effect’, ‘recapitulation’, ‘no QTL effect’, ‘unidirectional transgressive’, ‘bidirectional transgressive’, or ‘miscellaneous’ (**Figure 2-4**).

Of these 99 tests, 23 did not display a significant phenotypic difference between the parent strains (N2 and CB4856) in the NIL assay and were categorized as ‘no parental effect’ (Methods). The remaining 76 tests in which a significant parental difference was observed were classified further. We predicted that if a single QTL in the introgressed region contributed to the parental phenotypic difference, then each NIL would have a phenotype significantly different than the parental strain with the same genetic background. Furthermore, we expected each NIL to have a phenotype similar to the parental strain of its introgressed genomic region. This ‘recapitulation’ model was consistent for four tests. The normalized brood size trait in cisplatin (cisplatin.norm.n in cisplatin PC4) is one such

Table 2-2: All traits tested in NIL and CSS assays

PC	Hotspot	Correlated Traits	Correlation Range
carmustine.PC1	V	mean.EXT, mean.TOF, q75.EXT, median.EXT, median.TOF, q75.TOF, median.norm.EXT, q90.TOF, q90.EXT	0.72-0.95
carmustine.PC6	IVL	q25.norm.EXT, q10.norm.EXT	0.33-0.39
chlorothalonil.PC1	V	mean.EXT, q75.EXT, mean.TOF, median.EXT, median.TOF, q75.TOF	0.73-0.95
chlorothalonil.PC2	IVL	cv.TOF, cv.EXT	0.72-0.90
chlorothalonil.PC3	IVL, IVR	mean.norm.EXT, q75.norm.EXT, q90.norm.EXT, median.norm.EXT	0.50-0.65
cisplatin.PC1	IVL, V	mean.EXT, mean.TOF, median.EXT, median.TOF, q75.TOF, q75.EXT, q90.EXT, q90.TOF	0.78-0.97
cisplatin.PC3	IVL	var.TOF, var.EXT	0.38-0.54
cisplatin.PC4	V	norm.n, n	0.76-0.80
fluoxetine.PC1	IVR	mean.norm.EXT, q75.norm.EXT, mean.EXT, q75.EXT, q90.norm.EXT, q90.EXT	0.79-0.96
fluoxetine.PC5	IVR	q90.norm.EXT, q75.norm.EXT, mean.norm.EXT, q75.EXT, mean.EXT, q90.EXT	0.07-0.40
irinotecan.PC2	IVR	cv.TOF, cv.EXT	0.57-0.84
paraquat.PC1	V	median.EXT, mean.EXT, q25.EXT, q75.EXT, mean.TOF, q75.TOF, q10.EXT, q90.EXT, q90.TOF, median.TOF, q25.TOF, q10.TOF	0.75-0.95
silver.PC1	V	mean.EXT, median.EXT, q75.EXT, mean.TOF, q90.EXT, q90.TOF, median.TOF, q75.TOF	0.77-0.96
silver.PC3	IVL	q10.norm.EXT, q25.norm.EXT, mean.norm.EXT, median.norm.EXT, q75.norm.EXT, q90.norm.EXT	0.32-0.64
silver.PC4	IVL	n, norm.n	0.84-0.84
silver.PC5	IVL	n, norm.n	0.41-0.41
tunicamycin.PC1	IVL	median.EXT, q75.EXT, mean.TOF, q75.TOF, median.TOF, median.norm.EXT, q90.EXT, q90.TOF, mean.EXT, q75.norm.EXT, mean.norm.EXT, q25.norm.EXT, q90.norm.EXT, q10.norm.EXT	0.69-0.96
tunicamycin.PC3	IVL	norm.n, n	0.47-0.50

example of a trait in which the NILs on the center of chromosome V recapitulated the expected parental phenotype (**Figure 2-5A**). For 11 of the remaining 72 tests, the phenotype of each NIL was not significantly different from the phenotype of the parental strain sharing its background genotype. This phenotype indicates that the introgressed NIL region was not affecting the toxin-response phenotype. This lack of QTL effect suggests that the genetic architecture is more complex, we lacked sufficient statistical power to detect the QTL effect, or the real QTL is outside the introgressed region. The NILs on the center of chromosome V showed this result for median animal length in silver



Category	CSS	NIL
Recapitulation	Recapitulation	Recapitulation
No parental effect	No parental difference	No parental difference
No QTL effect	No QTL effect	No QTL effect
Inter-chromosomal external	Uni-directional or bi-directional transgressive phenotype	Recapitulation or no QTL effect
Inter-chromosomal internal	Uni-directional or bi-directional transgressive phenotype	Uni-directional or bi-directional transgressive phenotype
Intra-chromosomal	Recapitulation or no QTL effect	Uni-directional or bi-directional transgressive phenotype

Figure 2-4: NIL and CSS trait categorizations. **(A)** Flowchart for categorizing traits from the NIL or CSS tests of recapitulation of QTL effects. **(B)** Six potential categories for chromosome V traits tested in both NIL and CSS assays with a significant and consistent parental phenotypic split across both assays. The miscellaneous category is not depicted, but encompasses any other combination of NIL and CSS assay results.

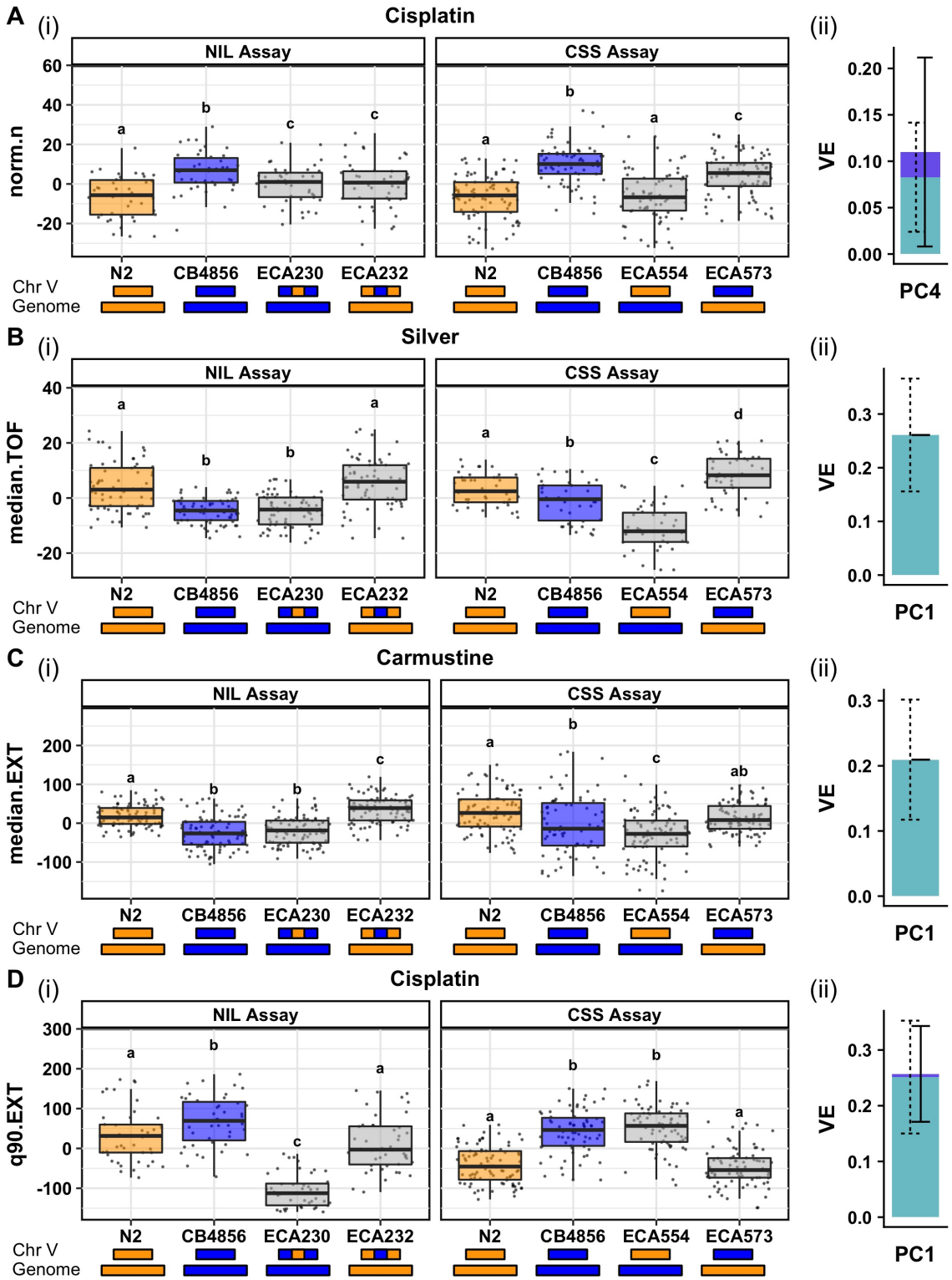
(silver.median.TOF in silver PC1) (**Figure 2-5B**). The phenotypes of the NILs for the remaining 61 tests cannot be explained by a single QTL model. For many of these tests, we observed NIL

phenotypes that are more sensitive or more resistant than both parental strains, suggesting that loci of opposite genotypes act additively or interact in the NILs to create transgressive phenotypes [180]. This finding was supported by the mixed-effects model, which suggested that both additive and interacting QTL remained undetected by linkage mapping (**Figure 2-2**). We further explored the results of these 61 tests by characterizing them based on the patterns of the transgressive phenotypes we observed.

For 38 of these 61 tests, only one NIL showed a transgressive phenotype. Some of these 38 'unidirectional transgressive' phenotypes seem to show an antagonism that counteracted the effect of the introgressed region (a predicted sensitive phenotype becomes hyper-resistant or a predicted resistant phenotype becomes hypersensitive, *e.g.* carmustine.median.EXT in carmustine PC1, **Figure 2-5C**). Other phenotypes displayed synergy that increased the effect of the introgressed region (a predicted sensitive phenotype becomes a hypersensitive phenotype or a predicted resistant phenotype becomes a hyper-resistant phenotype, *e.g.* cisplatin.q90.EXT in cisplatin PC1, **Figure 2-5D**). Interestingly, in most cases (82%), the transgressive phenotype was observed in the strain with the N2 genotype introgressed into the CB4856 background.

In addition to unidirectional transgressive phenotypes, we identified seven tests with suggested 'bidirectional transgressive' phenotypes in which both NILs showed an extreme phenotype compared to the parental strains. Some of these 'bidirectional transgressive' phenotypes were suggestive of purely antagonistic effects (*e.g.* tunicamycin.mean.norm.EXT), but others suggested an antagonistic effect in one NIL and a synergistic effect in the other (*e.g.* paraquat.median.TOF). We identified no cases of bidirectional synergistic effects. The remaining 16 tests of the 76 with a parental difference did not fall into any of the above categories and were classified as 'miscellaneous'.

The toxin-response traits tested above for recapitulation of QTL effects were selected to represent principal components that were mapped with linkage mapping. We wanted to compare the NIL assay categorizations for the toxin-response traits that underlie each principal component to analyze the overall QTL effect (**Figure 2-6**). For example, two traits, n and norm.n, were selected to represent cisplatin PC4 (**Table 2-2**). Both of these toxin-response traits were placed into the 'recapitulation' category from the NIL assay results (**Figure 2-6**). These results suggest that a single additive QTL underlies the brood size variation captured by PC4. Fourteen tunicamycin-response traits were


A
(i)
Cisplatin
NIL Assay
CSS Assay
norm.n
N2
CB4856
ECA230
ECA232
ECA554
ECA573
Chr V
Genome
(ii)
VE
PC4
B
Silver
median.TOF
PC1
C
Carmustine
median.EXT
D
Cisplatin
q90.EXT

selected to represent tunicamycin PC1 (**Table 2-2**). Eight of these 14 traits displayed unidirectional transgressive phenotypes, four traits displayed bidirectional transgressive phenotypes, and the remaining two traits did not have a significant parental phenotypic difference (**Figure 2-6**). Regardless of the classification, we see the same trend of resistance (ECA231 > N2 > CB4856 > ECA229) across 11 of the 14 traits representing this principal component. Therefore, our strict significance thresholds for categorization might have caused some phenotypes to be mis-categorized (usually into the miscellaneous or no parental/QTL effect categories). The prevalence of transgressive phenotypes in tunicamycin-response traits suggests that multiple QTL, acting additively or interacting, might impact tunicamycin responses.

We next sought to compare categorizations of toxin-response traits and QTL effect sizes of the PCs for those traits. The QTL underlying cisplatin PC4 explains about 7% of the total phenotypic variance. The traits selected to represent cisplatin PC4 were placed into the 'recapitulation' category, despite the small effect size of the QTL (**Figure 2-6**). In the other example above, the QTL underlying tunicamycin PC1 explains almost 16% of the total phenotypic variance, which is one of the highest effect sizes mapped in this study. The toxin-response traits selected to represent this principal component showed mostly transgressive phenotypes, indicating undetected additive or interacting QTL despite the seemingly large-effect additive QTL identified in linkage mapping (**Figure 2-6**).

Figure 2-5 *(preceding page)*: Results from near-isogenic line (NIL) and chromosome-substitution strain (CSS) tests of recapitulation of QTL effects are categorized based on potential genetic mechanisms implicated in toxin responses. A trait contributing to a mapped principal component for each category is reported: **(A)** Recapitulation (cisplatin norm.n, PC4), **(B)** Inter-chromosomal external bidirectional loci (silver median.TOF, PC1), **(C)** Inter-chromosomal internal unidirectional loci (carmustine median.EXT, PC1), and **(D)** Intra-chromosomal unidirectional loci (cisplatin q90.EXT, PC1). In each case, we show results from (i) the NIL assay (left) and CSS assay (right) plotted as Tukey box plots. The y-axis indicates residual phenotypic values for the given trait. Different letters (a-d) above each Tukey box plot represent significant differences ($p < 0.05$) while the same letter represents non-significant differences between two strains (Tukey HSD). The genotype of each strain on the x-axis is modeled by the colored rectangles beneath the plots (N2 genotypes are orange, CB4856 genotypes are blue). (ii) A stacked bar plot shows the the proportion of phenotypic variation attributable to additive (light blue with dashed error bars) and interactive (dark blue with solid error bars) genetic factors of the principal component represented by each trait, based on a mixed model.

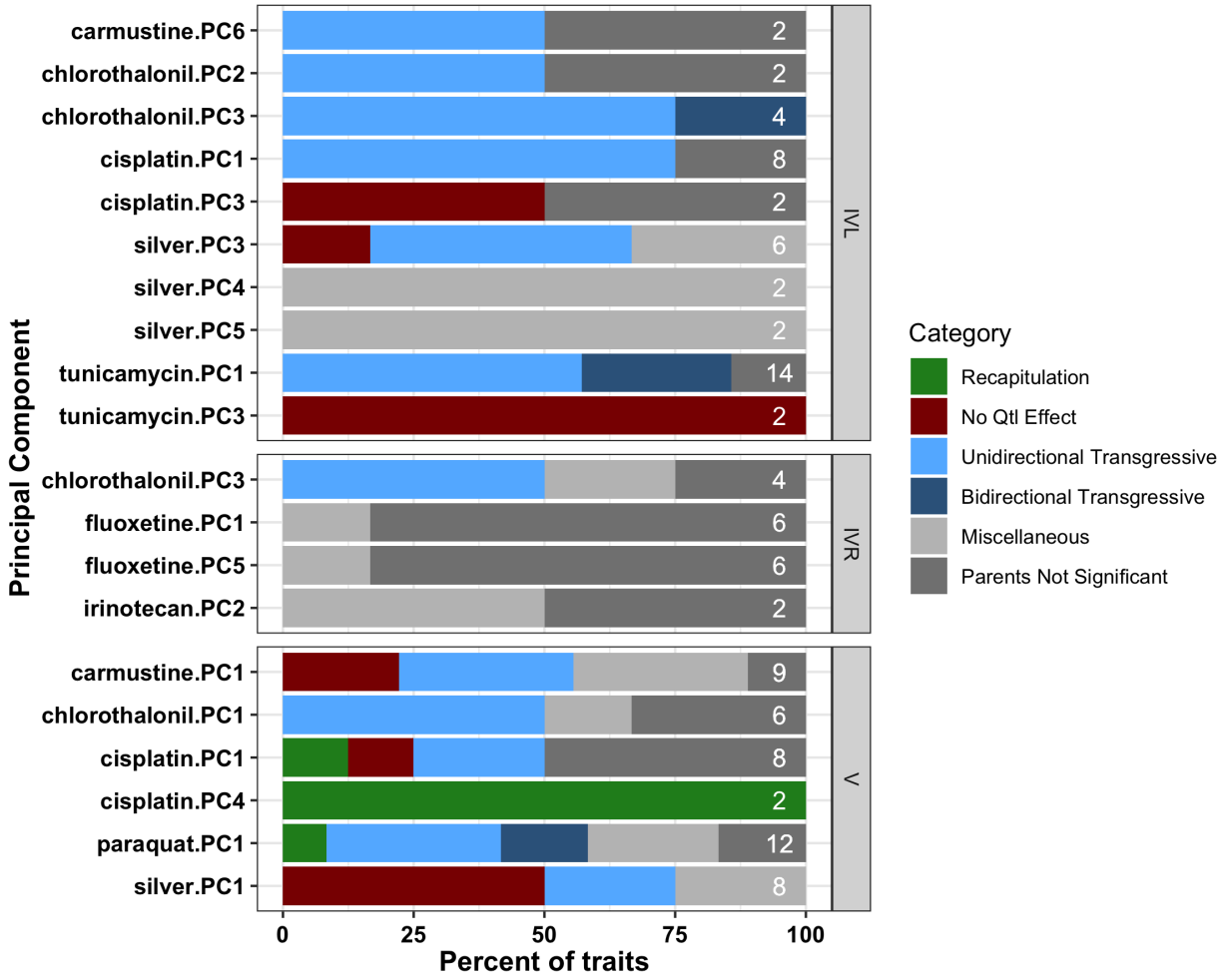


Figure 2-6: Categorization of correlated traits. The y-axis shows the principal component tested in NIL assays, split by hotspot. Each color represents the NIL assay categorization for each trait within a correlation cluster - either recapitulation (green), no QTL effect (red), unidirectional transgressive phenotype (light blue), bidirectional transgressive phenotype (dark blue), miscellaneous (light grey) or no significant parental difference (dark grey). The percent of all traits within the correlation cluster for each principal component that falls within a given category is shown on the y-axis. Numbers on the right of each bar indicate the number of traits within a correlation cluster.

2.5.5 Chromosome-substitution strains localize QTL underlying transgressive phenotypes

Because we found evidence of loci where opposite genotypes at each locus cause transgressive phenotypes, we attempted to further characterize these loci (**Figure 2-4, Figure 2-7**). To define each set of loci as either intra-chromosomal or inter-chromosomal, we built reciprocal chromosome-substitution strains (CSSs) for the hotspot on chromosome V that had the entire chromosome V introgressed from one parental strain into the genome of the opposite parental strain

(Methods). The hotspot on chromosome V was chosen to isolate the effects of one hotspot and avoid complications arising from traits whose confidence intervals might lie within both of the hotspots on chromosome IV. The CSSs were whole-genome sequenced and found to have the expected genotype at all markers (Methods), except for the chromosome I incompatibility locus [**?**] (Seidel *et al.* 2011; Seidel, Rockman, and Kruglyak 2008). We performed tests of recapitulation of QTL effects with the CSSs for each of the 45 toxin-response traits across the five toxins tested with the chromosome V NILs (**Table 2-2**).

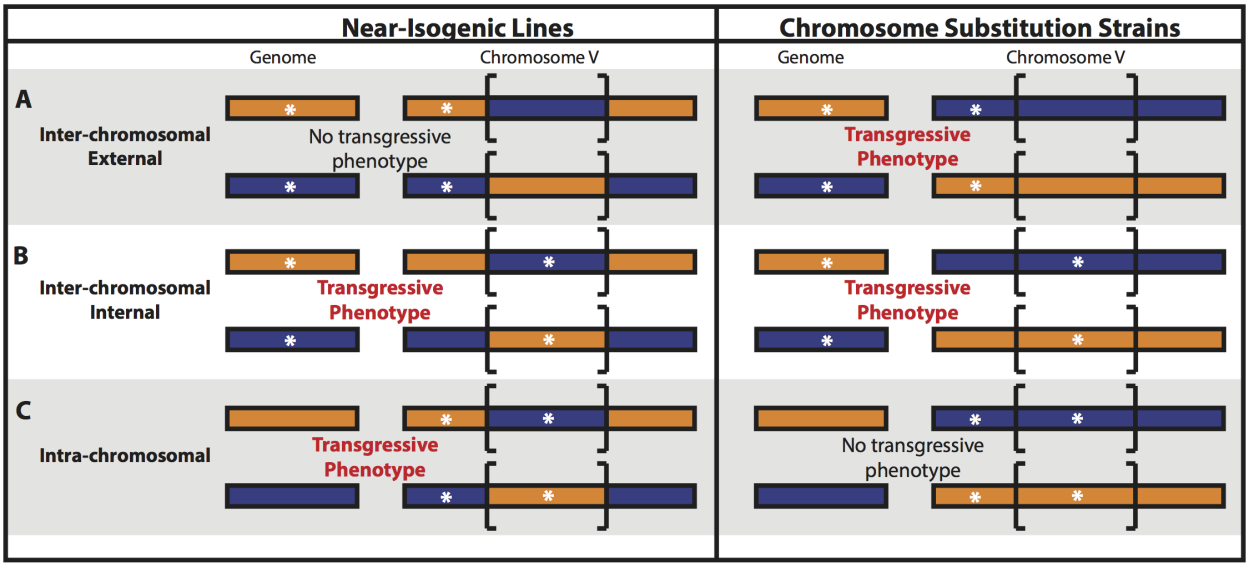


Figure 2-7: A model for potential locations of two loci is shown, according to toxin-response phenotypes of near-isogenic lines (NILs) and chromosome-substitution strains (CSSs). The NILs are represented on the left, and the CSSs are represented on the right. The strain genotype is indicated by colored rectangles. N2 is orange, and CB4856 is blue. Brackets indicate the genomic region that is introgressed in the NILs. White asterisks represent a potential location for additive or epistatic loci underlying transgressive phenotypes. Although bidirectional transgressive phenotype models are shown, each model could be bidirectional (both reciprocal introgressed strains show transgressive phenotypes) or unidirectional (only one reciprocal introgressed strain shows a transgressive phenotype). Models showing **(A)** inter-chromosomal external effects between a locus outside of the introgressed region in the NILs and a locus on another chromosome, **(B)** inter-chromosomal internal effects between a locus within the introgressed region in the NILs and a locus on another chromosome, and **(C)** intra-chromosomal effects between a locus within and a locus outside of the introgressed region in the NILs are drawn.

For traits in which the parental phenotypic difference was significant and consistent across the NIL and CSS tests, NIL and CSS phenotypes could be compared across assays. Eight traits across five toxins fit this criterion. One trait (cisplatin.norm.n) displayed phenotypic 'recapitulation' of the introgressed region in both the NIL and the CSS tests, suggesting a single QTL model (**Figure 2-5A**). Alternatively, transgressive phenotypes are indicative of a multi-QTL model, and locations of additive

or interacting QTL can be surmised by comparing results from the NIL and CSS tests. Transgressive phenotypes controlled by inter-chromosomal loci are defined by two loci on separate chromosomes that act additively or epistatically. Because NILs and CSSs have introgressed genotypes on chromosome V, we can deduce that at least one of the two inter-chromosomal loci is located on chromosome V. We further divided the inter-chromosomal class into two categories: 'inter-chromosomal external', in which the chromosome V locus is outside the region introgressed in the NILs (**Figure 2-7A**) and 'inter-chromosomal internal', in which the chromosome V locus is within the region introgressed in the NILs (**Figure 2-7B**). For an 'inter-chromosomal external' model, we expect only the CSSs to display hypersensitivity or hyper-resistance, because both loci share the same genotype in the NILs (**Figure 2-7A**) and would therefore not result in a more extreme phenotype than both parents. We found one such trait that fits a 'bidirectional inter-chromosomal external' loci model (silver.median.TOF) (**Figure 2-5B**). For an 'inter-chromosomal internal' model, we expected both the CSSs and the NILs to display the same hypersensitivity or hyper-resistance, because both strains share the same genotype across the introgressed region in the NILs (**Figure 2-7B**). We identified one such trait that fits a 'unidirectional inter-chromosomal internal' loci model (carmustine.median.EXT) (**Figure 2-5C**). To identify intra-chromosomal loci that underlie transgressive phenotypes in the remaining 10 traits, we searched for traits that display evidence of either a uni- or bidirectional transgressive phenotype in the NILs but not in the CSSs (**2-7C**). This result would suggest that two loci of opposite genotypes on chromosome V, one within and one outside the region introgressed in the NILs, act additively or epistatically to cause transgressive phenotypes. We found two examples of such 'unidirectional intra-chromosomal' loci models (*e.g.* cisplatin.q90.EXT, **Figure 2-5D**). The remaining three traits could not be characterized beyond their NIL assay characterization based on the results of the CSS assay.

We revisited the two-factor genome scan results for each of these eight empirically classified traits and compared the findings from these two independent methods used to identify multiple additive or epistatic QTL. No traits with significant interaction terms were identified by the two-factor genome scan. Although many other pairs of loci show suggestive evidence of additive or interacting effects, an increase in statistical power is required to definitively compare these suggestive findings to our

empirically derived model. Overall, this study highlights the benefits of leveraging both experimental and computational strategies to further dissect genetic components that underlie quantitative traits in a metazoan model.

2.6 DISCUSSION

Here, we show that three QTL hotspots underlie differences in responses to 16 diverse toxins. We further characterized these QTL using both modeling and empirical approaches. Through the use of near-isogenic lines and chromosome-substitution strains, we confirmed small-effect QTL and attempted to identify and localize genomic regions causing transgressive phenotypes. Finally, we used statistical analyses to computationally identify loci that might support some of our empirical findings. Although the number of biological replicates and recombinant strains in this study increased our power to detect QTL compared to previous studies, we are still too underpowered to definitively assess if missing heritability is composed of small additive effects or genetic interactions.

2.6.1 Pleiotropic regions underlie QTL shared between and among toxin classes

We performed principal component analysis on toxin-response phenotypes collected for a panel of RIAILs and used linkage mapping to identify 82 toxin-response QTL. Although some of these QTL are unique to one particular toxin, others suggest the existence of pleiotropic QTL that underlie responses to a diverse set of toxins. In particular, three QTL hotspots across chromosomes IV and V were enriched for toxin-response QTL and were investigated further. Because the molecular mechanisms implicated in responses to each toxin differ drastically, the notion that a single gene in each hotspot is regulating the response to several toxins is unlikely. However, the possibility exists that a single gene involved in drug transport could underlie one or several of these hotspots. More likely, multiple genes in close proximity, each regulating a process controlling cellular proliferation and survival, might underlie these hotspots. Notably, two of the three QTL hotspots are in swept regions with lower genetic diversity at the species level [79, 135, 63, 134]. The laboratory strain, N2, has experienced each of the selective sweeps, and CB4856 has not. Linkage mapping using a panel of RIAILs built between these two strains could identify QTL that underlie phenotypic differences between swept and

non-swept strains. Moreover, identifying QTL in these swept regions that underlie variation in fitness-related traits might indicate selective pressures that could have led to these chromosomal sweeps. For example, N2 is more resistant than CB4856 to tunicamycin, an antibiotic and chemotherapeutic produced by the soil bacterium *Streptomyces clavuligerus* [181]. This result might suggest that selective pressure toward responses to antibiotic compounds played a role in driving resistance-conferring alleles, such as those present in N2, to a high frequency. Alternatively, climate conditions could also impact local niche environments to sensitize toxin responses [136]. We observed that N2 is more resistant than CB4856 in responses to the majority of conditions, which could indicate that alleles present in swept strains confer robustness in responses to many conditions. This result emphasizes the importance of genetic background when considering toxin effects [182].

In addition to the three QTL hotspots, pleiotropic QTL across toxins within certain classes are suggested by our linkage mapping results. We observed an enrichment of QTL from the chemotherapeutic class on chromosome I, which could be representative of QTL that underlie a common mechanism targeted by these toxins, such as DNA damage or cell-cycle control. However, because many of these chemotherapeutics have distinct mechanisms of action and share these mechanisms with other toxin classes, this enrichment is likely caused by an overrepresentation of chemotherapeutics in our study. Direct comparisons of toxins with similar cellular mechanisms could provide more insights. For example, irinotecan and topotecan are both chemotherapeutics that cause DNA damage by inhibiting topoisomerase I [183] and share a QTL on the center of chromosome I. However, each of these chemotherapeutics also maps to distinct regions of the genome. For example, the irinotecan-response QTL on the right arm of chromosome V is not mapped for topotecan response and the topotecan-response QTL on the left arm of chromosome II is not mapped for irinotecan response. Vincristine also maps to this same region, however its mechanism of action is distinct from irinotecan and topotecan. The combination of overlapping and distinct genetic architectures underlying these highly similar compounds suggest that although some genetic variation implicated in responses to irinotecan and topotecan is shared, other QTL are specific to each compound and not representative of a general topoisomerase I inhibition mechanism. We have also observed this phenomenon of distinct genetic architectures underlying similar compounds for benzimidazole

responses [74].

2.6.2 A multi-faceted approach suggests that undetected epistatic loci impact toxin responses

To determine if we had sufficient power to experimentally validate even small-effect QTL, we constructed NILs for the three hotspots and assayed them in responses to multiple toxins. Because each principal component comprises multiple toxin-response traits, we measured NIL phenotypes for the most correlated toxin-response traits for each principal component to test recapitulation of QTL effects. For some of these tests of recapitulation for small-effect QTL, NILs showed a significant phenotypic effect. One such example is cisplatin.norm.n and cisplatin.n which represent the QTL mapped by cisplatin.PC4 that only explains 7% of the phenotypic variance. Our ability to recapitulate such a small effect suggests that our assay had sufficient power to detect small phenotypic effects in at least some cases. We postulated that our inability to recapitulate other QTL effects could be attributed to either insufficient power or additional additive or epistatic QTL that were undetected by linkage mapping. Particularly in cases where the NILs displayed transgressive phenotypes, undetected loci of opposite genotypes, acting additively or epistatically, likely caused these effects. Therefore, we investigated these interactions and found evidence for additional QTL that interact with the originally detected loci. However, we must note that whole-genome sequence data revealed that three of our NILs had a portion of the genome from the background of the starting RIAIL. Although we do not believe that these small regions are responsible for the unexpected phenotypes observed, this explanation could be a consideration for certain silver, cisplatin, carmustine, and chlorothalonil PCs, as they have significant QTL in these identified regions. This example emphasizes the importance of whole-genome sequencing NILs to verify the expected genotypes before making conclusions about phenotypic effects of a targeted QTL.

We used the results from the NIL assays to classify each test into a category that predicts a genetic model that might underlie NIL phenotypes. Categorizations were consistent across traits representing a principal component, with most of these traits falling into one or a few categorizations. This widespread consistency suggests that similar genetic architectures underlie phenotypes for these grouped traits. Furthermore, this consistency highlights the reproducibility of our high-throughput toxin response assay,

because results from independent assays (trait correlations, linkage mappings from RIAIL assays, and phenotype classifications from NIL assays) often align to support the same conclusion obtained from the individual experiments.

The majority of cases of transgressive phenotypes occurs when the N2 genotype is introgressed into the CB4856 genome. This trend might indicate allele-specific unidirectional incompatibilities between the two strains, and localizing these interactions could improve our understanding of the evolutionary processes driving such incompatibilities. However, identifying the loci that underlie these unidirectional transgressive phenotypes using a mixed-effect model or a two-factor genomic scan is difficult, because only a small number of the RIAILs have the required allelic combinations to quantify such an effect. For example, cisplatin.q90.EXT, a trait chosen to represent cisplatin PC1, fits a unidirectional intra-chromosomal model. The results of the NIL and CSS assays show that, although the CSSs seem to display no QTL effect, the NIL with the N2 genotype introgressed into the CB4856 genome displays strong hypersensitivity (**Figure 2-5D**). All of the narrow-sense heritability for cisplatin PC1 (25%) predicted by the mixed-effect model is explained by the three QTL identified through linkage mapping (the variance explained estimates of these three QTL add up to 26%). This finding suggests that most of the additive loci have been identified through linkage mapping. Therefore, the intra-chromosomal loci are likely acting epistatically to cause a unidirectional transgressive phenotype. However, using our mixed-model approach, we do not find a significant interaction component for cisplatin PC1, the principal component that is represented by cisplatin.q90.EXT. A two-dimensional genome scan for multiple loci that underlie cisplatin PC1 provides suggestive evidence for a two-QTL model over a one-QTL model, with or without interaction between the loci. These two loci are located on the left of chromosome V (outside the NIL interval) and in the center of chromosome V (inside the NIL interval) and match our empirical evidence of two intra-chromosomal loci underlying the transgressive phenotype observed (**Figure 2-7C**). Because the transgressive phenotype is unidirectional, RIAILs without the allelic combination that causes extreme phenotypes could dilute our power to detect the loci. For this reason, combining both computational models and empirical investigation facilitates the detection of loci that control transgressive phenotypes. Additionally, future studies should include even larger RIAIL panels than what we used here to empower approaches to

investigate the contributions of interactive loci.

Although we are statistically underpowered to identify some small-effect additive and interacting loci through modeling, the combination of three methods of searching for potential interactions suggests that not all fitness traits in *C. elegans* are composed of additive effects. Our two computational methods were used to identify additive and epistatic loci underlying many toxin responses, but their power was limited in cases of unidirectional transgressive phenotypes. Alternatively, the NIL and CSS phenotypic assays were able to identify unidirectional transgressive phenotypes, but they were restricted by their inability to distinguish between additive and epistatic loci. Constructing double CSS strains or multi-region NILs in which pairwise combinations of two genomic regions are introgressed within the opposite genotype could help to further define loci underlying transgressive phenotypes. However, each locus must be isolated to determine if the two loci act additively or epistatically. The results from the two-dimensional genome scan might provide insights into where to begin this approach. In cases where all three of our techniques suggested epistasis, we suspect that these QTL are not purely additive. Generating an even larger panel of recombinant strains and assaying a much larger number of biological replicates might allow us to further address the debate about how heritable loci contribute to trait variation in metazoans.

2.7 FUTURE DIRECTIONS

This project identified several traits with strong evidence of genetic interactions. Discovering both interacting loci has historically been difficult in most systems. However, our high-throughput platform in combination with the molecular and genetic tools available in *C. elegans* provides an opportunity for us to do just that. One option includes phenotyping over 600 new RILs (recombinant inbred lines) [72] in addition to the 296 RIAILs phenotyped in this study. Although time consuming, phenotyping an additional 600 RILs has two main advantages. First, we could use the phenotype data of all 896 recombinants to perform linkage mapping analysis. The addition of new breakpoints in the 600 RILs would likely result in smaller QTL intervals, making it easier to identify the causal gene or variant. Second, the increase in the number of strains might allow us to discover significant interactions computationally by performing a two-dimensional genome scan. This would provide an advantage to identify the interacting loci by narrowing down the search for the second locus to a specific

chromosomal region.

Regardless of if more RILs are phenotyped, a few traits that map to the center of chromosome V and show strong interactions could be selected for further analysis. The first step would be to identify the gene or variant underlying the QTL on chromosome V. We already have many NILs that tile across this region with varying lengths of introgressions. These NILs could be used to narrow the region of interest and CRISPR-Cas9 genome editing could be used to perform causality tests by deleting genes of interest or performing a single nucleotide variant exchange between N2 and CB4856 [72, 69, 70, 71]. Once the first locus has been isolated, the second locus can be flushed out in a similar manner. We already know whether or not the interacting locus is on chromosome V. For traits with inter-chromosomal interactions, chromosome substitution strains between N2 and CB4856 can be generated for each chromosome to identify on which chromosome the interacting locus lies. Because opposite genotypes are interacting to produce the transgressive phenotypes observed, we could perform crosses to remove the opposite genotype on the affected chromosome and look for the transgressive phenotype to disappear.

Another interesting idea to explore further is whether or not these QTL are pleiotropic (underlied by the same variant) or not pleiotropic (underlied by different, closely linked, variants). I have begun to address this question in a recent manuscript to ask if *scb-1* is a pleiotropic gene that is controlling responses to all chemotherapeutics with a QTL on the center of chromosome V [69]. I found evidence that *scb-1* was indeed a pleiotropic gene and I was able to show that deletion of *scb-1* resulted in increased sensitivity to four different chemotherapeutic drugs. It would be interesting to test the role of *scb-1* in other toxin responses in addition to chemotherapeutic drugs. This idea could also be expanded to identify pleiotropic genes that underlie the IVL and IVR QTL. It is particularly important to keep in mind that a Piwi-interacting RNA (piRNA) cluster exists within the IVR QTL interval and has been previously suggested to play a role in several benzimidazole drug responses [74]. Deletions of the piRNA-controlling gene *prg-1* already exist in the lab and could be used to implicate or exclude piRNAs from these drug responses.

2.8 CONTRIBUTIONS

Dr. Shannon Brady and I were equal partners throughout the design, experiments, analysis, and writing of this manuscript. Dr. Joshua Bloom performed the mixed-modeling approach to identify additive and epistatic loci that underlie toxin responses. Sarah Giuliani, Stephen Hippelheuser, Dr. Shannon Brady, and Dr. Mostafa Zamanian constructed NILs used in this project. Dr. Robyn Tanny created the DNA libraries for the whole-genome sequence data for these NILs and Dr. Daniel Cook analyzed this data. Tyler Shimko, Dr. Bryn Gaertner, and Sam Rosenberg assisted with the high-throughput phenotyping for QTL mapping. Dr. Erik Andersen conceptualized and supervised the entire project. Finally, I would like to thank all members of the Andersen Lab, particularly Dr. Stefan Zdraljevic and Dr. Daehan Lee, for their useful comments and insights throughout this project.

2.9 CITATION

Shared genomic regions underlie natural variation in diverse toxin responses

Kathryn S. Evans[*,1,2], Shannon C. Brady[*,1,2], Joshua S. Bloom[3,4,5], Robyn E. Tanny[1], Daniel E. Cook[1,2], Sarah E. Giuliani[1], Stephen W. Hippleheuser[1], Mostafa Zamanian[6], and Erik C. Andersen[1,7]

[1]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208
[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois 60208
[3]Department of Human Genetics, University of California, Los Angeles, California 90095
[4]Howard Hughes Medical Institute, University of California, Los Angeles, California 90095
[5]Department of Biological Chemistry, University of California, Los Angeles, California 90095
[6]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin, Madison, Wisconsin 53705
[7]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, Illinois 60611

[*]These authors contributed equally to this work

This manuscript was published in Genetics in October 2018 [73]

3 The gene *scb-1* underlies variation in *Caenorhabditis elegans* chemotherapeutic responses

3.1 PREFACE

This project was born late in my PhD career. Dr. Shannon Brady had discovered that variation in expression of the nematode-specific gene *scb-1* underlied responses to bleomycin [72]. Although we don't know the exact function of this gene, it is hypothesized that it plays a role in the stress response and/or acts as a hydrolase. Given these data combined with the knowledge that we have many chemotherapeutics that map to the same locus as bleomycin, we wanted to know if *scb-1* is specific to bleomycin response or if it controls nematode drug response generally. Around the same time, I was exploring mediation analysis, a new technique I brought to the lab, to see if we could pair overlapping gene expression QTL and drug QTL to identify high-priority candidate genes. This project is particularly special to me because I developed the mediation analysis for the lab and really pushed to see this project to completion. I was a first author on this manuscript published in *G3* in June 2020 [184], from which this chapter is based.

3.2 ABSTRACT

Pleiotropy, the concept that a single gene controls multiple distinct traits, is prevalent in most organisms and has broad implications for medicine and agriculture. The identification of the molecular mechanisms underlying pleiotropy has the power to reveal previously unknown biological connections between seemingly unrelated traits. Additionally, the discovery of pleiotropic genes increases our understanding of both genetic and phenotypic complexity by characterizing novel gene functions. Quantitative trait locus (QTL) mapping has been used to identify several pleiotropic regions in many organisms. However, gene knockout studies are needed to eliminate the possibility of tightly linked, non-pleiotropic loci. Here, we use a panel of 296 recombinant inbred advanced intercross lines of *Caenorhabditis elegans* and a high-throughput fitness assay to identify a single large-effect QTL on the center of chromosome V associated with variation in responses to eight chemotherapeutics. We

validate this QTL with near-isogenic lines and pair genome-wide gene expression data with drug response traits to perform mediation analysis, leading to the identification of a pleiotropic candidate gene, *scb-1* for some of the eight chemotherapeutics. Using deletion strains created by genome editing, we show that *scb-1*, which was previously implicated in response to bleomycin, also underlies responses to other double-strand DNA break-inducing chemotherapeutics. This finding provides new evidence for the role of *scb-1* in the nematode drug response and highlights the power of mediation analysis to identify causal genes.

3.3 INTRODUCTION

Pleiotropy refers to the well established notion that a single gene or genetic variant affects multiple distinct traits [185], and the discovery of pleiotropic genes can provide meaningful insights into the molecular mechanisms of these traits [186]. It has become easier to identify pleiotropic genes with the advent of reverse-genetic screens and quantitative trait locus (QTL) mapping [185]. For example, pleiotropic QTL for diverse growth and fitness traits have been identified in organisms such as yeast [187, 188, 189], *Arabidopsis* [190, 191, 192], *Drosophila* [193, 194], and mice [51, 195, 196]. These studies have led to important questions in the field of evolutionary genetics regarding the 'cost of complexity' [197, 198], as a single mutation might be beneficial for one trait and harmful for another [199]. Furthermore, human association studies have identified pleiotropic variants associated with different diseases [200, 201, 202], highlighting both the ubiquity and importance of certain immune-related genes and oncogenes across unrelated diseases [203, 204]. Perhaps the strongest evidence of pleiotropy exists for molecular phenotypes. Large-scale expression QTL (eQTL) mapping studies have identified single regulatory variants that control expression and likely the functions of hundreds of genes at once, opening a window into the mechanisms for how traits are controlled [115, 163, 164, 162, 143, 205].

The nematode *Caenorhabditis elegans* provides a tractable metazoan model to identify and study pleiotropic QTL [185]. A large panel of recombinant inbred advanced intercross lines (RIAILs) derived from two divergent strains, N2 and CB4856 [78, 68], has been leveraged in several linkage mapping analyses [130, 147, 133, 117, 70, 72, 71, 73, 100, 108, 35, 149, 125, 128, 115, 74, 127, 132, 112, 119,

81, 103, 102, 80, 161, 86, 85]. Quantitative genetic analysis using these panels and a high-throughput phenotyping assay [68] has facilitated the discovery of numerous QTL [74], several quantitative trait genes (QTG) [72] and quantitative trait nucleotides (QTN) [70, 71] underlying fitness-related traits in the nematode. Additionally, three pleiotropic genomic regions were recently found to influence responses to a diverse group of toxins [73]. However, overlapping genomic regions might not represent true pleiotropy but could demonstrate the co-existence of tightly linked loci [185].

Here, we use linkage mapping to identify a single overlapping QTL on chromosome V that influences the responses to eight chemotherapeutic compounds. We show that these drug-response QTL also overlap with an expression QTL hotspot that contains the gene *scb-1*, previously implicated in bleomycin response [72]. Although the exact mechanism of *scb-1* is yet unknown, it is hypothesized to act in response to stress [206] and has weak homology to a viral hydrolase [207, 208]. Together, these data suggest that the importance of *scb-1* expression might extend beyond bleomycin response. We validated the QTL using near-isogenic lines (NILs) and performed mediation analysis to predict that *scb-1* expression explains the observed QTL for four of the eight drugs. Finally, we directly tested the effect of *scb-1* loss of function on chemotherapeutic responses. We discovered that expression of *scb-1* underlies differential responses to several chemotherapeutics that cause double-strand DNA breaks, not just bleomycin. This discovery of pleiotropy helps to further define the role of *scb-1* by expanding its known functions and provides insights into the molecular mechanisms underlying the nematode drug response.

3.4 METHODS

3.4.1 Strains

Animals were grown at 20℃ on modified nematode growth media (NGMA) containing 1% agar and 0.7% agarose to prevent burrowing and fed OP50 [120]. The two parental strains, the canonical laboratory strain, N2, and the wild isolate from Hawaii, CB4856, were used to generate all recombinant lines. 208 recombinant inbred advanced intercross lines (RIAILs) generated previously by Rockman *et al.* [78] (set 1 RIAILs) were phenotyped for expression QTL mapping (detailed below). A second set of

296 RIAILs generated previously by Andersen *et al.* [68] (set 2 RIAILs) was used more extensively for drug phenotyping and linkage mapping. The set 2 RIAILs were used for linkage mapping because they addressed the three main disadvantages of the set 1 RIAILs detailed previously [68], namely a structured population, the laboratory-derived variant in *npr-1* [58], and the *peel-1; zeel-1* incompatibility [86, 85]. Because of these limitations, the set 2 RIAILs were generated using QX1430 and CB4856. QX1430 is from the N2 strain background but contains a transposon insertion in *peel-1* and the CB4856 *npr-1* allele introgressed on chromosome X [68]. Near-isogenic lines (NILs) were generated by backcrossing a selected RIAIL for several generations to the parent strain (N2 or CB4856) [72] using PCR amplicons for insertion-deletion (indels) variants to track the introgressed region. NILs were whole-genome sequenced to verify introgressions were only in the targeted genomic intervals. CRISPR-Cas9-mediated deletions of *scb-1* were described previously [72]. All strains are available upon request or from the *C. elegans* Natural Diversity Resource [63].

3.4.2 High-throughput fitness assays for linkage mapping

For dose responses and RIAIL phenotyping, we used a high-throughput fitness assay described previously [68]. In summary, populations of each strain were passaged and amplified on NGMA plates for four generations. In the fifth generation, gravid adults were bleach-synchronized and 25-50 embryos from each strain were aliquoted into 96-well microtiter plates at a final volume of 50 μL K medium [173]. The following day, arrested L1s were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA; [174]) at a final concentration of 5 mg/mL in K medium and were grown to the L4 larval stage for 48 hours at 20℃ with constant shaking. Three L4 larvae were sorted into new 96-well microtiter plates containing 10 mg/mL HB101 bacterial lysate, 50 μM kanamycin, and either diluent (1% water or 1% DMSO) or drug dissolved in the diluent using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica; Holliston, MA). Sorted animals were grown for 96 hours at 20℃ with constant shaking. The next generation of animals and the parents were treated with sodium azide (50 mM in 1X M9) to straighten their bodies for more accurate length measurements. Animal length (median.TOF), optical density integrated over animal length (median.EXT), and brood size (norm.n) were quantified for each well using the COPAS BIOSORT.

Nematodes get longer (animal length) and become thicker and more complex (optical density) over developmental time. Because animal length and optical density are highly correlated, we calculated a fourth trait (median.norm.EXT) that normalizes optical density by animal length (median.EXT / median.TOF). Phenotypic measurements collected by the BIOSORT were processed and analyzed using the R package *easysorter* [77] as described previously [72]. Differences among strains within the control conditions were controlled by subtracting the mean control-condition value from each drug-condition replicate for each strain using a linear model (drug_phenotype $\sim$ mean_control_phenotype). In this way, we are addressing only the differences among strains that were caused by the drug condition and the variance in the control condition does not affect the variance in the drug condition. For plotting purposes, these residual values (negative and positive residuals) were normalized from 0 to 1 where 0 refers to the smallest residual phenotypic value in that condition and 1 refers to the largest.

3.4.3 Dose-response assays

Four genetically divergent strains (N2, CB4856, JU258, and DL238) were treated with increasing concentrations of each of the eight drugs using the high-throughput fitness assay described above. The dose of each drug that provided a reproducible drug-specific effect that maximizes between-strain variation while minimizing within-strain variation across the four traits was selected for the linkage mapping experiments. The chosen concentrations are as follows: 100 μM amsacrine hydrochloride (Fisher Scientific, #A277720MG) in DMSO, 50 μM bleomycin sulfate (Fisher, #50-148-546) in water, 2 μM bortezomib (VWR, #AAJ60378-MA) in DMSO, 250 μM carmustine (Sigma, #1096724-75MG) in DMSO, 500 μM cisplatin (Sigma, #479306-1G) in K media, 500 μM etoposide (Sigma, #E1383) in DMSO, 500 μM puromycin dihydrochloride (VWR, #62111-170) in water, and 150 μM silver nitrate (Sigma-Aldrich, #S6506-5G) in water.

3.4.4 Linkage mapping

Set 1 and set 2 RIAILs were phenotyped in each of the eight drugs and controls using the high-throughput fitness assay described above. Linkage mapping was performed on each of the drug

and gene expression traits using the R package linkagemapping as described previously [72] (https://github.com/AndersenLab/linkagemapping). The cross object derived from the whole-genome sequencing of the RIAILs containing 13,003 SNPs was loaded using the function *load_cross_obj("N2xCB4856cross_full")*. The RIAIL phenotypes were merged into the cross object using the *merge_pheno* function with the argument *set = 1* for expression QTL mapping and *set = 2* for drug phenotype mapping. A forward search (*fsearch* function) adapted from the R/*qtl* package [176] was used to calculate the logarithm of the odds (LOD) scores for each genetic marker and each trait as $-n(ln(1 - R^2)/2ln(10))$ where R is the Pearson correlation coefficient between the RIAIL genotypes at the marker and trait phenotypes [27]. A 5% genome-wide error rate was calculated by permuting the RIAIL phenotypes 1000 times. The marker with the highest LOD score above the significance threshold was selected as the QTL then integrated into the model as a cofactor and mapping was repeated iteratively until no further significant QTL were identified. Finally, the annotate_lods function was used to calculate the effect size of each QTL and determine 95% confidence intervals defined by a 1.5 LOD drop from the peak marker using the argument *cutoff = proximal*.

3.4.5 Modified high-throughput fitness assay for NIL validation

NILs and *scb-1* deletion strains were tested using a modified version of the high-throughput fitness assay detailed above. Strains were propagated for two generations, bleach-synchronized in three independent replicates, and titered at a concentration of 25-50 embryos per well of a 96-well microtiter plate. The following day, arrested L1s were fed HB101 bacterial lysate at a final concentration of 5 mg/mL with either diluent or drug. After 48 hours of growth at 20℃ with constant shaking, nematodes were treated with sodium azide (5 mM in water) prior to analysis of animal length and optical density using the COPAS BIOSORT. As only one generation of growth is observed, brood size was not calculated. A single trait (median.EXT) was chosen to represent animal growth generally, as the trait is defined by integrating optical density over length. Because of the modified timing of the drug delivery, lower drug concentrations were needed to recapitulate the previously observed phenotypic effect. The selected doses are as follows: 12.5 μM amsacrine in DMSO, 12.5 μM bleomycin in water, 2 μM

bortezomib in DMSO, 250 μM carmustine in DMSO, 125 μM cisplatin in K media, 62.5 μM etoposide in DMSO, 300 μM puromycin in water, and 100 μM silver in water.

3.4.6 Expression QTL analysis

Microarray data for gene expression using 15,888 probes were previously collected from synchronized young adult populations of 208 set 1 RIAILs [115]. Expression data were corrected for dye effects and probes with variants were removed [100]. Linkage mapping was performed as described above for the remaining 14,107 probes, and a significance threshold was determined using a permutation-based False Discovery Rate (FDR). FDR was calculated as the ratio of the average number of genes across 10 permutations expected by chance to show a maximum LOD score greater than a particular threshold versus the number of genes observed in the real data with a maximum LOD score greater than that threshold. We calculated the FDR for a range of thresholds from 2 to 10, with increasing steps of 0.01, and set the threshold so that the calculated FDR was less than 5%.

Local eQTL were defined as linkages whose peak LOD scores were within 1 Mb of the starting position of the probe [115]. eQTL hotspots were identified by dividing the genome into 5 cM bins and counting the number of distant eQTL that mapped to each bin. Significance was determined as bins with more eQTL than the Bonferroni-corrected 99th percentile of a Poisson distribution with a mean of 3.91 QTL (total QTL / total bins) [209, 73, 115]. We identified nine eQTL hotspots (II, IVL, IVC, IVR, VL, VC, VR, XL, and XC). To avoid false positives, we increased the LOD threshold for QTL to be counted in the hotspot analysis to a LOD > 5 or LOD > 6. At a LOD > 5, six of the nine eQTL hotspots persist (IVL, IVR, VC, VR, XL, and XC), and at a LOD > 6, three persist (IVL, IVR, and XL). We further looked for spurious eQTL hotspots in ten permuted datasets. At a LOD > 5, we identified four hotspots, and at a LOD > 6, we identified one hotspot.

3.4.7 Mediation analysis

A total of 159 set 1 RIAILs were phenotyped in each of the eight drugs and controls using the standard high-throughput fitness assay described above. Mediation scores were calculated with bootstrapping using the mediate function from the mediation R package (version 4.4.7) [210] for each QTL identified

from the set 1 RIAILs and all 49 probes (including *scb-1*, A_12_P104350) that mapped to the chromosome V eQTL hotspot using the following models:

$$Mediator\_model : lm(expression \sim genotype)$$

$$Outcome\_model : lm(phenotype \sim expression + genotype)$$

The output of the mediate function can be summarized as follows: the total effect of genotype on phenotype, ignoring expression (*tau.coef*); the direct effect of genotype on phenotype, while holding expression constant (*z0*); the estimated effect of expression on phenotype (*d0*); the proportion of the total effect that can be explained by expression data (*n0*). This mediation proportion (n0) can be a useful way to identify the impact of gene expression on the overall phenotype. However, cases of inconsistent mediation (where the direct effect is either smaller than or in the opposite direction of the indirect mediation effect) render this measurement uninterpretable with values greater than one or less than zero [211]. We used the estimated effect of expression on phenotype (*z0*) as the final mediation score for this reason. Because the effect size can be positive or negative, mediation scores range from -1 to 1, and we evaluated the absolute value of mediation estimates to compare across traits. Each mediation estimate generated a *p*-value, indicating confidence in the estimate, derived from bootstrapping with 1000 simulations. The likelihood of *scb-1* mediating a given QTL effect was calculated relative to the other 48 probes with an eQTL in the region. Traits in which *scb-1* was at or above the 90th percentile of this distribution were prioritized over other traits.

3.4.8 Statistical analysis

Broad-sense heritability was calculated from the dose response phenotypes using the *lmer* function in the *lme4* R package [212] with the formula $phenotype \sim 1 + (1|strain)$ for each dose. For the NIL and *scb-1* deletion high-throughput assays, statistical significance of phenotypic differences between each strain pair was tested using the *TukeyHSD* function [175] on an ANOVA model with the formula $phenotype \sim strain$ to assess differences between strains in the control-regressed phenotype data.

3.5 RESULTS

3.5.1 Natural variation on chromosome V underlies differences in responses to several chemotherapeutics

We measured *C. elegans* development and chemotherapeutic sensitivity as a function of animal length (median.TOF), optical density (median.EXT), and brood size (norm.n) with a high-throughput assay developed using the COPAS BIOSORT (see Methods) [70, 72, 73, 71, 68]. Animal length and optical density (animal thickness and composition) are both measures of nematode development, and brood size is a measure of nematode reproduction [68]. Because optical density is calculated as a function of length and these traits are related, a fourth trait that captures the optical density normalized by length (median.norm.EXT) was also included. We exposed four genetically divergent strains (N2, CB4856, JU258, and DL238) to increasing doses of eight chemotherapeutic compounds. Five of these compounds (bleomycin, carmustine, etoposide, amsacrine, and cisplatin) are known to cause double-strand DNA breaks and/or inhibit DNA synthesis [213, 214, 215, 216, 217]. The remaining three compounds either inhibit protein synthesis (puromycin) [218], inhibit the proteosome and subsequent protein degradation (bortezomib) [219], or cause cellular toxicity in a poorly defined way (silver nitrate) [220] (**Table 3-1**). In the presence of each drug, nematodes were generally shorter, less optically dense, and produced smaller broods compared to non-treated nematodes. We observed significant phenotypic variation among strains and identified a substantial heritable genetic component for most traits (average $H^2 = 0.52 \pm 0.53$).

We exposed a panel of 296 RIAILs (set 2 RIAILs, see Methods) to all eight chemotherapeutics at a selected concentration that both maximizes among-strain and minimizes within-strain phenotypic variation. Linkage mapping for all four traits for each of the eight drugs (total of 32 traits) identified 79 QTL from 31 traits (one trait had no significant QTL), several of which have been identified previously [72, 73, 71]. Strikingly, a QTL on the center of chromosome V was linked to variation in responses to all eight compounds (**Figure 3-1**). In all cases, the CB4856 allele on chromosome V is associated with greater resistance to the drug than the N2 allele. We previously identified this genomic interval as a QTL hotspot, defined as a region heavily enriched for toxin-response QTL [73]. Because several of the

Table 3-1: Main mechanism of action for eight chemotherapeutic drugs

Drug	Drug class	Mechanism of action
Amsacrine	Topoisomerase inhibitors	DNA intercalation and inhibition of topoisomerase II, causing DNA double-strand breaks, cell cycle arrest, and cell death
Bleomycin	Antitumor antibiotic	Forms complexes with iron that reduce molecular oxygen to form free radicals which in turn cause DNA single- and double-strand breaks
Bortezomib	Proteosome inhibitors	Reversibly inhibits the 26S proteosome and inhibits nuclear factor (NF)-kappaB causing disruption of various cell signaling pathways, cell cycle arrest, and cell death.
Carmustine	Alkylating agents	Alkylates and cross-links DNA causing cell cycle arrest and cell death
Cisplatin	Alkylating agents	Alkylates and cross-links DNA causing cell cycle arrest and cell death
Etoposide	Topoisomerase inhibitors	Binds to and inhibits topoisomerase II causing an increase of DNA single- and double-strand breaks, cell cycle arrest, and cell death
Puromycin	Aminonucleoside antibiotic	Acts as analog of 3' terminal end of aminoacyl-tRNA and incorporates itself into growing polypeptide chain causing premature termination and inhibition of protein synthesis
Silver	*NA*	Multi-faceted induction of apoptosis

chemotherapeutics share a similar mechanism of action, a single pleiotropic gene might underlie the observed QTL for multiple drugs.

In order to isolate and validate the effect of this QTL, we constructed reciprocal near-isogenic lines (NILs) by introgressing a genomic region on chromosome V from the resistant CB4856 strain into the sensitive N2 background and vice versa. We used a modified high-throughput assay (see Methods) to measure length and optical density of a population of animals grown in the presence of the drug for 48 hours (from larval stages L1 to L4). In this modified assay, less drug was required to observe the same phenotypic effect as before. Statistical significance was calculated in a pairwise manner for each strain (see Methods). For all eight chemotherapeutics tested, the strain with the N2 introgression was significantly more sensitive than its CB4856 parent and/or the strain with the CB4856 introgression was significantly more resistant than its N2 parent (**Figure 3-2**). These data confirm that one or more genetic variant(s) within this region on chromosome V cause increased drug sensitivities in N2.

3.5.2 Expression QTL mapping identifies a hotspot on the center of chromosome V

Genetic variation can affect a phenotype most commonly through either modifications of the amino acid sequence that lead to altered protein function (or even loss of function) or changes in the expression

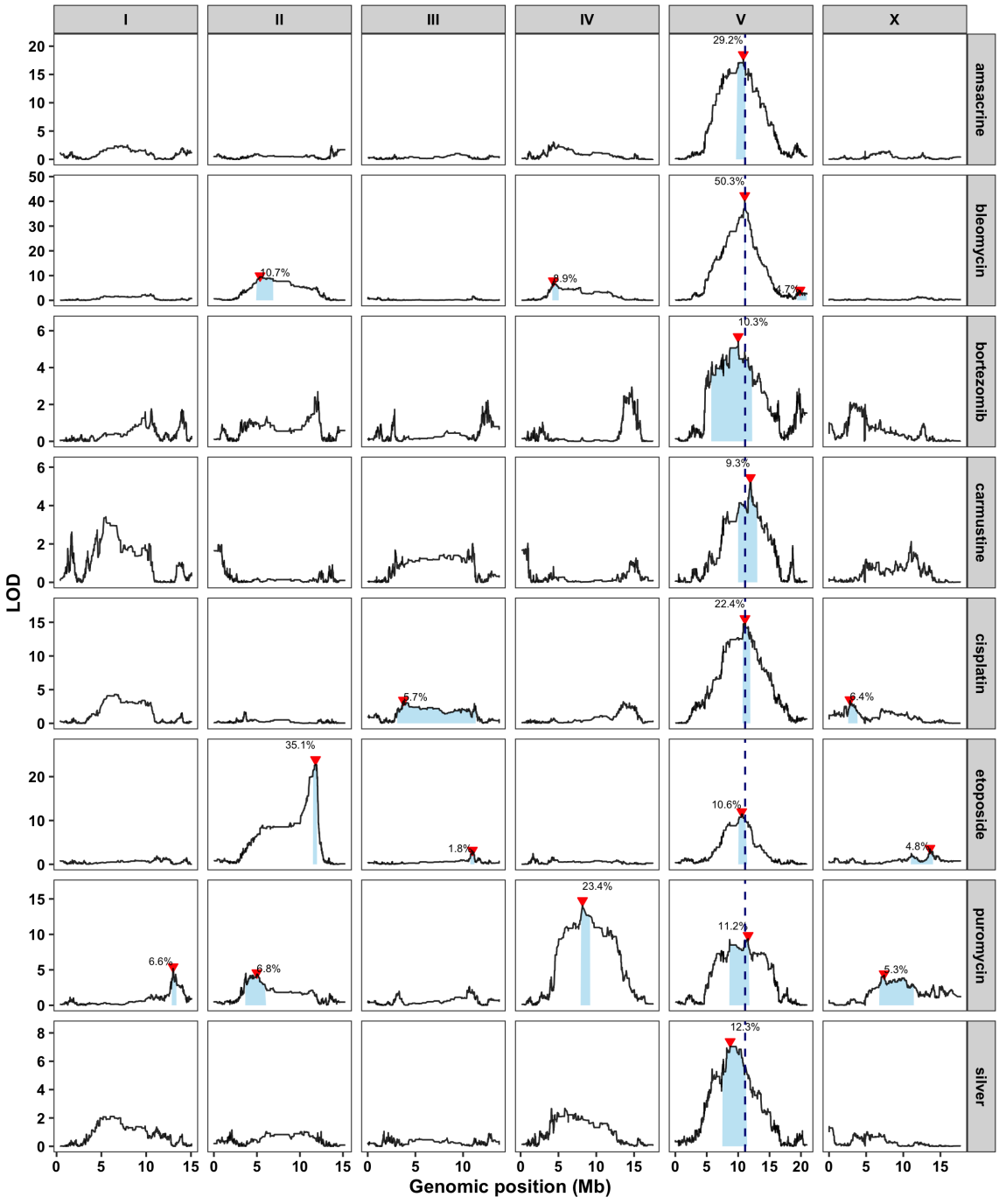


Figure 3-1: A large-effect QTL on the center of chromosome V underlies responses to several chemotherapeutics. Linkage mapping results with the set 2 RIAILs for a representative trait for each drug are shown (amascrine: median.norm.EXT, bleomycin median.TOF:, bortezomib: median.TOF, carmustine: norm.n, cisplatin: median.TOF, etoposide: norm.n, puromycin: median.TOF, silver: median.TOF). Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. The dotted vertical line represents the genomic position of *scb-1*.

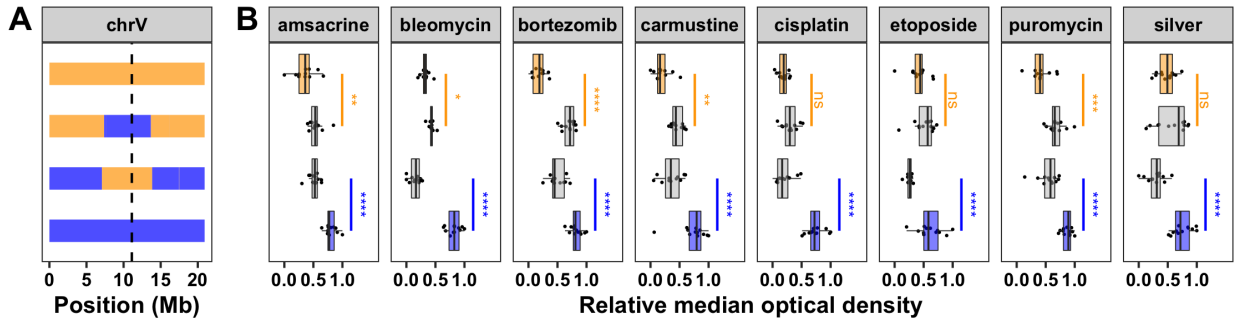


Figure 3-2: Near-isogenic lines validate the chromosome V QTL. (A) NIL genotypes on chromosome V are shown, colored orange (N2) and blue (CB4856). From top to bottom, strains are N2, ECA232, ECA1114, and CB4856. The dotted vertical line represents the location of *scb-1*. **(B)** NIL phenotypes in eight chemotherapeutics (12.5 μM amsacrine, 12.5 μM bleomycin, 2 μM bortezomib, 250 μM carmustine, 125 μM cisplatin, 62.5 μM etoposide, 300 μM puromycin, and 100 μM silver) are plotted as Tukey box plots with strain (y-axis) by relative median optical density (median.EXT, x-axis). Statistical significance was calculated for each strain pair. Significance of each strain compared to its parental strain (ECA232 to N2 and ECA1114 to CB4856) is shown above each strain pair and colored by the parent strain against which it was tested (ns = non-significant (p-value > 0.05); *, **, ***, and **** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

level of the protein. In the latter case, measuring the intermediate phenotype (gene expression) can be useful in elucidating the mechanism by which genetic variation causes phenotypic variation. More specifically, cases with overlap between expression QTL (eQTL) and drug-response QTL suggest that a common variant could underlie both traits and provide evidence in support of causality for the candidate gene in question [221, 222].

To identify such cases of overlap between expression QTL and the drug-response QTL on chromosome V, we need genome-wide expression data for the RIAILs. In a previous study, expression of 15,888 probes were measured using microarrays for a panel of 208 RIAILs (set 1 RIAILs, see Methods) between N2 and CB4856 [78]. This study used the variation in gene expression as a phenotypic trait to identify eQTL using linkage mapping with 1,455 variants [115]. They identified 2,309 eQTL and three regions with significantly clustered distant eQTL (eQTL hotspots), suggesting that these regions are pleiotropic, wherein one or more variant(s) are affecting expression of multiple genes. We recently performed whole-genome sequencing for these strains and identified 13,003 informative variants [72]. Using this new set of variants, we re-analyzed the eQTL mapping by performing linkage mapping analysis for a selected 14,107 of the 15,888 probes without genetic variation in CB4856 [100]. We identified 2,540 eQTL associated with variation in expression of 2,196 genes (**Figure 3-3A**). These eQTL have relatively large effect sizes compared to the drug-response

QTL. On average, each eQTL explains 23% of the phenotypic variance in gene expression among the RIAIL population. Half of the eQTL (50.2%; 1,276) mapped within 1 Mb of the gene whose expression was measured and were classified as local (see Methods) [205]. The other half (49.7%; 1,264) were found distant from their respective gene, and over a third (37%; 940) were found on different chromosomes entirely. In general, eQTL effect sizes increased, max LOD scores decreased, and confidence intervals became smaller compared to the original mapping results. These differences and the additional eQTL observed between this analysis and the original are possibly caused by the integration of new genetic markers. Additionally, we found several differences in methodology between the current approach and the previous one. These differences include ignoring the population structure of the set 1 RIAILs, adding the forward-search marker-regression linkage mapping, and altering the linkage mapping method itself (see Methods, [115]).

We noticed regions of the genome that appeared to be enriched for distant eQTL. We identified eQTL hotspots in a similar manner to the previous study (see Methods) and found a total of nine eQTL hotspots (**Figure 3-3B**). Six of the nine eQTL hotspots withstood more stringent filtering methods (see Methods), and three (left of chromosome IV, right of chromosome IV, and left of chromosome X) were the most significant. These three hotspots also overlap with the most significant eQTL hotspots in the previous study [115]. Notably, three of the eQTL hotspots (center of chromosome IV, right of chromosome IV, and center of chromosome V) overlap with the previously identified drug-response QTL hotspots on chromosomes IV and V (**Figure 3-3B**) [73]. The overlap of these eQTL and drug-response QTL hotspots could provide strong candidate genes whose expression underlies the differences in nematode drug responses generally. Expression of one gene of interest, *scb-1*, has been previously implicated in response to bleomycin [72] and resides within the eQTL hotspot region on the center of chromosome V. Although the exact mechanism of how *scb-1* responds to bleomycin is unknown, its putative hydrolase activity [207, 208, 72] suggests that it might act to break down chemotherapeutic compounds. These data suggest that variation in expression of *scb-1* and responses to these eight chemotherapeutics (including bleomycin) could be mechanistically linked through the metabolic breakdown of chemotherapeutic drugs.

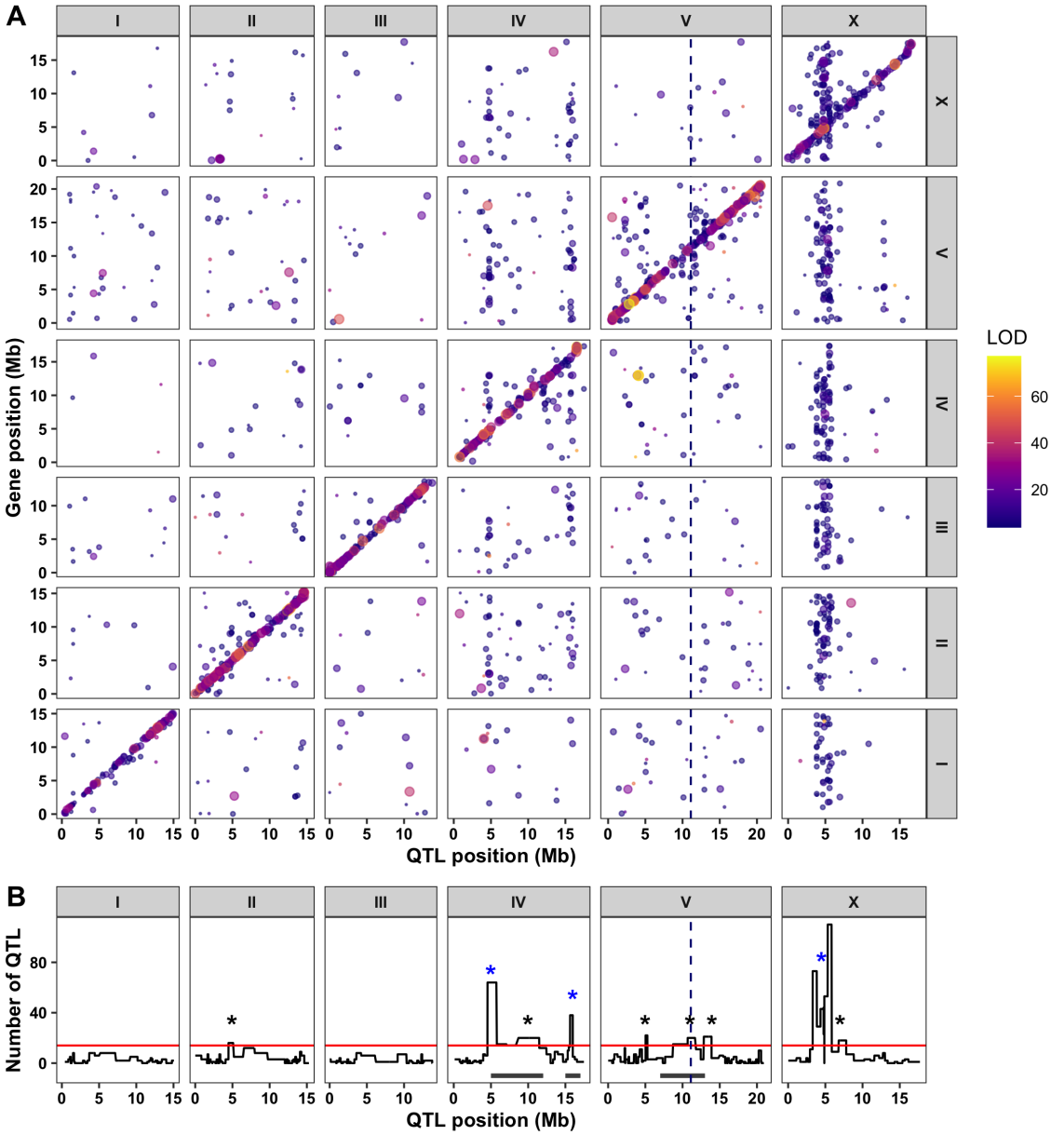


Figure 3-3: Expression QTL mapping identifies several hotspots. (A) The genomic locations of the eQTL peaks derived from linkage mapping using the set 1 RIAILs (x-axis) are plotted against the genomic locations of the probe (y-axis). The size of the point corresponds to the effect size of the QTL. eQTL are colored by the LOD score, increasing from purple to pink to yellow. The diagonal band represents local eQTL, and vertical bands represent eQTL hotspots. **(B)** Quantification of eQTL hotspots identified by overlapping distant eQTL. The number of distant eQTL (y-axis) in each 5 cM bin across the genome (x-axis) is shown. Bins above the red line are significant and marked with an asterisk. The bins with the blue asterisks are most significant and have been identified in a previous analysis. The dotted vertical line represents the genomic position of *scb-1*. Grey rectangles below the plot represent locations of the drug-response QTL hotspots previously identified.

3.5.3 Mediation analysis suggests that *scb-1* expression plays a role in responses to several chemotherapeutics

Mediation analysis seeks to explain the relationship between an independent and a dependent variable by including a third intermediary variable. We can use mediation analysis to understand how certain genetic variants on chromosome V (independent variable) affect drug responses (dependent variable) through differential gene expression of genes within the eQTL hotspot (mediator variable) (**Figure 3-4**). We measured brood size, animal length, and optical density in response to all eight chemotherapeutics in the set 1 RIAILs and performed linkage mapping for these traits. Although the power to detect QTL with these strains is lower than in our original mapping set (set 2 RIAILs; see Methods) [68], we still identified overlapping QTL on chromosome V for half of the drugs tested (bleomycin, cisplatin, silver, and amsacrine) (**Figure 3-5**).

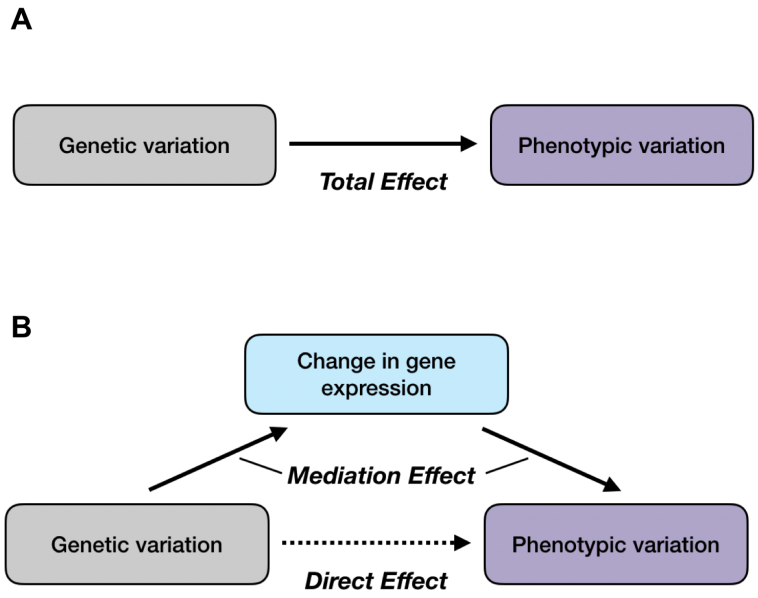


Figure 3-4: Model for gene expression as a mediator of a drug-response QTL. (A) The effect size of a QTL is calculated as the total effect a genetic variant has on the phenotypic variation of the population, regardless of the causal association of the genetic variant. **(B)** Mediation analysis tests the hypothesis that a genetic variant does not directly lead to phenotypic variation but rather causes a change in gene expression that further drives the change in phenotype observed. The "Direct Effect" can be calculated by including gene expression as a cofactor in the linear model between genotype and phenotype and extracting the partial coefficient of the effect of genotype on phenotype. The "Mediation Effect" can be calculated by subtracting the "Total Effect" - "Direct Effect". Mediation estimates are then calculated as the proportion of the "Total Effect" that can be explained by the "Mediation Effect" ("Mediation Effect" / "Total Effect").

We calculated the effect that variation in expression of *scb-1* had on drug-response traits compared

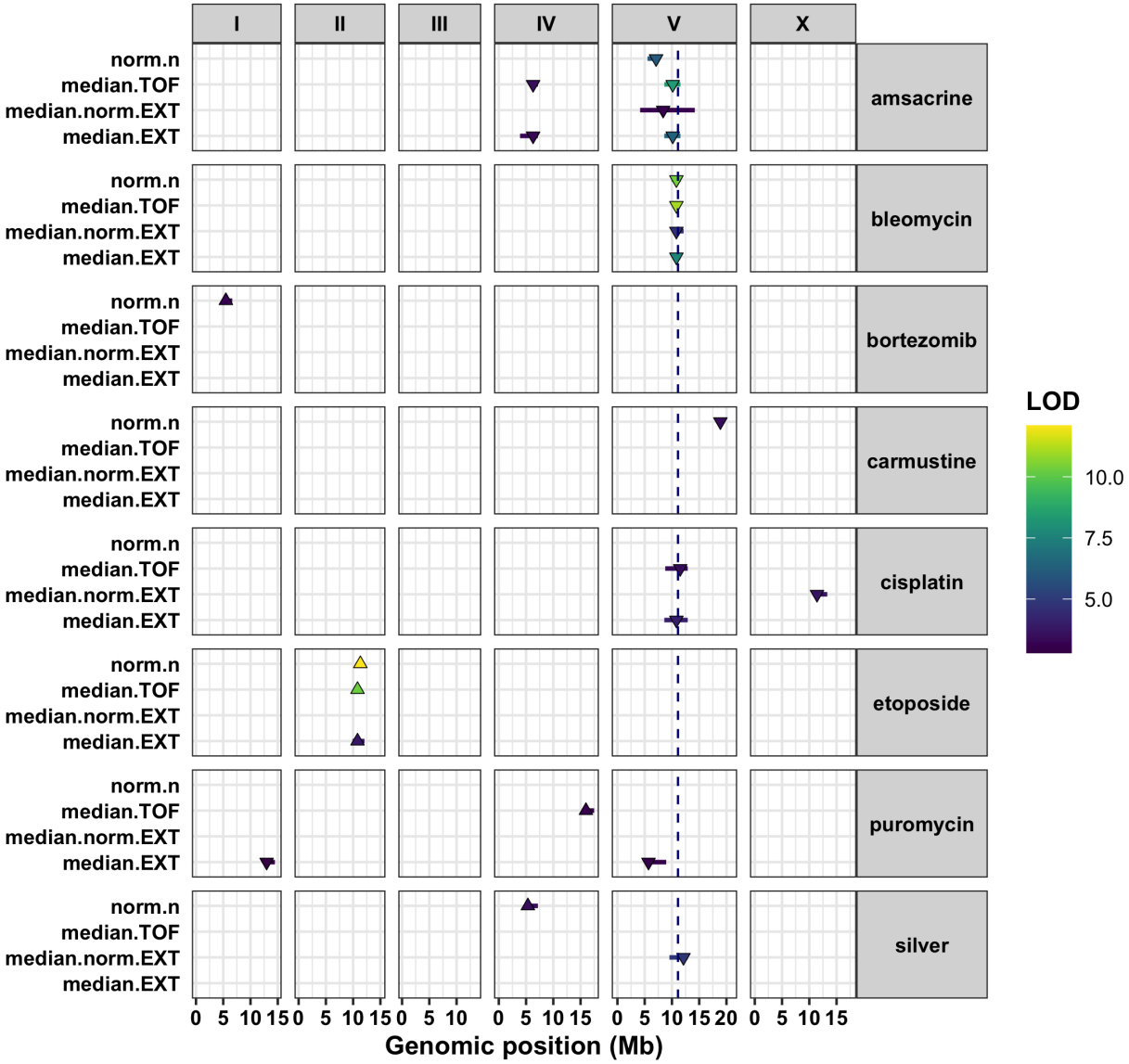


Figure 3-5: Linkage mapping summary for drug-response traits in the set 1 RIAILs. Genomic positions (x-axis) of all QTL identified from linkage mapping are shown for each drug-trait (y-axis). Each QTL is plotted as a triangle at the genomic location of the peak marker and a line that represents the 95% confidence interval. QTL with right side up triangles have a negative effect size (N2 allele is resistant), and QTL with upside down triangles have a positive effect size (CB4856 allele is resistant). QTL are colored by the logarithm of the odds (LOD) score, increasing in significance from purple to green to yellow. The dotted vertical line represents the genomic position of *scb-1*.

to the other 48 genes with an eQTL in the chromosome V eQTL hotspot using mediation analysis (see Methods). We estimated that the effect of expression variation of *scb-1* on bleomycin response is 0.65 (set 1 RIAILs, **Figure 3-6**). Moreover, out of all 49 genes with an eQTL in the region, *scb-1* was a clear mediation score outlier. All of the remaining three chemotherapeutics with a QTL on the center of

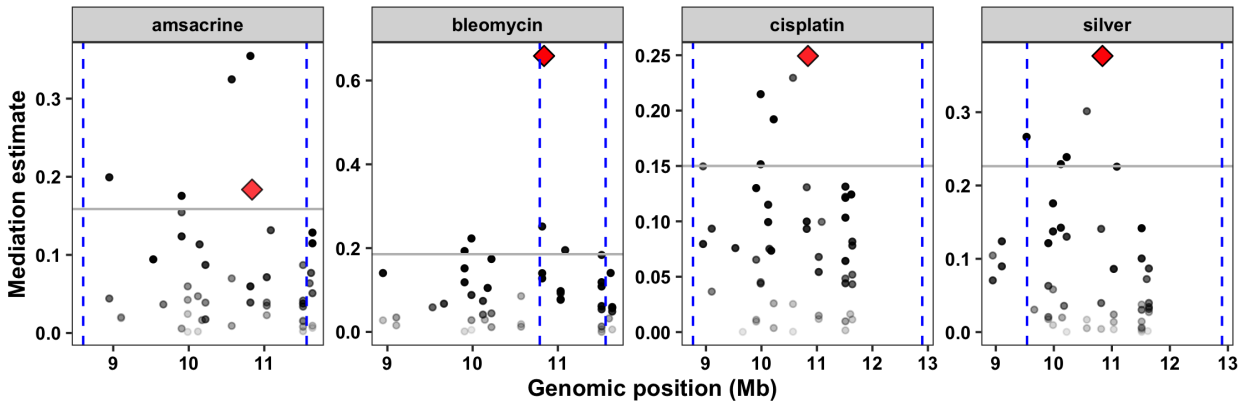


Figure 3-6: Mediation analysis for the eQTL hotspot on the center of chromosome V. Mediation estimates calculated as the indirect effect that differences in expression of each gene plays in the overall phenotype (y-axis) are plotted against the genomic position of the eQTL (x-axis) on chromosome V for 49 probes (including *scb-1* (red diamond)) that map to the chromosome V eQTL hotspot (set 1 RIAILs). A representative trait for each drug from the set 1 linkage mapping analysis are shown: amascrine (median.EXT), bleomycin (median.EXT), cisplatin (median.TOF), and silver (median.norm.EXT). The 90th percentile of the distribution of mediation estimates for each trait are represented by the horizontal grey lines. The confidence intervals for the QTL (set 1 RIAILs) are shown with the vertical blue dotted lines. The confidence of the estimate increases (*p*-value decreases) as points become less transparent.

chromosome V in the set 1 RIAIL mapping showed moderate evidence of *scb-1* mediation, with *scb-1* falling well above the 90th percentile of mediation estimates for all genes with an eQTL in this region (**Figure 3-6**). We further performed this mediation analysis on all 32 drug-response traits, regardless of the presence of a QTL in the set 1 RIAIL panel. Etoposide and puromycin also showed evidence of *scb-1* mediation. This in silico approach indicated that expression of *scb-1* might be an intermediate link between genetic variation on chromosome V and responses to several of the chemotherapeutic drugs tested.

3.5.4 Expression of *scb-1* affects responses to several chemotherapeutics that cause double-strand DNA breaks

To empirically test whether *scb-1* expression modulates the chromosome V QTL effect for each drug, we used the modified high-throughput assay (see Methods) to expose two independently derived strains with *scb-1* deletions [72] to each chemotherapeutic (**Figure 3-7**). Statistical significance was calculated in a pairwise manner for each strain (see Methods). Because RIAILs with the CB4856 allele on chromosome V express higher levels of *scb-1* than RIAILs with the N2 allele, we expect that loss of

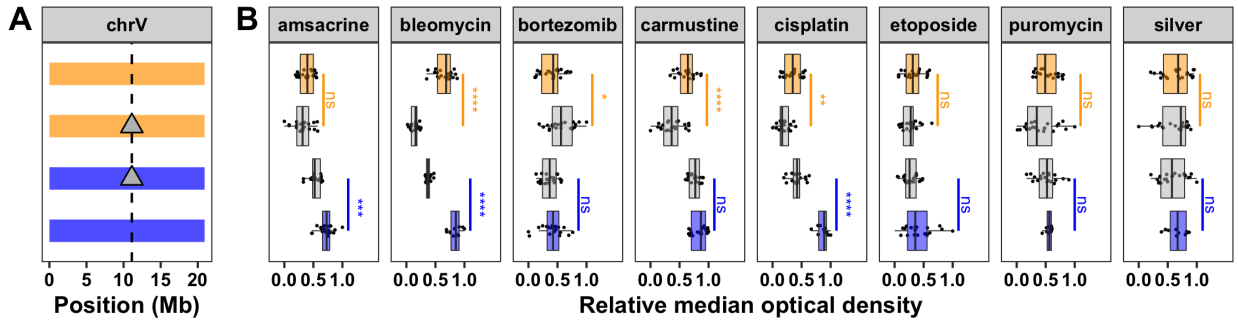


Figure 3-7: Testing the role of *scb-1* in drug responses. **(A)** Strain genotypes on chromosome V are shown, colored orange (N2) and blue (CB4856). From top to bottom, strains are N2, ECA1132, ECA1134, and CB4856. Deletion of *scb-1* is indicated by a grey triangle. The dotted vertical line represents the location of *scb-1*. **(B)** Phenotypes of strains in eight chemotherapeutics (12.5 μM amsacrine, 12.5 μM bleomycin, 2 μM bortezomib, 250 μM carmustine, 125 μM cisplatin, 62.5 μM etoposide, 300 μM puromycin, and 100 μM silver) are plotted as Tukey box plots with strain (y-axis) by relative median optical density (median.EXT, x-axis). Statistical significance was calculated for each strain pair. Significance of each deletion strain compared to its parental strain (ECA1132 to N2 and ECA1134 to CB4856) is shown above each strain pair and colored by the parent strain against which it was tested (ns = non-significant (p-value > 0.05); *, **, ***, and **** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

scb-1 will cause increased drug sensitivity in the CB4856 background but might not have an effect in the N2 background. We validated the results of Brady *et al.* and confirmed that ablated *scb-1* expression causes hyper-sensitivity to bleomycin in both N2 and CB4856 (**Figure 3-7**). We also observed similarly increased sensitivity to cisplatin with *scb-1* deletions in both backgrounds. Furthermore, removing *scb-1* shows moderately increased sensitivity in the CB4856 background for amsacrine and in the N2 background for carmustine. The remaining four drugs did not show a significantly different phenotype between the parental N2 and CB4856 strains, suggesting these traits might be less reproducible or that *scb-1* variation does not underlie these drug differences. Overall, these results provide evidence for the pleiotropic effect of *scb-1*, which appears to mediate responses to at least four of the eight chemotherapeutic drugs.

3.6 DISCUSSION

In this study, we identified overlapping QTL on the center of chromosome V that influence sensitivities to eight chemotherapeutic drugs. Because five of these drugs are known to cause double-strand DNA breaks, we hypothesized that this genomic region might be pleiotropic – a single shared genetic variant affects the responses to each drug. Because this variant might affect drug responses by

regulating gene expression levels, we looked for the co-existence of drug-response QTL and expression QTL on chromosome V. We identified 2,540 eQTL and nine eQTL hotspots, including a region on the center of chromosome V. We calculated the mediation effect of all 49 genes with an eQTL that maps to this hotspot region and identified *scb-1* as a candidate gene whose expression influences the responses to several chemotherapeutics. We used CRISPR-Cas9-mediated *scb-1* deletion strains to empirically validate the role of *scb-1* in the chemotherapeutic response. In addition to bleomycin [72], we discovered that responses to cisplatin, amsacrine, and carmustine are affected by *scb-1* expression. In this study, we found evidence that several overlapping QTL are representative of pleiotropy at the gene level and further elucidated the function of *scb-1* as a potential response to double-strand DNA break stress.

3.6.1 Mediation of drug-response QTL using gene expression to identify causal genes

Mediation analysis often suggests potential candidate genes that underlie different traits [222, 221] and could be applied to drug responses. Using *C. elegans* strains and high-throughput assays, we can rapidly validate hypotheses generated by mediation analysis. Three of the eight chemotherapeutics that map to an overlapping drug-response QTL and were potentially mediated by *scb-1* were validated using targeted deletion strains.

Although mediation analysis provided moderate evidence that expression of *scb-1* could also play a role in sensitivity to etoposide and puromycin, we observed no experimental evidence of this relationship. Additionally, we have evidence that expression of *scb-1* might mediate response to carmustine. However, mediation analysis disagrees. The discrepancy between the mediation analysis and validation of causality using targeted deletion strains could be partially explained by one of several possibilities. First, different traits were measured in each experiment. The mediation analysis used traits measured over 96 hours of growth in drug conditions spanning two generations, but the causality test used traits measured over 48 hours of growth in drug conditions within one generation. Second, the precision of our mediation estimates was likely reduced by the poor quality drug traits for the set 1 RIAIL panel [68]. Indeed, bortezomib, carmustine, etoposide, and puromycin did not map to the center of chromosome V using the set 1 RIAILs (**Figure 3-5**). Expression data for the set 2 RIAIL panel would

likely generate more accurate mediation estimates, especially if data were collected using RNA sequencing to avoid the inherent reference bias of microarray data [223]. Third, our mediation analysis was performed using expression data collected in control conditions and phenotype data collected in drug conditions. This analysis will only provide evidence of mediation if the baseline expression differences affect an individual's response to the drug. Collecting expression data from drug-treated nematodes could help us learn more about how gene expression varies in response to treatment with the chemotherapeutic. Finally, as we only directly assessed the complete loss of *scb-1* in drug sensitivity, it is still possible that reduction of function (or change in function) caused by a single nucleotide variant or other structural variation in CB4856 could validate the role of *scb-1* in responses to these drugs.

This study demonstrates the power of pairing genome-wide linkage mapping of gene expression and drug response data using simple colocalization as well as more complex mediation analysis techniques. In addition to providing a resource for candidate gene prioritization within a QTL interval, mediation analysis can help to identify the mechanism by which genetic variation causes phenotypic differences. This type of approach could be even more powerful using genome-wide association (GWA) where the lower linkage disequilibrium between variants also has smaller confidence intervals in some genomic regions. Smaller intervals have fewer spurious overlapping eQTL, which could help to narrow the list of candidate genes. Although mediation analysis is only effective if a change in expression is observed and might not be useful for identifying effects from protein-coding variation, many current studies show that the majority of genetic variants associated with complex traits lie in regulatory regions [224]. Whole-genome expression analysis could provide the missing link to the identification of causal genes underlying complex traits.

3.6.2 New evidence for the pleiotropic function of *scb-1*

We identified eight chemotherapeutics with a QTL that mapped to a genomic region defined as a QTL hotspot on the center of chromosome V [73]. Multiple genes in close proximity, each regulating an aspect of cellular growth and fitness, might underlie each QTL independently. Alternatively, genetic variation within a single gene might regulate responses to multiple (or all) of the eight drugs tested,

particularly if the gene is involved in drug transport or metabolism or if the drug mechanisms of action were shared (*e.g.* repair of double-strand DNA breaks). Expression of *scb-1*, a gene previously implicated in modulating responses to bleomycin, was found to reduce sensitivity to half of the drugs tested. This pleiotropic effect of *scb-1* provides new evidence for the function of the gene and possible molecular mechanisms underlying nematode drug responses. It is hypothesized that SCB-1 might function as a hydrolase that metabolizes compounds like bleomycin [72] or somehow plays a role in the nematode stress response [206]. Both hypotheses are consistent with our data, explaining why nematodes with low expression of *scb-1* are highly sensitive to the compound. Furthermore, all four of these chemotherapeutics, whose responses are mediated by expression of *scb-1*, are known to cause double-strand DNA breaks [213, 214, 216, 217]. Although the results for bortezomib, puromycin, and silver were inconclusive, we found no clear evidence that expression of *scb-1* dictates their responses. Together, these data suggest a potential role for *scb-1* specifically in response to stress induced by double-strand DNA breaks. However, the lack of sensitivity in etoposide, which also causes double-strand DNA breaks [215], indicates that this response might be more complex.

The exact variant that causes the differential expression of *scb-1* is still unknown. Importantly, *scb-1* lies within an eQTL hotspot region where it is hypothesized that genetic variation at a single locus might regulate expression of the 49 genes with an eQTL in this region. It is possible that the same causal variant that regulates expression of *scb-1* could also underlie the QTL for the remaining four chemotherapeutics through differential expression of other genes. For example, mediation analysis for both bortezomib and etoposide indicated that expression variation of a dehydrogenase (*D1054.8*) may underlie their differential responses. Alternatively, the causal variants underlying these drug-response QTL might be distinct but physically linked in the genome. This result would suggest a cluster of genes essential for the nematode drug response. Overall, our study highlights the power of using mediation analysis to connect gene expression to organismal traits and describes a novel function for the pleiotropic gene *scb-1*.

3.7 FUTURE DIRECTIONS

This study combined mediation analysis and CRISPR-Cas9-mediated deletions to implicate the gene *scb-1* in responses to cisplatin, amsacrine, and carmustine in addition to its previously known role in the bleomycin response. However, reciprocal hemizygosity tests must be performed to conclude that natural genetic variation in *scb-1* is responsible for the variation in drug sensitivity observed. Such a test has previously been performed for bleomycin [72] and could be replicated with the other three drugs. Because all the necessary strains already exist, this test would just require performing a series of crosses and phenotyping the heterozygous progeny. If the results indicate that natural genetic variation in *scb-1* between the N2 and CB4856 strains does not cause variation in drug sensitivity, it can be concluded that while *scb-1* plays a role in the drug responses, it is not the causal gene underlying the QTL on chromosome V. Alternatively, if the results indicate that natural genetic variation in *scb-1* between the N2 and CB4856 strains does cause variation in drug sensitivity, the search for the causal variant driving *scb-1* expression (and thus drug sensitivity) can commence. Brady *et al.* was not able to identify the exact genetic variant that causes increased sensitivity to bleomycin through decreased expression of *scb-1*, but lists 72 rare variants that could be causal, none of which are protein-coding variants [72]. Further analysis of each of these variants could be done by measuring *scb-1* expression after replacing the CB4856 allele with the N2 allele using CRISPR-Cas9 genome editing. Allele replacements could be grouped for testing efficiency. Alleles that cause a decrease in expression of *scb-1* would be prioritized for testing sensitivities to chemotherapeutics.

The function of *scb-1* is largely unknown, although it is predicted to function in the cellular stress response and might act as a hydrolase to break down molecular compounds. This study added to the hypothesized function of *scb-1* by showing that expression of this gene influences responses to several chemotherapeutics, not just bleomycin. Furthermore, we noticed that all of these drugs cause double strand DNA breaks, albeit via different mechanisms. Molecular assays to validate the function of *scb-1*, particularly how expression of *scb-1* mediates the *C. elegans* drug response, would be valuable. The hypothesis that SCB-1 acts as a hydrolase could be tested using a variety of in vitro assays to test its aminopeptidase function. Additionally, mass spectrometry could be used to detect bleomycin and its metabolites in nematodes with functional and non-functional copies of SCB-1. If SCB-1 acts as a

hydrolase, we would expect that bleomycin metabolites would be detected in animals with a functional copy of SCB-1 but not in animals with a non-functional copy of SCB-1. The alternative hypothesis that SCB-1 responds to double strand DNA breaks could be tested by treating with other compounds and/or stressors known to cause double strand DNA breaks. Additionally, a recent study showed that microirradiation in *C. elegans* allows researchers the opportunity to make precise double strand DNA breaks *in vivo* [225]. This technique could be used to cause double strand DNA breaks and observe if a fluorescently tagged SCB-1 is recruited following this particular stressor. Alternatively, similarly increased sensitivity to bleomycin and the other chemotherapeutics caused by deletion of a crucial component of the double strand DNA break repair machinery would provide evidence that SCB-1 might also be a component of this repair machinery. Better understanding the general function of SCB-1 will lead to a better understanding of how variation in *scb-1* causes variation in drug responses.

3.8 CONTRIBUTIONS

Dr. Shannon Brady made most of the strains (NILs and *scb-1* deletion strains) used to confirm the role of in responses to these drugs. This work was supported by an American Cancer Society Research Scholar Grant to Dr. Erik Andersen. Additionally, I received support from the NSF-Simons Center for Quantitative Biology at Northwestern University. I would like to thank the members of the Andersen Lab, particularly Dr. Daehan Lee and Dr. Tim Crombie, who acted as sounding boards throughout the development of the mediation analysis. I would also like to thank Dr. Shannon Brady who made this work possible by discovering *scb-1* and suggesting its potential role in responses to other chemotherapeutics that mapped to chromosome V.

3.9 CITATION

The gene *scb-1* underlies variation in *Caenorhabditis elegans* chemotherapeutic responses

Kathryn S. Evans[1,2] and Erik C. Andersen[1,3]

[1]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208
[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois 60208

[3]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, Illinois 60611

This manuscript was published in G3 in May 2020 [69]

4 Natural variation in the sequestosome-related gene, *sqst-5*, underlies zinc homeostasis in *Caenorhabditis elegans*

4.1 PREFACE

This chapter is what I consider to be the heart of my thesis. I have been working on this project from the very beginning and it is the only one that made it all the way through–not that there weren't hardships! The chromosome V QTL gave me so much grief over the last few years. Every time I thought I had narrowed the interval and chosen a candidate gene to test, I would run the assay again and the results would contradict my previous results. More than anything else, this project taught me about the limits of quantitative genetics. In theory, NILs are amazing and simple, what could go wrong? But in practice, it became clear to me that the traits we study are complex and there are limits to what we can discover, regardless of the amount of time and energy a PhD student spends on the project. The chromosome III QTL taught me to think outside the box and to never give up. I didn't find *sqst-5* until the very last experiment I ran before cutting my losses and publishing without a causal gene (six months before graduation). This chapter is based off my first-author manuscript submitted to PLOS Genetics in 2020 [123].

4.2 ABSTRACT

Zinc is an essential trace element that acts as a co-factor for many enzymes and transcription factors required for cellular growth and development. Altering intracellular zinc levels can produce dramatic effects ranging from cell proliferation to cell death. To avoid such fates, cells have evolved mechanisms to handle both an excess and a deficiency of zinc. Zinc homeostasis is largely maintained via zinc transporters, permeable channels, and other zinc-binding proteins. Variation in these proteins might affect their ability to interact with zinc, leading to either increased sensitivity or resistance to natural zinc fluctuations in the environment. We can leverage the power of the roundworm nematode *Caenorhabditis elegans* as a tractable metazoan model for quantitative genetics to identify genes that could underlie variation in responses to zinc. We found that the laboratory-adapted strain (N2) is

resistant and a natural isolate from Hawaii (CB4856) is sensitive to micromolar amounts of exogenous zinc supplementation. Using a panel of recombinant inbred lines, we identified two large-effect quantitative trait loci (QTL) on the left arm of chromosome III and the center of chromosome V that are associated with zinc responses. We validated and refined both QTL using near-isogenic lines (NILs) and identified a naturally occurring deletion in *sqst-5*, a sequestosome-related gene, that is associated with resistance to high exogenous zinc. We found that this deletion is relatively common across strains within the species and that variation in *sqst-5* is associated with zinc resistance. Our results offer a possible mechanism for how organisms can respond to naturally high levels of zinc in the environment and how zinc homeostasis varies among individuals.

4.3 INTRODUCTION

Heavy metals such as zinc, iron, and copper are known to play important roles in many biological systems [226, 227]. Of these metals, zinc is the most abundant and is essential for proper function of many proteins, including enzymes and transcription factors [228]. In addition to its function as a cofactor, zinc can act as a signaling molecule in neurons [229, 230, 231, 232] and is known to play a role in cell-fate determination [233, 234, 235, 236, 237]. Because of its many functions, zinc deficiency has been shown to cause major defects, including growth inhibition and death in several species [233, 238, 239, 240, 241]. On the other hand, excess zinc can also be toxic, displaying phenotypic effects similar to copper deficiency, anemia, and neutropenia [242]. Although the exact mechanisms are unknown, the data suggest that excess zinc might bind ectopically to other proteins, displacing similar metals such as copper or magnesium from these proteins [233]. Because of this need for intracellular zinc balance even though environmental zinc might fluctuate, biological systems must use proper zinc homeostasis mechanisms for uptake, distribution, efflux, and detoxification [238].

The nematode *Caenorhabditis elegans* is a tractable metazoan model for studying the molecular mechanisms of zinc homeostasis and toxicity [233, 243, 244, 245]. As observed in other organisms, zinc is essential for *C. elegans* growth [246]. In fact, it is estimated that about 8% of the *C. elegans* genes (1,600 genes) encode zinc-binding proteins [247]. However, zinc is also toxic to the nematode at higher concentrations [246]. High exogenous zinc can cause several defects including decreased

growth rate and survival, suppression of the multivulva phenotype, and formation of bilobed lysosome-related organelles in intestinal cells [233]. Genetic screens have identified several genes that act to increase sensitivity to high levels of zinc (*haly-1*, *natc-1*, and *daf-21*) [233, 248, 249, 250, 251]. However, mutations in these genes cause a change in response to multiple stressors (including metals, heat, and oxidation), suggesting they are not specific to zinc homeostasis [250, 233]. In addition to these nonspecific zinc proteins, *C. elegans* also has two complementary families (composed of 14 proteins each) of zinc transporters responsible for maintaining constant intracellular zinc concentrations via import and export [233]. Four of these zinc exporters (*cdf-1*, *cdf-2*, *sur-7*, and *ttm-1*) have been shown to promote resistance to high zinc toxicity [233, 248, 246, 252, 253, 234].

Although much is already known about zinc biology in *C. elegans*, previous studies were performed using a single laboratory-adapted strain (N2) that is known to differ significantly, both genetically and phenotypically, from wild isolates in the species [58]. As a complementary approach, we can leverage the power of natural genetic diversity among wild isolates [59, 79, 63] to identify novel mechanisms of zinc homeostasis and gain insights into the evolution of this process. We used a large panel of recombinant inbred advanced intercross lines (RIAILs) [78, 68, 72] constructed from a multi-generational cross between two genetically and phenotypically diverged strains, N2, the laboratory-adapted strain, and CB4856, a wild isolate from Hawaii [60]. This panel of RIAILs has been leveraged in several linkage mapping analyses, identifying hundreds of quantitative trait loci (QTL) [130, 147, 133, 117, 70, 72, 71, 73, 100, 108, 35, 149, 125, 128, 115, 74, 127, 132, 112, 119, 81, 103, 102, 80, 161, 86, 85, 69]. In combination with a high-throughput phenotyping assay to measure animal length, optical density, and brood size [68], several quantitative trait genes (QTG) [72, 69] and quantitative trait nucleotides (QTN) [70, 71] underlying fitness-related traits have been described.

Here, we use linkage mapping analysis to identify four QTL in response to high exogenous zinc. Several genes previously identified to be involved in the zinc response were found within the QTL on chromosomes V and X. However, no known zinc-related genes were located in the large-effect chromosome III QTL, suggesting a potentially novel mechanism of zinc homeostasis. We constructed reciprocal near-isogenic lines (NILs) for each QTL and used them to validate the two large-effect QTL on

chromosomes III and V. Expression QTL mapping and mediation analysis identified a single candidate gene, *sqst-5*, with predicted zinc ion-binding capability. We used CRISPR-Cas9 genome editing to show that strains without *sqst-5* were significantly more resistant to zinc supplementation than strains with a functional copy of *sqst-5*, suggesting a new role for this gene in zinc regulation. In addition to CB4856, several other wild isolates were found to share a 111 bp deletion in *sqst-5*. Moreover, we identified a second group of strains with a distinct haplotype of variation at *sqst-5* that was also associated with zinc resistance. Together, these data suggest that the regulation of zinc in nematodes is complex, but binding and accumulation of excess zinc might be a mechanism to respond to exogenous zinc.

4.4 METHODS

4.4.1 Strains

Animals were grown at 20℃ on modified nematode growth media (NGMA) containing 1% agar and 0.7% agarose to prevent burrowing and fed OP50 [100]. The two parental strains, the canonical laboratory strain, N2, and the wild isolate from Hawaii, CB4856, were used to generate all recombinant lines. 253 recombinant inbred advanced intercross lines (RIAILs) generated previously [68] (set 2 RIAILs) were used for zinc phenotyping and QTL mapping. A second set of 121 RIAILs generated previously [78] (set 1 RIAILs) were phenotyped for mediation analysis.

4.4.2 Standard high-throughput fitness assay

For dose responses and RIAIL phenotyping, we used a high-throughput fitness assay (HTA) described previously [68]. In summary, populations of each strain were passaged and amplified on NGMA plates for four generations without starvation. In the fifth generation, gravid adults were bleach-synchronized and 25-50 embryos from each strain were aliquoted into 96-well microtiter plates at a final volume of 50 μL K medium [173]. The following day, arrested L1s were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA; [174]) at a final concentration of 5 mg/mL in K medium and were grown to the L4 larval stage for 48 hours at 20℃ with constant shaking. Three L4 larvae were sorted into new 96-well microtiter plates containing 10 mg/mL HB101 bacterial

lysate, 50 μM kanamycin, and either 1% water or zinc sulfate dissolved in 1% water using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica; Holliston, MA). Sorted animals were grown for 96 hours at 20℃ with constant shaking. The next generation of animals and the parents were treated with sodium azide (50 mM in 1X M9) to straighten their bodies for more accurate measurements. Animal length (TOF) and optical density (EXT) were quantified for every animal in each well using the COPAS BIOSORT and the medians of each well population (median.TOF and median.EXT) were used to estimate these traits. Animal length and optical density are both measures of nematode development; animals get longer and more optically dense (thicker and denser body composition) as they develop [68]. However, the COPAS BIOSORT measures optical density as a function of length. Because these two traits are highly correlated, we also generated a fourth trait (median.norm.EXT) which normalizes the optical density by length (EXT/TOF) in order to provide a means to compare optical densities regardless of animal lengths. Finally, brood size (norm.n) was calculated as the total number of animals in the well normalized by the number of parents originally sorted and provides an estimate of nematode reproductive fitness [68].

4.4.3 Calculation of zinc-response traits

Phenotypic measurements collected by the BIOSORT were processed and analyzed using the R package *easysorter* [77] as described previously [72]. Briefly, raw data from the BIOSORT was read into R using the *read_data* function and contaminated wells were removed using the *remove_contamination* function. The sumplate function was used to calculate summary statistics per well and four main traits were output: median.TOF (animal length), median.EXT (animal optical density), median.norm.EXT (animal optical density normalized by animal length), and norm.n (brood size). When trait measurements were collected across multiple assay experiments, the *regress(assay = T)* function was used to fit the linear model $phenotype \sim assay$ to account for variation among assays. Outliers were pruned using *prune_outliers* to remove wells beyond two standard deviations of the mean for highly replicated assays. Alternatively, for assays with low replication (dose response and RIAIL phenotyping), *bamf_prune* was used to remove wells beyond two times the IQR plus the 75th quartile or two times the IQR minus the 25th quartile, unless at least 5% of the strains lie outside this

range. Finally, zinc-specific effects were calculated using the *regress(assay = FALSE)* function, which subtracts the mean water (control) value from each zinc replicate for each strain using a linear model $drug\_phenotype \sim control\_phenotype$. The residual phenotypic values were used as the zinc-response phenotype for all downstream analyses. In this way, we addressed only the differences among strains that were caused by treatment with zinc and ignored minor phenotypic variation among strains in the control condition. Pairwise tests for statistically significant differences in the zinc response between strains were performed using the *TukeyHSD function* [175] on an ANOVA model with the formula $phenotype \sim strain$. For plotting purposes, these residual values were normalized from zero to one with zero being the well with the smallest value and one the well with the largest value.

4.4.4 Zinc dose response

Four genetically divergent strains (N2, CB4856, JU258, and DL238) were treated with increasing concentrations of zinc using the standard high-throughput assay described above. A concentration of 500 μM zinc sulfate heptahydrate (Sigma #221376-100G) in water was selected for the linkage mapping experiments. This concentration provided a reproducible zinc-specific effect and maximizes between-strain variation and minimizes within-strain variation across the four traits. Broad-sense heritability was calculated from the dose response phenotypes using the *lmer* function in the *lme4* R package [212] with the formula $phenotype \sim 1 + (1|strain)$ for each dose.

4.4.5 Linkage mapping

253 RIAILs (set 2 RIAILs) were phenotyped in both zinc and water using the standard high-throughput assay described above. Linkage mapping was performed for all four zinc-response traits using the R package *linkagemapping* (https://github.com/AndersenLab/linkagemapping) as described previously [72]. The cross object derived from the whole-genome sequencing of the RIAILs containing 13,003 SNPs was loaded using the function *load_cross_obj("N2xCB4856cross_full")*. The RIAIL phenotypes were merged into the cross object using the *merge_pheno* function with the argument *set = 2*. A forward search (*fsearch* function) adapted from the *R/qtl* package [176] was used to calculate the logarithm of

the odds (LOD) scores for each genetic marker and each trait as $-n(ln(1-R2)/2ln(10))$ where R is the Pearson correlation coefficient between the RIAIL genotypes at the marker and trait phenotypes [27]. A 5% genome-wide error rate was calculated by permuting the RIAIL phenotypes 1000 times. The marker with the highest LOD score above the significance threshold was selected as the QTL then integrated into the model as a cofactor and mapping was repeated iteratively until no further significant QTL were identified. Finally, the *annotate_lods* function was used to calculate the effect size of each QTL and determine 95% confidence intervals defined by a 1.5 LOD drop from the peak marker using the argument *cutoff = chromosomal*. In the same manner, linkage mapping was performed for four other divalent cation metals: arsenic dibasic in water [70], copper in water, cadmium in water [73], and nickel in water [122].

4.4.6 Two-dimensional genome scan

A two-dimensional genome scan to identify interacting loci was performed for animal optical density (median.EXT) in zinc using the *scantwo* function from the *qtl* package [176] as described previously [73, 72]. Each pairwise combination of loci was tested for correlations with trait variation in the RIAILs. A summary of the maximum interactive LOD score for each chromosome pair can be output using the summary function. Significant interactions were identified by permuting the phenotype data 1000 times and determining the 5% genome-wide error rate. The significant interaction threshold for the zinc-response variation scantwo was 4.09.

4.4.7 Construction of near-isogenic lines (NILs)

NILs were generated as previously described [72, 73, 70, 71] by either backcrossing a selected RIAIL for six generations or de novo by crossing the parental strains N2 and CB4856 to create a heterozygous individual that is then backcrossed for six generations. PCR amplification of insertion-deletion (indel) variants between N2 and CB4856 were used to track the genomic interval. Smaller NILs to further break up the interval were created by backcrossing a NIL for one generation to create a heterozygous F1 individual. The F1 individuals were selfed, and the F2 population was scored for recombination events. NILs were whole-genome sequenced to verify introgressions and the absence of other introgressed

regions [73, 72].

4.4.8 Mediation analysis

121 RIAILs (set 1 RIAILs) were phenotyped in both zinc and water using the standard high-throughput assay described above. Microarray expression for 14,107 probes were previously collected from the set 1 RIAILs [115], filtered [100], and mapped using linkage mapping with 13,003 SNPs [69]. Mediation scores were calculated with bootstrapping using the *mediation* R package [210] as previously described [69] for each of the 19 probes (including *ver-2*/*sqst-5*, A_12_P104472) with an eQTL on the left arm of chromosome III. Briefly, a mediator model $expression \sim genotype$ and an outcome model $phenotype \sim expression + genotype$ were used to calculate the proportion of the QTL effect that can be explained by variation in gene expression. All expression and eQTL data can be found at https://github.com/AndersenLab/scb1_mediation_manuscript.

4.4.9 Generation of deletion strains

Deletion alleles for *sqst-5* and *ver-2* were generated as previously described using CRISPR-Cas9 genome editing [72, 75]. Briefly, 5' and 3' guide RNAs were designed with the highest possible on-target and off-target scores [254] and ordered from Synthego (Redwood City, CA). The following CRISPR injection mix was assembled and incubated for an hour before injection: 1 μM dpy-10 sgRNA, 5 μM of each sgRNA for the gene of interest, 0.5 μM of a single-stranded oligodeoxynucleotide template for homology-directed repair of dpy-10 (IDT, Skokie, IL), and 5 μM Cas9 protein (Q3B Berkeley, Berkeley, CA) in water. Young adults were mounted onto agar injection pads, injected in either the anterior or posterior arm of the gonad, and allowed to recover on 6 cm plates. After 12 hours, survivors were transferred to individual 6 cm plates and allowed to lay embryos. Two days later, the F1 progeny were screened and individuals with Rol or Dpy phenotypes were selected and transferred clonally to new 6 cm plates. After 48 hours, the F1 individuals were genotyped by PCR flanking the desired deletions. Individuals with heterozygous or homozygous deletions were propagated and genotyped for at least two additional generations to ensure homozygosity and to cross out the Rol mutation. Deletion amplicons were Sanger sequenced to identify the exact location of the

deletion.

4.4.10 Modified high-throughput fitness assay

Dominance and validation of candidate genes were tested using a modified version of the standard high-throughput assay detailed above as previously described [72, 69]. For candidate gene testing, strains were propagated for two generations, bleach-synchronized in six independent replicates, and titered at a concentration of 25-50 embryos per well of a 96-well microtiter plate. For dominance and hemizygosity assays, strains (males and hermaphrodites) were propagated and amplified for two generations. For each cross, 30 hermaphrodites and 60 males were placed onto each of four 6 cm plates and allowed to mate for 48 hours. Mated hermaphrodites were transferred to a clean 6 cm plate and allowed to lay embryos for eight hours. After the egg-laying period, adults were manually removed and embryos were collected by vigorous washing with 1X M9. Embryos were resuspended in K medium and titered to a concentration of 25 embryos per well of a 96-well microtiter plate. For both assays, arrested L1s were fed HB101 bacterial lysate the following day at a final concentration of 5 mg/mL with either water or zinc. After 48 hours of growth at 20℃ with constant shaking, nematodes were treated with sodium azide (5 mM in water) prior to analysis of animal length and optical density using the COPAS BIOSORT. Because only one generation of growth was observed, brood size was not calculated. Lower drug concentrations were needed to see the previous effect because of the modified timing of the drug delivery. A concentration of 250 μM zinc in water was used for these experiments.

4.4.11 Local alignment of the *sqst-5* region using long-read sequence data

To confirm the putative *sqst-5* deletion in CB4856, we aligned the long-read assembly for CB4856 [62] to the N2 reference genome using NUCmer (version v3.1) [255]. Using this alignment, we identified the coordinates of *sqst-5* in CB4856 and extracted the sequence using BEDtools (version v2.29.2) [256]. We aligned the unspliced N2 *sqst-5* transcript sequence (WormBase WS273) and the N2 SQST-5 protein sequence to this extracted CB4856 sequence using Clustal Omega [257] and GeneWise [258], respectively. Gene prediction in CB4856 was run using Augustus [259]. We visually inspected both alignments to identify the length of the deletion in CB4856 and identify the effect of the deletion of the

CB4856 SQST-5 protein sequence.

4.4.12 Assessment of strain relatedness through neighbor-joining tree

Variant data for dendrogram comparisons were assembled by constructing a FASTA file with the genome-wide variant positions across all strains and subsetting to keep only variants near *sqst-5* (III:145917-148620). Genotype data were acquired from the latest VCF release (release 20180517) from CeNDR. Multiple sequence comparison by log-expectation (MUSCLE, version v3.8.31) [260] was used to generate neighbor-joining trees. A second neighbor-joining tree was constructed with all the variants within the QTL confidence interval for comparison (III:4664-597553). Both trees were identical.

4.4.13 Genome wide association mapping

Eighty-one wild isolates were phenotyped in both zinc and water using the standard high-throughput assay described above. Genome-wide association (GWA) mappings were performed for all four traits using the R package *cegwas2* (https://github.com/AndersenLab/cegwas2-nf) as described previously [70]. Genotype data were acquired from the latest VCF release (release 20180517) from CeNDR. We used BCFtools [261] to filter variants with missing genotypes and variants below a 5% minor allele frequency. We used PLINK v1.9 [262, 263] to LD-prune genotypes. The additive kinship matrix was generated from the 64,046 markers using the *A.mat* function in the rrBLUP R package [264]. Because these markers have high LD, we performed eigen decomposition of the correlation matrix of the genotype matrix to identify 477 independent tests [70]. We used the GWAS function from the *rrBLUP* package to perform genome-wide mapping. Significance was determined in two ways: a strict Bonferroni threshold and a more lenient eigenvalue threshold set by the number of independent tests in the genotype matrix. Confidence intervals were defined as $\pm$ 150 SNVs from the rightmost and leftmost markers that passed the significance threshold.

4.4.14 Statistical analysis

All statistical tests of phenotypic differences between strains were performed using the *TukeyHSD* function [175] on an ANOVA model with the formula $phenotype \sim strain$. The *p*-values for individual pairwise strain comparisons were adjusted for multiple comparisons (Bonferroni). The datasets and code for generating figures can be found at https://github.com/AndersenLab/zinc_manuscript.

4.5 RESULTS

4.5.1 Natural genetic variation in response to zinc is complex

We exposed four genetically divergent strains of *C. elegans* (N2, CB4856, JU258, and DL238) to increasing concentrations of exogenous zinc and measured their development (animal length, optical density, and normalized optical density) and reproductive ability (brood size) with a high-throughput assay using the COPAS BIOSORT (see Methods) [70, 72, 73, 71, 68]. In the presence of high concentrations of zinc, animals of all strains had smaller broods, shorter lengths, and were less optically dense compared to non-treated animals (**Figure 4-1**). Because nematodes grow longer and become more optically dense as they develop, these results suggest a zinc-induced developmental delay. Furthermore, the lower brood size of animals treated with zinc suggest that exogenous zinc hinders reproductive ability in some way. In addition to these overall trends, we also observed significant phenotypic variation among strains. For example, although all strains had smaller lengths in the presence of exogenous zinc, animals of the N2 strain were the largest (most resistant to zinc), and animals of the CB4856 strain were smaller (more sensitive to zinc). At 500 µM zinc, a concentration that both maximizes among-strain and minimizes within-strain phenotypic variation, we identified a substantial heritable genetic component for two highly correlated developmental traits: animal length ($H^2 = 0.48$, 95% CI [0.30, 0.61]) and optical density ($H^2 = 0.48$, 95% CI [0.28, 0.59]).

To investigate the genetic basis of zinc response, we exposed a panel of 253 RIAILs derived from a cross between the N2 and CB4856 strains (set 2 RIAILs, see Methods) to high exogenous zinc. In these conditions, the N2 animals were longer (**Figure S4-1**) and more optically dense (**Figure 4-2A**) than the CB4856 animals, and were thus more resistant to high zinc supplementation. Interestingly,

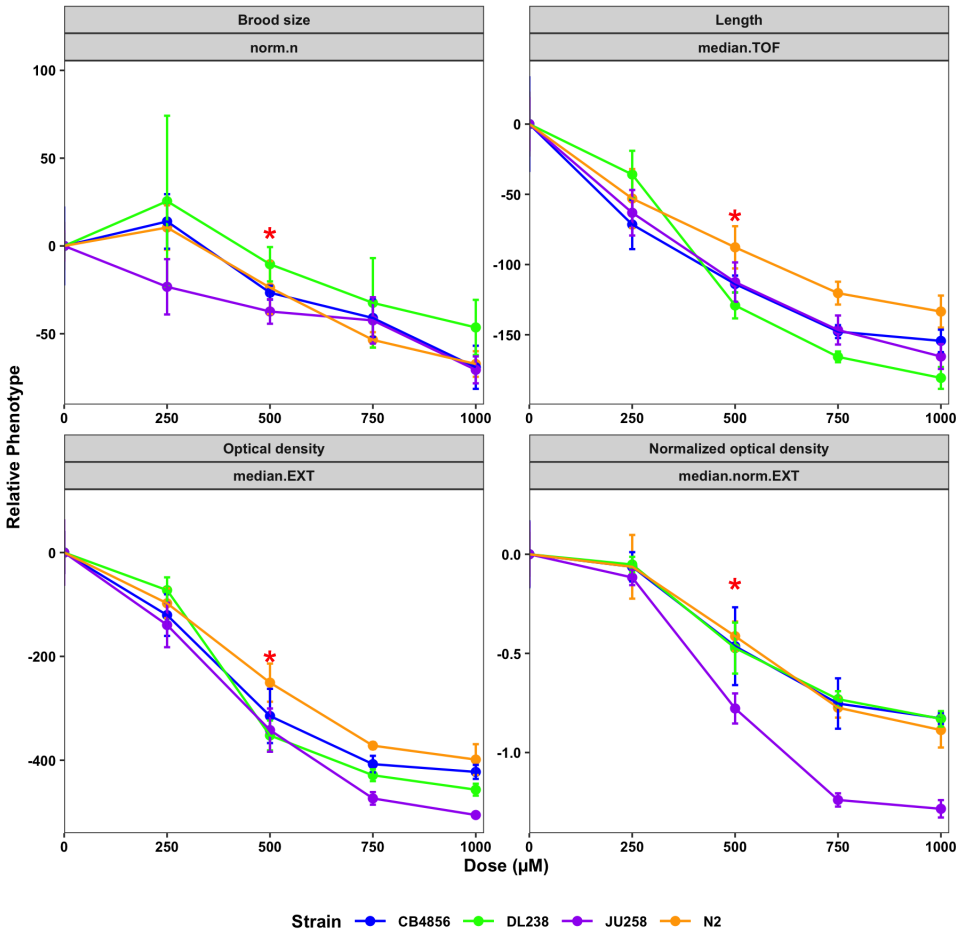


Figure 4-1: Dose response with four divergent wild isolates. Results from the zinc dose response HTA for brood size (norm.n), animal length (median.TOF), animal optical density (median.EXT), and normalized optical density (median.norm.EXT). For each trait, drug concentration (μM) (x-axis) is plotted against phenotype subtracted from control (y-axis), colored by strain (CB4856: blue, DL238: green, JU258: purple, N2: orange). A red asterisk indicates the dose selected for linkage mapping analysis.

many of the RIAILs were either more resistant than N2 or more sensitive than CB4856, suggesting that loci of opposite genotypes are either acting additively or interacting in the RIAILs to produce the observed transgressive phenotypes [180]. Linkage mapping analysis identified 12 QTL across all traits, representing five unique QTL on chromosomes III, IV, V, and X (**Figure S4-1**). Because genetic architectures looked similar across these traits, we chose to focus our analyses on optical density to avoid redundant analyses of correlated traits (**Figure 4-2B**). Together, the four QTL underlying animal optical density explain 40.5% of the phenotypic variation among the RIAILs. As expected, QTL of opposite effects were observed. Strains with the CB4856 allele on chromosome III were more

resistant to zinc than strains with the N2 allele at this locus. By contrast, strains with the CB4856 alleles on chromosomes IV, V, and X were more sensitive to zinc than strains with the N2 alleles at these loci (**Figure 4-2C**). We scanned the genome for interactions between pairs of genomic markers that might affect the phenotypic distribution of the RIAIL panel and identified no significant interactions (**Figure 4-3**). We further examined the additivity of the two QTL with the largest and opposite effect sizes (QTL on chromosomes III and V). We concluded that RIAILs with the CB4856 allele on chromosome III and the N2 allele on chromosome V were the most resistant, and RIAILs with the N2 allele on chromosome III and the CB4856 allele on chromosome V were the most sensitive (**Figure 4-4**). Furthermore, the effect size of the chromosome III locus was similar regardless of the genotype on chromosome V (**Figure 4-4**), and no significant interaction term was identified using a linear model (ANOVA, $p = 0.251$). These results suggest that multiple additive QTL rather than interacting loci affect animal optical densities in zinc.

4.5.2 Near-isogenic lines fractionate the chromosome V QTL into multiple additive loci

We first investigated whether any of the 28 known zinc transporters or any of the other 15 zinc-related genes [233] are located in one of the four detected QTL intervals. We discovered that three of these genes lie in the QTL on chromosome V and 11 lie in the QTL on chromosome X (**Table 4-1**). However, none of these zinc-related genes lie in either of the QTL on chromosomes III or IV (**Table 4-1**). To isolate and validate the effect of these four QTL, we constructed reciprocal near-isogenic lines (NILs) by introgressing a genomic region surrounding each of the QTL from the CB4856 strain into the N2 genetic background or vice versa. We then measured the animal optical densities in the presence of zinc for these strains to provide experimental evidence in support of each QTL independently. For the three QTL on chromosomes IV, V, and X, the N2 allele was associated with zinc resistance (**Figure 4-2C**). However, strains with the N2 allele crossed into a CB4856 genetic background on chromosomes IV and X were as sensitive as the CB4856 strain, and strains with the CB4856 allele crossed into an N2 genetic background were as resistant as the N2 strain (**Figure 4-5**). These two QTL have the smallest effect sizes among the four QTL detected and each explain only 5% of the total phenotypic variation among the RIAILs. The lack of a significant difference between the NILs and their respective parental

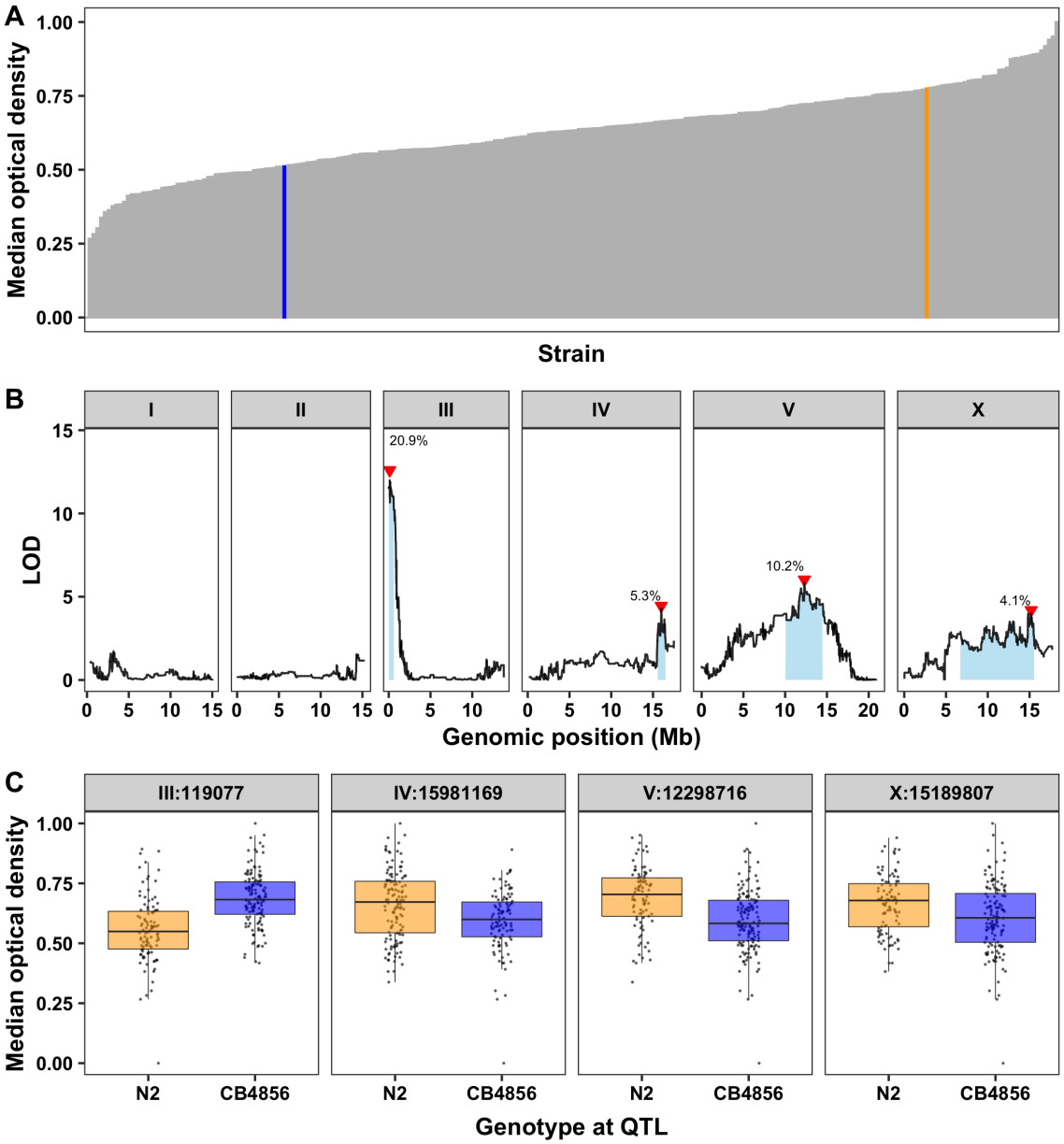


Figure 4-2: Linkage mapping identifies four QTL in response to high dietary zinc. (A) Normalized residual median optical densities (y-axis) of 253 RIAILs (x-axis) in response to zinc supplementation are shown. The parental strains are colored: N2, orange; CB4856, blue. **(B)** Linkage mapping results for optical density (median.EXT) is shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(C)** For each QTL, the normalized residual median optical densities (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange, and strains with the CB4856 allele are colored blue.

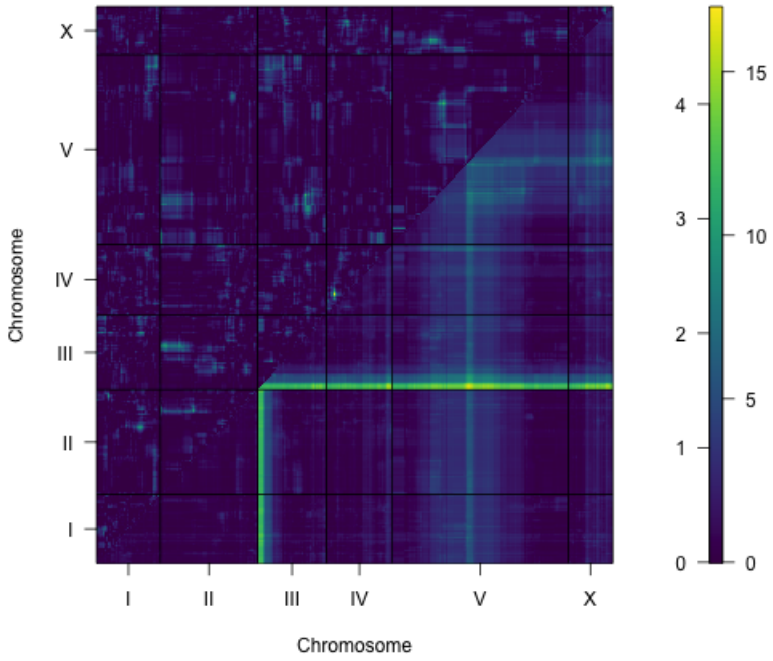


Figure 4-3: Two dimensional genome scan for median optical density (median.EXT) in zinc. Log of the odds (LOD) scores are shown for each pairwise combination of loci, split by chromosome. The upper-left triangle contains the epistasis LOD scores and the lower-right triangle contains the LOD scores for the full model. LOD scores are colored, increasing from purple to green to yellow. The LOD scores for the epistasis model are shown on the left of the color scale and the LOD scores for the full model are shown on the right.

strains suggests that the QTL effect might be smaller than 5% and we were underpowered to detect the difference. Alternatively, the interval might contain QTL of opposing effects requiring additional smaller NILs.

Table 4-1: Genetic variation in CB4856 for zinc-related genes in *C. elegans*

Gene ID	Gene name	Gene position	Variants[1]	Coding[2]	eQTL[3]	Overlap[4]
WBGene00013668	*Y105E8A.3*	I:14358502-14367594	Yes	No	No	No
WBGene00009286	*F31C3.4*	I:15050797-15052420	Yes	Yes	No	No
WBGene00019077	*F59A3.4*	I:5507123-5509288	Yes	No	No	No
WBGene00018897	*F55F8.9*	I:5669950-5673352	Yes	No	No	No
WBGene00010644	*K07G5.5*	I:7169997-7172189	No	No	No	No
WBGene00013207	*Y54G9A.4*	II:13712921-13715893	Yes	No	No	No

[1]Genetic variation in CB4856
[2]Protein-coding genetic variation in CB4856
[3]Does this gene have an eQTL between N2 and CB4856?
[4]Is the gene inside the confidence interval of one of the QTL identified?

WBGene00015940	*C18A3.2*	II:5718129-5720052	No	No	No	No
WBGene00012712	*ttm-1*	III:12952969-12966995	Yes	Yes	No	No
WBGene00022174	*Y71H2AM.9*	III:2690160-2698817	Yes	No	No	No
WBGene00006590	*toc-1*	III:5275225-5277797	Yes	No	No	No
WBGene00018948	*F56C9.3*	III:7311290-7313195	Yes	No	No	No
WBGene00014669	*C06G8.3*	IV:10799614-10802250	No	No	No	No
WBGene00044067	*hke-4.1*	IV:17285733-17287454	No	No	No	No
WBGene00006487	*tag-141*	IV:17364791-17366491	Yes	Yes	No	No
WBGene00017936	*F30B5.7*	IV:4229195-4231861	Yes	No	No	No
WBGene00044481	*ZK185.5*	IV:4541769-4545918	No	No	No	No
WBGene00021936	*Y55F3BL.2*	IV:807439-824954	Yes	No	Yes	No
WBGene00006486	*tag-140*	V:11462999-11466770	Yes	No	Yes	Yes
WBGene00011329	*T01D3.5*	V:13712854-13716624	Yes	No	No	Yes
WBGene00019841	*R02F11.3*	V:4802647-4805947	No	No	No	No
WBGene00007591	*C14H10.1*	X:10233739-10236944	Yes	No	No	Yes
WBGene00011821	*cdf-2*	X:12464640-12466917	No	No	No	Yes
WBGene00006494	*hke-4.2*	X:15506037-15507880	Yes	Yes	No	Yes
WBGene00006353	*sur-7*	X:16376595-16380047	Yes	No	No	No
WBGene00000393	*cdf-1*	X:6499928-6504613	Yes	No	No	No
WBGene00018280	*F41C6.7*	X:6882493-6885277	Yes	No	No	Yes
WBGene00019803	*PDB1.1*	X:7556788-7559085	Yes	No	No	Yes
WBGene00008954	*F19C6.5*	X:9982402-9985005	Yes	No	No	Yes
WBGene00004962	*spe-8*	I:118109-120870	Yes	Yes	No	No
WBGene00003996	*pgp-2*	I:5886613-5895970	Yes	Yes	No	No
WBGene00004966	*spe-12*	I:8224539-8226297	No	No	No	No
WBGene00007016	*mdt-15*	III:5828603-5833745	Yes	Yes	No	No
WBGene00004973	*spe-27*	IV:5864529-5865918	Yes	No	No	No
WBGene00004974	*spe-29*	IV:9933910-9934285	No	No	No	No
WBGene00003474	*mtl-2*	V:14018270-14018673	Yes	No	No	Yes
WBGene00000915	*daf-21*	V:14684918-14688543	Yes	No	No	No
WBGene00003473	*mtl-1*	V:6691371-6691863	No	No	No	No
WBGene00020719	*natc-1*	V:8462327-8465569	Yes	No	No	No
WBGene00001250	*elt-2*	X:10481240-10483446	Yes	Yes	No	Yes
WBGene00010341	*glo-3*	X:10533699-10537114	Yes	No	No	Yes
WBGene00009813	*haly-1*	X:10893567-10896598	Yes	No	No	Yes
WBGene00013976	*nhr-33*	X:11318278-11321085	Yes	No	No	Yes
WBGene00011083	*glo-1*	X:9872884-9882212	Yes	No	No	Yes

By contrast, the NIL with the N2 allele surrounding the QTL on chromosome V introgressed into the CB4856 genetic background is significantly more resistant than the sensitive CB4856 strain (**Figure 4-5**). This result confirms that genetic variation between the N2 and CB4856 strains on the center of chromosome V contributes to the differences in animal optical densities between the strains in the presence of high exogenous zinc. To further narrow this QTL, we created a panel of NILs with smaller regions of the N2 genome introgressed into the CB4856 genetic background. We exposed a subset of these NILs to zinc and measured their optical densities. We found strains with the resistant

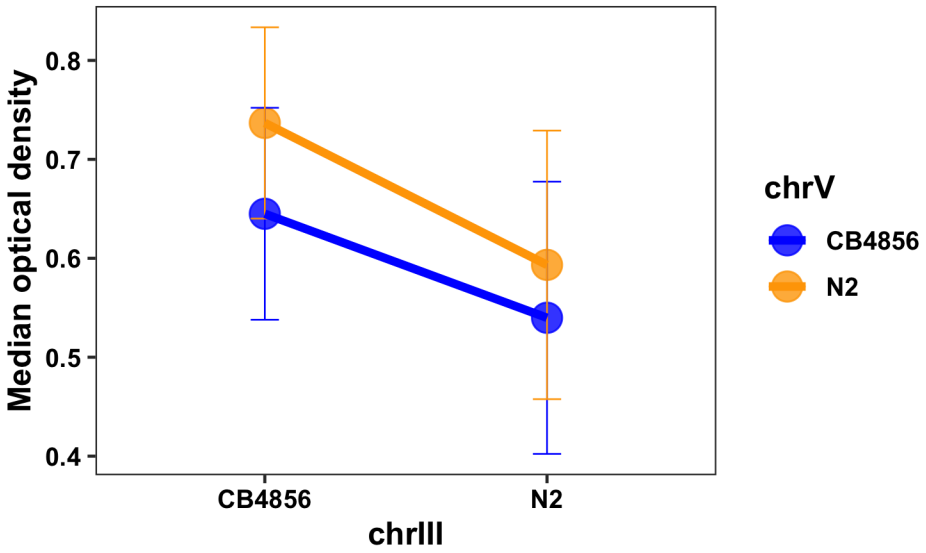


Figure 4-4: Reaction norm shows additive QTL effects between chromosome III and V. Normalized residual median optical density in zinc (median.EXT, y-axis) of RIAILs split by genotype at the chromosome III QTL (x-axis) are plotted as the mean of the population ± the standard deviation, colored by the genotype at the chromosome V QTL.

N2 phenotype (ECA481; V:9.6-13.8 Mb), strains with the sensitive CB4856 phenotype (ECA411; V:11.3-13.9 Mb), and strains with an intermediate phenotype (ECA437; V:10.5-13.8 Mb) (**Figure 4-6**). These data imply the existence of at least two loci (V:9.6-10.5 Mb and V:10.5-11.3 Mb) at which the N2 allele confers resistance to zinc. The intermediate strain (ECA437) contains one N2 locus and one CB4856 locus, and the resistant strain (ECA481) contains two N2 loci. Because this region is in the center of a chromosome where recombination frequency is lower [78], we were unable to generate NILs with a breakpoint to further narrow the QTL. Furthermore, it is possible that multiple small-effect loci are contributing to each of the two QTL, rendering it difficult to identify each causal gene or variant. Regardless, at least two novel loci on chromosome V were identified that influence zinc sensitivity in *C. elegans*.

4.5.3 Analysis of the chromosome III QTL suggests that a sequestosome-related gene, *sqst-5*, contributes to differences in zinc responses

The QTL on chromosome III accounts for 20% of the phenotypic variance in zinc response across the RIAIL population. In contrast to the previous three QTL, the CB4856 allele is associated with

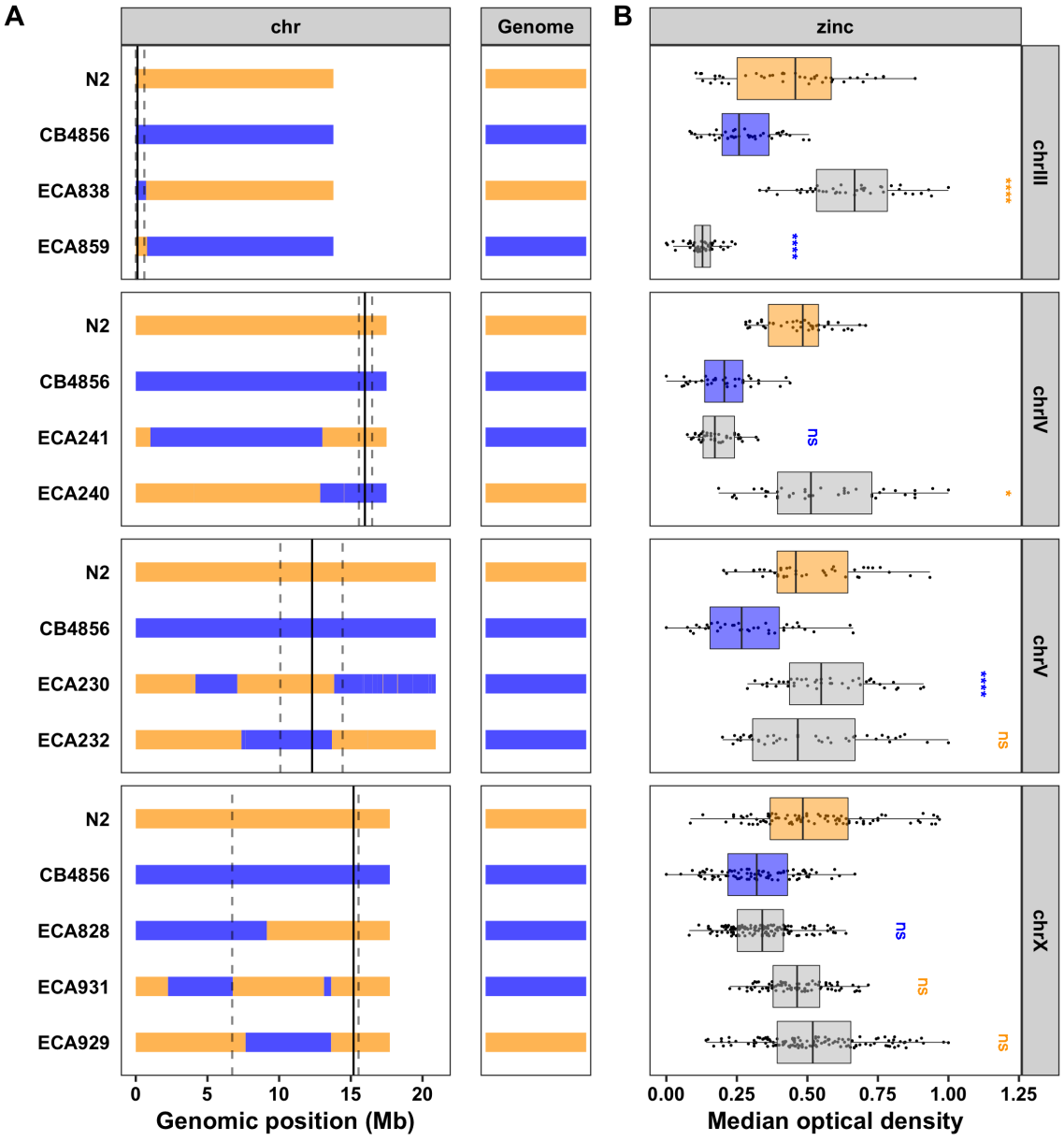


Figure 4-5: Validating QTL using near-isogenic lines (NILs). (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for each chromosome (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the QTL and the dashed vertical lines represent the confidence interval. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). The parental strains N2 and CB4856 are colored orange and blue, respectively. NILs are colored grey. Statistical significance of each strain compared to its parental strain (ECA838, ECA240, ECA232, ECA931, and ECA929 to N2 and ECA859, ECA241, ECA230, and ECA828 to CB4856) is shown above each strain and colored by the parent strain it was tested against (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

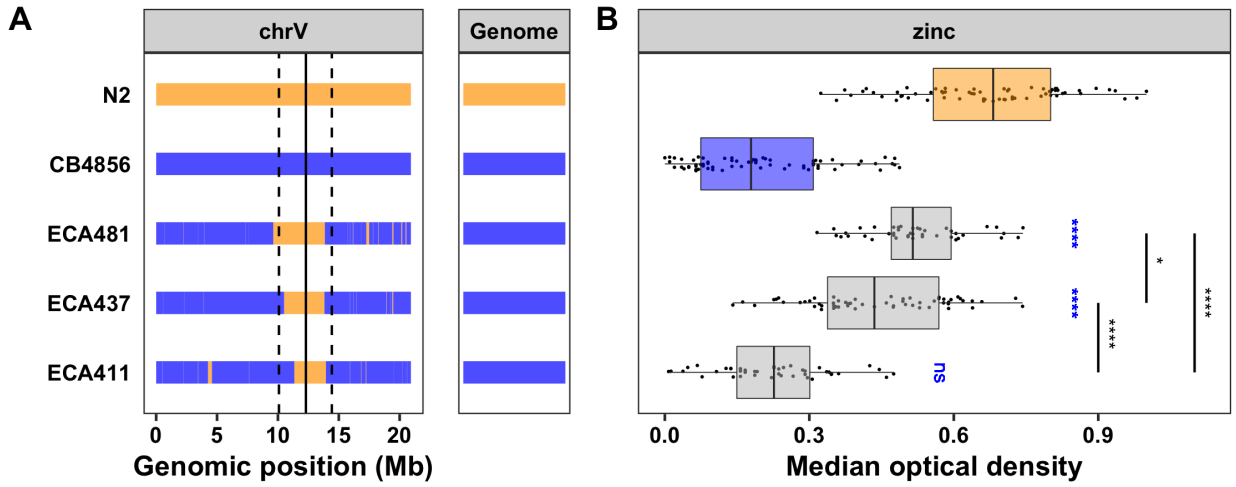


Figure 4-6: NILs identify multiple QTL on chromosome V. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the QTL, and the dashed vertical lines represent the confidence interval. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to CB4856 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

zinc resistance and the N2 allele is associated with zinc sensitivity (**Figure 4-2C**). The strain with the N2 allele on chromosome III crossed into a CB4856 genetic background (ECA859) was significantly more sensitive than the CB4856 strain (hyper-sensitive) and the strain with the opposite genotype (ECA838) was significantly more resistant than the N2 strain (hyper-resistant) (**Figure 4-5**). These results demonstrate that this locus contributes to the observed transgressive phenotypes in the RIAILs (**Figure 4-2A**). We also measured animal optical densities in zinc for individuals heterozygous for the chromosome III locus to determine whether the N2 or CB4856 allele confers the dominant phenotype. To analyze heterozygous individuals, we developed a modified high-throughput assay (see Methods, **Figure 4-7**). Individuals heterozygous for the chromosome III locus in the N2 genetic background (N2xECA838) were hyper-resistant similar to the NIL that is homozygous CB4856 for the chromosome III locus in the N2 genetic background (ECA838) (**Figure 4-8**). By contrast, individuals heterozygous for the chromosome III locus in the CB4856 genetic background (CB4856xECA859) were significantly more resistant than their hyper-sensitive NIL counterpart, which is homozygous N2 for the chromosome III locus in the CB4856 genetic background (ECA859). The phenotype of this heterozygous strain was also similar to that of the CB4856 strain. The results of these crosses validate that genetic variation

between N2 and CB4856 on the left arm of chromosome III contributes to the nematode zinc response and indicate that the CB4856 allele conferred a dominant phenotype.

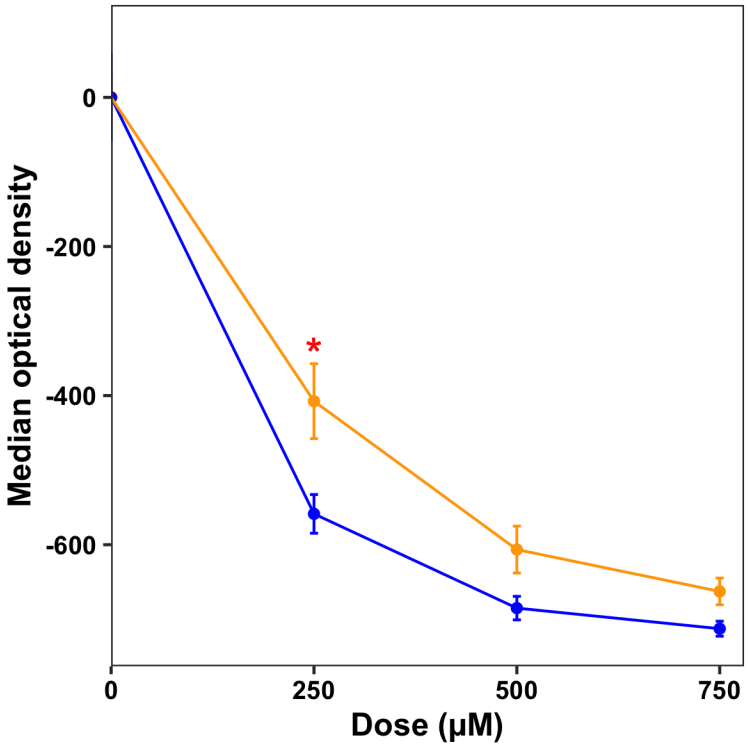


Figure 4-7: Dose response for modified HTA. Results from the zinc dose response with the modified HTA for median optical density (median.EXT). Drug concentration (μM) (x-axis) is plotted against phenotype subtracted from control (y-axis), colored by strain (CB4856: blue, N2: orange). A red asterisk indicates the dose selected for further analysis.

Because no previously identified zinc-related genes are in this interval, we investigated the composition of the genes in N2 to look for any obvious candidates that might underlie this QTL. We found 119 genes in this interval (**Table 4-2**). A change in phenotype is often observed when either genetic variation causes a change in the amino-acid sequence of the protein (protein-coding variation) or genetic variation causes a change in gene expression. Previously, whole-genome gene expression was measured in a set of 208 RIAILs derived from the N2 and CB4856 strains [115] and expression QTL (eQTL) mapping was performed [115, 69]. We used this dataset to find genes with an eQTL that maps to our region of interest. In total, we eliminated 19 genes that had no genetic variation in CB4856 and prioritized 62 genes that had protein-coding variation and/or an eQTL that mapped to this region (**Table 4-2**).

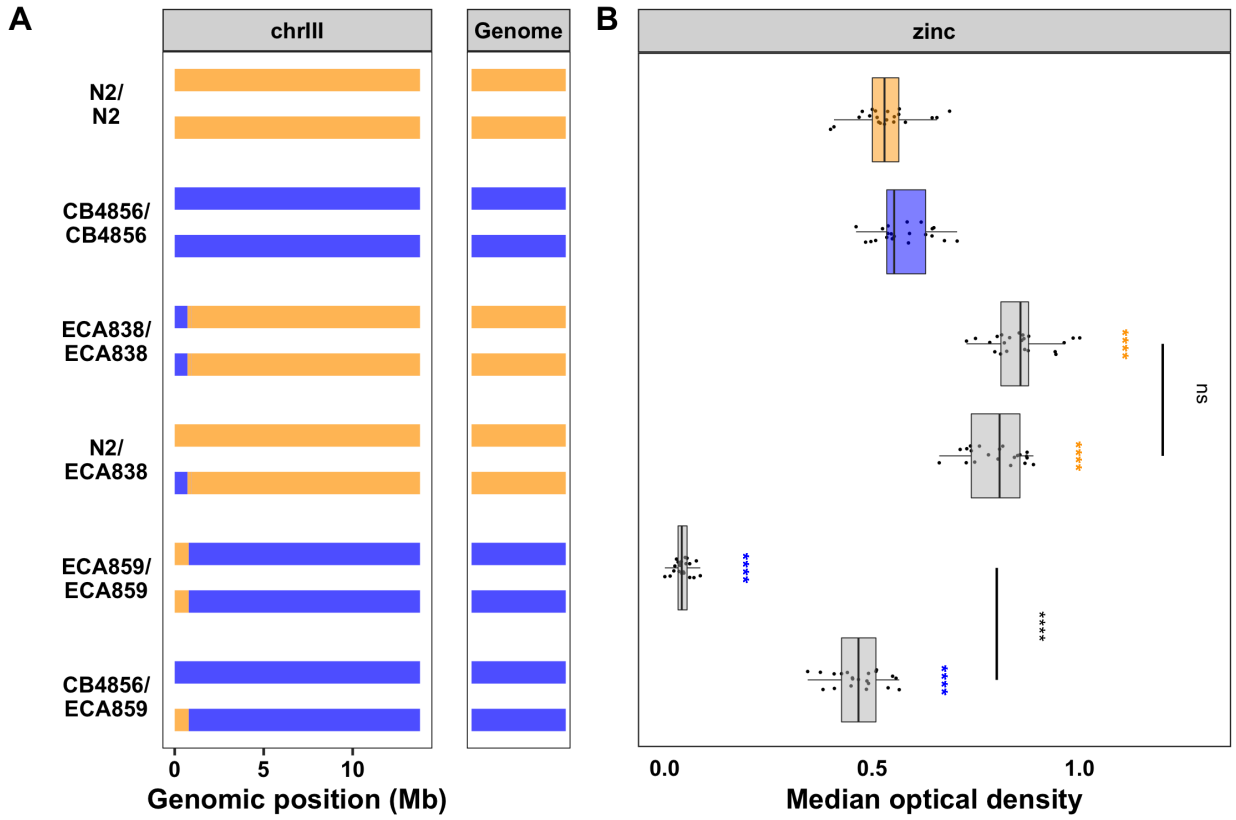


Figure 4-8: Dominance of chromosome III QTL. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome III (left) and in general for the rest of the chromosomes (right). Each rectangle represents a single copy of chromosome III. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). The N2 strain, which is normally resistant to zinc, was sick in this experiment. Statistical significance of each strain compared to its parental strain (ECA838/ECA838 and N2/ECA838 to N2 and ECA859/ECA859 and CB4856/ECA859 to CB4856) is shown above each strain and colored by the parent strain it was tested against (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

To narrow our list of genes further, we analyzed the functional descriptions and gene ontology (GO) annotations for all 62 candidate genes. A gene that is predicted to bind zinc and has protein-coding variation or variation in gene expression between N2 and CB4856 would be a high-priority candidate. We identified four genes that are predicted to bind zinc and a fifth gene that is regulated by a zinc finger transcription factor. Upon further inspection, one of these five genes (*sqst-5*) also had an eQTL that was originally assigned to the nearby pseudogene *ver-2* (**Figure 4-9A**, **Figure 4-10**). RIAILs with the N2 allele at the *sqst-5* locus have significantly higher expression of the gene than those with the CB4856 allele (**Figure 4-9B**). We previously showed that mediation analysis can be a useful tool to

Table 4-2: Genes in QTL intervals for chromosome III and V

QTL interval (bp)	No genetic variation[a]	Protein-coding variation and/or eQTL[b]	Other genetic variation[c]	Other eQTL that map to interval[d]	Total
III:4,664-597,553	19	55	45	7	**126**
V:9,620,518-10,511,995	183	49	97	24	**353**
V:10,511,995-11,345,444	215	45	115	19	**394**

[d]Genes within genomic interval with no genetic variation
[d]Genes within genomic interval with protein-coding variation and/or an eQTL that maps to this interval
[d]Genes within genomic interval with non-protein-coding variation and no eQTL that maps to this interval
[d]Genes outside genomic interval with an eQTL that maps to this interval

link variation in gene expression with drug-response phenotypes [69]. We used the standard high-throughput assay to measure zinc responses for 121 of the 208 RIAILs with gene expression data and performed mediation analysis for each of the 17 genes with an eQTL in the region. The mediation estimate for *sqst-5* was the strongest hit (**Figure 4-9C**). Together, these results suggest that genetic variation on chromosome III causes a decrease in expression of *sqst-5* that leads to increased zinc resistance.

4.5.4 Variation in *sqst-5* underlies differences in zinc responses

To test the function of *sqst-5* in the zinc response, we constructed two independently derived strains harboring large deletions of *sqst-5* in both the N2 and CB4856 genetic backgrounds. Because RIAILs with the CB4856 allele (which was associated with higher resistance to zinc) have lower expression of *sqst-5* (**Figure 4-9B**), we expected *sqst-5* deletions in the CB4856 genetic background might cause little or no change in zinc resistance. Alternatively, we expected *sqst-5* deletions in the N2 genetic background might cause increased zinc resistance. Surprisingly, we found that deletions of *sqst-5* had no effect in either background (**Figure 4-11**). However, the increased sensitivity of the N2 allele in the CB4856 genetic background (ECA859) always had a much larger effect than the increased resistance of the CB4856 allele in the N2 background (ECA838) (**Figure 4-8**, **Figure 4-5**). To take advantage of this sensitization, we deleted *sqst-5* in the hyper-sensitive NIL strain that contains the N2 *sqst-5* allele in the CB4856 genetic background (ECA859). We hypothesized that deleting *sqst-5* in the hyper-sensitive

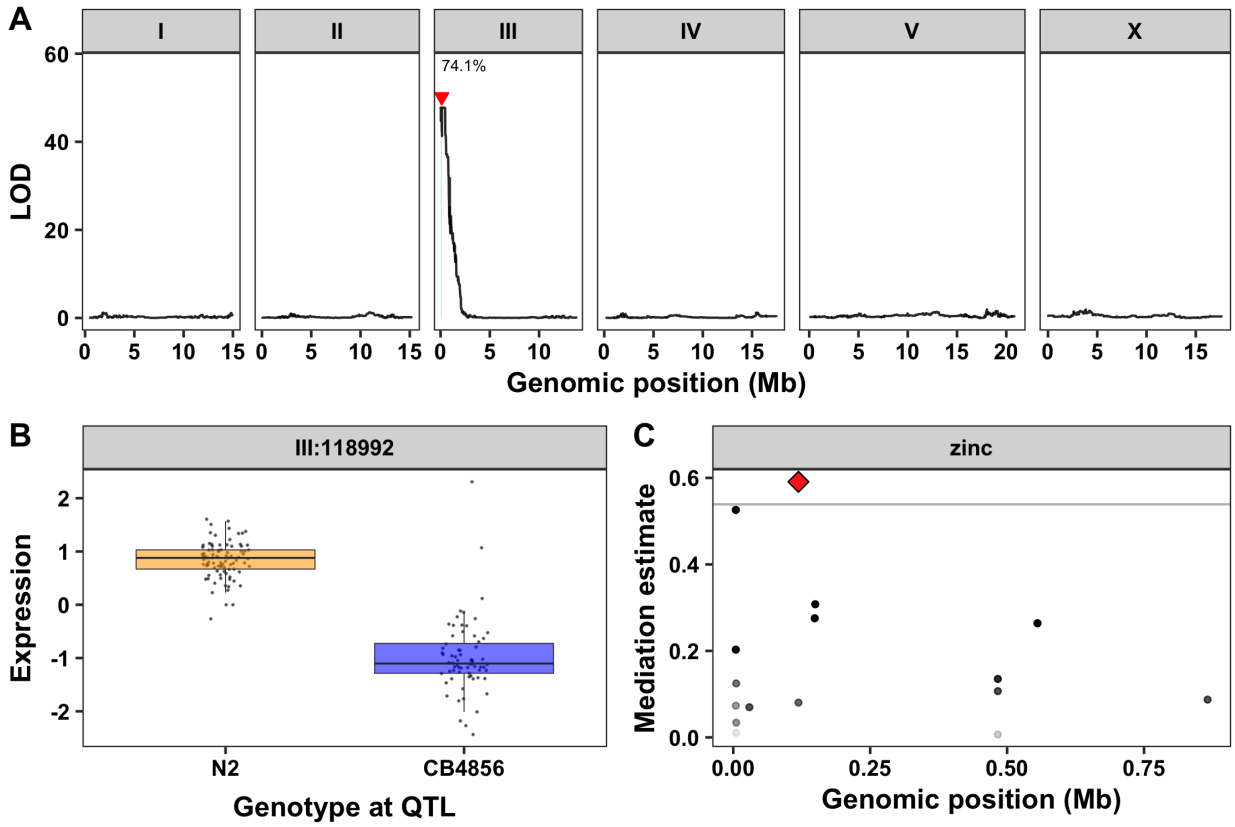


Figure 4-9: Expression QTL mapping and mediation analysis for *sqst-5*. (A) Results from the linkage mapping using expression of *sqst-5* as a quantitative trait. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. The significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by the QTL is shown above the QTL. **(B)** The expression of *sqst-5* (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange, and strains with the CB4856 allele are colored blue. **(C)** Mediation estimates calculated as the indirect effect that differences in expression of each gene plays in the overall phenotype (y-axis) are plotted against genomic position of the eQTL (x-axis) on chromosome III for al 17 genes with an eQTL in the drug-response QTL confidence interval. The 95th percentile of the distribution of mediation estimates is represented by the horizontal grey line. The confidence of the estimate increases (*p*-value decreases) as points become more solid. *sqst-5* is represented by a red diamond.

NIL would make this strain less sensitive to zinc (more similar to the CB4856 phenotype). As expected, we observed a significant increase in resistance for these deletions compared to the NIL (**Figure 4-12**), indicating a role for *sqst-5* in the *C. elegans* zinc response.

To provide further evidence that natural variation between N2 and CB4856 in *sqst-5* underlies the chromosome III QTL, we measured the optical density of individuals hemizygous for the N2 *sqst-5* allele in the hyper-sensitive NIL genetic background (ECA2517xECA859) in response to zinc. If a

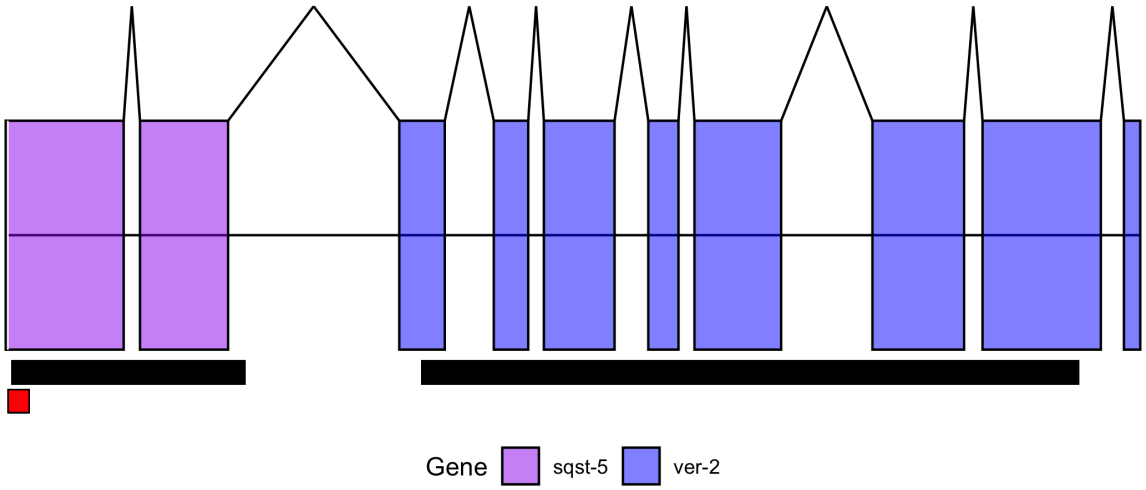


Figure 4-10: The gene *sqst-5*, not *ver-2*, has an eQTL. Original gene model for *ver-2* is shown with colored boxes representing exons connected by lines representing introns. Exons are colored blue for the new gene model for *ver-2* and purple for *sqst-5*. The black rectangles below represent approximate locations of CRISPR-mediated deletions of *ver-2* or *sqst-5*. The location of the microarray probe is designated as a red rectangle below the plot.

loss-of-function allele of *sqst-5* in the CB4856 strain is responsible for the variation in zinc response between N2 and CB4856 (**Figure 4-11**), then this hemizygous strain should show the same sensitivity as both the CB4856 strain and the strain with the homozygous deletion of *sqst-5* in the hyper-sensitive NIL genetic background (ECA2517). We observed that the strain hemizygous for the N2 *sqst-5* allele was indeed more resistant than the hyper-sensitive NIL and similar in sensitivity to the CB4856 strain (**Figure 4-13**). This result recapitulated the result of the dominance assay (**Figure 4-8**), suggesting that a loss-of-function allele of *sqst-5* conferred a dominant resistance phenotype. A dominant phenotype caused by a loss-of-function allele is, most times, caused by haploinsufficiency. Therefore, the zinc-response phenotype driven by the single functional N2 *sqst-5* allele is not sufficient to produce the hyper-sensitive phenotype of the NIL with two functional N2 alleles.

4.5.5 A natural deletion in *sqst-5* is conserved across wild isolates

We next searched for specific genetic variants in *sqst-5* that could lead to a loss-of-function allele in CB4856. We investigated the sequence read alignments of the N2 and CB4856 strains at the *sqst-5* locus using the Variant Browser on CeNDR (elegansvariaton.org) [63] and observed a putative large deletion in the second exon. We confirmed that this deletion is 111 bp (N2 coordinates: chrIII:147,076-

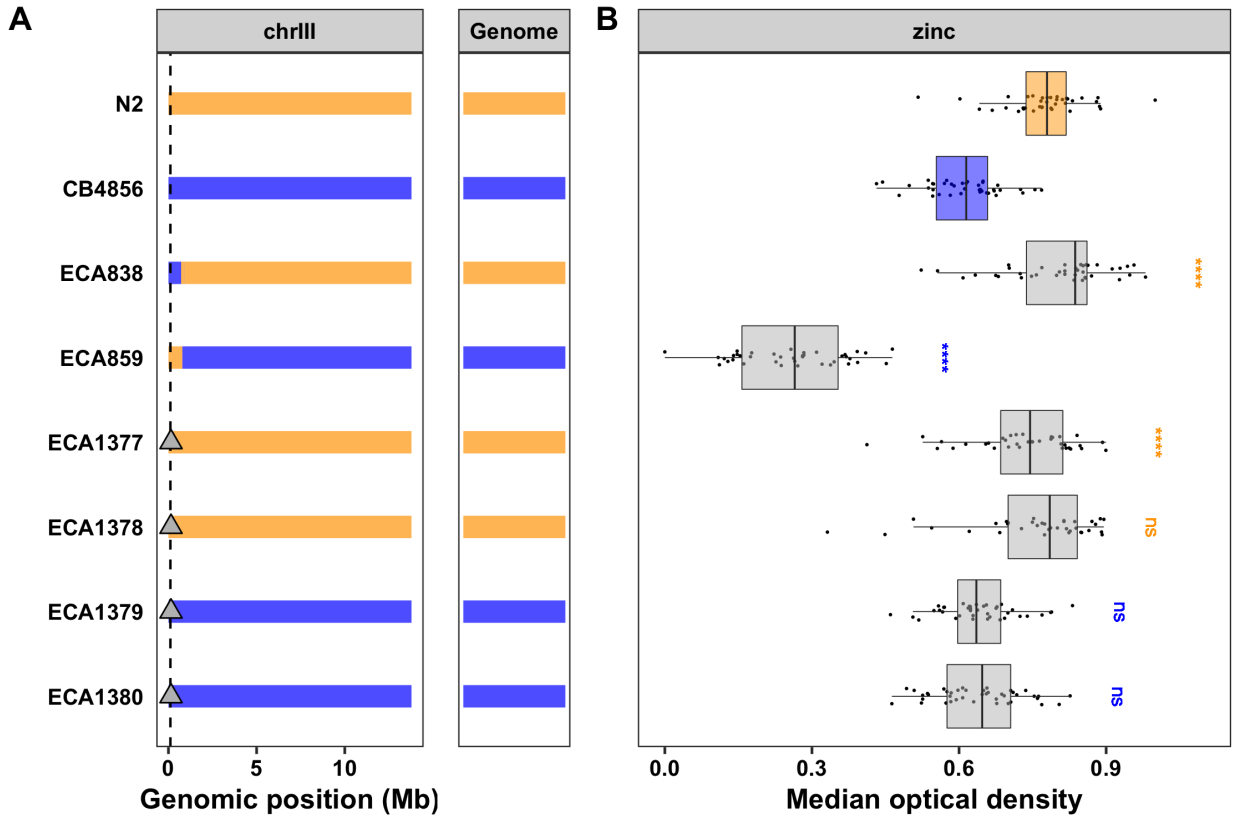


Figure 4-11: Testing the role of *sqst-5* in the zinc response. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome III (left) and in general for the rest of the chromosomes (right). The dashed vertical line represents the location of *sqst-5* and grey triangles represent *sqst-5* deletions. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each strain compared to its parental strain (ECA838, ECA1377, and ECA1378 to N2 and ECA859, ECA1379, and ECA1380 to CB4856) is shown above each strain and colored by the parent strain it was tested against (ns = non-significant (*p*-value > 0.05); *, **, ***, and *** = significant (*p*-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

147,186 bp) using a whole-genome alignment between the N2 reference genome and the CB4856 genome recently assembled using long-read sequencing [62] (**Figure 4-14**). We ran gene prediction algorithms on the CB4856 sequence, but no gene was predicted. The SQST-5 protein in the N2 strain has a single characterized protein domain: a Zinc finger, ZZ-type (Wormbase.org, WS275). The ZZ-type domains are predicted to bind two zinc ions using a repeated conserved motif of Cys-X2-Cys and are also important for protein-protein interactions [265, 266]. Interestingly, when we overlaid the location of the ZZ-type domain with the CB4856 alignment, we discovered that the 111 bp deletion spans most of the ZZ-type domain, including the essential Cys-X2-Cys motif (**Figure 4-14A**). Because this domain is important for binding zinc ions, this result suggests that even if CB4856 expresses low

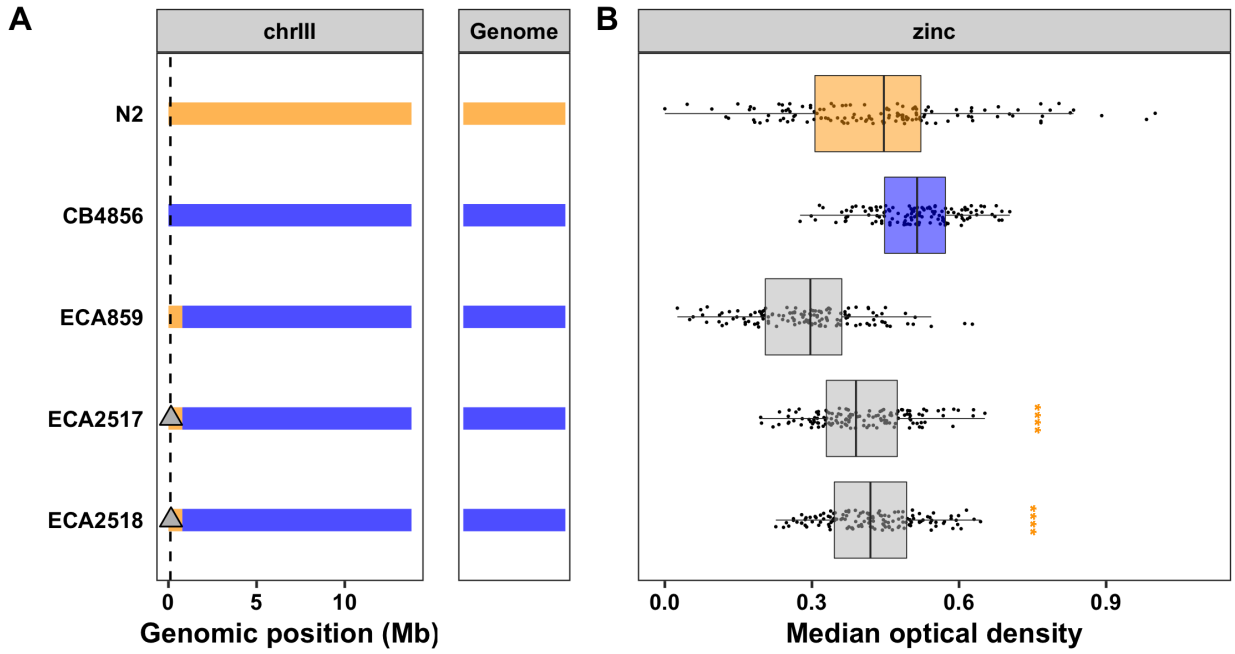


Figure 4-12: Isolating the effect of *sqst-5* in the zinc response. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome III (left) and in general for the rest of the chromosomes (right). The dashed vertical line represents the location of *sqst-5* and grey triangles represent *sqst-5* deletions. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). The N2 strain, which is usually resistant to zinc, was sick in this experiment. Statistical significance of each strain compared to ECA859 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

levels of SQST-5, it is unlikely to bind zinc at the same level as strains with a complete ZZ-type domain.

We next investigated structural variation across a panel of 328 wild isolates to ask if this deletion is unique to the CB4856 strain or common across many wild strains. We identified 31 additional strains with the same 111 bp deletion as CB4856 by manual inspection using the Variant Browser on CeNDR (**Table S4-1**). We also identified 25 strains that harbored low sequence identity with the N2 reference genome, indicating that these strains might contain structural variation different from the deletion in the CB4856 strain (**Table S4-1**). We assessed the genetic relatedness of these strains by constructing a neighbor-joining tree for all 328 wild isolates using the single nucleotide variants near and within *sqst-5*. All 32 strains with the predicted deletion in *sqst-5* cluster together (**Figure 4-14B**), suggesting these strains inherited this deletion from a common ancestor. These strains were not isolated from a single location, but rather spread geographically across Europe and the Pacific Rim (**Figure 4-15**). Furthermore, 24 of the 25 strains with other putative structural variation in *sqst-5* also cluster together

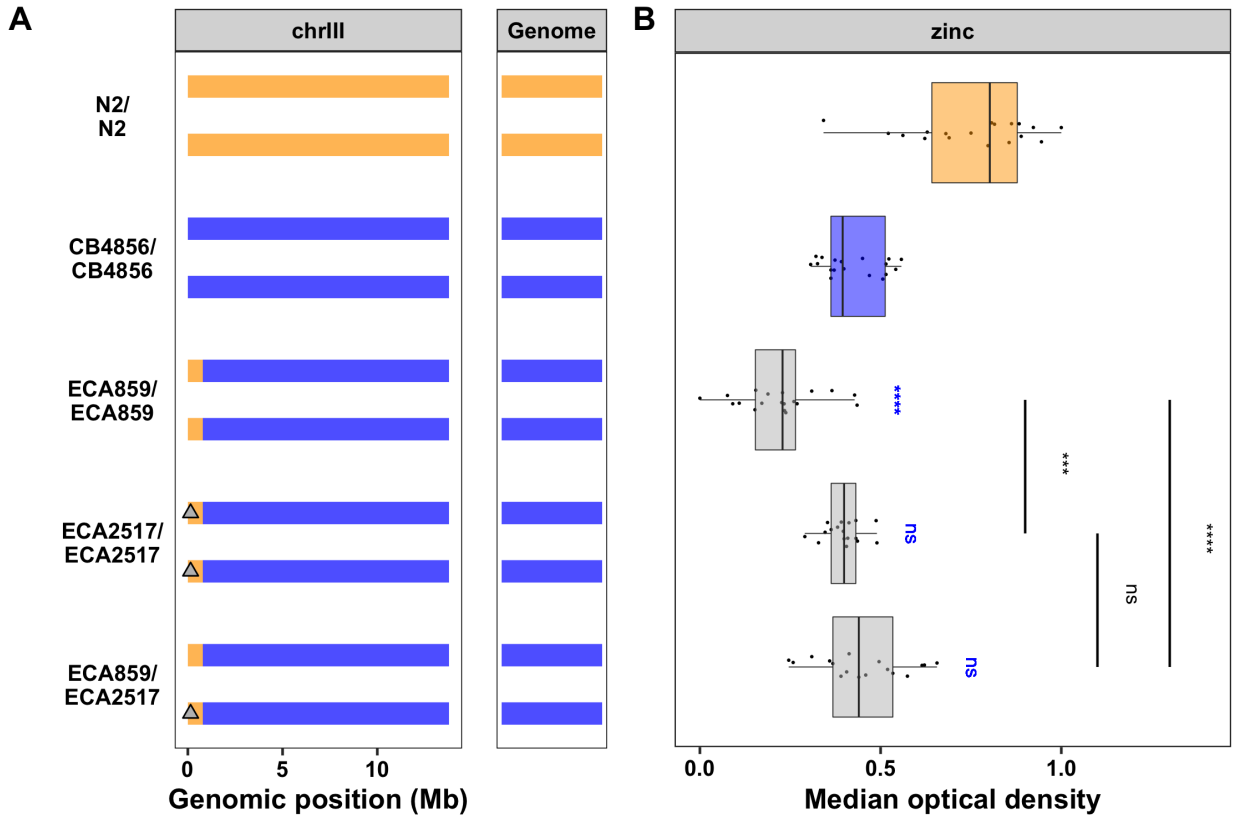


Figure 4-13: The gene *sqst-5* confers zinc sensitivity. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome III (left) and in general for the rest of the chromosomes (right). Each rectangle represents a single copy of chromosome III, and grey triangles represent *sqst-5* deletions. **(B)** Normalized residual median optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). The N2 strain, which is normally resistant to zinc, was sick in this experiment. Statistical significance of each strain compared to CB4856 is shown above each strain and NIL pairwise significance is shown as a bar above strains (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

separately from the strains with the 111 bp deletion (**Figure 4-14B**). This result suggests that this second group of strains also share a common ancestor that harbored variation in *sqst-5*. Additionally, strains with the deletion and strains with the other haplotype are sometimes found in nearby locations (**Figure 4-15**). Regardless, the 111 bp deletion and putative other structural variants might cause loss of *sqst-5* function.

To test if a loss-of-function allele of *sqst-5* correlates with zinc resistance among our panel of wild isolates, we measured animal development (length, optical density, and normalized optical density) and reproductive ability (brood size) of 81 strains in response to zinc. Including CB4856, we tested nine strains with variation in *sqst-5*: four strains with the 111 bp deletion and five strains with the other

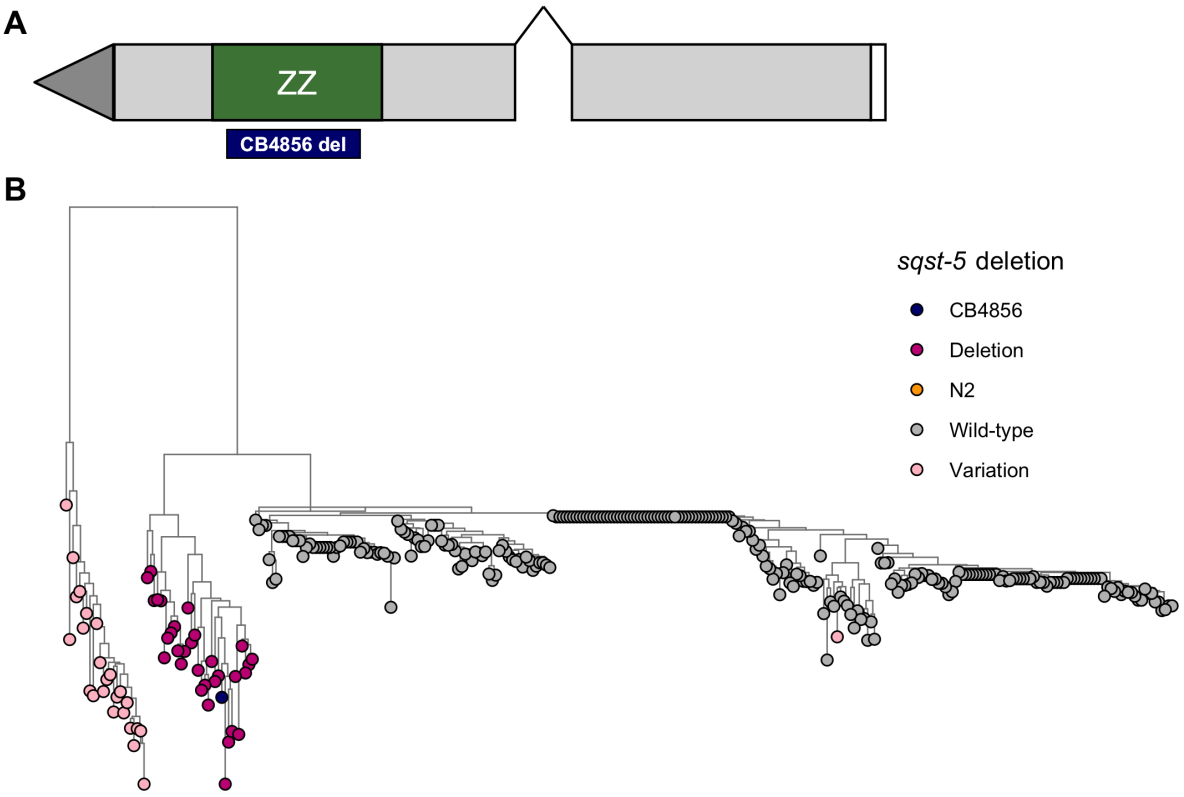


Figure 4-14: Natural genetic variation in *sqst-5*. (A) Gene model of *sqst-5*. Grey rectangles represent exons, and connecting black lines represent introns. The ZZ-type domain is indicated by a green rectangle, and the location of the natural 111 bp deletion in CB4856 is indicated below with a blue rectangle. **(B)** Neighbor-joining tree indicating genetic relatedness between wild isolates at the *sqst-5* locus. Branch lengths indicate the rate, the tree is midpoint rooted. Tips are colored by the variation haplotype at *sqst-5*: (Wild-type: grey, N2: orange, CB4856: navy, Deletion: magenta, Other putative structural variation: light pink).

putative structural variation. On average, these nine strains were more resistant than the rest of the population (**Figure 4-16A**, **Figure S4-2**), and variation in *sqst-5* explained up to 11.5% (median.EXT; *p*-value = 0.0019) of the total variation in zinc responses among the wild isolates. Genome-wide association mapping identified eight small-effect QTL across the four zinc-response traits (**Figure S4-2**). One trait, normalized optical density, had a QTL on the left arm of chromosome III (**Figure 4-16B,C**). The proximity of this QTL to *sqst-5* suggests that natural variation in *sqst-5* likely also contributes to variation in response to zinc among a panel of wild *C. elegans* strains.

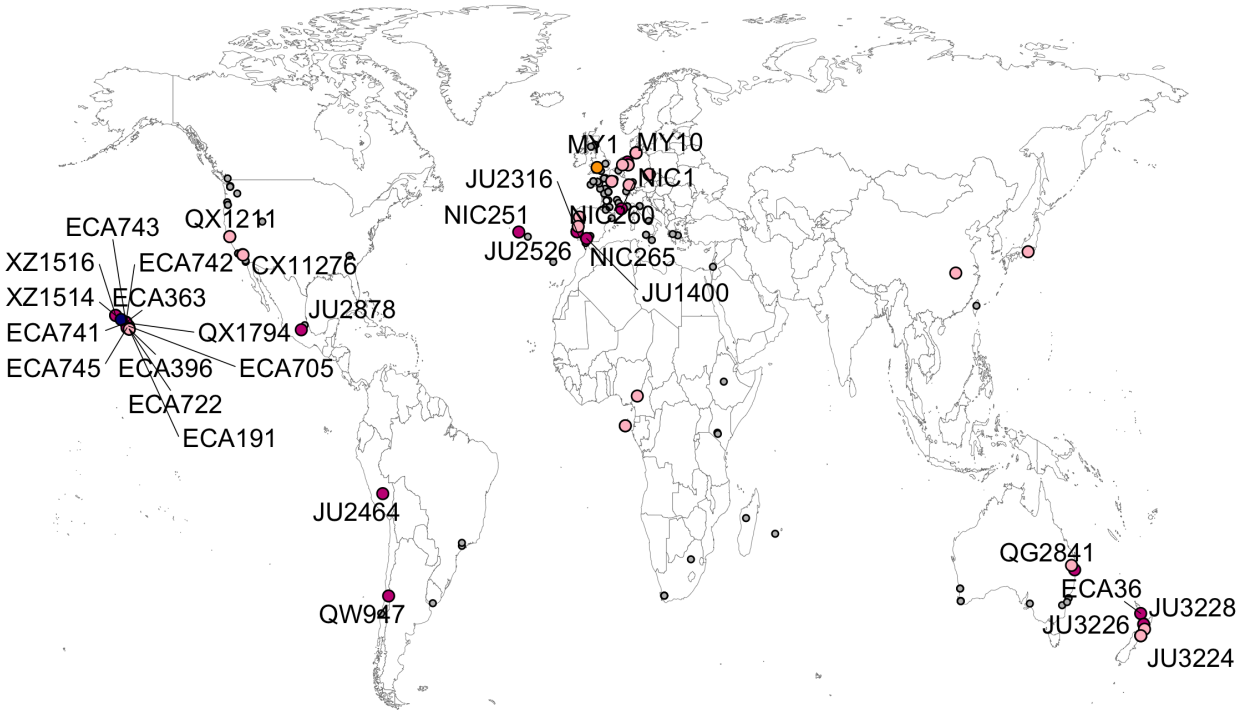


Figure 4-15: Map of wild isolates. Strains are colored by the variation haplotype at *sqst-5* (Wild-type: grey, N2: orange, CB4856: navy, Deletion: magenta, Other putative structural variation: light pink). Strains with the deletion are labeled.

4.6 DISCUSSION

We used linkage mapping to identify four QTL in response to high levels of exogenous zinc. We validated the two QTL with the largest effects on chromosomes III and V using near-isogenic lines. The QTL on chromosome V was further dissected into at least two additive loci that became difficult to narrow further. Mediation analysis was performed for the genes with expression variation that overlaps with the QTL on chromosome III, and a single gene (*sqst-5*) was identified whose variation in expression between N2 and CB4856 is correlated with differences in responses to zinc. The CB4856 strain harbors a natural deletion of this gene. We deleted the N2 version of *sqst-5* using CRISPR-Cas9 genome editing and showed that strains without *sqst-5* were significantly more resistant to exogenous zinc than strains with a functional copy of *sqst-5*, suggesting a new role for *sqst-5* in zinc homeostasis. In addition to CB4856, several other wild isolates were found to have a

putative independent structural variant in *sqst-5* that is loosely correlated with resistance to zinc. These strains cluster genetically into two distinct groups, suggesting that functional variation in *sqst-5* has emerged multiple times. These results demonstrate the power of leveraging natural genetic variation to identify novel genes in a toxin-response pathway and suggest mechanisms for how high exogenous zinc can be mitigated in natural environments.

4.6.1 A complex genetic architecture underlies differences in zinc responses

The identification of multiple QTL in response to excess zinc is not surprising, as zinc is an essential trace element and up to 8% of *C. elegans* genes have been predicted to encode proteins that bind zinc [247]. In particular, 28 genes encode putative zinc transporters and an additional 15 genes have been identified via mutagenesis screens in N2 that promote either zinc resistance or sensitivity [233, 243]. We identified two QTL that contain at least one gene that was previously found to be involved in the nematode zinc response and an additional two QTL that do not contain any genes previously found to affect zinc responses (**Table 4-1**). Several of the genes with previously described roles in the zinc response are located on chromosome X. However, only three of these genes (*hke-4.2*, *hizr-1*, and *elt-2*) have protein-coding variation in CB4856. The results of the linkage mapping experiment (**Figure 4-2B**) identified a broad peak on chromosome X, but we were unsuccessful validating this QTL using NILs (**Figure 4-5**) likely because the QTL could contain multiple small-effect loci that could each act in opposite directions. By contrast, we were able to validate that genetic variation on chromosome V contributes to the nematode zinc responses (**Figure 4-6**). However, as more NILs were tested with smaller introgressions, we observed a fractionation of the QTL into at least two small-effect loci. Several previous studies that aimed to deeply validate a single QTL have instead identified many tightly linked antagonistic QTL underlying the major QTL [144, 128, 145, 267, 131, 268, 269, 270, 122]. Unfortunately, as each QTL fractionates into several QTL, the individual effect sizes become smaller and our ability to accurately interpret signal from noise becomes more difficult. This polygenic nature of complex traits is a major roadblock in going from QTL to QTG [7, 22, 21].

Although we were unable to identify specific genes on chromosome V that contribute to the

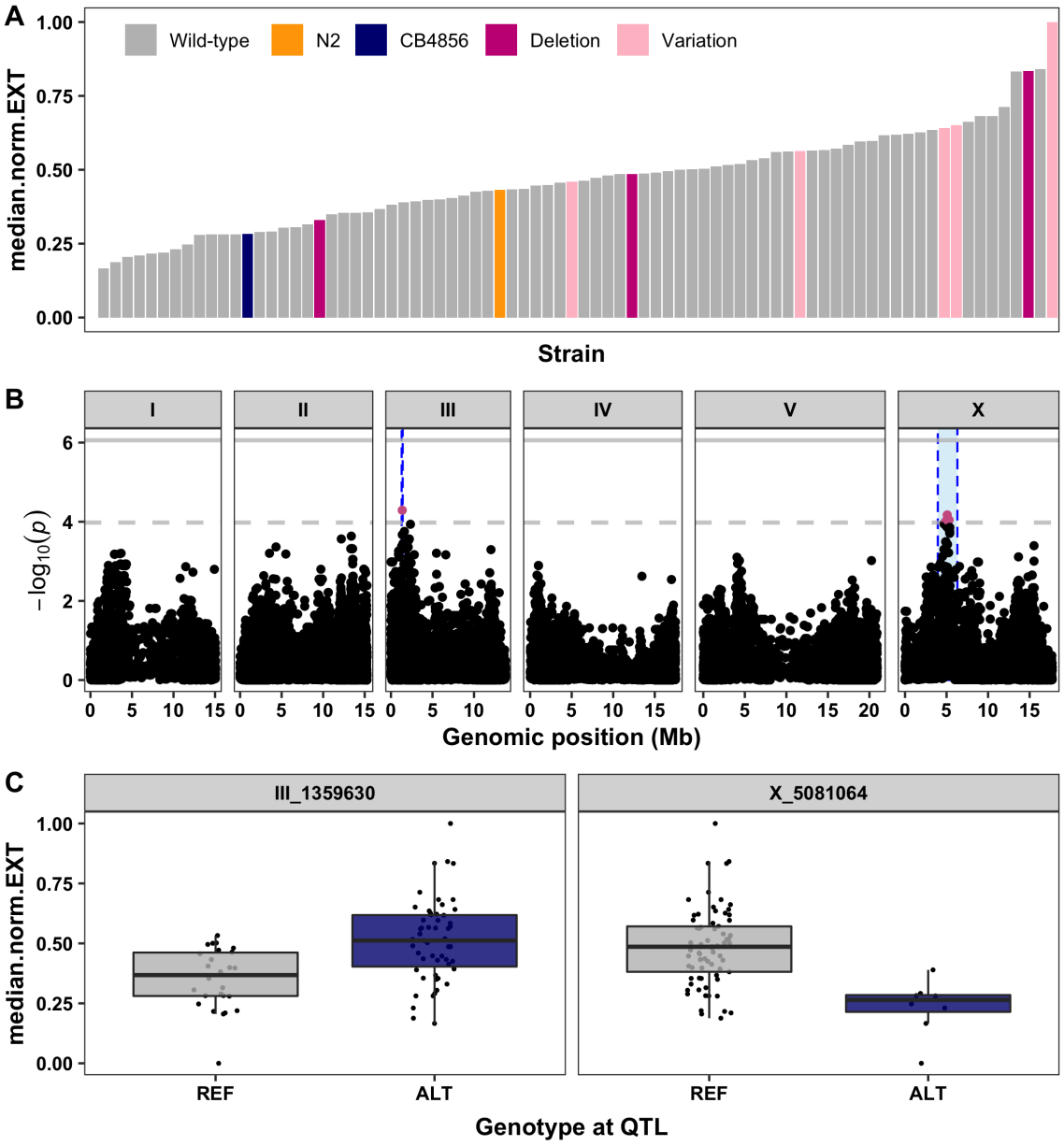


Figure 4-16: GWA mapping suggests common variation at *sqst-5* is associated with differences in zinc responses among wild isolates. **(A)** Normalized residual median normalized optical density (median.norm.EXT, y-axis) of 81 wild isolates (x-axis) in response to zinc supplementation. Strains are colored by the parental strains N2 (orange) and CB4856 (blue) or by the *sqst-5* variation (Deletion: magenta, other variation: light pink) **(B)** GWA results are shown. Genomic position (x-axis) is plotted against the -log10(*p*) value (y-axis) for each SNV. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance threshold designated by the dotted grey line. The solid grey line represents the more stringent Bonferroni significance threshold. The genomic regions of interest that pass the significance threshold are highlighted by blue rectangles. **(C)** For each QTL, the normalized residual median normalized optical density (median.norm.EXT, y-axis) of strains split by genotype at the peak marker (x-axis) are plotted as Tukey box plots. Each point corresponds to a wild isolate strain. Strains with the N2 reference allele are colored grey, and strains with an alternative allele are colored navy.

nematode zinc response, we were able to narrow the QTL interval from 4.3 Mb (chrV:10,084,029-14,428,285) to 1.2 Mb (chrV:10,084,029-11,345,444) containing at least two loci that underlie differential responses to zinc. Of the 573 genes of interest (**Table 4-2**), we identified two high-priority candidate genes (*zhit-3* and *H27A22.1*) that are predicted to bind zinc and have protein-coding variation in the CB4856 strain. The gene *zhit-3* encodes a protein that is an ortholog of the human protein ZNHIT6 and contains a zinc finger HIT-type domain (Wormbase.org, WS275). The gene *H27A22.1* encodes a protein that is an ortholog of the human protein QPCTL with glutaminyl-peptide cyclotransferase (Wormbase.org, WS275). It is possible that genetic variation in one or both of these genes underlies the QTL on chromosome V. However, future studies are needed to confirm the role of these genes in the nematode zinc response.

Although zinc is the most biologically relevant heavy metal, other divalent cations with similar chemistries also play important roles in biological systems (copper, nickel, and iron) or are highly toxic (cadmium) [226, 227, 70, 233]. We previously performed linkage mapping for three of these heavy metals and found that QTL for zinc, copper, and nickel overlap on the right arm of chromosome IV (**Figure 4-17**) [70, 73, 122]. The overlap of this QTL with other heavy metal QTL suggests that perhaps the molecular mechanisms underlying these QTL are not specific to zinc. Furthermore, this QTL is in regions previously defined as a QTL hotspot where a single pleiotropic gene might control several toxin responses or several independent yet tightly linked genes might control different traits [73]. Regardless, the high-effect QTL on chromosome III seems to be unique to the zinc response, as none of the other metals have a QTL on chromosome III (**Figure 4-17**).

4.6.2 Common genetic variation underlies differential responses to exogenous zinc

We discovered 31 additional wild strains with the same 111 bp deletion in *sqst-5* as found in the CB4856 strain and another 25 strains that show evidence of different structural variation. Long-read sequencing and local genome assembly of strains with this alternative variation are needed to fully define these haplotypes. Although these strains were isolated globally (**Figure 4-15**), phylogenetic analysis suggests that these strains comprise two common classes of variation at the *sqst-5* locus (**Figure 4-14B**). Strains from these two classes of variation are sometimes found in close geographical proximity, indicating a

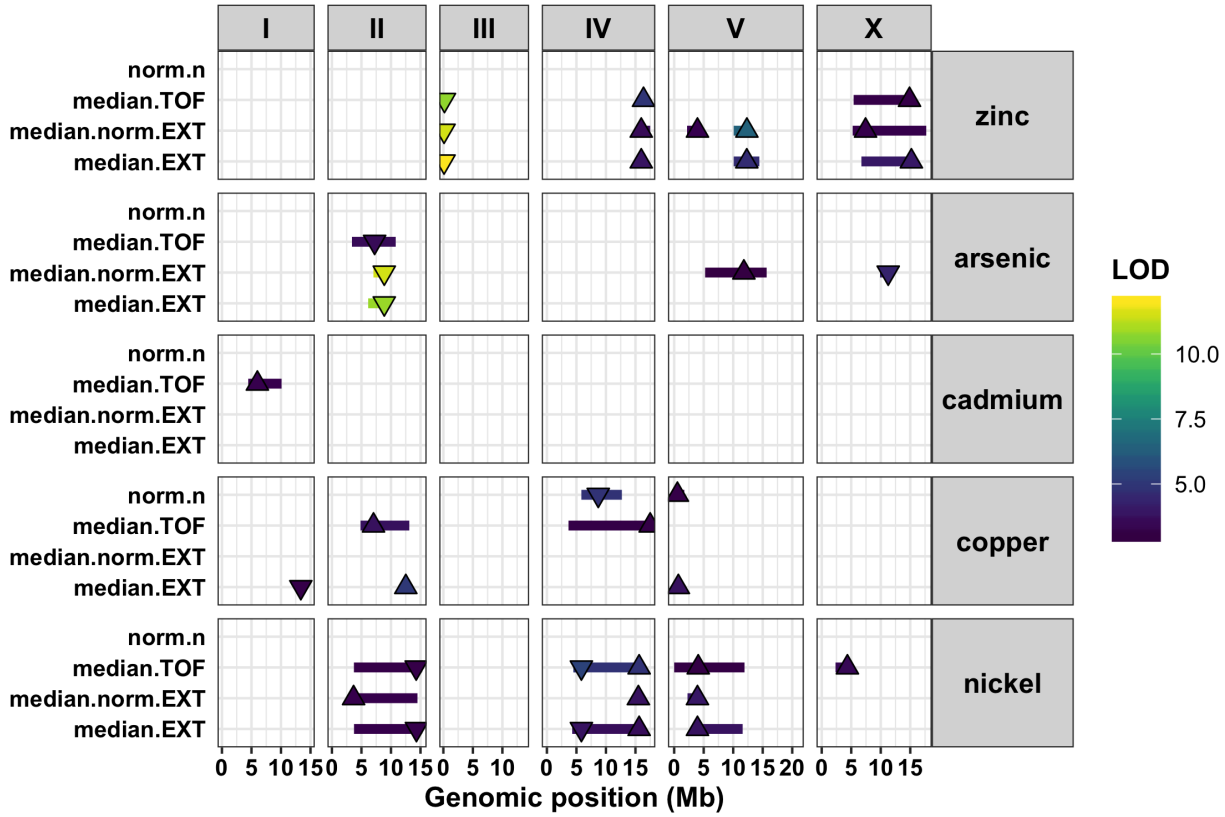


Figure 4-17: Linkage mapping summary for drug-response traits in response to five heavy metals. Genomic positions (x-axis) of all QTL identified from linkage mapping are shown for each drug-trait (y-axis). Each QTL is plotted as a triangle at the genomic location of the peak marker and a line that represents the 95% confidence interval. QTL with right side up triangles have a negative effect size (N2 allele is resistant), and QTL with upside down triangles have a positive effect size (CB4856 allele is resistant). QTL are colored by the logarithm of the odds (LOD) score, increasing in significance from purple to green to yellow.

possibility for convergent evolution in zinc resistance, perhaps in geographic regions with high levels of zinc in the environment. We require further investigation of metal contents in niches that contain *C. elegans* to connect environmental zinc levels to natural genetic variation.

We measured zinc responses for a subset of wild strains and found that variation in *sqst-5* could explain 11.5% of the total variation in response to zinc among the panel of wild isolates. Interestingly, N2 and CB4856 were among the more sensitive strains tested (**Figure 4-16**, **Figure S4-2**), suggesting the existence of several loci not found within the CB4856 strain that influence zinc responses. GWA mapping discovered several small-effect loci across the genome that were associated with zinc resistance or sensitivity (**Figure S4-2**). In particular, the QTL on chromosome III (nearby the *sqst-5*

locus) and chromosome X overlapped with QTL discovered using linkage mapping. Alternatively, a QTL on the right arm of chromosome III provides evidence of common genetic variation not present in the CB4856 strain that plays a role in the nematode zinc response. Because none of the known zinc-related genes are in this interval, this QTL might represent another novel gene that contributes to zinc resistance or sensitivity (**Table 4-1**). Our power to detect QTL would only improve with the phenotyping of more strains in the presence of high exogenous zinc.

4.6.3 SQST-5 might function to negatively regulate other sequestosome-related genes

We show that strains with a functional copy of *sqst-5* are more sensitive to zinc than strains with a large deletion of the gene, indicating that *sqst-5* negatively regulates the zinc response. Using BLASTp (Wormbase.org, WS275), we searched for paralogs and identified five other members of the sequestosome-related family (*sqst-1*, *sqst-2*, *sqst-3*, *sqst-4*, and *C06G3.6*) each containing a ZZ-type domain, like *sqst-5*. Two of these genes have been previously implicated in the nematode stress response. The gene *sqst-1* is upregulated in response to hormetic heat shock [271]. Both SQST-1 and the human ortholog, SQSTM1/p62, have been shown to bind to and target ubiquitinated proteins to an organelle (sequestosome) for subsequent degradation by autophagy [272, 271]. The ZZ-domain, particularly the zinc-coordinating Cys-X2-Cys residues, has been shown to be essential for this process [273, 274]. Additionally, *sqst-3* is expressed in response to exogenous cadmium [275], suggesting that the sequestosome-related family might be involved in divalent cation metal stress responses. If we connect these disparate results, then these genes could protect cells against zinc toxicity using sequestration and fusion with lysosomes. Alternatively, divalent cation metal stress could cause disruption of proteostasis and upregulation of sequestosome genes indirectly related to the specific metal stress. The role of this gene family in stress response has not been characterized using loss-of-function genetics, so we do not know whether the family is protective in response to exogenous stressors like zinc.

If the sequestosome-related genes do function to protect cells from high exogenous zinc, we would expect that loss-of-function of these genes should cause increased zinc sensitivity. However, loss-of-function of *sqst-5* caused increased zinc resistance, indicating that *sqst-5* might have a different

function than other sequestosome-related genes. Although the exact function of SQST-5 is unknown, it is predicted to have protein kinase C and K63-linked polyubiquitin modification-dependent protein binding activity (Wormbase.org, WS275). From previous studies, we know that the ZZ-type domain is important for protein-protein interactions [265], and we discovered that the natural 111 bp deletion in the CB4856 strain causes a loss of this zinc-binding domain. It is possible that SQST-5 could function as an inhibitor of mechanisms that mitigate exogenous zinc, potentially by binding to other sequestesome-related proteins and inhibiting their activities. When SQST-5 is removed, the protein partner is no longer inhibited and is available to bind or capture the excess zinc in the environment, thus reducing the toxicity induced by high exogenous zinc. Biological processes that are finely balanced in homeostasis often contain both positive and negative regulators. Functional studies to directly test the role of *sqst-5* in the autophagy pathway, both in control conditions and in the presence of exogenous zinc, are necessary to provide insights into its function as a negative regulator of zinc homeostasis. Likewise, zinc responses of animals with targeted deletions of the other sequestosome-related genes are needed to fully define the roles for these genes in the *C. elegans* zinc response pathway. Overall, this study leverages natural genetic variation to discover a novel gene that sensitizes nematodes to exogenous zinc, potentially by creating a negative feedback loop to regulate other sequestosome-related genes.

4.7 FUTURE DIRECTIONS

4.7.1 From QTL to QTG on chromosomes IV, V, and X

Although this study identified that genetic variation in the gene *sqst-5* causes differences in responses to exogenous zinc, there are still three other QTL that remain to be solved. We generated and tested NILs for the QTL on chromosomes IV and X, however we were unable to experimentally validate these genomic regions. This could be because the effect size of these loci are actually smaller than predicted or there could be multiple small-effect loci of opposite effects. Alternatively, the large-effect chromosome III QTL might be masking the effects of these loci. For example, ECA241 (which has the resistant N2 allele on chromosome IV introgressed into the sensitive CB4856 genetic background) actually contains two resistant loci (the CB4856 allele on chromosome III and the N2 allele on chromosome IV) and

two sensitive loci (the CB4856 alleles on chromosomes V and X). Just like the effect of *sqst-5* was heightened in the sensitive NIL genetic background, so too might this QTL. To test this theory, we could perform a genetic cross between the hyper-sensitive NIL ECA859 (N2 allele on chromosome III in a CB4856 genetic background) and the chromosome IV NIL ECA241. Selecting for the N2 alleles on both chromosome III and IV would create a strain in which the genetic background is completely sensitive to zinc except the region we are testing on chromosome IV. The same cross could be done with ECA828 and ECA859 to isolate the chromosome X QTL. However, generating a chromosome X substitution strain where the entire chromosome X is introgressed from the N2 strain into the CB4856 genetic background might be the first place to start.

If we are able to verify the QTL on chromosomes IV and X, there could be several followup experiments to identify the causal genes or variants underlying these QTL. For the chromosome X QTL, we could first test each of the three zinc-related genes within the QTL confidence interval that have protein-coding genetic variation in CB4856 (*hke-4.2*, *hizr-1*, and *elt-2*). The chromosome IV QTL has no known zinc-related genes, however it resides in a previously identified QTL hotspot where a single pleiotropic gene might control responses to several diverse toxins [73]. Furthermore, this region has also been shown to contain a large cluster of piwi-interacting RNAs (piRNAs) and variation in piRNAs between the N2 and CB4856 strains has been shown or suggested to play a role in nematode behavior [133] and drug responses [74]. It would be interesting to test the role of piRNAs in the zinc response by deleting a master regulator of piRNAs, *prg-1*. The biology of piRNAs is still understudied and connecting them to the zinc response, and potentially other toxin responses that also map to this region, would be novel and exciting.

There are also several candidate genes on chromosome V proposed in this study that could be explicitly tested for their role in the zinc response. Although there are hundreds of genes in our two narrowed QTL intervals, two strong candidate genes stand out. The gene *zhit-3* encodes a protein that is an ortholog of the human protein ZNHIT6 and contains a zinc finger HIT-type domain and the gene *H27A22.1* encodes a protein that is an ortholog of the human protein QPCTL with glutaminyl-peptide cyclotransferase. Both genes are predicted to bind zinc, however binding zinc does not ensure that these genes are involved in the nematode zinc response. Deletions of these genes using CRISPR-

Cas9 genome editing are needed. In addition to these candidate genes, this chromosome V QTL resides in a previously identified QTL hotspot [73]. We previously showed that the pleiotropic gene *scb-1* underlies responses to several double-strand-break-inducing chemotherapeutics with overlapping QTL on chromosome V [72, 69]. The gene *scb-1* is located within one of the two NIL-defined QTL on chromosome V (V:11.11 Mb), suggesting that variation in *scb-1* might also influence the zinc response. However, excess zinc is not known to cause double strand breaks. Furthermore, the direction of effect of the zinc-response QTL is opposite those underlied by *scb-1* (RIAILs with the N2 allele are resistant to zinc and sensitive to the double-strand-break-inducing chemotherapeutics like bleomycin). These results suggest that expression of *scb-1* might not be causal in the nematode zinc response. The only way to know for sure is to test *scb-1* deletions in response to high exogenous zinc.

4.7.2 Investigating the function of *sqst-5* and its mechanism of action in response to zinc

Although we identified that a deletion in *sqst-5* causes resistance to exogenous zinc, many questions still remain on the function of this gene and its mechanism of zinc resistance. This 111 bp deletion covers most of the ZZ-type domain, including the essential Cys-X2-Cys zinc-binding motif. It would be interesting to know if this deletion causes zinc resistance because it causes a loss-of-function allele and the gene is lowly expressed or because it removes the ZZ-type domain. Targeted genome-editing of one or more essential residues in the ZZ-type domain (perhaps a Cys > Ala) would eliminate the zinc-binding ability of *sqst-5* and would allow us to test if this function is necessary for responding to high exogenous zinc. To ensure that the targeted edits are reducing zinc binding, it would be essential to first measure zinc binding of the SQST-5 protein in both the N2 and CB4856 genetic backgrounds. Because the N2 strain has a functional copy of *sqst-5*, we would expect that the N2 copy of SQST-5 binds more zinc than the CB4856 copy of SQST-5.

The sequestosome-related gene family has been previously implicated in the nematode stress response. The gene *sqst-1* is upregulated in response to heat shock and *sqst-3* is upregulated in response to cadmium exposure. Interestingly, SQST-1 is known to bind to ubiquitinated proteins through its ZZ-type domain and deliver them to the autophagosome for degradation. Although no direct functional studies have been performed, it is thought that increased zinc would cause an

increase in SQST-1 which may sequester zinc, protecting cells from its toxicity. However, we observe the opposite interaction with SQST-5, as strains with a functional copy of this protein are more sensitive to high exogenous zinc. To fully understand this pathway, it is necessary to test how deletions of SQST-1 affect zinc response in our assay. Furthermore, it would be interesting to test the autophagy response in *sqst-5* mutants both in the presence and absence of zinc. One study showed that SQST-1 was upregulated in response to excess zinc in murine macrophage cells [276]. However, no previous studies have made a direct association between zinc and the sequestosome-related genes in autophagy.

4.7.3 Expanding natural variation studies to include more strains and provide evolutionary context of zinc resistance

We performed a GWA mapping for zinc resistance in 81 strains and found several QTL that both overlapped with linkage mapping or were unique. However, we now have 328 wild isolate strains with more genetic diversity than before [59]. Phenotyping more of these diverse strains, with and without variation in *sqst-5*, could tell us more about how *C. elegans* respond to zinc in their natural environments. Finally, it would be interesting to further investigate the evolutionary role of *sqst-5* in shaping zinc resistance in the wild. Using the 328 wild isolates as well as two or more closely related *Caenorhabditis* species such as *C. brigssae* and *C. remanei*, we could estimate inter- and intra-species selective pressure by calculating statistics such as Ka/Ks and Tajima's D [75]. This QTL lies in a highly divergent region of the *C. elegans* genome [60, 62, 277]. It is thought that these punctuated divergent regions might contribute to nematode diversity and evolution in a species with large chromosomal sweeps and relatively low genetic diversity [79, 277]. Because zinc is a natural environmental stressor, it would make sense that nematodes have evolved mechanisms to respond to high or low environmental zinc. Although limited data exists on zinc levels in the soil worldwide, perhaps future field collection trips could include a soil sample to test for levels of zinc and other metals.

Similarly, it could be interesting to correlate the natural levels of intracellular zinc with resistance to high exogenous zinc. The first step could be to test total zinc content in the N2 and CB4856 strains

using ICP-MS/MS and look to see if there is a significant difference. This method has been used to quantify differences in zinc content between control and mutant strains and could be easily adapted to test a diverse set of wild isolates in the presence and absence of high exogenous zinc [278]. If there is a difference in zinc content between the N2 and CB4856 strains, zinc content could next be measured for each of the NILs on chromosomes III and V and strains with *sqst-5* deletions. This experiment would determine if genetic variation at these loci control both intracellular zinc content and the individual's zinc response, suggesting the two might be functionally related. Alternatively, the zinc content for the RIAILs and/or the wild isolates could be measured and used as a phenotypic trait for QTL mapping using linkage or association mapping, respectively, to identify the genetic basis for intracellular zinc content. Together, these experiments could add to the knowledge of the mechanism behind which natural genetic variation contributes to the nematode zinc response.

4.8 CONTRIBUTIONS

During this work, I was funded by the Cell and Molecular Biology of Disease training program and the NSF-Simons Center for Quantitative Biology at Northwestern University. There are so many people to thank for the success of this project. Dr. Mostafa Zamanian, Dr. Stefan Zdraljevic, Kimberly Collins, and Dr. Shannon Brady all made NILs that were used in this project. I would also like to thank Katherine Introcaso and Britney Sun for their help and hard work cleaning NILs, genotyping CRISPR deletions, and setting up sorter assays. Dr. Stefan Zdraljevic also ran one of the early sorter assays for this project. Dr. Lewis Stevens performed sequence alignments and analysis. Dr. Robyn Tanny and Dr. Dan Cook did all the whole-genome sequencing and analysis of the NILs and RIAILs. Finally, Dr. Shannon Brady, Dr. Daehan Lee, Dr. Timothy Crombie, Dr. Erik Andersen, Katherine Introcaso, Clayton Dilks, and Loraina Stinson all played pivotal roles of support at varying times throughout this process.

4.9 CITATION

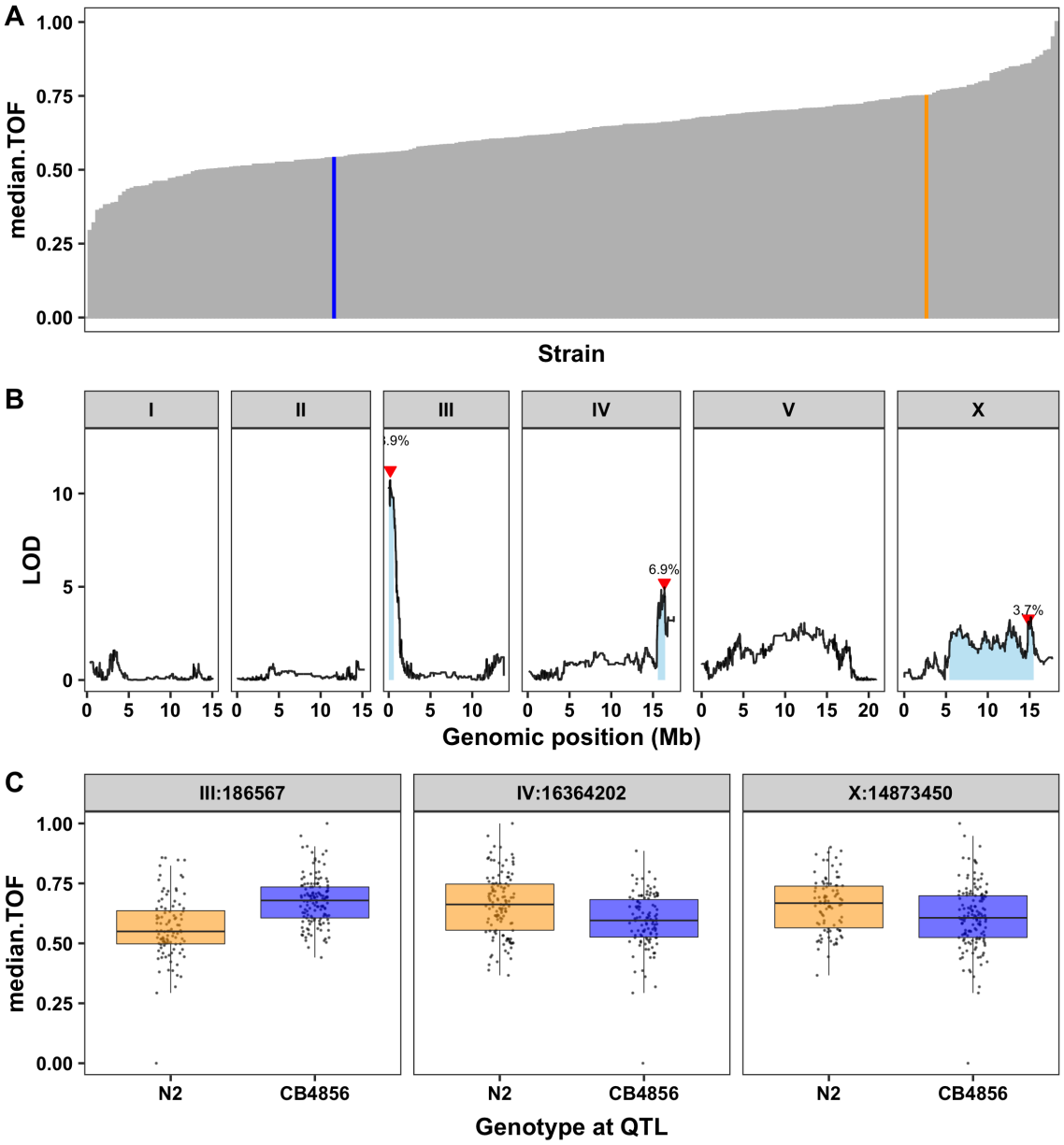
Natural variation in the sequestosome-related gene, *sqst-5*, underlies zinc homeostasis in *Caenorhabditis elegans*

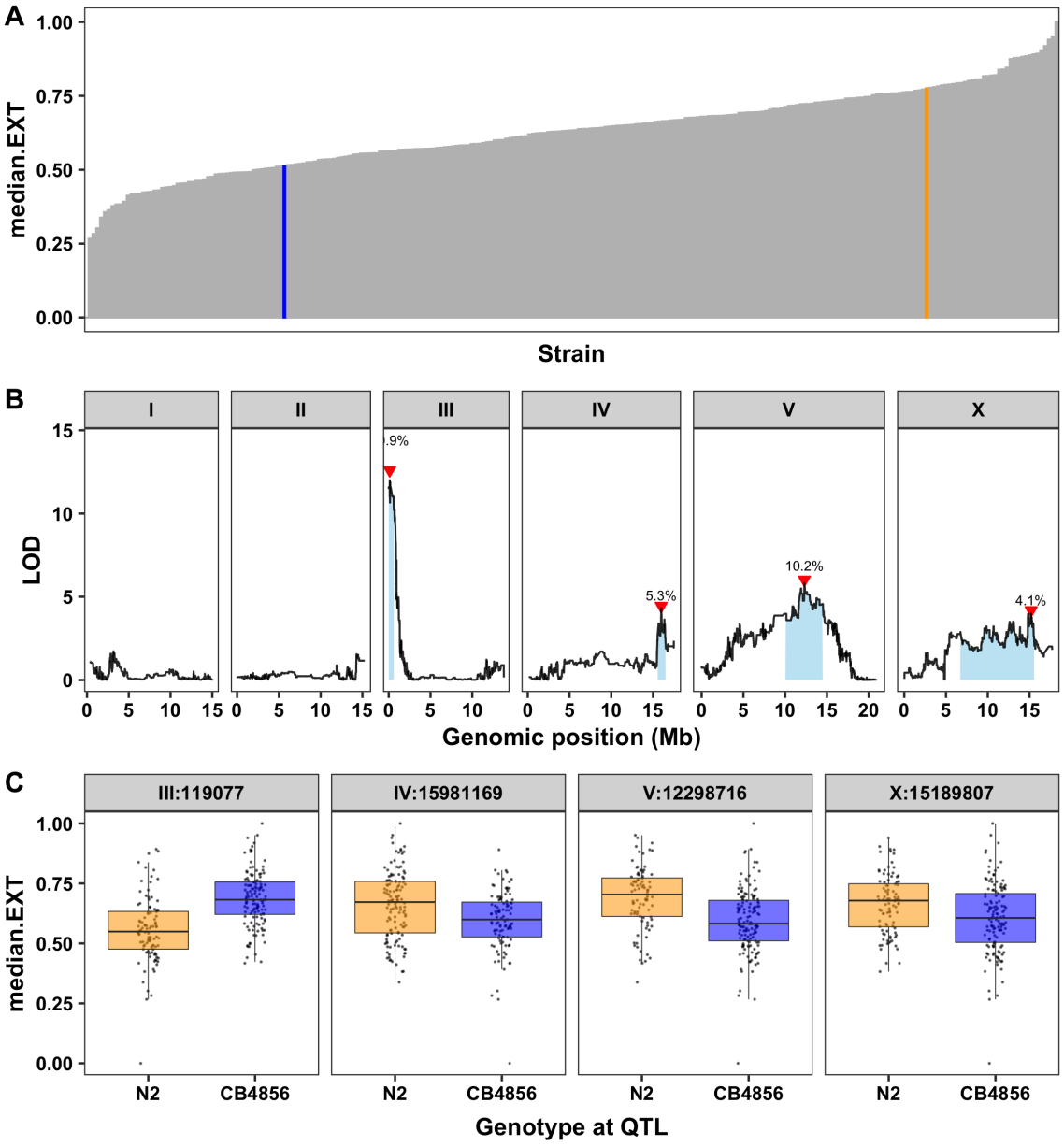
Kathryn S. Evans[1,2], Stefan Zdraljevic[3], Lewis Stevens[1], Kimberly Collins[1], Robyn E. Tanny[1], and Erik C. Andersen[1,3]

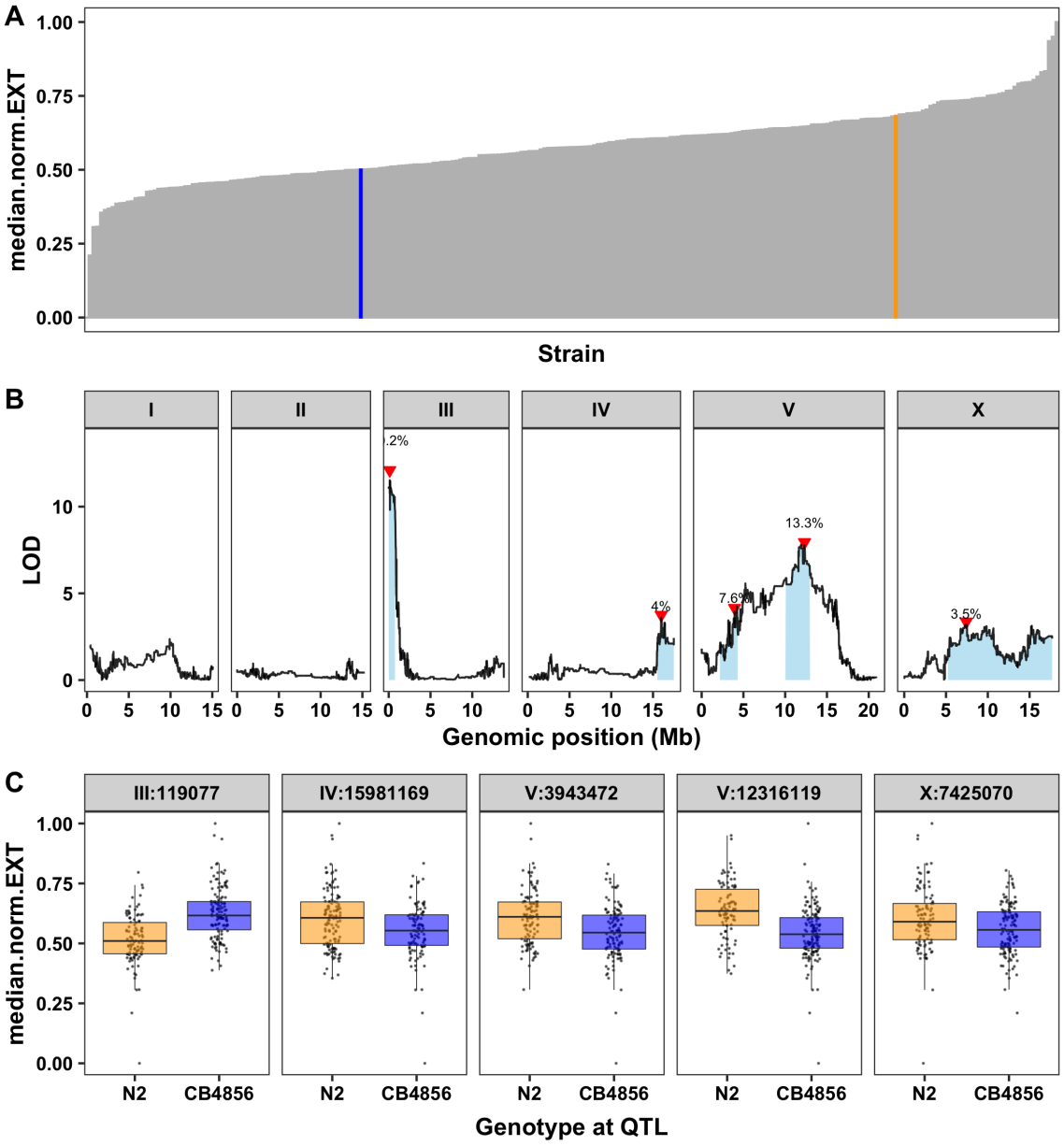
[1]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208
[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois 60208
[3]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Evanston, Illinois 60208

This manuscript was published on bioRxiv in July 2020 [123]

4.10 SUPPLEMENTAL FIGURES




A
median.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
LOD
15
10
5
0
.9%
5.3%
10.2%
4.1%
0 5 10 15
0 5 10 15
0 5 10
0 5 10 15
0 5 10 15 20
0 5 10 15
Genomic position (Mb)
C
III:119077
IV:15981169
V:12298716
X:15189807
median.EXT
1.00
0.75
0.50
0.25
0.00
N2
CB4856
Genotype at QTL


A
median.norm.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
LOD
10
5
0
.2%
4%
7.6%
13.3%
3.5%
0 5 10 15
0 5 10 15
0 5 10
0 5 10 15
0 5 10 15 20
0 5 10 15
Genomic position (Mb)
C
III:119077
IV:15981169
V:3943472
V:12316119
X:7425070
median.norm.EXT
1.00
0.75
0.50
0.25
0.00
N2
CB4856
Genotype at QTL

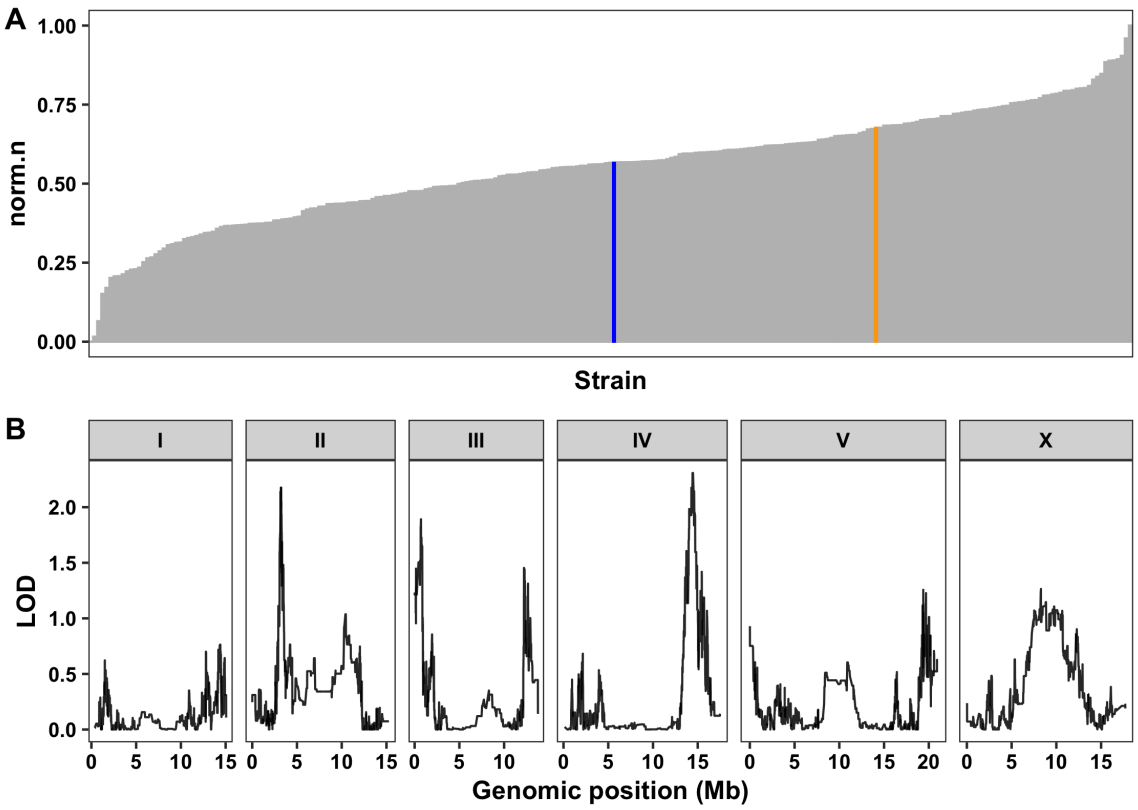
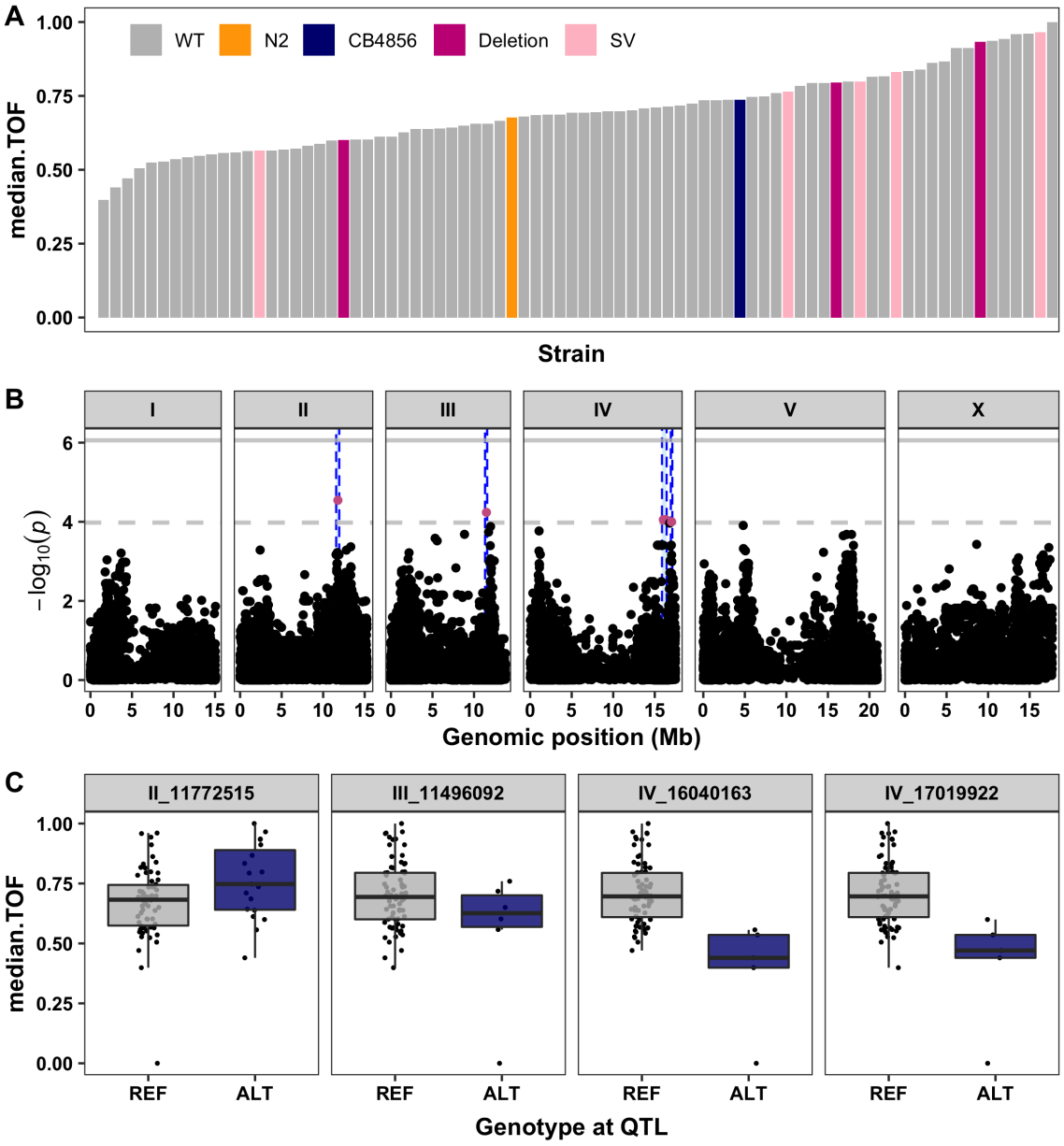
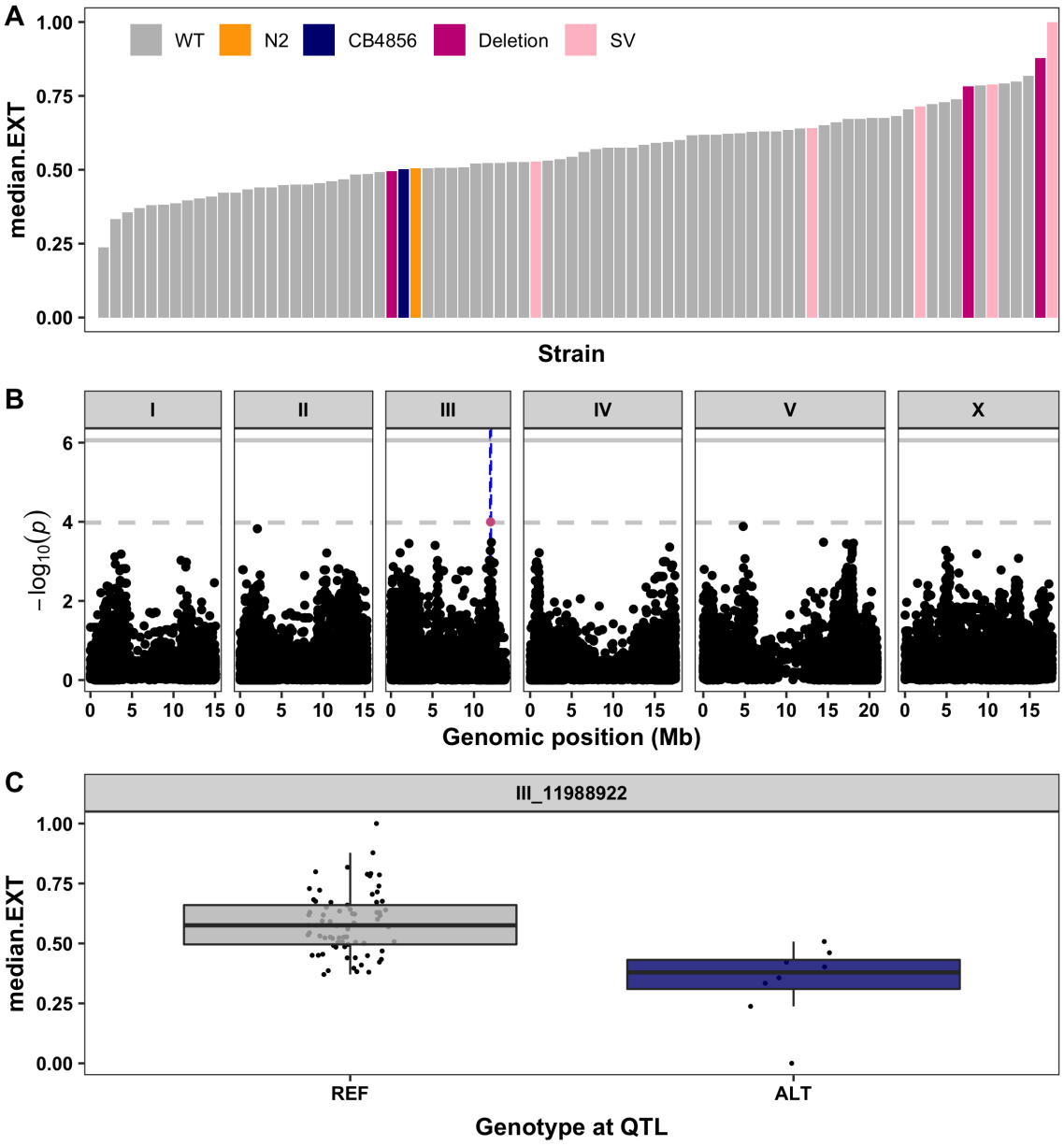
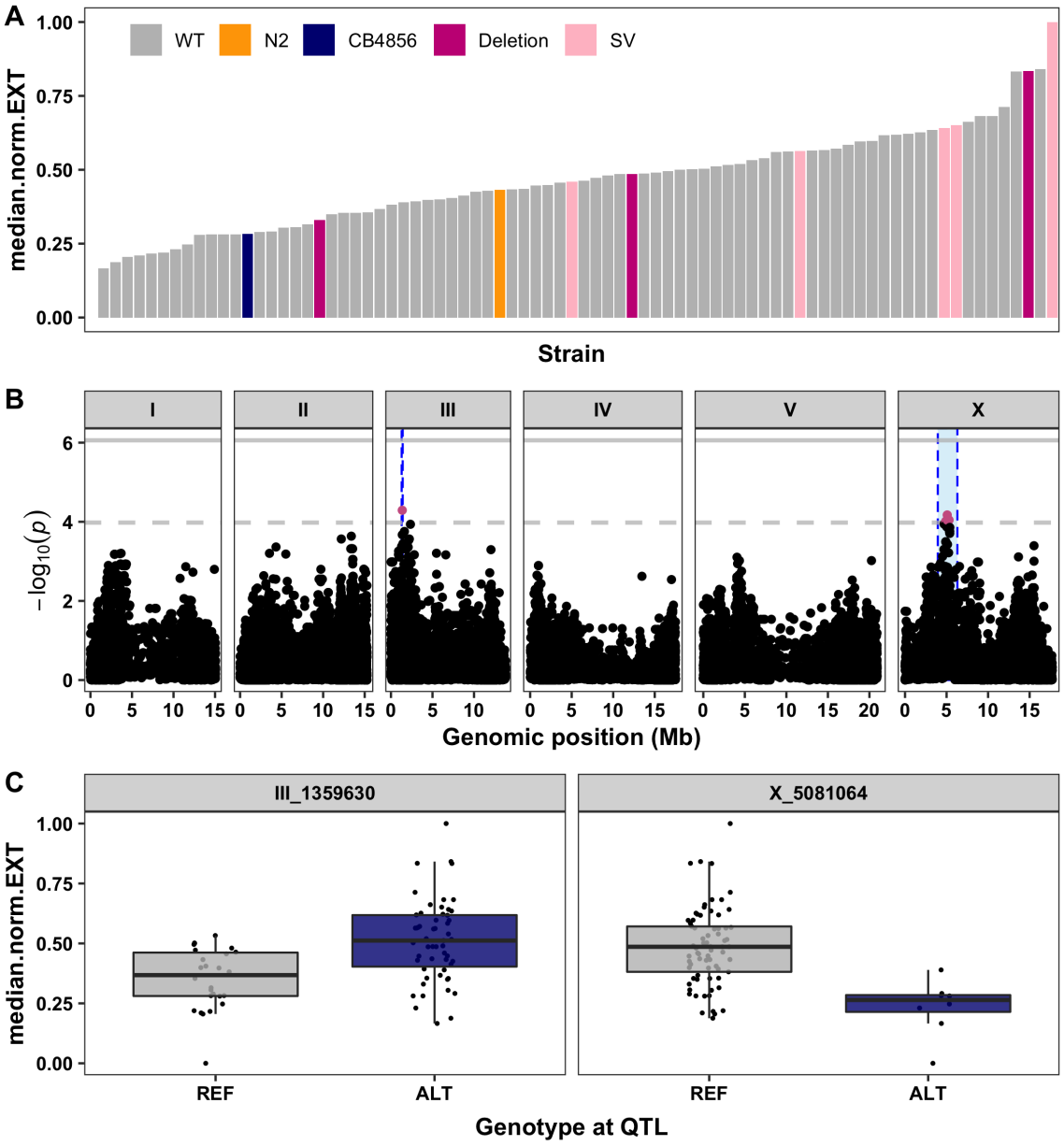


Figure S4-1: Linkage mapping identifies 12 QTL across three traits in response to high zinc. (A) Relative phenotype (y-axis) of 253 RIAILs (x-axis) in response to zinc supplementation. The parental strains are colored: N2, orange; CB4856, blue. **(B)** Linkage mapping results are shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(C)** For each QTL, the normalized residual phenotype (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange and strains with the CB4856 allele are colored blue.


A
WT
N2
CB4856
Deletion
SV
median.TOF
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
$-\log_{10}(p)$
6
4
2
0
0 5 10 15
0 5 10 15
0 5 10
0 5 10 15
0 5 10 15 20
0 5 10 15
Genomic position (Mb)
C
II_11772515
III_11496092
IV_16040163
IV_17019922
median.TOF
1.00
0.75
0.50
0.25
0.00
REF
ALT
Genotype at QTL


A
WT
N2
CB4856
Deletion
SV
median.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
$-\log_{10}(p)$
6
4
2
0
0 5 10 15
0 5 10 15
0 5 10
0 5 10 15
0 5 10 15 20
0 5 10 15
Genomic position (Mb)
C
III_11988922
median.EXT
1.00
0.75
0.50
0.25
0.00
REF
ALT
Genotype at QTL


A
WT
N2
CB4856
Deletion
SV
median.norm.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
$-\log_{10}(p)$
Genomic position (Mb)
C
III_1359630
X_5081064
median.norm.EXT
REF
ALT
Genotype at QTL

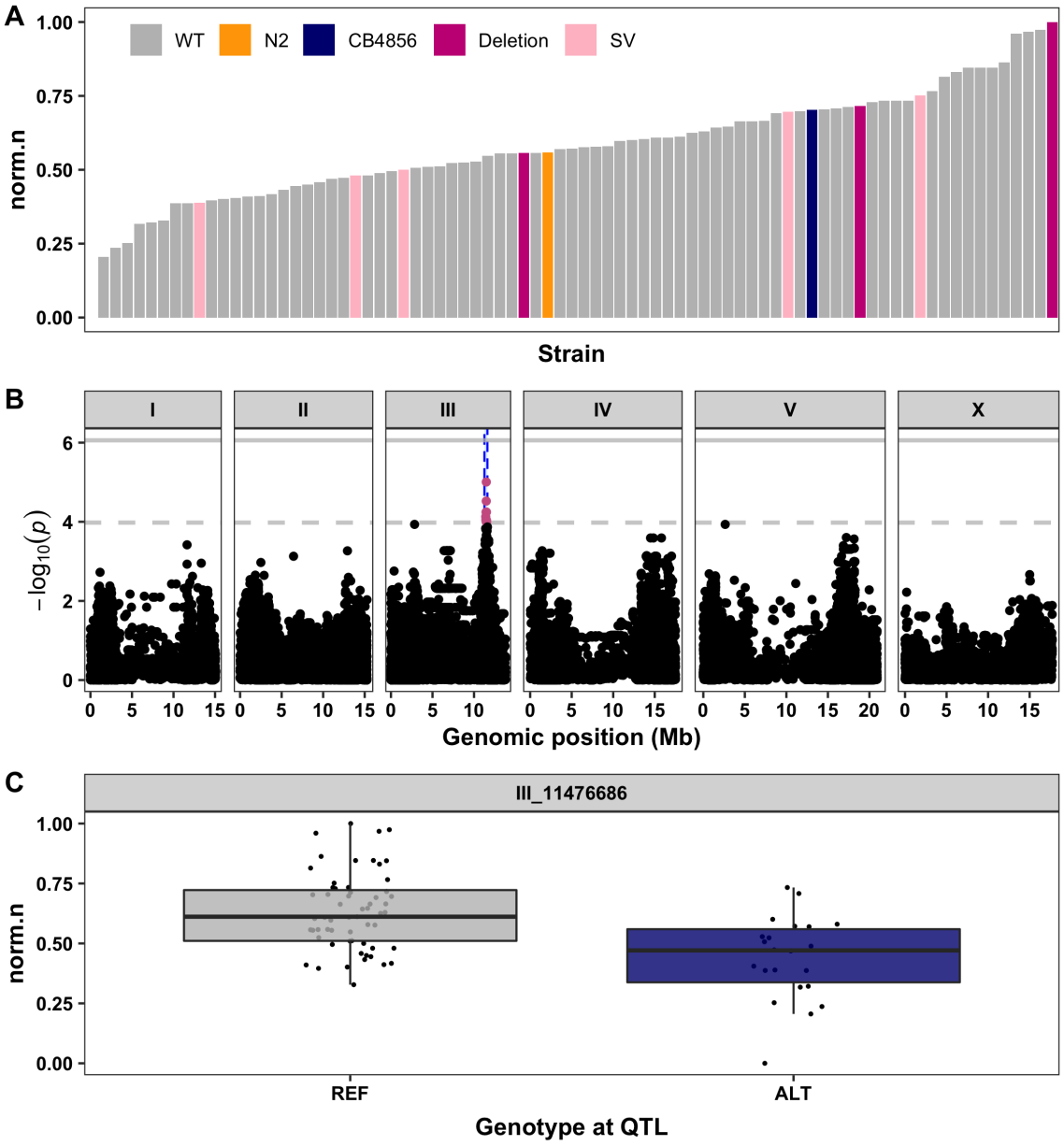


Figure S4-2: GWA mapping identifies QTL in response to high zinc. (A) Normalized residual phenotype (y-axis) of 81 wild isolates (x-axis) in response to zinc supplementation. Strains are colored by the parental strains N2 (orange) and CB4856 (blue) or by the *sqst-5* variation (Deletion: magenta, other variation: light pink). **(B)** GWA results are shown. Genomic position (x-axis) is plotted against the -log10(*p*) value (y-axis) for each SNV. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance threshold designated by the dotted grey line. The genomic regions of interest that pass the significance threshold are highlighted by blue rectangles. **(C)** For each QTL, the normalized residual phenotype (y-axis) of strains split by genotype at the peak marker (x-axis) are plotted as Tukey box plots. Each point corresponds to a wild isolate strain. Strains with the N2 reference allele are colored grey, and strains with an alternative allele are colored navy.

4.11 SUPPLEMENTAL TABLES

Table S4-1: Geographic locations of 57 wild isolates with putative structural variation in *sqst-5*

Strain	Latitude	Longitude	Deletion[5]
CB4856	21.33	-157.86	TRUE
CX11276	34.20111	-118.21198	TRUE
CX11307	34.12946	-118.10987	FALSE
ECA36	-36.893333	174.745529	TRUE
QX1794	20.70559	-156.35678	TRUE
ECA189	19.423889	-155.2225	FALSE
ECA191	19.424167	-155.22194	TRUE
ECA363	20.723611	-156.30444	TRUE
ECA396	19.449831	-155.23549	TRUE
XZ1514	22.149	-159.668	TRUE
XZ1516	22.149	-159.668	TRUE
ECA705	20.038831	-155.43892	TRUE
ECA722	19.423975	-155.22263	TRUE
ECA741	20.6762	-156.33832	TRUE
ECA742	20.80569	-156.2781	TRUE
ECA743	20.741871	-156.32317	TRUE
ECA745	19.700464	-155.95584	TRUE
EG4725	41.6288	-8.3476	FALSE
GXW1	30.542889	114.419828	FALSE
JU1088	34.7613	138.0149	FALSE
JU1400	37.3845	-5.988	TRUE
JU2316	38.7175	-9.1486	TRUE
JU2464	-13.155	-72.525	TRUE
JU2526	38.7175	-9.1486	TRUE
JU2841	-40	176	FALSE
JU2878	19.29732	-99.098769	TRUE
JU3134	48.7015	2.1725	FALSE
JU3135	48.7015	2.1725	FALSE
JU3144	6.18	10.52	FALSE
JU3166	0.28	6.59837	FALSE
JU3167	0.289	6.612	FALSE
JU3224	-41.296186	174.784857	TRUE
JU3225	-41.296186	174.784857	FALSE
JU3226	-41.29085	174.76872	TRUE
JU3228	-39.02256	175.71458	TRUE
JU3280	50.07182	14.42333	FALSE

[5]Does this strain have the same 111 bp deletion as CB4856?

JU3282	50.0711	14.42028	FALSE
MY1	52.54	7.31	TRUE
MY10	51.96	7.53	TRUE
MY2453	51.950777	7.536628	FALSE
MY2585	54.346355	10.117737	FALSE
MY2693	54.3491	10.11505	FALSE
MY2713	54.3491	10.11505	FALSE
MY2741	54.3491	10.11505	FALSE
NIC1	43.279	5.3543	TRUE
NIC251	38.66621	-28.151364	TRUE
NIC260	38.71829	-9.14875	TRUE
NIC265	38.71829	-9.14875	TRUE
NIC275	39.769459	-8.756356	FALSE
PB303	None	None	FALSE
QG2841	-28.2446	153.2063	TRUE
QG2875	-27.3672	152.1834	FALSE
QG536	37.7679	-122.4415	FALSE
QW947	-33.4213	-70.6106	TRUE
QX1211	37.7502	-122.4331	TRUE
RC301	47.99	7.84	FALSE
WN2002	51.975285	5.694834	FALSE

5 Gene-by-environment interactions influence docetaxel responses in *Caenorhabditis elegans*

5.1 PREFACE

When I joined the lab in the summer of 2016, we had a large collection of QTL from which I could choose one (or more) to follow up on in more detail. Throughout that summer and fall, I began testing a variety of these toxins to see if there was any "low-hanging fruit." Out of the 14 drugs I tested, I decided to write my qualification exam on docetaxel and zinc. I chose docetaxel because there was a large and reproducible parental split between N2 and CB4856 and both NILs validated the QTL effect on chromosome IV. I spent two years chasing down this QTL and getting pre-emptively excited about potential candidate genes (like *ttll-11*). Unfortunately, in June 2018 the worms stopped growing in our assays and no one could explain why. I spent my summer troubleshooting the issue and decided that the lysate we had been using since 2013 must have expired and is no longer supporting nematode growth. Lucky for us, our collaborator, Dr. Charlie Baer, had the potential to save us: 300 grams of the precious HB101 lysate from 2013, preserved in his freezer in Florida. In the fall of 2018 I tested my strains with his lysate and was ecstatic to find that the worms grew as I had seen before! However, they grew so well that the docetaxel effect I had been chasing disappeared. I developed what I called the "lysate hypothesis", that the quality of our precious lysate had been deteriorating over the years and had reached a point of no return. Although crushing, these results were transformative for the lab – we began to search for a new, reproducible food source for all future assays. More importantly, I believe that the end of this project marked the beginning of my scientific independence. It taught me to trust in myself and to always investigate inconsistencies, because they are likely telling us about biology!

5.2 ABSTRACT

Docetaxel is a commonly used chemotherapeutic drug in many different types of cancer and acts by inhibiting essential microtubule dynamics in actively growing cells that eventually leads to cell death. Although effective against cancer, docetaxel treatment can be associated with severe, patient-specific

side-effects that are difficult to predict. This variation in patient response to docetaxel can be attributed to both environmental factors, such as diet and exercise, and genetic factors, such as variation in proteins involved in drug transport or metabolism. Here, we use the roundworm nematode *Caenorhabditis elegans* to identify genetic variation on the center of chromosome IV that is associated with the nematode's variation in response to docetaxel. We used near-isogenic lines to validate a 1.8 Mb region on chromosome IV containing the microtubule-regulating gene *rmd-4* among other candidate genes. Finally, we show that environmental factors, such as food quality, strongly contribute to the docetaxel response in *C. elegans* and we further identified several important gene-by-environment interactions. This study highlights the importance of a constant environment when studying the genetic architecture of complex traits and provides evidence for the existence of strong gene-by-environment interactions in nematode drug responses.

5.3 INTRODUCTION

Cancer is among the leading causes of death worldwide and is commonly treated by surgery, radiation, and chemotherapy [279]. Many common chemotherapeutics induce apoptosis by targeting and disrupting essential cellular functions, often resulting in severe side-effects. Many drugs that are currently available are treated as "one size fits all", yet patients exhibit varying degrees of success that are difficult, or even impossible, to predict [280]. In most cases, both environmental and genetic factors contribute to this inter-individual variability in drug response. These environmental factors, such as diet and exercise, are often difficult to identify and control among patients. Alternatively, we can use several powerful genetic tools to correlate genetic variation to differences in drug responses across populations. The growing field of pharmacogenomics research attempts to identify genetic variation that underlies differences in drug import and export, metabolism, interaction with targets, excretion, and any other mechanism that can result in variable physiological responses [280]. Correlating this genetic variation to a specific response is the first step in the development of safe and effective treatments tailored to a person's genetic makeup.

Docetaxel is a chemotherapeutic drug with effective antitumor activity against many cancers including breast, head and neck, stomach, prostate, and non-small cell lung cancer [281]. Among

other functions, docetaxel promotes microtubule assembly and stabilization which inhibits microtubule dynamics essential for cell division and results in cellular apoptosis [282]. Despite its widespread use, variation in patient response to docetaxel is still a major limiting factor with dose-limiting side effects that include neutropenia, anemia, asthenia, and skin toxicity [283]. These side effects have been loosely linked to genetic variation in various drug transporters [283, 284, 285] and enzymes involved in the metabolism of docetaxel [283, 286, 287, 288, 289], however the correlation is not strong enough to merit FDA-approval of genetic biomarkers for docetaxel response.

The nematode roundworm *Caenorhabditis elegans* has numerous advantages that contribute to its power as a model organism to study genetic variation. *C. elegans* is easily maintained in the laboratory, has a quick generation time, and produces large amounts of genetically identical offspring [87]. *C. elegans* also has a small and highly annotated genome that contains homologs for 60-80% of human genes [55]. Importantly, many prominent pathways that are misregulated in cancer or targeted by therapeutics including Ras/MAPK, Wnt, Notch, TGF-β, and insulin signaling pathways are conserved in these nematodes [56]. In fact, many of these fundamental pathways were first identified and characterized in *C. elegans*, prompting future studies in mammalian systems [57]. Furthermore, access to 328 genetically divergent wild isolates provide high statistical power to connect quantitative genetic variation to phenotypic traits [120, 68, 134, 70, 71, 76].

Two strains of *C. elegans*, the Bristol laboratory-adapted strain, N2, and a wild isolate from Hawaii, CB4856, are highly genetically divergent, containing more than 300,000 single nucleotide variants (SNVs) and over 80,000 insertion or deletion events between the strains [60, 62]. A large panel of recombinant inbred advanced intercross lines (RIAILs) have been generated from a series of crosses between these two strains [68, 78], providing us with a powerful genetic tool to identify genomic regions that are correlated to a phenotype of interest. This tremendous genetic diversity has been found to affect many different traits including, but not limited to, aggregation behavior, growth rate, and response to various drugs [130, 147, 133, 117, 70, 72, 71, 73, 100, 108, 35, 149, 125, 128, 115, 74, 127, 132, 112, 119, 81, 103, 102, 80, 161, 86, 85].

Here, we show that phenotypic variation between N2 and CB4856 in response to the chemotherapeutic docetaxel correlates with genetic variation on the center of chromosome IV. We use

near-isogenic lines (NILs) to validate and narrow this QTL to a 1.8 Mb region containing over one hundred candidate genes including the putative microtubule-regulating gene *rmd-4*. However, throughout the course of the NIL validation experiments, the quality of the food source, otherwise held constant, was shown to deteriorate. We identified a significant interaction between genotype (N2 and CB4856) and food quality that severely impacted our ability to identify the quantitative trait gene underlying the chromosome IV locus. This study highlights the importance of a constant and reproducible environment, especially when studying traits with low effect sizes.

5.4 METHODS

5.4.1 Strains

Animals were grown at 20℃ on modified nematode growth media (NGMA) containing 1% agar and 0.7% agarose to prevent burrowing and fed OP50 [100]. The two parental strains, the canonical laboratory strain, N2, and the wild isolate from Hawaii, CB4856, were used to generate 247 recombinant inbred advanced intercross lines (RIAILs) [68]. Near-isogenic lines (NILs) were generated by backcrossing a selected RIAIL or NIL for several generations to the parent strain (N2 or CB4856) [72] using PCR amplicons for insertion-deletion (indels) variants to track the introgressed region. NILs were whole-genome sequenced to verify introgressions were only in the targeted genomic intervals. All strains are available upon request or from the *C. elegans* Natural Diversity Resource [63].

5.4.2 High-throughput phenotyping assay

For all phenotyping assays, we used a high-throughput fitness assay described previously [68]. In summary, populations of each strain were passaged and amplified on NGMA plates for four generations. In the fifth generation, gravid adults were bleach-synchronized and 25-50 embryos from each strain were aliquoted into 96-well microtiter plates at a final volume of 50 μL K medium [173]. The following day, arrested L1s were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA; [174]) at a final concentration of 5 mg/mL in K medium and were grown to the L4 larval stage for 48 hours at 20℃ with constant shaking. Three L4

larvae were sorted into new 96-well microtiter plates containing 10 mg/mL HB101 bacterial lysate, 50 μM kanamycin, and either 1% DMSO or docetaxel dissolved in 1% DMSO using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica; Holliston, MA). Sorted animals were grown for 96 hours at 20℃ with constant shaking. The next generation of animals and the parents were treated with sodium azide (50 mM in 1X M9) to straighten their bodies for more accurate length measurements. Animal length (median.TOF), optical density integrated over animal length (median.EXT), and brood size (norm.n) were quantified for each well using the COPAS BIOSORT. Nematodes get longer (animal length) and become thicker and more complex (optical density) over developmental time. Because animal length and optical density are highly correlated, we calculated a fourth trait (median.norm.EXT) that normalizes optical density by animal length (median.EXT / median.TOF). Phenotypic measurements collected by the BIOSORT were processed and analyzed using the R package *easysorter* [77] as described previously [72]. Differences among strains within the control conditions were controlled by subtracting the mean control-condition value from each drug-condition replicate for each strain using a linear model $drug\_phenotype \sim mean\_control\_phenotype$. In this way, we are addressing only the differences among strains that were caused by the drug condition and the variance in the control condition does not affect the variance in the drug condition. For NIL assays, complete pairwise strain comparisons were performed on drug residual phenotypes using a *TukeyHSD* function [175] on an ANOVA model with the formula $phenotype \sim strain$.

5.4.3 Dose response assay

Four genetically divergent strains (N2, CB4856, JU258, and DL238) were treated with increasing concentrations (0 μM, 2.5 μM, 5 μM, and 10 μM) of docetaxel (Fluka; #01885-5MG-F) in 1% DMSO using the high-throughput phenotyping assay described above. A dose of 5 μM provided a reproducible drug-specific effect that maximized between-strain variation while minimizing within-strain variation across the four traits and was selected for the linkage mapping experiments. Broad-sense heritability was calculated from the dose response phenotypes using the lmer function in the *lme4* R package [212] with the formula $phenotype \sim 1 + (1|strain)$ for each dose.

5.4.4 Linkage mapping

247 RIAILs were phenotyped docetaxel and DMSO using the high-throughput phenotyping assay described above. Linkage mapping was performed using the R package *linkagemapping* (https://github.com/AndersenLab/linkagemapping) as described previously [72]. The cross object derived from the whole-genome sequencing of the RIAILs containing 13,003 SNPs was loaded using the function *load_cross_obj("N2xCB4856cross_full")*. The RIAIL phenotypes were merged into the cross object using the merge_pheno function with the argument *set = 2*. A forward search (*fsearch* function) adapted from the *R/qtl* package [176] was used to calculate the logarithm of the odds (LOD) scores for each genetic marker and each trait as $-n(ln(1 - R2)/2ln(10))$ where R is the Pearson correlation coefficient between the RIAIL genotypes at the marker and trait phenotypes [27]. A 5% genome-wide error rate was calculated by permuting the RIAIL phenotypes 1000 times. The marker with the highest LOD score above the significance threshold was selected as the QTL then integrated into the model as a cofactor and mapping was repeated iteratively until no further significant QTL were identified. Finally, the *annotate_lods* function was used to calculate the effect size of each QTL and determine 95% confidence intervals defined by a 1.5 LOD drop from the peak marker using the argument *cutoff = "chromosomal"*.

5.4.5 Genotype-by-environment interactions

To identify significant components of the phenotypic variance, including GxE interactions, we performed a one-way analysis of variance (ANOVA) test using the *anova* function in the stats package in R. The results from each ANOVA analysis are listed in the Supplemental Tables (Tables S1-S3). We calculated the percent of the total phenotypic variation that could be explained by a single factor or interaction by dividing the sum squares of the factor by the total variance defined as the sum of the sum squares of all factors. Terms were considered significant with a p-value < 0.05, and all statistics were reported.

5.4.6 Modified high-throughput fitness assay

A modified version of the standard high-throughput assay detailed above was used to test docetaxel response as previously described [72, 69]. Briefly, strains were propagated for two generations, bleach-synchronized, and titered at a concentration of 25-50 embryos per well of a 96-well microtiter plate. Arrested L1s were fed HB101 bacterial lysate the following day at a final concentration of 5 mg/mL with either DMSO or docetaxel in DMSO. After 48 hours of growth at 20℃ with constant shaking, nematodes were treated with sodium azide (5 mM in water) prior to analysis of animal length and optical density using the COPAS BIOSORT. Because only one generation of growth was observed, brood size was not calculated.

5.5 RESULTS

5.5.1 Genetic variation on chromosome IV influences animal length in response to docetaxel

We measured *C. elegans* development and docetaxel sensitivity as a function of animal length (median.TOF), optical density (median.EXT), and brood size (norm.n) with a high-throughput assay developed with the COPAS BIOSORT (See Methods) [70, 72, 73, 71, 68]. Because optical density is calculated as a function of length, these traits are highly correlated. We calculated a fourth trait (normalized optical density; median.norm.EXT) that normalizes optical density by length, providing a value for the animal density that is independent of length. We exposed four genetically divergent strains (N2, CB4856, JU258, and DL238) to increasing doses of docetaxel. In the presence of the drug, nematodes were generally smaller, less optically dense, and produced smaller broods compared to non-treated nematodes (**Figure 5-1**). We also observed significant phenotypic variation among strains and identified a substantial heritable genetic component for most traits (average $H^2 = 0.55$).

We exposed a panel of 247 RIAILs generated from the N2 and CB4856 parental strains [68] to docetaxel at a concentration of 5 μM that both maximizes among-strain and minimizes within-strain phenotypic variation. Linkage mapping for all four traits identified a total of 13 QTL on chromosomes I, IV, V, and X. Brood size in response to docetaxel had a significant QTL on the center of chromosome I (**Figure S5-1**). The remaining three developmental traits shared a large-effect QTL on the center

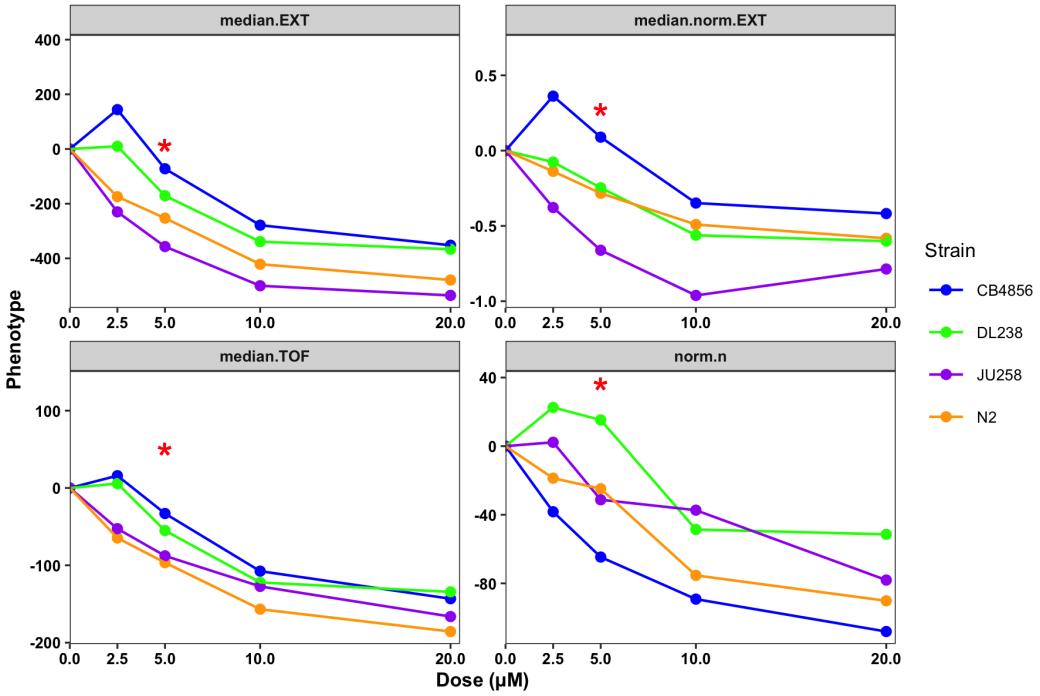


Figure 5-1: Docetaxel dose response. For each trait, drug concentration (μM) (x-axis) is plotted against drug phenotype subtracted from control phenotype (y-axis), colored by strain (CB4856: blue, DL238: green, JU258: purple, N2: orange). A red asterisk indicates the dose selected for linkage mapping analysis.

of chromosome IV and a low-effect QTL on the right of chromosome X (**Figure 5-2**, **Figure S5-1**). Additionally, normalized animal optical density had a large-effect QTL that mapped to the right arm of chromosome I, not overlapping with the QTL observed for brood size (**Figure S5-1**).

To isolate the large-effect QTL on chromosome IV, we generated near-isogenic lines (NILs) by introgressing the CB4856 genome on the center of chromosome IV into the N2 genetic background or vice versa. Although the N2 parent is shorter in docetaxel than the CB4856 parent (**Figure 5-1**), the N2 allele on chromosome IV is associated with longer animal lengths in the RIAIL panel compared to the CB4856 allele at this position (**Figure 5-2**). For this reason, we would expect that the NIL with the CB4856 genetic background and N2 introgression on chromosome IV would be hyper-resistant compared to both parental strains. We exposed both parental strains and the reciprocal NILs to docetaxel and measured animal lengths. Surprisingly, we saw a significant and reproducible switch in the animal lengths of the parental strains (**Figure 5-3**). The N2 strain now had a longer length than the CB4856 strain in response to docetaxel. Although unexpected, these results were reproduced several

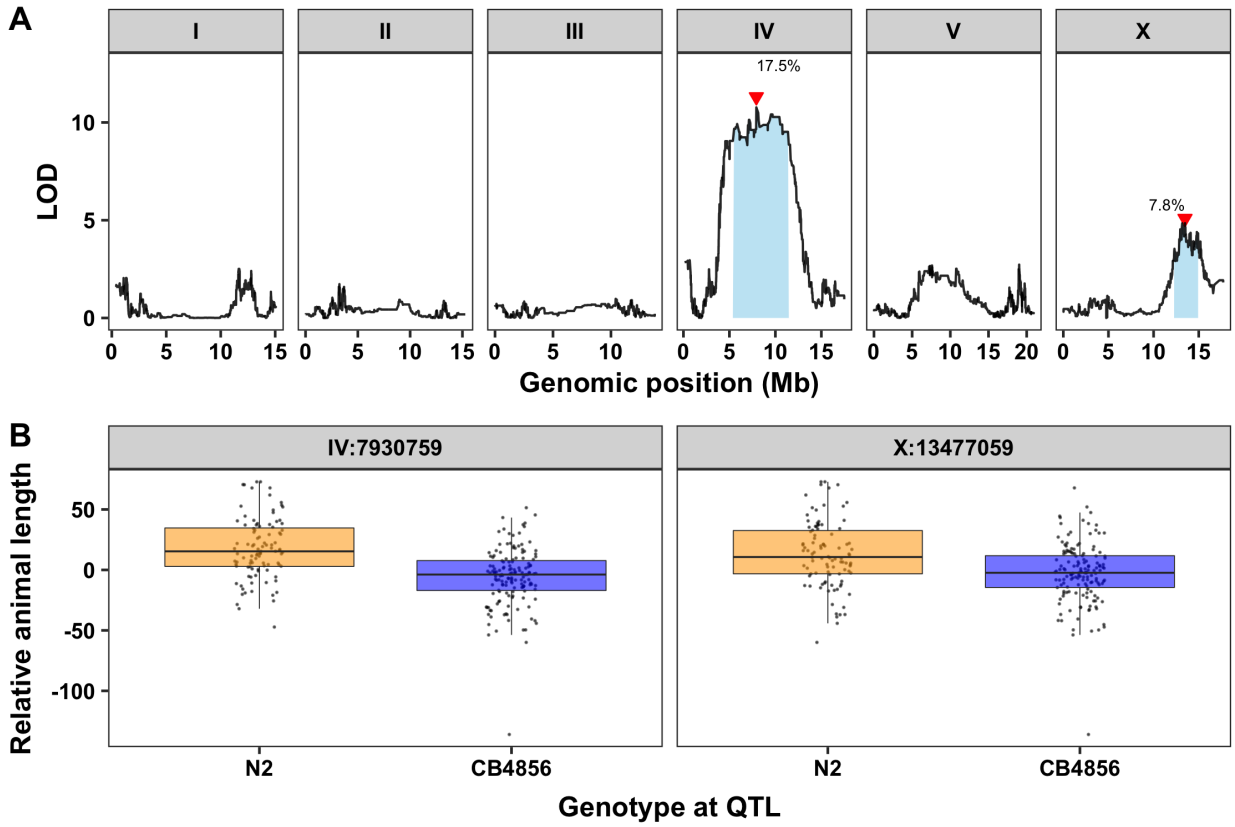


Figure 5-2: Linkage mapping results for animal length in docetaxel. (A) Linkage mapping results for animal length (median.TOF) is shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** For each QTL, the relative animal length (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange and strains with the CB4856 allele are colored blue.

times over different experimental conditions and with higher replication than the dose response assay. Regardless of the discrepancy between the parental effect, the NILs clearly showed evidence validating the QTL on chromosome IV. The strain with N2 introgressed into CB4856 (ECA599) is significantly longer than the CB4856 parent and the strain with CB4856 introgressed into N2 (ECA597) is significantly shorter than the N2 parent (**Figure 5-3**).

To narrow this QTL, we created a panel of NILs with smaller introgressions by backcrossing ECA599 to CB4856 and looking for recombination events. We selected three NILs that tile across the introgressed region in ECA599 and measured animal length of these strains in response to docetaxel (**Figure 5-4**). Strains with the N2 allele at the QTL should be longer than strains with the CB4856

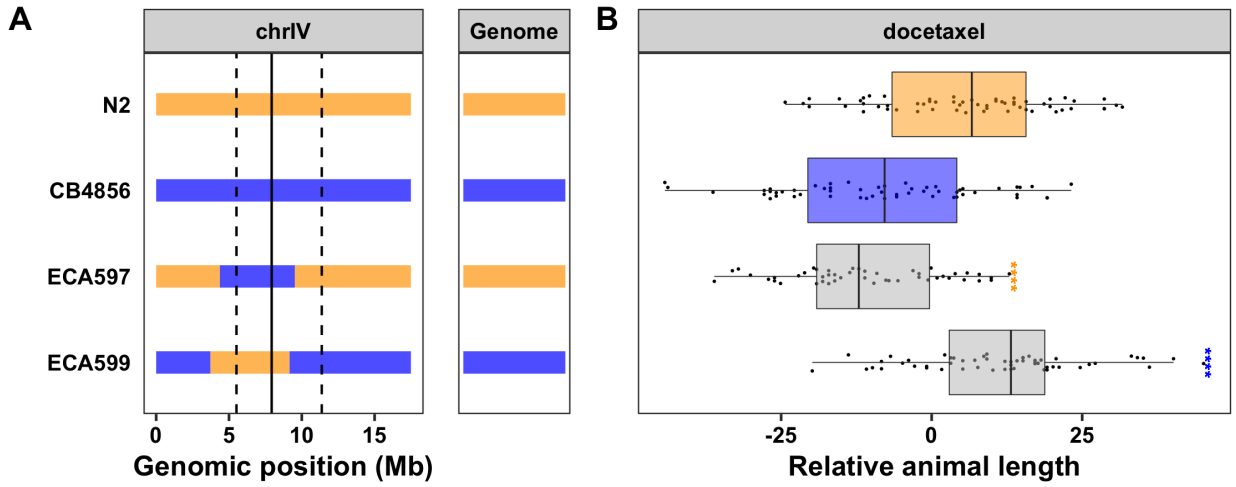


Figure 5-3: NILs validate that the N2 allele on chromosome IV confers resistance to docetaxel. **(A)** Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome IV (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the QTL and the dashed vertical lines represent the confidence interval. **(B)** Relative animal length in docetaxel (median.TOF, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to its parental strain (ECA597 to N2 and ECA599 to CB4856) is shown above each strain and colored by its parental strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

allele at this locus. We identified a 2 Mb region (IV:3,707,051-5,864,172) isolated within ECA675 that produced the same resistant phenotype as our original NIL, ECA599. Because the strain ECA673 has a sensitive phenotype like the CB4856 parental strain, we can narrow the QTL slightly to exclude the overlapping introgressed region between ECA675 and ECA673 resulting in a 1.8 Mb region (IV:3,707,051-5,531,970) containing 744 genes. Because we are assessing phenotypic variation caused by genetic variation, we can ignore 489 genes in this interval without genetic variation in CB4856. Of the 255 remaining genes, only 90 have variation that is predicted to cause a change in the amino acid sequence of the protein. In addition to protein-coding variation, phenotypic variation can also be caused by genetic variation leading to gene expression differences between strains. Using an expression QTL (eQTL) dataset previously generated from a separate set of N2xCB4856 RIAILs [78, 115, 184], we identified 115 genes with gene expression differences that mapped to genetic variation in this region. It is possible that one of these genes with protein-coding variation or an eQTL in this region is causing the differences observed in animal length in docetaxel between the N2 and CB4856 strains.

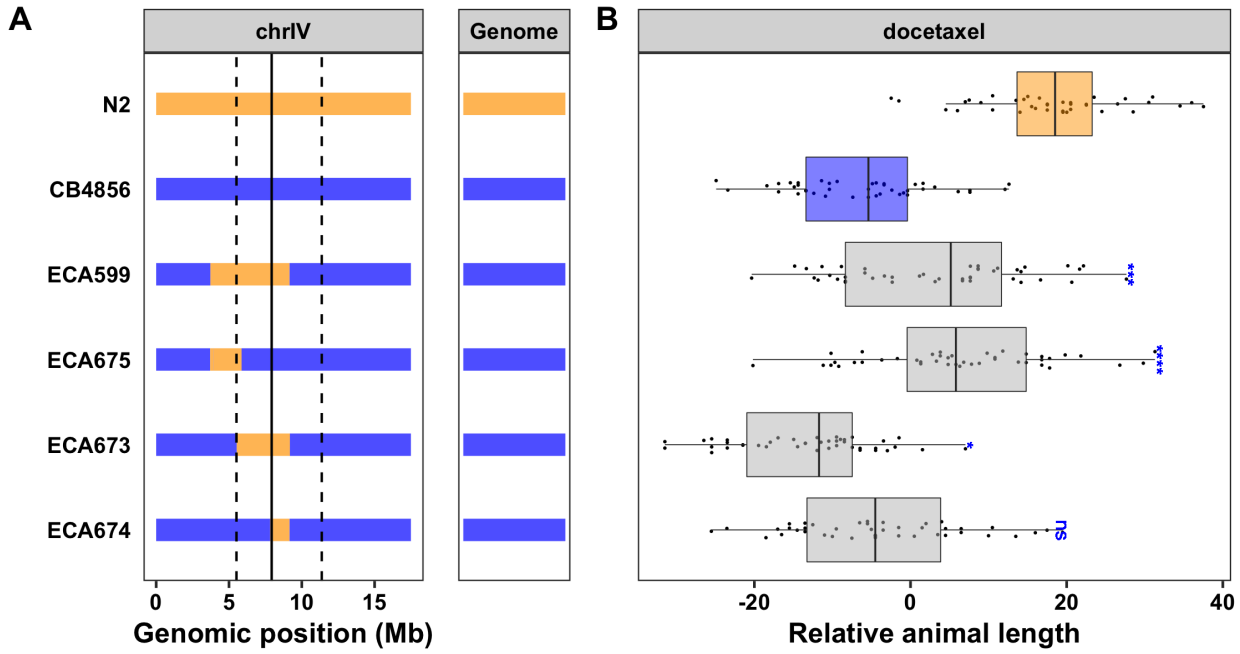


Figure 5-4: NILs narrow the QTL to a 2 Mb region. **(A)** Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome IV (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the QTL and the dashed vertical lines represent the confidence interval. **(B)** Relative animal length in docetaxel (median.TOF, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to CB4856 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

To narrow our list of candidate genes further, we next analyzed the functional descriptions and GO annotations for all genes in the region and all genes with an eQTL that maps to this region. Because docetaxel is known to inhibit microtubule dynamics, we first searched for genes related to microtubules. From this search, we found one gene in particular, *rmd-4*, that is an ortholog of the human genes RMDN2 and RMDN3 that function as regulators of microtubule dynamics (Wormbase.org). RIAILs with the N2 allele on chromosome IV have higher expression of *rmd-4* (**Figure 5-5**) and are more resistant to docetaxel (**Figure 5-2B**) than RIAILs with the CB4856 allele, suggesting that expression of *rmd-4* could be associated with increased resistance to docetaxel. Although the exact function of *rmd-4* is unknown, the homolog *rmd-1* has been shown to promote microtubule growth and plays a role in regulating proper chromosome segregation [290]. A review of the literature provides no evidence that expression of RMDN2 or RMDN3 has been previously associated with patient response to either docetaxel or the closely related drug, paclitaxel. Regardless, variation in expression of *rmd-4* might

influence the nematode's response to docetaxel.

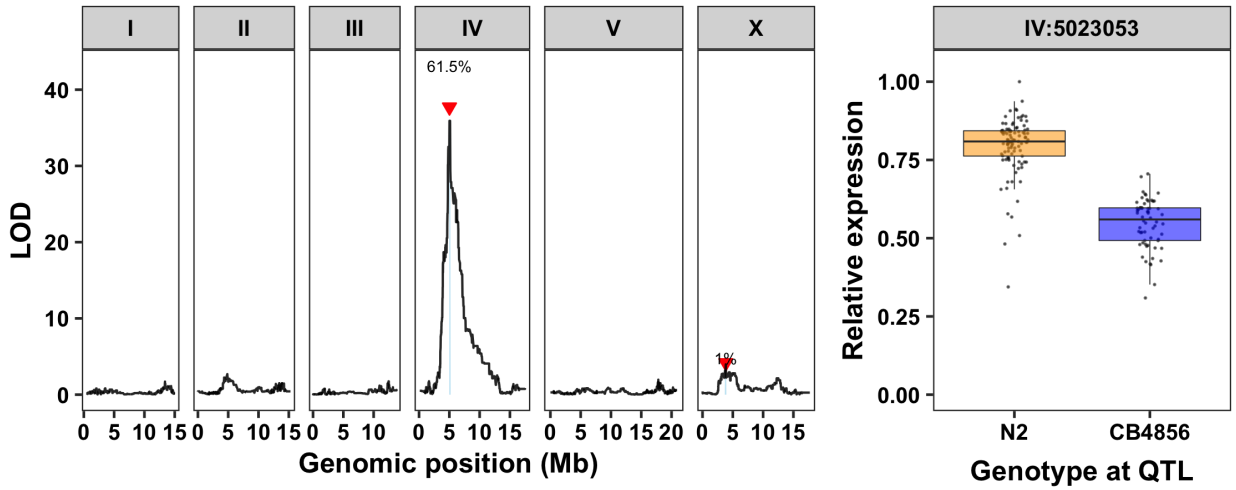


Figure 5-5: Expression QTL mapping for *rmd-4*. **(A)** Results from linkage mapping analysis using expression of *rmd-4* as a quantitative trait. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** For the chromosome IV QTL, the expression of *rmd-4* (y-axis) in RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to the relative expression of a unique recombinant strain. Strains with the N2 allele are colored orange, and strains with the CB4856 allele are colored blue.

5.5.2 Environmental variation of food quality dictates drug-response traits

For all experiments from dose responses to NIL validations, an HB101 lysate collected from a single prep at Pennsylvania State University's Shared Fermentation Facility (see Methods) was used in an attempt to eliminate environmental variation introduced by independent bacterial preps. However, in June 2018, this lysate prep appeared to expire and no longer produced healthy L4s within 48 hours after feeding a population of starved L1s (**Figure 5-6**). Because we know that different bacterial foods can drastically affect the *C. elegans* response to chemotherapeutics [174], we borrowed an aliquot of the same HB101 lysate from the Baer lab (University of Florida; Gainesville, FL) to continue performing experiments. We found that this lysate performed similarly to the lysate used in our previous experiments and produced healthy-sized L4 nematodes after 48 hours (**Figure 5-6**). We then used this lysate to measure animal lengths in docetaxel for N2, CB4856, and two of the NILs with a previously defined response (ECA599 and ECA672). Surprisingly, we discovered that the CB4856

strain was longer (more resistant) than the N2 strain in response to docetaxel (**Figure 5-7**). This parental effect matches that seen in the original dose response assay, but is opposite from what we previously observed with the NILs (**Figure 5-3**, **Figure 5-4**). More importantly, the NILs that used to confer resistance to docetaxel (**Figure 5-3**, **Figure 5-4**) were just as resistant as the CB4856 parent (**Figure 5-7**). If CB4856 is resistant and the N2 allele on chromosome IV is also resistant, we would expect the NILs with the N2 introgression to be hyper-resistant compared to CB4856. This suggests a lack of evidence for the validation of the chromosome IV QTL in response to docetaxel.

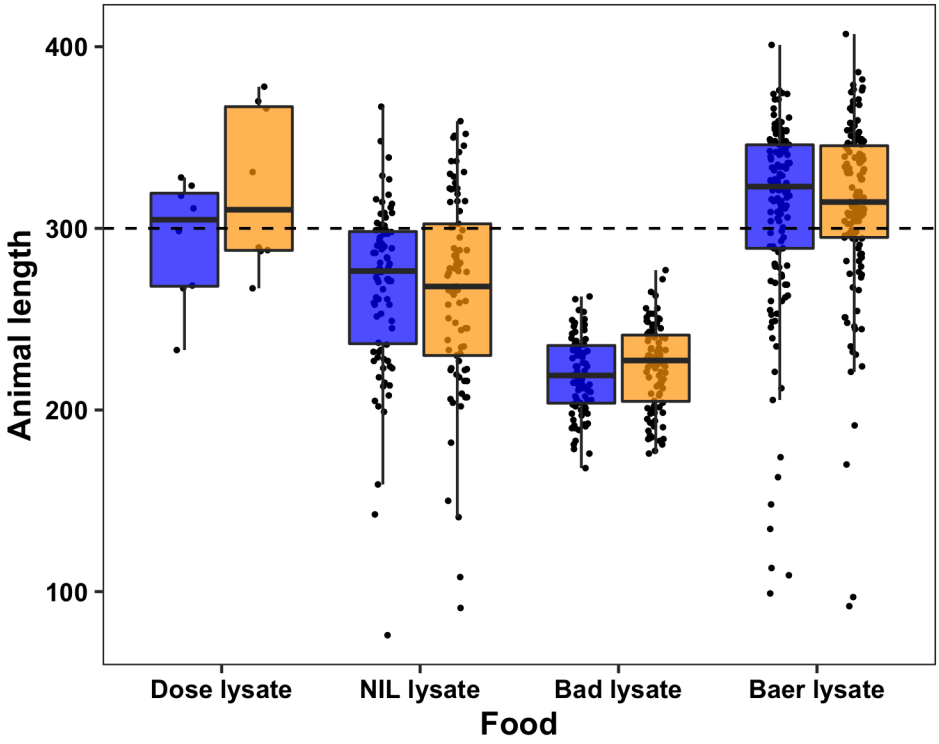


Figure 5-6: Nematodes grown in the bad lysate do not reach L4 stage within 48 hours after feeding. Animal length 48 hours post feed (y-axis) is plotted as Tukey box plots for N2 (orange) and CB4856 (blue) across three experimental conditions (x-axis). The horizontal dashed line represents the average length of an L4 nematode.

The inconsistency in the parental effect in response to docetaxel could be a direct result of a genotype-by-environment (GxE) interaction with the quality of the HB101 lysate. The parental effect in the “good quality” HB101 lysate used in the dose response and linkage mapping experiment matches the parental effect in the “good quality” HB101 lysate borrowed from the Baer lab. We hypothesized that sometime between the linkage mapping experiment (2014) and the NIL validation (2016), the

quality of our HB101 lysate began to deteriorate, resulting in “medium quality” lysate. With this “medium quality” lysate, we saw an opposite parental effect compared to the “good quality” lysate. In June 2018, this deterioration hit a new low at which point the nematodes were no longer able to grow, resulting in “bad quality” lysate. This hypothesis is summarized in **Figure 5-8**.

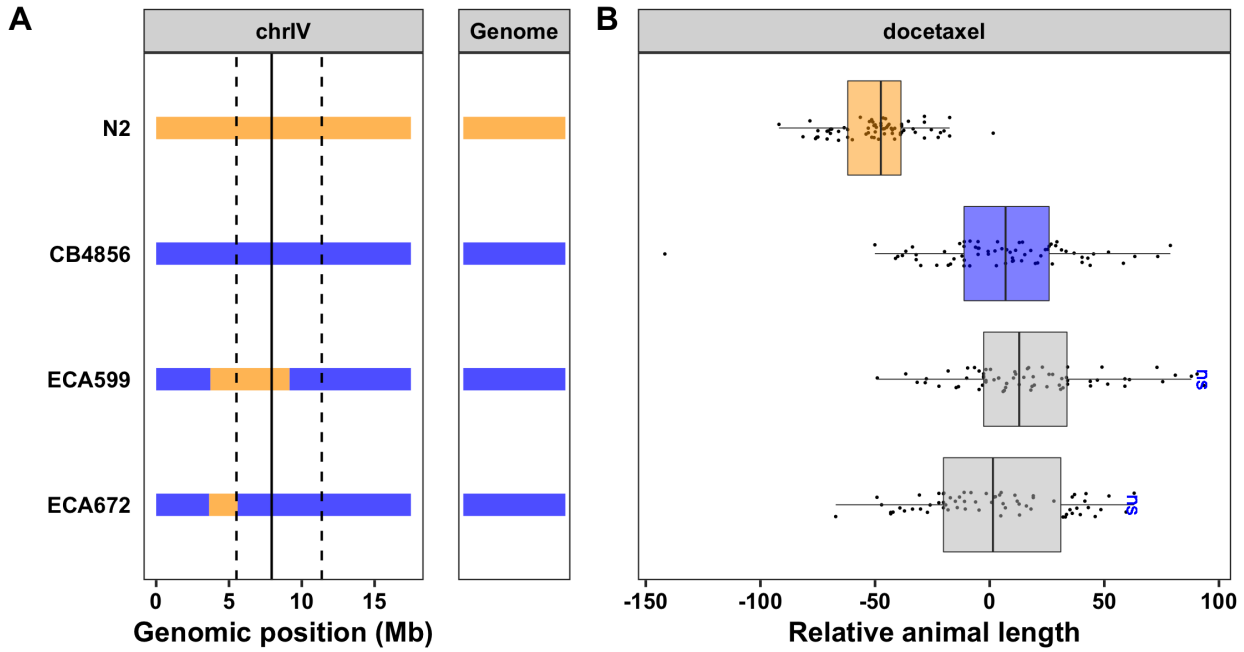


Figure 5-7: An older aliquot of the same HB101 lysate changes the parental response to docetaxel. **(A)** Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome IV (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the QTL and the dashed vertical lines represent the confidence interval. **(B)** Relative animal length in docetaxel (median.TOF, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to CB4856 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

5.5.3 The growth-response QTL on chromosome IV responds to environmental variation

Previous studies have reported a growth-response QTL between the N2 and CB4856 strains on the center of chromosome IV [68]. We hypothesized that this growth-response QTL might be driving the growth dynamics in the different food qualities. To test if the same growth-response QTL is present in the DMSO control for our assay, we performed linkage mapping analysis for the four primary growth-response traits (median.TOF, median.EXT, norm.n, and median.norm.EXT) as well as summary statistics of the population (q10, q25, median, mean, q75, and q90). Of these 20 traits, 11

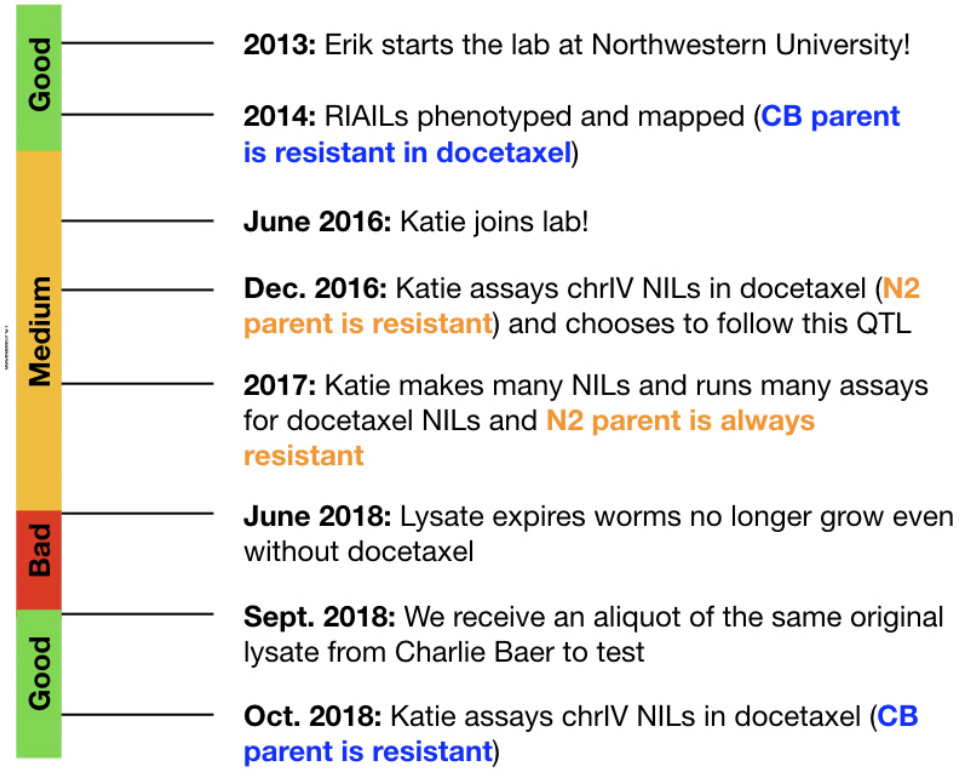


Figure 5-8: The lysate hypothesis (GxE interactions). Timeline of docetaxel experiments showing the parental effect and a proposed lysate quality (good, medium, and bad).

have a significant QTL in one of four distinct regions on chromosomes II, IV, V, and X (**Figure 5-9**). In particular, the QTL on chromosome IV overlaps with the docetaxel-response QTL. The growth-response QTL shows that RIAILs with the N2 allele have a higher brood, smaller length, and are less optically dense compared to RIAILs with the CB4856 allele (**Figure 5-9**). Although differences between strains in DMSO are accounted for in our experiments with a linear regression (see Methods), this overlap could still make it difficult to correctly interpret our results.

Because genetic variation on the center of chromosome IV controls both responses to docetaxel and the general growth response in control conditions, we hypothesized that the observed GxE interaction could exist in either control or drug conditions. We investigated the potential GxE interactions in both DMSO and docetaxel for three of our previous assays (dose response (2014), NIL validation (2017), and NIL validation using lysate aliquot from the Baer lab (2018)) using the different HB101 lysates as different environmental conditions. In the DMSO control, we see decreasing animal length for the

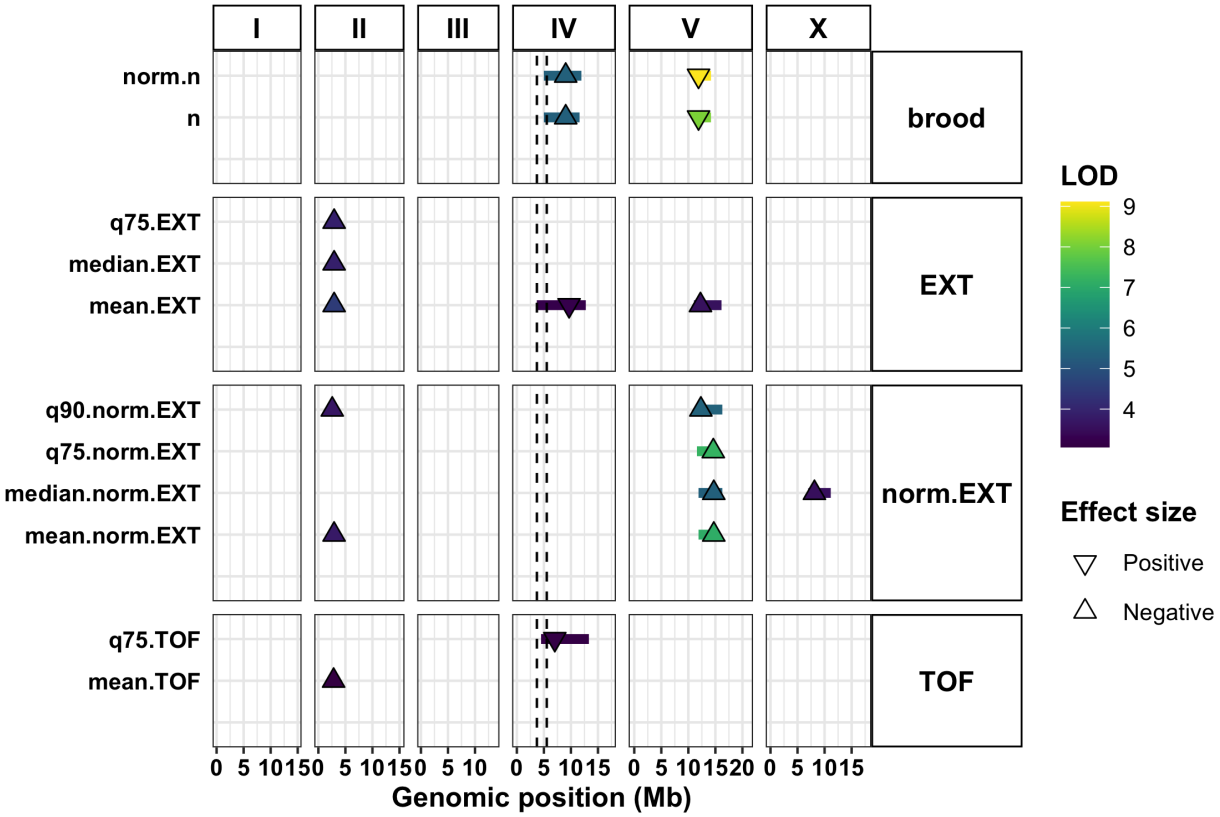


Figure 5-9: Growth-response QTL in DMSO control. Linkage mapping results for 20 growth-response traits identifying 18 QTL from 11 traits in response to DMSO are plotted. Genomic position (Mb) is shown along the x-axis, split by chromosome, and each of the traits with a significant QTL is plotted along the y-axis. Each QTL is plotted as a triangle at the location of the most significant genetic marker and a line indicating the 95% confidence interval. QTL with right side up triangles have a negative effect size (N2 allele is resistant), and QTL with upside down triangles have a positive effect size (CB4856 allele is resistant). QTL are colored by the logarithm of the odds (LOD) score, increasing in significance from purple to green to yellow. Dotted vertical lines represent the NIL-defined confidence interval of the docetaxel-response QTL.

CB4856 strani from the dose lysate to the NIL lysate, and again to the Baer lysate (**Figure 5-10A**). Interestingly, we see a different response from the N2 strain. Although N2 also decreases in length between the dose and NIL lysates, it does so at a greater slope than the CB4856 strain. The N2 strain was longer than the CB4856 strain in the dose lysate and significantly shorter than the CB4856 strain in the NIL lysate (**Figure 5-10A**). Furthermore, the N2 strain again performs better than the CB4856 strain in the Baer lysate, falling intermediate to its dose and NIL phenotypes (**Figure 5-10A**).

We next performed an analysis of variance (ANOVA) test to compare the effect of genotype, food, and the interaction between genotype and food on the nematode growth in DMSO conditions (**Table 5-1**, see Methods). In this model, the effect of genotype explained only 1% of the variation in phenotype

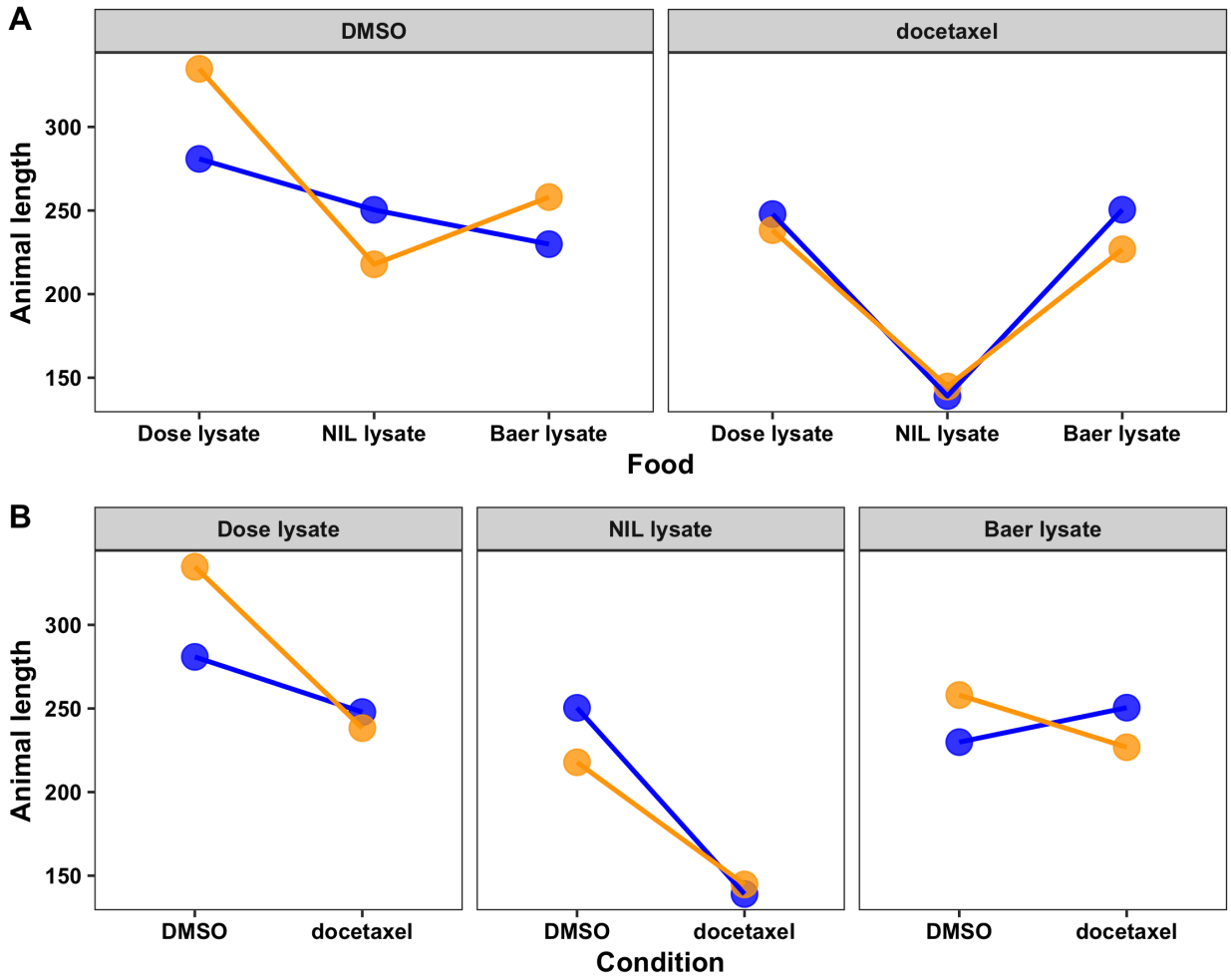


Figure 5-10: Reaction norm plots identify significant GxE interactions. **(A)** Average animal length (y-axis) in response to DMSO (left) and docetaxel (right) is plotted for N2 (orange) and CB4856 (blue) across multiple experimental conditions (x-axis). **(B)** Average animal length (y-axis) in response to DMSO and docetaxel (x-axis) is plotted for N2 (orange) and CB4856 (blue) across multiple experimental conditions. GxE interactions can be visualized as a difference in slopes between two conditions for N2 and CB4856.

(*p*-value = 0.065). Alternatively, 15% of the variation in phenotype can be explained by food quality (*p*-value = 6.5e-11) and 19% of the variation in phenotype can be explained by the GxE interaction between food quality and genotype (*p*-value = 2.7e-13). Together, these data indicate a statistically significant GxE interaction between the animal length of the N2 and CB4856 strains in DMSO and the quality of the lysate.

Table 5-1: Results of ANOVA test[a] of genotype, food quality, and genotype-by-food interaction on length in DMSO

	DF	Sum Sq	Mean Sq	F value	P value
Strain	1	2685	2685.2	3.4357	0.06514
Food	2	40863	20431.7	26.1420	6.517e-11
Strain:Food	2	51705	25852.5	33.0779	2.756e-13
Residuals	220	171944	781.6	*NA*	*NA*

[a] $anova(lm(phenotype \sim strain + food + strain * food, data = DMSO\_phenotypes))$

The same general shape is seen for N2 in response to docetaxel, with the NIL lysate producing the smallest animal lengths (**Figure 5-10A**). Surprisingly, this same shape is also observed in CB4856, differing from the strain's response to the DMSO control. In fact, the animal length of CB4856 in docetaxel is actually greater than the animal length of CB4856 in DMSO with the Baer lysate (**Figure 5-10A**). We again performed a one-way ANOVA to test the effect of genotype, food quality, and interaction between genotype and food on animal length in docetaxel (**Table 5-2**). We discovered that 81% of the variation in phenotype in docetaxel could be explained by the food quality (*p*-value = 2.2e-16). Unlike in DMSO conditions, the effect of the GxE interaction between strain and food quality was negligible (1.8%, *p*-value = 6.1e-6). The observed difference in responses between the DMSO and docetaxel conditions provide evidence that the response to docetaxel might still be separate from the response to DMSO.

Table 5-2: Results of ANOVA test[a] of genotype, food quality, and genotype-by-food interaction on length in docetaxel

	DF	Sum Sq	Mean Sq	F value	P value
Strain	1	7611	7611	17.904	3.428e-05
Food	2	484376	242188	569.705	<2.2e-16
Strain:Food	2	10797	5399	12.700	6.085e-06
Residuals	217	92249	425	*NA*	*NA*

[a] $anova(lm(phenotype \sim strain + food + strain * food, data = docetaxel\_phenotypes))$

We next looked for GxE interactions between DMSO and docetaxel across the three environmental conditions and found evidence of an interaction in every case (**Figure 5-10B**). If the N2 strain was resistant in DMSO, the CB4856 strain would be resistant in docetaxel. Similarly, if the CB4856 strain was resistant in DMSO, the N2 strain would be resistant in docetaxel. A one-way ANOVA (**Table 5-3**) identified a statistically significant interaction between food quality and drug condition that can explain 19% of the total phenotypic variation (*p*-value < 2.2e-16). Together, these data suggest that the

variation in DMSO phenotypes seen across the different food qualities is directly influencing the docetaxel response. Overall, these results indicate that environmental conditions such as food quality can drastically affect drug-response phenotypes. Because we can no longer replicate the food conditions from the original mapping, it is impossible to validate and narrow this QTL in response to docetaxel.

Table 5-3: Results of ANOVA test[a] of genotype, food quality, drug condition, and various interactions

	DF	Sum Sq	Mean Sq	F value	P value
Strain	1	340	340	0.4696	0.493548
Condition	1	161549	161549	223.4059	<2.2e-16
Food	2	330317	165159	228.3970	<2.2e-16
Strain:Food	2	7521	3761	5.2005	0.005858
Strain:Condition	1	11197	11197	15.4839	9.671e-05
Condition:Food	2	195136	97568	134.9261	<2.2e-16
Residuals	439	317450	723	*NA*	*NA*

[a] $anova(lm(phenotype \sim strain + condition + food + strain * food + strain * condition + food * condition))$

5.5.4 Modified high-throughput assay does not facilitate candidate gene testing

Although the bacterial lysate could no longer provide satisfactory nutrition for 96 hours of growth, we showed that it was still sufficient to produce healthy L4 nematodes after 48 hours (**Figure 5-6**). In attempt to circumvent the problem observed with growth over 96 hours, we developed a modified assay where animals were grown for 48 hours, from L1 to L4, in the presence of docetaxel and then analyzed for the developmental traits animal length (median.TOF), optical density (median.EXT), and normalized optical density (median.norm.EXT) [69]. We treated the parental strains (N2 and CB4856) and two NILs with previously defined phenotypes (ECA599 and ECA672) to increasing concentrations of docetaxel (**Figure 5-11**). In this modified assay, we observed a dose-dependent response to docetaxel for all three traits. Interestingly, the N2 strain showed higher resistance to docetaxel compared to the CB4856 strain, as seen with our original NIL assays. If the N2 strain is resistant to docetaxel and the N2 allele on chromosome IV is also associated with docetaxel resistance, we would expect that the NILs with the N2 introgressions in the CB4856 genetic background would be intermediate in resistance compared to the N2 and CB4856 strains. However, these NILs were both more sensitive than either parental strain (**Figure 5-11**). These results suggest that this modified assay would not be useful in testing candidate

genes, like *rmd-4*, underlying the docetaxel response.

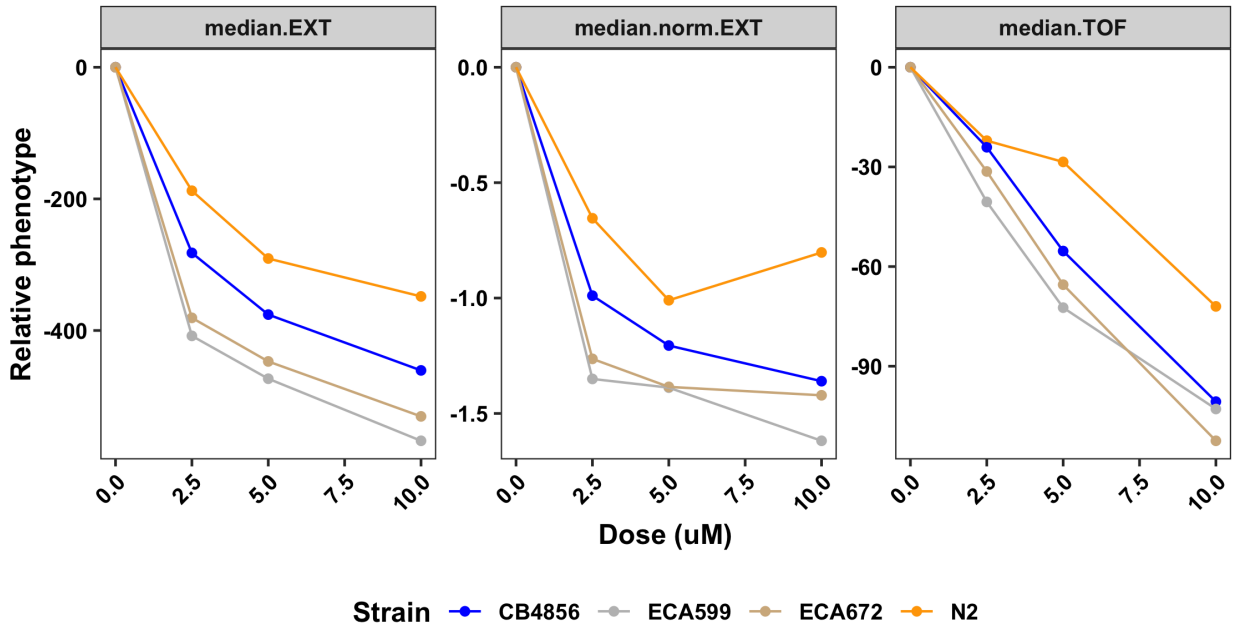


Figure 5-11: Modified high-throughput assay does not capture phenotypic differences between N2 and CB4856 strains on chromosome IV. Results from the docetaxel dose response modified HTA for animal length (median.TOF), animal optical density (median.EXT), and normalized optical density (median.norm.EXT) are shown. For each trait, drug concentration (uM) (x-axis) is plotted against the phenotype subtracted from control (y-axis) and colored by strain.

5.6 DISCUSSION

Here, we showed that phenotypic variation in response to the chemotherapeutic docetaxel maps to genetic variation on the center of chromosome IV. Using a medium quality of our HB101 lysate, we were able to narrow the QTL to a 1.8 Mb region containing the microtubule-regulating gene *rmd-4*. However, we discovered that this trait is highly susceptible to changes in the environment, such as food quality. In a post hoc analysis, we retrospectively identified significant GxE interactions between animal length in the DMSO control across three different lysate qualities for the N2 and CB4856 strains. We hypothesize that this variation in phenotype in DMSO is causing the inconsistencies seen in the docetaxel response between assays. However, we also provide evidence of a docetaxel response independent of the DMSO response. Without access to the same quality of lysate used in the linkage mapping experiments, we were not able to validate and narrow this QTL further. Regardless, we stress

the importance of using a stable food source when performing quantitative experiments.

The overlap between the QTL for control and drug conditions might create added complexity when trying to identify causal genetic variation. We attempt to account for any strain-to-strain variation in control conditions by performing a linear regression that subtracts the control phenotype from the drug phenotypes. In the RIAILs, there is a weak correlation between animal optical densities (mean.EXT: $rho = -0.18$, p = 0.0036) in DMSO and docetaxel. This correlation, although weak, might make it difficult to analyze the effect of docetaxel outside of the DMSO effect. This is certainly seen in our NIL assays where major variation in animal length in DMSO might influence the docetaxel phenotype. Isolating the DMSO QTL could be useful to further investigate docetaxel response. However, preliminary studies using NILs on the center of chromosome IV suggest that several loci contribute to the differential growth response.

Regardless of any GxE interactions observed, it is likely that genetic variation on the center of chromosome IV does impact docetaxel responses in *C. elegans*. However, it is obvious that genes play different roles in the docetaxel response based on the environmental conditions (*i.e.* food quality). Several studies have investigated the role of diet on drug metabolism and transport, particularly implicating that cytochrome P450 genes might be sensitive to changes in diet [291]. There are about 80 cytochrome P450 genes in *C. elegans*, four of which are contained within the conservative genomic interval defined by linkage mapping and two (*cyp-25A6* and *cyp-31A3*) are found inside of the NIL-defined confidence interval. Variation at one of these loci might be causing differential responses to docetaxel.

5.7 FUTURE DIRECTIONS

Although unintended, this project highlights the very real and very interesting aspects of gene-by-environment interactions. A great next step would be to directly test the impact of diet on growth and drug responses (like docetaxel) in different genetic backgrounds (such as N2 and CB4856) in an otherwise controlled environment. This different diet could be different strains of *E. coli* or the same strain of *E. coli* with certain parameters altered. One could also start to think about changing other environmental variables such as temperature or oxygen, however these could be more difficult.

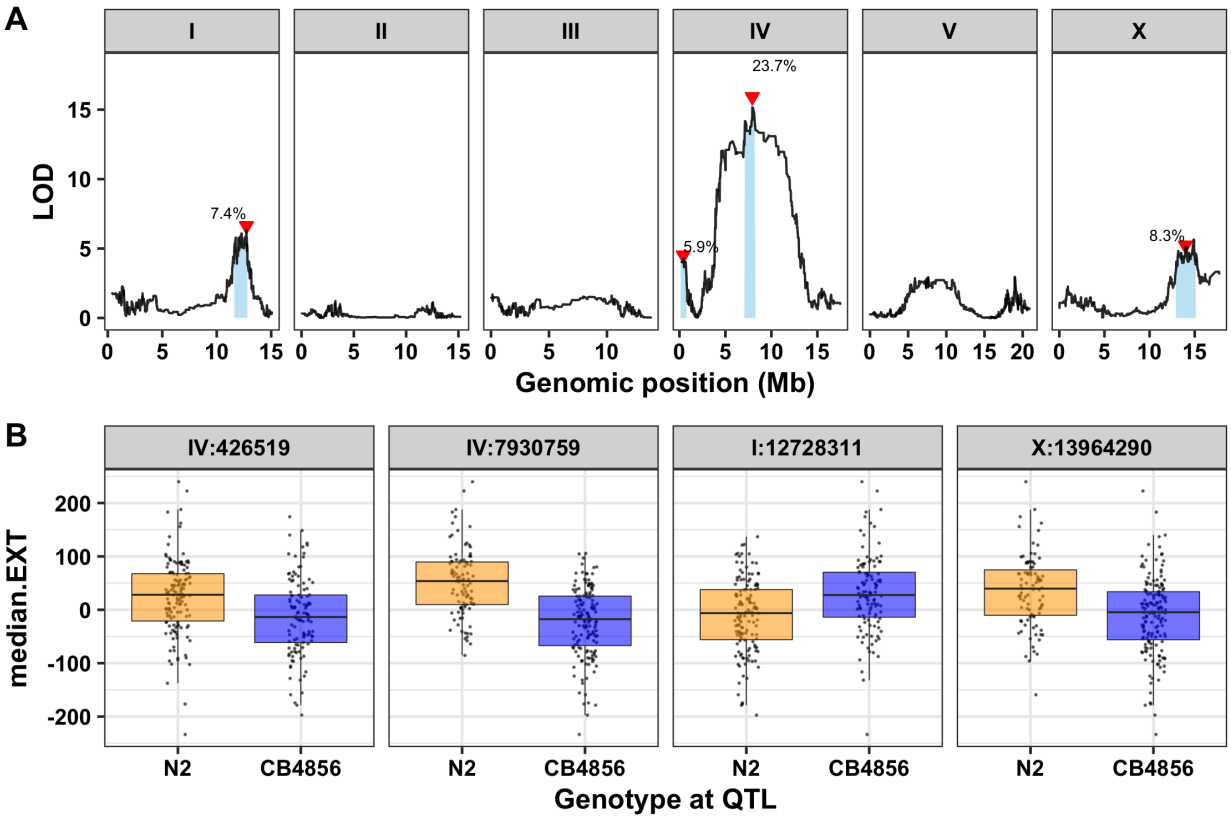
With reliable and replicable GxE interactions, one could perform a mapping to identify QTL that overlap between environments compared to QTL that are different between environments. These environment-specific QTL might provide additional insight into the complexity of inter-individual drug responses.

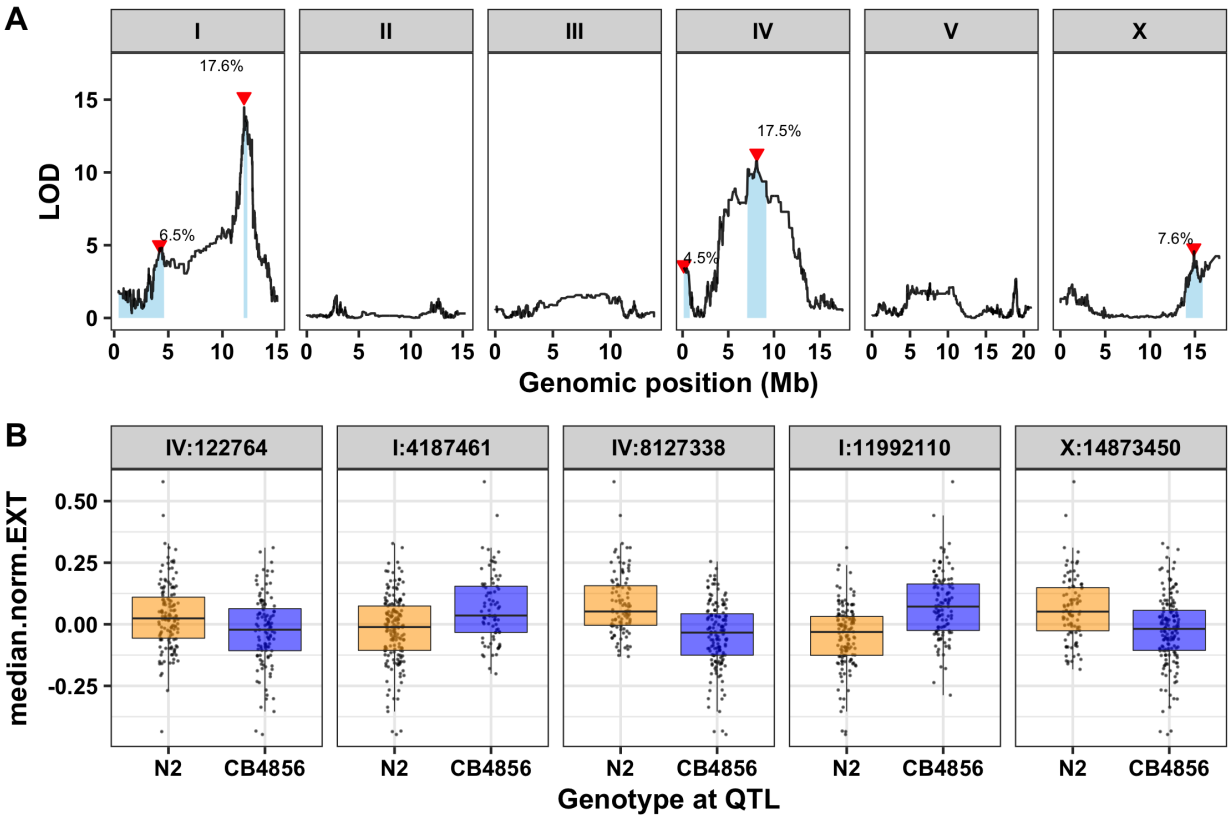
Unfortunately, because this phenotype seemed to be dependent on the food source and we no longer have the same food, it does not make sense to continue to follow this QTL. In order to learn more about how genetic variation impacts docetaxel responses in *C. elegans*, I believe it would be necessary to redo the QTL mapping fresh with a new and reproducible food source. Members of the Andersen Lab (mostly myself, Joy Nyaanga, and Dr. Tim Crombie) have worked hard to develop a new protocol for growing large-scale HB101 *E. coli* at a stage and density that is favorable for nematode growth. Using this protocol, new mappings could be performed with RIAILs and/or wild isolates to identify QTL. Ideally, we would discover a QTL on the center of chromosome IV. This result would reassure us that we were in fact chasing a docetaxel effect, not a food effect. No matter the location of the resulting QTL, it might be important to also identify any growth QTL in control conditions and isolate this region before fine-mapping with NILs.

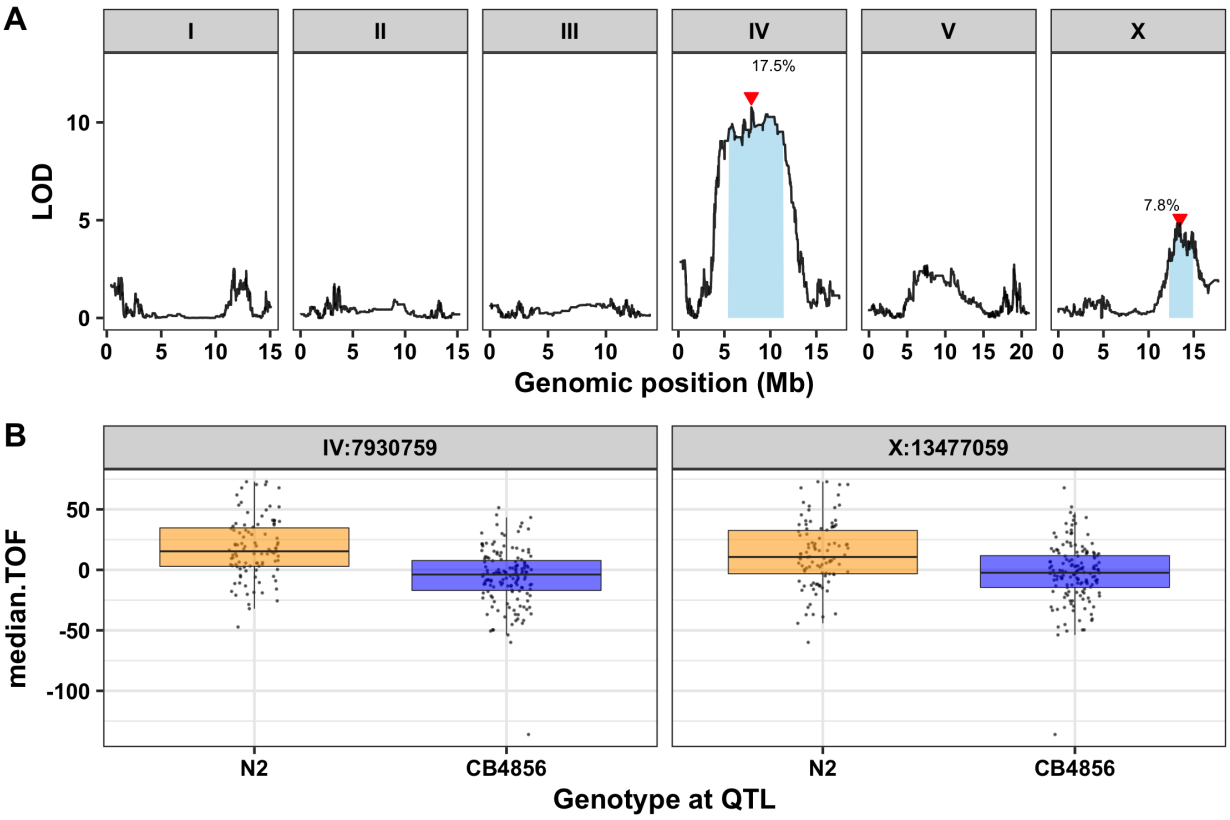
5.8 CONTRIBUTIONS

During this work, I was partially funded by the Cell and Molecular Biology of Disease training program. All NILs in this chapter were generated by myself and a visiting masters student, Joost van der Zwagg. I want to specially thank Dr. Charlie Baer for sending us his unused HB101 lysate, without which I never would have completed this project. I further want to thank all members of the Andersen Lab, particularly Dr. Shannon Brady and Dr. Daehan Lee, for helping me to see this through and realizing that, although the end result was not what I had hoped for, it is interesting and important nonetheless. Finally, thank you to Dr. Erik Andersen for his constant support throughout this project.

5.9 SUPPLEMENTAL FIGURES




A
I
II
III
IV
V
X
17.6%
6.5%
17.5%
4.5%
7.6%
LOD
Genomic position (Mb)
B
IV:122764
I:4187461
IV:8127338
I:11992110
X:14873450
median.norm.EXT
N2
CB4856
Genotype at QTL


A
I
II
III
IV
V
X
17.5%
7.8%
LOD
Genomic position (Mb)
B
IV:7930759
X:13477059
median.TOF
N2
CB4856
Genotype at QTL

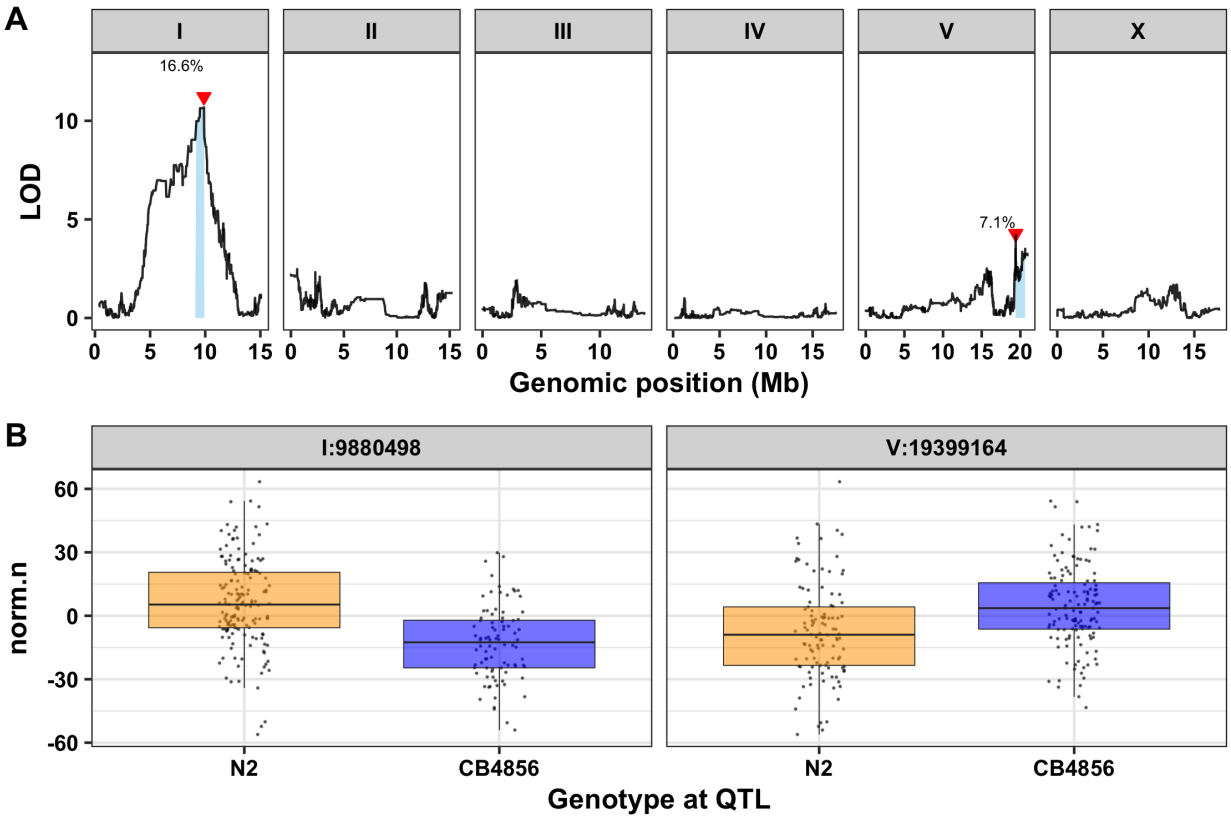


Figure S5-1: Linkage mapping results for docetaxel response. For each trait: **(A)** Linkage mapping results for is shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** For each QTL, the phenotype (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange and strains with the CB4856 allele are colored blue.

6 Genetic mapping of avermectin resistance in *Caenorhabditis elegans* identifies two novel loci that overlap with resistance loci in *Haemonchus contortus*

6.1 PREFACE

This project has been an orphan project in the lab for many years, passed between many hands. A previous postdoc in the lab (Dr. Mostafa Zamanian) first started working on it with his highly talented undergraduate, Grace Park. When he left to start his own lab, Grace continued to test NILs with the help of Erik, myself, and Briana Rodrigez. Because progress was slow without someone working on it full time, Dr. Shannon Brady and her talented undergraduate Ellen Chao began working on it alongside Grace and I. I think this is a really exciting project because of the real-world applicability. Many of my projects in the lab focus on cancer chemotherapeutics, which may or may not translate to human medicine. However, much of what we know about parasite resistance to anthelmintics was discovered in *C. elegans*. Furthermore, avermectins (abamectin for plants and ivermectin for humans and animals) are one of the main classes of anthelmintics. My grandpa even treats his cattle with Ivermectin! Resistance to these drugs is a growing problem and as there have been very few new drugs on the market in the last few decades, discovering loci that confer resistance is a primary focus of many research groups. I was lucky enough to be invited to give a talk about my contributions to this project at the Anthelmintics IV: From Discovery to Resistance conference held in Santa Monica, CA (Feb 2020). It was my first (and only) conference talk in graduate school and it was a great experience that deepened my appreciation for this work.

6.2 ABSTRACT

Parasitic nematodes cause a huge economic burden to humans, livestock, and agriculture worldwide. Anthelmintics have been widely successful in treating parasitic nematodes. However, resistance is increasing, and little is known about the molecular and genetic causes of resistance. The free-living roundworm *Caenorhabditis elegans* provides a tractable model to identify genomic regions that

underlie resistance. Unlike parasitic nematodes, *C. elegans* is easy to maintain in the laboratory, has a complete and well annotated genome, and has many genetic tools available for molecular assays. Using a panel of recombinant inbred lines constructed from crosses of two genetically and phenotypically divergent strains of *C. elegans*, we identified three genomic regions or quantitative trait loci (QTL) on chromosome V that underlie variation in response to the avermectin abamectin. Each of these loci acts additively on abamectin resistance. One QTL (*glc-1*) was identified previously and encodes an alpha subunit of a glutamate-gated chloride channel. A second QTL overlaps with a genomic region implicated in ivermectin resistance in the parasitic nematode *Haemonchus contortus*, demonstrating the relevance of our approach. The third QTL does not overlap with any known anthelmintic resistance loci and, therefore, could be unique to *C. elegans* or represent previously unidentified resistance alleles shared with parasitic nematodes. Here, we validated and narrowed the remaining two QTL using near-isogenic lines (NILs) and generated a list of prioritized candidate genes to evaluate using CRISPR-Cas9-mediated genome editing. Our work highlights the advantages of using *C. elegans* as a model to better understand avermectin resistance in parasitic nematodes.

6.3 INTRODUCTION

Parasitic nematodes, commonly termed helminths, pose a significant health and economic threat for countries around the world [292, 293, 294]. A number of serious diseases have been linked to parasitic nematode infections, many of which are traditionally underfunded. These diseases vary in both symptoms and severity but commonly affect developing nations in tropical and subtropical regions. It is estimated that almost two billion people suffer from infections with one or more species of helminth [295], and the loss of disability-adjusted life years caused by helminths ranks among the top of all Neglected Tropical Diseases [292]. In addition to their devastating impact on human health, several parasitic nematode species can infect a variety of key crops and livestock. These infections cause severe economic burdens worldwide [296].

Parasitic nematodes are largely treated using a limited number of anthelmintic drugs from one of the three main drug classes: benzimidazoles, nicotinic agonists, and macrocyclic lactones. Most anthelmintics are cheap and can efficiently eliminate parasitic nematodes from an infected individual

[297]. However, the introduction of mass drug administration to treat populations in endemic regions, combined with the lack of drug alternatives, has led to increased drug resistance [295]. This problem is highly prevalent for species that infect various livestock, where several studies report a majority of farms surveyed had anthelmintic-resistant nematodes [298, 299, 300]. In many cases, resistance has been shown to have a highly heritable component, suggesting the evolution of anthelmintic-resistant nematodes might occur under drug selection [141]. We must understand the drug mechanisms and identify the genetic loci that contribute to drug resistance in parasitic nematodes to provide effective long-term helminth treatments.

The avermectins abamectin and ivermectin are two common anthelmintic drugs used to treat agricultural and veterinary or human parasitic nematode infections, respectively [297, 140]. However, widespread resistance to avermectins has been reported and remains a significant concern [141]. Previous research has shown that avermectins target ligand-gated channels such as the glutamate-gated (GluCl) and GABA-gated (GABACl) chloride channels [140, 141]. Studies in the free-living nematode *Caenorhabditis elegans* have shown evidence that mutations in GluCl channels confer drug resistance [141]. Despite these findings, several resistant parasitic nematode isolates have been discovered and these isolates do not have mutations in genes that encode GluCl subunits, suggesting that alternative mechanisms of resistance to avermectins must exist. Recently, advances in quantitative trait loci (QTL) mapping in numerous species have identified several genomic regions of interest containing genetic variation that confers drug resistance [301, 302, 303, 304, 74, 75]. However, the identification of specific genes or variants can be challenging in most species [22].

It is difficult to identify the molecular basis of drug resistance in parasitic nematodes for several reasons. First, their life-cycles are long, complicated, and costly, often requiring the use of a host organism [305, 304]. Second, most species do not have chromosome-level and annotated genomes. To date, the most complete genome is from *Haemonchus contortus*, enabling genetic mapping and comparative genomic approaches [301]. Finally, most species lack several key molecular and genetic tools such as CRISPR-Cas9 genome editing [305]. By contrast, the free-living nematode *C. elegans* has a short life cycle that is easy to grow in the laboratory, a well annotated reference genome, and a plethora of molecular and genetic tools to characterize anthelmintic responses [75, 120, 74, 306, 99].

Genetic screens and selections performed in the laboratory-adapted reference strain, N2, have identified three GluCl channels (*glc-1*, *avr-14*, and *avr-15*) that mediate avermectin resistance in *C. elegans* [141]. Access to 328 genetically and phenotypically diverged wild strains of *C. elegans* collected around the world could identify other loci that contribute to avermectin resistance across natural populations [120, 75, 63].

Here, we use linkage mapping analysis to identify three large-effect QTL on chromosome V that contribute to resistance to abamectin. One of these QTL was previously identified and is known to be caused by variation in the GluCl channel *glc-1* [120, 99]. The remaining two QTL are novel and might overlap with QTL identified for ivermectin resistance in *H. contortus* and *Teladorsagia circumcincta* [301, 302]. We used near-isogenic lines (NILs) to validate and narrow each QTL independently and suggest candidate genes in regions of interest to further test using CRISPR-Cas9 genome editing. This study demonstrates the value of combining data collected from parasitic nematodes with molecular and quantitative assays performed in *C. elegans* to collectively work toward understanding the molecular mechanisms of anthelmintic resistance.

6.4 METHODS

6.4.1 Strains

Animals were grown on modified nematode growth media (NGMA) containing 1% agar and 0.7% agarose at 20℃ and fed the *E. coli* strain OP50 [100]. A total of 225 recombinant inbred advanced intercross lines (RIAILs) were generated previously [68] and assayed here. The two parental strains of the recombinant lines were a derivative of the canonical laboratory strain N2, QX1430 (which contains the CB4856 allele at the *npr-1* locus and a transposon insertion in *peel-1*), and the wild isolate from Hawaii, CB4856. Near-isogenic lines (NILs) were generated by backcrossing a selected RIAIL to either N2 or CB4856 for several generations [72] using PCR amplicons for insertion-deletion events (indels) to track the introgressed region. NILs were whole-genome sequenced to verify clean introgressions. *lgc-54* mutants (FX3448 and FX3518) were obtained from the National BioResource Project (Japan) - *C. elegans*. All strains are available upon request or from the *C. elegans* Natural

Diversity Resource [63].

6.4.2 High-throughput fitness assays for linkage mapping

For all phenotyping assays, we used a high-throughput fitness assay described previously [68]. In summary, populations of each strain were passaged and amplified on NGMA plates for four generations, bleach-synchronized, and 25-50 embryos were aliquoted into 96-well microtiter plates at a final volume of 50 μL K medium [173]. After 12 hours, arrested L1s were fed HB101 bacterial lysate (Pennsylvania State University Shared Fermentation Facility, State College, PA; [174]) at a final concentration of 5 mg/mL in K medium and were grown for 48 hours to the L4 larval stage at 20℃ with constant shaking. Three L4 larvae were sorted into new 96-well microtiter plates containing 10 mg/mL HB101 bacterial lysate, 50 μM kanamycin, and either 1% DMSO or abamectin dissolved in 1% DMSO using a large-particle flow cytometer (COPAS BIOSORT, Union Biometrica; Holliston, MA). Sorted animals were grown for 96 hours at 20℃ with constant shaking. The next generation of animals and the parents were treated with sodium azide (50 mM in 1X M9) to straighten their bodies for more accurate length measurements. Animal length (mean.TOF), optical density integrated over animal length (mean.EXT), and brood size (norm.n) were quantified for each well using the COPAS BIOSORT. Nematodes get longer (animal length) and become thicker and more complex (optical density) over developmental time. Phenotypic measurements collected by the BIOSORT were processed and analyzed using the R package *easysorter* [77] as described previously [72]. Differences among strains within the control conditions were controlled by subtracting the mean control-condition value from each drug-condition replicate for each strain using a linear model $drug\_phenotype \sim mean\_control\_phenotype$. In this way, we are addressing only the differences among strains that were caused by the drug condition and the variance in the control condition does not affect the variance in the drug condition.

6.4.3 Abamectin dose response

Four genetically divergent strains (N2, CB4856, JU775, and DL238) were treated with increasing concentrations of abamectin using the standard high-throughput assay described above. A concentration of 5 μM abamectin (Sigma, #31732-100MG) in DMSO was selected for the linkage

mapping experiments and 7.5 nM abamectin in DMSO was selected for the genome-wide association mapping and NIL experiments. These concentrations provided a reproducible abamectin-specific effect that maximizes between-strain variation and minimizes within-strain variation across the three traits. The discrepancy between these two concentrations is likely caused by independent drug orders and dissolutions.

6.4.4 Linkage mapping

A total of 225 RIAILs [68] were phenotyped in abamectin and DMSO using the HTA described above. Linkage mapping was performed on the measured traits using the R package *linkagemapping* (https://github.com/AndersenLab/linkagemapping) as described previously [72, 69]. The cross object derived from the whole-genome sequencing of the RIAILs containing 13,003 SNVs was merged with the RIAIL phenotypes using the *merge_pheno* function with the argument *set = 2*. A forward search (*fsearch* function) adapted from the *R/qtl* package [176] was used to calculate the logarithm of the odds (LOD) scores for each genetic marker and each trait as $-n(ln(1-R2)/2ln(10))$ where R is the Pearson correlation coefficient between the RIAIL genotypes at the marker and trait phenotypes [27]. A 5% genome-wide error rate was calculated by permuting the RIAIL phenotypes 1000 times. QTL were identified as the genetic marker with the highest LOD score above the significance threshold. This marker was then integrated into the model as a cofactor and mapping was repeated iteratively until no further QTL were identified. Finally, the *annotate_lods* function was used to calculate the effect size of each QTL and determine 95% confidence intervals defined by a 1.5 LOD drop from the peak marker using the argument *cutoff = "proximal"*.

6.4.5 Genome-wide association mapping

A total of 210 wild isolates were phenotyped in both abamectin and DMSO using the standard high-throughput assay described above. A genome-wide association (GWA) mapping was performed for animal optical density (mean.EXT), length (mean.TOF), and brood size (norm.n) using the R package *cegwas2* (https://github.com/AndersenLab/cegwas2-nf) as described previously [70, 123]. Genotype data were acquired from the latest VCF release (release 20180517) from CeNDR. We used

BCFtools [261] to filter variants below a 5% minor allele frequency and variants with missing genotypes and used PLINK v1.9 [262, 263] to LD-prune genotypes. The additive kinship matrix was generated from the 83,019 markers using the *A.mat* function in the *rrBLUP* R package [264]. Because these markers have high LD, we performed eigen decomposition of the correlation matrix of the genotype matrix to identify 1185 independent tests [70]. We performed genome-wide association mapping using the *GWAS* function from the *rrBLUP* package. Significance was determined by an eigenvalue threshold set by the number of independent tests in the genotype matrix [70]. Confidence intervals were defined as $\pm$ 150 SNVs from the rightmost and leftmost markers that passed the significance threshold.

6.4.6 Identification of orthologous genes between *C. elegans* and *H. contortus*

A list of *H. contortus* genes within the QTL interval was created from the gene annotation (GFF3) file downloaded from Wormbase Parasite (parasite.wormbase.org; PRJNA205202). Protein sequences for the 384 genes were extracted from the protein FASTA file (parasite.wormbase.org) and compared to the *C. elegans* protein database (PRJNA13758; WBPS12) using BLASTp [307]. Orthologs with the highest percent identity match between species were selected for further analysis. *C. elegans* genes were filtered based on genomic position and analyzed for functional descriptions and GO annotations.

6.4.7 Modified high-throughput fitness assay

A modified version of the standard high-throughput assay detailed above was used to test docetaxel response as previously described [72, 69]. Briefly, strains were propagated for two generations, bleach-synchronized, and titered at a concentration of 25-50 embryos per well of a 96-well microtiter plate. Arrested L1s were fed HB101 bacterial lysate the following day at a final concentration of 5 mg/mL with either DMSO or docetaxel in DMSO. After 48 hours of growth at 20°C with constant shaking, nematodes were treated with sodium azide (5 mM in water) prior to analysis of animal length and optical density using the COPAS BIOSORT. Because only one generation of growth was observed, brood size was not calculated.

6.4.8 Statistical analysis

All gene position data for the *C. elegans* genome was collected using WormBase WS273. For NIL assays, complete pairwise strain comparisons were performed on drug residual phenotypes using a *TukeyHSD* function [175] on an ANOVA model with the formula $phenotype \sim strain$.

6.5 RESULTS

6.5.1 Three additive loci on chromosome V underlie differential responses to abamectin

We measured *C. elegans* development and anthelmintic sensitivity as a function of animal length (mean.TOF), optical density (mean.EXT), and brood size (norm.n) with a high-throughput assay developed with the COPAS BIOSORT (See Methods) [70, 72, 73, 71, 68, 69, 123]. We exposed four genetically divergent strains (N2, CB4856, JU775, and DL238) to increasing doses of abamectin. In the presence of abamectin, nematodes were generally smaller, less optically dense, and produced smaller broods compared to non-treated nematodes, suggesting an abamectin-induced reproductive and developmental delay. (**Figure 6-1**). In addition to this general trend, we also observed significant phenotypic variation among strains, indicating that genetic variation between strains might control the abamectin-response phenotype.

To investigate the genetic basis of natural abamectin resistance, we exposed 210 wild isolates to abamectin and measured their developmental rates and brood sizes. We performed genome-wide association (GWA) mapping and identified a total of 15 QTL across the three traits (animal length, optical density, and brood size) on chromosomes I, II, III, and V (**Figure S6-1**). In particular, all three traits were correlated with genetic variation on chromosome V. Based on QTL overlap among the traits, we estimate that three independent loci on chromosome V (referred to as VL (left-most), VC (center), and VR (right-most)) underlie responses to abamectin (**Figure 6-2A**). To avoid the redundant analysis of correlated traits, we will focus on mean animal optical density, for which all three QTL on chromosome V were identified.

In parallel, we measured animal length, optical density, and brood size in response to abamectin for a panel of 225 RIAILs generated by a cross between the N2 and CB4856 strains [68]. Linkage

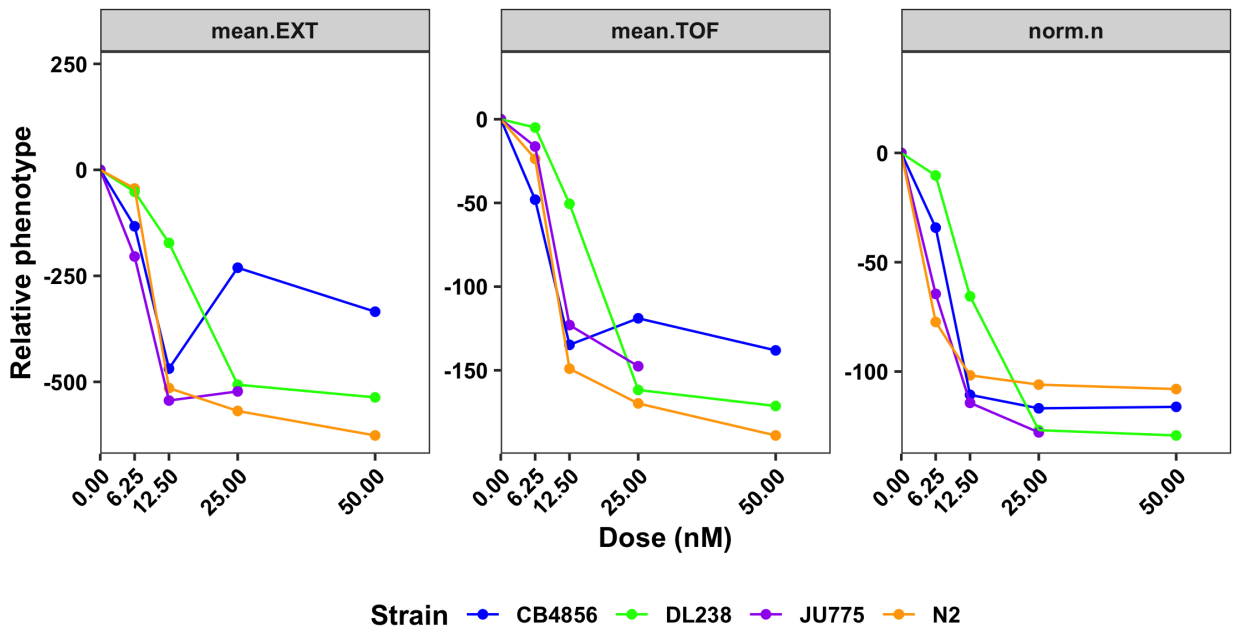
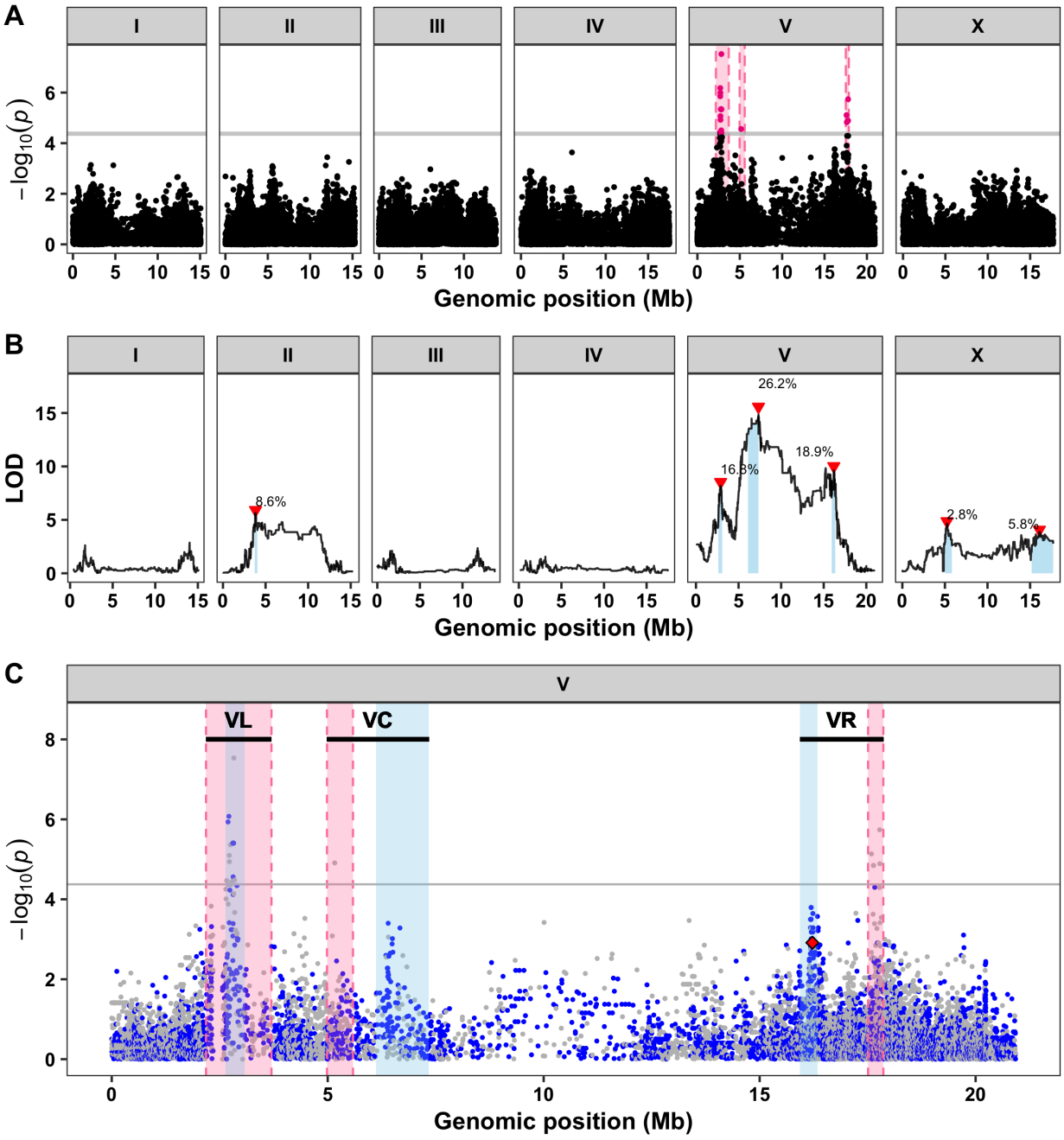


Figure 6-1: Dose response with four divergent wild isolates. Results from the abamectin dose response HTA for brood size (norm.n), animal length (mean.TOF), and animal optical density (mean.EXT). For each trait, drug concentration (nM) (x-axis) is plotted against phenotype subtracted from control (y-axis) and colored by strain (CB4856: blue, DL238: green, JU775: purple, N2: orange). A concentration of 7.5 nM was chosen for future experiments.

mapping analysis for all three traits identified a total of ten QTL on chromosomes II, V, and X including three distinct QTL on chromosome V (**Figure S6-2**). At each of the three loci identified on chromosome V, the CB4856 allele was correlated with resistance (denser animals) in abamectin compared to the N2 allele (**Figure S6-2**). To investigate genetic interactions among the three QTL on chromosome V and other loci throughout the genome, we performed a two-dimensional genome scan and found evidence of additive QTL on chromosome V but no significant interactions (**Figure 6-3**). To more specifically address the additivity of these three QTL, we compared phenotypes of RIAILs with each of the eight unique genotype combinations (two possible genotypes, N2 or CB4856, across three loci). As expected, strains with the N2 allele at all three loci were the most sensitive to abamectin and strains with the CB4856 allele at all three loci were the most resistant (**Figure 6-4**). Furthermore, strains with one CB4856 allele, regardless of the locus, were equally more resistant than strains with no CB4856 alleles and equally less resistant than stains with two CB4856 alleles. This result suggests that the effect of each locus is comparable and that there are no significant interactions between any


A
I
II
III
IV
V
X
$-\log_{10}(p)$
Genomic position (Mb)
B
LOD
8.6%
26.2%
16.3%
18.9%
2.8%
5.8%
Genomic position (Mb)
C
V
VL
VC
VR
$-\log_{10}(p)$
Genomic position (Mb)

two loci.

The combination of both association and linkage mapping provides a method to narrow the causal variants by looking for overlapping QTL. Overlapping QTL between the two methods, as seen with the VL QTL (**Figure 6-2C**, **Figure 6-5**), suggest that a common variant present in the CB4856 strain might cause differences in abamectin responses in both mapping populations. Alternatively, non-overlapping QTL, as seen with the VC and VR QTL (**Figure 6-2C**, **Figure 6-5**), suggests that a rare variant in the CB4856 strain drives the QTL identified from linkage mapping and a common variant not found in the CB4856 strain drives the QTL identified from association mapping. It is somewhat surprising that the VR locus does not overlap between the two methods because a relatively common four-amino-acid deletion in the glutamate-gated chloride (GluCl) channel *glc-1* had been previously discovered to underlie phenotypic differences in both swimming paralysis [120] and survival [99] in the presence of abamectin using three different mapping populations. Here, we test a third trait, nematode development, in two additional mapping populations, but only identified the *glc-1* locus with the RIAILs (**Figure 6-2B**). These results could suggest that other genes, in addition to *glc-1*, contribute to abamectin resistance on the right of chromosome V in the wild isolate population. However, given prior knowledge about the causal gene *glc-1*, it is likely that the differences in population structure between the wild isolates and RIAILs cause the same QTL to be identified in two distinct, yet linked, regions (**Figure 6-2C**).

Figure 6-2 *(preceding page)*: Three large-effect QTL on chromosome V control differences in abamectin responses. **(A)** Linkage mapping results for optical density (mean.EXT) is shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle covers the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** Genome-wide association mapping results is shown. Genomic position (x-axis) is plotted against the -log10(*p*) value (y-axis) for each SNV. SNVs are colored pink if they pass the genome-wide significance threshold designated by the grey line. The genomic regions of interest that pass the significance threshold are highlighted by pink rectangles. **(C)** Fine mapping of all common variants on chromosome V is shown. Genomic position (x-axis) is plotted against the -log10(*p*) values (y-axis) for each variant and colored by the genotype of the variant in the CB4856 strain (grey = reference, blue = alternative). Genomic regions identified from linkage mapping analysis are highlighted in blue and genomic regions identified from association mapping are highlighted in pink. The horizontal grey line represents the significance threshold. The red diamond represents the most significant variant in the gene *glc-1*.

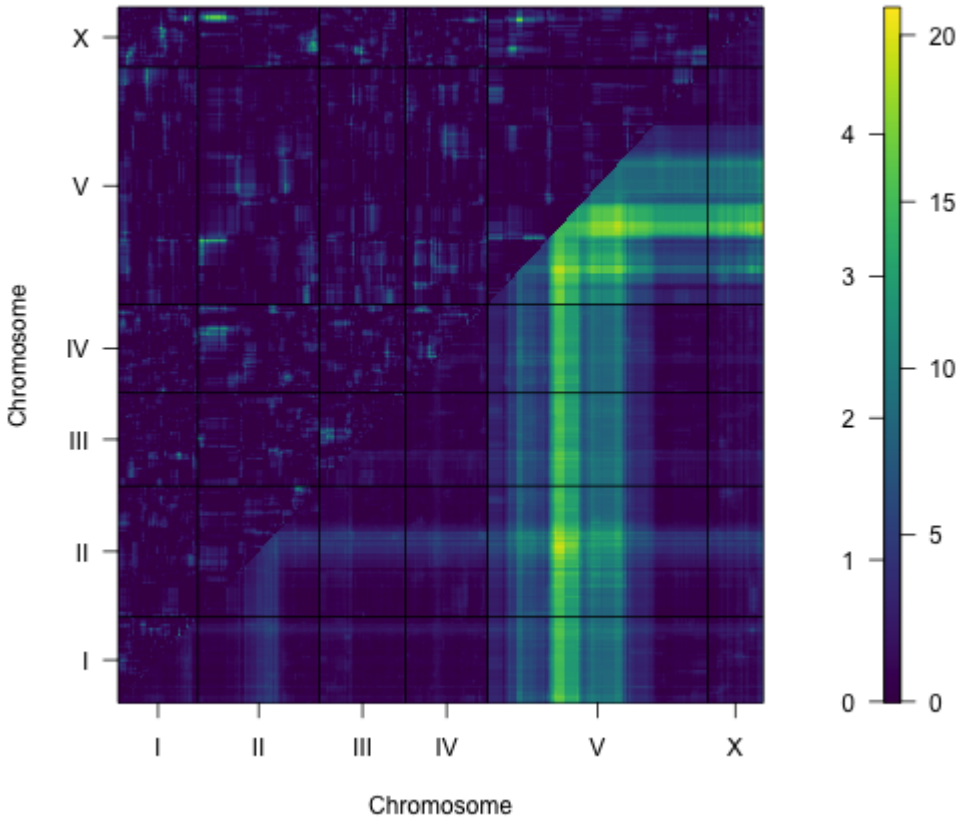


Figure 6-3: Two-dimensional genome scan for mean optical density (mean.EXT) in abamectin. Log of the odds (LOD) scores are shown for each pairwise combination of loci, split by chromosome. The upper-left triangle contains the epistasis LOD scores (interaction effects), and the lower-right triangle contains the LOD scores for the full model (both interaction and additive effects). LOD scores are colored by significance, increasing from purple to green to yellow. The LOD scores for the epistasis model are shown on the left of the color scale, and the LOD scores for the full model are shown on the right.

6.5.2 Near-isogenic lines validate effects of all three loci

To validate that genetic variation on chromosome V between the N2 and CB4856 strains contributes to abamectin resistance, we first generated chromosome substitution strains in which the entire chromosome V from the CB4856 strain was introgressed into the N2 genetic background and vice versa. We measured the animal lengths, optical densities, and brood sizes of these strains and observed that the genotype on chromosome V significantly contributed to differences in abamectin resistance. The strains with the CB4856 genotype on chromosome V in the N2 genetic background

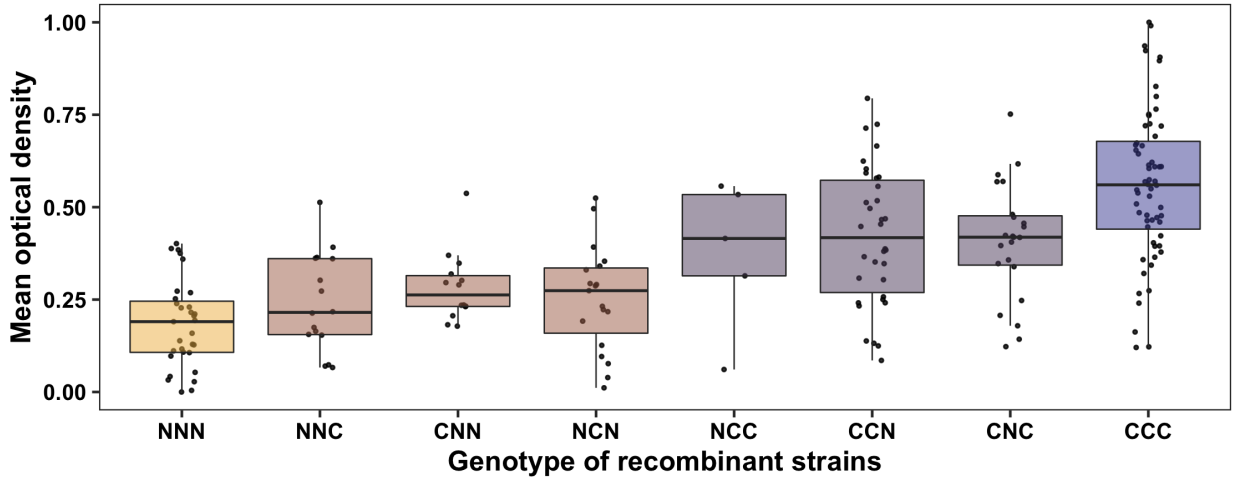


Figure 6-4: The three loci on chromosome V are independent and additive. Relative mean optical density in abamectin (mean.EXT, y-axis) for each RIAIL is plotted as Tukey box plots by genotype at the three QTL on chromosome V (y-axis). The x-axis labels are written with the N2 allele as represented by an "N" and the CB4856 allele as represented by a "C" in the order left-most (VL), central (VC), and right-most (VR) QTL. For example, the group "NCN" has the N2 allele at both the left-most (VL) and right-most (VR) loci and the CB4856 allele at the central (VC) locus. Groups are colored by the number of N2 and CB4856 alleles so that more orange colors represent more N2 alleles and more blue for more CB4856 alleles at the three QTL.

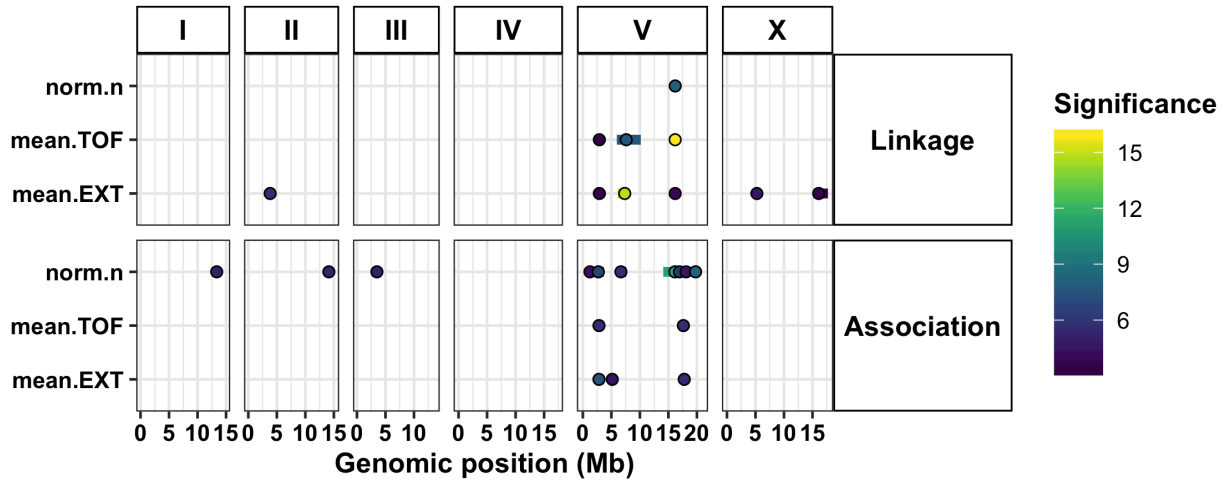


Figure 6-5: Summary of QTL mapping for responses to abamectin. Genomic positions (x-axis) of all QTL identified from linkage mapping (top) and association mapping (bottom) are shown for each drug-trait (y-axis). Each QTL is plotted as a point at the genomic location of the peak marker and a line that represents the confidence interval. QTL are colored by the significance of the LOD score (linkage) or -log10(*p*) value (association), increasing from purple to green to yellow.

were significantly more resistant than the sensitive N2 strain (*p*-value = 2.47e-06) (**Figure 6-6**). Similarly, the strains with the N2 chromosome V in the CB4856 genetic background were significantly more sensitive to abamectin compared to the resistant CB4856 strain (*p*-value = 0.0041, **Figure 6-6**).

These results demonstrate that genetic variation between the N2 and CB4856 strains at one or more loci on chromosome V contribute to the difference in abamectin sensitivity between these strains.

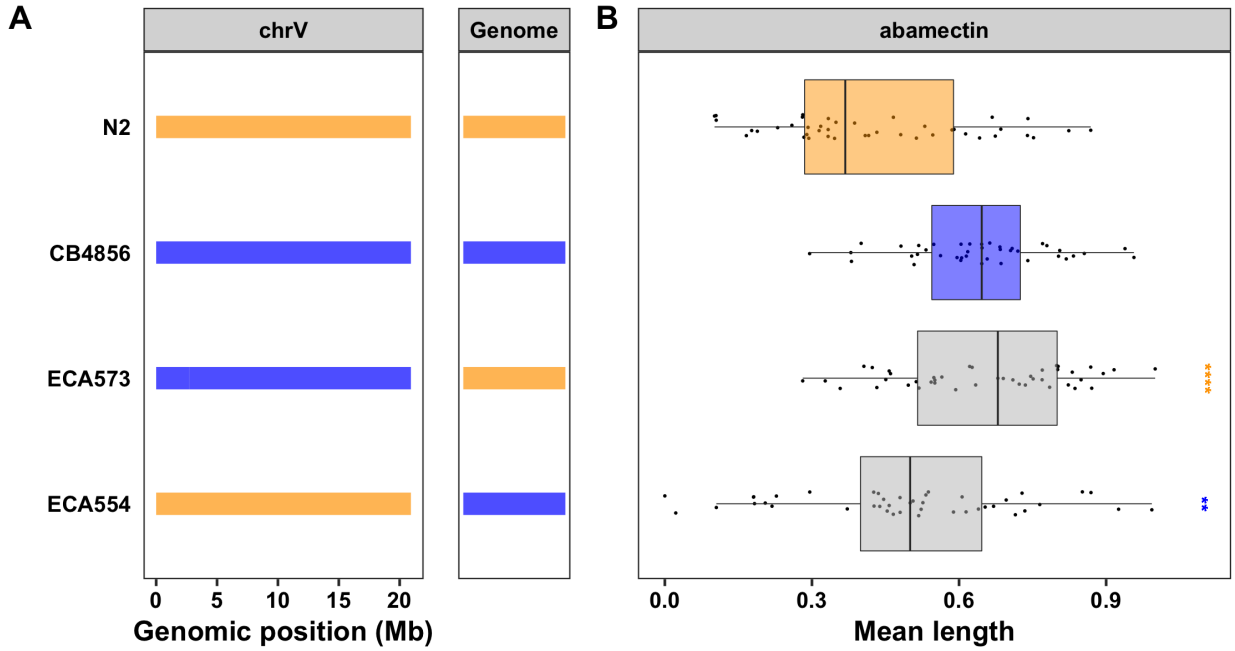


Figure 6-6: Chromosome substitution strains validate the existence of one or more resistance loci on chromosome V. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (right). **(B)** Relative mean lengths in abamectin (mean.TOF, x-axis) are plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL as compared to N2 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

The strong effect of the *glc-1* locus has been previously demonstrated [120, 99]. Therefore, it was important that we show that the two novel QTL on the left and center of chromosome V also contribute to the overall resistance phenotype observed in the chromosome substitution strain. We generated a near-isogenic line (NIL) that contains the resistant CB4856 alleles at both the VL and VC loci and the sensitive N2 allele at the VR *glc-1* locus (**Figure 6-7A**). When tested, we observed that this strain (ECA1059) was significantly more resistant to abamectin than the N2 strain (p-value = 8.83e-14) and less resistant than the CB4856 strain (p-value = 1.57e-13, **Figure 6-7B**). This result indicates that genetic variation besides *glc-1* on chromosome V contributes to the differences in abamectin sensitivity between the N2 and CB4856 strains.

To further isolate each QTL independently, we generated three NILs containing approximately 5 Mb of the CB4856 genome introgressed into the N2 genetic background at different locations on

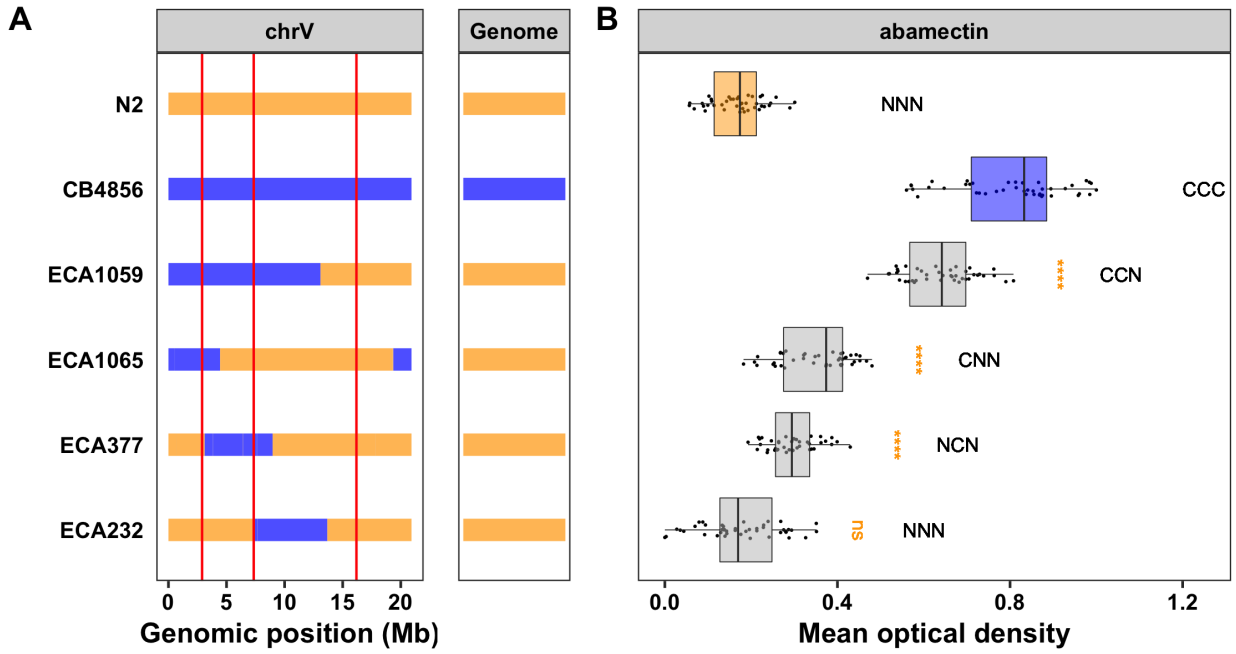


Figure 6-7: Near-isogenic lines confirmed the additive effects of all three QTL. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left box) and in general for the rest of the chromosomes (right box). The solid vertical lines represent the peak marker of each QTL. **(B)** Relative mean optical density in abamectin (mean.EXT, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to N2 is shown above each strain (ns = non-significant (p-value > 0.05), *** = significant (p-values < 0.0001). Predicted genotypes at the three QTL are shown above each strain from VL to VC to VR (N = N2 allele, C = CB4856 allele).

chromosome V so that they tile across the introgressed region in ECA1059 (**Figure 6-7A**). The strain ECA232 was not significantly more resistant to abamectin compared to the N2 strain (p-value = 0.997), suggesting that this NIL has the sensitive N2 alleles at all three loci on chromosome V (**Figure 6-7B**). Alternatively, both ECA1065 and ECA377 were significantly more resistant to abamectin than the N2 strain (p-values = 1.45e-13 and 5.44e-09, respectively), suggesting that the introgressed regions in both of these NILs contain one or more resistant loci (**Figure 6-7B**). Because both strains are less resistant than the NIL with two CB4856 alleles (ECA1059-ECA1065 p-value = 8.83e-14, ECA1059-ECA377 p-value = 8.83e-14), we can deduce that ECA1065 and ECA377 each contain one resistant locus (**Figure 6-7B**). Because the introgressions in these two NILs overlap by 1.3 Mb, this leaves two possibilities: either this overlapped region (V:3,120,168-4,446,729) contains a single QTL shared by the two NILs or each NIL validates a separate QTL within the non-overlapping regions. Because we identified three QTL from both linkage and association mapping, we believe the latter

Table 6-1: Genomic regions significantly correlated with abamectin resistance[a]

QTL	Association mapping	Linkage mapping	NIL-defined interval
VL	V:2,182,093-3,698,359	V:2,629,324-3,076,312	V:1-3,120,167
VC	V:4,983,265-5,585,155	V:6,118,360-7,342,129	V:5,260,997-5,906,132
VR	V:17,510,186-17,863,725	V:15,933,659-16,336,743	13,678,801-19,303,558 (*glc-1*[b])

[a]Mean animal optical density
[b]From previous data [120]

case that ECA1065 has the CB4856 allele for the VL locus and ECA377 has the CB4856 allele for the VC locus (**Table 6-1**, **Figure 6-8**).

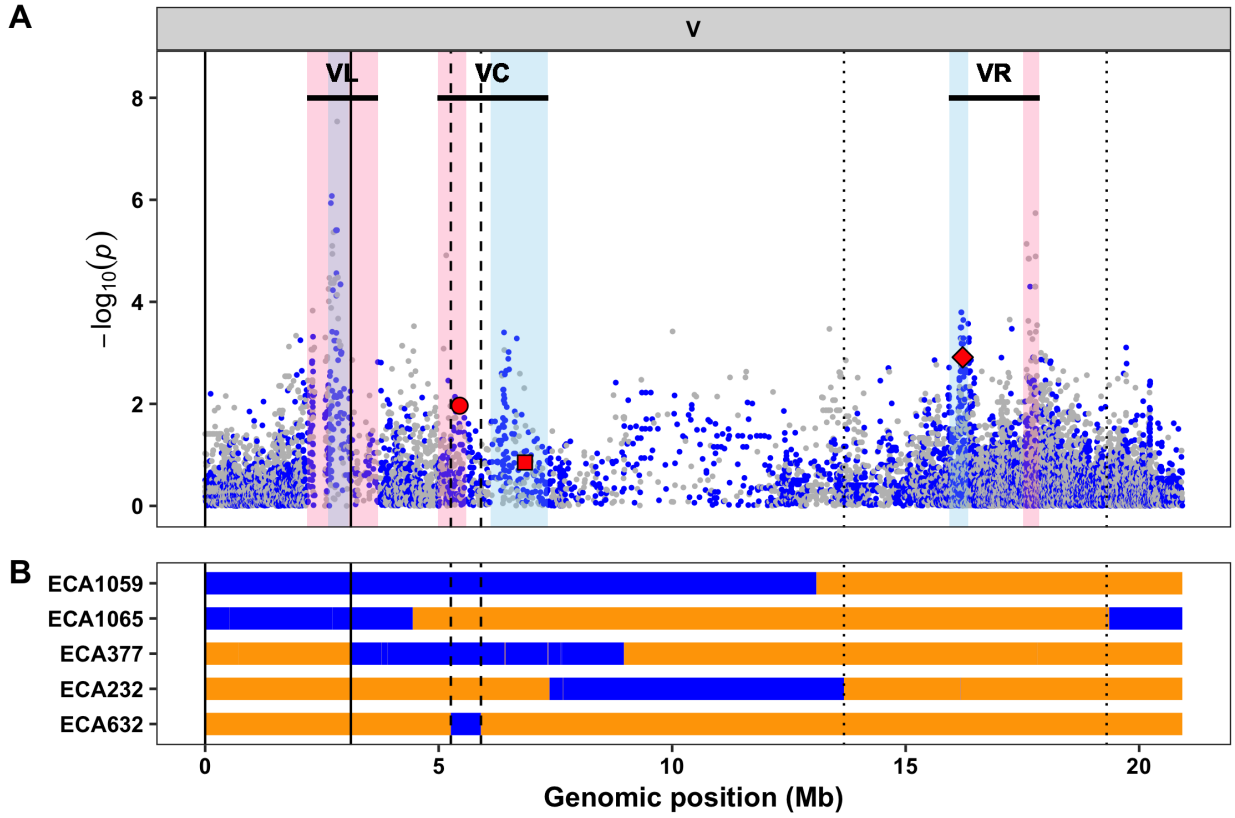


Figure 6-8: Refining QTL positions with NILs. **(A)** Fine mapping of all common variants on chromosome V is shown. Genomic position (x-axis) is plotted against the -log10(*p*) values (y-axis) for each variant and colored by the genotype of the variant in the CB4856 strain (grey = reference, blue = alternative). Genomic regions identified from linkage mapping analysis are highlighted in blue and genomic regions identified from association mapping are highlighted in pink. The horizontal grey line represents the genome-wide eigen-decomposition significance threshold. The red points represent the approximate locations of *glc-1* (diamond), *glc-3* (circle), and *lgc-54* (square). The vertical lines represent the smallest NIL-defined genomic region for the VL (solid), VC (dashed), and VR (dotted) QTL. **(B)** Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V. The vertical lines represent the smallest NIL-defined genomic region for the VL (solid), VC (dashed), and VR (dotted) QTL.

The VC QTL is contained within the CB4856 introgression in ECA377 and is defined by the CB4856 introgressions in the NILs ECA1065 and ECA232 (V:4,446,729-7,374,928) (**Figure 6-7**, **Figure 6-8**). This large 3 Mb interval encompasses both of the genomic regions identified using linkage and association mappings (**Figure 6-8**, **Table 6-1**). We next attempted to narrow this region further by generating additional NILs with smaller introgressions and measuring the lengths and optical densities of these strains. All four NILs with the CB4856 introgression on the center of chromosome V were significantly longer in abamectin than the N2 strain (*p*-values < 2.83e-06, **Figure 6-9**). This result suggests that the QTL position was contained within the smallest introgression, ECA632 (V:5,260,997-5,906,132) (**Table 6-1**, **Figure 6-8**). The results for optical density were similar, but the variation within strains was higher (**Table 6-10**). Interestingly, this 645 kb region overlaps with the genomic interval defined by association mapping but not the confidence interval identified from linkage mapping (**Figure 6-8**). However, the broad peak on the center of chromosome V observed from linkage mapping might suggest that the confidence interval is underestimated, possibly because of the linkage between the three loci on chromosome V. Alternatively, other rare variants within the CB4856 strain on the center of chromosome V might contribute marginal effects to abamectin resistance that could result in this discrepancy.

This NIL-defined genomic interval contains 206 genes. Of those genes, 116 harbor genetic variation in CB4856 that is shared with at least 5% of the mapping population and 45 of these genes have common genetic variation that causes a change in the amino acid sequence of the protein (protein-coding variation). Notably, the glutamate-gated chloride channel, *glc-3*, resides within this narrowed region (V:5,449,287, **Figure 6-8**). GLC-3 has been previously implicated in avermectin resistance *in vitro* [308]. However, it has yet to be identified in mutant screens nor tested for avermectin resistance *in vivo*. The CB4856 strain harbors 10 common variants in *glc-3*, including a single missense variant (I493F) that is also present in 10 other wild isolates. It is possible that genetic variation in *glc-3* is causing increased resistance to abamectin in the CB4856 strain.

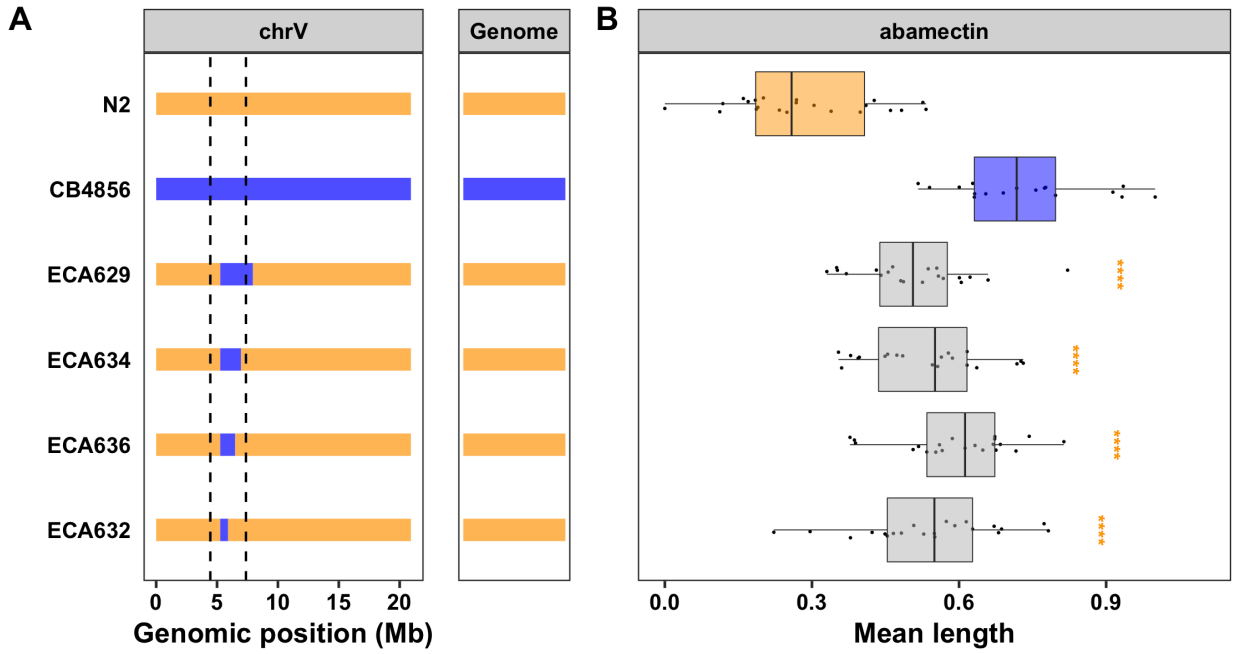


Figure 6-9: NILs isolate and narrow the VC QTL. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (right). The dashed vertical lines represent the previous NIL-defined QTL interval for VC. **(B)** Relative mean animal length in abamectin (mean.TOF, x-axis) is plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL compared to N2 is shown above each strain (ns = non-significant (p-value > 0.05), *** = significant (p-values < 0.0001).

6.5.3 Overlap of *C. elegans* and parasitic nematode candidate genes for avermectin resistance

Several previous studies have identified QTL in parasitic nematode models that provide candidate genes that might underlie responses to ivermectin, an avermectin closely related to abamectin. Introgression mapping in the sheep parasite *T. circumcincta* identified several potential candidate genes, including the ortholog of the *C. elegans* gene *lgc-54* [302], but this genome is highly fragmented and genomic locations are likely not correct. Regardless, this gene encodes a ligand-gated chloride channel, but it has yet to be directly implicated in avermectin resistance. Interestingly, *lgc-54* resides on the *C. elegans* chromosome V near 6.8 Mb. Although this gene is outside of the NIL-defined interval (**Figure 6-9**, **Table 6-1**, **Figure 6-8**), it is well within the confidence interval defined by the linkage mapping experiment (**Table 6-1**). Furthermore, the CB4856 strain harbors four genetic variants in this gene, including a common nonsense variant in the first exon. To test if *lgc-54* plays a role in abamectin resistance in *C. elegans*, we exposed two independent *lgc-54*

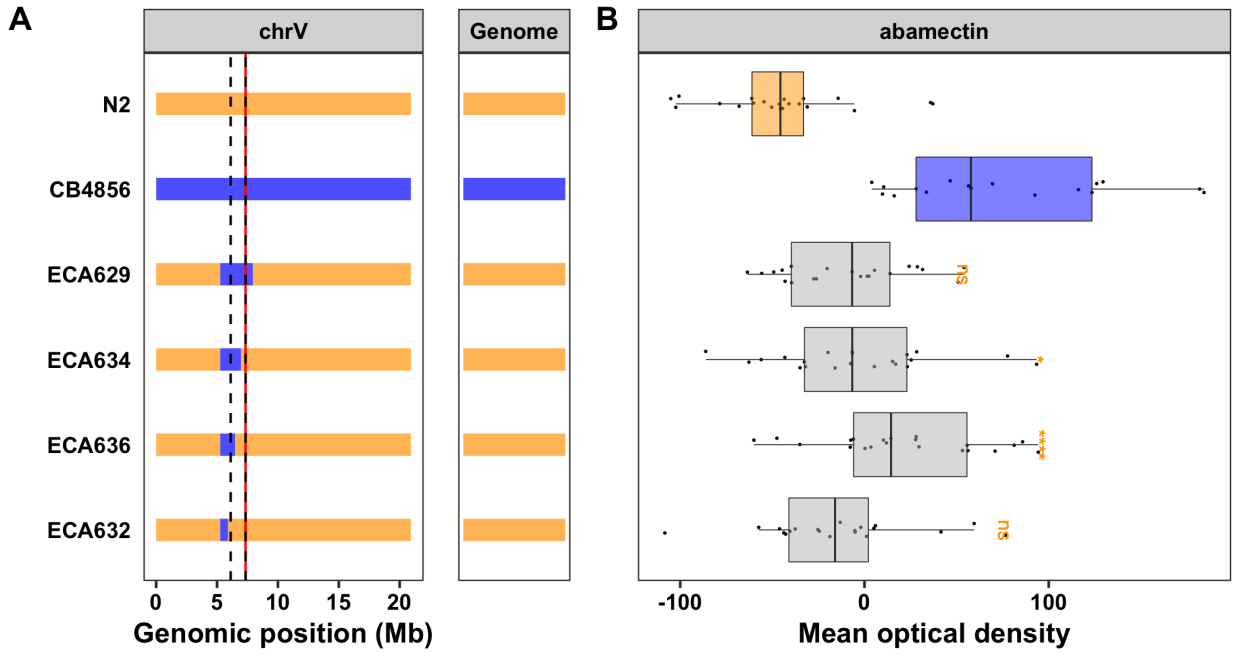


Figure 6-10: NILs validate and narrow the VC QTL. (A) Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (right). The solid vertical lines represent the peak marker of each QTL. **(B)** Relative mean optical densities in abamectin (mean.EXT, x-axis) are plotted as Tukey box plots against strain (y-axis). Statistical significance of each NIL as compared to N2 is shown above each strain (ns = non-significant (p-value > 0.05); *, **, ***, and *** = significant (p-value < 0.05, 0.01, 0.001, or 0.0001, respectively).

mutants to abamectin and measured animal length, optical density, and brood size. Both *lgc-54* mutants were significantly more resistant than the N2 strain (p-values < 0.0004), which would provide evidence for the role of *lgc-54* in abamectin resistance (**Figure 6-11A**). However, we noticed that these mutants grew much slower in the control conditions than both the N2 and CB4856 strains (**Figure 6-11B**). This growth defect makes it difficult to compare abamectin sensitivities, as the observed resistance could be an artifact of the statistical regression analysis.

More recently, a large-effect QTL on chromosome V (37-42 Mb) was identified in response to ivermectin treatment in *H. contortus* [301]. The authors did not provide candidate genes, but stated that no previously identified candidate genes for ivermectin resistance were found within this region. The *H. contortus* ortholog of *glc-3* is also found on chromosome V but left of the defined QTL region (27.6 Mb). Interestingly, they also identified a smaller-effect QTL nearby (V:45-48 Mb). It is possible that these two QTL are comparable to the two QTL we identified (VL and VC). To test this hypothesis, we first compared the synteny of genes in these regions between the two species. We identified the

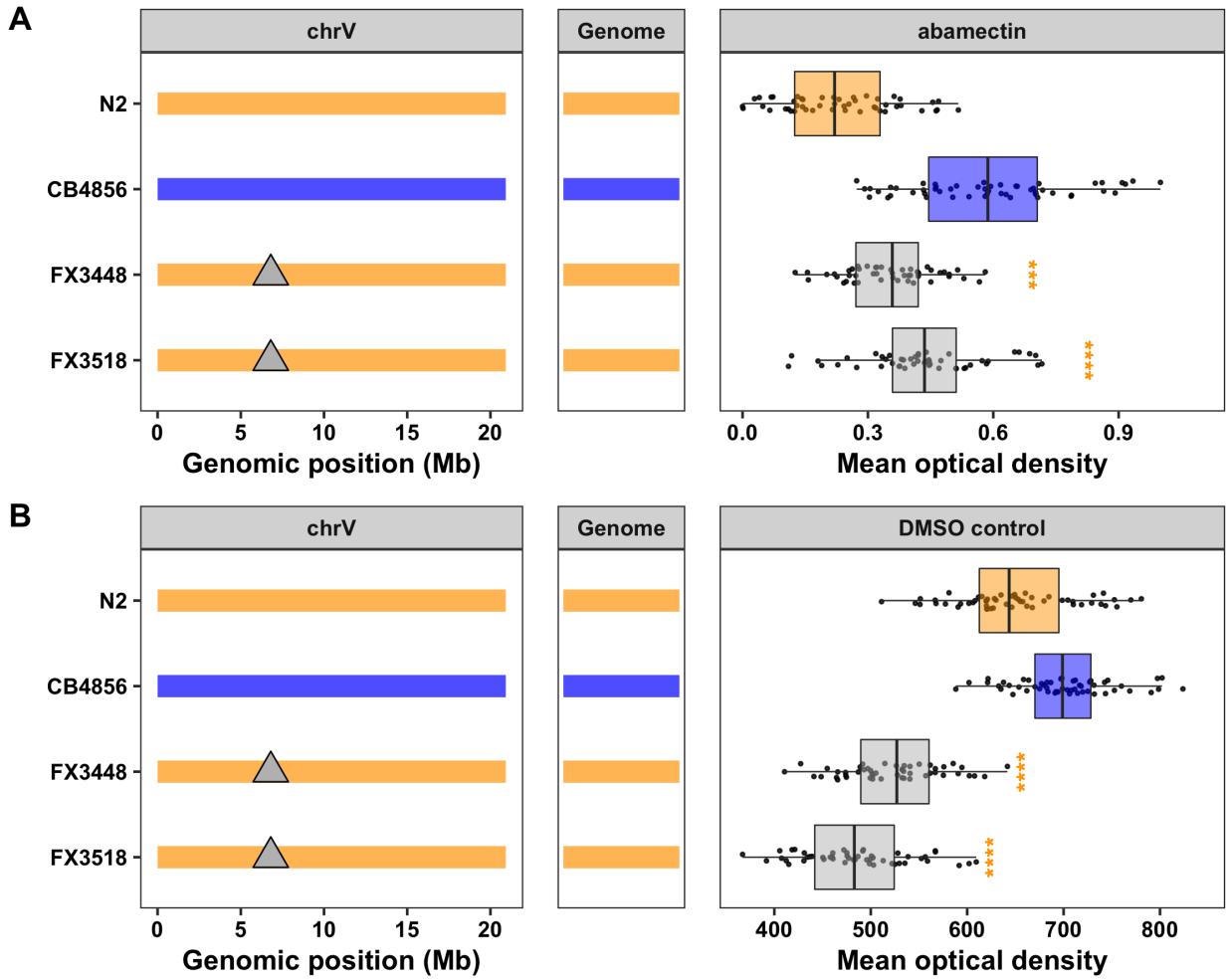


Figure 6-11: Testing the role of *lgc-54* in the *C. elegans* abamectin response. Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (center). Grey triangles represent mutations in the *lgc-54* gene. On the right, relative mean optical density in abamectin (mean.EXT, x-axis) **(A)** or mean optical density in DMSO control **(B)** is plotted as Tukey box plots against strain (y-axis). Statistical significance of each deletion strain compared to N2 is shown above each strain (ns = non-significant (p-value > 0.05), **** = significant (p-value < 0.0001).

384 genes in the large-effect QTL for *H. contortus* and identified the *C. elegans* orthologs for each gene using a best-match BLASTp [307] search. We identified a total of 347 *H. contortus* genes with 340 *C. elegans* orthologs. Of these *C. elegans* genes, 115 also reside on chromosome V and eight reside within our narrowed VC QTL interval. Two of these genes, *pgph-1* and *pgph-2*, are homologs of phosphoglycolate phosphatase (PGP). Several PGP genes have been previously implicated in avermectin resistance including, but not limited to, *pgp-1* and *pgp-2* [301, 141]. This analysis indicates

that it is likely that these QTL from differing species overlap, suggesting the potential for shared mechanisms of drug resistance between *C. elegans* and parasitic nematodes.

6.5.4 Modified high-throughput assay does not facilitate candidate gene testing

In the summer of 2018, the bacterial lysate used in all our high-throughput experiments expired and could no longer provide satisfactory nutrition for 96 hours of growth (see Chapter 5). To circumvent this problem, we developed a modified assay where animals were grown for 48 hours, from L1 to L4, in the presence of abamectin and then analyzed for the developmental traits animal length (mean.TOF) and optical density (mean.EXT) [69]. We treated the N2 and CB4856 strains to increasing concentrations of abamectin while feeding either HB101 lysate or HB101 live bacteria (**Figure 6-12**). We observed a dose-dependent response to abamectin under both conditions. However, we did not observe significant phenotypic variation between the N2 and CB4856 strains. These results indicate that this modified assay would not be useful in testing candidate genes underlying the abamectin response.

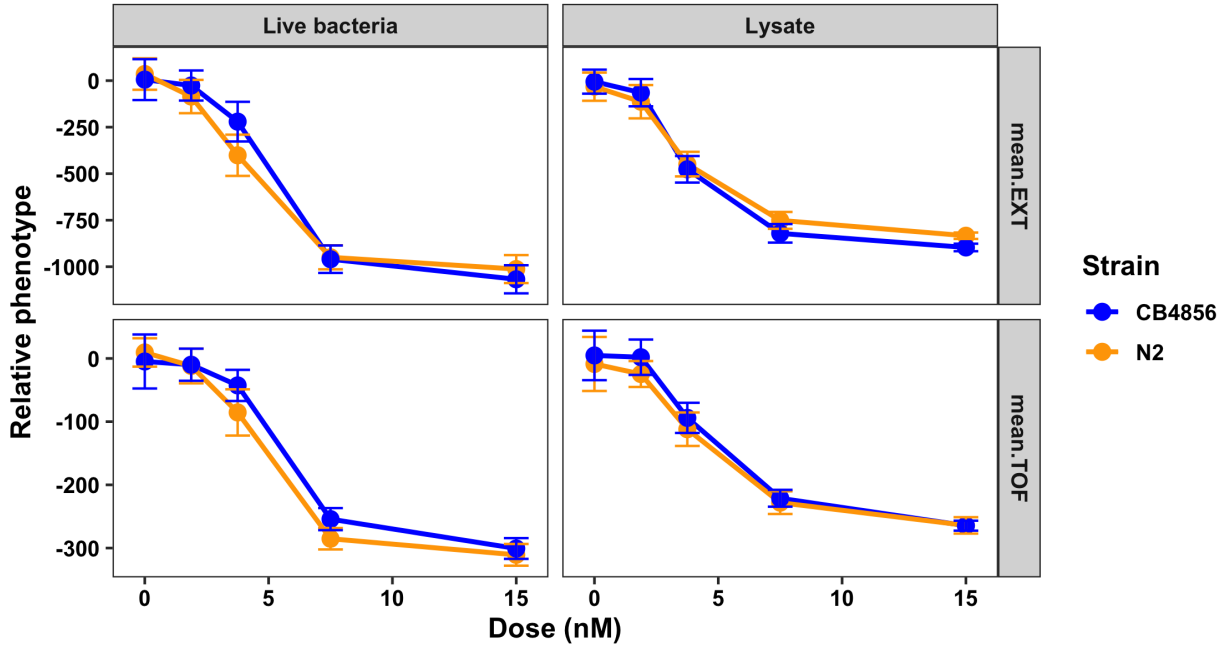


Figure 6-12: Modified high-throughput assay does not capture phenotypic differences between N2 and CB4856 strains. Results from the abamectin dose response modified HTA for animal length (mean.TOF) and animal optical density (mean.EXT) in both lysate and live bacteria. For each trait, drug concentration (nM) (x-axis) is plotted against the phenotype subtracted from control (y-axis) and colored by strain (CB4856: blue, N2: orange).

6.6 DISCUSSION

In this study, we used linkage mapping and genome-wide association analysis to identify three QTL on chromosome V that influence responses to the anthelmintic abamectin in *C. elegans*. One of these QTL overlaps with the previously identified GluCl gene *glc-1* [120, 99]. However, the remaining two QTL are novel and might overlap with ivermectin QTL from parasitic nematodes, suggesting validity of using *C. elegans* to discover anthelmintic drug resistance genes. We used NILs to validate and narrow each QTL independently. Although we were unable to discover the specific causal genes or variants, we suggest the GluCl channel *glc-3* within our narrowed genomic interval could play a role in abamectin resistance.

6.6.1 Linkage and association mapping identify similar, yet different, QTL

Although the QTL on the left of chromosome V overlapped between the two methods, QTL on the center and right of chromosome V were separated by up to 1 Mb (**Figure 6-4**). These results could suggest that there is a different, rare variant that confers resistance in the CB4856 strain compared to the common variant in the mapping population. However, the *glc-1* variant is common and has been discovered previously from association mapping with wild isolates [120] and both linkage mapping (**Figure 6-2B**, [120]) and bulk segregant analysis [99] with recombinants derived from the N2 and CB4856 strains. Because this GluCl gene has been identified across different traits, populations, and mapping methods, it suggests that the VR locus in our mapping might also be driven by variation in this gene. In the fine mapping of all variants across chromosome V, we observe a sharper, yet less significant, peak at the *glc-1* locus (**Figure 6-2C**). Furthermore, both the *glc-1* locus at 16.1 Mb and the unidentified locus at 17.7 Mb are detected in the association mapping for brood size (**Figure S6-1**, **Figure 6-5**), suggesting they might be linked. The particular causal variant, the four-amino-acid deletion in *glc-1*, is also not one of the genetic markers used in our study, as we mostly focused on single nucleotide variants (SNVs). Therefore, it is likely that SNVs linked to the *glc-1* allele are causing the observed association. In addition, this region on the right of chromosome V overlaps with several known divergent regions in both the CB4856 strain and other strains [277]. The high density of SNVs in divergent regions might

increase the possibility of detecting a QTL and artificially narrow the confidence interval (arbitrarily defined as 150 SNVs on either side of the marker with the strongest association).

A similar issue is also seen with the QTL on the center of chromosome V. Association and linkage mapping identify different regions and the fine mapping shows a stronger, yet less significant, peak in the region defined from linkage mapping (**Figure 6-2C**). Because no genes in this combined region have been previously implicated in abamectin resistance, we generated NILs to cover both the regions identified from association and linkage mapping. Surprisingly, we found that in this case the NILs supported the region identified from association mapping, not linkage (**Figure 6-9**). Because we used NILs generated from the N2 and CB4856 parents, this result suggests that the two QTL identified are likely the same locus. The VC QTL in linkage mapping is the first QTL to be identified from a forward search algorithm and has a large, broad peak. It is possible that the confidence interval was miscalculated due to combined effects from the flanking QTL and low recombination frequency, and thus low resolution, in the center of chromosomes. However, it is also possible that association mapping identified a causal variant that is commonly found in the population, including in the CB4856 strain, and linkage mapping identified a causal variant that is unique to CB4856. Regardless, this study emphasizes that, although powerful, QTL mapping is ultimately a statistical method that can be influenced by population structure, among other things, and that it is essential to validate QTL before drawing conclusions about the causal variant.

6.6.2 Power of QTL mapping in *C. elegans* to identify causal genes underlying anthelmintic resistance

This study highlights the benefits of mutual communication between the parasite and *C. elegans* communities. Genetic mappings, screens, and selections are more easily performed in the free-living nematode to help discover the drug targets and mechanisms of action. However, it is important that these findings are then translated back to parasitic nematodes to confirm that these genes found in *C. elegans* are responsible for drug resistance in parasites both in the lab and in the field. The suggested overlap between QTL for avermectin resistance in *C. elegans* and *H. contortus* [301] strengthens the validity of our approach and indicates that the causal variants in our mapping population might also

confer resistance in *H. contortus* and perhaps other parasitic nematode species. Future studies to discover the causal genes and variants underlying our two novel QTL (VL and VC) could be informative to parasitologists and help treat infected individuals smarter and more effectively.

In addition to identifying genes in *C. elegans* that might cause resistance to anthelmintics in parasites, hypotheses generated about candidate genes and variants that confer resistance in parasitic nematode species can be more easily validated in the experimentally tractable *C. elegans* model. Here, we tested the role of *lgc-54*, a gene predicted to be involved in ivermectin resistance in *T. circumcincta*. Unfortunately, the results of this assay suggest that loss of function of *lgc-54* does not cause increased resistance to abamectin in *C. elegans*. However, the decreased fitness of these mutants make the results more difficult to interpret. Targeted deletions or allele editing of *lgc-54* using CRISPR-Cas9 genome editing might provide more evidence for the role of *lgc-54* in abamectin resistance. However, because *lgc-54* is not within our NIL-defined genomic interval, it is not likely to be the causal gene underlying our mapping results, regardless of its role in avermectin resistance. Regardless, it is possible that *lgc-54* might cause resistance to ivermectin in parasitic nematodes even if it does not function in the same way in *C. elegans*. Although predicting candidate genes based on the gene function can be extremely useful, this study demonstrates the power of functional validation in model systems like *C. elegans* to experimentally prove or disprove hypotheses.

6.6.3 Shared niches provide the same selective pressures for soil transmitted helminths and *C. elegans*

The potential overlap of QTL for avermectin resistance between *C. elegans* and parasitic nematodes suggests that the loci that confer resistance to avermectins are conserved across several nematode species. Among parasitologists, it is believed that parasitic nematodes gain anthelmintic resistance due to standing genetic variation in a population (or novel mutations) in combination with exposure to selective pressures in their environment [309, 310]. Soil transmitted helminths such as *H. contortus*, spend part of their life cycle in soil or rotting vegetation–an environment that overlaps with the niche associated with the free-living *C. elegans* [311, 59]. Selective pressures in this environment could originate from natural toxic compounds produced by soil-dwelling bacteria and fungi from which many

anthelmintic drugs are derived [312, 313, 314]. Additionally, synthetic anthelmintic compounds are a common soil and water pollutant in some areas and can be found in runoff from farms that use anthelmintics to treat agriculture or livestock [315, 316]. This exposure to the same selective pressures and the known genetic diversity in the *C. elegans* species suggests a method for how the free-living nematode might evolve the same resistance alleles as other parasitic nematodes. In one example, we showed that recent selective pressures have likely acted on the *C. elegans ben-1* locus, resulting in an excess of putative loss-of-function alleles across the population despite *ben-1* being an evolutionarily constrained locus [75]. This conclusion again highlights the relevance of using the experimentally tractable *C. elegans* as a model to study anthelmintic resistance in parasitic nematode species.

6.7 FUTURE DIRECTIONS

We suggested the GluCl channel *glc-3* as a candidate gene for the VC QTL. This hypothesis could be explicitly tested by first deleting *glc-3* in both the N2 and CB4856 strains using CRISPR-Cas9 genome editing. If a loss of function of *glc-3* causes resistance to abamectin in the CB4856 strain, we would expect the deletion in the CB4856 strain to have no change in its resistance phenotype, but the deletion in the N2 strain would cause an increased resistance to abamectin compared to the N2 strain. Alternatively, CRISPR could be used to edit the single nucleotide variant to change the amino acid in the 439th position from an isoleucine to phenylalanine in the N2 strain and from phenylalanine to isoleucine in the CB4856 strain. If this variant (I439F) is responsible for increased resistance to abamectin in the CB4856 strain, we would expect that the strains with the isoleucine (N2 and CB4856[F439I]) would be sensitive to abamectin and the strains with the phenylalanine (CB4856 and N2[I439F]) would be resistant to abamectin. If the deletion has an effect but the SNV doesn't, reciprocal hemizygosity tests could be used to determine if natural variation in *glc-3* between the N2 and CB4856 strains underlies this QTL.

In addition to testing the candidate gene *glc-3*, the *lgc-54* mutants should be re-tested to verify the growth defect seen in control conditions. It is possible that these strains are sick because of other genetic mutations that exist outside of *lgc-54* since they are mutants derived from EMS mutagenesis. If this growth defect is still present, targeted deletions of *lgc-54* using CRISPR could remove concern

about background mutations. Alternatively, editing a single variant instead of deleting the entire gene might allow the worms to grow normally while still testing the function of the variant in abamectin response.

The VL QTL has a strong peak that overlaps between linkage and association mapping and has not been previously identified in other studies. To begin narrowing this QTL, we have generated 17 genetically distinct strains derived from ECA1065 containing increasingly smaller introgressions that tile across this interval (**Figure 6-13**). These strains should be tested for their response to abamectin and these phenotypes can then be used to narrow the genomic interval similarly to the VC QTL. Of the 160 genes inside the confidence interval defined by linkage mapping, there are no known GluCl channels nor other ligand-gated or ion channels. This result suggests that the resistant-causing locus on the left of chromosome V might not be related to the drug target, but perhaps is associated with drug import, export, or metabolism. In particular, three cytochrome P450 genes have expression variation between the N2 and CB4856 strains that is linked to this region on chromosome V.

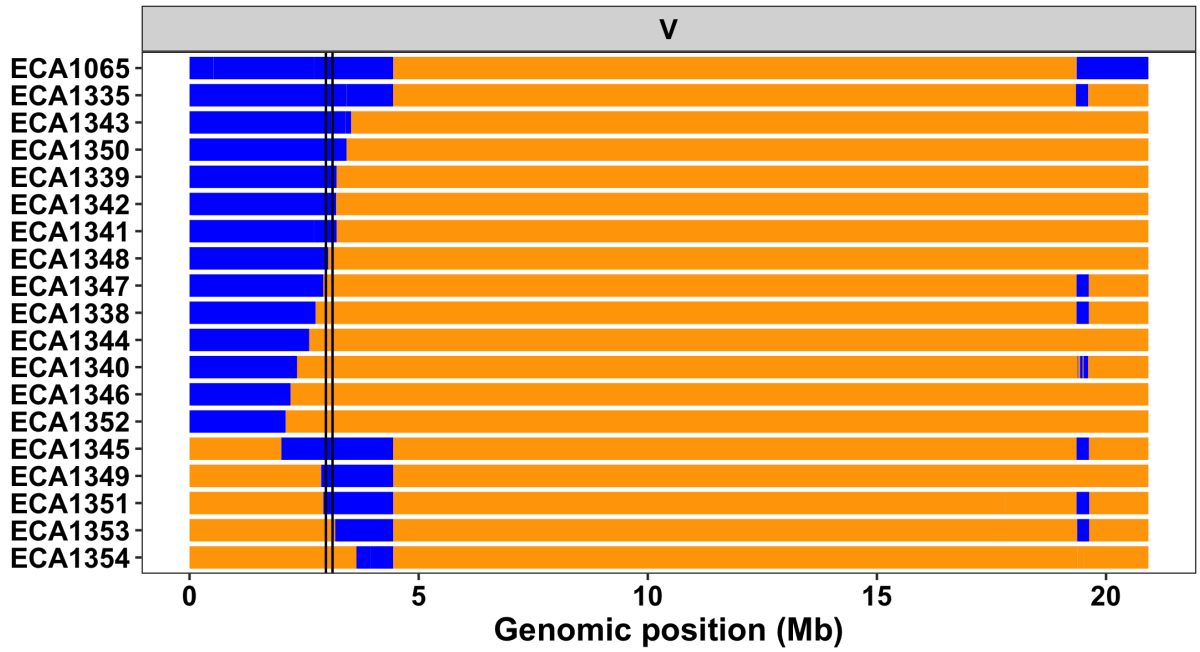


Figure 6-13: NILs to isolate and narrow the VL QTL. Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V. All strains are in the N2 genetic background. The solid vertical lines represent the previous NIL-defined interval.

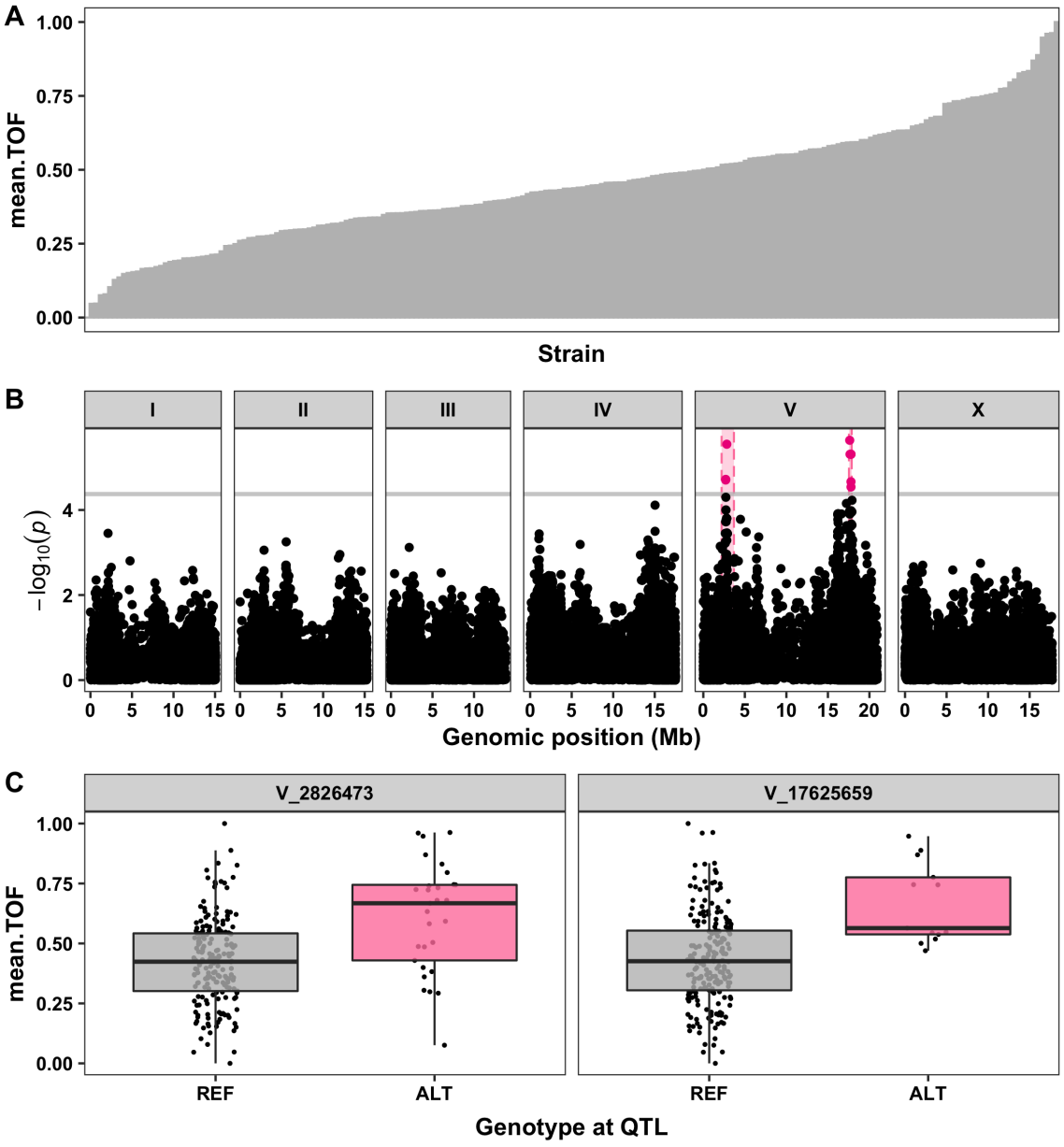
One major caveat to exploring the causality of the aforementioned candidate genes is that we

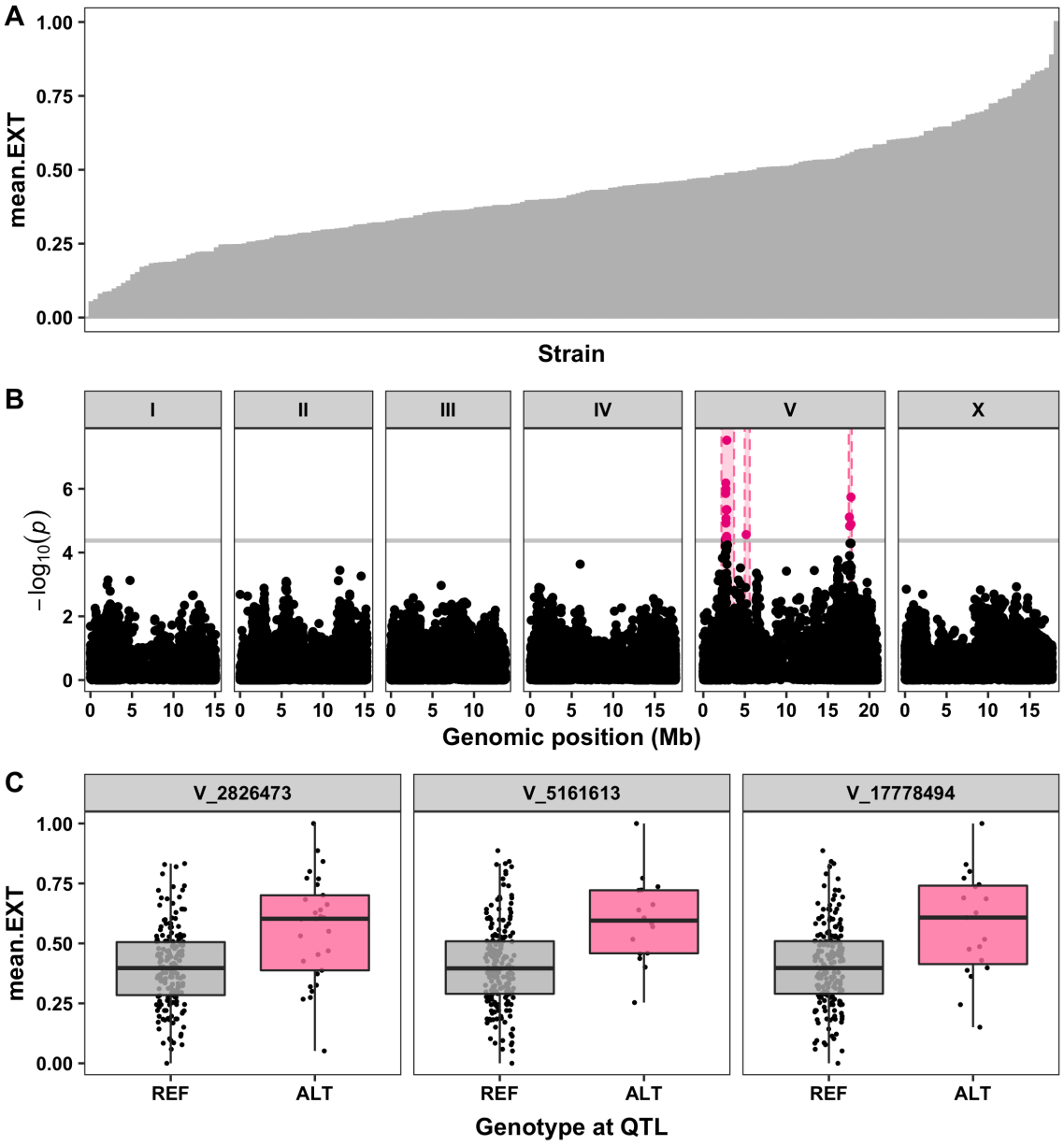
no longer have a reliable assay that can reproduce the QTL effects. This must be the first problem solved to continue studying abamectin resistance in *C. elegans*. Using the modified HTA, live HB101 bacteria, and low concentrations of abamectin, we were able to observe a difference in abamectin response between the N2 and CB4856 parents with the CB4856 strain conferring resistance (**Figure 6-12**). Although the difference is small, this is a promising start, if reproducible. The next step would be to test a finer-scale range of abamectin concentrations with NILs to see if the effect of genetic variation on chromosome V can be captured with this assay. Alternatively, it is possible that 48 hours of growth from L1 to L4 is not enough to see a difference in abamectin responses between the N2 and CB4856 strains. The original high-throughput assay could be performed with live HB101 bacteria that promotes nematode growth over 96 hours to see if the resistant phenotype can be rescued. However, the live HB101 bacteria used here has been optimized for 48 hours of nematode growth, so it would be necessary to first optimize a new batch of bacteria that can reproduce the effect of the lysate over 96 hours of growth.

6.8 CONTRIBUTIONS

During this work, I was funded by the Cell and Molecular Biology of Disease training program and the NSF-Simons Center for Quantitative Biology at Northwestern University. I would also like to thank the National BioResource Project (Japan) - *C. elegans* for *lgc-54* mutant strains (FX3448 and FX3518). Many people have contributed to this project throughout the last several years. Dr. Mostafa Zamanian began this work making NILs with Grace Park. Ellen Chao and Katie Introcaso have also contributed NILs for this project. Dr. Steffen Hahnel and Briana Rodriguez performed the GWA experiment. Briana Rodriguez, Dr. Shannon Brady, Grace Park, and Ellen Chao haved worked under my guidance to help perform sorter assays to test a variety of NILs. I performed all of the data analysis. I would further like to thank Dr. Janneke Wit and Katie Introcaso for their help trying to further narrow the QTL using the new V3 sorter assay and, of course, Dr. Erik Andersen for his mentorship and guidance throughout the project.

6.9 SUPPLEMENTAL FIGURES




A
mean.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
$-\log_{10}(p)$
0
2
4
6
0 5 10 15
0 5 10 15
0 5 10
0 5 10 15
0 5 10 15 20
0 5 10 15
Genomic position (Mb)
C
V_2826473
V_5161613
V_17778494
mean.EXT
1.00
0.75
0.50
0.25
0.00
REF
ALT
REF
ALT
REF
ALT
Genotype at QTL

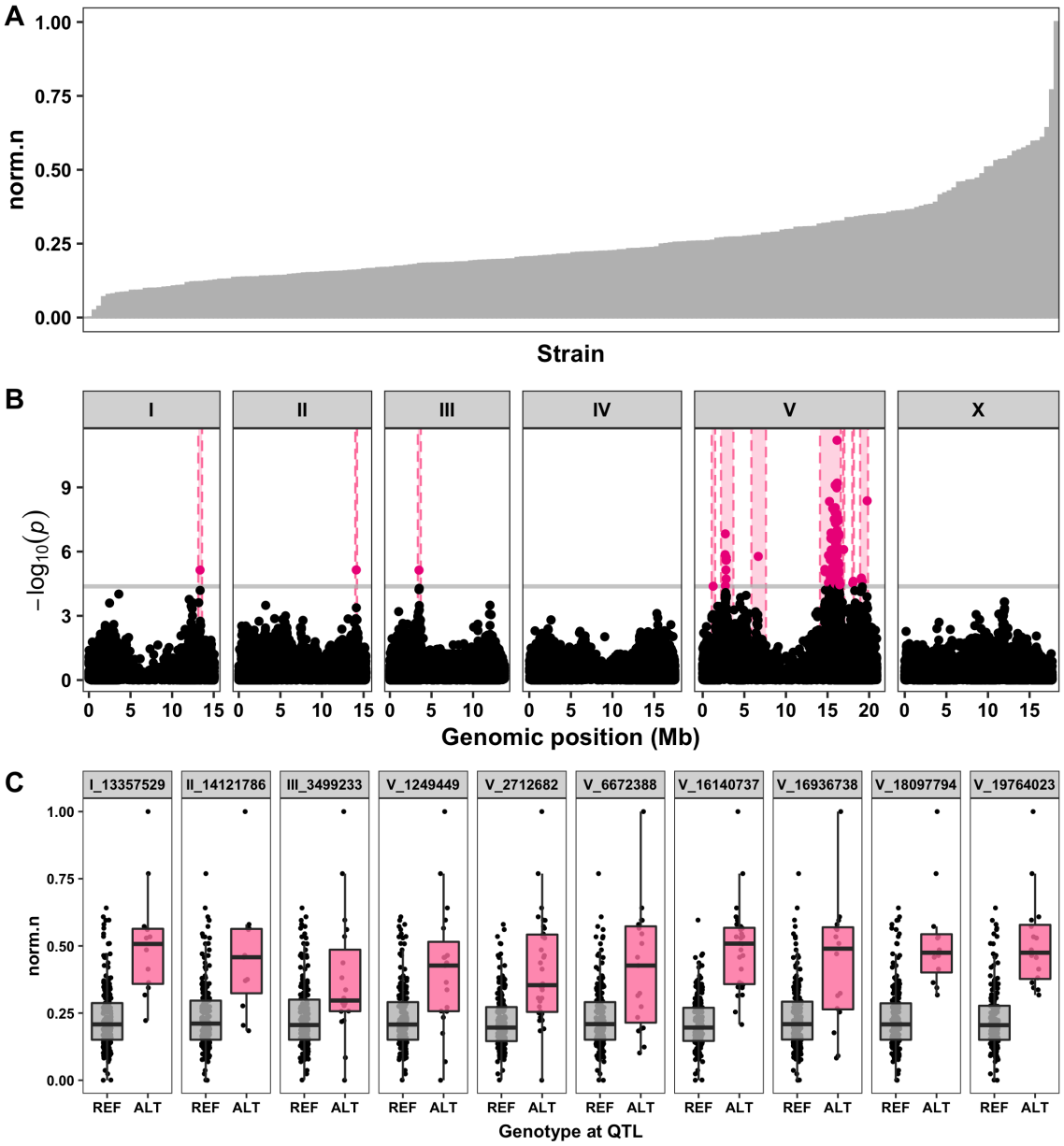
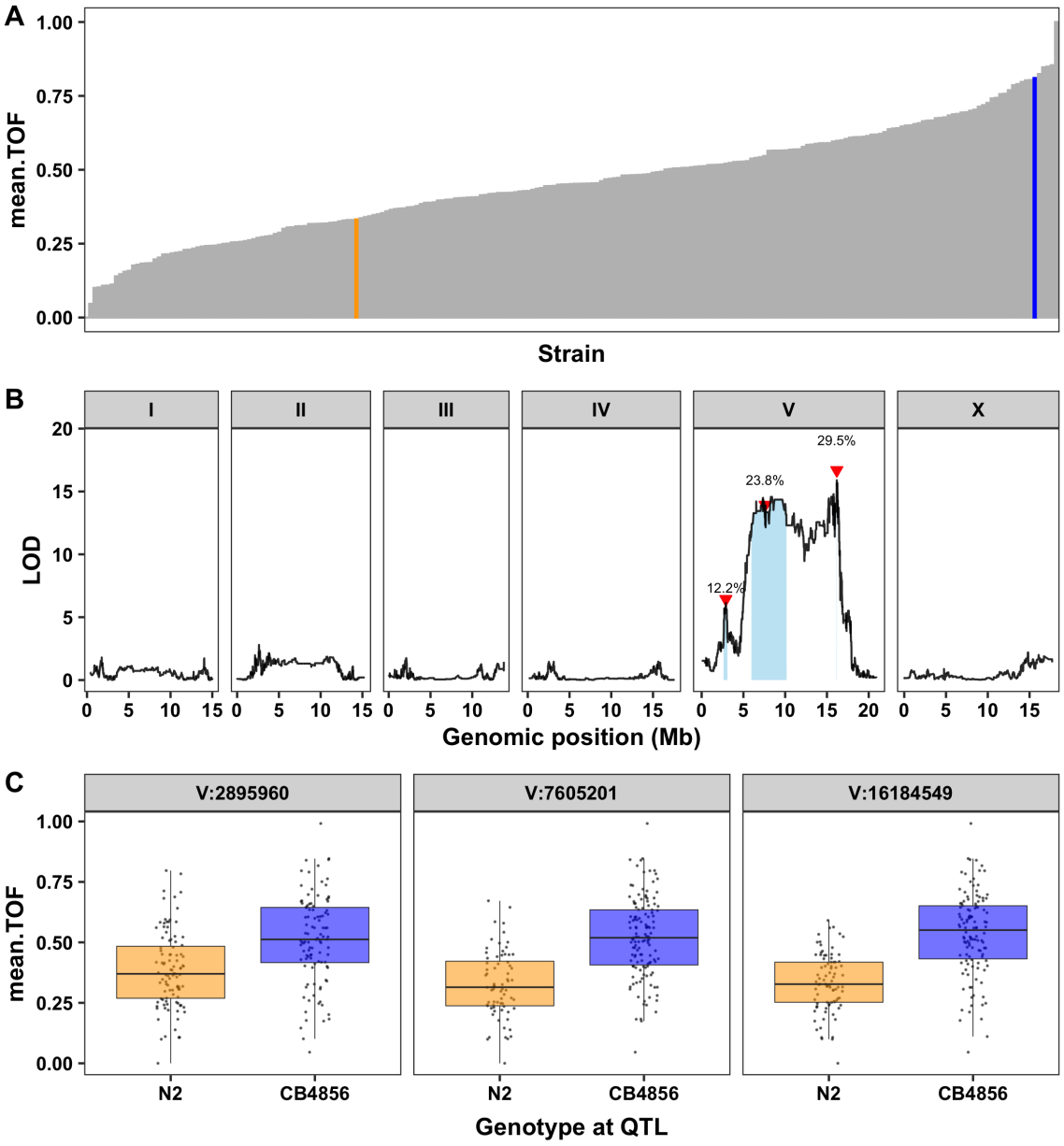
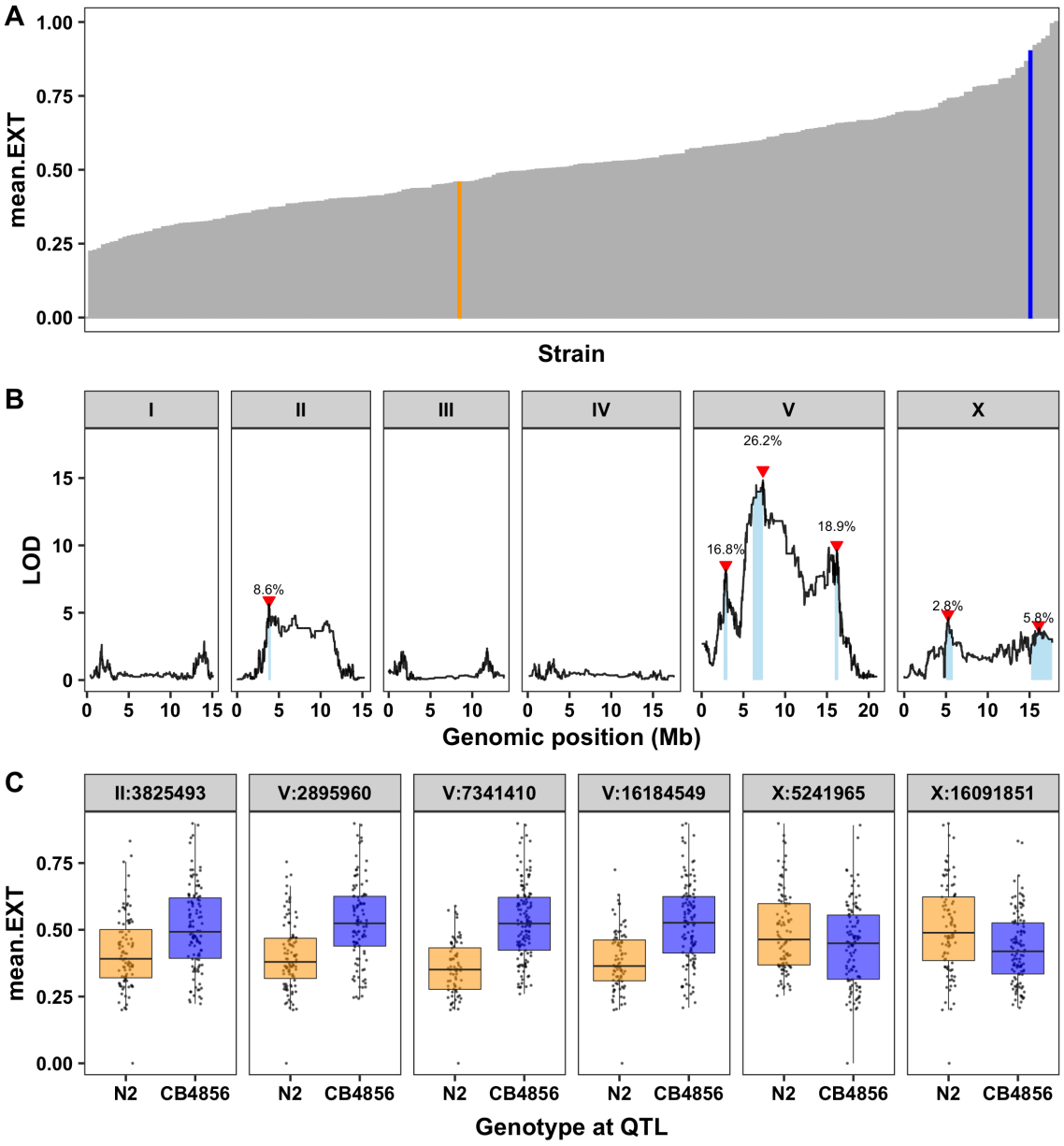


Figure S6-1: Genome-wide association (GWA) mapping identifies 15 QTL across three traits in response to abamectin. **(A)** Normalized residual phenotype (y-axis) of 210 wild isolates (x-axis) in response to abamectin. Strains are colored by the parental strains N2 (orange) and CB4856 (blue). **(B)** GWA results are shown. Genomic position (x-axis) is plotted against the -log10(*p*) value (y-axis) for each SNV. SNVs are colored pink if they pass the genome-wide eigen-decomposition significance threshold designated by the grey line. The genomic regions of interest that pass the significance threshold are highlighted by pink rectangles. **(C)** For each QTL, the normalized residual phenotype (y-axis) of strains split by genotype at the peak marker (x-axis) are plotted as Tukey box plots. Each point corresponds to a wild isolate strain. Strains with the N2 reference allele are colored grey, and strains with an alternative allele are colored pink.


A
mean.TOF
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
LOD
20
15
10
5
0
29.5%
23.8%
12.2%
0 5 10 15 20
Genomic position (Mb)
C
V:2895960
V:7605201
V:16184549
mean.TOF
1.00
0.75
0.50
0.25
0.00
N2
CB4856
Genotype at QTL


A
mean.EXT
1.00
0.75
0.50
0.25
0.00
Strain
B
I
II
III
IV
V
X
LOD
26.2%
18.9%
16.8%
8.6%
2.8%
5.8%
Genomic position (Mb)
C
II:3825493
V:2895960
V:7341410
V:16184549
X:5241965
X:16091851
mean.EXT
N2
CB4856
Genotype at QTL

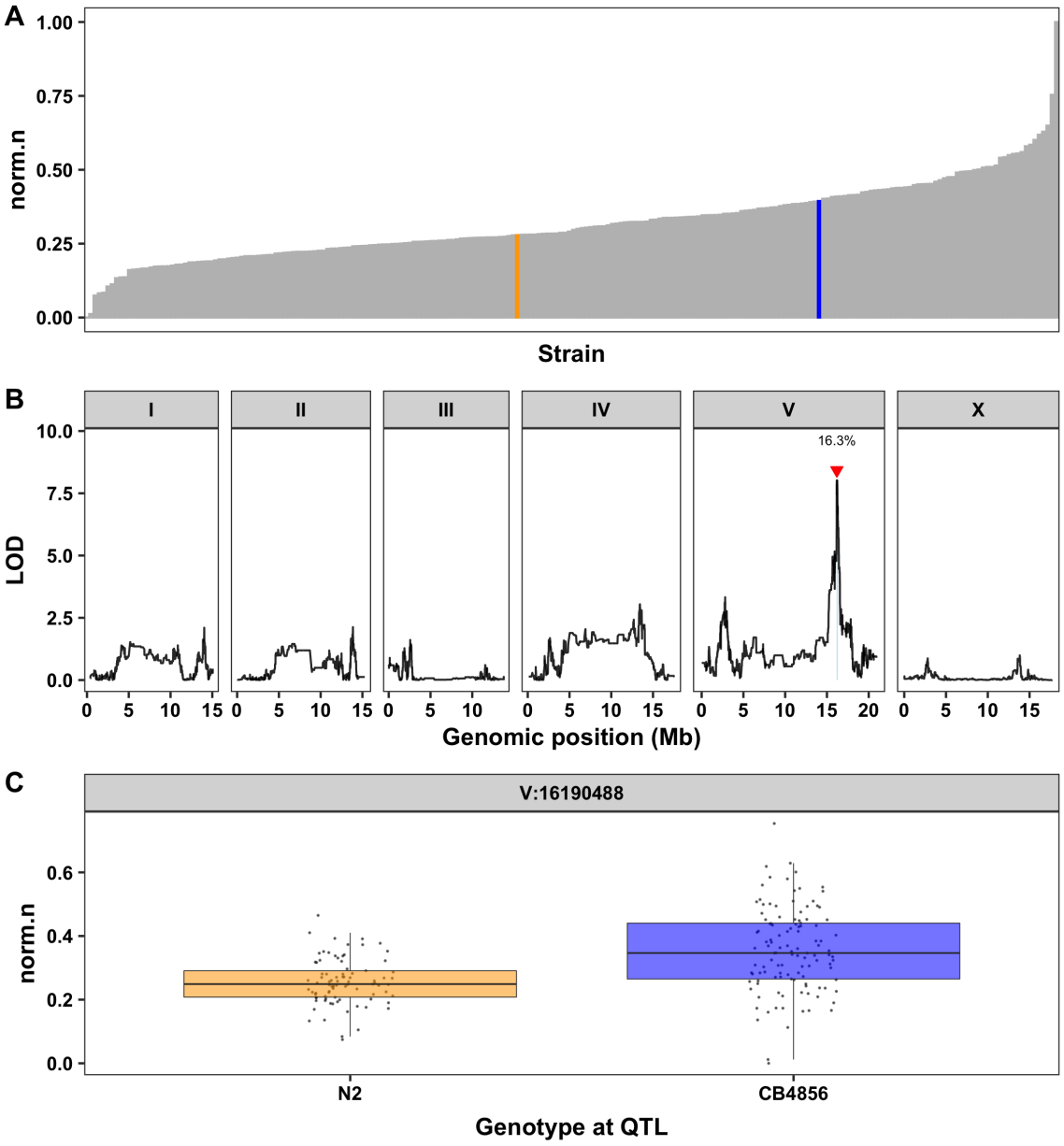


Figure S6-2: Linkage mapping identifies 10 QTL across three traits in response to abamectin. **(A)** Normalized residual phenotype (y-axis) of 296 RIAILs (x-axis) in response to abamectin. The parental strains are colored: N2, orange; CB4856, blue. **(B)** Linkage mapping results are shown. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(C)** For each QTL, the normalized residual phenotype (y-axis) of RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to a unique recombinant strain. Strains with the N2 allele are colored orange, and strains with the CB4856 allele are colored blue.

7 Correlation of genotype with climate parameters suggest *Caenorhabditis elegans* niche adaptation

7.1 PREFACE

This project was my rotation project from the Fall of 2015 and the first project I took on as a graduate student. In Erik's previous paper from 2012 [79], he identified a QTL on the left arm of chromosome II using strain latitude from 97 wild isolates as a phenotype. In this manuscript, Erik notes that this association could reflect population structure or it might implicate this region of the genome in an unknown ecological niche, such as temperature. Erik pitched this idea to me in my rotation to use the GPS coordinates and date of isolation from our wild isolate collection, now 152 isotypes, to extract detailed weather and climate parameters to see if any associations were observed between genetic variation and environment. This was a great opportunity for me as a first-year graduate student to learn to code (and work with extremely messy datasets from the internet...), learn to write a manuscript, and discover an interest in the power of genome-wide association studies. This chapter is based on my first first-author manuscript published in *G3* in 2017 that was the result of my successful rotation project [136].

7.2 ABSTRACT

Species inhabit a variety of environmental niches, and the adaptation to a particular niche is often controlled by genetic factors, including gene-by-environment interactions. The genes that vary in order to regulate the ability to colonize a niche are often difficult to identify, especially in the context of complex ecological systems and in experimentally uncontrolled natural environments. Quantitative genetic approaches provide an opportunity to investigate correlations between genetic factors and environmental parameters that might define a niche. Previously, we have shown how a collection of 208 whole-genome sequenced wild *Caenorhabditis elegans* can facilitate association mapping approaches. To correlate climate parameters with the variation found in this collection of wild strains, we used geographic data to exhaustively curate daily weather measurements in short-term (three

month), middle-term (one year), and long-term (three year) durations surrounding the date of strain isolation. These climate parameters were used as quantitative traits in association mapping approaches where we identified 11 QTL for three climatic variables: elevation, relative humidity, and average temperature. We then narrowed the genomic interval of interest to identify gene candidates with variants potentially underlying phenotypic differences. Additionally, we performed two-strain competition assays at high and low temperatures to validate a QTL that could underlie adaptation to temperature and found suggestive evidence supporting that hypothesis.

7.3 INTRODUCTION

Ecological niches describe how individuals of a species respond to and alter the distribution of resources and competitors within their environment [317]. These resources could include food availability, soil type, short-term weather conditions, and long-term climate. Often, a species can be found in multiple distinct geographic areas that all share a common set of environmental factors and resources. For example, a plant that thrives at high temperature might grow equally well anywhere along the equator. An organism's ability to survive in a specific niche is driven by both environmental and genetic factors. Genetic variation between species, and among individuals within a species, contributes to the wide variety of niches observed [318]. A genetic variant could result in an increased affinity for an individual to its environment. This individual will be selected and, over time, evolution will favor the successful variant. This phenomenon, known as gene-by-environment interactions, refers to phenotypes in which different genotypes respond to environmental variation in diverse ways.

Previous studies in model organisms, particularly *Drosophila* and *Arabidopsis*, have investigated gene-by-environment interactions with clinal variation. In *Drosophila*, selection on body size is correlated with temperature [319], and survival is affected by climate change [320]. Machado *et al.* performed a longitudinal study of *Drosophila* collected at differing latitudes during a two-year time span and compared physiological traits of two different species – *D. melanogaster* and *D. simulans* [321]. Other studies have gone further by identifying quantitative trait loci (QTL) for body size and cold tolerance traits involved in adaptation to seasonally varying environments [322, 323]. Gerken *et al.* found substantial heritable variation in both short-term and long-term acclimation [324]. They then

performed genome-wide association (GWA) mappings on these traits and found the QTL for short-term and long-term adaptations did not overlap, but each resulted in a set of gene candidates sharing similar functions in apoptosis, autophagy, cytoskeletal and membrane structural components, or ion binding and transport. Additionally, GWA studies have been performed in *Arabidopsis* for various adaptive traits such as seed oil melting point [325]. Another *A. thaliana* study performed QTL mapping using recombinant inbred lines produced from two strains isolated from different climates and found that relatively few QTL explain much of the adaptive divergence between them [326]. Furthermore, Fournier-Level *et al.* provide evidence for broad scale local adaptation in *A. thaliana* by using a GWA mapping approach that combines fitness traits measured in multiple natural environments and geographic and climatic analyses [327].

C. elegans is a free-living nematode often found in microorganism-rich organic material such as rotting fruits and compost heaps in temperate and humid environments [311, 328]. The first studied *C. elegans* strain, N2, was isolated from mushroom farm compost in Bristol, England in 1951 [329]. Since that time, N2 has been used as the wild-type strain for *C. elegans* laboratory research. This strain was cultured for many years in the laboratory, potentially resulting in selection for alleles favorable in that environment [58]. To study natural variation and the role of niche specification on this species, we require a worldwide collection of wild strains. Our research group acquired a large collection of 208 wild strains and sequenced the whole genomes of these strains [63, 60, 134]. By comparing the genomes of the 208 strains, we found that some strains from similar geographic locations have nearly identical genome sequences. This analysis resulted in 152 unique genome-wide haplotypes or isotypes. This large pool of genetic information provides us with the statistical power to make connections between genotype and phenotype using GWA studies.

A handful of studies have addressed differences in temperature sensitivities across *C. elegans* strains, and many of these studies show that temperature affects the lifetime fecundity and reproductive timing of *C. elegans*. Two separate groups used a small subset of wild isolates to assay thermal tolerance [330] and thermal sensitivities for fitness traits [331]. Other studies performed linkage mappings using a recombinant inbred line panel of *C. elegans* strains to map life history traits such as fertility, growth rate, and egg size at both low and high temperatures to various locations

across the genome [103, 102]. However, the use of population-scale *C. elegans* data to map differences found in a large number of wild isolates for natural environmental conditions has yet to be addressed. In this study, we correlate natural genetic variation among 152 wild *C. elegans* strains with climate measurements of their environmental niches as quantitative traits. We mapped traits that describe the niche of the isolation location for each strain, including geographic parameters, seasonal weather patterns, and climate variables. We find significant associations for elevation, relative humidity, and temperature. These findings suggest genetic control of niche specification. Additionally, we tested the QTL associated with temperature and found possible evidence of adaptation to lower temperatures based on genetic background.

7.4 METHODS

7.4.1 *C. elegans* wild isolate collection and sequencing

A collection of 208 wild *C. elegans* strains have been previously isolated worldwide and annotated for each strain's geographic location and date of isolation [63]. Members of the Andersen Lab have carefully and manually curated this dataset to offer the most accurate information possible while accounting for sometimes imprecise sample recording. Whole-genome sequence (WGS) data were collected from all 208 strains [63]. The raw Illumina data are deposited with the Short Read Archive under project PRJNA318647. WGS data were analyzed as previously described [63]. In brief, after alignment with BWA [332] and variant calling using Samtools [333], strains with a concordance of 99.93% or higher were grouped as a genome-wide haplotype or isotype. This analysis resulted in 152 unique isotypes.

7.4.2 Weather and climate data acquisition

For each wild strain with a known isolation location, elevation was estimated with the *geosphere* package in R [334] using the geographic coordinates of strain isolation. A correlation test using 74 points of known elevation were used to verify accuracy of the elevation function resulting in a correlation of 0.998. Weather data were downloaded from the Integrated Surface Data (ISD) FTP

server (ftp://ftp.ncdc.noaa.gov/pub/data/noaa/) managed by the National Oceanic and Atmospheric Administration (NOAA) and the National Climatic Data Center (NCDC). The ISD dataset comprises worldwide surface weather observations from over 27,447 stations managed by the Automated Weather Network (AWN), the Global Telecommunications System (GTS), the Automated Surface Observing System (ASOS), and others. Data are collected once every three hours for some stations. Some parameters include air quality, atmospheric pressure, atmospheric temperature, dew point, atmospheric winds, clouds, precipitation, ocean waves, and tides.

Three distinct sets of weather station data were collected for analysis: a three-month window, a one-year window, and a three-year window. These data were filtered to include values centered around the date of isolation. For the three-month set, wild isolates with only a known month or year of isolation were not considered. Exact day of isolation is necessary to understand the seasonal environment in which an animal was isolated. For the one-year dataset, wild isolates with a month or day of isolation were used. For the three-year dataset, strains with only a year of isolation were used in addition to those strains with more defined dates of isolation. If only the year was known, the date of isolation defaulted to January 1 of that year, and data were collected surrounding that date. If only the month of isolation was known, the date of isolation was defaulted to the first of that month for data collection.

The 27,447 NOAA weather stations were filtered by their availability of data collected within the years of interest. Stations that had less than ten recordings of any type for any month within the time period of data collection were excluded to avoid misrepresentation by datasets that were averaged from only a few data points. Stations were then filtered by location, and the closest station to location of isolation for each wild strain was selected and downloaded using the *stationaRy* package available at https://github.com/rich-iannone/stationaRy [335]. We performed a rank-correlation test for temperature, relative humidity, and atmospheric pressure between two neighboring weather stations (ranging from 0.93-153 km apart) and found high correlations regardless of distance between stations ($rho = 0.920, 0.913$, and 0.707, respectively). All station-isotype pairs were included in our analysis regardless of the distance between them. The primary fields as well as all additional quantitative data available for each station were downloaded. Some fields (*e.g.* “AT1” or “Present-weather-observation”) were not downloaded from the NOAA database because the traits are qualitative and would not be

conducive to quantitative analyses. The station data were filtered to contain only information from the months surrounding the date of isolation. This process was repeated for each dataset (three-month, one-year, and three-year) in case a closer weather station contained only data for the three-month set but not for the one-year or three-year sets. The station data were meticulously checked by manually removing missing values from each weather category independently and, in certain cases, converting fields to uniform units that can be averaged to form a trait value. For example, precipitation ("AA1") was downloaded in two columns: 1) time period; 2) depth of precipitation recorded during that time period. Because the variable time periods in which data were recorded, averaging precipitation would lead to skewed results. Precipitation was changed to adapt a "precipitation per hour" model that would be more permissive to our analyses. The daily average of each trait was averaged over the time span collected (three months, one year, or three years), and this value was designated as the phenotypic value for that strain. Furthermore, the minimum and maximum daily values and total variance were averaged over the time span collected. Only traits with values in more than 90% of strains were analyzed further.

7.4.3 Association mapping

Genome-wide association (GWA) mapping was performed using 152 genome-wide *C. elegans* isotypes using the *cegwas* R package found at https://github.com/AndersenLab/cegwas. This package uses the EMMA algorithm for performing association mapping and correcting for population structure [336], which is implemented by the GWAS function in the rrBLUP package [264]. The kinship matrix used for association mapping was generated using a whole-genome high-quality single nucleotide variant (SNV) set [63] and the *A.mat* function from the rrBLUP package. Single-nucleotide variants identified using RAD-marker sequencing [79] that had at least 5% minor allele frequency in the 152 isotype set were used for performing GWA mappings. Association mappings that contained at least one SNV that had a -log(*p*-value) greater than the Bonferroni-corrected *p*-value were processed further using fine mapping, which entails a Spearman's rank correlation test with variants from the whole-genome sequence data of moderate to severe predicted effects as determined by the *SnpEff* function [337].

7.4.4 Temperature competition assays

We chose two strains, CX11314 and JU847, that had different alleles for the peak QTL marker (chrV: 14,822,276; JU847: T, CX11314: A) in our three-year temperature GWA mapping. JU847 has the reference allele for the peak marker and was isolated at a low temperature, whereas CX11314 has the alternative allele for the peak marker and was isolated at a higher temperature. We designed a Taqman probe (5'-[A]CCGTTTTTTTT[T/A]AATTTT-3') to measure each of these two alleles from mixed samples of nematodes using the standard software from Applied Biosystems and a corresponding primer set to amplify the region of interest (below).

$$F : 5' - AAACCCAAGATTTTTATGGTTACTTTAAGATTTGT - 3'$$

$$R : 5' - ATCTATAGTTAACTTGGATATATTGTTTGTTTTCGGT - 3'$$

These two strains were chunked to fresh 10 cm NGMA plates seeded with OP50. 48 hours later, seven L4s from each strain were added to each of 45 6 cm NGMA plates seeded with OP50 for both 15℃ and 25℃ competition experiments. The 45 plates at each temperature represent nine experimental replicates each composed of five independent populations. Plates were placed at either 15ºC or 25ºC and grown to starvation. After one week for 25℃ competition experiments or 10 days for 15℃ competition experiments, nematodes were transferred to fresh NGMA plates by cutting a 0.5 cm x 0.5 cm square of agar (containing approximately 100 worms) and replaced at the appropriate temperature. After culture transfers two, four, and six, starved animals were washed off the plates with M9, and DNA was collected using the Qiagen DNeasy Kit. Genomic DNA from each time point was digested with the EcoRI enzyme and purified using the Zymo DNA Clean Concentrator Kit. The concentration of fragmented genomic DNA was adjusted to 2 ng/μL by Qubit assay. The number of JU847 and CX11314 alleles in each replicate population was measured using Taqman analysis in a Biorad QZ200 digital droplet PCR system. Digital PCR was performed following the standard protocol provided by Biorad with the absolute quantification method. The proportion of the JU847 allele and the relative selection coefficients were calculated.

7.5 RESULTS

7.5.1 Genome-wide association of geographic traits

The location where a *C. elegans* strain was identified could reflect the process of selection for a particular genotype in a specific niche. For this reason, we investigated correlations between genetic variation in the *C. elegans* population and parameters describing the geographic locations of isolation as quantitative traits. Previous work with a smaller set of strains (97 wild isolates) detected a significant QTL on the left arm of chromosome II associated with the latitude where strains were isolated [79]. To evaluate this trait and other parameters describing the location where each strain was isolated in our larger strain set (152 strains), we curated the isolation location information and defined several traits based on geographic data for each strain (**Table 7-1**), namely latitude, longitude, elevation, the absolute value of latitude, and the absolute value of longitude. We performed genome-wide association mappings with 149 wild strains with known isolation locations (**Figure 7-1**) to correlate these trait values with common genetic variation (see Methods). Using this strain set, only the mapping of the elevation of the isolation location identified a significant QTL on the left arm of chromosome III (**Figure 7-2A**). When we divided the population by the genotype at the peak marker, we found that the elevation values for these two sets of strains were similar with a few outliers (**Figure 7-2B**), suggesting that the outliers were causing the detection of a QTL. It is possible that the association is spurious and driven by outliers or that the outliers are strains harboring rare alleles in the *C. elegans* species that impact this trait. The outlier strains in our mapping could share some genetic similarities as they were all collected within the last 15 years, and most were collected in France or elsewhere in Northern Europe. We did not recapitulate the QTL for latitude observed in the previous study [79] likely because it also appears to be driven by strains with extreme latitude values. Again, these outlier strains could be highly related as seven of the 15 strains with the alternate genotype at the peak marker position originated from South Africa or Kenya. Our larger strain set reduces the effect of these outliers on the GWA mapping, and the previously detected QTL is no longer significant. These results suggest that common variation in the *C. elegans* species does not correlate with geographic parameters describing the location of strain isolation. However, rare variants

might control whether strains can colonize and/or proliferate in specific geographic locations.

Table 7-1: Definition of geographic and weather traits[a]

Trait name	Variable[b]	Strains[c]	Description
Latitude	latitude	149	Latitude coordinate where the nematode strain was collected (degrees)
Abs. value of latitude	abslat	149	Absolute value of the latitude coordinate where the nematode strain was collected (degrees)
Longitude	longitude	149	Longitude coordinate where the nematode strain was collected (degrees)
Abs. value of longitude	abslong	149	Absolute value of the longitude coordinate where the nematode strain was collected (degrees)
Elevation	elevation	149	Calculated elevation based on latitude/longitude coordinates using “geosphere” package (m)
Temperature	temp	145	Average daily[d] temperature of the air (℃)
Relative humidity	rh	129	Amount of water vapor present in air expressed as a percentage of the amount needed for saturation at the same temperature (%)
Wind direction	wd	134	Angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing (angular degrees)
Wind speed	ws	136	Rate of horizontal travel of air past a fixed point (m/s)
Cloud height[e]	ceil_hgt	122	Height above the ground level of the lowest cloud or obscuring phenomena layer with 5/8 or more summation total sky cover, which may be predominantly opaque, or the vertical visibility into a surface-based obstruction (m)
Dew point	dew_point	129	Temperature to which a given parcel of air must be cooled at constant pressure and water vapor content in order for saturation to occur (℃)

[a]Definitions for traits obtained from the Federal Climate Complex Data Documentation for Integrated Surface Data (August 20, 2015)

[e]Abbreviation for weather traits obtained from the raw data file format

[e]Number of wild isolates with data for each geographic and weather trait. Number of strains for the weather traits was obtained from the three-month mapping dataset

[e]All weather traits were defined by averaging the average daily values (see Methods)

[e]All weather traits were defined by averaging the average daily values (see Methods)

7.5.2 Weather conditions and climate parameters can be determined using the geographic location of the site of strain isolation

It is likely that the possible genetic association we observed between the elevation of strain isolation and a region on chromosome III is correlated with weather patterns and/or climate variables at specific geographic locations. Strain latitude and longitude can be used to determine the weather or climate

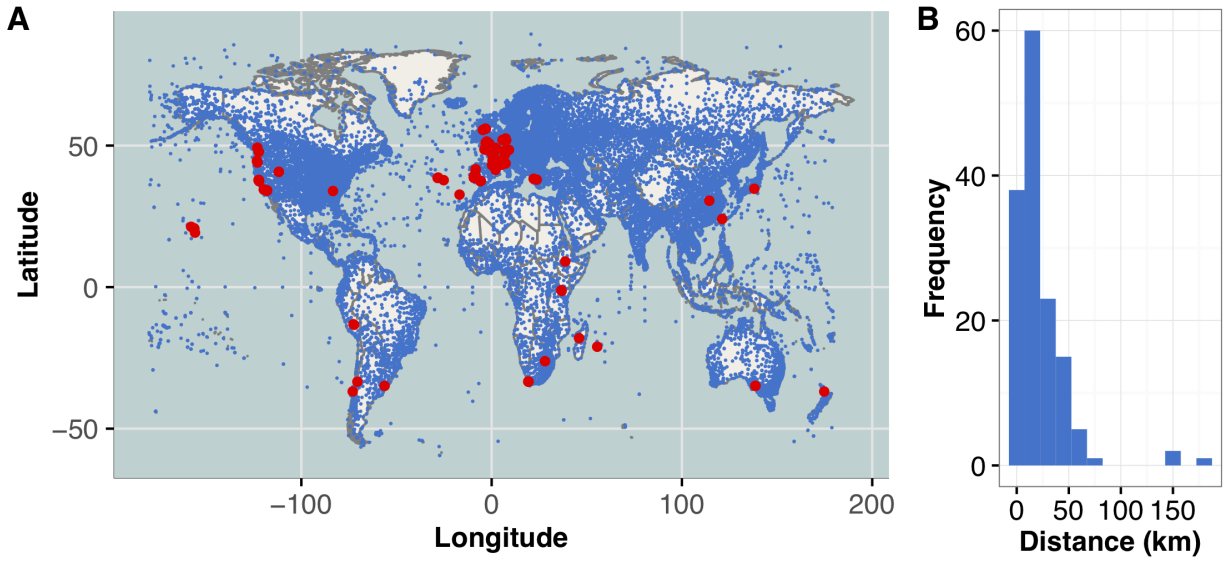


Figure 7-1: Global distribution of wild isolates and NOAA weather stations. (A) Map of 27,447 ISD NOAA weather stations (blue) and 149 *C. elegans* wild isolate locations (red). Three isotypes are not depicted as they have no known location of isolation. **(B)** Histogram of station distance from wild isolate, measured in kilometers (km). Data from the three-year weather dataset are shown.

at the time and location from which each wild *C. elegans* strain was isolated. The short-term weather surrounding the day of isolation as well as the long-term climate of the geographic location for each strain could improve our understanding of adaptation to niches for specific strains or for the species as a whole. Furthermore, correlating these climate parameters with whole-genome sequence data could identify potential alleles that might contribute to adaptation of a wild isolate to certain environmental factors.

The National Oceanic and Atmospheric Administration (NOAA) collects and provides multiple datasets related to weather and climate information, including the Integrated Surfaces Data (ISD). The ISD dataset is archived at the National Climatic Data Center (NCDC) and is composed of worldwide surface weather observations from over 27,000 stations managed by different global institutions [338]. First, we manually curated the isolation information for the 152 strains, including the date of isolation, location, and sampling information. Then, we overlaid the locations of the 27,447 ISD weather stations with the isolation locations of the 149 *C. elegans* wild strains with complete sampling data (**Figure 7-1A**). Using the date and isolation location for each wild strain, we identified the closest weather station with available data and collected weather observations in three time periods surrounding the

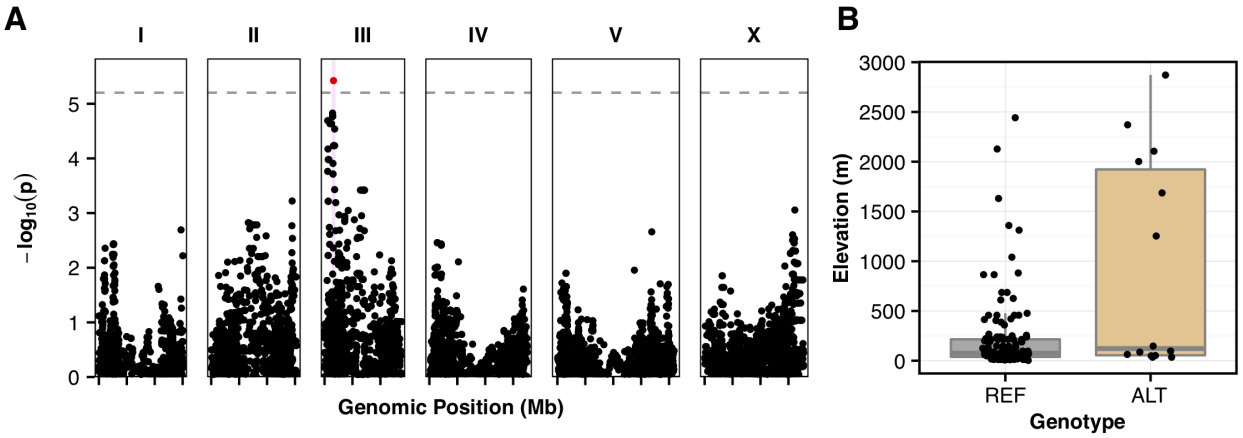


Figure 7-2: Genome-wide association of elevation. (A) Genome-wide association of elevation of strain isolation shown as Manhattan plots. Genomic position is plotted on the x-axis against the negative log-transformed *p*-value on the y-axis. Single nucleotide variants (SNVs) that are above the Bonferroni-corrected significance threshold, indicated by the dotted grey line, are shown in red and SNVs below the Bonferroni threshold are shown in black. Confidence intervals are represented by the pink bars. **(B)** Box plots show strain isolation elevations, measured in meters (m), separated by the genotype at the peak marker location. Each point represents one strain. The reference genotype (REF) refers to strains that share the genotype of the reference strain, N2. The alternative genotype (ALT) refers to all other strains that do not have the reference genotype at the peak marker position.

date of nematode isolation: three months, one year, and three years. Most strains were found less than 60 km from a weather station (**Figure 7-1B**). However, we found a strong correlation (*rho* = 0.707) between weather stations up to 150 km apart, suggesting the weather station assigned to each strain is representative of the weather and climate from which the strain was isolated. Of the 149 strains with known isolation locations, we knew at least the year of isolation for 145 strains, the month for 138 strains, and the day for 122 strains. For the three-month period, we analyzed weather station data only for strains with known days of isolation to provide a precise account of the daily weather experienced immediately surrounding the date of isolation for each strain. For the one-year period, we used data from strains with known day or month of isolation. For the three-year period, we used data from strains with a known day, month, or year of isolation to provide an estimated overall climate of the strain isolation location. Not every weather station sampled contained data for each weather parameter. Additionally, only quantitative weather parameters that were measured at locations shared in a majority of the wild isolate population (more than 90% of the strains) were considered for further analysis (**Table 7-1**). The daily averages of all observations for each weather parameter were averaged over the given time period, and this averaged value was used as the trait measurement for

each strain (see Methods).

We evaluated all weather observations over the three months, one year, or three years surrounding the date of nematode isolation for each of the 149 wild strains to define the weather or climate experienced by each strain. These data were mapped using GWA to define ten QTL for two distinct traits – relative humidity and temperature.

7.5.3 Genome-wide association of average relative humidity

C. elegans is found at locations with various average daily relative humidity ranging from 34% to 89%, with an average of 70%. This estimate of the average relative humidity of locations harboring *C. elegans* is in agreement with previous studies that show *C. elegans* is often found in humid environments [311]. To determine if variation in relative humidity at isolation location is correlated with genetic variation, we performed GWA mapping for relative humidity of isolation location over three months, one year, and three years surrounding the date of isolation (see Methods). We found nine significant QTL in three distinct regions of the *C. elegans* genome: the left arm of chromosome II, the right arm of chromosome III, and the right arm of chromosome V (**Figure 7-3**). The mapping for average relative humidity over three months surrounding the date of isolation (**Figure 7-3A**) resulted in two distinct QTL: one on the left arm of chromosome II and the other the right arm of chromosome V ($LD = 0.341$). The mapping for average relative humidity over one year surrounding the date of isolation (**Figure 7-3B**) resulted in one QTL on the left arm of chromosome II, two linked QTL on the right arm of chromosome V ($LD = 0.536$), and one barely significant QTL on the right arm of chromosome III that is highly linked to the other three QTL ($LD = 0.774, 0.777, 0.548$). The position of the chromosome II and V QTL are the same as those QTL observed for the mapping of three-month humidity. Finally, the mapping of average relative humidity over the three years surrounding the date of isolation (**Figure 7-3C**) resulted in the same two QTL on chromosome V and the QTL on chromosome III as found for the mapping of one-year relative humidity. For each QTL, strains with the reference allele at the peak marker tend to be isolated at higher relative humidity, and strains with an alternative allele tend to be isolated at lower relative humidity. This evidence of a phenotypic split dependent on genotype of the peak marker suggests that at least one variant could contribute to the adaptation of *C. elegans* to different relative humidity.

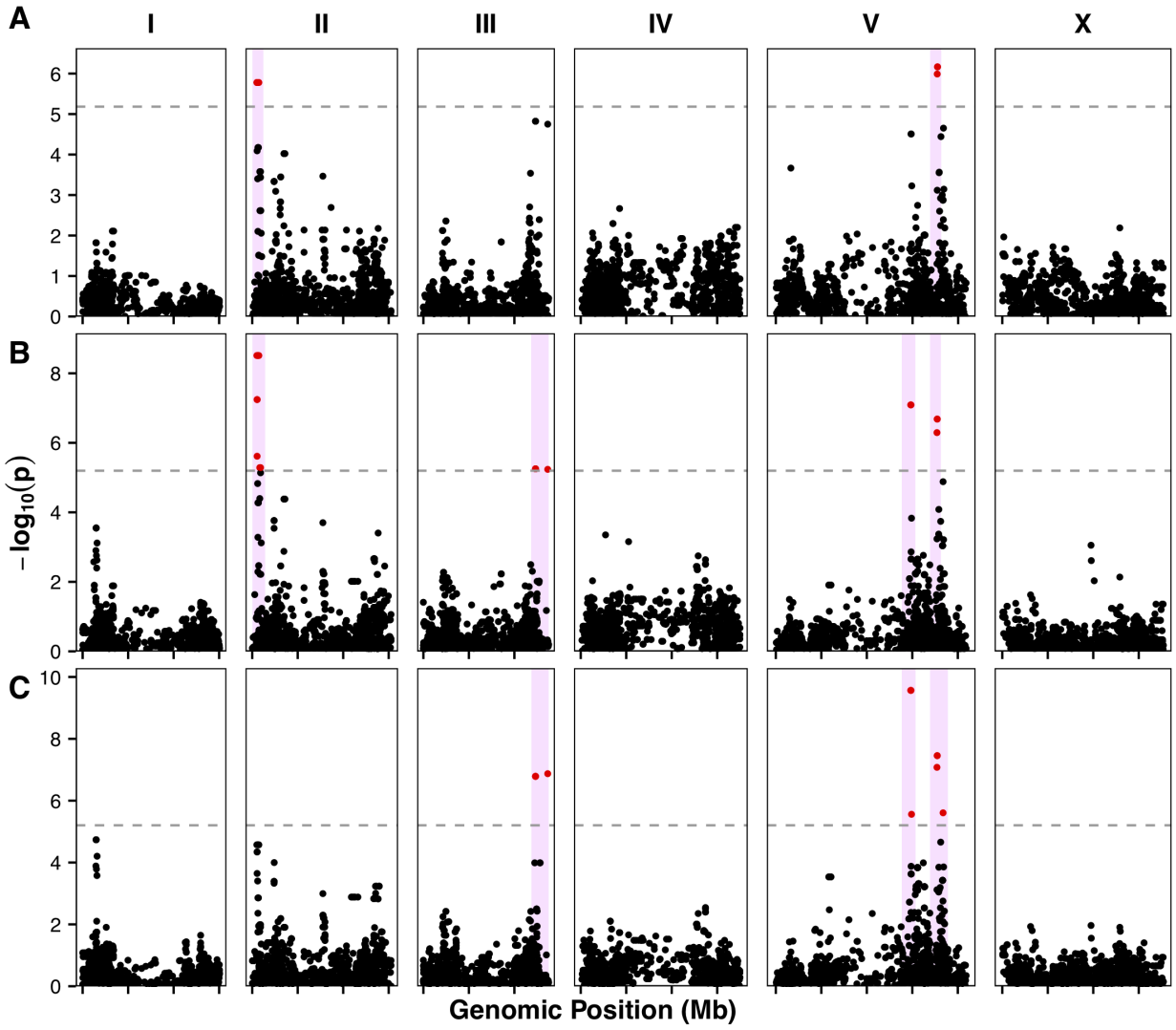


Figure 7-3: Genome-wide association of humidity traits. Genome-wide association of relative humidity for three different time periods are visualized as Manhattan plots: three-month **(A)**, one-year **(B)**, three-year (c) durations. Genomic position is plotted on the x-axis against the negative log-transformed *p*-value on the y-axis. SNVs that are above the Bonferroni-corrected significance threshold, indicated by the dotted grey line, are shown in red and SNVs below the Bonferroni threshold are shown in black. Confidence intervals are represented by the pink bars.

7.5.4 Genome-wide association of average temperature

C. elegans are also found at a variety of average temperatures ranging from 7℃ to 25℃, with an average of 15.3℃. To determine if temperature is associated with genetic variation, we performed a GWA mapping for the average daily temperature of isolation location over three months, one year, and three years surrounding the date of isolation (see Methods). We found one significant QTL just right

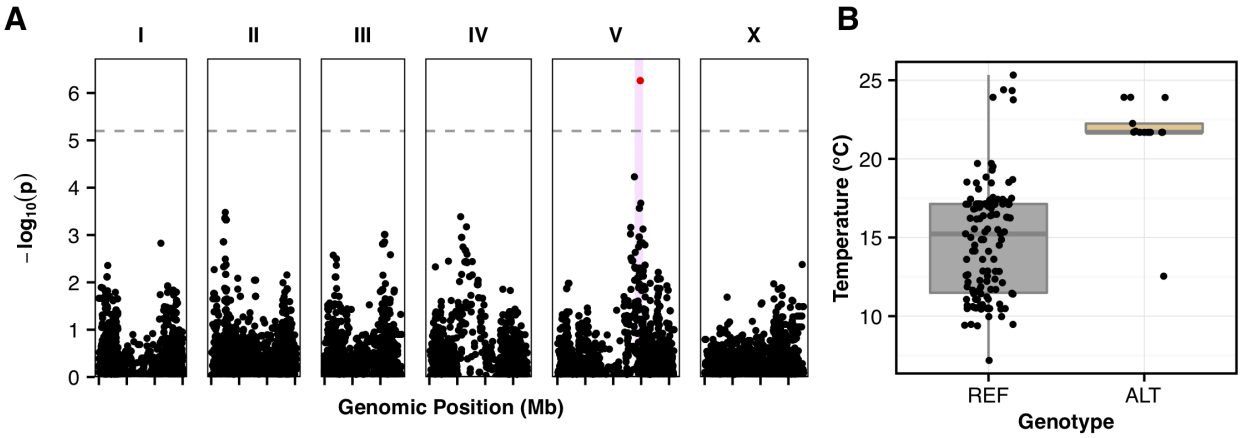


Figure 7-4: Genome-wide association of temperature. (A) Genome-wide association of three-year average temperature is visualized as a Manhattan plot. Genomic position is plotted on the x-axis against the negative log-transformed *p*-value on the y-axis. SNVs that are above the Bonferroni-corrected significance threshold, indicated by the dotted grey line, are shown in red, and SNVs below the Bonferroni threshold are shown in black. Confidence intervals are represented by the pink bars. **(B)** Box plots show the strain three-year average temperatures, measured in degrees Celsius, separated by genotype at the peak marker locus. Each point represents one strain. The reference genotype (REF) refers to strains that share the genotype of the reference strain, N2. The alternative genotype (ALT) refers to all other strains that do not have the reference genotype at the peak marker position.

of the center of chromosome V (**Figure 7-4A**). This QTL is in the same location as that observed for relative humidity (**Figure 7-3**). We found that strains with the reference (N2) allele at the peak marker tend to be isolated from geographic locations with lower temperatures, and strains with an alternative allele at this position tend to be isolated from geographic locations with higher temperatures (**Figure 7-4B**). This QTL suggests that an allele, or alleles nearby this marker, could confer fitness advantages to strains that experience different temperatures. To identify the variant(s) that could underlie this QTL, we investigated a region on chromosome V (V:13,845,281-15,332,878) defined by 1.48 Mb that contains 619 total genes. The genes with predicted functional variants are most likely to cause phenotypic differences among diverse strains in species. Therefore, we focused on 363 genes within this region predicted to harbor functional variants of moderate or severe effects on gene function, as determined by *SnpEff* [337]. We narrowed the list of candidate genes further by identifying 27 genes that are highly correlated with differences in temperature. Although an investigation of these 27 genes did not identify an obvious candidate related to temperature regulation, one or more of these variant genes could explain adaptation to specific temperature.

Because *C. elegans* are isolated at locations with variable temperatures, we wanted to investigate

the variance of temperatures experienced by each strain. We found one significant QTL associated with the variance in annual temperature on chromosome V and another on the right arm of chromosome III. These QTL appear to be linked ($LD = 0.661$) but act in opposing manners. On chromosome III, strains with the reference allele at the peak marker tend to experience lower variance in annual temperature than strains with the alternative allele at this position. However, the strains with the reference allele at the peak marker for the QTL on chromosome V tend to experience higher variance in annual temperature than those strains with the alternative allele. Furthermore, the average minimum daily temperature for three years experienced by each strain also maps to the same chromosome V QTL, while the variance in temperature over the three-year window maps to the same QTL previously identified on chromosome III. Although the same strains have both the alternative allele at the peak marker for average temperature and minimum temperature (and are mostly found in either Southern California or Hawaii), only a few of these strains also have the alternative allele for variance in temperature. Taken together, these data are suggestive of genotypic differences that exist within the strains harboring the alternative allele that highly correlate with being found in locations with higher temperatures and lower temperature variation.

7.5.5 Strains from divergent climates might be adapted to specific temperatures

Although we have GWA mappings for various weather conditions, validating these QTL would provide more evidence for *C. elegans* selection of niche based on environmental and geographic factors. Because temperature can be controlled easily and survival at defined temperatures can be tested experimentally, we decided to determine whether two strains from divergent climates are adapted to the respective temperatures nearby their isolation locations using a competition assay. Strains were chosen that had both a different genotype at the peak marker of the QTL on chromosome V identified in the three-year temperature mapping experiment and a large difference in the temperatures nearby the isolation locations. JU847, a strain isolated from Northern France in 2005, has the reference genotype at the three-year temperature QTL peak marker and was found at a low three-year average temperature (11.3℃). CX11314, a strain isolated from Southern California, USA in 2003, has the alternative genotype at the three-year temperature QTL peak marker and was found at a higher

three-year average temperature (20.9℃). High fitness at or around the temperature nearby the isolation location of one strain and lower fitness at or around the temperature nearby the isolation location for the other strain would suggest potential adaptive alleles that contribute to better survival at a specific temperature.

Replicate cultures were initiated with an equal number of animals from each strain at either 15℃ or 25℃ and allowed to compete for at least six generations. After culture transfers two, four, and six, we analyzed the ratios of the two strains found at each temperature (see Methods). CX11314 was found to have higher fitness than JU847 at both temperatures tested (**Figure 7-5**). At high temperature, CX11314 had a clear selective advantage compared to JU847 (for fitness = 1, relative CX11314 fitness $s = 2.29$), resulting in JU847 alleles comprising less than 1% of the alleles measured after six culture transfers. However, JU847 performed better at the lower temperature than at the higher temperature, comprising almost 8% of the total nematode population after six culture transfers (relative CX11314 fitness $s = 1.57$). These data suggest that JU847, although not more fit than CX11314 at either temperature, is more fit at 15ºC than at 25ºC.

7.6 DISCUSSION

In this study, we have defined geographic, weather, and climate variables over three different time periods as phenotypic traits for 149 unique wild *C. elegans* strains and performed genome-wide association mapping for 11 traits. Each phenotype described in this study was obtained using the location, date, and weather of the known isolation location of each isotype in our collection. We found significant correlations between genotype and phenotype for three traits and a total of 11 QTL. However, only temperature displayed strong phenotypic separation associated with genotypic variation that was likely not driven by outlier strains.

Although we found a significant QTL associated with the elevation of isolation location, we did not recapitulate the QTL for latitude observed in the previous study [79]. This difference is likely because the previous mapping appears to be driven by strains with extreme latitude values, similar to our elevation mapping. The larger strain set in this analysis reduces the effect of these outliers on the GWA mapping, and thus the QTL is no longer significant. However, we did observe a QTL associated

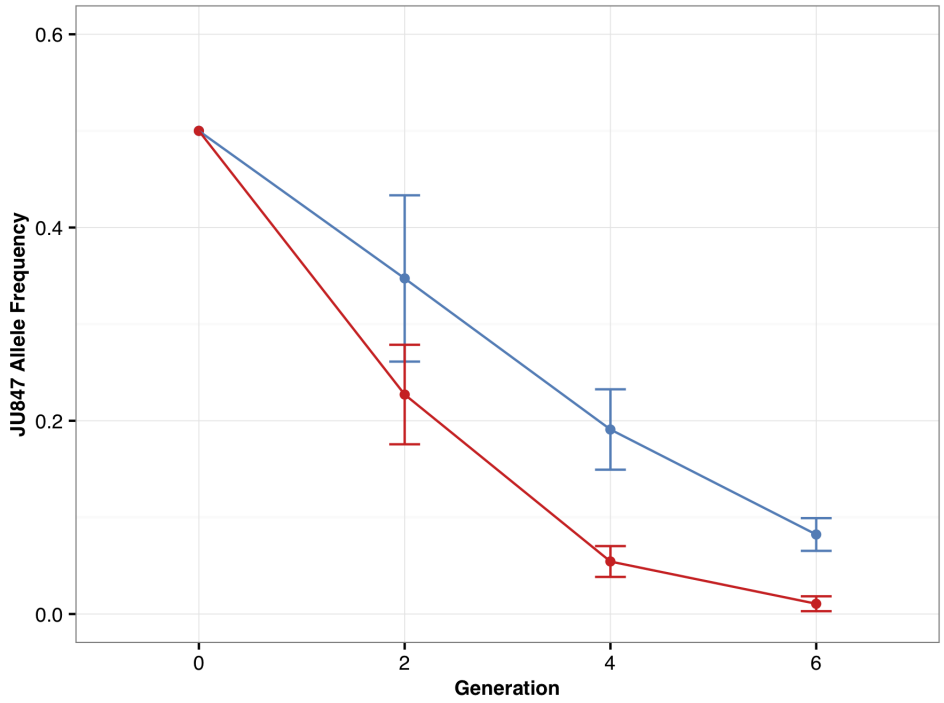


Figure 7-5: Two-strain temperature competition assay. U847 (isolated at low temperature) was competed against CX11314 (isolated at high temperature) at both 15℃ (indicated in blue) and 25℃ (indicated in red). The mean frequency of the JU847 allele in the population is plotted on the y-axis. Error bars represent one standard deviation from the mean. Data were collected from nine experimental and five technical replicates.

with latitude just below the Bonferroni-corrected significance threshold on the right arm of chromosome V. This suggestive QTL is in close proximity to the QTL for temperature. It is possible that the moderate association seen with latitude is representative of a real association with temperature, as the two traits are highly correlated. In our dataset, we see a high negative correlation between the absolute value of latitude and temperature ($rho = -0.827$) and between latitude and temperature ($rho = -0.795$). Unfortunately, the latitude and longitude recorded for a particular strain are not always precise, especially for strains with older isolation dates, which were not recorded as accurately. One tenth of a degree of latitude can distinguish between large cities, but could cover up to 11.1 km of distance. The function we used to determine elevation of strain isolation used these imprecise latitude and longitude coordinates, potentially resulting in a range of small to sizeable errors. These estimations could affect not just the geographic traits, but the weather traits as well, which are calculated based on the geographic coordinates. However, we do not expect this estimation to have a

large effect, as we found weather data between stations up to 150 km apart to be highly correlated ($rho = 0.707$; see Methods).

We chose to assess weather conditions for each of the 149 strains over three time periods: three months, one year, and three years. A duration of three months was chosen to define the weather contemporary to the date of isolation. One year was chosen to evaluate the average weather patterns a strain must be able to survive in nature. Three years was chosen to define the climate of the isolation location. Using data from three years could help us understand the average long-term climate conditions for each strain by eliminating any unusual weather patterns during the year of isolation that are not representative of the overall average environmental conditions. The relative humidity trait maps to the same chromosome positions for all three time periods assayed. Previous studies have also shown that *C. elegans* tend to be found in humid regions [311]. The same QTL on the right arm of chromosome V was observed for both relative humidity and temperature. Because relative humidity depends on air temperature, these traits are expected to be correlated. In our dataset, we see a high correlation between temperature and relative humidity ($rho = -0.675$). Although the average temperature maps with only the three-year data set, we observed a QTL at the same position for minimum daily temperature and another QTL just below the significance threshold for the average one-year temperature mapping dataset. Additionally, this QTL was mapped for other weather traits such as maximum daily dew point and variance of wind direction. Similar to relative humidity, dew point is dependent on (and highly correlated with) ambient temperature. However, wind direction and temperature are more difficult to relate to one another. Perhaps, this QTL represents residual population structure shared by strains at this location and not suggestive of underlying mechanisms impacting fitness in differing environmental conditions. Volkers *et al.* attempted to characterize such genomic and phenotypic diversity in wild *C. elegans* populations and discovered major hotspots of polymorphic genes on the left arm of chromosome II and the right arm of chromosome V [339]. Maybe the higher genetic diversity observed at these positions is controlling the observed QTL for temperature and other weather traits. Regardless, we observe several QTL that provide evidence that one or more variants on chromosome V are associated with differences in long-term average temperature where *C. elegans* strains are isolated from nature.

The QTL for temperature was evaluated further because it could be controlled in a laboratory setting. The competition assay between JU847 and CX11314 showed that CX11314 had a higher selective advantage than JU847 at both high and low temperatures. This result might be observed because the laboratory environment cannot completely recapitulate conditions experienced in the wild. Additionally, it is possible that CX11314 is more fit than JU847 regardless of temperature. To test this QTL more thoroughly and eliminate this possibility, it would be necessary to compete multiple high and low temperature strains. Alternatively, because JU847 and CX11314 differ throughout their genomes in unknown ways that could affect overall fitness, testing near-isogenic lines (NILs), in which only the region surrounding the QTL is different between the two strains, would be a better way to test temperature-dependent fitness effects of this QTL. Regardless, we found that the low temperature strain, JU847, was more competitive with the high temperature strain, CX11314, at the lower temperature. This result provides evidence that one or more genetic variants within the QTL could contribute to a higher fitness at specific temperatures. Furthermore, the duration of this experiment was only six weeks, while the QTL was identified for a time span of three years. At the lower temperature, it is possible we could observe a stronger competitive advantage for JU847 over a longer time period.

Although our analyses were unable to identify a single gene or variant that could underlie potential differences in niche specification, our conclusions suggest that different strains are found in unique niches and at least some of the environmental differences in niches are related to genetic variation among strains. As we expand our collection of wild *C. elegans* strains, we will be able to better define these weather and climate differences. Additionally, longitudinal collection studies with dense sampling, especially in a location with known high species diversity (such as the Hawaiian Islands) would give us more valuable data about how genetic variation in *C. elegans* is related to environmental conditions. We expect that similar data could be analyzed for other species and allow for investigation of niche specification, specifically in plant species where dense sampling and whole-genome datasets are available.

7.7 FUTURE DIRECTIONS

This study identified genetic variation on chromosome V that was associated with the temperature and humidity of the climate in which strains were isolated. We attempted to validate this QTL using a competition assay between two strains isolated from different environments and found some evidence that the strain isolated from cooler temperatures performed better in cooler temperatures than the strain isolated from warmer temperatures. However, only one strain per environment was chosen and there is a lot of genetic variation between these two strains, not just on chromosome V. A more thorough experiment would be to create near-isogenic lines (NILs) isolating the region on chromosome V. For example, if CX11314 were to show a higher fitness at warmer temperatures than a strain with the CX11314 genetic background and a JU847 introgression on chromosome V, then we would know that genetic variation in this region contributes to the nematode's adaptation to different temperatures. It would also be best to create three or four pairs of reciprocal NILs from strains with different haplotypes at the chromosome V region to best capture the variation among the GWA panel.

In addition to testing NILs, this study could be repeated as more wild strains are collected. Additional QTL might be discovered as we collect more divergent strains from different climates. We already noticed that there was a change in the genetic architecture of latitude of isolation from the 96 strain set to the 152 strain set containing more genetic diversity [79]. This likely indicates that the population structure might influence association mapping results. To help improve estimation of the nematode environmental niche, newer collections even include documented environmental parameters measured at the time of collection that could be used for association mapping [59]. Finally, adapting this study to other available datasets could be enlightening. For example, a dataset detailing soil environmental parameters such as water, oxygen, or carbon content could be more informative than traits like cloud cover, as *C. elegans* are soil dwelling nematodes.

7.8 CONTRIBUTIONS

I would like to thank Dr. Patrick McGrath and his students Yuehui Zhao and Lijiang Long from the Georgia Institute of Technology for their help designing, implementing, and analyzing the competition

assay to test if genetic variation on chromosome V leads to fitness advantages at different temperatures. I would further like to thank Dr. Shannon Brady for her help setting up the competition assay and Dr. Daniel Cook for his help learning R and analyzing the GWAS results. Lastly, I would like to thank Dr. Erik Andersen for his mentorship and encouragement to publish results from my rotation project.

7.9 CITATION

Correlations of genotype with climate parameters suggest *Caenorhabditis elegans* niche adaptations

Kathryn S. Evans[1,2], Yuehui Zhao[3], Shannon C. Brady[1,2], Lijiang Long[3], Patrick T. McGrath[3], and Erik C. Andersen[1,4,5]

[1]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208
[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, Illinois 60208
[3]School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332
[4]Chemistry of Life Processes Institute, Northwestern University, Evanston, Illinois 60208
[5]Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208

This manuscript was published in G3 in November 2016 [136]

8 Discussion

The majority of my graduate work was focused on identifying causal genetic variants that underlied a variety of toxin-response traits in *C. elegans*. The goal of this work is three-fold: to better understand the genetic architecture of complex traits, to discover the mechanism by which natural genetic variation influences a phenotype, and to hypothesize how the forces of evolution acted to produce the observed phenotypic variation. This feat is surprisingly more difficult than it seems due simply to the complexity of quantitative traits. We learned early on that traits with overlapping QTL from both linkage and association mapping often converged relatively easily on a small number of candidate variants [70, 71]. However, QTL in densely populated genomic regions (such as the chromosomal arms) or genomic regions with extremely high genetic diversity (divergent regions) present even more difficult problems [123]. Because of several expected and unexpected challenges, I spent a lot of time trying to identify the causal variant, leaving little time to explore the more interesting mechanistic and evolutionary questions. Here, I will discuss what I have learned from my experiences performing quantitative genetics in *C. elegans* and provide insight that will hopefully propel future research in the field.

8.1 Functional validation to identify the causal variant underlying a QTL

One of the many benefits of performing quantitative genetics in *C. elegans* over other species is the ability to functionally validate detected QTL. NILs are an extremely useful tool in this process because they allow testing of genomic regions (small or large) without first choosing specific candidate genes. This unbiased approach to identifying causal variants is important because prior knowledge of gene function might not be indicative of the mechanism by which natural genetic variation influences phenotypic traits. Furthermore, many genes in the *C. elegans* genome remain unannotated with no known function. A panel of NILs with increasingly smaller introgressed regions can also be generated and phenotyped relatively quickly to narrow the QTL by identifying the smallest genomic region that produces the NIL phenotype. However, if the QTL confidence interval is already small or your QTL is in the center of a chromosome where little recombination occurs, this method might not be beneficial.

Targeted double strand breaks with CRISPR-Cas9 machinery in heterozygous individuals followed by homology directed repair might improve the speed and precision of NIL construction. Although it has not been used yet in *C. elegans*, this method (sometimes called 'loss of heterozygosity') has been successfully used in yeast to generate a panel of recombinants [340] and offers hope for a new era of fine-mapping with well-designed recombinants.

NILs are quite effective for simple QTL with a single causal variant. However, there are several cases in which QTL are more complex than they seem, in which case NILs might yield more questions than answers. In Chapter 2, we observed many examples of multiple additive or interacting loci in the NILs that were undetected in the original QTL scan. This result offers two main questions: why were these secondary loci not detected with linkage mapping and why do we see them with NILs? To answer the first, the power to detect QTL with linkage mapping is increased with the addition of more strains. Because many of the putative interactions we identified were uni-directional, we likely did not have enough strains with the correct genotype combination to detect a signal. In answer to the second question, NILs remove the genetic heterogeneity of the RIAIL background, making it easier to observe genetic interactions between two loci. In Chapter 4, we saw a similar, yet distinct, issue where a single QTL fractionates into two or more small-effect loci. This can make it laborious to identify all small-effect variants contributing to an already small-effect phenotype. Several quantitative genetics studies do not even attempt to map QTL to QTN for this very reason and focus instead on understanding the overall genetic architecture of a trait.

There is no way to escape these issues of uncovering fractionating and undetected QTL because these results are a product of the underlying genetics of a complex trait. In fact, in these cases, NILs provide more information than could be gleaned from a CRISPR-targeted knockout of a specific candidate gene. However, there are certain steps that can be taken to decrease frustration induced by complicated NIL behavior. First, for all QTL mapping studies, but particularly for those with small-effect fractionating loci, it is essential to have a high throughput and reproducible phenotyping assay that is sensitive enough to detect small differences and robust enough to distinguish signal from noise. Any precautions that can be taken to reduce environmental variation will increase the chance of success. It is essential to be confident that small differences in NIL phenotypes are in fact derived from

differences in NIL genotypes and are not an artifact of the assay. Second, efforts to improve fine-mapping and candidate gene prioritization would lessen the need for a large panel of NILs to narrow the QTL. This is already often the case for QTL identified with association mapping due to the decreased linkage disequilibrium between variants. Furthermore, mapping with two or more populations (*e.g.* linkage and association) could help to implicate or exclude genomic intervals. And finally, it might be necessary to sensitize a NIL if there are multiple QTL of opposite effect sizes for a given trait (see Chapter 4, *sqst-5* deletions).

Similar to NILs, genome editing with the CRISPR-Cas9 system is another way to perform functional validation in *C. elegans*, but often for specific candidate genes of interest. CRISPR can easily be used to generate full deletions of genes leading to a loss-of-function mutant. However in some cases, large insertions or deletions spanning several genes might be possible. To aid the study of natural variation, CRISPR can also be used to introduce a genomic region or even a single nucleotide variant from one strain (such as CB4856) into another (such as N2). Although extremely powerful, CRISPR relies on having a prioritized list of candidate genes. However, QTL often contain hundreds of genes, further necessitating a method for narrowing a region and/or prioritizing candidate genes. An ambitious student could methodically delete ten or more top candidate genes, but this takes lots of time and money and doesn't even guarantee success. Sometimes a loss-of-function mutation can have an effect on the phenotype without being the causal variant that underlies the QTL [72]. Othertimes, the loss-of-function mutation might not have an effect even though variation in the gene does cause the phenotypic variation observed [123]. Additionally, some loss-of-function mutations cause other phenotypes like slow growth (Chapter 6, *lgc-54* mutants) or even lethality. In these examples, we have shown that a reciprocal hemizygosity test would be necessary to show causation [72, 123].

8.2 Benefits and limitations of high-sensitivity phenotyping

The sorter platform in the Andersen Lab is great for quantitative genetics because it allows us to discern small differences that we can't observe with the human eye. The ability to detect small differences is necessary for accurately describing quantitative variation. Using this platform, we were able to identify QTL that explained less than 5% of the phenotypic variation and in some cases we were able to validate

these loci using NILs [73, 123]. This is really important because it gets us closer to identifying all loci that contribute to the heritable phenotypic variation for a single trait. Identifying all loci allows us to begin answering questions about the components of missing heritability. What is the genetic architecture of complex traits? To what extent do epistatic interactions between loci contribute to complex traits? In Chapter 2, we attempted to answer these questions using the sorter platform and linkage mapping analysis. We show evidence for many small-effect QTL but, even with our 296 RIAILs, we still do not have the power to distinguish genetic interactions in the recombinant panel.

When I joined the Andersen Lab, most of the large-effect "low-hanging" QTL were already being pursued, which made me the first to attempt to identify causal genes underlying medium- and small-effect QTL (Chapters 4-6). With this, came a new set of challenges: in order to trust such small phenotypic differences, it is essential that the environmental variation within an assay is minimized. For example, with the chromosome V QTL for zinc response in Chapter 4, I observed significant variation between experiments that made it difficult to pinpoint the causal loci. Furthermore, in Chapter 5, I provide a clear example of a trait (docetaxel response) controlled by a genetic locus whose effect is drowned out almost entirely by large environmental variation (food quality). Although nothing could be done to replicate the original food source to continue studying the docetaxel response, I was able to minimize the effect of background variation in the zinc response by searching for the experimental variable that introduced the most variation (bleach synchronization) and then replicating my experiments across this variable. Moving forward, I believe that this is the best way to feel confident in results for small-effect loci. Regardless, there is still a limit to the smallest QTL effect size that can be accurately and robustly validated in our assay, currently estimated near 5% of variation in the RIAILs explained by the genetic locus. As new phenotyping assays are developed, it will be crucial to fully evaluate the assay's sensitivity and specificity and only attempt to validate QTL that fit these specifications, as disregard for the influence of environmental variation can lead to a misinterpretation of results.

8.3 Representing quantitative traits for mapping and validation

One of the most important parts of QTL mapping is selecting a trait that not just describes the biology you are looking to study, but also is quantitative, varies across the population, and is heritable.

Sometimes this is simple: if you are studying lifespan, your trait could be the number of days a worm lives for. Other times, the trait you are looking to study is more complex or cannot be measured directly. Our current approach in the Andersen Lab is to describe toxin responses as a function of animal length, optical density, and brood size. These traits have been successfully used to discover the causal gene or variant underlying several toxin responses [70, 71, 72, 123, 69], demonstrating that we have captured at least some of the nematode's toxin-response biology with these traits. One of the main challenges, but also one of the main advantages, to using these traits is that because each of these trait measurements are taken at a population-wide level, we have replicated distributions for each trait rather than a single value. Early in the development of this assay, we described how summary statistics describing these distributions can be used to more thoroughly investigate the complexity of these traits [68]. For example, instead of just taking the mean animal length of the population, we could also take the 10^{th} and 90^{th} quantiles to look at the range of phenotypic variation.

Despite the strengths of more thoroughly describing a population, there are several weaknesses to this method. First, this method leads to multiple correlated traits describing one population, and treating them each as independent traits can lead to an oversampling bias. To address this issue, we have occasionally used principal component analysis (Chapter 2) to condense traits and identify the main components of phenotypic variance [73, 70]. However, the main disadvantage to this method is that these new traits lose biological relevance. Second, how can the trait biology be correctly interpreted if the median animal length has a QTL that the mean animal length does not? I worry that having so many traits to choose from could result in "cherry picking" the trait based on the expected result. Therefore, when describing results for a particular trait, it is important to remember the biology behind the numbers. Finally, increased environmental variation could alter the calculated quantiles. For this reason, I believe that the median values are often the most robust and reproducible. However, simplifying the distributions by describing them with a single value removes potentially valuable information hidden within the distribution itself. In the future, especially as new assays are developed that have more trait measurements, it would be great to see an implementation of multivariate trait mappings as seen in other systems [341, 342, 343]. However, like principal component analysis, it is unclear how to best functionally validate these complex traits post-mapping.

As mentioned before, our toxin-response assay measures indirect, organismal phenotypes that are often very complex and might be driven by the combination of numerous loci. By contrast, molecular phenotypes such as gene expression are often more simplistic–a single locus may contribute most of the heritable variation for a gene's expression, while the expression of this gene might only partly contribute to the organismal phenotype. Although several molecular phenotypes have been mapped in *C. elegans* in our lab and others [112, 113, 80, 108, 114, 115, 116, 117, 118], very few have been analyzed to identify the underlying causal variant. In one example, a variant in the neuropeptide receptor *npr-1* was shown to control expression of 247 genes [100]. There are a plethora of expression QTL (eQTL) to be further explored, including at least two additional eQTL hotspots [115, 69] with potential candidate genes. Furthermore, the lab could begin to explore other molecular phenotypes. In Chapter 4, I suggested measuring the basal zinc content in the N2 and CB4856 strains to investigate whether differences in natural zinc content in the nematode contributes to their variable tolerance to high exogenous zinc. Connecting molecular phenotypes to organismal phenotypes through methods like mediation analysis (Chapter 3) could further enhance our knowledge on the genetic and molecular mechanisms underlying phenotypic variation.

8.4 Mediation analysis to improve candidate gene prioritization

Although any genetic variation might cause phenotypic variation, it is most likely that the causal genetic variation will impact gene expression or function. For this reason, we can begin to narrow a list of candidate genes to only include genes in the genomic region of interest with genetic variation that causes a change in the amino acid sequence of the protein (protein-coding variation) and genes with expression variation linked to variation in the genomic region of interest. To further prioritize these genes, we can analyze the functional descriptions and GO terms associated with each gene to evaluate if a gene fits the functional profile associated with the phenotype. However, doing this introduces bias into the candidate screening and must be interpreted with caution as there are still many genes in the *C. elegans* genome with unknown functions. Furthermore, some genes with an obvious functional connectivity might not be causal such as *jmjd-5* for bleomycin [72] and *cdr-6* for zinc (Appendix C).

In Chapter 3, to further investigate genes with expression variation, I adapted a method for causal

mediation analysis to suggest links between genetic variation, gene expression, and drug responses. The simple overlap of expression QTL and drug-response QTL could indicate that a single variant controls gene expression and the drug-response phenotype. However, wide confidence intervals for the drug-response QTL results in a large number of overlapping eQTL. Mediation analysis uses a series of linear regression models to deduce the probability that a single variant drives variation in gene expression, which in turn affects the drug response. Mediation analysis is not a conclusive test to identify causal genes. However, it can be very effective to help identify the causal gene, especially if the gene function is unknown and/or the gene is distantly localized to the QTL. Although this method has the potential to be extremely powerful, the tools currently available in the Andersen Lab are fairly weak and introduce several caveats. Most importantly, we only have genome-wide expression data for the set 1 RIAILs, although these strains have lowly heritable drug responses. The results of mediation analysis would be much more robust and interpretable if we had genome-wide expression data for the set 2 RIAILs that were primarily used for QTL mapping of drug responses. In addition to using mediation analysis to prioritize candidate genes in linkage mapping, genome-wide expression data collected from the 328 wild isolates would facilitate mediation for association mapping. Due to the differences in population structure and the inherently lower linkage disequilibrium in the wild isolates compared to the recombinants, mediation might be even more advantageous in wild isolates. In addition to investigating other methods to fine-map a QTL, I believe that incorporating mediation analysis into the usual Andersen Lab method for prioritizing candidate genes will prove to be an investment worth making.

8.5 QTL in divergent regions present new challenges that need to be addressed

In addition to large genomic regions with many genes, divergent regions (like the left arm of chromosome III) complicate the discovery of the causal gene or variant underlying a QTL. Most work in *C. elegans* has traditionally been performed in the N2 genetic background for which we have a reasonably large wealth of knowledge about the genomic features such as gene content and annotations. By contrast, relatively little has been studied about the genome of other wild isolates. We know that the CB4856 strain is highly genetically divergent from the genome of the N2 strain and that

this diversity is largely contained in punctuated divergent regions [60, 62]. Recently, our lab has shown that these punctuated divergent regions are a common feature across all wild isolates [277]. Using long-read sequencing data from 15 wild isolates, we showed that genetic variation inside the divergent regions rivals that seen between species of *Caenorhabditis* and that the gene content in these regions are not even conserved between strains [277].

With this in mind, inferring gene annotations in the CB4856 strain from the N2 reference strain, at least in divergent regions, is unreliable. In Chapter 4, we were lucky that the causal gene was present in both the N2 and CB4856 strains. However, large structural variation in the CB4856 strain did result in a lack of gene prediction for *sqst-5* using *de novo* assembly from long-read sequencing [123]. To improve the chance of success in the future, we first need to first know what genes are in the region for other strains besides the N2 strain. To accomplish this, long-read sequencing is needed to perform *de novo* genome assemblies for each strain. Second, accurate gene prediction models are necessary to compare gene content between strains. Finally, we need curated gene annotations similar to what we have for the N2 reference to help identify potential candidate genes based on function. We now have long-read sequence data for 15 strains, including N2 and CB4856 [277, 62], and I am hopeful that the inclusion of these data will improve our knowledge about the *C. elegans* genome and strengthen our ability to discover the causal variant underlying a QTL.

One potential alternative to using CRISPR to delete specific genes in divergent regions that might not be conserved between strains is fosmid rescue. Fosmid libraries can be generated and used to incorporate genomic regions of approximately 40 kb from one strain (such as CB4856) into another (such as N2). The benefit of fosmids lies in using genomic DNA; it is impartial to the genetic variation or difference in gene content between strains. This is particularly important for narrowing QTL in divergent regions. If a particular fosmid generated from the CB4856 strain is found to rescue the CB4856 phenotype in an N2 strain, this fosmid can be re-sequenced or re-analyzed to identify all differences between strains in this small genomic region. The major downfall for fosmid rescue is in the nature of overexpression. A negative result might not be informative, however a positive result could drive forward the search for the causal gene. Libraries currently exist for both the N2 and CB4856 strains, and I have recently re-mapped the CB4856 fosmid library to the newly assembled

CB4856 genome to aid future fosmid rescue experiments. However, QTL identified using a different set of strains (besides N2 and CB4856) would require first generating the fosmid library, adding another laborious step to this process.

8.6 A look forward: the future of quantitative genetics

In conclusion, better tools and newer technology have led to an increase in the number of QTL detected and QTN identified over the last decade. In the coming years, I expect more QTN with medium-to-small effect sizes to be discovered due to the implementation of suggestions outlined in this chapter in addition to utilizing more powerful mapping methods such as bulk-segregant analysis, multi-parental RILs, or a combination of both linkage and association mapping. The research in this field is vital and cutting-edge. The more QTL we detect, the better our understanding of the genetic factors that contribute to phenotypic variation. Additionally, the more QTN we identify from these QTL, the better our understanding of the genetic mechanisms and the evolutionary forces that shape phenotypic variation in natural populations.

References

[1] Byers, D. L. Components of phenotypic variance. https://www.nature.com/scitable/topicpage/adaptation-and-phenotypic-variance-1132/. Accessed: 2020-8-24.

[2] Roser, M., Appel, C. & Ritchie, H. Human height. *Our world in data* (2013).

[3] Nathans, J., Piantanida, T. P., Eddy, R. L., Shows, T. B. & Hogness, D. S. Molecular genetics of inherited variation in human color vision. *Science* **232**, 203–210 (1986).

[4] Gregor mendel and the principles of inheritance. https://www.nature.com/scitable/topicpage/gregor-mendel-and-the-principles-of-inheritance-593/. Accessed: 2020-8-23.

[5] Lamichhaney, S. *et al.* Evolution of darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).

[6] Walser, T. *et al.* Smoking and lung cancer: the role of inflammation. *Proc. Am. Thorac. Soc.* **5**, 811–815 (2008).

[7] Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* **10**, 565–577 (2009).

[8] Sturtevant, A. H. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *J. Exp. Zool.* **14**, 43–59 (1913).

[9] Mohan, M. *et al.* Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol. Breed.* **3**, 87–103 (1997).

[10] Jorgensen, E. M. & Mango, S. E. The art and design of genetic screens: caenorhabditis elegans. *Nat. Rev. Genet.* **3**, 356–369 (2002).

[11] Forsburg, S. L. The art and design of genetic screens: yeast. *Nat. Rev. Genet.* **2**, 659–668 (2001).

[12] St Johnston, D. The art and design of genetic screens: Drosophila melanogaster. *Nat. Rev. Genet.* **3**, 176–188 (2002).

[13] Patton, E. E. & Zon, L. I. The art and design of genetic screens: zebrafish. *Nat. Rev. Genet.* **2**, 956–966 (2001).

[14] Kile, B. T. & Hilton, D. J. The art and design of genetic screens: mouse. *Nat. Rev. Genet.* **6**, 557–567 (2005).

[15] Page, D. R. & Grossniklaus, U. The art and design of genetic screens: Arabidopsis thaliana. *Nat. Rev. Genet.* **3**, 124–136 (2002).

[16] Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to huntington's disease. *Nature* **306**, 234–238 (1983).

[17] McKusick, V. A. Current trends in mapping human genes. *FASEB J.* **5**, 12–20 (1991).

[18] Kingston, H. M., Thomas, N. S., Pearson, P. L., Sarfarazi, M. & Harper, P. S. Genetic linkage between becker muscular dystrophy and a polymorphic DNA sequence on the short arm of the X chromosome. *J. Med. Genet.* **20**, 255–258 (1983).

[19] Zielenski, J. & Tsui, L. C. Cystic fibrosis: genotypic and phenotypic variations. *Annu. Rev. Genet.* **29**, 777–807 (1995).

[20] White, R. *et al.* A closely linked genetic marker for cystic fibrosis. *Nature* **318**, 382–384 (1985).

[21] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

[22] Rockman, M. V. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1–17 (2012).

[23] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).

[24] Mäki-Tanila, A. & Hill, W. G. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355–367 (2014).

[25] Ehrenreich, I. M. Epistasis: Searching for interacting genetic variants using crosses. *G3* **7**, 1619–1622 (2017).

[26] Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

[27] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234–237 (2013).

[28] Bloom, J. S. *et al.* Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nat. Commun.* **6**, 8712 (2015).

[29] Nelson, R. M., Pettersson, M. E. & Carlborg, Ö. A century after fisher: time for a new paradigm in quantitative genetics. *Trends Genet.* **29**, 669–676 (2013).

[30] Mackay, T. F. C. Epistasis for quantitative traits in drosophila. *Methods Mol. Biol.* **1253**, 47–70 (2015).

[31] Lachowiec, J., Shen, X., Queitsch, C. & Carlborg, Ö. A Genome-Wide association analysis reveals epistatic cancellation of additive genetic variance for root length in arabidopsis thaliana. *PLoS Genet.* **11**, e1005541 (2015).

[32] Malmberg, R. L., Held, S., Waits, A. & Mauricio, R. Epistasis for fitness-related quantitative traits in arabidopsis thaliana grown in the field and in the greenhouse. *Genetics* **171**, 2013–2027 (2005).

[33] Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–1198 (2012).

[34] Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet.* **5**, e1000459 (2009).

[35] Doroszuk, A., Snoek, L. B., Fradin, E., Riksen, J. & Kammenga, J. A genome-wide library of CB4856/N2 introgression lines of caenorhabditis elegans. *Nucleic Acids Res.* **37**, e110 (2009).

[36] Wong, G. T. Speed congenics: applications for transgenic and knock-out mouse strains. *Neuropeptides* **36**, 230–236 (2002).

[37] Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

[38] Paix, A. *et al.* Scalable and versatile genome editing using linear DNAs with microhomology to cas9 sites in caenorhabditis elegans. *Genetics* **198**, 1347–1356 (2014).

[39] Doudna, J. A. & Charpentier, E. Genome editing. the new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).

[40] Frøkjær-Jensen, C. Exciting prospects for precise engineering of caenorhabditis elegans genomes with CRISPR/Cas9. *Genetics* **195**, 635–642 (2013).

[41] Dickinson, D. J. & Goldstein, B. CRISPR-Based methods for caenorhabditis elegans genome engineering. *Genetics* **202**, 885–901 (2016).

[42] Stern, D. L. Identification of loci that cause phenotypic variation in diverse species with the reciprocal hemizygosity test. *Trends Genet.* **30**, 547–554 (2014).

[43] Brenner, S. The genetics of caenorhabditis elegans. *Genetics* **77**, 71–94 (1974).

[44] Paix, A., Folkmann, A., Rasoloson, D. & Seydoux, G. High efficiency, Homology-Directed genome editing in caenorhabditis elegans using CRISPR-Cas9 ribonucleoprotein complexes. *Genetics* **201**, 47–54 (2015).

[45] Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, caenorhabditis elegans. *Dev. Biol.* **56**, 110–156 (1977).

[46] Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

[47] White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 1–340 (1986).

[48] Hall, D. H. & Russell, R. L. The posterior nervous system of the nematode caenorhabditis elegans: serial reconstruction of identified neurons and complete pattern of synaptic interactions. *J. Neurosci.* **11**, 1–22 (1991).

[49] Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H. & Chklovskii, D. B. Structural properties of the caenorhabditis elegans neuronal network. *PLoS Comput. Biol.* **7**, e1001066 (2011).

[50] Jarrell, T. A. *et al.* The connectome of a decision-making neural network. *Science* **337**, 437–444 (2012).

[51] White, J. K. *et al.* Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**, 452–464 (2013).

[52] C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).

[53] Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

[54] Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19428–19433 (2007).

[55] Kaletta, T. & Hengartner, M. O. Finding function in novel targets: C. elegans as a model organism. *Nat. Rev. Drug Discov.* **5**, 387–398 (2006).

[56] Shaye, D. D. & Greenwald, I. OrthoList: a compendium of c. elegans genes with human orthologs. *PLoS One* **6**, e20085 (2011).

[57] Wang, Z. & Sherwood, D. R. Chapter 5 - dissection of genetic pathways in c. elegans. In Joel H. Rothman and Andrew Singson (ed.) *Methods in Cell Biology*, vol. 106, 113–157 (Academic Press, 2011).

[58] Sterken, M. G., Snoek, L. B., Kammenga, J. E. & Andersen, E. C. The laboratory domestication of caenorhabditis elegans. *Trends Genet.* **31**, 224–231 (2015).

[59] Crombie, T. A. *et al.* Deep sampling of hawaiian caenorhabditis elegans reveals high genetic diversity and admixture with global populations. *Elife* **8**, e50465 (2019).

[60] Thompson, O. A. *et al.* Remarkably divergent regions punctuate the genome assembly of the caenorhabditis elegans hawaiian strain CB4856. *Genetics* **200**, 975–989 (2015).

[61] Koch, R., van Luenen, H. G., van der Horst, M., Thijssen, K. L. & Plasterk, R. H. Single nucleotide polymorphisms in wild isolates of caenorhabditis elegans. *Genome Res.* **10**, 1690–1696 (2000).

[62] Kim, C. *et al.* Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in c. elegans. *Genome Res.* **29**, 1023–1035 (2019).

[63] Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* (2016).

[64] Gosai, S. J. *et al.* Automated high-content live animal drug screening using c. elegans expressing the aggregation prone serpin α1-antitrypsin Z. *PLoS One* **5**, e15460 (2010).

[65] Leung, C. K., Deonarine, A., Strange, K. & Choe, K. P. High-throughput screening and biosensing with fluorescent c. elegans strains. *J. Vis. Exp.* (2011).

[66] Leung, C. K. *et al.* An ultra high-throughput, whole-animal screen for small molecule modulators of a specific genetic pathway in caenorhabditis elegans. *PLoS One* **8**, e62166 (2013).

[67] Mondal, S. *et al.* Large-scale microfluidics providing high-resolution and high-throughput screening of caenorhabditis elegans poly-glutamine aggregation model. *Nat. Commun.* **7**, 13023 (2016).

[68] Andersen, E. C. *et al.* A powerful new quantitative genetics platform, combining caenorhabditis elegans High-Throughput fitness assays with a large collection of recombinant strains. *G3* **5**, 911–920 (2015).

[69] Evans, K. S. & Andersen, E. C. The gene scb-1 underlies variation in caenorhabditis elegans chemotherapeutic responses. *G3* (2020).

[70] Zdraljevic, S. *et al.* Natural variation in c. elegans arsenic toxicity is explained by differences in branched chain amino acid metabolism. *Elife* **8** (2019).

[71] Zdraljevic, S. *et al.* Natural variation in a single amino acid substitution underlies physiological responses to topoisomerase II poisons. *PLoS Genet.* **13**, e1006891 (2017).

[72] Brady, S. C. *et al.* A novel gene underlies Bleomycin-Response variation in caenorhabditis elegans. *Genetics* **212**, 1453–1468 (2019).

[73] Evans, K. S. *et al.* Shared genomic regions underlie natural variation in diverse toxin responses. *Genetics* (2018).

[74] Zamanian, M. *et al.* Discovery of genomic intervals that underlie nematode responses to benzimidazoles. *PLoS Negl. Trop. Dis.* **12**, e0006368 (2018).

[75] Hahnel, S. R. *et al.* Extreme allelic heterogeneity at a caenorhabditis elegans beta-tubulin locus explains natural resistance to benzimidazoles. *PLoS Pathog.* **14**, e1007226 (2018).

[76] Lee, D. *et al.* Selection and gene flow shape niche-associated variation in pheromone response. *Nat Ecol Evol* **3**, 1455–1463 (2019).

[77] Shimko, T. C. & Andersen, E. C. COPASutils: an R package for reading, processing, and visualizing data from COPAS large-particle flow cytometers. *PLoS One* **9**, e111090 (2014).

[78] Rockman, M. V. & Kruglyak, L. Recombinational landscape and population genomics of caenorhabditis elegans. *PLoS Genet.* **5**, e1000419 (2009).

[79] Andersen, E. C. *et al.* Chromosome-scale selective sweeps shape caenorhabditis elegans genomic diversity. *Nat. Genet.* **44**, 285–290 (2012).

[80] Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in c. elegans. *PLoS Genet.* **2**, e222 (2006).

[81] 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.* **3**, e34 (2007).

[82] Rockman, M. V. & Kruglyak, L. Breeding designs for recombinant inbred advanced intercross lines. *Genetics* **179**, 1069–1078 (2008).

[83] Darvasi, A. & Soller, M. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**, 1199–1207 (1995).

[84] Palopoli, M. F. *et al.* Molecular basis of the copulatory plug polymorphism in caenorhabditis elegans. *Nature* **454**, 1019–1022 (2008).

[85] Seidel, H. S., Rockman, M. V. & Kruglyak, L. Widespread genetic incompatibility in c. elegans maintained by balancing selection. *Science* **319**, 589–594 (2008).

[86] Seidel, H. S. *et al.* A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in c. elegans. *PLoS Biol.* **9**, e1001115 (2011).

[87] Gaertner, B. E. & Phillips, P. C. Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation. *Genet. Res.* **92**, 331–348 (2010).

[88] Harvey, S. C., Shorto, A. & Viney, M. E. Quantitative genetic analysis of life-history traits of caenorhabditis elegans in stressful environments. *BMC Evol. Biol.* **8**, 15 (2008).

[89] McGrath, P. T. *et al.* Parallel evolution of domesticated caenorhabditis species targets pheromone receptor genes. *Nature* **477**, 321–325 (2011).

[90] Frézal, L., Demoinet, E., Braendle, C., Miska, E. & Félix, M.-A. Natural genetic variation in a multigenerational phenotype in c. elegans. *Curr. Biol.* **28**, 2588–2596.e8 (2018).

[91] Torres Cleuren, Y. N. *et al.* Extensive intraspecies cryptic variation in an ancient embryonic gene regulatory network. *Elife* **8** (2019).

[92] Greene, J. S. *et al.* Balancing selection shapes density-dependent foraging behaviour. *Nature* **539**, 254–258 (2016).

[93] Duveau, F. & Félix, M.-A. Role of pleiotropy in the evolution of a cryptic developmental variation in caenorhabditis elegans. *PLoS Biol.* **10**, e1001230 (2012).

[94] Noble, L. M. *et al.* Natural variation in plep-1 causes Male-Male copulatory behavior in c. elegans. *Curr. Biol.* **25**, 2730–2737 (2015).

[95] Noble, L. M. *et al.* Polygenicity and epistasis underlie Fitness-Proximal traits in the caenorhabditis elegans multiparental experimental evolution (CeMEE) panel. *Genetics* **207**, 1663–1685 (2017).

[96] Snoek, B. L. *et al.* A multi-parent recombinant inbred line population of c. elegans allows identification of novel QTLs for complex life history traits. *BMC Biol.* **17**, 24 (2019).

[97] Steve A Quarrie, Vesna Lazic-Jancic, Dragan Kovacevic, Andy Steed, and Sofija Pekic. Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany* **50**, 1299–1306 (1999).

[98] Ehrenreich, I. M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039–1042 (2010).

[99] Burga, A., Ben-David, E., Lemus Vergara, T., Boocock, J. & Kruglyak, L. Fast genetic mapping of complex traits in c. elegans using millions of individuals in bulk. *Nat. Commun.* **10**, 2680 (2019).

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

[101] Gimond, C. *et al.* Natural variation and genetic determinants of caenorhabditis elegans sperm size. *Genetics* **213**, 615–632 (2019).

[102] Gutteling, E. W. *et al.* Environmental influence on the genetic correlations between life-history traits in caenorhabditis elegans. *Heredity* **98**, 206–213 (2007).

[103] Gutteling, E. W., Riksen, J. A. G., Bakker, J. & Kammenga, J. E. Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in caenorhabditis elegans. *Heredity* **98**, 28–37 (2007).

[104] Large, E. E. *et al.* Selection on a subunit of the NURF chromatin remodeler modifies life history traits in a domesticated strain of caenorhabditis elegans. *PLoS Genet.* **12**, e1006219 (2016).

[105] Zhu, Z. *et al.* Identification of specific nuclear genetic loci and genes that interact with the mitochondrial genome and contribute to fecundity in caenorhabditis elegans. *Front. Genet.* **10**, 28 (2019).

[106] Large, E. E. *et al.* Modeling of a negative feedback mechanism explains antagonistic pleiotropy in reproduction in domesticated caenorhabditis elegans strains. *PLoS Genet.* **13**, e1006769 (2017).

[107] Stastna, J. J., Snoek, L. B., Kammenga, J. E. & Harvey, S. C. Genotype-dependent lifespan effects in peptone deprived caenorhabditis elegans. *Sci. Rep.* **5**, 16259 (2015).

[108] Viñuela, A., Snoek, L. B., Riksen, J. A. G. & Kammenga, J. E. Genome-wide gene expression regulation as a function of genotype and age in c. elegans. *Genome Res.* **20**, 929–937 (2010).

[109] Viñuela, A., Snoek, L. B., Riksen, J. A. G. & Kammenga, J. E. Aging uncouples heritability and Expression-QTL in caenorhabditis elegans. *G3* **2**, 597–605 (2012).

[110] Zhu, Z., Lu, Q., Zeng, F., Wang, J. & Huang, S. Compatibility between mitochondrial and nuclear genomes correlates with the quantitative trait of lifespan in caenorhabditis elegans. *Sci. Rep.* **5**, 17303 (2015).

[111] Webster, A. K. *et al.* Population selection and sequencing of caenorhabditis elegans wild isolates identifies a region on chromosome III affecting starvation resistance. *G3* **9**, 3477–3488 (2019).

[112] Schmid, T. *et al.* Systemic regulation of RAS/MAPK signaling by the serotonin metabolite 5-HIAA. *PLoS Genet.* **11**, e1005236 (2015).

[113] Sterken, M. G. *et al.* Ras/MAPK modifier loci revealed by eQTL in caenorhabditis elegans. *G3* **7**, 3185–3193 (2017).

[114] Li, Y. *et al.* Global genetic robustness of the alternative splicing machinery in caenorhabditis elegans. *Genetics* **186**, 405–410 (2010).

[115] Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in c. elegans. *Science* **330**, 372–376 (2010).

[116] Snoek, B. L. *et al.* Contribution of trans regulatory eQTL to cryptic genetic variation in c. elegans. *BMC Genomics* **18**, 500 (2017).

[117] Singh, K. D. *et al.* Natural genetic variation influences protein abundances in c. elegans developmental signalling pathways. *PLoS One* **11**, e0149418 (2016).

[118] Gao, A. W. *et al.* Natural genetic variation in c. elegans identified genomic loci controlling metabolite levels. *Genome Res.* **28**, 1296–1308 (2018).

[119] Balla, K. M., Andersen, E. C., Kruglyak, L. & Troemel, E. R. A wild c. elegans strain has enhanced epithelial immunity to a natural microsporidian parasite. *PLoS Pathog.* **11**, e1004583 (2015).

[120] Ghosh, R., Andersen, E. C., Shapiro, J. A., Gerke, J. P. & Kruglyak, L. Natural variation in a chloride channel subunit confers avermectin resistance in c. elegans. *Science* **335**, 574–578 (2012).

[121] Nakad, R. *et al.* Contrasting invertebrate immune defense behaviors caused by a single gene, the caenorhabditis elegans neuropeptide receptor gene npr-1. *BMC Genomics* **17**, 280 (2016).

[122] Bernstein, M. R., Zdraljevic, S., Andersen, E. C. & Rockman, M. V. Tightly linked antagonistic-effect loci underlie polygenic phenotypic variation in c. elegans. *Evol Lett* **3**, 462–473 (2019).

[123] Evans, K. S. *et al.* Natural variation in the sequestosome-related gene, sqst-5, underlies zinc homeostasis in caenorhabditis elegans (2020).

[124] Na, H., Zdraljevic, S., Tanny, R. E., Walhout, A. J. M. & Andersen, E. C. Natural variation in a glucuronosyltransferase modulates propionate sensitivity in a c. elegans propionic acidemia model (2020).

[125] Rodriguez, M., Snoek, L. B., Riksen, J. A. G., Bevers, R. P. & Kammenga, J. E. Genetic variation for stress-response hormesis in c. elegans lifespan. *Exp. Gerontol.* **47**, 581–587 (2012).

[126] Elvin, M. *et al.* A fitness assay for comparing RNAi effects across multiple c. elegans genotypes. *BMC Genomics* **12**, 510 (2011).

[127] Bendesky, A., Tsunozaki, M., Rockman, M. V., Kruglyak, L. & Bargmann, C. I. Catecholamine receptor polymorphisms affect decision-making in c. elegans. *Nature* **472**, 313–318 (2011).

[128] Glater, E. E., Rockman, M. V. & Bargmann, C. I. Multigenic natural variation underlies caenorhabditis elegans olfactory preference for the bacterial pathogen serratia marcescens. *G3* **4**, 265–276 (2014).

[129] Harvey, S. C. Non-dauer larval dispersal in caenorhabditis elegans. *J. Exp. Zool. B Mol. Dev. Evol.* **312B**, 224–230 (2009).

[130] McGrath, P. T. *et al.* Quantitative mapping of a digenic behavioral trait implicates globin variation in c. elegans sensory behaviors. *Neuron* **61**, 692–699 (2009).

[131] Gaertner, B. E., Parmenter, M. D., Rockman, M. V., Kruglyak, L. & Phillips, P. C. More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in caenorhabditis elegans. *Genetics* **192**, 1533–1542 (2012).

[132] 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.* **8**, e1003157 (2012).

[133] Lee, D. *et al.* The genetic basis of natural variation in a phoretic behavior. *Nat. Commun.* **8**, 273 (2017).

[134] Cook, D. E. *et al.* The genetic basis of natural variation in caenorhabditis elegans telomere length. *Genetics* **204**, 371–383 (2016).

[135] Laricchia, K. M., Zdraljevic, S., Cook, D. E. & Andersen, E. C. Natural variation in the distribution and abundance of transposable elements across the caenorhabditis elegans species. *Mol. Biol. Evol.* **34**, 2187–2202 (2017).

[136] Evans, K. S. *et al.* Correlations of genotype with climate parameters suggest caenorhabditis elegans niche adaptations. *G3* **7**, 289–298 (2017).

[137] Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75**, 311–335 (2012).

[138] Bock, K. W. The UDP-glycosyltransferase (UGT) superfamily expressed in humans, insects and plants: Animal-plant arms-race and co-evolution. *Biochem. Pharmacol.* **99**, 11–17 (2016).

[139] Russell, R. J. *et al.* The evolution of new enzyme function: lessons from xenobiotic metabolizing bacteria versus insecticide-resistant insects. *Evol. Appl.* **4**, 225–248 (2011).

[140] McCavera, S., Walsh, T. K. & Wolstenholme, A. J. Nematode ligand-gated chloride channels: an appraisal of their involvement in macrocyclic lactone resistance and prospects for developing molecular markers. *Parasitology* **134**, 1111–1121 (2007).

[141] Wolstenholme, A. J., Fairweather, I., Prichard, R., von Samson-Himmelstjerna, G. & Sangster, N. C. Drug resistance in veterinary helminths. *Trends Parasitol.* **20**, 469–476 (2004).

[142] Wallace, J. G., Rodgers-Melnick, E. & Buckler, E. S. On the road to breeding 4.0: Unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* **52**, 421–444 (2018).

[143] Albert, F. W., Bloom, J. S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. *Elife* **7** (2018).

[144] Green, J. W. M., Snoek, L. B., Kammenga, J. E. & Harvey, S. C. Genetic mapping of variation in dauer larvae development in growing populations of caenorhabditis elegans. *Heredity* **111**, 306–313 (2013).

[145] Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).

[146] Ben-David, E., Burga, A. & Kruglyak, L. A maternal-effect selfish genetic element in caenorhabditis elegans. *Science* **356**, 1051–1055 (2017).

[147] Bendesky, A. & Bargmann, C. I. Genetic contributions to behavioural diversity at the gene-environment interface. *Nat. Rev. Genet.* **12**, 809–820 (2011).

[148] Chandler, C. H. Cryptic intraspecific variation in sex determination in caenorhabditis elegans revealed by mutations. *Heredity* **105**, 473–482 (2010).

[149] Snoek, L. B. *et al.* Widespread genomic incompatibilities in caenorhabditis elegans. *G3* **4**, 1813–1823 (2014).

[150] Mardis, E. R. DNA sequencing technologies: 2006-2016. *Nat. Protoc.* **12**, 213–218 (2017).

[151] Easton, D. F., Bishop, D. T., Ford, D. & Crockford, G. P. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium. *Am. J. Hum. Genet.* **52**, 678–701 (1993).

[152] Cowley, A. W., Jr. The genetic dissection of essential hypertension. *Nat. Rev. Genet.* **7**, 829–840 (2006).

[153] Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).

[154] Rothschild, M. F., Hu, Z.-L. & Jiang, Z. Advances in QTL mapping in pigs. *Int. J. Biol. Sci.* **3**, 192–197 (2007).

[155] Johnsson, M., Jonsson, K. B., Andersson, L., Jensen, P. & Wright, D. Quantitative trait locus and genetical genomics analysis identifies putatively causal genes for fecundity and brooding in the chicken. *G3* **6**, 311–319 (2015).

[156] Leal-Bertioli, S. C. M. *et al.* Genetic mapping of resistance to meloidogyne arenaria in arachis stenosperma: A new source of nematode resistance for peanut. *G3* **6**, 377–390 (2015).

[157] Shang, L. *et al.* Main effect QTL with dominance determines heterosis for dynamic plant height in upland cotton. *G3* **6**, 3373–3379 (2016).

[158] Mackay, T. F. Quantitative trait loci in drosophila. *Nat. Rev. Genet.* **2**, 11–20 (2001).

[159] Peng, W. *et al.* An ultra-high density linkage map and QTL mapping for sex and growth-related traits of common carp (cyprinus carpio). *Sci. Rep.* **6**, 26693 (2016).

[160] Simon, M. *et al.* Quantitative trait loci mapping in five new large recombinant inbred line populations of arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers. *Genetics* **178**, 2253–2264 (2008).

[161] Reddy, K. C., Andersen, E. C., Kruglyak, L. & Kim, D. H. A polymorphism in npr-1 is a behavioral determinant of pathogen susceptibility in c. elegans. *Science* **323**, 382–384 (2009).

[162] Keurentjes, J. J. B. *et al.* Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1708–1713 (2007).

[163] Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* **4**, e1000232 (2008).

[164] Hasin-Brumshtein, Y. *et al.* Hypothalamic transcriptomes of 99 mouse strains reveal trans eQTL hotspots, splicing QTLs and novel non-coding genes. *Elife* **5** (2016).

[165] Ehrenreich, I. M. *et al.* Genetic architecture of highly complex chemical resistance traits across four yeast strains. *PLoS Genet.* **8**, e1002570 (2012).

[166] Singh, U. M. *et al.* QTL hotspots for early vigor and related traits under dry Direct-Seeded system in rice (oryza sativa l.). *Front. Plant Sci.* **8**, 286 (2017).

[167] Knoch, D. *et al.* Genetic dissection of metabolite variation in arabidopsis seeds: evidence for mQTL hotspots and a master regulatory locus of seed metabolism. *J. Exp. Bot.* **68**, 1655–1667 (2017).

[168] Najarro, M. A. *et al.* Identifying loci contributing to natural variation in xenobiotic resistance in drosophila. *PLoS Genet.* **11**, e1005663 (2015).

[169] Marriage, T. N., King, E. G., Long, A. D. & Macdonald, S. J. Fine-mapping nicotine resistance loci in drosophila using a multiparent advanced generation inter-cross population. *Genetics* **198**, 45–57 (2014).

[170] Highfill, C. A. *et al.* Naturally segregating variation atUgt86DdContributes to nicotine resistance indrosophila melanogaster. *Genetics* **207**, 311–325 (2017).

[171] Bubier, J. A. *et al.* Identification of a QTL in mus musculus for alcohol preference, withdrawal, and ap3m2 expression using integrative functional genomics and precision genetics. *Genetics* **197**, 1377–1393 (2014).

[172] Crusio, W. E., Dhawan, E., Chesler, E. J. & Delprato, A. Analysis of morphine responses in mice reveals a QTL on chromosome 7. *F1000Res.* **5**, 2156 (2016).

[173] Boyd, W. A., Smith, M. V. & Freedman, J. H. Caenorhabditis elegans as a model in developmental toxicology. *Methods Mol. Biol.* **889**, 15–24 (2012).

[174] García-González, A. P. *et al.* Bacterial metabolism affects the c. elegans response to cancer chemotherapeutics. *Cell* **169**, 431–441.e8 (2017).

[175] R Core Team. R: A language and environment for statistical computing (2017).

[176] Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).

[177] Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1572–1577 (2005).

[178] David Clifford And. The regress function. *R News* **6**, 6–10 (2006).

[179] Cook, D. E. & Andersen, E. C. VCF-kit: assorted utilities for the variant call format. *Bioinformatics* **33**, 1581–1582 (2017).

[180] Dittrich-Reed, D. R. & Fitzpatrick, B. M. Transgressive hybrids as hopeful monsters. *Evol. Biol.* **40**, 310–315 (2013).

[181] Price, N. P. J. & Tsvetanova, B. Biosynthesis of the tunicamycins: a review. *J. Antibiot.* **60**, 485–491 (2007).

[182] Zdraljevic, S. & Andersen, E. C. Natural diversity facilitates the discovery of conserved chemotherapeutic response mechanisms. *Curr. Opin. Genet. Dev.* **47**, 41–47 (2017).

[183] Pommier, Y. Topoisomerase I inhibitors: camptothecins and beyond. *Nat. Rev. Cancer* **6**, 789–802 (2006).

[184] Andersen, E. & Evans, K. S. A single locus underlies variation in caenorhabditis elegans chemotherapeutic responses (2020).

[185] Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).

[186] Tyler, A. L., Crawford, D. C. & Pendergrass, S. A. The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief. Bioinform.* **17**, 13–22 (2016).

[187] Peltier, E., Friedrich, A., Schacherer, J. & Marullo, P. Quantitative trait nucleotides impacting the technological performances of industrial saccharomyces cerevisiae strains. *Front. Genet.* **10**, 683 (2019).

[188] Jerison, E. R. *et al.* Genetic variation in adaptability and pleiotropy in budding yeast. *Elife* **6** (2017).

[189] Cubillos, F. A. *et al.* Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol. Ecol.* **20**, 1401–1413 (2011).

[190] Fusari, C. M. *et al.* Genome-Wide association mapping reveals that specific and pleiotropic regulatory mechanisms Fine-Tune central metabolism and growth in arabidopsis. *Plant Cell* **29**, 2349–2373 (2017).

[191] McKay, J. K., Richards, J. H. & Mitchell-Olds, T. Genetics of drought adaptation in arabidopsis thaliana: I. pleiotropy contributes to genetic correlations among ecological traits. *Mol. Ecol.* **12**, 1137–1151 (2003).

[192] El[U+2010]Assal, S. E. D., Alonso[U+2010]Blanco, C., Hanhart, C. J. & Koornneef, M. Pleiotropic effects of the arabidopsis cryptochrome 2 allelic variation underlie fruit Trait[U+2010]Related QTL. *Plant Biol.* **6**, 370–374 (2004).

[193] McGuigan, K. *et al.* The nature and extent of mutational pleiotropy in gene expression of male drosophila serrata. *Genetics* **196**, 911–921 (2014).

[194] Brown, E. B., Layne, J. E., Zhu, C., Jegga, A. G. & Rollmann, S. M. Genome-wide association mapping of natural variation in odour-guided behaviour in drosophila. *Genes Brain Behav.* **12**, 503–515 (2013).

[195] Leamy, L. J. *et al.* Quantitative trait loci for energy balance traits in an advanced intercross line derived from mice divergently selected for heat loss. *PeerJ* **2**, e392 (2014).

[196] Lin, C. H. S. *et al.* Endostatin and kidney fibrosis in aging: a case for antagonistic pleiotropy? *Am. J. Physiol. Heart Circ. Physiol.* **306**, H1692–9 (2014).

[197] Fisher, R. A. *The genetical theory of natural selection* (The Clarendon Press, 1930).

[198] Orr, H. A. Adaptation and the cost of complexity. *Evolution* **54**, 13–20 (2000).

[199] Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12**, 204–213 (2011).

[200] Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).

[201] Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. *Hum. Genet.* **137**, 39–44 (2018).

[202] Pavlides, J. M. W. *et al.* Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* **8**, 84 (2016).

[203] Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).

[204] Borrello, M. G., Degl'Innocenti, D. & Pierotti, M. A. Inflammation and cancer: the oncogene-driven connection. *Cancer Lett.* **267**, 262–270 (2008).

[205] Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

[206] Riedel, C. G. *et al.* DAF-16 employs the chromatin remodeller SWI/SNF to promote stress resistance and longevity. *Nat. Cell Biol.* **15**, 491–501 (2013).

[207] Zhang, L. *et al.* Structural and biochemical characterization of endoribonuclease nsp15 encoded by middle east respiratory syndrome coronavirus. *J. Virol.* **92** (2018).

[208] Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

[209] Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).

[210] Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R package for causal mediation analysis. *Journal of Statistical Software, Articles* **59**, 1–38 (2014).

[211] MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. Mediation analysis. *Annu. Rev. Psychol.* **58**, 593–614 (2007).

[212] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear Mixed-Effects models using lme4 (2014). 1406.5823.

[213] Dorr, R. T. Bleomycin pharmacology: mechanism of action and resistance, and clinical pharmacokinetics. *Semin. Oncol.* **19**, 3–8 (1992).

[214] Dasari, S. & Tchounwou, P. B. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur. J. Pharmacol.* **740**, 364–378 (2014).

[215] Montecucco, A., Zanetta, F. & Biamonti, G. Molecular mechanisms of etoposide. *EXCLI J.* **14**, 95–108 (2015).

[216] Ketron, A. C., Denny, W. A., Graves, D. E. & Osheroff, N. Amsacrine as a topoisomerase II poison: importance of drug-DNA interactions. *Biochemistry* **51**, 1730–1739 (2012).

[217] Nikolova, T., Roos, W. P., Krämer, O. H., Strik, H. M. & Kaina, B. Chloroethylating nitrosoureas in cancer therapy: DNA damage, repair and cell death signaling. *Biochim. Biophys. Acta* **1868**, 29–39 (2017).

[218] Azzam, M. E. & Algranati, I. D. Mechanism of puromycin action: fate of ribosomes after release of nascent protein chains from polysomes. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3866–3869 (1973).

[219] Piperdi, B., Ling, Y.-H., Liebes, L., Muggia, F. & Perez-Soler, R. Bortezomib: understanding the mechanism of action. *Mol. Cancer Ther.* **10**, 2029–2030 (2011).

[220] Kaplan, A., Akalin Ciftci, G. & Kutlu, H. M. Cytotoxic, anti-proliferative and apoptotic effects of silver nitrate against h-ras transformed 5RP7. *Cytotechnology* **68**, 1727–1735 (2016).

[221] Sasaki, E., Frommlet, F. & Nordborg, M. GWAS with heterogeneous data: Estimating the fraction of phenotypic variation mediated by gene expression data. *G3* **8**, 3059–3068 (2018).

[222] Huang, Y.-T., Liang, L., Moffatt, M. F., Cookson, W. O. C. M. & Lin, X. iGWAS: Integrative Genome-Wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genet. Epidemiol.* **39**, 347–356 (2015).

[223] Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644 (2014).

[224] Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).

[225] Harrell, K. E., Koury, E. & Smolikove, S. Microirradiation for precise, double-strand break induction in vivo in caenorhabditis elegans. *Bio Protoc* **8** (2018).

[226] Rosen, B. P. Metals in biology: past, present, and future. In Tamas, M. J. & Martinoia, E. (eds.) *Molecular Biology of Metal Homeostasis and Detoxification: From Microbes to Man*, 485–505 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).

[227] Oleszkiewicz, J. A. & Sharma, V. K. Stimulation and inhibition of anaerobic processes by heavy metals—a review. *Biological Wastes* **31**, 45–67 (1990).

[228] Vallee, B. L. & Falchuk, K. H. The biochemical basis of zinc physiology. *Physiol. Rev.* **73**, 79–118 (1993).

[229] Frederickson, C. J. & Bush, A. I. Synaptically released zinc: physiological functions and pathological effects. *Biometals* **14**, 353–366 (2001).

[230] Colvin, R. A., Fontaine, C. P., Laskowski, M. & Thomas, D. Zn2+ transporters and zn2+ homeostasis in neurons. *Eur. J. Pharmacol.* **479**, 171–185 (2003).

[231] Wall, M. J. A role for zinc in cerebellar synaptic transmission? *Cerebellum* **4**, 224–229 (2005).

[232] Yamasaki, S. *et al.* Zinc is a novel intracellular second messenger. *J. Cell Biol.* **177**, 637–645 (2007).

[233] Dietrich, N., Tan, C.-H., Cubillas, C., Earley, B. J. & Kornfeld, K. Insights into zinc and cadmium biology in the nematode caenorhabditis elegans. *Arch. Biochem. Biophys.* **611**, 120–133 (2016).

[234] Yoder, J. H., Chong, H., Guan, K.-I. & Han, M. Modulation of KSR activity in caenorhabditis elegans by zn ions, PAR-1 kinase and PP2A phosphatase. *EMBO J.* **23**, 111–119 (2004).

[235] Bruinsma, J. J., Jirakulaporn, T., Muslin, A. J. & Kornfeld, K. Zinc ions and cation diffusion facilitator proteins regulate ras-mediated signaling. *Dev. Cell* **2**, 567–578 (2002).

[236] Samet, J. M., Dewar, B. J., Wu, W. & Graves, L. M. Mechanisms of zn(2+)-induced signal initiation through the epidermal growth factor receptor. *Toxicol. Appl. Pharmacol.* **191**, 86–93 (2003).

[237] Wu, W. *et al.* Activation of the EGF receptor signaling pathway in human airway epithelial cells exposed to metals. *Am. J. Physiol.* **277**, L924–31 (1999).

[238] Roohani, N., Hurrell, R., Kelishadi, R. & Schulin, R. Zinc and its importance for human health: An integrative review. *J. Res. Med. Sci.* **18**, 144–157 (2013).

[239] Fernandes, G. *et al.* Impairment of cell-mediated immunity functions by dietary zinc deficiency in mice. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 457–461 (1979).

[240] Zhao, X.-Q. & Bai, F.-W. Zinc and yeast stress tolerance: micronutrient plays a big role. *J. Biotechnol.* **158**, 176–183 (2012).

[241] Navarro, J. A. & Schneuwly, S. Copper and zinc homeostasis: Lessons from drosophila melanogaster. *Front. Genet.* **8**, 223 (2017).

[242] Fosmire, G. J. Zinc toxicity. *Am. J. Clin. Nutr.* **51**, 225–227 (1990).

[243] Warnhoff, K. *et al.* The nuclear receptor HIZR-1 uses zinc as a ligand to mediate homeostasis in response to high zinc. *PLoS Biol.* **15**, e2000094 (2017).

[244] Zhao, Y. *et al.* The zinc transporter ZIPT-7.1 regulates sperm activation in nematodes. *PLoS Biol.* **16**, e2005069 (2018).

[245] Dietrich, N., Schneider, D. L. & Kornfeld, K. A pathway for low zinc homeostasis that is conserved in animals and acts in parallel to the pathway for high zinc homeostasis. *Nucleic Acids Res.* **45**, 11658–11672 (2017).

[246] Davis, D. E. *et al.* The cation diffusion facilitator gene cdf-2 mediates zinc metabolism in caenorhabditis elegans. *Genetics* **182**, 1015–1033 (2009).

[247] Andreini, C., Banci, L., Bertini, I. & Rosato, A. Zinc through the three domains of life. *J. Proteome Res.* **5**, 3173–3178 (2006).

[248] Bruinsma, J. J., Schneider, D. L., Davis, D. E. & Kornfeld, K. Identification of mutations in caenorhabditis elegans that cause resistance to high levels of dietary zinc and analysis using a genomewide map of single nucleotide polymorphisms scored by pyrosequencing. *Genetics* **179**, 811–828 (2008).

[249] Murphy, J. T. *et al.* Histidine protects against zinc and nickel toxicity in caenorhabditis elegans. *PLoS Genet.* **7**, e1002013 (2011).

[250] Warnhoff, K. *et al.* The DAF-16 FOXO transcription factor regulates natc-1 to modulate stress resistance in caenorhabditis elegans, linking insulin/IGF-1 signaling to protein n-terminal acetylation. *PLoS Genet.* **10**, e1004703 (2014).

[251] Wang, Y. & Ezemaduka, A. N. Combined effect of temperature and zinc on caenorhabditis elegans wild type and daf-21 mutant strains. *J. Therm. Biol.* **41**, 16–20 (2014).

[252] Roh, H. C. *et al.* ttm-1 encodes CDF transporters that excrete zinc from intestinal cells of c. elegans and act in a parallel negative feedback circuit that promotes homeostasis. *PLoS Genet.* **9**, e1003522 (2013).

[253] Roh, H. C., Collier, S., Guthrie, J., Robertson, J. D. & Kornfeld, K. Lysosome-related organelles in intestinal cells are a zinc storage site in c. elegans. *Cell Metab.* **15**, 88–99 (2012).

[254] Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

[255] Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

[256] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

[257] Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* **7**, 539 (2011).

[258] Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).

[259] Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–7 (2005).

[260] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

[261] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

[262] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

[263] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

[264] Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).

[265] Ponting, C. P., Blake, D. J., Davies, K. E., Kendrick-Jones, J. & Winder, S. J. ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins. *Trends Biochem. Sci.* **21**, 11–13 (1996).

[266] Legge, G. B. *et al.* ZZ domain of CBP: an unusual zinc finger fold in a protein interaction module. *J. Mol. Biol.* **343**, 1081–1093 (2004).

[267] Metzger, B. P. H. & Wittkopp, P. J. Compensatory trans-regulatory alleles minimizing variation in TDH3 expression are common within saccharomyces cerevisiae. *Evol Lett* **3**, 448–461 (2019).

[268] Shao, H. *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19910–19914 (2008).

[269] Kroymann, J. & Mitchell-Olds, T. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**, 95–98 (2005).

[270] Steinmetz, L. M. *et al.* Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).

[271] Kumsta, C. *et al.* The autophagy receptor p62/SQST-1 promotes proteostasis and longevity in c. elegans by inducing autophagy. *Nat. Commun.* **10**, 5648 (2019).

[272] Gatica, D., Lahiri, V. & Klionsky, D. J. Cargo recognition and degradation by selective autophagy. *Nat. Cell Biol.* **20**, 233–242 (2018).

[273] Zhang, Y., Mi, W., Xue, Y., Shi, X. & Kutateladze, T. G. The ZZ domain as a new epigenetic reader and a degradation signal sensor. *Crit. Rev. Biochem. Mol. Biol.* **54**, 1–10 (2019).

[274] Cha-Molstad, H. *et al.* p62/SQSTM1/Sequestosome-1 is an n-recognin of the n-end rule pathway which modulates autophagosome biogenesis. *Nat. Commun.* **8**, 102 (2017).

[275] Cui, Y., McBride, S. J., Boyd, W. A., Alper, S. & Freedman, J. H. Toxicogenomic analysis of caenorhabditis elegans reveals novel genes and pathways involved in the resistance to cadmium toxicity. *Genome Biol.* **8**, R122 (2007).

[276] Hirano, S. & Kanno, S. Relevance of autophagy markers to cytotoxicity of zinc compounds in macrophages. *Toxicol. In Vitro* **65**, 104816 (2020).

[277] Lee, D. *et al.* Balancing selection maintains ancient genetic diversity in c. elegans (2020).

[278] Baesler, J. *et al.* Zn homeostasis in genetic models of parkinson's disease in caenorhabditis elegans. *J. Trace Elem. Med. Biol.* **55**, 44–49 (2019).

[279] American cancer society .

[280] Reference, G. H. What is pharmacogenomics? `https://ghr.nlm.nih.gov/primer/genomicresearch/pharmacogenomics`. Accessed: 2017-4-2.

[281] FDA approval for docetaxel. `https://www.cancer.gov/about-cancer/treatment/drugs/fda-docetaxel`. Accessed: 2017-4-9.

[282] Herbst, R. S. & Khuri, F. R. Mode of action of docetaxel – a basis for combination with novel anticancer agents. *Cancer Treat. Rev.* **29**, 407–415 (2003).

[283] Choi, J. R. *et al.* Genetic variations of drug transporters can influence on drug response in patients treated with docetaxel chemotherapy. *Cancer Res. Treat.* **47**, 509–517 (2015).

[284] Jabir, R. S. *et al.* Pharmacogenetics of taxanes: impact of gene polymorphisms of drug transporters on pharmacokinetics and toxicity. *Pharmacogenomics* **13**, 1979–1988 (2012).

[285] Lewis, L. D. *et al.* The relationship of polymorphisms in ABCC2 and SLCO1B3 with docetaxel pharmacokinetics and neutropenia: CALGB 60805 (alliance). *Pharmacogenet. Genomics* **23**, 29–33 (2013).

[286] Baker, S. D., Sparreboom, A. & Verweij, J. Clinical pharmacokinetics of docetaxel : recent developments. *Clin. Pharmacokinet.* **45**, 235–252 (2006).

[287] Lehmann, J. M. *et al.* The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. *J. Clin. Invest.* **102**, 1016–1023 (1998).

[288] Miyoshi, Y. *et al.* Prediction of response to docetaxel by CYP3A4 mRNA expression in breast cancer tissues. *Int. J. Cancer* **97**, 129–132 (2002).

[289] Tran, A. *et al.* Pharmacokinetics and toxicity of docetaxel: role of CYP3A, MDR1, and GST polymorphisms. *Clin. Pharmacol. Ther.* **79**, 570–580 (2006).

[290] Oishi, K., Okano, H. & Sawa, H. RMD-1, a novel microtubule-associated protein, functions in chromosome segregation in caenorhabditis elegans. *J. Cell Biol.* **179**, 1149–1162 (2007).

[291] Harris, R. Z., Jang, G. R. & Tsunoda, S. Dietary effects on drug metabolism and transport. *Clin. Pharmacokinet.* **42**, 1071–1088 (2003).

[292] Hotez, P. J. *et al.* The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. *PLoS Negl. Trop. Dis.* **8**, e2865 (2014).

[293] Kahn, L. P. & Woodgate, R. G. Integrated parasite management: products for adoption by the australian sheep industry. *Vet. Parasitol.* **186**, 58–64 (2012).

[294] Sutherland, I. A. & Leathwick, D. M. Anthelmintic resistance in nematode parasites of cattle: a global issue? *Trends Parasitol.* **27**, 176–181 (2011).

[295] Lustigman, S. *et al.* A research agenda for helminth diseases of humans: the problem of helminthiases. *PLoS Negl. Trop. Dis.* **6**, e1582 (2012).

[296] Charlier, J., van der Voort, M., Kenyon, F., Skuce, P. & Vercruysse, J. Chasing helminths and their economic impact on farmed ruminants. *Trends Parasitol.* **30**, 361–367 (2014).

[297] Yadav, B., Wennerberg, K., Aittokallio, T. & Tang, J. Searching for drug synergy in complex Dose–Response landscapes using an interaction potency model. *Comput. Struct. Biotechnol. J.* **13**, 504–513 (2015).

[298] Rose, H. *et al.* Widespread anthelmintic resistance in european farmed ruminants: a systematic review. *Vet. Rec.* **176**, 546 (2015).

[299] Crook, E. K. *et al.* Prevalence of anthelmintic resistance on sheep and goat farms in the mid-atlantic region and comparison of in vivo and in vitro detection methods. *Small Rumin. Res.* **143**, 89–96 (2016).

[300] Howell, S. B. *et al.* Prevalence of anthelmintic resistance on sheep and goat farms in the southeastern united states. *J. Am. Vet. Med. Assoc.* **233**, 1913–1919 (2008).

[301] Doyle, S. R. *et al.* Population genomic and evolutionary modelling analyses reveal a single major QTL for ivermectin drug resistance in the pathogenic nematode, haemonchus contortus. *BMC Genomics* **20**, 218 (2019).

[302] Choi, Y.-J. *et al.* Genomic introgression mapping of field-derived multiple-anthelmintic resistance in teladorsagia circumcincta. *PLoS Genet.* **13**, e1006857 (2017).

[303] Stear, M. J., Boag, B., Cattadori, I. & Murphy, L. Genetic variation in resistance to mixed, predominantly teladorsagia circumcincta nematode infections of sheep: from heritabilities to gene identification. *Parasite Immunol.* **31**, 274–282 (2009).

[304] Doyle, S. R. *et al.* Genome-wide analysis of ivermectin response by onchocerca volvulus reveals that genetic drift and soft selective sweeps contribute to loss of drug sensitivity. *PLoS Negl. Trop. Dis.* **11**, e0005816 (2017).

[305] Hunt, P. W. *et al.* The use of DNA markers to map anthelmintic resistance loci in an intraspecific cross of haemonchus contortus. *Parasitology* **137**, 705–717 (2010).

[306] Dilks, C. M. *et al.* Quantitative benzimidazole resistance and fitness effects of parasitic nematode beta-tubulin alleles (2020).

[307] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

[308] Horoszok, L., Raymond, V., Sattelle, D. B. & Wolstenholme, A. J. GLC-3: a novel fipronil and BIDN-sensitive, but picrotoxinin-insensitive, l-glutamate-gated chloride channel subunit from caenorhabditis elegans. *Br. J. Pharmacol.* **132**, 1247–1254 (2001).

[309] Redman, E. *et al.* The emergence of resistance to the benzimidazole anthlemintics in parasitic nematodes of livestock is characterised by multiple independent hard and soft selective sweeps. *PLoS Negl. Trop. Dis.* **9**, e0003494 (2015).

[310] Gilleard, J. S. & Redman, E. Chapter two - genetic diversity and population structure of haemonchus contortus. In Gasser, R. B. & Samson-Himmelstjerna, G. V. (eds.) *Advances in Parasitology*, vol. 93, 31–68 (Academic Press, 2016).

[311] Frézal, L. & Félix, M.-A. C. elegans outside the petri dish. *Elife* **4** (2015).

[312] Campbell, W. C. History of avermectin and ivermectin, with notes on the history of other macrocyclic lactone antiparasitic agents. *Curr. Pharm. Biotechnol.* **13**, 853–865 (2012).

[313] Campbell, W. C. Serendipity and new drugs for infectious disease. *ILAR J.* **46**, 352–356 (2005).

[314] Alivisatos, S. G., Lamantia, L. & Matijevitch, B. L. Imidazolytic processes. VI. enzymic formation of benzimidazole and 5,6-dimethylbenzimidazole containing dinucleotides. *Biochim. Biophys. Acta* **58**, 209–217 (1962).

[315] Cycoń, M., Mrozik, A. & Piotrowska-Seget, Z. Bioaugmentation as a strategy for the remediation of pesticide-polluted soil: A review. *Chemosphere* **172**, 52–71 (2017).

[316] Taube, J., Vorkamp, K., Förster, M. & Herrmann, R. Pesticide residues in biological waste. *Chemosphere* **49**, 1357–1365 (2002).

[317] Hutchinson, G. E. Cold spring harbor symposium on quantitative biology. *Concluding remarks* (1957).

[318] Saltz, J. B. & Nuzhdin, S. V. Genetic variation in niche construction: implications for development and evolutionary genetics. *Trends Ecol. Evol.* **29**, 8–14 (2014).

[319] Taylor, M. L., Skeats, A., Wilson, A. J., Price, T. A. R. & Wedell, N. Opposite environmental and genetic influences on body size in north american drosophila pseudoobscura. *BMC Evol. Biol.* **15**, 51 (2015).

[320] Bozinovic, F., Medina, N. R., Alruiz, J. M., Cavieres, G. & Sabat, P. Thermal tolerance and survival responses to scenarios of experimental climatic change: changing thermal variability reduces the heat and cold tolerance in a fly. *J. Comp. Physiol. B* **186**, 581–587 (2016).

[321] Machado, H. E. *et al.* Comparative population genomics of latitudinal variation in drosophila simulans and drosophila melanogaster. *Mol. Ecol.* **25**, 723–740 (2016).

[322] Tyukmaeva, V. I. *et al.* Localization of quantitative trait loci for diapause and other photoperiodically regulated life history traits important in adaptation to seasonally varying environments. *Mol. Ecol.* **24**, 2809–2819 (2015).

[323] Hangartner, S. B., Hoffmann, A. A., Smith, A. & Griffin, P. C. A collection of australian drosophila datasets on climate adaptation and species distributions. *Sci Data* **2**, 150067 (2015).

[324] Gerken, A. R., Eller, O. C., Hahn, D. A. & Morgan, T. J. Constraints, independence, and evolution of thermal plasticity: probing genetic architecture of long- and short-term thermal acclimation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4399–4404 (2015).

[325] Branham, S. E., Wright, S. J., Reba, A., Morrison, G. D. & Linder, C. R. Genome-Wide association study in arabidopsis thaliana of natural variation in seed oil melting point: A widespread adaptive trait in plants. *J. Hered.* **107**, 257–265 (2016).

[326] Ågrena, J., Oakley, C. G., McKay, J. K., Lovell, J. T. & Schemske, D. W. Genetic mapping of adaptation reveals fitness tradeoffs in arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 21077–21082 (2013).

[327] Fournier-Level, A. *et al.* A map of local adaptation in arabidopsis thaliana. *Science* **334**, 86–89 (2011).

[328] Félix, M.-A. & Braendle, C. The natural history of caenorhabditis elegans. *Curr. Biol.* **20**, R965–9 (2010).

[329] Hodgkin, J. & Doniach, T. Natural variation and copulatory plug formation in caenorhabditis elegans. *Genetics* **146**, 149–164 (1997).

[330] Petrella, L. N. Natural variants of c. elegans demonstrate defects in both sperm function and oogenesis at elevated temperatures. *PLoS One* **9**, e112377 (2014).

[331] Anderson, J. L., Albergotti, L., Ellebracht, B., Huey, R. B. & Phillips, P. C. Does thermoregulatory behavior maximize reproductive fitness of natural isolates of caenorhabditis elegans? *BMC Evol. Biol.* **11**, 157 (2011).

[332] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

[333] Li, H. *et al.* The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[334] Hijmans, R. J., Williams, E. & Vennes, C. geosphere: Spherical trigonometry. R package version 1.2–28. *CRAN. R-project. org/package= geosphere* (2012).

[335] Iannone, R. stationary: Get hourly meteorological data from global stations. R package version 0.4.1. *CRAN. R-project. org/package= stationaRy* (2015).

[336] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

[337] Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

[338] Integrated surface global hourly data - NOAA data catalog. https://data.noaa.gov/dataset/integrated-surface-global-hourly-data. Accessed: 2016-9-2.

[339] Volkers, R. J. M. *et al.* Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild caenorhabditis elegans populations. *BMC Biol.* **11**, 93 (2013).

[340] Sadhu, M. J., Bloom, J. S., Day, L. & Kruglyak, L. CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science* **352**, 1113–1116 (2016).

[341] Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* **9**, e95923 (2014).

[342] Weighill, D. *et al.* Multi-Phenotype association decomposition: Unraveling complex Gene-Phenotype relationships. *Front. Genet.* **10**, 417 (2019).

[343] Oladzad, A. *et al.* Single and multi-trait GWAS identify genetic factors associated with production traits in common bean under abiotic stress environments. *G3* **9**, 1881–1892 (2019).

[344] Dong, J., Boyd, W. A. & Freedman, J. H. Molecular characterization of two homologs of the caenorhabditis elegans cadmium-responsive gene cdr-1: cdr-4 and cdr-6. *J. Mol. Biol.* **376**, 621–633 (2008).

[345] Dong, J., Song, M. O. & Freedman, J. H. Identification and characterization of a family of caenorhabditis elegans genes that is homologous to the cadmium-responsive gene cdr-1. *Biochim. Biophys. Acta* **1727**, 16–26 (2005).

[346] Hall, J., Haas, K. L. & Freedman, J. H. Role of MTL-1, MTL-2, and CDR-1 in mediating cadmium sensitivity in caenorhabditis elegans. *Toxicol. Sci.* **128**, 418–426 (2012).

Appendices

A Co-authored publications

I had the pleasure of collaborating with members of my own laboratory as well as members of other laboratories on several papers. Here, I describe the papers I co-authored.

A.1 Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in *C. elegans*

Chuna Kim[1,2], Jun Kim[2,3], Sunghyun Kim[1,4], Daniel E. Cook[5], **Kathryn S. Evans**[5], Erik C. Andersen[5], and Junho Lee[1,2,3]

[1]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea 08826
[2]Department of Biological Sciences, Seoul National University, Seoul, Korea 08826
[3]Research Institute of Basic Sciences, Seoul National University, Seoul, Korea 08826
[4]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA
[5]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208, USA

This manuscript was published in Genome Research in May 2019 [62].

A.1.1 ABSTRACT

Long-read sequencing technologies have contributed greatly to comparative genomics among species and can also be applied to study genomics within a species. In this study, to determine how substantial genomic changes are generated and tolerated within a species, we sequenced a *C. elegans* strain, CB4856, which is one of the most genetically divergent strains compared to the N2 reference strain. For this comparison, we used the Pacific Biosciences (PacBio) RSII platform (80$\times$, N50 read length 11.8 kb) and generated de novo genome assembly to the level of pseudochromosomes containing 76 contigs (N50 contig = 2.8 Mb). We identified structural variations that affected as many as 2694 genes, most of which are at chromosome arms. Subtelomeric regions contained the most extensive genomic rearrangements, which even created new subtelomeres in

some cases. The subtelomere structure of Chromosome VR implies that ancestral telomere damage was repaired by alternative lengthening of telomeres even in the presence of a functional telomerase gene and that a new subtelomere was formed by break-induced replication. Our study demonstrates that substantial genomic changes including structural variations and new subtelomeres can be tolerated within a species, and that these changes may accumulate genetic diversity within a species.

A.1.2 CONTRIBUTIONS

I analyzed the copy number of the TALT region in all wild strains and created a phylogenetic tree for this region to show that all strains with high copy number are closely related (Figure 4C, E)

A.2 Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations

Tim A. Crombie[1], Stefan Zdraljevic[1,2], Daniel E. Cook[,21], Robyn E. Tanny[1], Shannon C. Brady[1,2], Ye Wang[1], **Kathryn S. Evans**[1,2], Steffen Hahnel[1], Daehan Lee[1], Briana C. Rodriguez[1], Gaotian Zhang[1], Joost van der Zwagg[1], Karin Kiontke[3], Erik C. Andersen[1]

[1]Department of Molecular Biosciences, Northwestern University, Evanston, United States
[2]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, United States
[3]Department of Biology, New York University, New York, United States

This manuscript was published in eLIFE in December 2019 [59].

A.2.1 ABSTRACT

Hawaiian isolates of the nematode species *Caenorhabditis elegans* have long been known to harbor genetic diversity greater than the rest of the worldwide population, but this observation was supported by only a small number of wild strains. To better characterize the niche and genetic diversity of Hawaiian *C. elegans* and other *Caenorhabditis* species, we sampled different substrates and niches across the Hawaiian islands. We identified hundreds of new *Caenorhabditis* strains from known species and a new species, *Caenorhabditis oiwi*. Hawaiian *C. elegans* are found in cooler climates at

high elevations but are not associated with any specific substrate, as compared to other *Caenorhabditis* species. Surprisingly, admixture analysis revealed evidence of shared ancestry between some Hawaiian and non-Hawaiian *C. elegans* strains. We suggest that the deep diversity we observed in Hawaii might represent patterns of ancestral genetic diversity in the *C. elegans* species before human influence.

A.2.2 CONTRIBUTIONS

I was part of the original team to plan the trip to Hawaii and collect samples in August 2017. I further worked with the members of the Andersen lab to clean and analyze data from collections, including looking for enrichment of *C. elegans* on specific substrate types (Figure 2) and at different temperature and elevation (Figure 3)

B R Shiny application development

I learned that one of my favorite parts of lab was writing code to make pretty figures or writing functions to help make analyzing data easier, both for me and for the lab! Here, I describe some of the R Shiny applications I designed.

B.1 High-throughput assay dilutions

B.1.1 PREFACE

I created this shiny application to help members of the lab (including myself) easily and accurately perform drug dilutions for the high-throughput sorter assay. When I first joined the lab, Erik ran a script that performed the necessary calculations for each person before every assay. Over time, it became necessary to calculate my own dilutions. I first created a markdown script that eventually evolved to this user-friendly application. The code used to generate this app can be found here: https://github.com/AndersenLab/HTA_dilutions.

B.1.2 EXPLANATION OF FUNCTIONALITY

An R shiny web app (version 1.4.0.2) was developed to calculate drug dilutions for the high-throughput drug-response assays and can be found here: https://katiesevans9.shinyapps.io/HTA_V3_dilution/. The user selects the type of assay (V2 or V3) and the type of food (lysate or live bacteria), indicates whether or not they will be performing a dose response (serial dose response or regular) with several drug doses or a standard assay with just one drug dose, and sets the minimum volume to pipette and the number of drugs in the assay (**Figure B-1A**). Once these settings are complete, the user will press the “Setup Complete” button and a new window will appear to the right under the “setup” tab. In the setup tab, the user will input the information for the assay including the name and stock concentration of the drug, the final concentration(s) of the drug, and the number of plates to be run in the assay (**Figure B-1B**).

When complete, the user pushes the “Calculate” button and the application automatically calculates the drug dilutions based on pre-set specifications and transfers the user to the “dilution” tab. Here, the application details the necessary dilutions for 1) food, 2) drug stock and 3) drug plate dilutions. A

printable PDF can be generated by clicking the “Download Dilutions” button at the bottom (**Figure B-1C**).

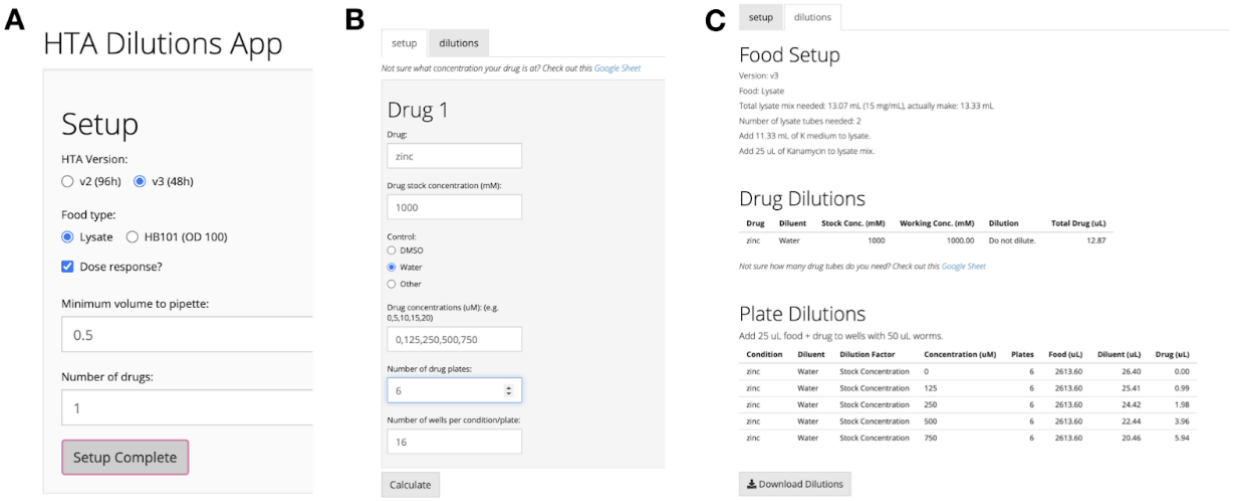


Figure B-1: Screenshots of the HTA dilutions app.

B.2 Fine-map QTL with NIL phenotypes

B.2.1 PREFACE

I wanted to create an application that could be used to easily view and interact with NIL phenotype data generated from the sorter. Every time that I ran an assay I had to look at 5-20 different traits, in both control and drug conditions and drug conditions with and without regression. It soon became an exhausting task. I developed this app to help myself and others analyze NIL data quickly and use the “show QTL?” button to narrow a complex NIL interval. The code used to generate this app can be found here: https://github.com/katiesevans/finemap_NIL.

B.2.2 EXPLANATION OF FUNCTIONALITY

An R shiny web app (version 1.4.0.2) was developed to visualize the results from the high-throughput assays and can be found here: https://katiesevans9.shinyapps.io/QTL_NIL/. To begin analysis, the user can find all data controls in a panel on the left-hand side of the screen. A test dataset is provided (user

should check “Use sample data” checkbox) or the user can upload a file from their local computer. Input files should be in the R data file format (.Rda or .RData) or a CSV file and should be the pruned output from the easysorter pipeline. It is important that the “condition” column of the dataframe contains both a drug and a control and the “control” column contains either the control or “None” (for the control of the control) [77]. Control regression will be performed in the application. The user should select the control condition and the drug condition. In some cases, the user might want to further choose a specific assay to view if multiple options are available. The user also has the option to view one of many drug-response traits by selecting a trait from the drop-down menu. Finally, the user should choose a chromosome to view, generally the chromosome which contains the highlighted QTL for that particular assay (**Figure B-2**).

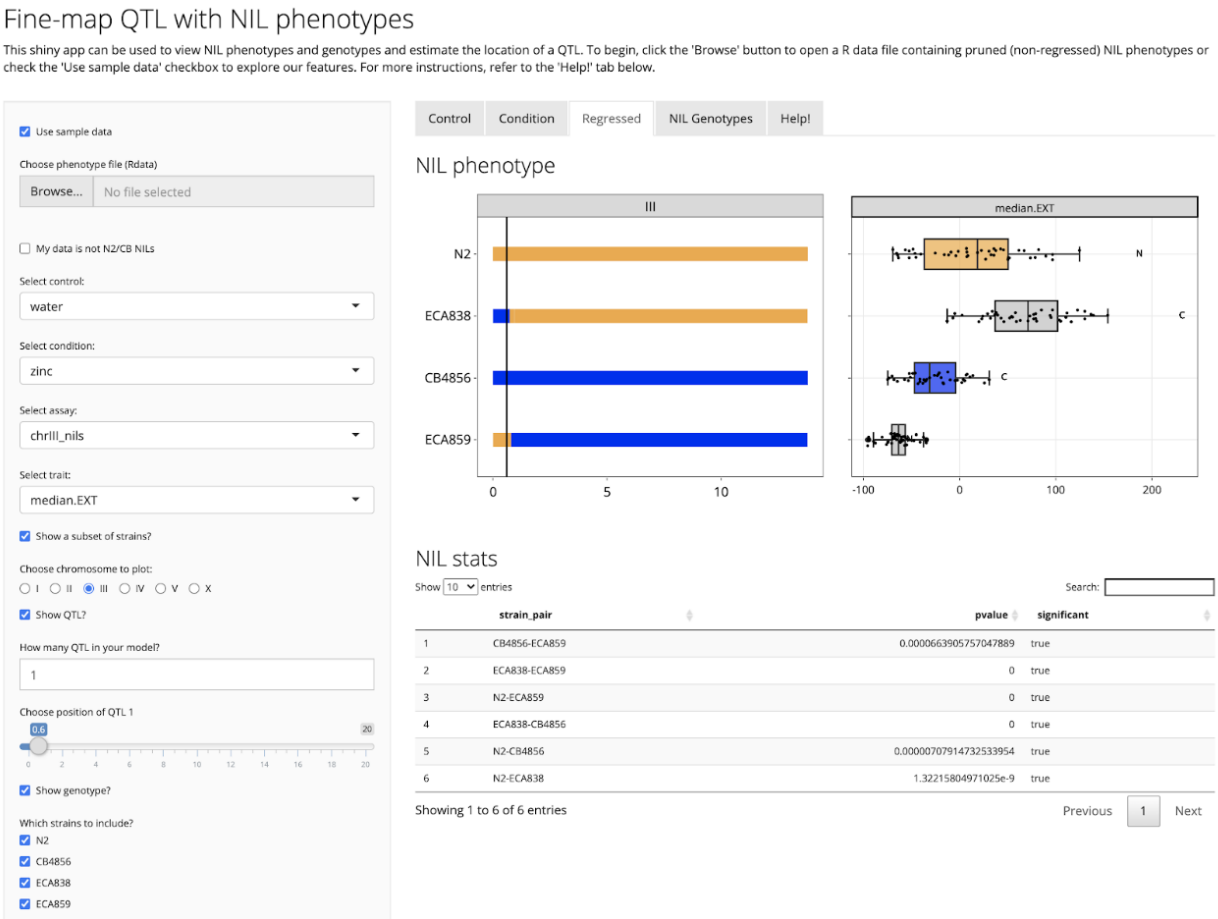


Figure B-2: Screenshot of the Finemap QTL Shiny application.

Along the top of the main panel, the user can navigate several tabs including “Control”, “Condition”, “Regressed”, “NIL Genotypes”, and “Help!”. The “Control”, “Condition”, and “Regressed” tabs each show the NIL genotypes along the selected chromosome (left) and the NIL phenotypes for the selected trait (right) in the control condition (control), raw drug condition (condition), or regressed drug condition (regressed). Genotypes for strains that are not NILs generated from the N2 and CB4856 strains (such as NILs generated from different parents or CRISPR-generated deletion strains) are represented by a grey bar. The user can hover their mouse above the NIL genotype or phenotype plots to see more information or zoom in on a specific area of the plot. Below this plot is an interactive datatable containing the pairwise strain comparisons for this condition and trait. The “NIL Genotypes” tab shows a plot of the NIL genotypes across all chromosomes, not just the chromosome selected by the user (top) and an interactive datatable with the genotypes of each strain across all chromosomes (bottom). The final tab, “Help!” provides the user with the instructions detailed here to help them use the Shiny App (**Figure B-2**).

In addition to these basic controls, the user also has access to several advanced features. The user can choose a subset of strains to plot by checking the box labeled “Show a subset of strains?” and unchecking the boxes next to strains the user wishes to omit. Additionally, the user can plot the location of one or more QTL as a vertical line on the NIL genotype plot by checking the “Show QTL?” box. The user then chooses how many QTL to show and uses the appropriate slider input below to designate the genomic positions of each QTL. Finally, if the “Show genotype?” box is checked, the genotype of each strain at each QTL position will be shown on the phenotype plot as an orange “N” representing N2 and a blue “C” representing CB4856 (**Figure B-2**).

B.3 PCR calculator

B.3.1 PREFACE

We do a LOT of PCR in the Andersen lab and instead of having to calculate how much “master mix” to make for each different-sized PCR by hand (or using *EXCEL*), I decided to design a shiny app that could be used on my phone at my bench for extra ease! It is extremely useful for me, probably the most

useful thing I have done in graduate school (and it only took ∼30 minutes!).

B.3.2 EXPLANATION OF FUNCTIONALITY

An R shiny web app (version 1.4.0.2) was developed for calculating PCR reagents. For ease of use, this app (found here: https://katiesevans9.shinyapps.io/pcr_calculator/) can be accessed via phone or computer. User simply inputs 1) the exact number of lysis reactions, 2) the number of different primer sets and 3) the volume of the PCR. Optionally, the user can change the volume of lysis to add to each PCR (default is 1 uL lysis in 20 uL PCR volume).

After clicking the "Calculate" button, the application will output the amount of all reagents needed to make a "master mix" (containing everything but the primers). Below, the application outputs the "Primer mix". The user will add the "mix" from the finished "master mix" above plus both primers unique to that PCR. In the example below I indicated there were two primer sets, so I will have one large master mix and two different primer mixes. Importantly, the application automatically adjusts the final volumes to allow for pipetting error. The instructions at the bottom remind the user how much lysis and primer mix to add for the final PCR product (**Figure B-3**).

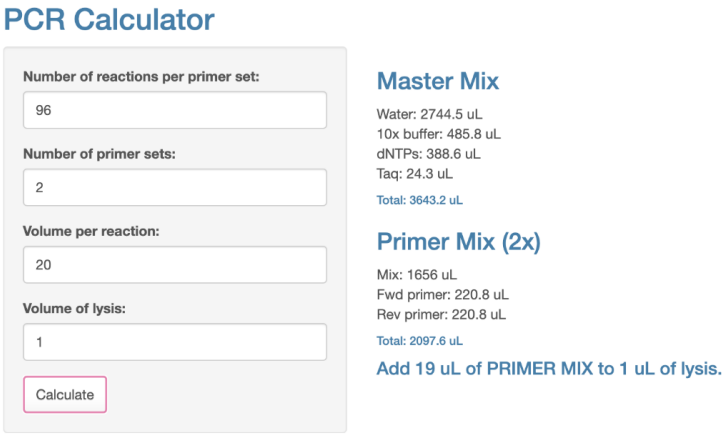


Figure B-3: Screenshot of the PCR Calculator Shiny application.

C Expression of the cadmium-responsive gene, *cdr-6*, does not influence responses to high exogenous zinc in *Caenorhabditis elegans*

C.1 PREFACE

As part of the zinc project for Chapter 4, I tested the gene *cdr-6* for causality of the chromosome V QTL. It ended up not being the causal gene and didn't really fit in the flow of the manuscript, however Erik and I decided that we still wanted to share the results with the scientific community. I plan to publish this short appendix as a micropublication in the fall of 2020.

C.2 RESULTS AND DISCUSSION

Regulation of the essential trace element zinc is necessary to avoid the toxic consequences caused by too little or too much of this metal [226, 228]. The zinc-response pathway has been extensively studied in the nematode roundworm *Caenorhabditis elegans* and several genes have been discovered that function to modulate sensitivity to both high and low zinc concentrations [233]. Recently, we identified a quantitative trait locus (QTL) on the center of chromosome V, indicating that natural genetic variation between the laboratory strain, N2, and a genetically divergent wild isolate from Hawaii, CB4856, contributes to differential responses to excess zinc [123].

The confidence interval for this QTL is 1.6 Mb and contains 629 genes (WS263). Of these genes, 113 have one or more genetic variants predicted to modify the amino-acid sequence of the protein. However, protein-coding variation is just one of the ways that genetic variation can cause phenotypic variation. Another is variation in gene expression, which is hypothesized to be important in the majority of complex traits [224]. We identified 83 expression QTL (eQTL) within this 1.6 Mb region using the eQTL dataset that mapped expression differences among a panel of RIAILs also derived from N2 and CB4856 [115, 69]. The most significant eQTL in this region caused a change in expression of the gene *cdr-6* (**Figure C-1A**). This gene, a homolog of the cadmium-response gene *cdr-1*, is downregulated after treatments with arsenic, cadmium, or zinc [344]. Furthermore, *cdr-1* was previously shown to

mitigate cadmium toxicity in *C. elegans* [345, 346]. Together, these data suggest that expression of *cdr-6* might be toxic to *C. elegans* in the presence of heavy metals. Additionally, RIAILs with the CB4856 allele on chromosome V naturally express higher levels of *cdr-6* (**Figure C-1B**) and are also more sensitive to zinc than RIAILs with the N2 allele [123]. This result indicates that strains with naturally low expression of *cdr-6* are more resistant to excess zinc than strains with naturally high expression of this gene.

To test this hypothesis, we used CRISPR-Cas9 genome editing to create strains with large deletions of *cdr-6* in both the N2 and CB4856 genetic backgrounds (**Figure C-1C,D**). Because expression of *cdr-6* was higher in RIAILs with the CB4856 allele (associated with zinc sensitivity) than in RIAILs with the N2 allele (associated with zinc resistance) (Figure A-4B), we expected that a knockout of *cdr-6* in the CB4856 genetic background would cause increased resistance to excess zinc. Alternatively, if variation in expression of *cdr-6* underlies the zinc-response QTL on chromosome V, a knockout of *cdr-6* in the N2 genetic background should not cause an increase in zinc resistance. We exposed N2, CB4856, and two strains with independently derived *cdr-6* deletion alleles in each genetic background to elevated zinc and measured their optical densities using a high-throughput assay with the COPAS BIOSORT [69, 123, 68]. We found that strains with a deletion of *cdr-6* phenocopied the strain with the same genetic background (**Figure C-1E**), suggesting that differences in expression of *cdr-6* do not underlie zinc responses.

This study not only provides evidence against *cdr-6* as the causal gene underlying differences in zinc resistance between the N2 and CB4856 strains, but also indicates that *cdr-6* does not influence zinc resistance in *C. elegans*. These results are not in disagreement with the previous study that showed that expression of *cdr-6* was downregulated in response to zinc [344]. In fact, the authors also showed that the accumulation of fluid-filled droplets in the pseudocoelom, a phenotype observed in strains with inhibited *cdr-6* expression using RNAi, is not increased in response to zinc or cadmium exposure [344]. Taken together, we conclude that expression of *cdr-6* decreases in response to zinc but animal development in the presence of zinc is not affected by *cdr-6* function. It is likely that *cdr-6* does not function in the nematode zinc response but rather is downregulated as an indirect effect of zinc exposure.

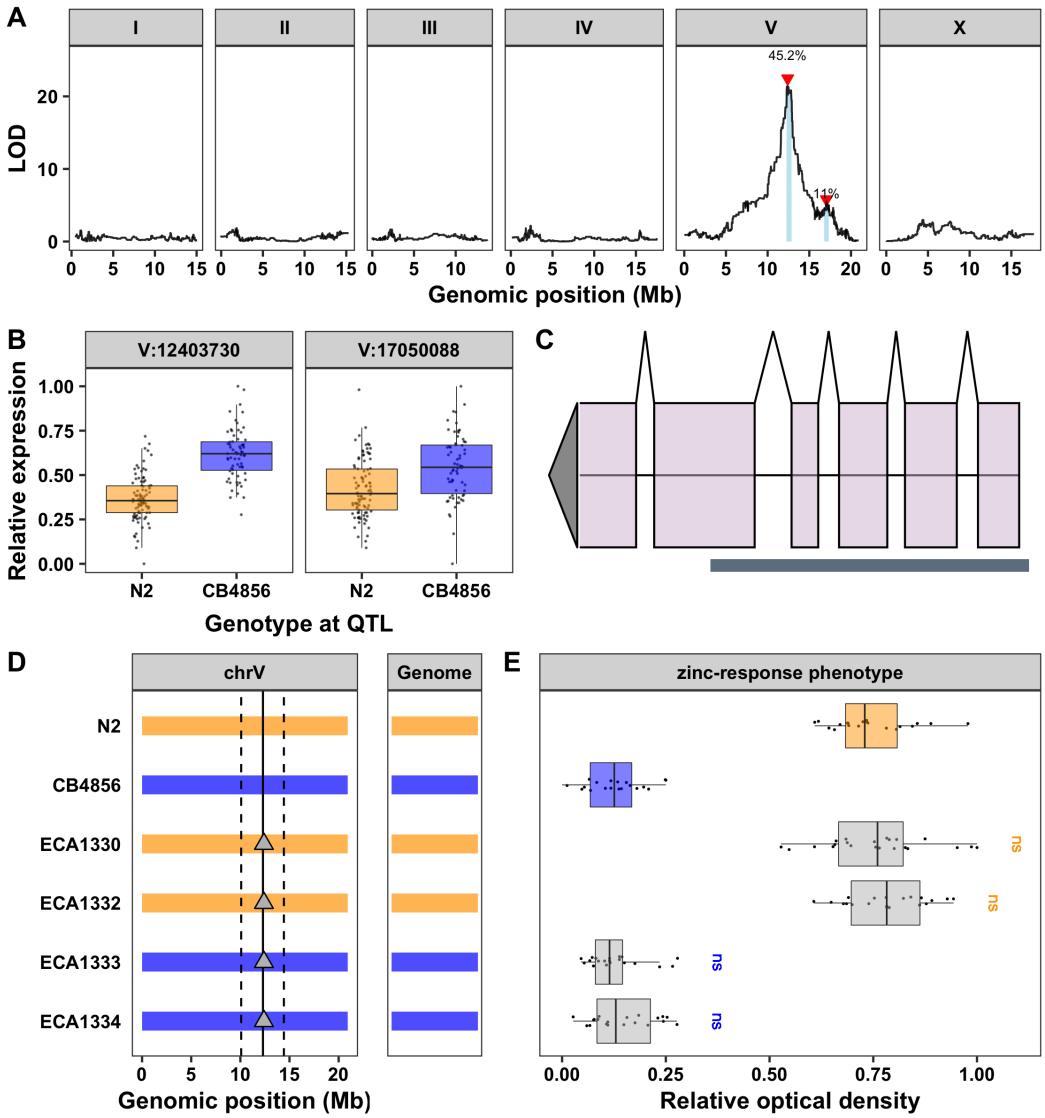


Figure C-1: Testing the role of *cdr-6* in the *C. elegans* zinc response. (A) Linkage mapping analysis using expression of *cdr-6* as a quantitative trait. Genomic position (x-axis) is plotted against the logarithm of the odds (LOD) score (y-axis) for 13,003 genomic markers. Each significant QTL is indicated by a red triangle at the peak marker, and a blue rectangle shows the 95% confidence interval around the peak marker. The percentage of the total variance in the RIAIL population that can be explained by each QTL is shown above the QTL. **(B)** For each QTL, the expression of *cdr-6* (y-axis) in RIAILs split by genotype at the marker with the maximum LOD score (x-axis) are plotted as Tukey box plots. Each point corresponds to the relative expression of a unique recombinant strain. **(C)** Gene model for *cdr-6* is shown. Exons are represented by light purple rectangles and introns are represented by connecting lines. Location of the CRISPR-Cas9 deletion is shown with the grey box below the gene model. **(D)** Strain genotypes are shown as colored rectangles (N2: orange, CB4856: blue) in detail for chromosome V (left) and in general for the rest of the chromosomes (right). The solid vertical line represents the peak marker of the zinc-response QTL, and the dashed vertical lines represent the confidence interval. Grey triangles represent *cdr-6* deletions. **(E)** Relative animal optical density in zinc (median.EXT, x-axis) is plotted as Tukey box plots against strain (see C, y-axis). Statistical significance of each strain compared to its parental strain (ECA1330 and ECA1332 to N2 and ECA1333 and ECA1334 to CB4856) is shown above each strain and colored by the parent strain it was tested against (ns = non-significant, *p*-value > 0.05; * = significant, *p*-value < 0.05).